

UNIVERSITÀ DEGLI STUDI DI CASSINO E DEL LAZIO MERIDIONALE
CORSO DI DOTTORATO IN METODI, MODELLI e TECNOLOGIE PER L'INGEGNERIA
DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL'INFORMAZIONE



Handwriting Analysis for the Development of a system to support the Diagnosis of Neurodegenerative Diseases

Tiziana D'Alessandro

tiziana.dalessandro@unicas.it

In Partial Fulfillment of the Requirements for the Degree of
PHILOSOPHIAE DOCTOR in
Electrical and Information Engineering

16/01/2024

TUTOR
Prof. Claudio De Stefano

COORDINATOR
Prof. Fabrizio Marignetti

UNIVERSITÀ DEGLI STUDI DI CASSINO E DEL LAZIO MERIDIONALE
CORSO DI DOTTORATO IN METODI, MODELLI E TECNOLOGIE PER
L'INGEGNERIA

Date: **16/01/2024**


Author: **Tiziana D'Alessandro**

Title: **Handwriting Features Analysis for the Development of a system to support
the Diagnosis of Neurodegenerative Diseases**

Department: **DIPARTIMENTO DI INGEGNERIA ELETTRICA E
DELL'INFORMAZIONE**

Degree: **PHILOSOPHIAE DOCTOR**

Permission is herewith granted to university to circulate and to have copied for non-commercial purposes,
at its discretion, the above title upon the request of individuals or institutions.


Signature of Author

THE AUTHOR RESERVES OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR
EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE
AUTHOR'S WRITTEN PERMISSION.

THE AUTHOR ATTESTS THAT PERMISSION HAS BEEN OBTAINED FOR THE USE OF ANY
COPYRIGHTED MATERIAL APPEARING IN THIS THESIS (OTHER THAN BRIEF EXCERPTS REQUIRING
ONLY PROPER ACKNOWLEDGEMENT IN SCHOLARLY WRITING) AND THAT ALL SUCH USE IS CLEARLY
ACKNOWLEDGED.

Dedica

A Sperso.

Alle nostre interminabili avventure.

Acknowledgements

The essence of research lies in its contribution to advancing human knowledge and addressing challenges by providing new information, diverse perspectives, and innovative solutions. My research centres explicitly on employing an artificial intelligence-based support system for Alzheimer's diagnosis. Over the years, I have developed a profound passion for this work, driven by the belief in its potential for improvement and the community benefits achievable through an accessible, non-invasive, and cost-effective screening tool. I emphasize the significance of preventive measures for neurodegenerative diseases within this context.

This thesis marks the culmination of three years characterized by dedication, curiosity, learning, and passionate exploration. I am deeply grateful to those who initiated me into this journey, guiding me through the highs and lows of research. My pride in this period extends beyond the production of scientific research to the profound lessons I've gained. Research, I've discovered, transcends mere study; it demands sharing, interaction, dedication, courage, and a boundless curiosity. Remarkably, throughout these years, I've never experienced boredom, and even in the face of unexpected challenges, I've found inspiration to persevere. After all, even negative results contribute valuable information.

I've been fortunate never to walk this journey alone, always having a reliable team by my side. My heartfelt gratitude goes to my supervisors, Professor Claudio De Stefano and Professor Francesco Fontanella, who consistently motivated and believed in me. They encouraged me to leave my comfort zone, trusting in me even when I was afraid to fail. I am also thankful to my entire work team and the professors of DIEI.

During my time abroad on the beautiful island of Gran Canaria, collaborating with the esteemed Professors of the University of Las Palmas de Gran Canaria, IDETIC, was a transformative experience. The warm days spent there provided a different perspective on my research, encouraging me to evaluate the pros and cons and fostering continuous improvement. Muchas Gracias to Professor Miguel Angel Ferrer, Professor Cristina Carmona-Duarte, and Professor Moisés Diaz.

While facing occasional challenges, I always found support from my family, especially my sisters and Andrea, who never stopped believing in me. I am grateful to all of them, including my little Giulio, who provided much-needed carefree moments.

Special appreciation goes to the friends from Stanza32, a second family where I always felt at home. I thank those friends who never gave up on me, contributing significantly to the richness of my journey.

Contents

Acknowledgements	iv
List of figures	vii
List of tables	ix
Acronyms	xi
1 Introduction	1
1.1 Aims and Motivation	2
1.2 Overview of Contents	3
1.3 Scientific Production	4
2 Theoretical Background	6
2.1 Neurodegenerative Diseases	6
2.1.1 Alzheimer’s Disease	7
2.2 Artificial Intelligence in Healthcare	10
2.3 The Role of Handwriting	11
2.4 Related Work	13
2.4.1 Handwriting Analysis	13
2.4.2 Speech Analysis	17
2.4.3 Gait Analysis	18
2.4.4 NeuroImaging Analysis	19
2.4.5 Other Techniques	19
3 Data	21
3.1 Data Acquisition	21
3.2 Image Generation	25
3.2.1 Binary Synthetic Images	26
3.2.2 RGB Synthetic Images	27
3.2.3 Multi-Channel	28
3.2.4 Offline	28
3.3 Features	30
3.3.1 Handcrafted Features	30
3.3.2 Lognormal Features	32

4	Results and Findings	36
4.1	Baseline Experimental Setting	36
4.2	Comparison among Binary, RGB on-paper and Handcrafted Features on Graphic Tasks	38
4.3	Comparison among RGB on-paper, Multichannel and Handcrafted Features on Graphic Tasks	45
4.4	Comparison among RGB on paper, Offline and Handcrafted Features on Writing Tasks	51
4.5	Exploiting Lognormal Features to Support the Diagnosis of AD through Handwriting Analysis	56
4.6	A Machine Learning Approach to Analyze the Effects of AD on Handwriting through Lognormal Features	59
4.7	Complementary Results	67
4.7.1	Comparison between Deep results achieved from Binary, RGB and Offline Images on Graphic Tasks	69
4.7.2	Comparison between Deep results achieved from Multi-Channel Images	70
4.7.3	Comparison between Deep and Machine Learning results achieved from RGB Images	70
5	Evolutionary Algorithms	72
5.1	Using Genetic Algorithms to Optimize a DL-based System for the Prediction of Cognitive Impairments	72
5.2	An evolutionary approach for Neural Architecture Search	74
5.3	Integrating Data Augmentation in Evolutionary Algorithms for Feature Selection	79
6	Conclusions and Future Work	83
	Bibliography	87

List of Figures

3.1	Example of generated strokes	26
3.2	Example of a trait generated for binary images.	26
3.3	Example of colour encoding for the traits generation.	28
3.4	Example of encoding for the trait generation in an MC image.	29
3.5	An example of lognormal.	33
4.1	Basic experimental setting.	36
4.2	Examples of RGB images generated from the online handwriting data acquired from a participant while performing the selected graphic tasks	38
4.3	Average accuracy achieved by the classifiers (a) and the CNNs (b) for features extracted by RGB images.	41
4.4	Accuracy for each task averaged over the results of five classifiers.	43
4.5	Comparison results between deep and handcrafted features.	43
4.6	Accuracy for each task averaged over the results of the five classifiers using merged features, compared with deep features.	44
4.7	Examples of tasks performed by a participant involved in the experiments.	46
4.8	Average accuracy achieved by the classifiers (a) and the CNNs (b) using RGB images.	49
4.9	Average accuracy achieved by the classifiers (a) and the CNNs (b) using MC images.	49
4.10	Accuracy for each task averaged over the results of five classifiers.	50
4.11	Comparison results between Deep-RGB and handcrafted features.	51
4.12	Example of offline images.	52
4.13	Average accuracy achieved by the classifiers (a) and the Convolutional Neural Networks (CNNs) (b) using offline images.	54
4.14	Comparison between accuracy achieved by the ML classifiers from RGB, offline and Handcrafted features.	55
4.15	Examples RGB in-air on-paper images generated from the execution of tasks by a participant involved in the experiments.	57
4.16	Comparison of ROC curves obtained from RF, SVM and VGG19 for every task.	60
4.17	Workflow Representation.	61
4.18	Averaged evaluation metrics achieved on 30 runs for every ML algorithm on lognormal features, extracted from the execution of task 23.	63
4.19	Comparison between average Accuracy achieved on 30 runs for every task with the best-performing ML algorithm for Dynamic and Lognormal features.	64
4.20	Box plots showing how the contact time feature is related to age and education. The first age range is from 44 to 66 years old, while the second is from 67 to 88 years old.	68
4.21	Experimental setting.	69
4.22	Experimental setting.	70
4.23	Experimental setting.	71
5.1	The layout of the proposed system. Note that in our case $n = 34$	73
5.2	Evolution of accuracy and average number of selected tasks for every model of CNN, considering weighted predictions.	74
5.3	Evolution of accuracy and average number of selected tasks for every model of CNN, considering non-weighted predictions.	74
5.4	Comparison of the number of occurrences of the selected tasks for the three CNN between the normal and the weighted approach.	75
5.5	Best individuals.	77

LIST OF FIGURES

5.6	MNIST results.	78
5.7	CIFAR-10 results.	78
5.8	Training and validation accuracy of the best individuals for each task	79
5.9	Experimental workflow.	80

List of Tables

3.1	Average demographic data of participants. Standard deviations are shown in parenthesis .	22
3.2	Protocol handwriting tasks	24
3.3	Feature list. Feature types are dynamic (D), static (S) and personal (P).	31
3.4	Summary of computed Lognormal Features, first set.	33
3.5	Temporal and SNR related features.	35
3.6	Geometrical features.	35
4.1	Number of parameters and input/output size of the CNN used in the experiments.	37
4.2	Values of the Machine Learning (ML) classifiers hyperparameters used in the experiments.	39
4.3	Classification results achieved on deep features extracted by binary images.	40
4.4	Classification results achieved with deep features extracted by RGB images.	41
4.5	Summary of the best experiments obtained on RGB images. The table also shows, for each task, the other considered performance measures	42
4.6	Results of classification with handcrafted features. Bold values highlight the overall best performance achieved on each task.	42
4.7	Classification results achieved using the weighted majority vote rule.	44
4.8	Classification results achieved using RGB features.	47
4.9	Classification results achieved using MC features. Bold values highlight the overall best performance achieved on each task.	48
4.10	Classification results achieved by the FC classifier, using RGB and MC features.	50
4.11	Results of classification with the handcrafted features. Bold values highlight the overall best performance achieved on each task.	51
4.12	ML Classifiers and their hyperparameters involved in the Grid search process	53
4.13	Performance comparison on all the feature sets considered.	54
4.14	Combining results.	56
4.15	Average Accuracy achieved on 30 runs for every ML algorithm on lognormal features . . .	57
4.16	Average Sensitivity achieved on 30 runs for every ML algorithm on lognormal features . .	58
4.17	Average Specificity achieved on 30 runs for every ML algorithm on lognormal features . .	58
4.18	Average Precision achieved on 30 runs for every ML algorithm on lognormal features . . .	58
4.19	Average FNR achieved on 30 runs for every ML algorithm on lognormal features	59
4.20	Average AUC achieved on 30 runs for every ML algorithm on lognormal features	59
4.21	Comparison results.	59
4.22	Average Accuracy achieved on 30 runs for every ML algorithm on lognormal features . . .	62
4.23	Average results achieved on 30 runs for every ML algorithm on lognormal features, extracted from the execution of task 23, i.e. the one that reached the best performance according to Table 4.22	63
4.24	Comparison between average Accuracy achieved on 30 runs for every task with the best-performing ML algorithm for Dynamic and Lognormal features	64
4.25	Stacking results averaged over thirty runs with XGB classifiers, with the output of first-step classifiers	65
4.26	Tasks ranking for each ML Classifier	66
4.27	Majority vote to a different set of ranked tasks.	66
4.28	Distribution of people according to Education level and Age.	67
4.29	Comparison among the evaluated deep approaches, considering binary, RGB and Offline images.	69
4.30	Results of applying a majority vote rule with reject.	69

4.31	Accuracy and FNR results from applying the majority vote rule on all the tasks or subsets of them.	70
4.32	Results of the majority vote rule with rejection.	71
5.1	The values of the parameters used for the GA.	73
5.2	Comparison of results obtained by applying the majority vote rule on a subset of tasks selected by the Genetic Algorithm (GA) or on all the tasks.	75
5.3	CNN layers and their hyperparameters.	76
5.4	Benchmark datasets.	77
5.5	Accuracy comparison among CNNs and ENAS	79
5.6	The datasets used in the experiments.	80
5.7	Baseline experiment results in average accuracy and standard deviation computed over 20 runs.	80
5.8	PSO parameters setting.	81
5.9	GA parameters setting.	81
5.10	Feature selection results in average and standard deviation accuracy and average number of selected features computed over 20 runs.	81
5.11	Data augmentation and feature selection results in average accuracy and standard deviation computed over 20 runs for every FS technique and DA percentage.	82
5.12	Comparison between baseline experiment and best performance achieved with FS and FS combined with DA.	82

Acronyms

AD	Alzheimer's Disease.
AI	Artificial Intelligence.
ALS	Amyotrophic Lateral Sclerosis.
ANN	Artificial Neural Network.
AUC	Area Under the Curve.
C	Classifier.
CAD	Computer Aided Diagnosis system.
CDT	Clock Drawing Test.
CNN	Convolutional Neural Network.
CT	Computed Tomography.
DL	Deep Learning.
DNN	Deep Neural Network.
DT	Decision Tree.
EC	Evolutionary Computation.
EEG	Electroencephalography.
FAB	Frontal Assessment Battery.
FC	Fully Connected.
FE	Feature Extractor.
FNR	False Negative Rate.
GA	Genetic Algorithm.
GB	Gradient Boosting.
KNN	K-Nearest Neighbors.
LR	Logistic Regression.
MC	Multi-Channel.
MCI	Mild Cognitive Impairment.
ML	Machine Learning.
MLP	Multilayer Perceptron.
MMSE	Mini-Mental State Examination.
MoCA	Montreal Cognitive Assessment.
MRI	Magnetic Resonance Imaging.
NAS	Neural Architecture Search.
ND	Neurodegenerative Disease.
PD	Parkinson's Disease.
PET	Positron Emission Tomography.
PSO	Particle Swarm Optimization.
RF	Random Forest.
RFE	Recursive Feature Elimination.
ROI	Region Of Interest.

SGD	Stochastic Gradient Descent.
SNR	Signal-to-Noise Ratio.
SPECT	Single-Photon Emission Computed Tomography.
SVM	Support Vector Machine.
TNR	True Negative Rate.
TPR	True Positive Rate.
XGB	Extreme Gradient Boosting.

Chapter 1

Introduction

The prevalence of neurodegenerative diseases has been steadily increasing in recent years, underscoring a concerning trend. A rising number of individuals are facing these debilitating conditions, reflecting the complex challenges posed by ageing populations and changing lifestyles. The growth in the incidence of neurodegenerative diseases highlights the pressing need for continued research, heightened awareness, and enhanced support systems to address the evolving healthcare landscape and provide better care for those affected by these conditions.

Another aspect underscoring the importance of research in this field is that neurodegenerative diseases currently lack a cure. They can cause cognitive impairments manifesting as difficulties in memory, language, thinking, judgment, and motor skills. Individuals exhibiting a combination of these symptoms face a significantly heightened risk of developing dementia and, in more severe cases, Alzheimer's Disease (AD) or Parkinson's Disease (PD).

Given the progressive nature of AD, early detection becomes crucial to initiate therapies to mitigate its effects. Early diagnosis is a fundamental prerequisite for the effectiveness of these treatments, aimed at slowing down disease progression. This early intervention not only helps extend the life expectancy of patients but also enhances their overall quality of life. Once the signs of the disease manifest, substantial and irreversible damage may have already occurred.

The impact of neurodegenerative diseases on handwriting is a notable concern, as these conditions can compromise fine motor control and cognitive functions. Individuals affected by neurodegenerative diseases often experience changes in their handwriting, such as altered penmanship, irregular letter shapes, and diminished overall legibility. This decline in handwriting proficiency can be studied as a tangible manifestation of the broader cognitive challenges associated with these diseases.

Diagnosing neurodegenerative diseases like AD involves a comprehensive physician assessment, using several sources of information and incorporating various tools and tests to evaluate cognitive function, neurological health, and overall well-being.

The involvement of Artificial Intelligence (AI) in supporting the diagnosis of neurodegenerative diseases has been a progressively evolving field over the past couple of decades. The application of AI techniques gained momentum in the 21st century with advances in computational power, the availability of data, and improvements in algorithmic approaches. Since the early to mid-2000s, researchers began exploring the potential of AI in analyzing various data types associated with neurodegenerative diseases, including medical images, genetic information, and clinical data. The use of AI in supporting the diagnosis of Neurodegenerative Diseases (NDs) represents a promising frontier in healthcare. AI applications, such as machine learning algorithms and deep learning models, analyze vast datasets to identify patterns and indicators associated with neurodegenerative conditions. These technologies offer the potential for earlier and more accurate detection of NDs, facilitating timely intervention and personalized treatment plans.

In the past ten years, the research community agreed that the application of artificial intelligence to handwriting analysis holds great potential for supporting the diagnosis of NDs. AI algorithms can discern subtle changes in handwriting patterns, offering valuable insights into cognitive decline associated with conditions like AD or PD. By analyzing features such as pressure, speed, and stroke dynamics, AI may contribute to the early detection and monitoring of NDs, providing a non-invasive and cost-effective diagnostic tool. This approach enables a better understanding of neurological changes and complements traditional diagnostic methods. Ongoing research in AI-driven handwriting analysis underscores its promise in enhancing the accuracy and efficiency of NDs diagnosis.

1.1 Aims and Motivation

The diagnosis of NDs is assessed by physicians and experts in the medical field. They gather information about the patient's medical history, including family history, lifestyle factors, and the onset of symptoms. This helps in understanding potential risk factors and the progression of cognitive decline. Cognitive tests commonly assess memory, language, and other cognitive functions. A thorough physical examination helps identify any underlying health issues, and a neurological exam assesses functions like reflexes, coordination, and sensory abilities. Blood tests are often conducted to rule out other conditions that may mimic symptoms of neurodegenerative diseases. Advanced imaging techniques provide detailed images of the brain, evaluated by physicians to detect anomalies. Occasionally, cerebrospinal fluid analysis may be performed to detect specific biomarkers associated with AD.

While diagnostic tests for NDs play a crucial role in early detection and management, they also have certain limitations and considerations, especially for impaired individuals, they can present some potential drawbacks. Some diagnostic tests, such as advanced brain imaging, can be expensive. The financial burden may be a concern, especially for individuals without comprehensive insurance coverage. The invasiveness is another aspect to consider, as certain tests, like lumbar punctures for cerebrospinal fluid analysis, can be invasive and uncomfortable. The procedure involves inserting a needle into the spinal canal to collect fluid. Invasive tests may pose challenges for individuals with cognitive impairment, as they may find it distressing or be unable to cooperate. Cognitive tests and neurological exams, though generally non-invasive, may cause discomfort or anxiety for individuals with cognitive impairment. These individuals may find it challenging to follow instructions or become agitated during testing. Some individuals may have difficulty cooperating during cognitive tests or imaging procedures, potentially affecting the accuracy of the results. This limitation underscores the need for alternative, patient-friendly diagnostic approaches.

Moreover, access to certain diagnostic tests may be limited based on geographical location, healthcare infrastructure, or financial constraints. This can impact the timely diagnosis and intervention for individuals with cognitive impairment. While cerebrospinal fluid analysis for biomarkers is a valuable tool, it is not as widely available as other diagnostic tests. Limited accessibility may hinder its use in routine clinical practice.

It's essential for healthcare providers to carefully consider the individual needs, preferences, and limitations of each patient, particularly those with cognitive impairment. Additionally, ongoing research and technological advancements aim to address some of these limitations and improve the accessibility and accuracy of diagnostic tools for NDs. AI applications may assist in analyzing and interpreting cognitive test results more efficiently. Recognizing the importance of an early diagnosis, it is widely acknowledged in the medical community that handwriting is among the first skills impacted by the onset of cognitive disorders. This is because cognitive diseases can impact motor activities, including handwriting, which relies on cognitive, kinaesthetic, and perceptive motor skills. The intricate connection between cognitive decline and changes in handwriting serves as a valuable early indicator, emphasizing the importance of timely diagnosis to initiate interventions that can potentially alleviate the impact of these debilitating conditions.

Recent research has highlighted specific aspects of the writing process that exhibit greater vulnerability and could serve as diagnostic indicators in signal and image processing. For instance, dysgraphia, a writing impairment, manifests early in the course of AD, even during its initial phase. Consequently, over the past two decades, researchers have devised writing tests to examine the progression of cognitive impairment more effectively, aiming to establish a means for early predictions. It's worth noting that most studies exploring the impact of cognitive impairment and neurodegenerative disorders on handwriting kinematics have been conducted within the medical field, utilizing statistical tools for analysis.

Recognizing the significance of early diagnosis in the medical domain, researchers have turned to AI techniques to identify potential features that can enhance the recognition mechanism across a broad spectrum of illnesses. Diseases such as AD and other cognitive impairments fall within this category. While the traditional diagnostic process relies on medical expertise, advancements in technology, including AI, are playing an increasing role in aiding diagnosis. AI is increasingly being explored for its potential in ND diagnostics. Machine learning algorithms can analyze large datasets, including medical records, imaging data, and genetic information, to identify patterns and predict disease risk or progression. AI applications, including deep learning models, are being developed to assist in interpreting medical imaging, such as identifying subtle changes in brain structures indicative of neurodegenerative diseases. AI-powered decision support systems may aid physicians by providing additional insights based on a comprehensive

analysis of diverse data sources. While AI holds promise in supporting the diagnostic process, it is crucial to note that the final diagnosis is typically made by experienced healthcare professionals who consider a wide view of the patient’s clinical presentation. AI tools are valuable aids, assisting healthcare providers in more efficient and accurate decision-making.

Handwriting analysis holds promising potential as a non-invasive and accessible tool for developing a system to support the diagnosis of neurodegenerative diseases using AI. By examining subtle changes in handwriting patterns, this innovative approach seeks to detect early signs of cognitive decline associated with diseases such as AD and PD. AI algorithms play a crucial role in processing vast datasets of handwritten samples, extracting intricate features related to motor control, velocity, and pressure. The goal is to create a sophisticated diagnostic system to discern variations indicative of neurodegenerative disorders. Integrating AI-driven handwriting analysis into diagnostic protocols offers a patient-friendly alternative, particularly for those facing challenges with traditional testing methods. This approach can potentially enhance early detection, paving the way for timely interventions and improved patient outcomes. Ongoing research in this field explores the correlation between handwriting alterations and neurodegenerative changes, contributing to the advancement of precision medicine in diagnosing and managing these complex conditions. The last two decades of research have demonstrated that applying AI techniques to handwriting analysis can yield positive outcomes.

This thesis aims to leverage AI to develop a Computer Aided Diagnosis system (CAD) specifically designed to support Alzheimer’s diagnosis. There are several motivations behind this goal. First, a timely diagnosis is crucial for effective intervention in AD. AI-powered CAD systems have the potential to detect subtle cognitive changes and can analyze complex patterns and data, providing a more accurate and objective assessment of cognitive decline based on various inputs. This has the potential to enhance diagnostic precision compared to traditional methods. Moreover, using AI can ease the diagnostic process, reducing the time and resources required. This efficiency is particularly important given the increasing prevalence of neurodegenerative diseases and the need for scalable and accessible diagnostic solutions. AI systems also offer an objective and standardized evaluation of biomarkers, reducing the potential for subjective interpretation.

Finally, insights gained from the system can potentially lead to new research directions, therapeutic targets, and a deeper comprehension of the disease’s progression. This thesis’s contributions include evaluating AD diagnosis systems based on dynamic and shape information and comparing the results achieved through different experimental settings. The study assessed CNNs’ ability as automatic feature extractors, tested long-term motor planning ability through new tasks, and compared different classification approaches. The findings shed light on the effectiveness of combining shape and dynamic information for AD diagnosis using machine learning techniques.

The final objective is to provide cost-effective, non-invasive, and easily accessible support for AD diagnosis by integrating AI with handwriting analysis. This innovative approach seeks to harness the unique patterns in handwriting to detect early signs of cognitive decline associated with AD. By utilizing AI algorithms, the system aims to process handwriting samples efficiently, extracting subtle features related to motor control and cognitive function. The emphasis on low cost and non-invasiveness ensures broader accessibility, particularly for individuals facing challenges with conventional diagnostic methods. The goal is to create a user-friendly and widely applicable tool to facilitate timely AD detection, fostering better patient outcomes. Ongoing research in this field underscores the potential for AI-driven handwriting analysis to revolutionize the diagnostic landscape for neurodegenerative diseases, offering an accessible and scalable solution for widespread use and implementation.

1.2 Overview of Contents

In this context, preliminary studies explored the combined use of shape and dynamic features extracted from handwriting for AD diagnosis. In this thesis, an experimental handwriting task protocol was considered for the data acquisition step, exploiting the ability of the Wacom Bamboo Folio graphic tablet to record essential information. After the execution of the protocol, many types of data were achieved from handwriting tasks, namely static and dynamic features directly derived from the sequences of points provided by the tablet and the handwriting images recorded or generated at the end of each task.

Features were processed and used for ML classification algorithms to distinguish between healthy people and patients. Concerning images, several variants have been investigated. Deep learning techniques were considered, and a set of CNNs was empirically selected to extract features automatically.

Promising results were achieved, but challenges remained in distinguishing healthy subjects from those

with AD in the early stages. To address this point, the study expanded, incorporating more complex tasks demanding higher fine motor control and more significant cognitive load. The choice of unfamiliar graphic tasks aimed to accentuate differences in writing characteristics between healthy subjects and those with NDs. Various classification schemes were employed to assess the performance of different feature representations.

The remainder of the thesis is organized as follows. Chapter 1 introduces the objective of the work and the issue it means to solve. Many concepts addressed in this part will be deeply explained in the remainder of this work. Chapter 2 describes the NDs with a particular emphasis on AD and resumes the last advancements in the employment of AI in healthcare. It also contains a section dedicated to the handwriting process. The second part of this chapter consists of a rich selection of research involved in the diagnosis or monitoring of NDs from different information sources and technological approaches. It provides an overview of the literature regarding the examined topic. Chapter 3 is designed to describe the data acquisition step followed by an image generation and a feature computation phase. Chapter 4 discusses the experimental architectures, results and findings, while Chapter 5 introduce an additional research path involving the evolutionary theory. Finally, the last Chapter 6 is devoted to a final discussion and considering possible future works.

1.3 Scientific Production

The contents addressed in this thesis focus on the following publications.

Journal papers

- N. D. Cilia, T. D'Alessandro, C. De Stefano, F. Fontanella and M. Molinara.
From Online Handwriting to Synthetic Images for Alzheimer's Disease Detection Using a Deep Transfer Learning Approach
in IEEE Journal of Biomedical and Health Informatics, Dec. 2021, vol. 25, no. 12, pp. 4243-4254.
- N. D. Cilia, T. D'Alessandro, C. De Stefano and F. Fontanella.
Deep transfer learning algorithms applied to synthetic drawing images as a tool for supporting Alzheimer's disease prediction
Mach. Vision Appl, May 2022, vol. 33, no. 3.
- N. D. Cilia, T. D'Alessandro, C. De Stefano, F. Fontanella, and A. Scotto di Freca.
Comparing filter and wrapper approaches for feature selection in handwritten character recognition
Pattern Recognition Letters, 2023, vol. 168, pp. 39-46.

Conference Proceeding

- N.D. Cilia, T. D'Alessandro, C. De Stefano, F. Fontanella, A. Scotto di Freca.
Using Genetic Algorithms to Optimize a Deep Learning Based System for the Prediction of Cognitive Impairments
International Workshop on Artificial Life and Evolutionary Computation (WIVACE), 2021, vol. 1722.
- N.D. Cilia, T. D'Alessandro, C. De Stefano and F. Fontanella.
Offline handwriting image analysis to predict Alzheimer's disease via deep learning
International Conference on Pattern Recognition (ICPR), 2022, pp. 2807-2813.
- N.D. Cilia, C. Carmona-Duarte, T. D'Alessandro, C. De Stefano, M. Diaz, M.A. Ferrer, F. Fontanella.
Lognormal Features for Early Diagnosis of Alzheimer's Disease Through Handwriting Analysis
Intertwining Graphonomics with Human Movements (IGS), 2022, vol. 13424.
In collaboration with University of Las Palmas de Gran Canaria
- F. Fontanella, S. Pinelli, C. Babiloni, R. Lizio, C. Del Percio, S. Lopez, G. Noce, F. Giubilei, F. Stocchi, G.B. Frisoni, F. Nobili, R. Ferri, T. D'Alessandro, N.D. Cilia, and C. De Stefano.
Machine Learning to Predict Cognitive Decline of Patients with Alzheimer's Disease

Using EEG Markers: A Preliminary Study

International Conference on Image Analysis and Processing (ICIAP), 2022, vol. 13231.

- A. Bria, P. De Ciccio, T. D'Alessandro, F. Fontanella.
A Novel Evolutionary Approach for Neural Architecture Search
International Workshop on Artificial Life and Evolutionary Computation (WIVACE), 2022, vol. 1780.
- T. D'Alessandro, C. De Stefano, F. Fontanella, E. Nardone, A. Scotto di Freca.
Feature Evaluation in Handwriting Analysis for Alzheimer's Disease using Bayesian Network
International conference of the International Graphonomics Society (IGS), 2023, vol. 14285.
- T. D'Alessandro, C. Carmona-Duarte, C. De Stefano, M. Diaz, M.A. Ferrer, F. Fontanella.
A Machine Learning Approach to Analyze the Effects of Alzheimer's Disease on Handwriting Through Lognormal Features
International conference of the International Graphonomics Society (IGS), 2023, vol. 14285.
In collaboration with University of Las Palmas de Gran Canaria

Unpublished research

- N.D. Cilia, T. D'Alessandro, C. De Stefano, F. Fontanella, I. Marthot-Santaniello, M. Molinara, A. Scotto Di Freca.
A Novel Writer Identification Approach for Greek Papyri Images
Workshop International Workshop on Pattern Recognition for Cultural Heritage (PatReCH), 2023.
In collaboration with University of Basel
- T. D'Alessandro, C. De Stefano, F. Fontanella, E. Nardone.
Integrating Data Augmentation in Evolutionary Algorithms for Feature Selection: A Preliminary Study
Evostar 2024.

Chapter 2

Theoretical Background

Neurodegenerative diseases are characterized by the progressive degeneration of the structure and function of the nervous system, leading to cognitive and motor impairment over time. In this context, AD is the most common, predominantly affecting older individuals. Most of these conditions lack a cure, so timely diagnosis is crucial to start treatments as soon as possible, aiming at slowing down the effect of the onset of such diseases, prompting the search for effective automated methods.

Numerous research efforts propose innovative approaches to support the diagnosis of NDs. The research on using automated methods, particularly ML and Deep Learning (DL), to support the diagnosis of NDs began gaining significant traction in the last couple of decades. The importance of this research lies in the potential for early and accurate detection of these diseases, which can significantly impact patient outcomes and the development of effective treatment strategies. In detail, the exploration of automated methods for NDs diagnosis started to emerge in the early 2000s. Researchers began to recognize the potential of computational approaches in analyzing complex patterns and biomarkers associated with these diseases. From the mid to late 2000s, a growing emphasis was on leveraging machine learning techniques, such as Support Vector Machine (SVM) and Artificial Neural Networks (ANNs), to analyze medical data, including neuroimaging and clinical information. Researchers aimed to identify distinctive patterns that could aid in early diagnosis.

Regarding the use of DL, it gained prominence in the 2010s. DL, with its ability to automatically learn hierarchical representations from data, became a powerful tool for analyzing large and complex datasets, including those related to neurodegenerative diseases. Nowadays, it is generally agreed in the research community that automated methods allow for a more personalized and precise approach to diagnosis, considering individual variations and patterns in patient data.

NDs often present complex and subtle patterns that may not be readily apparent through traditional diagnostic methods. ML and DL offer the capability to discern intricate relationships within large datasets. Automated techniques contribute to identifying biomarkers associated with these diseases, understanding their underlying mechanisms and facilitating the development of targeted treatments. In summary, automated methods for supporting the diagnosis of NDs started gaining momentum in the early 2000s and have continued to evolve, offering promising avenues for early detection, personalized medicine, and a deeper understanding of these complex conditions.

However, advancing more efficient learning techniques requires a comprehensive understanding of existing work in the field. This Chapter is devoted to describing and analysing the ongoing research in the context of automated methods to support the diagnosis of NDs. The subsequent sections provide a comprehensive overview of this subject, drawing insights from diverse information sources and exploring various facets, including feature extraction and automated methodologies. The analysis deeply examines these techniques, offering valuable perspectives on future directions.

2.1 Neurodegenerative Diseases

Neurological disorders are recognized to be the primary cause of disability and the second leading cause of death on a global scale [51]. Over the last decades, there has been a significant increase in the number of individuals facing disabilities due to neurological conditions [37]. This surge is particularly relevant in low-income and middle-income countries, and it is expected to increase globally due to population growth and ageing.

According to the Global Status Report, it is estimated that around 55 million individuals worldwide were

impacted by dementia in 2019. Projections from the same report indicate a substantial increase, with an anticipated rise to 139 million people by the year 2050. This alarming trend underscores the pressing need for advancements in research, care, and support to address the growing challenges posed by these conditions.

NDs specifically refer to a subset of neurological disorders commonly characterized by the progressive degeneration of the structure and function of the nervous system, typically involving the gradual loss of neurons in specific areas of the brain or spinal cord.

Although the precise cause remains elusive, several prevalent risk factors have been discerned [3]. It's crucial to emphasize that these factors can differ based on the particular disease. They are usually associated with ageing as the primary risk factor, but it's not the only one. Family history and genetic factors play a role in neurodegeneration, as some genetic mutations or variations can increase susceptibility. Exposure to certain environmental toxins and pollutants may contribute to the development of NDs. This includes exposure to heavy metals, pesticides, and other environmental toxins, so a dangerous environment raises the risk. Head trauma and some cardiovascular problems have been associated with the development of neurodegeneration. Other causes are related to an unhealthy lifestyle, hormonal changes, viral infections and even sleep disorders. Nevertheless, it is important to note that research has identified oxidative stress and inflammation as the two major contributors to neurodegeneration [124, 128].

Biomedical research has revealed diseases belonging to this realm share striking similarities at the sub-cellular level, including the formation of atypical protein assemblies (proteinopathy) and induced cell death. The compromised integrity of neurons, leading to their death, is a significant contributor to the progression of neurodegenerative conditions.

In [140], the authors describe some key, recurring features observed in various forms of these diseases in addition to pathological protein aggregation, neuron death and aberrant proteostasis. They are related to dysfunctions in the synaptic and neuronal network, cytoskeleton abnormalities, altered energy homeostasis and DNA and RNA defects. By targeting multiple hallmarks simultaneously, personalized therapies can be developed to halt or slow down neurodegenerative disease progression effectively. These shared characteristics suggest that progress in therapies for one neurodegenerative disease could potentially yield positive effects for other diseases within this category. Among the most known are amyotrophic lateral sclerosis, multiple sclerosis, Parkinson's disease, Alzheimer's disease, Huntington's disease, multiple system atrophy, tauopathies, and prion diseases.

NDs are generally considered irreversible, meaning that the damage to the nervous system and the loss of neurons that occur during the progression of these diseases cannot be fully reversed or restored to normal function. This irreversibility is a significant challenge in the field and limits current therapeutic interventions. Currently, there's no resolute cure, but only treatments aiming to slow down the progression of these diseases, manage symptoms, and improve the quality of life for affected individuals and their families and caregivers.

While a complete reversal of the damage may be difficult, there is hope that early detection and intervention strategies may help delay the onset of severe symptoms and provide a better quality of life for individuals living with NDs. Additionally, emerging technologies, such as regenerative medicine and gene therapies, hold promise in exploring more advanced treatment approaches.

2.1.1 Alzheimer's Disease

AD is recognized as one of the most prevalent neurodegenerative disorders, characterized by progressive cognitive decline and is estimated to be the cause of 60 – 70% of dementia cases. Globally, it is estimated that a minimum of 55 million individuals are currently living with Alzheimer's disease or other forms of dementia. Without significant breakthroughs, this number is poised to nearly double every 20 years, projected to reach 78 million in 2030 and a staggering 152 million in 2050 [10]. A poignant statistic reveals that one out of every three seniors ultimately succumbs to Alzheimer's or a related form of dementia. This underscores these conditions' pervasive and profound impact on the ageing population.

This disorder is named after Dr. Alois Alzheimer, a German psychiatrist who first identified the distinctive brain abnormalities associated with the disease in the early years of the XX century [129]. After this discovery, in 1984, the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association proposed the NINCDS-ADRDA Alzheimer's Criteria for diagnosis, then widely revised in 2007. According to these criteria, a clinical diagnosis of AD has to confirm the existence of cognitive impairment and a suspected dementia syndrome through neuropsychological testing. A conclusive diagnosis, however, requires histopathologic confirmation in-

volving a microscopic examination of brain tissue. The diagnostic criteria demonstrate strong statistical reliability and validity, aligning well with definitive histopathological confirmation. It's crucial to note that a definitive diagnosis of Alzheimer's disease can only be established post-mortem, emphasizing the need for a thorough examination of brain tissue after death.

Dementia is one of the most economically burdensome conditions for societies globally [64]. As populations age, the associated costs are anticipated to rise, posing a significant social and economic challenge. Expenses related to AD encompass both direct and indirect medical costs, with variations observed among countries based on the level of social care provided. Direct costs encompass expenditures on doctor visits, hospital care, medical treatments, nursing home services, specialized equipment, and household expenses. Indirect costs encompass the financial impact of informal caregiving and the productivity loss experienced by informal caregivers.

AD is thought to develop when abnormal accumulations of amyloid beta, forming as amyloid plaques externally and as tau proteins internally in the brain, disrupt neuronal functioning and connectivity [132]. This disruption leads to a progressive loss of brain function. The altered ability to clear these proteins is age-related and influenced by brain cholesterol, and it is also linked to other neurodegenerative diseases. The cause of most Alzheimer's cases remains largely unknown and not entirely understood, except for deterministic genetic differences identified in only 1 – 2% of cases [85]. Concerning this, while most Alzheimer's cases are not directly inherited, genetic factors can play a role. Mutations in specific genes, such as the Amyloid Precursor Protein gene on chromosome 21, can increase the risk of developing Alzheimer's. The APOE gene, especially the APOE4 variant, is a well-established genetic risk factor for late-onset Alzheimer's. The cholinergic hypothesis [10] suggests that a deficiency in the neurotransmitter acetylcholine may contribute to AD. Acetylcholine is involved in memory and learning, and its levels are reduced in the brains of individuals with Alzheimer's. Chronic inflammation in the brain and dysregulation of the immune system have been implicated in Alzheimer's disease. Another risk is related to conditions that affect the cardiovascular system, such as high blood pressure, diabetes, and high cholesterol. Many other risk factors for developing this disease are the same as described for the NDs in Section 2.1, like advanced age, environmental and lifestyle factors. It's important to note that AD likely results from a combination of genetic, environmental, and lifestyle factors, and ongoing research aims to unravel the intricate interplay among these elements.

This degenerative disorder primarily affects brain regions responsible for thought, memory, and language and disrupts various cognitive functions, leading to thinking, reasoning, and language challenges. Early symptoms are often mistaken for brain ageing or stress, but as the disease advances, it significantly impairs a person's capacity to perform routine daily activities. The primary and most evident manifestation is short-term memory loss, characterized by difficulty recalling recently acquired information and an inability to grasp new facts. The early phases of Alzheimer's disease may also manifest subtle challenges in executive functions such as attentiveness, planning, flexibility, and abstract thinking. Additionally, impairments in semantic memory involving the recall of meanings and conceptual relationships can become apparent. Apathy and depression often surface during this stage, with apathy persisting as the most enduring symptom throughout the disease progression.

Mild Cognitive Impairment (MCI) frequently serves as a transitional phase between typical ageing and dementia [109]. MCI can manifest with various symptoms, and when memory loss takes precedence, it is labelled as amnesic MCI. This form is frequently recognized as a prodromal stage of AD, with a greater than 90% likelihood of being associated with AD.

Individuals with Alzheimer's may struggle to recall information, navigate familiar spaces, and express themselves verbally. As the disease gradually erodes the individual's cognitive and functional independence, comprehensive care strategies are required, including supportive environments, adapted living spaces, and person-centred care approaches. AD patients often face challenges in maintaining a sense of purpose and dignity and performing basic daily living, such as dressing, bathing, grooming, and preparing meals. The disease's progressive nature can result in the need for assistance with these tasks.

AD progresses through distinct stages, each characterized by specific symptoms and impairment levels, though the disease's progression can vary from person to person. The most common symptoms of the early stage refer to minor events of memory loss and challenges in executive functions and language. People in this stage can still perform routine daily activities independently, though social skills may start to decline. Regarding the emotional aspects, apathy may be persistent and can be registered in mild personality changes and depression phenomena.

During the middle stage, the memory decline becomes more important, affecting both short-term and long-term memory. Communication becomes more unstable with a decrement in vocabulary and word

fluency. Issues with spatial orientation and perception may arise. Behavioural changes become more debilitating and noticeable, including agitation and irritability, wandering and hallucinations. Typically, individuals in this stage require assistance as their motor skills are impaired.

The late stage is the phase where the symptoms become unbearable, consisting of severe memory loss, communication breakdown and a profound impairment in all cognitive functions. Individuals become increasingly dependent on others for all aspects of care and emotionally unstable, also manifesting aggressive behaviour. They become unable to perform basic tasks requiring continued assistance. Mobility and motor skills decline significantly, and swallowing difficulties may also appear. Caregivers and healthcare professionals need to provide appropriate support and address the unique challenges associated with each stage of the disease. Additionally, providing emotional support to both individuals with AD and their caregivers is crucial in navigating the complexities associated with the progressive nature of the disease.

Currently, there is no cure for AD, and available treatments provide relatively small symptomatic benefits while maintaining a palliative nature. These treatments fall into three categories: pharmaceutical, psychosocial, and caregiving.

Concerning the pharmaceutical group, medications used to address cognitive symptoms include acetylcholinesterase inhibitors (tacrine, rivastigmine, galantamine, and donepezil) and memantine, an NMDA receptor antagonist [6]. Acetylcholinesterase inhibitors are designed for mild to severe Alzheimer's, while memantine is intended for moderate to severe cases [127]. However, the benefits from these medications are modest. Acetylcholinesterase inhibitors aim to slow down the breakdown of acetylcholine in the brain, counteracting its loss due to the death of cholinergic neurons. Memantine acts on the glutamatergic system to prevent excitotoxicity. Ginkgo biloba extract and atypical antipsychotics are also explored for their potential benefits [74].

Psychosocial interventions complement pharmaceutical treatment and include behaviour-, emotion-, cognition-, or stimulation-oriented approaches. Behavioural interventions focus on identifying and reducing antecedents and consequences of problem behaviours. Emotion-oriented interventions involve reminiscence therapy and simulated presence therapy [126]. Cognition-oriented treatments, such as reality orientation and cognitive retraining, aim to reduce cognitive deficits. Stimulation-oriented treatments encompass art, music, pet therapies, and exercise. Regarding the last group, as Alzheimer's disease has no cure and gradually hinders individuals' ability to care for themselves, caregiving becomes a crucial aspect of treatment. During the early and moderate stages, living environment and lifestyle modifications can enhance safety and reduce caregiver burden. In the final stages, treatment focuses on relieving discomfort until death, often with the assistance of hospice care.

A conclusive diagnosis of AD is typically established through autopsy findings; in the absence of autopsy, clinical diagnoses are designated as "possible" or "probable" based on other indicators. As many as 23% of individuals clinically diagnosed with AD may receive a misdiagnosis, with pathology indicative of a different condition sharing symptoms similar to those of AD. Clinical diagnosis of AD commonly relies on the person's medical history, information from relatives, and behavioural observations. The presence of distinctive neurological and neuropsychological features and the absence of alternative conditions support the diagnosis. Advanced medical imaging, such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Single-Photon Emission Computed Tomography (SPECT), or Positron Emission Tomography (PET), aids in ruling out other cerebral pathology or dementia subtypes. These imaging techniques can also predict the progression from prodromal stages (MCI) to full-fledged AD.

Assessing intellectual functioning, including memory testing, further contributes to characterizing the disease's status. Medical organisations have developed diagnostic criteria to facilitate and standardize the diagnostic process for physicians. Nevertheless, a definitive diagnosis can only be confirmed through post-mortem evaluations involving examination of brain material for senile plaques and neurofibrillary tangles.

Neuropsychological assessments are useful in diagnosing cognitive impairments like AD. Essential components of this diagnostic process include widely employed cognitive tests like the Mini-Mental State Examination (MMSE) [54], the Montreal Cognitive Assessment (MoCA) [95] and the Frontal Assessment Battery (FAB) [70]. Notably, these assessments may have limitations in accuracy, particularly regarding sensitivity to mild cognitive impairment and vulnerability to biases stemming from language or attention issues. It is crucial to underscore that certain tests, such as those evaluating handwriting alterations recognized by early AD researchers like Alois Alzheimer, play a distinctive role. Recognizing the nuances of these assessments is imperative to enhance reliability, especially in the early stages of the disease.

In addition, complementary neurological examinations are crucial for distinguishing AD from other con-

ditions. Family interviews provide valuable insights into daily living abilities and changes in mental function. The caregiver’s perspective is particularly significant, given that individuals with Alzheimer’s disease often lack awareness of their deficits. Families, however, may struggle to identify initial dementia symptoms and may not accurately convey information to physicians.

Supplemental testing serves to exclude other potentially treatable diagnoses and prevent misdiagnoses. Blood tests, thyroid function tests, assessments of vitamin B12 levels, screenings for neurosyphilis, and evaluations for metabolic problems (including kidney function, electrolyte levels, and diabetes) are commonly employed. Imaging techniques such as MRI or CT scans can help rule out alternative causes such as tumours or strokes. Psychological tests for depression are integral to the diagnostic process. Depression can coexist with AD, serve as an early sign of cognitive impairment, or even be the underlying cause. Recognizing and addressing psychological factors is essential to comprehensively understand the individual’s cognitive health.

Early diagnosis becomes crucial for implementing interventions and support strategies that can enhance the quality of life for affected individuals. Ongoing research aims to uncover effective treatments and preventative measures, recognizing the growing impact of Alzheimer’s in an ageing global population. The multifaceted nature of the disease underscores the need for holistic approaches that encompass medical, social, and psychological dimensions to address the complex challenges posed by Alzheimer’s. Caregiver support and public awareness are pivotal in fostering understanding, empathy, and a supportive environment for those affected by this debilitating condition. Currently, as for the other NDs described in Section 2.1, there isn’t a cure for AD.

2.2 Artificial Intelligence in Healthcare

AI is fundamentally reshaping the healthcare landscape, offering a wide set of advantages that significantly enhance patient care and the efficiency of healthcare systems.

It employs intricate algorithms and software to replicate human cognition for analyzing complex medical data or inferring diagnoses by observing specific aspects of a subject. AI’s ability to analyze vast datasets enables the early detection of diseases, leading to timely interventions and improved patient outcomes. Through the personalized analysis of patient data, including genetic information and lifestyle factors, AI tailors treatment plans, increasing effectiveness. Nevertheless, the escalating volume of clinical data and health records poses a significant challenge for healthcare professionals, and AI has emerged as a valuable ally in managing this burgeoning complexity. Its capacity to process, analyze, and derive meaningful insights from vast datasets solves clinicians’ information overload.

By leveraging AI-driven algorithms, healthcare providers can sift through extensive electronic health records and clinical data more efficiently. This expedites decision-making processes and allows clinicians to focus on patient care rather than being overwhelmed by the sheer volume of information. Furthermore, AI contributes to organising and structuring diverse data types, facilitating interoperability between different systems. This interoperability is crucial for creating a comprehensive and cohesive view of a patient’s medical history, enabling more informed diagnoses and personalized treatment plans.

AI techniques are applied across various domains, such as diagnosis, treatment recommendations, drug development, patient monitoring, and administrative tasks. Numerous research studies demonstrate AI’s effectiveness in medicine, often performing on par with or outperforming humans in certain tasks. Its role is complementary, enhancing diagnosis accuracy rather than replacing medical expertise. Recent research has shown substantial progress in diagnosing various illnesses, especially in radiology, imaging, psychiatry, and disease diagnosis.

Data-driven techniques lead to many advantages in the medical-scientific and academic worlds. Among the most important advantages is the notable capacity for accuracy and speed of diagnosis on a large amount of data, even about patients in the early stages. A further advantage lies in suggesting personalized treatment plans based on individual patient data, improving outcomes and reducing trial-and-error approaches. Moreover, AI-powered tools enable remote monitoring of patients’ health, facilitating telemedicine, extending healthcare access to remote areas and expediting the drug discovery process by analyzing molecular structures, predicting potential drug candidates, and accelerating research phases.

Nevertheless, some cons have to be considered. Among the most discussed are data privacy and security concerns, as long as sensitive patient data raises concerns about privacy breaches and security vulnerabilities in AI systems. Another critical aspect lies in ethical consideration, as integrating AI systems with existing healthcare infrastructure can be complex, requiring upgrades and adaptations, which may disrupt workflows. All this is concerning regulatory and safety aspects.

Otherwise, some technical aspects have to be considered. AI models may exhibit biases based on the data they were trained on, leading to errors or inaccurate predictions, especially in underrepresented populations. While AI holds immense potential in revolutionizing healthcare, addressing these challenges is crucial to ensure responsible and effective integration into the healthcare ecosystem.

Many applications have been implemented involving AI in the medical field in the last decade, ranging from medical imaging analysis, clinical decision support systems, genetic analysis, genomics, and diagnostic assistance. In drug discovery, AI accelerates the process by analyzing complex biological data, reducing the time and costs of bringing new medications to market. Predictive analytics powered by AI predicts disease progression and patient outcomes, aiding healthcare providers in making informed decisions for better patient management. AI facilitates remote patient monitoring, providing real-time data for proactive care outside traditional healthcare settings. This continuous monitoring allows for early intervention and reduces the need for frequent hospital visits. Additionally, AI optimizes hospital operations, streamlining workflows and allocating resources efficiently.

Enhancements in medical imaging analysis, particularly in radiology and pathology, contribute to more accurate diagnoses. Of particular interest has been the application of these methodologies to analyze images from MRI and CT scans to detect and characterize tumours in various organs or tissues and to assist radiologists by highlighting suspicious areas that might indicate the presence of tumours, aiding in early detection. The high algorithmic capacity is employed to analyze patient data, medical records, lab results, and symptoms to provide clinicians with insights, potential diagnoses, and treatment recommendations.

Following the canonical division, the supervised learning methodology has been commonly used for cancer diagnosis, organ segmentation, radiotherapy dose denoising and prediction. On the contrary, as regards the domain adaptation task, classification of patient groups and image reconstructions, the unsupervised learning approach (autoencoders, dimensionality reduction, clustering) led to better results. Addressing issues such as tumour segmentation and treatment planning has been extensively tested by reinforcement learning (Q-learning, Markov Decision Process).

Natural language processing by AI extracts valuable insights from unstructured medical records, supporting clinical decision-making. In surgical settings, AI assists in robotic-assisted surgeries, enhancing precision and minimizing invasiveness. Furthermore, AI applications empower patients with personalized health information, fostering engagement and promoting preventive care.

Integrating of AI in healthcare signifies a transformative shift, improving diagnostics, treatment strategies, and overall patient care. These advancements underscore the potential of AI to revolutionize the healthcare landscape, offering unprecedented benefits for individuals globally.

2.3 The Role of Handwriting

Handwriting [139] is a complex motor skill and form of communication where individuals use a writing instrument to create marks on a surface, typically paper. It involves the coordination of fine motor movements and cognitive processes. Each person's handwriting is unique, influenced by muscle control, hand-eye coordination, and individual style.

Handwriting can be analyzed in various ways and for different purposes, including forensic investigations, personality assessments, and medical diagnoses. The study of handwriting is known as graphology. It encompasses the formation of letters, spacing, slant, size, and overall visual appearance. Handwriting can convey personal traits, emotions, and cultural nuances. It has historical significance, preserving information in personal letters, documents, and manuscripts. Forensic contexts focus on authenticating documents by scrutinizing specific characteristics like letter formations, slant, and pressure. This deep comparison aids in determining the authorship or authenticity of handwritten materials.

Handwriting analysis becomes a valuable tool for identifying neurological or motor disorders in medical diagnosis. By scrutinizing patterns such as tremors, micrographia, and inconsistencies, practitioners can detect early signs of conditions like PD. Biometric authentication leverages the unique features of an individual's handwriting, such as pressure, speed, and stroke sequence, to verify identity for security purposes. Integrating computerized systems into this process adds a layer of sophistication, emphasizing the adaptability of handwriting analysis in contemporary technological contexts.

Advancements in ML and computer vision further propel the field forward. These technologies allow for automated analysis, enabling signature verification and sentiment analysis tasks. Such automation enhances efficiency and contributes to more objective and standardized handwriting analysis. Kinematic analysis looks at the motor control and movement patterns involved in handwriting. Researchers gain

insights into the intricate dynamics of the writing process by measuring aspects like speed, acceleration, and pressure using specialized tools. This approach is particularly relevant for understanding the physiological aspects of handwriting. Cognitive assessment through handwriting analysis delves into the study of writing-related cognitive processes. Analyzing elements such as pen pressure, stroke duration, and pauses provides valuable insights into cognitive functions, attention, and memory. This interdisciplinary approach bridges the realms of psychology and neurology.

Emotion recognition, an intriguing facet of handwriting analysis, seeks to understand emotional states by examining writing characteristics. Handwriting changes may reflect mood or psychological well-being shifts, offering a unique avenue for emotional insight. In summary, the versatility of handwriting analysis is evident in its applications across forensic, medical, biometric, and cognitive domains.

As technology advances, the field benefits from increased efficiency and objectivity, particularly with ML and computer vision integration. These developments open up new possibilities for handwriting analysis as a valuable tool in various professional and clinical settings. The advent of digital technology has led to a decline in the emphasis on handwriting, with keyboards and touchscreens becoming dominant tools for written communication. Cursive writing, a script with connected letters, is a traditional form of handwriting that has seen variations in education curricula. Handwriting can have therapeutic benefits, promoting cognitive function and mindfulness.

Many applications have been developed based on the handwriting. Handwriting recognition technology facilitates the conversion of handwritten text into digital formats. Signature analysis is a specialized handwriting examination used for authentication and security purposes. Handwriting can evolve, influenced by age, health, and external factors, and it is a cultural and educational skill that evolves across societies and generations. It engages multiple brain regions and is linked to improved learning and memory retention. Despite technological advances, the personal touch and individuality conveyed through handwriting remain valued in certain contexts.

Research, exemplified by studies [147, 90], indicates that intricate tasks like drawing and handwriting demand a combination of graphomotor skills. These encompass visual-perceptual maturity, spelling coding, motor planning and execution, kinesthetic feedback, and visual-motor coordination. The integrity of these skills is crucial, as any dysfunction linked to brain disorders can significantly impair an individual's drawing and handwriting performance. The analysis of graphomotor impressions has proven valuable as a psychometric tool for identifying various neuropsychological and neurological disorders, including apraxia, visuospatial neglect, dysgraphia, and dementia [125]. This underscores the significance of handwriting as a potential diagnostic indicator. Studies such as [82] also establish a connection between writers' emotional states and handwriting, further highlighting the intricate interplay of cognitive and emotional factors in writing. This holistic understanding of the cognitive components involved in handwriting emphasizes its role as a motor skill and a reflection of both neurological function and emotional states.

Handwriting analysis holds utility in diagnosing AD due to its sensitivity to cognitive decline and motor skill changes associated with the condition [35, 138, 88]. As Alzheimer's progresses, individuals often experience deterioration in their handwriting, characterized by irregular size, spacing, and letter formation. These alterations reflect cognitive decline, affecting the intricate coordination required for precise writing. Handwriting analysis is a non-invasive and cost-effective tool for early detection, potentially allowing for timely intervention and support.

The degradation of fine motor control and visuospatial abilities in Alzheimer's patients manifests in the written script, making handwriting a valuable indicator of disease progression. Researchers have identified specific features such as micrographia (reduced letter size) and dysfluency in writing patterns that correlate with cognitive decline in Alzheimer's. By examining these nuances, handwriting analysis can contribute to a more comprehensive diagnostic approach, complementing other clinical assessments. Moreover, writing engages multiple cognitive processes, including memory, attention, and executive functions. Changes in these cognitive domains, common in Alzheimer's, often manifest in written expression. Analyzing handwriting may provide insights into the evolving cognitive profile of an individual, aiding in the differentiation between Alzheimer's and other forms of dementia.

While handwriting analysis alone is not diagnostic, its integration into a broader assessment toolkit enhances the accuracy and sensitivity of Alzheimer's detection. The non-intrusive nature of this method is particularly valuable for monitoring cognitive decline over time, facilitating personalized care strategies, and improving the quality of life for individuals affected by AD.

2.4 Related Work

2.4.1 Handwriting Analysis

Section 2.3 widely describes the handwriting process and how it is sensitive to the symptoms of NDs. Handwriting can be analyzed in various ways depending on the type of investigation intended. In the context of NDs, different types of analysis can be performed:

- **Cognitive Assessment:** aims to evaluate cognitive processes related to writing by analysing elements such as pen pressure, stroke duration, and pauses to obtain detailed information about cognitive processes, attention, and memory.
- **Kinematic Analysis:** aims to assess motor control and movement patterns. It involves measurement of the kinematics of writing [1], including speed, acceleration, and pressure, often recorded with specialized equipment such as digitizing tablets.
- **Image Analysis:** consists of graphical analysis of handwriting images, using techniques including computer vision algorithms and ML to recognize and analyze unique patterns in writing.

Cognitive assessment tests play a crucial role in diagnosing and monitoring NDs, offering clinicians valuable insights into the cognitive functioning of individuals. Among these, the MMSE is a widely employed screening tool, evaluating aspects like orientation, memory, attention, and language to derive an overall cognitive impairment score. Another commonly used assessment, the MoCA, extends beyond the MMSE by incorporating tasks that assess executive functions and visuospatial abilities. Its increased sensitivity makes it particularly useful in detecting mild cognitive impairment, a crucial stage in the progression of neurodegenerative diseases. For AD specifically, the Alzheimer’s Disease Assessment Scale-Cognitive Subscale is tailored to assess cognitive dysfunction, including memory, language, and praxis. This focused approach aids in diagnosing and tracking cognitive decline associated with AD. The Clinical Dementia Rating offers a more comprehensive evaluation, considering multiple cognitive and functional domains. Widely used for staging dementia severity, the Clinical Dementia Rating provides an understanding of the impact of NDs on an individual’s daily life. Addenbrooke’s Cognitive Examination takes a different approach by aiming to detect and differentiate between various types of dementia, including Alzheimer’s and frontotemporal dementia. Its versatility makes it a valuable tool for identifying specific cognitive impairments associated with different neurodegenerative conditions. For a more focused assessment of executive functions linked to the frontal lobes, the Frontal Assessment Battery proves valuable. Tasks targeting mental flexibility, motor programming, and inhibitory control contribute to a nuanced evaluation of cognitive functions related to the frontal lobes. The Repeatable Battery for the Assessment of Neuropsychological Status provides a comprehensive battery covering various cognitive domains, offering a more detailed understanding of an individual’s cognitive strengths and weaknesses. The Trail Making Test stands out in assessing cognitive flexibility and visual attention with its two parts (A and B). This test is particularly valuable for evaluating how individuals switch between tasks and manage complex cognitive processes. While intelligence tests like the Wechsler Adult Intelligence Scale are primarily designed to assess general intelligence, specific subtests within the Wechsler Adult Intelligence Scale can offer insights into cognitive functions affected by neurodegenerative diseases, such as working memory and processing speed.

In conclusion, the selection of cognitive assessment tests depends on the specific diagnosis goals and the suspected neurodegenerative condition. Combining these tests with clinical evaluation and neuroimaging often provides a more comprehensive picture, enabling healthcare professionals to make informed decisions regarding diagnosis and treatment. Cognitive assessment tests are typically administered and evaluated by trained healthcare professionals, such as neuropsychologists, neurologists, or clinical psychologists.

Integrating advanced techniques, such as ML and DL methodologies, allows these analyses’ automation and improved objectivity. AI algorithms can analyze handwriting patterns to detect subtle changes related to cognitive impairment that may escape manual assessment, providing quantitative measures that may aid in early diagnosis and monitoring disease progression.

While many studies in the field of handwriting analysis still rely on conventional statistical techniques, like [73, 115], there is a growing body of literature that embraces ML methodologies for data exploration, in particular in the context of PD [7, 84, 39].

In [58], an investigation into handwriting time series encompassed parameters such as horizontal and vertical velocity, absolute velocity, acceleration, pressure, and trajectory curvature. The researchers adopted

a noise-robustness paradigm employing the singular value decomposition technique and a sparse, non-negative least-square classifier. In [2], the authors utilized features extracted from the Attentional Matrices Test to assess selective attention. Using various machine learning algorithms and an ensemble scheme, these features were employed to classify subjects into distinct groups, such as AD patients or healthy controls. In [47], the authors introduce the PaHaW Parkinson’s disease handwriting database, comprising samples from 37 PD patients and 38 healthy controls across various handwriting tasks. Kinematic and pressure features were examined for their potential in PD diagnosis. Three classifiers were compared, demonstrating that the analysis of kinematic and pressure features in handwriting can effectively discern subtle characteristics, aiding in differentiating PD patients from healthy controls. The research in [21] introduces the DARWIN dataset for studying AD. It is the largest publicly available dataset, featuring handwriting data from individuals affected by Alzheimer’s and a control group. With 174 participants, the dataset follows a specific protocol designed for early Alzheimer’s detection. The study assesses the effectiveness of proposed tasks and features in capturing distinctive handwriting aspects for Alzheimer’s diagnosis. The research addresses the need for standardized experimental protocols and datasets to explore handwriting dynamics as a tool for early neurodegenerative disease diagnosis.

Additionally, in [142], the authors employed semi or unsupervised learning techniques to unveil homogeneous clusters of subjects. The analysis focused on understanding the information carried by these clusters regarding cognitive profiles. The researchers introduced a novel temporal representation learning approach from handwriting trajectories, uncovering a comprehensive set of features such as the complete velocity profile, size and slant, fluidity, and shakiness. This approach revealed how these handwriting features collectively characterize cognitive profiles. In the study [137], kinematic measures of the handwriting process were conducted to evaluate the significance of features in distinguishing groups and assessing handwriting characteristics across five distinct functional tasks of copying. The findings indicated that kinematic measures effectively differentiated between patients in different groups in conjunction with the MMSE score. Notably, pressure and time-in-air emerged as the top-performing features in the analysis. Similarly, [71] focused on analyzing the stability of the offline handwritten word “mamma” (meaning ‘mum’ in Italian) to discern AD patients from healthy controls. The stability of the word was quantified by segmenting its image into elementary parts and measuring the similarity among adjacent segments. The authors employed the Yoshimura approach as a classification algorithm, comparing stability features between the sample to be recognized and training samples. In a unique approach presented in [106], the authors explored the early diagnosis of AD by analysing handwritten signatures. Patients’ signatures were represented using the Sigma-Lognormal model [41, 97], incorporating twelve features.

The objective of the research detailed in [56] was to differentiate participants from three distinct groups (AD, MCI, and a control group) by comparing their handwriting kinematics. Discriminant analysis served as the classification algorithm, and a protocol consisting of seven tasks, including copying and drawing tasks, was adopted. The study investigated the most discriminating features for the same task and identified that discriminating features were group-dependent. Some tasks, such as the clock drawing test, facilitated effective discrimination between certain groups with high scores of evaluation metrics.

In [50], the author proposed a cost-effective, rapid, and accurate CNN model for early AD diagnosis. Using the DARWIN dataset, they transformed handwriting features into 2D RGB images, and their model achieved remarkable accuracy. The study in [105] explores CNNs to analyze images of handwritten dynamics for PD diagnosis. The provided dataset, containing images and signal-based data, supports research on computer-aided PD diagnosis. The proposed CNN-based approach yielded promising results, especially in early-stage detection compared to raw data and texture-based descriptors. The findings suggest that leveraging deep learning on handwritten dynamics is valuable for automatic PD identification, potentially surpassing traditional handcrafted features. The work in [14] introduces a decision-aid tool for early AD detection using Archimedes spiral drawings on a Wacom digitizer. The study explores transfer learning to address sparse data, embedding kinematic time functions in spiral trajectory images. Through experiments on 30 AD patients and 45 healthy controls, the extracted features significantly improved sensitivity and accuracy compared to raw images. The research identifies intermediate-level features as the most discriminant, and the decision fusion of experts trained in these features outperforms low-level fusion.

In [89], the focus is on handwriting tremors prevalent in NDs. The research collects image-based handwriting trajectories from individuals with mild and severe PD and those without tremors. Image features are extracted, and a corner detection method is employed to assess trajectory fluctuations. The trajectories are then transformed into frequency-domain space using a 2-dimensional Discrete Fourier Transform, and texture features from the amplitude spectrum are extracted through the grey-level co-occurrence

matrix. Finally, a ML approach is used to classify these features, enabling the diagnosis of diseases. The work in [93] introduces the use of recurrent neural networks for early stage AD classification based on handwriting. The study compares Bidirectional Long Term Short Term and CNN methods. The focus extends beyond accuracy alone, considering the energy costs incurred during training for a comprehensive accuracy-efficiency trade-off analysis. The study emphasizes the importance of examining accuracy-efficiency trade-offs in neural network models to mitigate environmental impacts during training.

The study in [31] aims to detect and classify early-stage Alzheimer’s patients using online handwriting loop patterns. It addresses limited training data challenges by employing data augmentation techniques, including a variant of Generative Adversarial Networks, DoppelGANger, for synthesising realistic online handwriting sequences. Methods involve data preprocessing, traditional augmentation, and DoppelGANger for synthetic data. A 1D CNN is chosen for classification, with feature selection and evaluation metrics applied. Results show DoppelGANger effectively generates synthetic data, leading to state-of-the-art Alzheimer’s classification performance. In [5], the authors developed an early diagnostic method for Parkinson’s disease using artificial intelligence and the spiral test. Patients’ spiral drawings are analysed, employing an Echo State Network and a Multilayer Perceptron (MLP) layer for classification. Various algorithms, including boosting and decision trees, validate the approach.

In [81], the authors aimed to develop an efficient early diagnosis method for PD using off-line handwriting analysis, specifically focusing on spiral hand drawings. A Continuous Convolution Network was employed to overcome limitations in existing AI-based methods. The research in [119] aimed to enhance PD diagnosis using dynamic handwriting analysis through a three-stage fuzzy classifier method. It extracted features based on kinematic characteristics and pen pressure using public datasets. The fuzzy classifier construction involved generating the structure, feature selection, and tuning parameters with fuzzy logic rules. In [112], a PD detection system was developed using machine learning and handwriting analysis, focusing on spiral and wave drawings from healthy individuals and PD patients. They used a Histogram of Oriented Gradients for feature extraction and a Random Forest (RF) for classification. An interesting study in [94] describe the development of a vision-based system using commodity cameras and RGB video to capture and analyse handwriting kinematics for NDs screening. Using a smartphone camera and digitising tablet, the method compares kinematic data from both sources, achieving good results demonstrating the system’s potential as an accurate and accessible screening tool for NDs.

The research in [101, 120] compares AI classification methods for PD diagnosis using handwriting samples. Cartesian Genetic Programming outperforms white-box approaches in accuracy and black-box methods in interpretability by providing explicit rules. The findings suggest that the proposed approach offers a twofold benefit: supporting PD diagnosis and generating explicit classification models, aiding in designing non-invasive and cost-effective diagnostic protocols. The comparison involved machine learning techniques on handwriting samples from benchmark datasets (PaHaW and NewHandPD), highlighting its superior performance in accuracy and interoperability. The study in [103] addresses the development of a ML tool for PD diagnosis that provides accurate results and explains its decisions in an understandable way for clinicians. The Decision Tree (DT) is chosen for its transparency in presenting decision criteria based on relevant features. The evaluation on a public dataset demonstrates that the decision tree-based system achieves comparable or superior results to state-of-the-art solutions and stands out as the only approach providing a clear description of decision criteria based on observed features and their values. In [32], the authors propose a multi-classifier approach, employing as many classifiers as there are tasks (handwriting and drawing) for discrimination. The method combines outputs through a majority vote. Experiments using six popular ML techniques demonstrate that selecting task-specific classifiers and combining their outputs achieve superior results.

The study in [102] explores neurodegenerative disease diagnosis through handwriting and drawing analysis, employing one-class classifier models. These models, requiring only healthy subject data for training, eliminate the need for patient data collection. In this article [57], the authors introduce an innovative approach for diagnosing PD using CNNs applied directly to handwriting images. Unlike traditional frameworks, this method eliminates the need for specialized devices or feature engineering, offering an end-to-end solution. The proposed architecture employs multiple fine-tuning steps and an ensemble.

Various handwriting modalities for PD diagnosis are assessed in [45, 43, 44], including on-surface, in-air, and pressure on the tablet surface. It emphasizes the significance of in-air movement and pressure-based features. Including entropy and empirical mode decomposition features alongside traditional kinematic and spatio-temporal features enhances diagnostic capabilities. The study in [48] focuses on micrographia, a common clinical sign of PD, characterized by reduced letter size and altered kinematic aspects in

handwriting. The research introduces a template to capture handwriting during various tasks, including established PD diagnosis tasks like the Archimedean spiral and new tasks addressing micrographia aspects. Another study, [46], aims to identify a subset of handwriting features for effective PD diagnosis, extracting various kinematic measures. Novel measures based on entropy, signal energy, and empirical mode decomposition were computed. In [38], the authors explore the potential of dynamically enhanced static images of handwriting in CAD systems. The enhanced images are synthetically generated, incorporating both static and dynamic properties of handwriting to improve discrimination. The proposed representation involves drawing points of the samples without linking them and adding pen-ups to retain temporal/velocity information.

The impact of advanced online handwriting parameterization using fractional-order derivatives is analysed in [92] for PD dysgraphia. The research explores the relationship between the newly designed features and clinical data through partial correlation analysis. Binary classification analysis evaluates the discrimination power of these features, and regression models assess their ability to gauge the progress and severity of PD, comparing results with a baseline of conventional online handwriting features. The computed features demonstrate stronger correlations with clinical characteristics and more accurate assessments of PD severity, suggesting potential improvement in computerized PD severity assessment when combined with specific tasks. The study in [87] explores the dynamics of signatures based on recent motor learning findings, suggesting that signatures are stored in the brain as both trajectory and motor plans. The research proposes that the stored representation may focus on specific, more learned parts of the signature, executed more automatically and less prone to variations. The study discusses experiments using an algorithm to identify and utilize these stable regions in signature ink for automatic verification. Finally, the study in [121] explores handwriting movements as a non-invasive tool for early screening of PD, particularly focusing on cases where the disease affects the contralateral side of writing. The research identifies distinctive signs in early-stage PD handwriting, indicating the potential for early disease detection. The study analyzed handwriting samples from PD patients and healthy subjects using a novel protocol with various complexity levels. Findings reveal that specific features during the execution of handwriting tasks can contribute to early PD detection, offering guidelines for designing a diagnostic tool and suggesting conditions that benefit patients' performance.

Many researches focus on studying how gender, age and environmental factors can influence one's handwriting. The research described in [55] aims to compare the classification performance of automatically extracted features by a pre-trained CNN and handcrafted features in detecting PD dysgraphia. The multilingual dataset includes Parkinson's patients and healthy controls from various countries. Three analysis scenarios explore the impact of language on classification. Results indicate that handcrafted features slightly outperform CNN-extracted features in all language scenarios for sentence writing tasks, while for spiral drawing tasks, CNN-extracted features show competitive results.

The study described in [63] aimed to enhance PD diagnosis accuracy by considering inherent neurological differences between genders and age groups. Using online handwriting data from individuals with Parkinson's and healthy controls, a sex-specific and age-dependent classifier outperformed the generalized classifier. Combining age and sex information proved beneficial, revealing distinct features for higher accuracy in different classification categories. In [4], the authors investigated age-related changes in handwriting among healthy individuals using ML. Subjects were categorized into younger adults, middle-aged adults, and older adults. Handwriting tasks were digitized and analyzed with a CNN and DBNet algorithm for stroke sizes. The CNN effectively distinguished age groups with a valuable performance, highlighting the model's robustness in classifying age-related handwriting changes.

The research in [16] presents a method for synthesizing the temporal evolution of handwriting from childhood to adulthood for biometric applications. The approach includes online and offline handwriting, utilizing parameters to manage the synthesized handwriting evolution. The methodology simplifies text trajectory plans and handwriting dynamics using a modified kinematic theory and a neuromotor-inspired synthesizer. Realism is evaluated through quantitative tests measuring letter variability and stroke count and a subjective evaluation by 30 individuals assessing the perceived realism of the synthetic handwriting's evolution. In [107], the authors investigate the complexity of handwriting generation as a neuromotor skill, exploring the interaction of cognitive processes involved in producing ink traces on a writing medium. The Kinematic Theory of rapid human movements and its lognormal models offer analytical representations, considering strokes as fundamental handwriting units. The lognormality of velocity patterns is interpreted as a reflection of subjects in perfect control of their movements, supported by experimental confirmation and physiological tests. The paper explores how software tools can leverage these models to analyze ideal lognormal behaviours and investigate the operational convergence hypothesis through

studies on motor learning in children and the impact of ageing on handwriting.

The study in [114] investigates age-related changes in executive functions and handwriting performance in 80 healthy participants. Using the Behavioral Assessment of the Dysexecutive Syndrome and the Computerized Penmanship Evaluation Tool, the research reveals significant differences in executive functions and temporal/spatial handwriting measures across age groups.

In conclusion, integrating AI and handwriting analysis has demonstrated promising advancements in supporting the diagnosis of NDs. The synergy between cutting-edge technology and the intricate patterns within handwriting offers a unique avenue for early detection and accurate assessment. This section also underscored the importance of collaboration between AI, neuroscience, and psychology experts, which becomes imperative to implement innovative solutions that hold the potential to revolutionize the diagnosis and understanding of NDs, ultimately contributing to improved therapeutic interventions and enhanced quality of life for individuals affected by these conditions.

2.4.2 Speech Analysis

Speech analysis has shown promise as a non-invasive and cost-effective tool for the early diagnosis and monitoring of NDs. Several NDs, such as AD, PD, and Amyotrophic Lateral Sclerosis (ALS), can affect speech patterns and vocal characteristics. Changes in speech may occur in the early stages of NDs, often before other noticeable symptoms. Speech analysis provides an objective and quantifiable way to measure changes in speech parameters, such as pitch, intensity, rate, and pauses. This objectivity can be particularly valuable for tracking disease progression over time. It can monitor the progression of NDs, providing insights into how the diseases impact different aspects of speech over time. This longitudinal data can be valuable for both clinicians and researchers. Moreover, speech analysis can be performed remotely, continuously monitoring individuals in their natural environments. This is particularly beneficial for individuals who may face challenges with regular clinic visits. Analyzing speech characteristics can help in early detection and intervention, potentially allowing for more effective disease management. Advances in technology, including machine learning and natural language processing, have facilitated more sophisticated analysis of speech patterns. These technologies can identify subtle changes that may not be easily discernible to the human ear. Ongoing research in speech analysis for NDs contributes to developing new diagnostic tools and technologies. This includes the exploration of voice-based biomarkers and the integration of speech analysis with other types of data. In [15], the authors introduce a novel speech kinematics-based model for studying and analyzing complex speech movements. Unlike previous speech motor models, this model employs the kinematic theory of rapid human movements and the Sigma-lognormal model, similar to approaches used in handwriting studies. The method parameterizes the neuromuscular response to a neuromotor command, allowing for the derivation of muscular response parameters and the subject's age from continuous speech.

The study in [86] aims to address the challenge of detecting cognitive impairment, particularly in AD, by employing automatic speech analysis as a non-invasive screening tool. The research presents a non-linear multi-task approach analyzing three tasks with varying language complexity levels, incorporating features such as linear features, perceptual features, Castiglioni fractal dimension, and Multiscale Permutation Entropy. Instead, in [9, 111] are presented systematic reviews related to the speech analysis. In particular [9] explores hypokinetic dysarthria in PD, focusing on early diagnosis, disease progression monitoring through acoustic voice and speech analysis, neural correlates investigated via functional imaging, and the impact of dopaminergic medication and brain stimulation. The review identified 14 recommended combinations of speech tasks and acoustic features for describing this disorder in PD. In [111], instead, a state-of-the-art review of automatic speech and voice analysis techniques for monitoring patients with AD is presented. It focuses on feature extraction techniques, classification methods, and frequently used data repositories. The review aims to guide researchers in the field, highlighting clinically relevant results and current developments.

The research in [99] introduces a methodology for automatically detecting pathologies in the phonatory system using continuous speech records. Based on estimating nonlinear dynamics features, the approach enables the segmentation and characterization of voice registers without relying on pitch period estimation, making it independent of gender and intonation. Related to this research, the work in [66] focus on assessing voice quality by employing objective nonlinear measures, departing from traditional linear techniques. Six chaotic measures based on nonlinear dynamics theory were applied to discriminate between healthy and pathological voice qualities.

DL techniques are also considered in this field, like in [67], which describes an ensemble of CNNs for the

computerized detection of Parkinson’s disease (PD) based on voice recordings.

The growing research in machine learning and deep learning for diagnosing NDs through speech analysis holds significant promise, offering a non-invasive and accessible avenue for early detection and monitoring, thereby contributing to advancements in timely and effective clinical interventions.

2.4.3 Gait Analysis

Gait analysis systematically studies an individual’s walking pattern, including step length, walking speed, stride duration, and other related parameters. It is typically conducted through observational methods, wearable sensors, or specialized equipment like pressure-sensitive walkways. While gait analysis is often associated with musculoskeletal system conditions, it can also be valuable in the context of NDs, including AD.

Many studies suggest that alterations in gait patterns may be observed in individuals with AD, particularly in the later stages of the condition. Gait analysis may provide insights into motor function and coordination, potentially contributing to a more comprehensive understanding of the disease progression. In summary, while gait analysis is not a primary tool for diagnosing AD, it can offer supplementary information about motor function and may be considered as part of a comprehensive assessment, especially in research studies focused on understanding the broader impact of NDs on movement.

However, Gait Analysis is more commonly utilized to assess and diagnose movement disorders such as PD [76]. It is particularly useful for both monitoring and diagnosing PD at different stages, as it is a disorder that affects movement, where changes in gait are common symptoms. Individuals with Parkinson’s often exhibit specific gait abnormalities, including shuffling steps, reduced arm swing, and a stooped posture. Analyzing these gait characteristics and other clinical assessments can aid in the accurate diagnosis of PD. Moreover, changes in gait patterns can indicate disease progression and regular gait assessments can help healthcare professionals track the evolution of motor symptoms. This information is crucial for adjusting treatment plans and interventions accordingly. By regularly assessing gait parameters, healthcare providers can determine whether interventions are helping to alleviate symptoms and improve overall mobility.

One of the primary data acquisition methodologies involves utilizing visual information, where patients perform movement tests [62]. In this context, gait analysis systems based on vision can be categorized into two main types: marker-based and markerless systems. In gait analysis, systems considered gold standards for gait assessment are marker-based systems. Multi-Camera Motion Capture systems are commonly used in clinical settings due to their high tracking accuracy and sampling frequency [91, 100, 144].

In marker-based systems, spherical or reflective markers are detected by cameras or motion sensors, and the collected data are used to reconstruct the three-dimensional kinematics of joints and body segments during walking. However, these systems have limitations, such as being expensive, requiring significant setup time, and potentially influencing the naturalness of movement during data acquisition. Additionally, specialized personnel are needed to position markers on patients correctly. For these reasons, significant efforts have been dedicated in recent years to study and implement markerless systems.

The evolution of computer vision has represented a significant advancement in this field, opening new possibilities for gait analysis. Using ML and DL techniques, combined with gait feature recognition algorithms, has led to new markerless solutions that leverage information extracted from videos. These systems can detect and track anatomical features and body landmarks without markers. Thanks to recent technological innovations, key joint positions can be directly inferred from colour or depth images through 2D prediction algorithms, such as OpenPose [13], or 3D prediction algorithms [30, 122].

In [60], a 3D CNN model was proposed, utilizing spatiotemporal saliency maps of RGB images. In another case [116, 117], 2D and 3D skeletons of PD patients were extracted using multivariate ordinal Logistic Regression (LR) models and the SpatioTemporal Graph Convolutional Network to predict PD severity from joint trajectories.

In conclusion, the fusion of artificial intelligence and gait analysis stands at the forefront of transformative advancements in diagnosing NDs. AI-driven gait assessments offer a unique lens into these disorders’ subtle yet significant markers. This highlights the importance of the collaboration between AI experts, biomechanics specialists, and healthcare professionals.

2.4.4 NeuroImaging Analysis

Neuroimaging in detecting NDs represents a crucial advancement in research and clinical applications. These imaging techniques, including MRI, PET, Electroencephalography (EEG) and SPECT, provide a window into the structural and functional changes occurring in the brain. One of the primary advantages of neuroimaging is its role in early detection, which is vital for timely intervention and effective management. Structural changes in the brain, such as atrophy patterns and the presence of abnormal protein deposits, can be visualized through MRI and PET scans.

Functional imaging techniques delve into the dynamics of brain activity and connectivity. Changes in regional cerebral blood flow, glucose metabolism, and neural network functioning offer valuable insights into disease progression. The ability to identify biomarkers associated with specific diseases, such as beta-amyloid plaques in AD or abnormal protein aggregates in PD, contributes to a more accurate and targeted diagnosis. Moreover, neuroimaging plays an important role in research and drug development. In clinical trials, these techniques help monitor changes in the brain over time, assess the efficacy of treatments, and deepen our understanding of the underlying mechanisms of NDs. Tracking disease progression longitudinally provides critical data for predicting trajectories and evaluating treatment outcomes. Neuroimaging facilitates a more individualized approach to diagnosis and treatment as we move towards personalised medicine. By assessing the unique brain characteristics of each patient, healthcare providers can tailor interventions to specific needs, leading to more effective and patient-centric care. Despite these advancements, challenges remain, including accessibility to advanced imaging technologies, standardization of imaging protocols, and the need for further research to uncover additional biomarkers.

The integration of AI with neuroimaging techniques revolutionizes the field of neuroscience. AI algorithms, particularly DL models, enhance the analysis of vast and complex neuroimaging datasets, enabling more accurate and efficient detection of abnormalities associated with NDs. In neuroimaging, AI facilitates automated segmentation of brain structures, precisely quantifying volumes and abnormalities. ML algorithms applied to functional MRI data contribute to the identification of unique patterns of brain activity, supporting personalized diagnostics and treatment strategies. The synergy between AI and neuroimaging is promising for advancing our understanding of brain disorders and improving early detection and intervention.

Many researchers focus on applying AI techniques to MRI images to diagnose neurodegenerative diseases [118]. The most used techniques involve 2D and 3D CNNs [72], in some cases in conjunction with the slice-level attention mechanism [68, 18]. The study in [12] addresses the challenge of predicting cognitive performance in AD using MRI measures employing a non-linear, norm-regularized multi-kernel multi-task feature learning formulation.

In [79], the authors propose a novel framework for monitoring AD, utilizing longitudinal neuroimaging data for clinical score prediction. The research in [145] focuses on enhancing the CAD for AD diagnosis through the automatic detection of dementia in MRI brain data. The study employs established techniques such as registration, slicing, and classification, introducing deep convolutional models and transformer-based architectures.

Moreover, various AI models have been developed to address the complexity of NDs diagnosis, utilizing clinical data and medical imaging like PET [110, 75] and EEG [8, 53]. In closing, the combined use of artificial intelligence and neuroimaging offers a promising frontier in NDs diagnosis.

2.4.5 Other Techniques

Despite the aforementioned studies on AI applications to support the diagnosis of NDs, the research community also uses various techniques focusing on genetic data, multimodal approaches, and daily activities.

The integration of genetic data and AI holds significant promise for advancing the diagnosis of NDs. Researchers and clinicians can use AI algorithms to analyze genetic information to identify patterns, mutations, and genetic markers associated with NDs. ML models, including DL, excel at detecting complex relationships within large genetic datasets, identifying genetic risk factors for diseases like Alzheimer’s and Parkinson’s [136, 69, 80]. The combination of genetic data and AI enhances the precision of diagnostics, enabling earlier and more accurate detection of NDs, potentially facilitating personalized treatment approaches based on an individual’s genetic profile. This interdisciplinary approach represents a cutting-edge avenue in medical research, potentially revolutionising our understanding and management of NDs. Many other researchers adopt a multimodal approach to analyse the disease from different points of view, considering a combination of data from neuroimaging, handwriting, wearable sensors, video and audio

recordings and other sources. For example, in [133], researchers introduced a multimodal dataset encompassing online handwriting, speech signals, and eye movement recordings, while in [134], three approaches were compared for PD detection: wearable sensors, video recordings, and handwriting samples; instead in [135] imaging, genetic, and clinical test data were analysed for AD and mild cognitive disorders. Wearable Devices and Sensors or patient and caregiver interviews are usually used to monitor daily activities, including movement patterns, sleep quality, and behavioural changes. These data can provide valuable insights into the early signs of neurodegenerative disorders [104, 96].

Chapter 3

Data

In Chapter 2, Section 2.3 discusses how changes in handwriting can be considered as observable indicators due to the cognitive and motor alterations associated with the AD. The degradation of fine motor control, visuospatial skills, and memory functions, hallmarks of Alzheimer’s, can manifest in the act of writing. Interest in handwriting analysis for diagnosing AD has increased over the past few decades, aligning with advancements in neuroscientific research and technology. The following Section 3.1 describes the data acquisition phase and presents an experimental protocol for handwriting tasks conceived in 2018 by a group of researchers from the University of Cassino and Southern Lazio [23]. The remainder of this chapter aims to describe the data involved in the experimental part of this research. It provides a detailed description of the image generation in Section 3.2 and ends with features calculation, described in Section 3.3.

3.1 Data Acquisition

In the context of AD, as stated in Section 2.3, alterations in handwriting are often observable due to the cognitive and motor changes associated with the condition. The degradation of fine motor control, visuospatial skills, and memory functions, which are characteristic of Alzheimer’s, can manifest in the act of writing.

The interest in handwriting analysis for diagnosing AD has grown over the past few decades, aligning with advancements in neuroscientific research and technology. While early studies exploring the link between handwriting changes and cognitive decline date back to the late 20th century, significant attention to this area has emerged in the 21st century. In the 2000s, researchers began to recognize the potential of handwriting analysis as a non-invasive and accessible tool for early detection. With the increasing prevalence of Alzheimer’s and the global ageing population, there has been a heightened emphasis on developing innovative diagnostic approaches. Technological advancements, including sophisticated imaging techniques and computer-based analysis tools, have facilitated more precise and quantitative assessments of handwriting changes associated with cognitive decline. This intersection of neuroscience, technology, and a growing awareness of the importance of early detection has fueled the surge in interest in handwriting analysis for Alzheimer’s diagnosis.

In 2016, the Department of Electrical and Information Engineering at the University of Cassino and Southern Lazio began investigating this topic. In 2018, the work [35] was published, providing a concise compilation of the research involved in this subject and highlighting the presence of issues and needs in this research field. Researchers discussed how NDs like Alzheimer’s and Parkinson’s impact patients’ lives and proposed methods for early diagnosis. They highlighted the relationship between these diseases’ symptoms and the gradual deterioration of motor skills, leading to difficulties in handwriting. They aimed to survey the state-of-the-art work on diagnosing NDs by handwriting analysis, showcasing achieved results and advocating for classification systems.

In particular, they discussed the absence of a well-designed dataset for NDs as a significant concern. First, they noticed that the available datasets regarding NDs comprised a very small number of participants. Limited data availability hampers the efficacy of classifier-based approaches, as they are known to be data-hungry. Secondly, they noticed that many approaches focus on offline acquisitions in the literature thanks to the large availability of handwritten documents.

In Section 2.1, it is explained how NDs impair not only the motor assessment of a person but also the cognitive aspect. This encompasses the need to define a new protocol of tasks, considering the en-

tire spectrum of consequences that neurological degeneration causes. In addition, new acquisition tools were necessary for an online characterization of the movements performed. Given these assumptions, in 2018, these researchers published an experimental protocol to support cognitive impairment through handwriting analysis [23] to strengthen standard diagnosis techniques with research on handwriting and neuro-muscular diseases. This study introduced an experimental protocol to address the aforementioned challenges, aiming to construct a database with hundreds of samples from subjects with NDs and healthy controls. This extensive database aims to enhance the performance of classifier-based approaches, enabling more effective training of the underlying classification algorithms.

Thus, in 2018, researchers from The University of Cassino and Southern Lazio decided to start a meticulous data acquisition campaign crucial for the validity and reliability of the data involved in the study. The acquisition was done by administering the experimental protocol. Participant selection was conducted following specific criteria defined through collaboration with the geriatric ward’s Alzheimer unit at the “Federico II” hospital in Naples. The selection process incorporated clinical assessments and standard cognitive evaluations such as the MMSE, FAB, and MoCA. These assessments span diverse cognitive domains, encompassing aspects such as orientation in time and place and registration recall. Healthy controls were chosen based on demographic and educational characteristics to ensure equitable comparisons. Both patients and controls underwent scrutiny for medication use, with an exclusion criterion for individuals using psychotropic drugs or other substances that could impact cognitive abilities.

Participants were informed about the research objectives and provided informed consent for participation. The research included a total of 174 participants, with 89 patients diagnosed with AD and 85 serving as healthy control subjects. Table 3.1 shows participants’ personal information, like age, years of school and number of female and male people for every class of the dataset, healthy controls (HC) and patients (PT). Moreover, this information is systematically recorded because writing skills may be influenced by factors such as age, educational background, and occupational type. By capturing these additional demographic and contextual details, the study aims to comprehensively understand how these variables may contribute to variations in writing performance.

People	Age	Education	#Women	#Men
HC	63.8 (11.0)	13.2 (4.3)	49	36
PT	71.7 (9.5)	10.7 (5.0)	46	43

Table 3.1: Average demographic data of participants. Standard deviations are shown in parenthesis

The proposed protocol aims to investigate the distinctive features in handwriting dynamics that can differentiate individuals affected by AD from healthy ones. For this reason, tasks were designed to increase in difficulty, targeting specific cognitive functions progressively. These tasks included graphic challenges, copy and reverse copy exercises, memory assessments, and dictation activities. Tasks aimed to evaluate motor control, coordination, memory, and spatial organization. Graphical tasks and free spaces were employed to evaluate the spatial organization skills of patients. Copy and dictation tasks enabled the comparison of writing variations in response to different stimuli (visual or auditory). Tasks involving different pen-ups allowed the analysis of air movements, known to be altered in patients with AD, while tasks with varied graphic arrangements, such as words with ascenders/descenders or complex shapes, assessed fine motor control capabilities. Task intensity and duration are varied to test patient responses under different fatigue conditions.

The tasks include copying letters with different graphic compositions, copying letters on adjacent rows to test spatial organization abilities, continuous cursive writing of single letters and bigrams, and word copying tasks.

The study also explores word copying with variations in spatial organization, introducing cues to observe the impact. For instance, patients were asked to sign their names, draw circles, copy letters and words, and perform memory tests. The tasks ranged from basic motor activities to more complex activities like copying a paragraph or performing the Clock Drawing Test (CDT) to assess cognitive functions associated with mild AD. The researchers introduced variations in word copying tasks, considering spatial organization and cues.

The protocol also incorporated tasks related to daily activities, such as copying details from a postal order. The researchers utilized a systematic approach, considering both quantitative and qualitative data. Tasks were carefully structured to avoid influencing patient performance, with the experimenter playing a critical role in guiding and ensuring accurate data collection. Overall, the detailed experimental protocol

aimed to uncover specific features in the handwriting of subjects affected by NDs, providing valuable insights into the cognitive aspects related to these diseases.

The protocol includes 25 writing tasks to highlight a potential deterioration in motor, spatial, and cognitive skills commonly compromised by AD. The protocol's tasks are organized in ascending order of difficulty, considering the cognitive functions required for task execution. Based on their objectives, tasks are divided into four categories, as follows:

- Graphic tasks: assess the patient's proficiency in producing basic strokes, connecting specific points, creating figures, both simple and intricate, and adjusting their proportions.
- Copy and Reverse Copy tasks: aim to evaluate the patient's capacity to replicate complex graphic movements with semantic significance, such as reproducing letters, words, and numbers of varying lengths and spatial arrangements.
- Memory tasks: focus on highlighting changes in the graphic component while retaining in memory a word, a letter, a graphic gesture, or a motor plan.
- Dictation tasks: aim to explore how writing in the task context (involving phrases or numbers) varies when utilizing working memory is necessary.

Table 3.2 defines every protocol task with the corresponding description and belonging group. Task 1 involves executing one's signature, a gesture frequently encountered in literature; individuals must perform a motion they have repeated multiple times throughout their lives. Tasks 2 and 3 investigate the wrist joint and finger joint motor abilities, respectively, while Tasks 4 and 5 test the movements' automaticity, coordination and spatial organization. Task 6 is a copy task of letters that present ascender and descender traits in their execution. Task 7 is a copy task to test the spatial organization. Tasks 8 and 9 evaluate the motion control alternation. Additionally, tasks from 10 to 13 mean checking the spatial organization. Task 14 is a short memory test, while 15 and 16 are reverse copy tasks inspired by the MMSE. Task 17 tests the handwriting of different types of words, with or without semantic sense, in defined boxes. Task 18 is a memory task; instead, task 19 requires performing a complex but daily activity. Task 20 involves dictation, so the person has to write without the stimulus of visualization. Task 21 is a complex graphic task to evaluate the person's fine and long motor control abilities. Tasks 22 and 23 are copy and dictation tasks involving numbers, respectively, which require a different motor planning from the one used for words. The CDT, task 24, is particularly useful for mild AD. In the last task, number 25, the person has to copy a short paragraph consisting of 110 characters. The table shows nine additional tasks obtained by considering parts of others.

This protocol aimed at recording handwriting samples and their dynamics to understand whether there were significant features to support the diagnosis of AD. Digital tools were employed to collect writing samples, ensuring the standardization of the process. Each participant was invited to perform the experimental protocol by using the WACOM Bamboo Folio graphic tablet, enabling participants to write on standard A4 white paper sheets using a pen that appears typical. This pen not only produced ink traces on the sheet but also generated digital information recorded by the tablet in the form of spatial coordinates and pressure for each point (x , y , and z). The data were acquired at a frequency of $200Hz$. The tablet additionally captured in-air movements, allowing tracking of motions up to a maximum height of 3cm from the tablet surface.

All participants were positioned comfortably, approximately 70cm from the sheet, and all individuals included in the study were right-handed. It is noteworthy that under these conditions, participants were instructed to maintain their natural writing movements, avoiding alterations commonly observed when using an electronic stylus on the surface of a tablet. For patients with AD or elderly individuals, using traditional paper and pen for tasks might be more intuitive and familiar than a digital tablet. Consequently, this choice ensures that the collected data remains free from biases from the invasiveness or unfamiliarity associated with using a less comfortable tool.

A computer application, developed in C#, accompanied the study, streamlining the uniform collection of data by automatically storing information generated by the tablet onto the computer's storage. The experimental procedure involved presenting visual and auditory stimuli to guide participants in task execution. Task instructions and the specified letters/words/phrases for copying were presented on the white sheets used by the subjects. Moreover, participants were instructed to adhere to the experimenter's guidance throughout the experiment. After completing each handwriting task, the specifically developed software automatically saved a *.csv* file in the computer's memory. Each file comprised four columns

Task	Description	Type
1	Signature	Memory
2	join two points horizontally (x4)	Graphic
3	join two points vertically (x4)	Graphic
4	trace a circle continuously (x4, d = 6cm)	Graphic
5	trace a circle continuously (x4, d = 3cm)	Graphic
6	copy "l, m, p"	Copy
7	copy "n, l, o, g" on adjacent rows	Copy
8	write continuously "l" (x4)	Copy
9	write continuously "le" (x4)	Copy
10	word copy: "foglio"	Copy
11	word copy with a cue: "foglio"	Copy
12	word copy: "mamma"	Copy
13	word copy with cue: "mamma"	Copy
14	memorize and then write: "telefono, cane, negozio"	Memory
15	reverse copy "bottiglia"	Reverse Copy
16	reverse copy "casa"	Reverse Copy
17	copy words in boxes: "pane, mela, prosciutto, ciliegia, taganaccio, lonfo"	Copy
18	write the name of the object shown (a chair)	Memory
19	copy the details of a postal order	Copy
20	write a simple sentence under dictation	Dictation
21	retracing a complex form	Graphic
22	copy a telephone number	Copy
23	write a telephone number under dictation	Dictation
24	Clock Drawing Test	Graphic
25	write a short paragraph from a FAB story	Copy
Additional Tasks		
26	Telefono	Memory
27	Cane	Memory
28	Negozio	Memory
29	Pane	Copy
30	Mela	Copy
31	Prosciutto	Copy
32	Ciliegia	Copy
33	Taganaccio	Copy
34	Lonfo	Copy

Table 3.2: Protocol handwriting tasks

documenting the timestamp, spatial coordinates (x, y) , and pressure (z) . This meticulous recording resulted in a compilation of individual *.csv* files, each corresponding to a distinct task performed by a participant. Subsequently, these files played a central role in generating new data, producing images, and computing features, as elaborated in the subsequent sections.

The subsequent discussion is divided into two sections for a more comprehensive exploration of the subject matter. The first section, Section 3.2, is specifically dedicated to explain the image generation process. This portion aims to provide a detailed and thorough examination of the steps involved in generating images within the context of the broader topic.

Concurrently, the second section, Section 3.3, is crafted to offer an understanding of the computation of features. This segment is strategically designed to delve into the methodologies and computations associated with extracting features within the specified framework.

3.2 Image Generation

This study delves into the analysis of handwriting as a component in aiding the diagnostic process of AD. The execution of the designed protocol not only produces a comprehensive *.csv* file containing pertinent information but also captures handwritten traits recorded on paper sheets affixed to a graphic tablet. Recognizing the dual significance of these information sources, the aim is to harness their potential by extracting and generating images.

These images, containing the subtleties of individual handwriting characteristics, serve as the input data for subsequent processing through ANNs. This approach seeks to leverage the amalgamation of quantitative data stored in the *.csv* files and the qualitative intricacies encapsulated in the handwritten traits, providing a multi-dimensional dataset for more robust and insightful analysis through the lens of AI. The interplay between traditional data and image-based information allows for a holistic exploration of handwriting features, contributing to a wide understanding of their potential role as diagnostic markers for AD.

Analyzing images of handwriting traits instead of dynamic features in the context of AD has advantages. First, basing the analysis on static traits, such as size, slant, pressure, and spacing, can be instrumental in the early detection of cognitive decline associated with AD. Changes in these static features may manifest before significant alterations in dynamic features. In addition, continuous monitoring of static traits provides a longitudinal perspective and enables the identification of subtle changes over time that may indicate cognitive decline. It's important to note that while analyzing static handwriting traits may offer valuable insights, a broader approach that considers both static and dynamic features could provide a more robust assessment of cognitive function in AD. Combining various methodologies may enhance the accuracy and reliability of early detection and monitoring efforts. This is why this research focuses on analysing several aspects of handwriting by considering multiple sources of information and many processing and AI techniques to support the diagnosis.

This section describes, in the following, three different processes of image generation. Every generated image must adhere to the standards specified by the ANNs employed for their processing. Ensuring conformity to these standards is paramount for seamless integration into the subsequent phases of the AI framework. The adherence to specific requirements is not merely a procedural formality; it directly influences the accuracy and efficiency of the image processing pipeline. The reliability and consistency of the analytical process are upheld by aligning generated images with the prescribed standards of the neural networks. Contemporary understanding underscores the indispensability of a rich and well-structured dataset for experiments leveraging AI and deep transfer learning techniques. Furthermore, considering the constraints imposed by the adopted CNN models and their respective input size requirements, a meticulous resizing of the original x, y coordinates into the range $[0, 299]$ was implemented for each image. This proactive adjustment ensures that the generated images conform to the designated size criteria, minimizing the potential loss of information associated with zoom-in/out effects.

Python, as the programming language, facilitated the development of specialized software tailored to generate diverse image types. Pillow and OpenCV, prominent libraries in the Python ecosystem, played a crucial role in managing image files. Pillow is particularly known for its image processing capabilities, while OpenCV is widely recognized for its computer vision and image manipulation functionalities. The integration of these libraries ensured a robust and efficient workflow for handling image-related tasks during the image generation process.

3.2.1 Binary Synthetic Images

Any endeavour rooted in AI, particularly when centred around DL methodologies, necessitates a preliminary data processing phase for systematically collecting and organising data. The intricacies of this phase are contingent upon the nature of the data being handled. In the context of this study, data processing has been executed through a dedicated Python script, distinctly separated from the model construction. This segregation ensures that the neural network can seamlessly load the pre-processed dataset, streamlining its operational commencement.

Within the initial data category, generating synthetic images is a process starting from the information encapsulated in *.csv* files. Specifically, these files contain different data, but this process required focusing on timestamps and spatial coordinates represented as (x, y) pairs. Each pair of coordinates corresponds to a discrete point captured by the system at a frequency of $200Hz$. This high acquisition rate translates to registering points at intervals as brief as $5ms$.

Therefore, the *.csv* files serve as repositories of temporally stamped spatial data, allowing for the reconstruction of the subject’s trajectory with a good approximation. The $200Hz$ sampling rate ensures that the synthetic images encapsulate spatial information, faithfully representing the dynamic evolution of the subject’s movements over time.

The generation process starts by treating points (x_i, y_i) as vertices of a polygonal structure, closely approximating the original curve. Each image undergoes a reconstruction process where the subject’s traits are delineated by interpolating consecutive points acquired during the execution of tasks. This binary representation, exemplified in Figure 3.1, depicts the subject’s trace with a constant line width of 5 pixels. This approach contributes to creating images that align with the specifications of the neural network models and strives to minimize any potential loss of essential information.

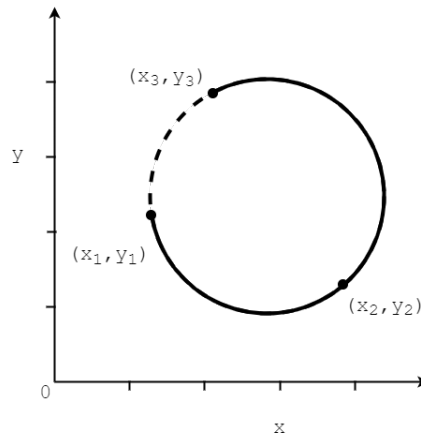


Figure 3.1: Example of generated strokes

Figure 3.2 illustrates a concrete instance of the synthesized strokes falling under this category.

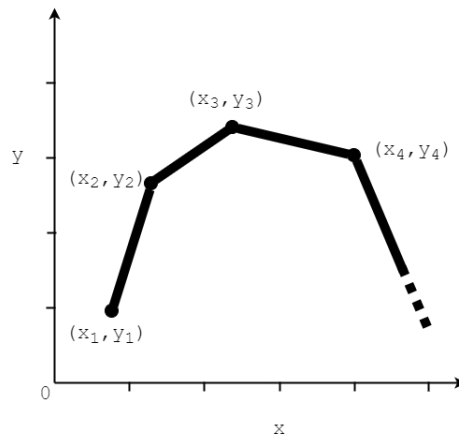


Figure 3.2: Example of a trait generated for binary images.

The selected deep architectures for the research experiments can technically process binary images as input. However, it is worth noting that these models were initially pre-trained on RGB colour images sourced from the ImageNet dataset [36]. Consequently, the architectures were designed to receive inputs with three colour channels (depth = 3). To align binary images with these architectural constraints, a requisite adjustment involved replicating the single channel across all three channels, thereby creating a three-channel image (RGB) for compatibility with the selected architectures. This adjustment, commonly called channel replication, is a standard practice to tailor models pre-trained on RGB images to process grayscale or binary image inputs effectively. These images predominantly unveil the morphological and personal details of the traced pattern, offering a focused depiction of the subject’s distinctive writing characteristics. The described approach ensures that the generated images meet neural network standards and convey a rich portrayal of the subject’s writing gesture by encapsulating information intricately tied to the form and structure of the trace.

3.2.2 RGB Synthetic Images

The generation process of the second category of synthetic images incorporates kinematic information. Analogous to the generation process for binary images, this procedure begins with the data stored in the *.csv* files obtained post-protocol execution. Specifically, these synthetic images are constructed by considering three key factors:

1. Similar to the binary case, the points (x_i, y_i) are considered vertices forming a polygonal structure approximating the original curve.
2. The triplet of values (z_i, v_i, j_i) is assigned as the RGB colour component for the i -th trait, delimited by the point pairs (x_i, y_i) and (x_{i+1}, y_{i+1}) .
3. Movements in air and on paper are directly recorded by the acquisition device.

This approach ensures a comprehensive integration of kinematic properties into the synthetic images, enriching their representation across various features and aspects of the writing dynamics derived from the outcomes of data acquisition and feature extraction phases. The triplet of values (z_i, v_i, j_i) encoded in the RGB colour channels contains dynamic information about the writing process’s actual motion, speed, and acceleration characteristics. In this way, it is possible to describe the dynamic aspects of the motion involved in handwriting rather than relying only on static or spatial characteristics. The generation of these triplets involves the following computations:

- z_i : Represents the pressure value at the point (x_i, y_i) directly acquired from the graphic tablet, assumed constant along the i -th trait.
- v_i : Denotes the velocity of the i -th trait, calculated as the ratio of the length of the i -th trait to the interval time of 5ms given by the tablet’s acquisition frequency of $200Hz$.

$$v_i = \frac{\Delta S}{\Delta t} = \frac{\sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}}{t_{i+1} - t_i} \quad (3.1)$$

- j_i : Represents the jerk of the i -th trait, defined as the second derivative of v_i .

$$j_i = \frac{d^2 v_i}{dt^2} \quad (3.2)$$

The triplets’ values (z_i, v_i, j_i) have undergone normalization to fit within the range $[0, 255]$, aligning with the standard colour scale. This normalization was achieved by considering these three quantities’ minimum and maximum values across the set. Figure 3.3 provides an illustrative example of a trait generated from these images. As for the binary images, the subject’s trace is also represented with a constant line width of 5 pixels. In this representation, the colour of the first trait corresponds to the triplet $(z = 166, v = 128, j = 184)$, while the colour of the second trait corresponds to the triplet $(z = 103, v = 171, j = 159)$. This normalization process ensures uniformity and compatibility with the standard colour scale conventions.

Notably, the z coordinate stored in the *.csv* files refers to pressure. Due to the acquisition tool’s capability to capture air movements within a 3cm range above the tablet surface, the z coordinate assumes a null value during in-air movements and takes a value greater than zero otherwise. This information allows to generate three different RGB image datasets:

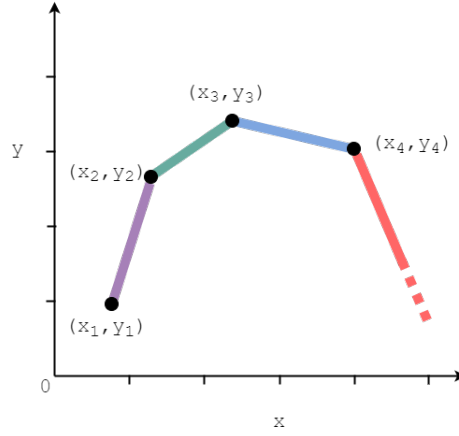


Figure 3.3: Example of colour encoding for the traits generation.

1. RGB on-paper, showing only on-paper movements;
2. RGB in-air, showing only in-air traits;
3. RGB in-air on-Paper, showing both in-air and on-paper traits.

This approach, indeed, not only captures the essential spatial and dynamic characteristics of handwriting but also incorporates insights into the patient’s hesitations during the writing process. The images encompass a comprehensive understanding of the physical act of writing and the pauses or hesitations that may indicate underlying conditions.

3.2.3 Multi-Channel

Multi-Channel (MC) TIFF images were generated to consolidate four representations (frames) of a single handwriting sample into a unified image file to enhance the dynamic information encoded. Each frame portrays a grayscale depiction of traits acquired through a process similar to that elucidated for RGB images. In detail, considering the points (x_i, y_i) as vertices of the polygon approximating the original curve, pixel values in each frame are assigned based on the following criteria:

- The first frame encapsulates the acceleration feature: the acceleration of the i -th trait is delineated as the derivative of v_i :

$$a_i = \frac{dv_i}{dt} \quad (3.3)$$

- The second frame encodes the jerk feature: the jerk of the i -th trait is articulated as the second derivative of v_i .
- The third frame encodes the velocity feature: the velocity of the i -th trait is computed as the ratio between the length of the i -th trait and the interval time of $5ms$, corresponding to the tablet’s acquisition period.
- The fourth frame encodes the pressure feature, presumed to be constant along the i -th trait.

Figure 3.4 shows an example of the generation of MC images.

As for the RGB images described in Section 3.2.2, also MC images were generated in three variants, containing only in-air or on-paper traits or depicting both of them. Moreover, each segment is replicated with a constant line width of 5 pixels.

3.2.4 Offline

Concerning the last category of images, these were straightforwardly derived through the segmentation of the original traits executed on the paper sheets during the performance of each task. This is why those images are named "offline"; the trait is original and not rebuilt by software, as is the case with synthetic binary or RGB images. The significance of studying offline images of handwriting in supporting the diagnosis of AD lies in the rich information embedded in the traces. Offline images encapsulate natural handwriting features, comprehensively reflecting the participant’s unique writing dynamics, pressure

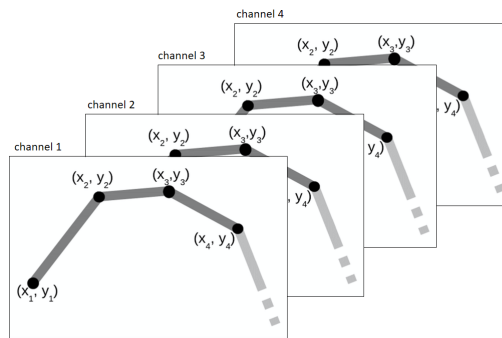


Figure 3.4: Example of encoding for the trait generation in an MC image.

variations, and movement characteristics. Analyzing these offline images offers a wide perspective on the subtleties of the writing process, potentially uncovering subtle changes or patterns that may serve as early indicators of cognitive decline associated with AD.

However, constructing a dataset can be challenging, particularly when dealing with paper documents. For each participant involved in the acquisition step, their task-execution paper sheets were stored. Offline images were generated by scanning these A4 paper sheets using the same scanner tool for all the documents. Thus, instead of relying on online data, images are sourced directly from the physical paper sheets utilized in the handwriting test, scanned and converted into *.tif* files. Each subject’s test on different fascicles yields a *.tif* image comprising multiple frames, one for each task.

Post-data gathering, the need for processing and manipulation arose to enhance readability and organize the dataset effectively. Starting from these considerations and the *.tif* images, various operations were executed to construct a well-structured offline dataset with high-quality images, optimizing the performance of the deep transfer learning strategy. This meticulous approach aims to ensure the dataset’s integrity and enhance the efficacy of subsequent analyses and classification tasks.

First, it was necessary to define and extract the Region Of Interest (ROI) from every frame of the *.tif* fascicle. In this context, the ROI was the portion of the image corresponding to the executed task. However, not all content in these images corresponded to the executed task. The elements in the images can be categorized into three types:

- **Artefacts:** Unnecessary features caused by mistakes during task execution or scanning operations. These artefacts need to be removed.
- **Command of the task:** Every image contains a command providing instructions. While this is part of the protocol, it is not useful and needs to be removed.
- **Region Of Interest:** The actual task execution is represented by the pen stroke drawn by the subject. This information is crucial.

A Python script was developed to automatically extract the ROI from each image. The script utilizes the OpenCV library to perform preprocessing operations and extract the region containing the task contours. The algorithm includes steps such as image blurring, conversion to grayscale, thresholding, dilation, and contour identification. Despite the algorithm’s effectiveness, challenges arose in some cases because of random strokes, unfulfilled requests, and varying image dimensions. These issues were particularly prominent in non-graphical tasks. A semi-automatic approach was adopted to address these challenges.

Every image underwent resizing to adapt to the distinct input formats required by the deep neural networks employed in the study. This resizing was executed with precision to ensure the centred alignment of the trace, thereby minimizing any potential loss of information attributable to zoom-in/out effects. This meticulous approach guarantees that the intrinsic details of the handwriting traces are preserved and effectively utilized in subsequent analyses, contributing to the robustness and accuracy of the diagnostic support system.

As a result of the whole process, in each image, the trace accurately represents the participant’s handwriting during the task, with pixel values capturing the natural shades of grey left by the ink on the paper. These pixel values and the traits’ width are influenced by both the pressure applied and the dynamics of the movements involved, creating an authentic representation of the handwriting process.

3.3 Features

The analysis of handwriting features has emerged as a promising avenue in supporting the diagnosis of AD. Researchers seek valuable insights into cognitive health by delving into the dynamic, kinematic, and personal aspects of handwriting. Dynamic features, evaluating the fluency and rhythm of writing, offer subtle indicators of cognitive function, with disruptions potentially signifying early cognitive decline.

Kinematic features, encompassing velocity, acceleration, and pen pressure, provide an understanding of fine motor control. Changes in these parameters may serve as sensitive markers for cognitive alterations associated with AD progression. Personal features, including individual writing styles, deviations from established norms, and the consistency of handwriting, contribute to a more personalized diagnostic approach. Advanced technologies, like digital tablets and specialized tools, facilitate the precise measurement and analysis of these features. Integrating ML algorithms further refine diagnostic accuracy by discerning patterns and abnormalities within the intricate data sets. As a non-invasive and cost-effective method, handwriting analysis holds promise for early AD detection, potentially allowing for timely intervention and improved patient outcomes. Ongoing research in this domain continues to unlock the full diagnostic potential of handwriting features in the context of AD. The data acquisition phase is crucial for any research involving handwriting analysis. In the context of this work, CSV files serve as a rich repository of crucial information obtained during the acquisition campaign. Starting from the valuable information stored in these files, extracting and computing interesting features to characterize handwriting is possible, providing a broad range of kinematic and dynamic information.

Various approaches for extracting additional features indicate an advanced methodology to refine and broaden the understanding of writing regarding specific traits. This diversification of approaches can include statistical methodologies, ML algorithms, or specific analyses for segmenting and interpreting handwriting features. In summary, the wealth of information collected during the data acquisition phase provides the basis for feature extraction and the flexibility to adopt diversified approaches for a more profound understanding of writing and its distinctive traits. The following sections define the approaches employed to compute handcrafted features in Section 3.3.1; and lognormal features in Section 3.3.2.

3.3.1 Handcrafted Features

The investigation into cognitive impairment among subjects involved an in-depth analysis of features extracted during the handwriting process. Following the completion of data cleaning and modelling to refine information stored in CSV files, the subsequent phase revolves around feature engineering. This process involves computing features from the .csv files and organizing them for classification. On-paper and in-air traits were processed, segmenting them into elementary strokes as single, connected, and continuous components of the handwritten trait. Identifying segmentation points crucially relies on events such as pen-down and pen-up occurrences, coupled with the zero-crossing of the vertical velocity profile. These segmentation points delineate each stroke's boundaries, capturing the handwriting's distinctive and coherent elements.

This approach ensures precision in the segmentation process and aligns with the fundamental notion that strokes encapsulate the essence of a continuous handwritten sequence. For each stroke, feature values were computed and averaged across all strokes about a specific task. Considering the observed differences in patients' motor performance between in-air and on-paper traits, each feature was calculated separately for these conditions. In particular, four groups of features were computed:

- In air: considering only in air strokes, specifically those recorded when the pen tip is elevated from the surface within the permissible maximum distance. These movements signify motor planning activities associated with positioning the pen tip between consecutively written traits;
- On paper: considering only on-paper strokes, encompassing the pen-down and the subsequent pen-up phases.
- In air on paper: putting together in-air and on-paper features. Consequently, the total number of features equals the sum of the in-air and on-paper features;
- All: considering all the traits, independently of whether they are executed in the air or on paper.

Additionally, personal factors such as the subjects' age, education level and gender were incorporated into the final features. It is worth noticing that every task corresponds to a dataset of features. Table 3.3 lists the computed features, providing the corresponding description and the type.

#	Name	Description	Type
1	Duration	Time interval between the first and the last points in a stroke	D
2	Start Vertical Position	Vertical start position relative to the lower edge of the active digitizer area	S
3	Vertical Size	Difference between the highest and lowest y coordinates of the stroke	S
4	Peak vertical velocity	Maximum value of vertical velocity among the points of the stroke	D
5	Peak vertical acceleration	Maximum value of vertical acceleration among the points of the stroke	D
6	Start horizontal position	Horizontal start position relative to the lower edge of the active tablet area	S
7	Horizontal size	Difference between the highest (rightmost) and lowest (leftmost) x coordinates of the stroke	S
8	Straightness error	It is calculated estimating the length of the straight line, fitting the straight line, estimating the (perpendicular) distances of each point to the fitted line, estimating the standard deviation of the distances and dividing it by the length of the line between beginning and end	D
9	Slant	Direction from the beginning point to endpoint of the stroke, in radiant	S
10	Loop Surface	Area of the loop enclosed by the previous and the present stroke	S
11	Relative initial slant	Departure of the direction during the first 80 ms to the slant of the entire stroke.	D
12	Relative time to peak vertical velocity	Ratio of the time duration at which the maximum peak velocity occurs (from the start time) to the total duration	D
13	Absolute size	Calculated from the vertical and horizontal sizes	S
14	Average absolute velocity	Average absolute velocity computed across all the samples of the stroke	D
15	Road length	length of a stroke from beginning to end, dimensionless	S
16	Absolute y jerk	The root mean square (RMS) value of the absolute jerk along the vertical direction, across all points of the stroke	D
17	Normalized y jerk	Dimensionless as it is normalized for stroke duration and size	D
18	Absolute jerk	The Root Mean Square (RMS) value of the absolute jerk across all points of the stroke	D
19	Normalized jerk	Dimensionless as it is normalized for stroke duration and size	D
20	Number of peak acceleration points	Number of acceleration peaks both up-going and down-going in the stroke	S
21	Pen pressure	Average pen pressure computed over the points of the stroke	D
22	#strokes	Total number of strokes of the task	S
23	Sex	Subject's gender	P
24	Age	Subject's age	P
25	Work	Type of work of the subject (intellectual or manual)	P
26	Education	Subject's education level, expressed in years	P

Table 3.3: Feature list. Feature types are dynamic (D), static (S) and personal (P).

3.3.2 Lognormal Features

The Sigma-Lognormal model served as the foundation for the computation of two distinct sets of lognormal features. These sets were derived through a comprehensive model analysis, capturing multiple aspects of movement characteristics. The computation process involved intricate calculations based on the lognormal parameters defined by the model. This dual set of lognormal features offers a better understanding of the underlying dynamics, allowing for a comprehensive examination of movement patterns and behaviours. Utilizing these feature sets adds depth to the analysis, providing researchers with a more refined toolkit for studying and characterizing movements within the Sigma-Lognormal framework.

Sigma-Lognormal Model

Grounded in lognormal movement decomposition, numerous studies have explored the normative range of variations in lognormal parameters, providing insights into the ideal characteristics of a movement [108]. The Kinematic Theory, employed to parameterize human movement velocity and trajectory, has spurred the development of diverse algorithms, such as Robust XZERO [98, 42] and IDeLog [52]. For this investigation, I used the IDeLog algorithm [52]. The Sigma-Lognormal model conceptualizes the velocity of each simple, fast movement primitive as a lognormal function (Λ), with each velocity peak between two-speed minima modelled by a lognormal. The lognormal parameters, t_{0_j}, μ_j and σ_j^2 , are calculated by minimizing the error between the velocity profile and the lognormal obtained through successive interactions. This includes the comparison between the original trajectory profile and the reconstructed one. The lognormal function defining each velocity peak, termed a "simple movement" or "stroke," is expressed as:

$$v_j(t; t_{0_j}, \mu_j, \sigma_j^2) = D_j \Lambda(t; t_{0_j}, \mu_j, \sigma_j^2) = \frac{D_j}{\sigma_j \sqrt{2\pi}(t - t_{0_j})} \exp\left\{\frac{[-\ln(t - t_{0_j}) - \mu_j]^2}{2\sigma_j^2}\right\} \quad (3.4)$$

where time t , amplitude D_j , time of occurrence t_{0_j} , time delay μ_j , and response time σ_j operate on a logarithmic time scale. In complex movements, characterized by a succession of simple movements or strokes, as illustrated in Figure 3.5, the velocity profile $v_n(t)$ emerges from the time superposition of the M preceding lognormals.

$$v_n(t) = \sum_{j=1}^M v_j(t; t_{0_j}, \mu_j, \sigma_j^2) = \sum_{j=1}^M D_j \begin{bmatrix} \cos(\Phi_j(t)) \\ \sin(\Phi_j(t)) \end{bmatrix} \Lambda(t; t_{0_j}, \mu_j, \sigma_j^2) \quad (3.5)$$

where $\Phi_j(t)$ is the angular position given by:

$$\Phi_j(t) = \Theta_{s_j} + \frac{(\Theta_{e_j} - \Theta_{s_j})}{2} \left[1 + \operatorname{erf}\left(\frac{\ln(t - t_{0_j}) - \mu_j}{\sigma_j \sqrt{2}}\right)\right] \quad (3.6)$$

being Θ_{s_j} and Θ_{e_j} the starting and the end angular direction of the j^{th} simple movement or stroke. This comprehensive approach outlines the Sigma-Lognormal model and sheds light on the intricacies of parameter calculation and the dynamic interplay between velocity and trajectory profiles in both simple and complex movement scenarios. The following sections describe two approaches to feature engineering from the sigma-lognormal extracted parameters.

Lognormal Features: First set

The first feature engineering process enabled the identification of a set of features strategically chosen to discern the handwriting patterns of individuals affected by AD from those of the healthy controls group. The Sigma-Lognormal model, detailed in Section 3.3.2, was applied to obtain from every task execution the set of sigma-lognormal parameters. Given the information stored in the *.csv* files, the outcome of this application was the decomposition of each task into a vector summation of simple time-overlapped movements, facilitating the extraction of Sigma-Lognormal parameters $P_j = [D_j, t_{0_j}, \mu_j, \sigma_j, \Theta_{s_j}, \Theta_{e_j}]$. Specifically, for every point (x, y) recorded during task execution, multiple overlapping lognormals were identified, with their respective parameters and the percentage of contribution meticulously stored for each point.

The term "First lognormal" denotes the lognormal that primarily contributes to a specific point. Subsequently, fourteen features were computed with the Sigma-Lognormal parameters obtained for every task and participant, as detailed in Table 3.4. These features encapsulate essential characteristics derived from

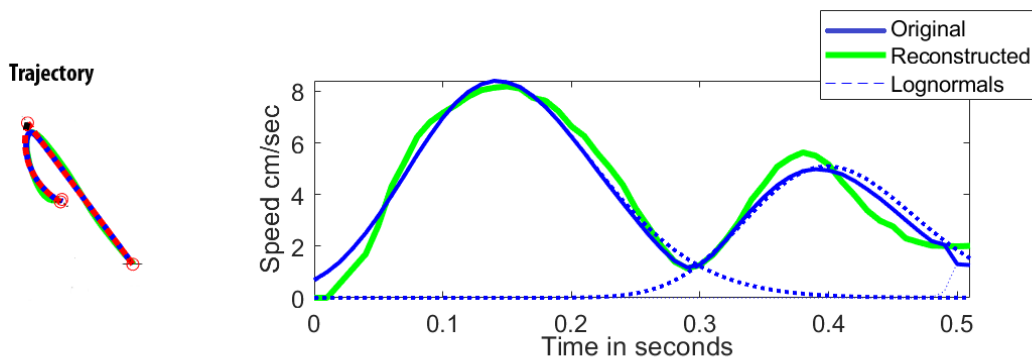


Figure 3.5: An example of lognormal.

Name	Description
<i>num_seg</i>	Total number of segments generated by the execution of the task
<i>avg_log</i>	Average of the number of overlapped lognormals for every point
<i>tot_log</i>	Total number of lognormals extracted from the entire trace of the task
<i>avg_D</i>	Average of D parameter of the first lognormal for every point
<i>D_max</i>	Max of D parameter found among the first lognormals of all the points
<i>P_first_log</i>	Average of the percentage of contribution of the first lognormal for all the points
<i>σ_stability</i>	Variance of the sigma parameter of the first lognormal for all the points
<i>diff_logs</i>	Average of the differences between the percentage of contribution of the first and the second lognormal on all points
<i>var_log</i>	Variance of the percentage of contribution of the first lognormal on all points
<i>avg_t_o</i>	Average of the t_o parameter of the first lognormal on all points
<i>avg_σ</i>	Average of the σ parameter of the first lognormal on all points
<i>avg_μ</i>	Average of the μ parameter of the first lognormal on all points
<i>avg_Θ_s</i>	Average of the Θ_s parameter of the first lognormal on all points
<i>avg_Θ_e</i>	Average of the Θ_e parameter of the first lognormal on all points

Table 3.4: Summary of computed Lognormal Features, first set.

the decomposition process, offering a comprehensive quantitative representation of handwriting traits for further analysis and discrimination between individuals affected by AD and healthy individuals.

This first set of Lognormal Features serves as a key indicator for subsequent analyses aimed at characterizing and distinguishing handwriting patterns within the studied context.

Lognormal Features: Second Set

The second phase of lognormal feature engineering, centred around the Sigma-Lognormal model, incorporated insights from [146, 40] and various studies on the normative range of lognormal parameters [108], offering an appropriate understanding of the characteristics of an ideal movement through lognormal movement decomposition. This comprehensive approach aimed to extract diverse features from those presented in Section 3.3.2, providing an exhaustive characterization of individual handwriting traits. The goal was to accentuate potential disparities between the handwriting of individuals affected by AD and that of healthy controls.

As for the previous set of lognormal features computed, initiating this process involved the application of the Sigma-Lognormal model, outlined in Section 3.3.2, to the data acquired by the procedure detailed in Section 3.1. Consequently, each task underwent decomposition into a vector summation of simple time-overlapped movements, with each associated lognormal function generating a distinct set of Sigma-Lognormal parameters. To refine the dataset for analysis, only points corresponding to the initial and final pen-down events were processed for each task execution. This step aimed to exclude extraneous movements recorded when the person approached or departed from the paper, focusing solely on the genuine handwriting gestures under examination.

After extracting Sigma-Lognormal parameters, three distinct groups of features were computed. These features encompass various aspects and measurements related to the execution of handwriting, contributing to a multifaceted analysis of individual writing styles and behaviours. They are divided into three categories:

1. Temporal features: related to the temporal aspects of the execution. Among them is the total time, representing the overall time taken to execute a task. The contact time refers to the duration during which movements were executed without losing contact with the tablet surface, where the pen remained within a maximum distance of 3cm. The remaining time, denoted as losing time, accounts for the instances when the pen exceeded this threshold. The summation of the contact and the losing time gives the total time. Some features within this category also correlate with the number of lognormals identified in the reconstructed velocity profile, providing proportional insights into task execution.
2. Signal-to-Noise Ratio (SNR): Indicates the quality of the reconstructed trace (SNR_t) and velocity profile (SNR_v) derived from the Sigma-Lognormal model. Features associated with SNR offer valuable information regarding the fidelity of the reconstructed data.
3. Geometrical features: These features are related to the geometric shapes of the reconstructed speed profile. Insightful for comprehending movement velocity, stability, and fluency, these features are derived from lognormal parameters like D and σ , as well as geometrical shapes (area, height, and width) of lognormals within the reconstructed velocity profile. Specifically, "area" refers to the overlapping area between consecutive lognormals, "height" represents the maximum, and "width" denotes the base of a lognormal function, as outlined in [40].

Tables 3.5 and 3.6 present a comprehensive display of the computed features, each denoted by the nomenclature $f\#$ and accompanied by its corresponding explanation. This crucial step aims to ascertain the potential for estimating AD by leveraging features extracted through the Sigma-Lognormal model applied to handwriting movements. In addition to the previously discussed feature groups, my experiments incorporated personal features, including age, gender, education, and type of profession, recognizing the potential impact of Alzheimer's on various facets of an individual.

Temporal features bear significance in this exploration, as individuals affected by AD may exhibit prolonged task execution times and increased losing time, representing instances when the pen is lifted too far from the tablet, possibly due to fatigue or distraction. Parameters such as the number of lognormals generated from the velocity profile and the count of segments, where each corresponds to an entire trace acquired without losing contact, are also considered. Anticipated results suggest elevated values for all temporal features among individuals affected by AD. The signal-to-noise ratio serves as a critical metric

for assessing reconstruction quality. When normalized by the number of lognormals, it provides insights into the fluency of movements, a key aspect of handwriting analysis [146]. Geometrical features, derived from sequences involving overlapping areas, heights, and widths of lognormal functions, offer valuable insights into handwriting fluency. Larger overlapping areas signify smoother handwriting, height correlates with speed, and width indicates movement pace. Geometrical features related to lognormal parameters D and σ provide information on the lognormal distance covered in kinematic space and the lognormal response time. Understanding the dynamics of these measures during a handwriting task or establishing correlations with temporal features can yield valuable information, contributing to our comprehension of the impact of Alzheimer’s on handwriting behaviours.

Features			
TEMPORAL		SNR	
f1	number of lognormals	f15	mean(SNRt)
f2	number of segments	f16	std(SNRt)
f3	task total time	f17	mean(SNRv)
f4	contact time	f18	std(SNRv)
f5	losing time	f19	sum(SNRt)/f1
f6	standard deviation of seg. time	f20	f15/f1
f7	f3/f2	f21	f16/f1
f8	f3/f1	f22	sum(SNRv)/f1
f9	f4/f2	f23	f17/f1
f10	f4/f1	f24	f18/f1
f11	f5/f2		
f12	f5/f1		
f13	mean(number of log.s per seg.)		
f14	std(number of log.s per seg.)		

Table 3.5: Temporal and SNR related features.

Features					
GEOMETRICAL					
f25	std(areas)	f39	mean(areas)/f25	f53	dif(widths)/1
f26	std(heights)	f40	mean(heights)*exp(f26)	f54	dev(widths)/1
f27	std(widths)	f41	mean(heights)*ln(f26)	f55	'seg_difA_div_nlog'
f28	sum(areas)/f3	f42	mean(heights)/f26	f56	'std_seg_difA_div_nlog'
f29	sum(areas)/f4	f43	mean(widths)*exp(f27)	f57	dif(sigma)/f1
f30	sum(areas)/f1	f44	mean(widths)*ln(f27)	f58	std(sigma)/f1
f31	sum(heights)/f3	f45	mean(widths)/f27	f59	dif(sigma)/f4
f32	sum(heights)/f4	f46	f25/f4	f60	std(sigma)/f4
f33	sum(heights)/f1	f47	f26/f4	f61	dif(D)/f1
f34	sum(widths)/f3	f48	f27/f4	f62	std(D)/f1
f35	sum(widths)/f4	f49	dif(areas)/f1	f63	dif(D)/f4
f36	sum(widths)/f1	f50	std(areas)/f1	f64	std(D)/f4
f37	mean(areas)*exp(f25)	f51	dif(heights)/f1		
f38	mean(areas)*ln(f25)	f52	std(heights)/f1		

Table 3.6: Geometrical features.

Chapter 4

Results and Findings

Delving into the insightful outcomes of this research work, the Results and Findings section presents a comprehensive analysis of the experimental methods deployed, leveraging different data sources as outlined in Chapter 3. This chapter aims to provide a wide understanding of fundamental patterns, trends, and noteworthy discoveries essential for developing a robust system supporting the diagnosis of NDs through handwriting analysis.

The following sections highlight the implications and contributions of this research, facilitating a comparative assessment of various AI-based mechanisms and techniques. Through a wide investigation and analysis, the following sections represent a detailed exploration of this research endeavour, discussing results and consequential findings.

4.1 Baseline Experimental Setting

During my research, I conducted several experiments considering different configurations of tasks and data. Many experimental settings have been deployed but share a common baseline architecture, allowing me to compare the obtained results and derive interesting findings. The implemented baseline experimental architecture is visually presented in Figure 4.1. It comprises four steps:

1. Data acquisition: it resumes the data acquisition and image generation phase described in Section 3.1.
2. Feature extraction/engineering: this step shows the features used. They usually are features from a feature engineering process, like handcrafted or lognormal features; instead, for images, features are automatically extracted by CNNs.
3. Classification: the classification step involves using ML algorithms and a fully connected classifier.
4. Combining rule: a combination rule, like a majority vote, is usually applied according to tasks, classifiers or deep networks.

The figure shows a baseline, meaning adjustments and changes were adopted according to the specific experiment, though some choices remained the same.

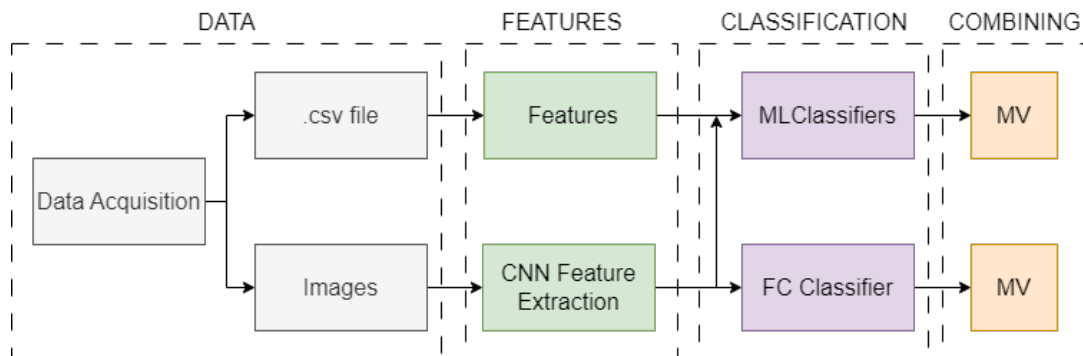


Figure 4.1: Basic experimental setting.

Following the data acquisition, many data types were produced and can be divided into tabular features and images. Once the data collection was complete, the images were utilized to input into four well-established CNNs. The initial part of these architectures served as a feature extractor, enabling the extraction of features from images, resulting in a feature vector for each image. After selecting four CNN models, four sets of deep features were obtained for each task and image type.

For the classification step, two distinct approaches were implemented. In the first approach, standard ML classifiers were employed, utilizing every feature, whether from images or not. Conversely, the second approach utilized the classifier made of fully connected layers of the CNN. In the first approach, many well-established classification schemes were considered: RF [11], DT, MLP, SVM [17], K-Nearest Neighbors (KNN), LR, Gradient Boosting (GB) and Extreme Gradient Boosting (XGB). These represent diverse model types, with RF being an ensemble of DT, MLP being a connectionist network, KNN being an instance-based non-parametric regression algorithm, and SVM being kernel-based. In detail, thirty runs were performed for every ML algorithm, and the final result was computed as an average. This is a very common practice, as averaging the results over multiple runs provides a more stable and reliable estimate of a model’s performance and helps to assess the robustness and generalization ability of the model.

Regarding the second classification approach, the architecture of the classifier included two hidden layers with 2048 neurons and a dropout between them, named Fully Connected (FC) classifier from now on. The FC classifier was exclusively applied to the feature sets directly obtained from the convolutional layers. Regarding CNNs, four models were employed: VGG19 [123], ResNet50 [65], InceptionV3 [131], and InceptionResNetV2 [130]. These models underwent enhancements over the years by introducing new structural elements and increasing the number of layers. This augmentation led to a parameter rise, ranging from twenty-five million in VGG19 to sixty-two million in InceptionResNetV2, as detailed in Table 4.1. The table also shows the input size required and the output size, which refers to the dimension of the feature vector at the bottleneck.

Table 4.1: Number of parameters and input/output size of the CNN used in the experiments.

Model	Parameters	Input size	Output size
VGG19	25M	256x256	512
ResNet50	32M	224x224	2048
InceptionV3	30M	299x299	2048
InceptionResNetV2	62M	299x299	1536

All the CNN architectures adopted in this study consist of two main parts: the convolutional part designed for feature extraction (Feature Extractor (FE)) directly from input images and the classification part (Classifier (C)). CNN models are known to be very data-hungry, as not only the quality but also the quantity of data can affect their performance. A transfer learning technique was considered because the available data were insufficient to train those models. Every model was pre-trained on the public dataset ImageNet [36], but this training involved only the FE part, as the weights of the C part were frozen. Following a re-training process involved both parts, FE and C, using the fine-tuning (FT) approach. Notably, all models’ original C layers were replaced with the FC classifier, specifically adapted for the context, i.e., classifying two classes (healthy control or patient).

Following the training phase, the CNN networks served a dual purpose: deep feature extraction and classification using the final fully connected layers (the classifier section of the deep network). Deep features were obtained by pruning the network after the FE, often called the "bottleneck". Each model generated a flattened vector of varying size, as indicated in Table 4.1. A preliminary experimental phase was undertaken to assess the CNN architectures, involving minimising the accuracy of all models. The selected settings and hyper-parameters included Stochastic Gradient Descent (SGD) with a learning rate of 0.001 and momentum of 0.9 as the optimization method, categorical cross-entropy as the loss function, a batch size of 16, and a maximum of 2,000 epochs. The training process employed a patience value of 200, whereby if the validation accuracy did not improve for 200 epochs, the training was interrupted. The principal evaluation metric for performance was accuracy, but many other metrics were computed according to the experiment, like True Positive Rate (TPR), True Negative Rate (TNR), False Negative Rate (FNR) and Area Under the Curve (AUC). The training phase incorporated a validation set to mitigate the undesired overfitting phenomena and followed a 5-fold cross-validation strategy. Each fold utilized a test set comprising 20% of the images, with the remaining images divided into a 70% training set and a 10% validation set. It is important to note that the images in the validation set were randomly

selected from folds that were not used as the test set. The experiments were conducted on a computing system featuring an Intel Core i7-7700 CPU @3.60GHz with 32GB of RAM and a GPU Titan Xp, and Keras 2.2.2 and TensorFlow 1.10.0 were utilized as the software framework.

The following subsections comprehensively describe each experiment, discussing the related results.

4.2 Comparison among Binary, RGB on-paper and Handcrafted Features on Graphic Tasks

The current section shows the development of a system for AD diagnosis based on dynamic features like speed and jerk acceleration and morphological information from handwriting [22].

This experiment’s contributions include assessing a diagnostic system’s impact by investigating the combined use of shape and dynamic features. Different models of CNNs were tested as automatic feature extractors, comparing their performance with traditional ML algorithms. The following subsections show the data used, the details of the experimental framework implemented, and the performance obtained.

Data

In order to test the experimental setting, participants’ handwriting was examined as they drew lines or circles to predict their cognitive status, focusing on evaluating fine motor control without requiring cognitive or memory skills. The handwriting tasks comprising the experimental protocol are thoroughly described in Section 3.1. Four tasks out of 25 were selected for this experiment; in particular, they belong to the subset of graphic tasks from the second to the fifth. The initial two tasks involved connecting two points 5cm apart with a straight continuous line either horizontally (task 2) or vertically (task 3), repeated four times. These tasks aimed to assess elementary motor functions, with horizontal movements emphasizing arm movements and fixed finger positions, while vertical movements required smaller finger and wrist motions. Drawing a single continuous line four times also assessed long-term motor planning, a function often compromised in individuals with cognitive impairments. The subsequent two tasks involved retracing a 6cm (task 4) or 3cm (task 5) wide circle continuously four times. These tasks focus on demonstrating the continuity of the line by repetitively retracing a circular shape of varying dimensions. The consistency and distance from the background shape traced were indicators of cognitive deterioration. Additionally, these tasks allowed evaluation of the automaticity, regularity and coordination of the sequence of movements.

Three data types were selected to understand the discriminatory power of different handwriting aspects: Binary images, RGB on paper images and handcrafted features. The generation of the synthetic images is exhaustively described in Sections 3.2.1 and 3.2.2, while the computation of handcrafted features is depicted in Section 3.3.1. In particular, binary images represent an approximation of the real handwritten trace executed by a person. RGB on paper images approximate the real handwritten trace and encode kinematic information in the colour channels, i.e., pressure, velocity, and jerk. Finally, the handcrafted features refer, in this case, to the on-paper acquired points and are grouped into static, dynamic and personal characteristics. These representations were organized into three datasets, and their performance was evaluated using various classification schemes. Examples of RGB on paper images of the selected tasks are illustrated in Figure 4.2.

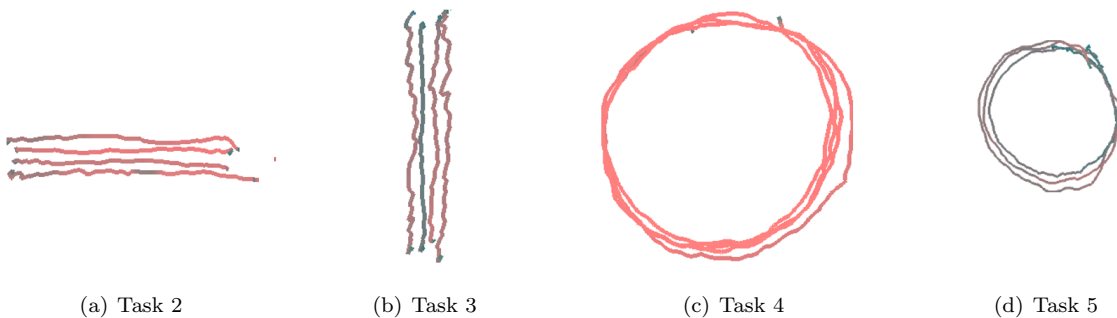


Figure 4.2: Examples of RGB images generated from the online handwriting data acquired from a participant while performing the selected graphic tasks

Experimental Setting

The implemented comprehensive system architecture is exactly the baseline architecture shown in Figure 4.1 and described in Section 4.1. Following acquisition, three types of data were procured: handcrafted features, synthetic binary images, and synthetic RGB on paper images. Once the data collection was complete, the images were utilized to input into the four CNNs selected. I used these models to extract features from images, so four sets of deep binary features and four sets of deep RGB features were obtained for each task, amounting to a total number of nine sets of features to be evaluated.

For the classification step, two distinct approaches were implemented. In the first approach, four well-established classification schemes were considered: RF, MLP, SVM and KNN. Regarding the second approach, I used a fully connected classifier, which was exclusively applied to the feature sets directly obtained from the convolutional layers.

Experimental Results

In the classification step, four standard classification schemes (RF, MLP, KNN and SVM) and a modified fully connected network FC were employed, as mentioned in the previous section. The parameter values used in this experiment are detailed in Table 4.2.

Table 4.2: Values of the ML classifiers hyperparameters used in the experiments.

Classifier	Hyperparameter	Value
RF	trees	100
K-NN	K	3
MLP	Learning rate	0.3
	Momentum	0.2
	Hidden Neurons	$(\#features + \#classes)/2$
SVM	Epochs	500
	Kernel	RBF
	C	1.0
	γ	0.5

In the case of the FC classifier, since its training required substantial resources and time, the FC results were averaged based on accuracy obtained from the 5-fold cross-validation strategy; no multiple runs were performed. The aforementioned feature extraction procedure was applied to data from four handwriting tasks described in Section 4.2. Regarding deep features, considering that their extraction requires a training phase, the approach involved the utilization of "test deep features" to prevent bias. In practical terms, the feature vector obtained from the feature extraction part was employed for each sample when that sample was in the test set.

Three experiments were conducted to assess the system's effectiveness according to the procedures outlined in the previous sections. First, features extracted from binary images containing only an approximation of the real morphological traits were tested. Then, it was evaluated whether there was a performance improvement, considering features extracted from RGB on-paper images containing dynamic information with respect to the binary ones. The third set involved comparing the classification performance of the proposed approach with that achieved using handcrafted features, only concerning the dynamic and static characteristics of the movement, without considering the information related to the shape. Finally, based on the analysis of the results, an additional set of experiments was executed to verify if performance could be enhanced by combining the classifier responses for each subject across all tasks or by fusing handcrafted and deep features.

Regarding the first experiment, Table 4.3 displays the classification outcomes on the binary images for individual tasks using the five classifiers applied to the deep features extracted by the CNNs. The table provides a comprehensive account of the results, focusing on accuracy, which quantifies the number of correct predictions (for both patients and healthy controls) relative to the total number of subjects in consideration. A notable observation from the table is the substantial variability in performance across different classifiers when extracting features with the same CNN for each task. This pattern is also evident in the performance disparities observed across tasks. Looking at the overall results it is also possible to point out that features extracted from some CNNs outperform others. For each task, bold values highlight the best accuracy. Thus, it is easy to notice that the best-performing configurations are

given by features extracted from InceptionV3 and InceptionResNetV2 models and classified by SVM, RF and MLP algorithms. The worst result is usually obtained by the FC classifier, which also shows a greater standard deviation, but this can be explained by the fact that no multiple runs were performed. SVM obtains the best result with an accuracy of 70.8% on the fifth task with features extracted from InceptionV3.

	Task 2		Task 3		Task 4		Task 5	
	ACC	SD	ACC	SD	ACC	SD	ACC	SD
VGG19								
RF	64.8	2.8	59.4	2.5	68.0	1.8	67.1	2.0
K-NN	62.2	1.8	56.9	3.0	64.0	2.2	64.8	2.6
SVM	60.3	2.9	58.8	2.5	68.6	1.6	66.1	1.5
MLP	57.4	2.5	57.8	2.9	61.7	2.9	61.9	3.2
FC	50.1	8.1	48.1	7.6	42.6	6.5	50.6	4.4
ResNet50								
RF	62.3	2.8	60.0	2.2	68.0	2.0	67.7	2.0
K-NN	58.5	2.2	53.9	2.7	59.5	1.8	59.5	1.8
SVM	61.3	2.8	58.0	1.9	65.0	1.7	68.9	1.5
MLP	52.8	3.0	51.9	2.7	55.9	5.4	53.7	0.7
FC	45.2	8.4	50.9	10.3	47.2	6.7	47.0	7.8
InceptionV3								
RF	65.5	2.3	55.1	3.0	68.9	2.0	68.4	2.2
K-NN	63.1	1.9	51.3	2.8	66.4	2.1	63.2	2.7
SVM	66.2	2.2	56.0	3.0	69.5	1.3	70.8	1.5
MLP	67.4	2.1	49.6	3.2	61.0	3.8	63.4	2.5
FC	52.4	7.6	51.9	11.3	44.4	9.1	45.2	11.8
Inc.ResNetV2								
RF	64.9	2.8	61.5	2.2	67.8	1.5	66.5	2.7
K-NN	59.1	2.5	57.6	2.6	61.2	1.8	55.5	2.1
SVM	66.0	1.9	60.6	2.3	67.7	2.0	65.9	2.0
MLP	66.8	3.5	61.2	4.1	63.5	3.0	58.3	1.5
FC	49.1	10.9	48.2	5.1	46.8	8.4	49.6	10.6

Table 4.3: Classification results achieved on deep features extracted by binary images.

Concerning the second experiment, Table 4.4 shows the performance obtained on the RGB on paper images. Here, the same observations can be applied to the variability of the outcomes considering different configurations of CNNs and classifiers. Regarding the classification algorithms, RF and SVM consistently outperform the others in most cases. Concluding these results, it is possible to make the following considerations: the ensemble-based strategy of RF, along with the kernel-based approach tailored for two-class problems, yields the best performance, and the effectiveness of features extracted by CNNs remains independent of the classifier used, enabling even better results compared to those achieved by the fully connected layer of the CNN. Concerning the CNNs, InceptionResNetV2 generally achieves the best results, except for task 5, where InceptionV3 outperforms. This outcome may be attributed to the complexity of task 5, allowing for better discrimination between healthy controls and patients. Consequently, a simpler CNN like InceptionV3, facilitating more practical training on available data, yields superior results. The best result is given by MLP on the second task, reaching an accuracy of 74.6% with features extracted from InceptionResNetV2.

In summarizing the outcomes presented in Table 4.4, Figure 4.3 shows two vertical bar graphs. The objective is to quickly discern whether a particular CNN or classifier outperforms the others. In particular, Figure 4.3 (a) displays, for each task, the mean accuracy of each classifier, averaged across the features of the four CNNs. This plot confirms what was said previously, that RF and SVM are the best classifiers for the features in the exam. Similarly, 4.3 (b) shows the mean accuracy obtained with the features of each CNN for each task, averaged across the results of all the classifiers. This plot shows that deeper CNN models perform better in this case. Upon examination of the figure, it is evident that task 3 yielded the lowest performance, though it can be considered very similar to the second. This outcome is understandable, considering task 2 imposes a greater motor load than task 3. The nature of task 2, which involves executing the activity without moving the arm but with small movements of both fingers and wrist, makes it comparatively more challenging than task 3.

	Task 2		Task 3		Task 4		Task 5	
	ACC	SD	ACC	SD	ACC	SD	ACC	SD
VGG19								
RF	67.6	2.3	62.8	2.8	70.0	2.4	68.2	1.6
K-NN	64.3	2.2	56.7	2.7	61.0	2.0	61.6	2.1
SVM	61.7	2.4	55.4	2.0	67.6	1.4	66.9	1.3
MLP	63.5	3.3	56.6	2.8	58.1	3.4	58.2	3.4
FC	64.1	15.2	59.0	5.1	70.7	7.5	64.7	11.0
ResNet50								
RF	69.7	2.1	60.3	2.4	71.2	2.4	71.7	2.2
K-NN	66.8	2.1	57.0	2.4	66.3	2.6	57.9	1.4
SVM	72.0	1.7	58.9	1.9	67.3	2.2	72.6	1.6
MLP	53.8	2.1	50.6	3.4	60.3	8.7	52.9	3.5
FC	61.1	8.0	53.5	9.5	68.8	7.7	64.4	12.7
InceptionV3								
RF	68.6	2.1	54.5	3.8	70.5	2.5	71.7	1.9
K-NN	66.5	1.8	56.0	2.8	62.7	2.5	66.8	2.0
SVM	71.1	1.5	55.4	2.7	71.1	2.0	73.1	1.9
MLP	67.5	2.5	53.5	1.5	63.9	1.7	62.8	4.7
FC	64.2	9.0	57.1	4.8	68.0	6.6	65.7	11.1
Inc.ResNetV2								
RF	70.4	2.4	65.4	2.0	73.0	2.1	69.2	2.3
K-NN	68.4	1.8	62.8	2.4	65.0	2.0	64.6	2.0
SVM	71.1	2.2	61.2	2.4	71.8	1.6	67.5	1.4
MLP	74.6	2.3	63.3	1.5	70.4	2.1	52.6	2.1
FC	60.0	15.9	62.7	7.3	68.0	10.6	65.3	10.1

Table 4.4: Classification results achieved with deep features extracted by RGB images.

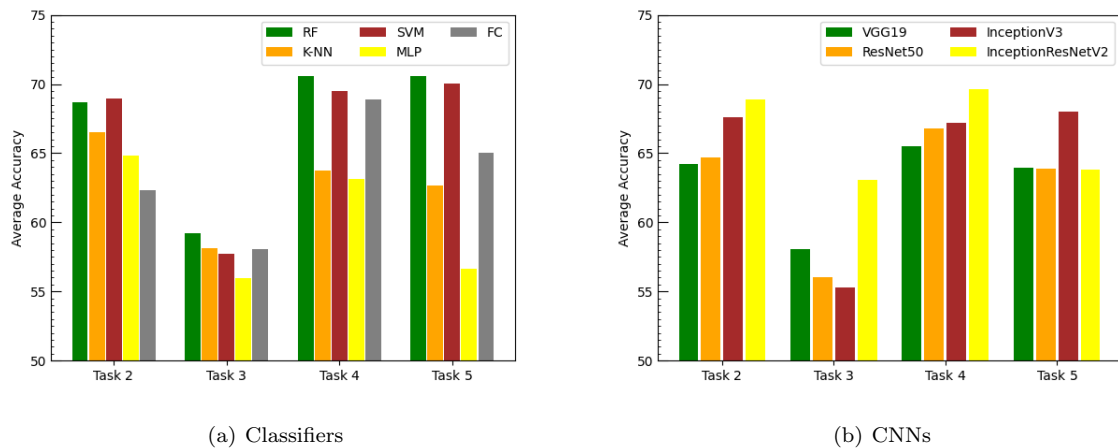


Figure 4.3: Average accuracy achieved by the classifiers (a) and the CNNs (b) for features extracted by RGB images.

	Task 2	Task 3	Task 4	Task 5
ACC	74.65	65.42	73.04	73.14
ERR	25.35	34.58	26.96	26.86
TPR	72.36	63.46	72.19	67.05
TNR	75.08	65.06	73.05	81.42
FNR	27.64	36.54	27.81	32.95
AUC	0.81	0.71	0.78	0.75

Table 4.5: Summary of the best experiments obtained on RGB images. The table also shows, for each task, the other considered performance measures

To have a closer look at the results obtained, Table 4.5 displays a set of evaluation metrics for experiments providing the best results on RGB images for each task (Table 4.4). The set comprises accuracy, error, sensitivity (TPR), specificity (TNR), FNR and the AUC. FNR is a very important metric in the medical field, particularly in the context of AD, as false negatives occur when a diagnostic test incorrectly indicates that a patient does not have AD. If the FNR is high, it means that a significant number of individuals with Alzheimer’s may go undetected, delaying appropriate care and support. Every value is expressed in percentage except the AUC. The analysis reveals that task 2 performs best as it reaches the highest accuracy and AUC values, affirming the effectiveness of predictions. Moreover, this task reports the highest TPR and the lowest FNR, meaning that most patients are correctly classified. Task 5, while displaying the highest accuracy, exhibits a lower AUC value and a significantly lower TPR than TNR, indicating that a higher number of patients go unidentified.

In almost every case, the performance obtained from RGB images outperforms that obtained from binary images. This was expected as RGB images contain more information than binary ones, thanks to the dynamic information encoded in their colour channels.

Figure 4.4 compares these approaches. A vertical bar graph was created for each CNN to illustrate the accuracy per task, computed by averaging the results from the five classifiers. The graphs indicate that performances on RGB images consistently surpass those achieved with binary images. This finding reaffirms that relying solely on shape information in binary images is inadequate for distinguishing patients’ handwriting from that of the control group.

For the third series of experiments, Table 4.6 presents the accuracy achieved with handcrafted features for each task and the four classification algorithms selected. The table shows that RF and KNN deliver the best performance when utilizing handcrafted features. This reinforces the efficacy of the RF ensemble-based strategy and highlights the satisfactory results obtained by KNN, contrasting with the deep features scenario. In fact, unlike in the deep features case, the KNN algorithm effectively estimated the probability distributions represented by the handcrafted features in this instance. The best result is given by RF on task 5, with an accuracy of 68.3%. Notably, task 3 contributed to good performance, suggesting that, unlike the deep feature scenario, certain handcrafted features facilitated effective discrimination between the handwriting of patients and that of the control group.

	Task 2		Task 3		Task 4		Task 5	
	ACC	SD	ACC	SD	ACC	SD	ACC	SD
RF	61.3	2.5	66.4	1.8	53.0	3.2	68.3	1.5
K-NN	58.1	3.4	64.3	1.7	57.9	2.9	63.7	2.3
SVM	52.1	0.1	51.7	0.1	51.3	1.0	51.0	0.4
MLP	57.3	2.7	66.3	1.8	55.0	3.6	63.4	2.2

Table 4.6: Results of classification with handcrafted features. Bold values highlight the overall best performance achieved on each task.

To summarize the comparison between deep and handcrafted features, Figure 4.5 shows a vertical bar graph depicting the best overall accuracy achieved for each task. The graph includes the best overall classification performance obtained using deep-RGB features (bold values in Table 4.4) and handcrafted features (bold values in Table 4.6) for each task. The plot illustrates that the deep-based approach outperforms the handcrafted feature-based approach, except for task 3. These findings affirm the effectiveness of the proposed approach in combining shape and dynamic information. The marginal performance difference on task 3 is likely due to the task’s low complexity, which restricts the selection of discriminant features, as evidenced by the generally poor classification results obtained.

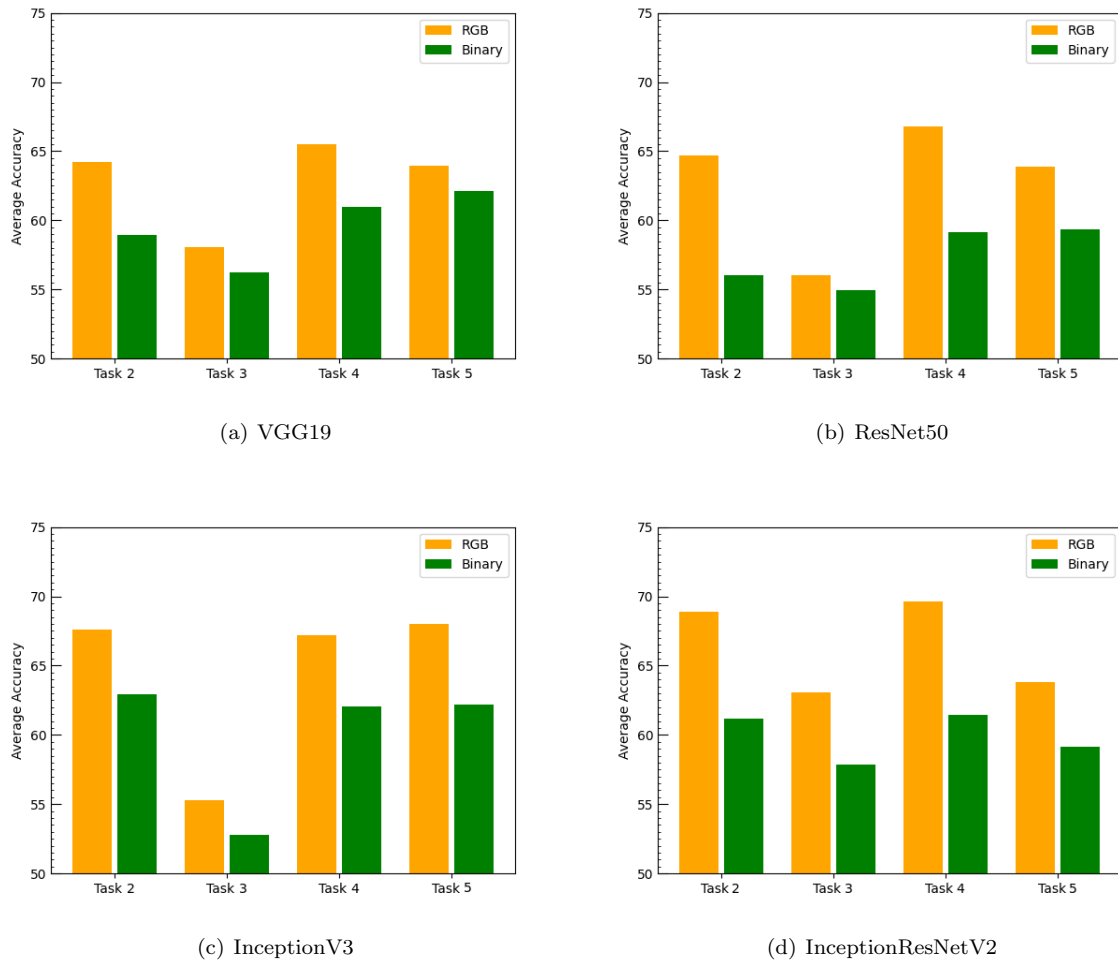


Figure 4.4: Accuracy for each task averaged over the results of five classifiers.

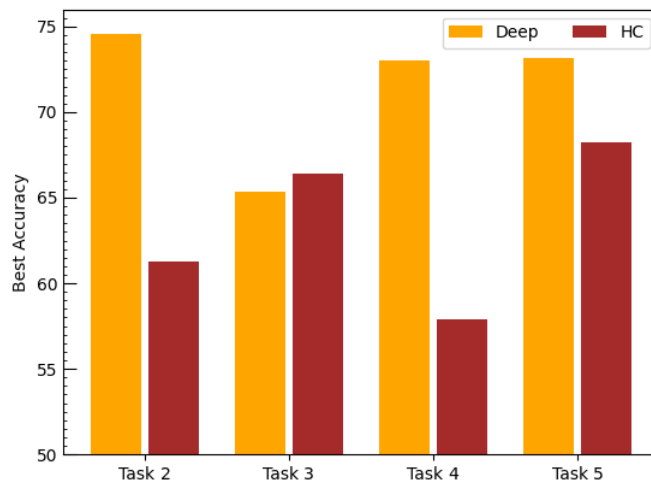


Figure 4.5: Comparison results between deep and handcrafted features.

The final set of experiments aimed to enhance the system’s performance, employing two distinct approaches. The first approach focused on enriching the feature space by combining deep and handcrafted features. These feature sets, originating from synthetic off-line RGB images and engineered handwriting features, were theoretically largely uncorrelated. The new feature representation was created by concatenating handcrafted and deep features from RGB images for each sample. The results, summarized in Figure 4.6, displayed a slight performance increment in most tests.

The second approach sought performance improvement by combining responses for each task and classifier on data related to each subject. A weighted majority vote rule was applied, incorporating the confidence degree provided by each classifier as weights. The results, reported in Table 4.7, revealed significant performance increments in many cases compared to the best results from single classifiers. Notably, the best result was achieved by combining classifiers using features from ResNet50, yielding an accuracy of 81.03%.

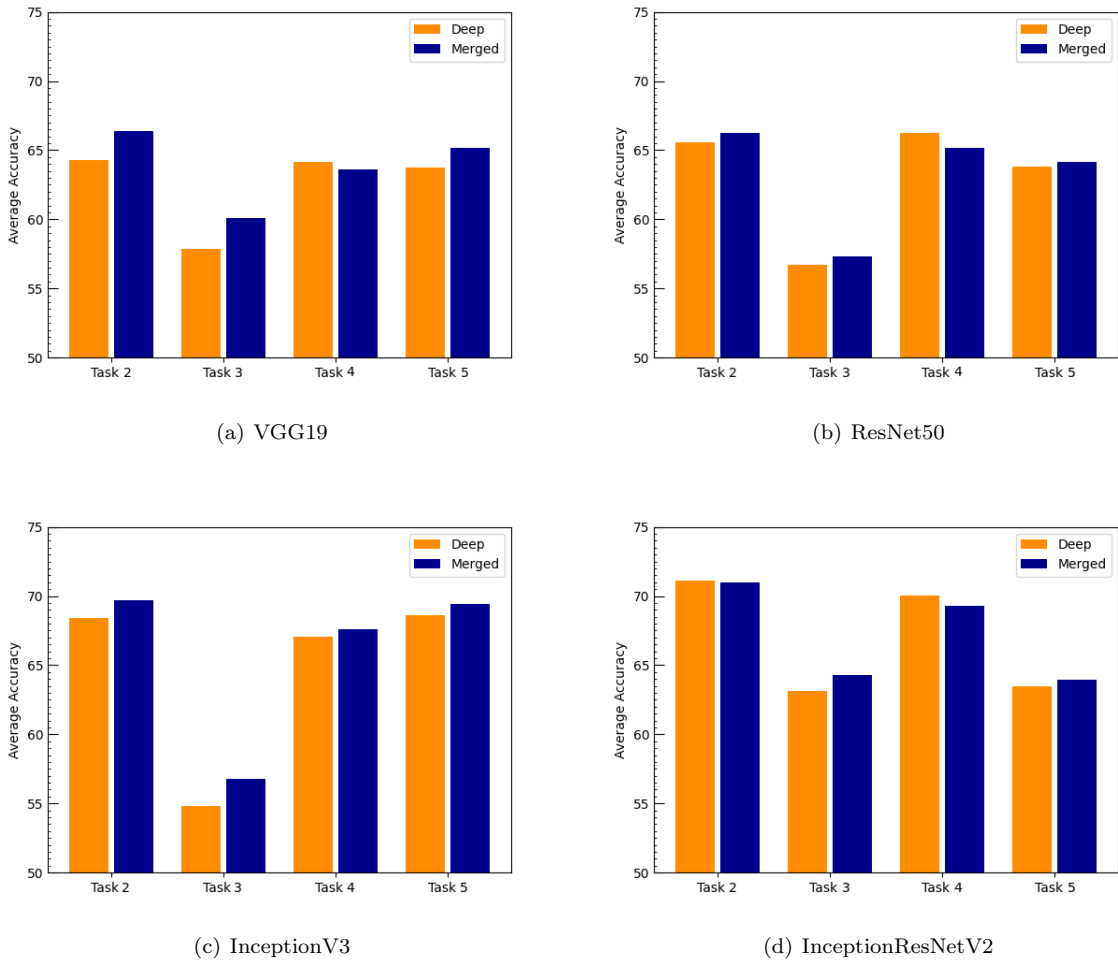


Figure 4.6: Accuracy for each task averaged over the results of the five classifiers using merged features, compared with deep features.

	VGG19	ResNet50	InceptionV3	Inc.Res.V2	All Nets
All classif.	74.13	81.03	77.01	79.31	74.13
	RF	K-NN	SVM	MLP	FC
All nets	71.83	71.25	76.21	69.54	68.96

Table 4.7: Classification results achieved using the weighted majority vote rule.

In conclusion, this research presented a comprehensive comparison between classifiers based on handcrafted and deep features for Alzheimer’s diagnosis from handwriting. Three feature sets were extracted for each handwriting sample based on handcrafted features, features from CNNs applied to synthetic

colour images (RGB), and features from CNNs applied to synthetic binary images. The comparison allowed for evaluating the role played by shape and the combined use of shape and dynamic information. The deep features demonstrated more promise than handcrafted ones, reaching the best performance with the RF classifier. Additionally, the contribution of shape information was noted to be significant for subject classification, especially when combined with dynamic information from RGB channels.

4.3 Comparison among RGB on-paper, Multichannel and Handcrafted Features on Graphic Tasks

The study presented in this section stems from considerations from the study described in Section 4.2. In particular, the previous section demonstrated the effectiveness of a hybrid approach based on feature extraction through deep neural networks and classification using machine learning algorithms from RGB images containing dynamic information about writing within the colour channels. These reasons led me to perform new experiments to compare standard handcrafted features with those derived from a feature extraction approach using RGB on paper and multichannel images [20]. Building upon these considerations, the set of experiments outlined in Section 4.2 was expanded by incorporating additional graphic tasks from the protocol, characterized by a higher difficulty level. In summary, nine distinct datasets were obtained for each task: one was based on standard dynamic features, four were based on features provided by CNNs applied to synthetic RGB images, and four were provided by CNNs applied to the MC images. Furthermore, the performance for each task and dataset was evaluated using the same classification schemes, namely Random Forest, K-Nearest Neighbor, Multi-Layer Perceptron, Support Vector Machines and a further comparison by considering the classification results directly provided by the fully connected layer of CNN. This approach facilitated a straightforward comparison of experimental results related to different feature vector representations, emphasizing the roles of shape and the combined use of shape and dynamic information. The primary contributions of this study can be summarized as follows:

- Evaluation of the contribution of dynamic information encoded in RGB channels of specifically generated images for an AD diagnosis system’s performance.
- Comparison of results achieved using these images with those obtained from multi-channel images, which include an additional dynamic feature in the fourth channel.
- Assessment of CNNs’ ability as automatic feature extraction tools, comparing their performance with widely-used handcrafted features.
- Evaluation of the method presented in Section 4.2 on additional tasks, providing insights into participants’ long-term motor planning ability.
- Comparison of two classification approaches: one employing handcrafted features with well-known machine learning algorithms and the other using features automatically extracted by CNNs from RGB and multi-channel images. The classification results from the fully connected layers of CNNs were also considered for comparison.

Data

Similar to the previous study, this work focuses on graphic tasks requiring subjects to produce handwritten forms less familiar than characters and words in their native language. This choice stemmed from the rationale that the habitual writing of individuals with neurodegenerative disorders might make alterations in their handwriting less conspicuous, rendering it more akin to that of healthy subjects with limited writing habits. In essence, we selected writing tasks that subjects were not accustomed to, making them less automated from a neuromotor control perspective. This approach aimed to highlight distinctions in writing characteristics between healthy subjects and those affected by neurodegenerative disorders more prominently. The tasks considered for this work are the same as described in 4.2 with two additional tasks. Specifically, tasks demanding increased fine motor control and those imposing a higher cognitive load and greater complexity in spatial organization. The fifth task involved reproducing a complex figure to assess the participant’s motor control abilities. This task examines changes in handwritten traits independently of letters, words, or related semantic meanings. Retracing the form necessitates constant motor re-modulation, as the form comprises a continuous line with varying curvature radii to evaluate

both fine control and long-term motor motion planning. Moving on to the sixth task, participants were asked to execute the well-known clock drawing test: drawing a clock face, including numbers, and positioning the hands at five past eleven. The clock-drawing test (CDT) is a screening tool for cognitive impairments and dementia, assessing spatial dysfunction and attention deficits. Initially designed for evaluating visuoconstructive abilities, it has been observed that abnormal clock drawing occurs in various cognitive impairments. The test requires verbal comprehension, memory, spatially coded knowledge, and constructive skills. Figure 4.7 shows sample images of the selected tasks.

As the previous study enhanced the inefficacy of using binary images, only approximating the actual shape of the handwritten traits, they have been discarded from this new set of experiments. A new procedure for generating synthetic images was adopted by incorporating a fourth channel alongside the previously considered three to assess the significance of dynamic information associated with each handwritten trait. According to this new procedure, a MC TIFF image was generated for each handwritten sample. The first three channels encoded the dynamic information used in our previous study, namely velocity, jerk, and pressure, while the fourth channel encoded acceleration. This kind of image is described in Section 3.2.3. In this context, MC on-paper images were considered. A feature extraction on TIFF images was employed using CNN’s ability to extract features automatically. More information about this kind of image can be found in Section 3.2.3.

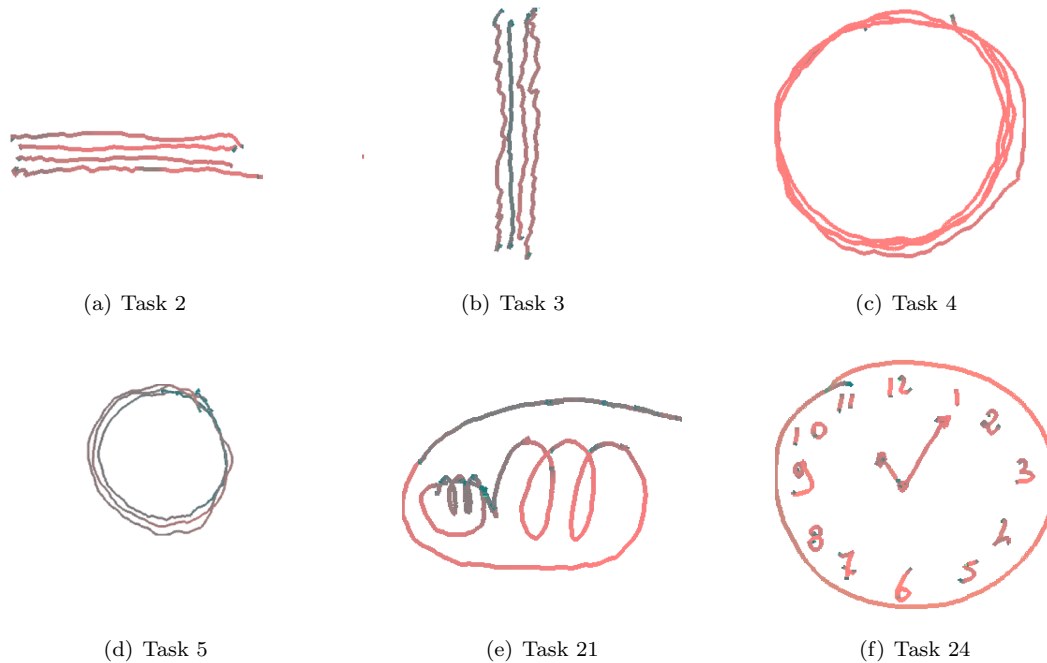


Figure 4.7: Examples of tasks performed by a participant involved in the experiments.

Experimental Setting

The experimental setting employed in this case is the same as described in Section 4.2, and its workflow is shown in Figure 4.1. RGB-deep and MC-deep features are obtained by feeding four models of CNNs. A difference from the previous studies is that results haven’t been combined in this case, so the last part of the system is missing. For comparison, the selected models are the same exploited in the previous research: VGG19, ResNet50, InceptionV3 and InceptionResNetV2, pre-trained on ImageNet and fine-tuned using the handwriting images. Each set of features is employed individually in the classification stage by various ML classification schemes, with a fully connected layer classifier of the CNNs also contributing to the classification process. After the classification step, there is a comparison of the performances obtained.

For what concerns hyperparameters of CNNs used for feature extraction and the ML algorithms for the classification, they were the same as the previous study, detailed in Section 4.2, for comparison sake. This comprehensive approach enables the analysis of participants’ handwriting across various tasks to predict cognitive status.

Experimental Results

To assess the system’s efficacy, I conducted three sets of experiments. First, I evaluated the performance of features extracted from the RGB on-paper images. Then, in the same way, I extracted features from MC images and evaluated the outcomes. Finally, I compared the results obtained from the proposed approach with those from the handcrafted features. A comprehensive account of the results from these experiments is presented in the following. To provide an exhaustive overview of the results for the first two experiments, as they show common patterns, Tables 4.8 and 4.9 present a detailed breakdown of the performance for each task by the five classifiers. The assessment involved the use of both RGB-deep features (Table 4.8) and MC-deep features (Table 4.9) extracted from different CNNs. Analysis of the tables reveals substantial variations in classifier performance for each task. Similarly, the performance fluctuates significantly for each task when utilizing features extracted from different CNNs. Moreover, within each classifier, the performance varies notably based on the features extracted from different CNNs and the tasks at hand. Looking at Table 4.8, the best accuracy value (74.6%) was achieved by MLP with features extracted by InceptionResNetV2 from task 2.

	Task 2		Task 3		Task 4		Task 5		Task 21		Task 24	
	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD
VGG19												
RF	67.6	2.3	62.8	2.8	70.0	2.4	68.2	1.6	63.7	2.4	73.2	2.1
K-NN	64.3	2.2	56.7	2.7	61.0	2.0	61.6	2.1	64.3	2.0	72.8	1.9
SVM	61.7	2.4	55.4	2.0	67.6	1.4	66.9	1.3	60.1	2.2	70.8	1.2
MLP	63.5	3.3	56.6	2.8	58.1	3.4	58.2	3.4	60.5	2.8	72.6	1.8
FC	64.1	15.2	59.0	5.1	70.7	7.5	64.7	11.0	64.8	11.7	70.4	13.5
ResNet50												
RF	69.7	2.1	60.3	2.4	71.2	2.4	71.7	2.2	66.0	2.2	69.1	2.3
K-NN	66.8	2.1	57.0	2.4	66.3	2.6	57.9	1.4	59.6	2.0	62.5	2.0
SVM	72.0	1.7	58.9	1.9	67.3	2.2	72.6	1.6	65.1	1.4	67.9	1.8
MLP	53.8	2.1	50.6	3.4	60.3	8.7	52.9	3.5	53.0	2.7	56.2	4.2
FC	61.1	8.0	53.5	9.5	68.8	7.7	64.4	12.7	64.3	9.3	66.8	11.4
InceptionV3												
RF	68.6	2.1	54.5	3.8	70.5	2.5	71.7	1.9	66.13	2.93	67.69	3.09
K-NN	66.5	1.8	56.0	2.8	62.7	2.5	66.8	2.0	64.79	2.06	58.58	2.43
SVM	71.1	1.5	55.4	2.7	71.1	2.0	73.1	1.9	69.18	1.32	69.84	2.28
MLP	67.5	2.5	53.5	1.5	63.9	1.7	62.8	4.7	58.12	5.55	63.24	3.87
FC	64.2	9.0	57.1	4.8	68.0	6.6	65.7	11.1	58.86	11.15	62.41	7.00
Inc.ResNetV2												
RF	70.4	2.4	65.4	2.0	73.0	2.1	69.2	2.3	65.2	2.1	65.3	2.4
K-NN	68.4	1.8	62.8	2.4	65.0	2.0	64.6	2.0	64.5	2.0	64.1	2.3
SVM	71.1	2.2	61.2	2.4	71.8	1.6	67.5	1.4	68.7	1.9	58.6	1.7
MLP	74.6	2.3	63.3	1.5	70.4	2.1	52.6	2.1	57.3	4.9	51.6	2.0
FC	60.0	15.9	62.7	7.3	68.0	10.6	65.3	10.1	62.0	8.5	55.3	9.4

Table 4.8: Classification results achieved using RGB features.

The best result achieved on MC images, as shown in Table 4.9, is an accuracy of 72.8% by the FC classifier, with features extracted by ResNet50 from task 4.

Two vertical bar graphs were generated for each feature type to summarize the outcomes in Tables 4.8 and 4.9, in Figure 4.8, regarding RGB-deep features, and Figure 4.9, concerning MC-deep features. For both figures, the left plot (a) showcases the mean accuracy of each classifier for each task, averaged across results obtained from the four CNNs. Conversely, the right plot (b) displays the mean accuracy of each CNN, averaged across the results from the five classifiers. These figures aim to facilitate a quick assessment of superior performance among CNNs, classifiers and tasks. Observing the figures and the tables, it becomes apparent that task 3 exhibits the poorest performance for both RGB and MC features. This outcome is understandable and already explained in the previous study in Section 4.2. The best results are achieved for the RGB images on the second task and for the MC images on the fourth task. The additional tasks (21 and 24) usually didn’t perform better from these experiments, though they require a major cognitive skill. Instead, concerning ML classifiers, the best performing are RF and SVM, confirming the effectiveness of the ensemble-based strategy of RF and that of the SVM kernel-based approach. These results also highlight the effectiveness of RGB features extracted by the CNNs, inde-

	Task 2		Task 3		Task 4		Task 5		Task 21		Task 24	
	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD
VGG19												
RF	64.0	2.6	60.9	2.5	68.6	1.5	64.3	2.4	66.4	2.4	69.0	2.2
K-NN	56.9	2.5	55.5	2.3	59.7	2.4	59.5	2.0	62.6	2.9	65.4	1.8
SVM	59.7	1.8	55.1	2.5	64.5	2.1	62.5	2.0	61.9	2.0	68.3	1.7
MLP	58.2	3.3	56.6	2.8	61.5	2.9	62.1	2.4	61.3	2.3	65.6	2.9
FC	66.2	10.7	62.7	10.0	69.5	10.5	64.2	11.1	64.7	7.4	64.5	12.3
ResNet50												
RF	60.1	2.6	56.3	3.0	72.0	2.4	66.8	2.7	65.0	2.2	66.0	2.7
K-NN	59.3	2.3	54.9	1.7	61.1	1.5	62.8	2.1	58.3	2.8	58.3	1.9
SVM	57.4	2.5	53.7	2.3	72.2	1.7	67.2	1.9	66.6	1.8	69.9	2.1
MLP	55.6	2.8	49.7	1.0	64.0	5.4	57.7	4.8	57.2	5.2	58.3	4.3
FC	61.7	12.1	59.0	8.8	72.8	9.0	67.7	13.7	65.6	10.3	68.6	8.3
InceptionV3												
RF	67.5	2.2	57.9	2.5	68.8	2.5	67.6	2.2	65.6	3.0	62.2	2.8
K-NN	63.5	1.9	54.1	2.2	57.1	1.7	61.4	2.0	59.6	2.5	60.6	2.3
SVM	67.7	2.2	61.8	2.0	68.3	2.3	68.7	2.0	66.5	2.4	64.5	2.8
MLP	66.3	2.0	56.8	3.2	62.7	4.0	63.7	4.0	55.4	3.5	61.8	2.2
FC	67.2	8.9	58.9	13.1	70.5	2.6	71.2	13.0	61.1	7.1	65.4	13.6
Inc.ResNetV2												
RF	65.2	2.3	61.4	2.7	67.6	2.1	59.8	2.6	67.6	2.2	66.4	2.2
K-NN	58.9	2.9	56.9	2.1	63.0	2.1	57.9	2.2	59.6	2.5	60.7	2.2
SVM	63.3	2.7	59.9	2.6	65.6	1.8	58.2	2.6	67.9	1.7	66.8	1.7
MLP	66.7	2.5	58.5	2.3	59.7	3.0	58.2	3.3	68.4	2.8	61.4	2.6
FC	61.6	11.8	58.7	10.4	67.8	11.3	64.8	10.9	70.7	3.3	55.0	17.1

Table 4.9: Classification results achieved using MC features. Bold values highlight the overall best performance achieved on each task.

pendent of the classification algorithm used to implement the classification layer. Furthermore, RF and SVM performance is better than that of the FC classifier trained during the process for feature extraction. The same does not occur for MC features. Looking at the CNNs (right images), the deepest architectures usually perform better. Figure 4.8, which illustrates CNN performances using RGB features, shows that InceptionResNetV2 demonstrated superior results for the first three tasks. However, different CNNs yielded the best performance for the subsequent tasks. InceptionV3 and VGG19 outperformed others in tasks 5 and 24, respectively. This divergence is likely attributed to the increased complexity of these tasks compared to the initial three. Consequently, simpler CNN architectures facilitated more effective training on the available data, leading to better outcomes. Conversely, for task 21, CNNs exhibited comparable performances, suggesting that the number of parameters did not significantly impact the training process. Turning attention to Figure 4.9, which depicts CNN performance on MC features, it is noteworthy that, except tasks 3 and 21, InceptionResNetV2 did not achieve the best performance. This outcome implies that employing CNNs with higher complexity may not necessarily result in improved system performance.

Furthermore, as depicted in Figure 4.9 (a), the FC classifier, in the case of MC features, demonstrated slightly superior or comparable performance compared to the other considered classifiers. This finding reinforces the notion that, during the training phase, addressing the heightened complexity of MC images necessitated leveraging the interaction between the feature extractor and the classification layer. It is worth noting that the results provided by the FC classifier exhibit higher standard deviation values than those of the other classifiers. This is likely attributable to the fact that, as previously mentioned, FC results were computed by averaging accuracy over the five test folders, while the results of the other classifiers were averaged over 30 runs. To highlight these aspects, Table 4.10 specifically compares the performance of RGB and MC features when utilizing the FC classifier.

Figure 4.10 and Table 4.10, in fact, compare the classification performance between RGB and MC features. The objective is to assess the contribution of the fourth channel in MC images in terms of performance. Figure 4.10 illustrates the contrast in classification performance between MC and RGB features. For this purpose, I computed the average accuracy across the five classifiers and represented a vertical bar for each task. Analysis of the graphs reveals that, in most instances, the performance obtained with RGB features is marginally superior to or on par with that obtained with MC features.

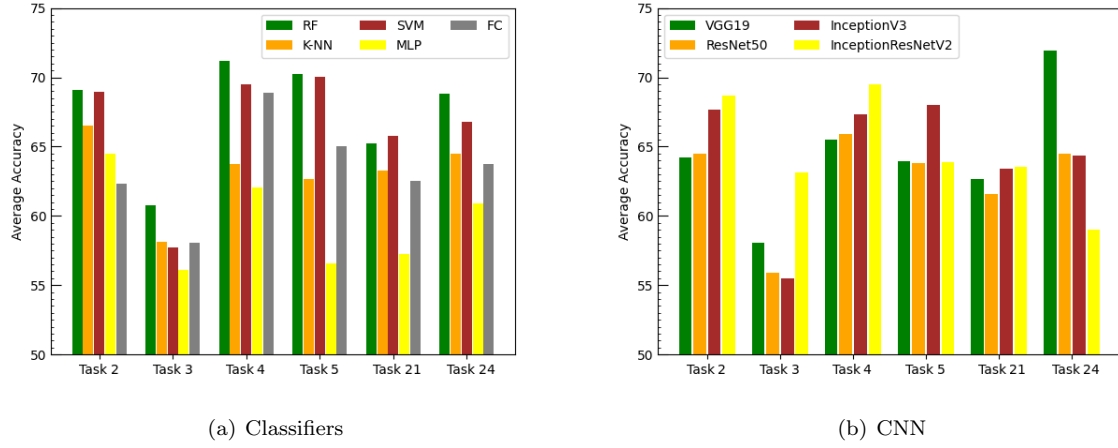


Figure 4.8: Average accuracy achieved by the classifiers (a) and the CNNs (b) using RGB images.

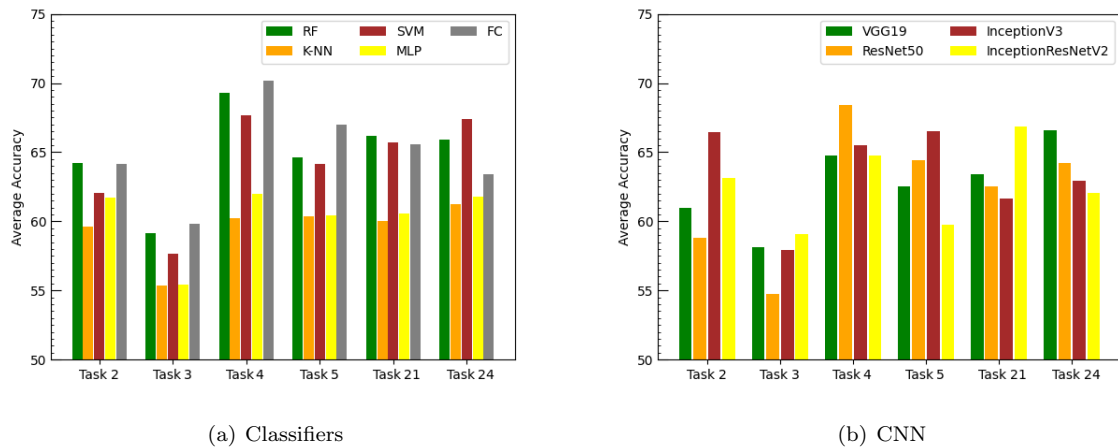


Figure 4.9: Average accuracy achieved by the classifiers (a) and the CNNs (b) using MC images.

This outcome supports the conclusion that the additional information from the fourth channel did not significantly enhance the system’s performance.

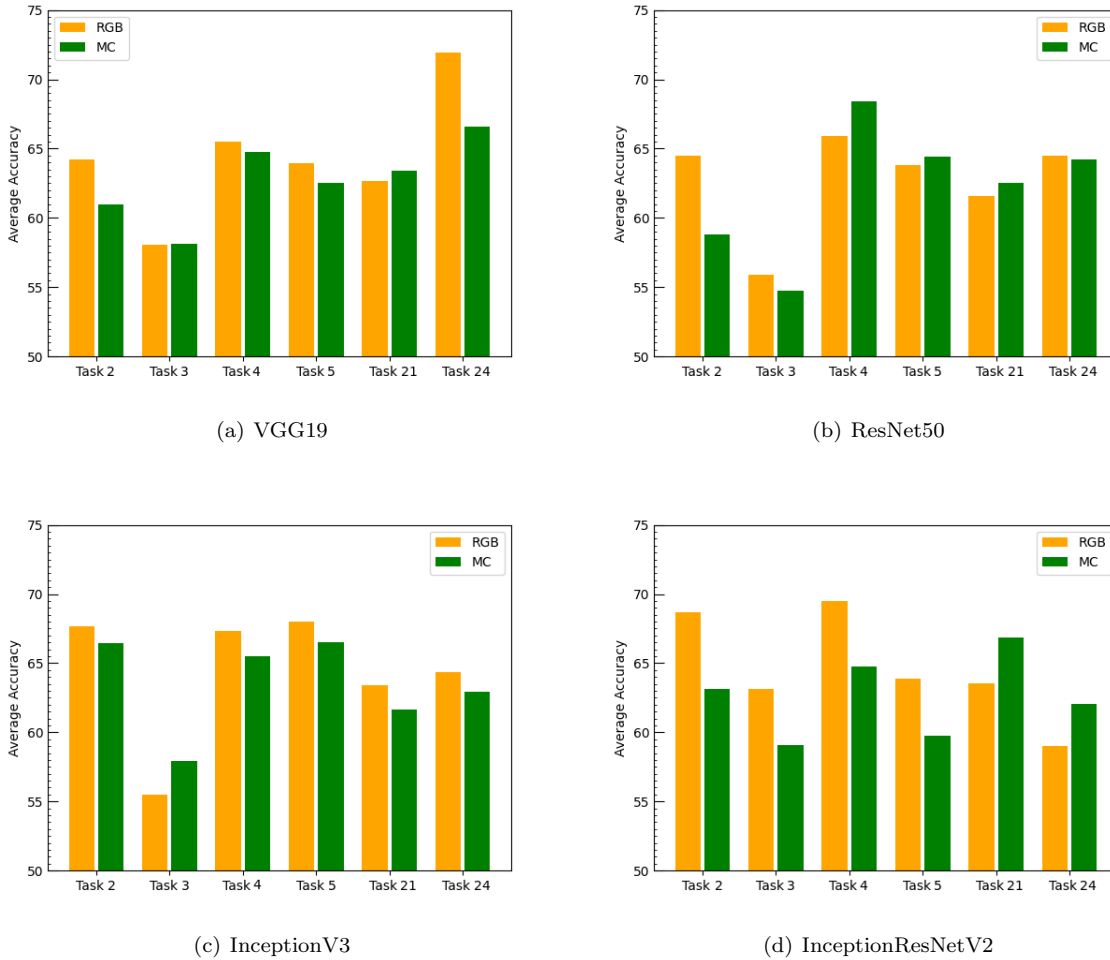


Figure 4.10: Accuracy for each task averaged over the results of five classifiers.

	Task 2		Task 3		Task 4		Task 5		Task 21		Task 24	
	RGB	MC	RGB	MC	RGB	MC	RGB	MC	RGB	MC	RGB	MC
VGG19	64.0	66.2	59.0	62.7	70.7	69.5	64.7	64.2	64.8	64.7	70.4	64.5
ResNet50	61.1	61.7	53.5	59.0	68.8	72.8	64.4	67.7	64.3	65.6	66.8	68.6
InceptionV3	64.2	67.2	57.1	58.9	68.0	70.5	65.7	71.2	58.9	61.1	62.4	65.4
Inc.ResNetV2	60.0	61.6	62.7	58.7	68.0	67.8	65.3	64.8	62.0	70.7	55.2	55.0

Table 4.10: Classification results achieved by the FC classifier, using RGB and MC features.

In the final experiments, the performance of handcrafted features is evaluated and compared to the RGB results. Table 4.11 presents the accuracy obtained using the handcrafted features. Analysis of the table reveals that the RF and KNN classifiers yield the best performance, affirming the effectiveness of the RF ensemble-based strategy, with KNN, in contrast to the deep features scenario, delivering satisfactory results. Moreover, in this context, task 3 demonstrates notable performance. These results suggest that, unlike the situation with deep features, certain handcrafted features contribute information, enabling effective differentiation between the handwriting of patients and that of the control group. Conversely, tasks 4 and 24 exhibit poor performance with these features, implying that they do not sufficiently capture the shape and dynamics of handwriting to distinguish between samples of cognitively impaired individuals and those of the control group.

To summarize the comparison between deep RGB and handcrafted features, Figure 4.11 shows a vertical bar plot representing the best overall accuracy achieved for each task in a vertical bar graph. The plot showcases the superior performance of our deep-based approach in combining shape and dynamic

information, except for task 3, where the slight difference in performance may be attributed to the task’s low complexity, hindering the selection of discriminant features. This is consistent with the generally poor classification results obtained using both deep and handcrafted features in this specific task.

	Task 2		Task 3		Task 4		Task 5		Task 21		Task 24	
	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD
RF	61.3	2.5	66.4	1.7	53.0	3.2	68.2	1.5	64.9	2.9	55.9	2.4
K-NN	58.1	3.4	64.3	1.7	57.9	2.9	63.7	2.3	61.1	2.3	54.2	2.9
SVM	52.1	0.1	51.7	0.0	51.3	1.0	51.0	0.4	51.7	0.9	52.3	2.6
MLP	57.3	2.7	66.3	1.8	55.0	3.6	63.4	2.2	63.2	3.5	53.3	3.3

Table 4.11: Results of classification with the handcrafted features. Bold values highlight the overall best performance achieved on each task.

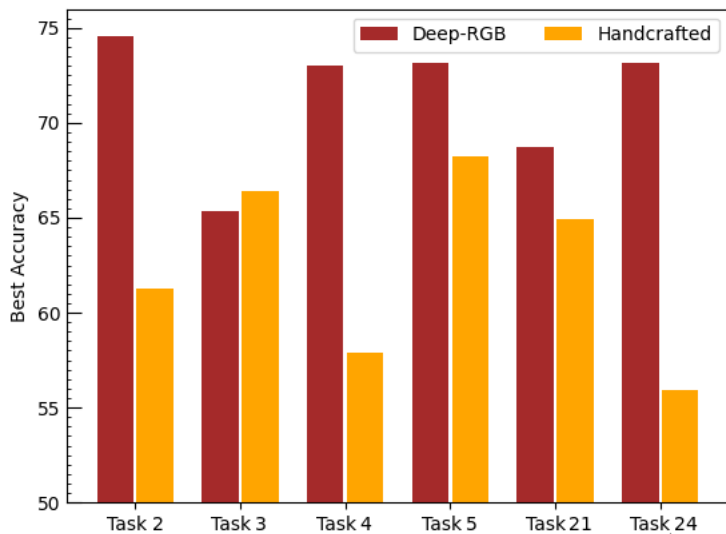


Figure 4.11: Comparison results between Deep-RGB and handcrafted features.

This work reinforces the outcomes of the previous one as also in this case, it is apparent that deep features show more promise than their handcrafted counterparts, consistently achieving superior accuracy. For each task and classification approach, there is consistently a CNN model whose features outperform those obtained with the handcrafted ones. The lone exception is task 3, where the performance with handcrafted features was marginally better than that achieved with deep features. Regarding the comparison between RGB and MC deep features, the analysis of results indicates that adding an extra channel in generating multi-channel images does not enhance feature extraction significantly. Classification results using RGB deep features are nearly always superior to MC deep features. The only anomaly is observed in task 21, where the FC classifier, trained with MC deep features using InceptionResNetV2, yielded slightly better results.

4.4 Comparison among RGB on paper, Offline and Handcrafted Features on Writing Tasks

In prior investigations outlined in 4.2 and 4.3, an attempt was made to distinguish AD patients from healthy control with a system based on ML and DL techniques by integrating shape-related information with the dynamics of the handwriting process. Given the outcomes from the previous section, it seems that RGB on-paper images are the most suitable for analysis by the proposed experimental workflow. Building upon these findings, the current study aims to investigate whether performance can be enhanced by utilizing original offline images obtained by digitizing text written on paper sheets during the administration of our protocol, as detailed in Section 3.2.4. The rationale behind this approach is to consider the

handwriting samples' shape, size, and actual thickness. Once again, I leveraged the capability of Deep Neural Networks (DNNs) to extract features from raw images automatically, following the transfer learning approach [25]. It is important to note that an advantageous aspect of this approach is the potential to use parts of text previously written by subjects for diagnostic purposes. This enables the examination of whether initial signs of the disease were already present and facilitates the analysis of its progression. The drawback, of course, is the loss of dynamic information directly derived from online data. However, based on the preliminary results, this loss does not appear significant. The following Sections describe the data used, the system workflow and the experimental results.

Data

Three kinds of data are considered in this research: RGB on paper images, Section 3.2.2; offline images, Section 3.2.4; and handcrafted features, Section 3.3.1. The proposed protocol encompasses various tasks designed to assess participants' motor control, memory, and cognitive capacity, including writing letter groups, words, and graphic exercises. However, for the purposes of this study, only writing tasks were considered. Specifically, the following tasks: signature (Task 1), continuously write the cursive bigram "le" four times (Task 9), and write the Italian word for sheet, "foglio" (Task 10). The first task is a well-established activity often found in literature, typically executed with a highly automated gesture. The ninth task evaluates fine motor control by repeating a consistent pattern sequentially but with varying sizes. Lastly, task 10 involves copying a common word with an interesting graphic composition, incorporating ascending and descending traits. These specific tasks were chosen because they facilitate the examination of automatism, regularity, coordination of motor sequences, and spatial organization. Graphic tasks were deliberately excluded from our selection. This decision aligns with one of the study's objectives: to determine if features automatically extracted from offline handwriting images effectively predict AD. Under this hypothesis, earlier handwriting examples, which usually lack graphic patterns, could be valuable for detecting early signs of cognitive impairment and assessing their progression. Figure 4.12 shows examples of offline images from the selected tasks.

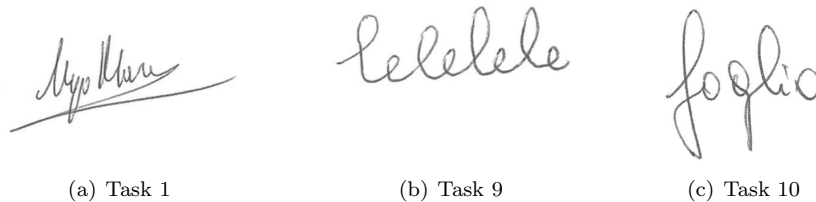


Figure 4.12: Example of offline images.

Experimental Setting

The experimental setting for these experiments focuses on the baseline workflow shown in Figure 4.1; which is the same as described for the previous studies, but this time, binary and MC images were substituted with offline ones and different tasks were evaluated. First, I evaluated the system on features automatically extracted from RGB images, then those extracted from Offline images and in the end, I tested handcrafted features. A final experiment involved the application of the majority vote rule. The extraction of deep features from images was the same as the previous study, as the same CNNs were used without changing their hyperparameters. New ML algorithms were introduced to assess the efficacy of the features extracted, while the KNN algorithm used in the previous works was discarded. The overall selected algorithms are XGB, RF, DT, SVM, and MLP. A 5-fold cross-validated grid search was conducted to optimise each classifier's performance to select the most suitable set of hyperparameters. This involved defining a range of values for each parameter and exhaustively testing all possible combinations, as depicted in Table 4.12. The principal metric to evaluate the system's performance is accuracy, but other metrics were also computed, like sensitivity, specificity, and precision.

Classifier	Hyperparameters	Constraints
XGB	min child weight	1, 5, 10
	gamma	0.5, 1, 1.5, 2
	subsample	0.6, 0.8, 1
	colsample bytree	0.6, 0.8, 1
	max depth	3, 4
RF	bootstrap	True, False
	max depth	10, 20, 50
	max features	auto, sqrt
	min samples leaf	1, 2, 4
	min samples splir	2, 5, 10
	n estimators	100, 200
Tree	criterion	gini, entropy
	min samples split	2, 10
	max depth	2, 5, 10
	min samples leaf	1, 5, 10
	max leaf nodes	2, 5, 10
SVM	C	0.1, 1, 10, 100
	gamma	1, 0.1, 0.01, 0.001
	kernel	rbf
	class weight	balanced, None
MLP	hidden layer sizes	50, 100, 200
	activation	tanh, relu
	solver	lbfgs, SGD
	alpha	0.0001, 0.05
	learning rate	constant, adaptive

Table 4.12: ML Classifiers and their hyperparameters involved in the Grid search process

Experimental Results

To ease the comparison, Table 4.13 shows the results obtained by each set of features in terms of Accuracy, Sensitivity, Specificity and Precision expressed in percentage for every task taken into account in this work. It is worth noting that for every CNN feature extractor, only two classification results are reported, one for the FC classifier and one for one ML algorithm out of the five tested. Concerning the ML evaluation, only the performance achieved by the best algorithm was reported to reduce the complexity of the table. For each task, the best values of the computed metrics are in bold.

The table shows a high variability according to the experiment implemented, as the performance widely differs across tasks and feature sets. Independent of the kind of image, the best classifiers are XGB and RF. Looking at the outcomes on handcrafted images, instead, the best classifier is always RF. Moreover, ML algorithms outperform the FC classifier both for RGB and offline deep features. This result demonstrates that ensemble-based architectures more effectively capture the differences between patients and healthy people. Another aspect that contributed to these outcomes is the grid-search procedure, which optimised the ML classifiers' performance. Differently from the previous studies, in this case, it is not easy to assess which CNN is the best, though, also in this case, the best results are obtained from features extracted from deepest models like InceptionResNetV2. The table shows that the best accuracy was achieved for the first task by the RF classifier with the handcrafted features (65.86%), while for both the ninth and the tenth task by the XGB classifier with the offline features (74.7% and 74.4%, respectively). The first task is the one that achieves the worst result among all the tasks, while the highest performing is the ninth task for every evaluated set of features. The motivation behind this result can be explained by analysing the task: the first task requires a well-known kinematic gesture, becoming almost an automatic graphic task that doesn't require significant motor or cognitive attention. The ninth and tenth tasks, instead, thanks to their characteristic of including descending and ascending traits and requiring greater coordination and control skills, are the most useful to highlight the difference between patients and healthy controls. Looking at the set of features, it seems that features extracted from offline images perform better than the others, but for the first task, they don't achieve the best result as handcrafted features outperform them. The table shows that offline features outperformed the RGB features in most cases. These performance differences are more important for the ML classifiers, confirming that the latter could better exploit the information contained in the offline features. This is

an interesting result, as these features were extracted from offline images containing the original traits of the participants’ handwriting without any information regarding the dynamics of the movement.

	Task 1				Task 9				Task 10			
	Acc	Sen	Spe	Pre	Acc	Sen	Spe	Pre	Acc	Sen	Spe	Pre
RGB features												
VGG19												
ML	56.6 (RF)	61.6	51.6	59.3	68.6 (MLP)	67.1	64.7	68.7	66.5 (XGB)	67.1	66.5	67.7
FC	56.3	50.6	62.7	59.2	68.8	58.7	79.6	75.3	67.8	58.7	77.2	72.8
ResNet50												
ML	56.4 (XGB)	59.1	54.8	58.9	66.6 (RF)	69.9	64.6	68.6	65.3 (XGB)	67.9	64.1	66.8
FC	58.7	74.2	42.2	57.8	62.5	62.1	62.7	63.5	59.7	71.3	46.9	58.5
Inc.V3												
ML	59.97 (RF)	64.7	55.2	62.8	68.3 (XGB)	70.5	66.6	69.9	65.1 (RF)	69.9	61.6	66.8
FC	52.5	66.3	37.4	53.2	62.5	60.9	63.8	63.8	65.5	72.5	57.8	64.3
Inc.Res.V2												
ML	56.8 (RF)	60.8	54.7	59.5	69.6 (XGB)	69.3	66.1	69.5	63.3 (RF)	65.5	61.9	65.6
FC	55.1	58.4	51.8	56.5	68.8	60.9	77.1	73.6	60.7	90.8	28.9	57.3
Offline features												
VGG19												
ML	61.1 (XGB)	67.3	54.2	65.5	71.2 (XGB)	73.7	66.5	74.2	68.1 (MLP)	69.2	64.3	70.2
FC	66.3	77.2	53.8	65.3	70.4	73.8	66.6	71.4	67.6	69.3	65.3	69.3
ResNet50												
ML	61.2 (XGB)	67.3	54.8	64.4	70.6 (XGB)	73.3	68.0	74.9	74.4 (XGB)	75.7	70.5	77.8
FC	59.3	57.9	60.2	62.1	65.9	88.6	41.1	62.9	58.9	82.9	32.1	57.9
Inc.V3												
ML	56.4 (XGB)	74.2	54.4	61.1	75.4 (XGB)	80.3	69.5	76.2	68.6 (RF)	80.3	51.7	66.5
FC	60.3	62.4	50.8	61.8	71.1	88.6	51.2	67.2	67.3	72.7	61.5	68.0
Inc.Res.V2												
ML	59.6 (XGB)	63.3	55.6	66.4	75.7 (XGB)	78.5	72.8	78.6	68.4 (MLP)	73	62.8	71.5
FC	62.6	77.2	46.1	61.8	67.7	92.1	41.1	63.7	67.2	75	58.9	67.3
Handcrafted features												
ML	65.8 (RF)	67.6	63.8	64.2	74.7 (RF)	78.8	70.8	72.2	69.2 (RF)	69.1	69.4	69.0

Table 4.13: Performance comparison on all the feature sets considered.

To have a better understanding of the performance achieved by the offline images, Figure 4.13 (a) compares the results of FC and ML classifiers, while Figure 4.13 (b) compares the results from the point of view of CNNs. In the first case, the plot shows the average accuracy of every task over the CNNs. The second case shows the average accuracy over the classifiers for every task. The first illustrates how, for two tasks out of three, the ML algorithms outperformed the FC classifier, confirming the effectiveness of the ML in characterizing the handwriting of people affected by AD. The second shows a great variability of the CNNs’ performance, and no CNN outperformed the others on all three tasks. However, these plots also confirmed that the best performance came from the second task.

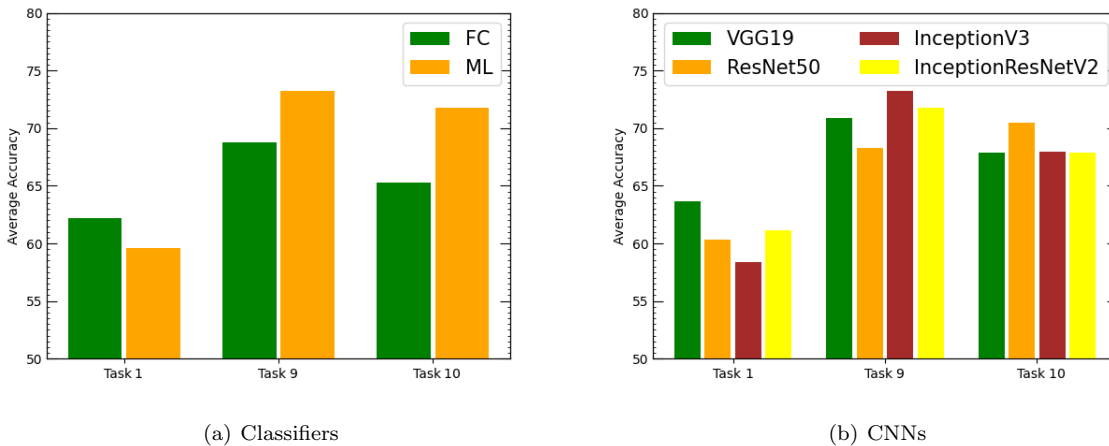


Figure 4.13: Average accuracy achieved by the classifiers (a) and the CNNs (b) using offline images.

To further compare the effectiveness of the three types of features extracted, Figure 4.14 plots for each CNN the average accuracy reached by the ML algorithms. Looking at the figure, it is worth noting that the HC features, in most cases, outperformed the others, while offline features always outperformed the RGB ones, except for the first task using InceptionV3. Comparable performance is observed between offline and HC features in Task 9 and Task 10, both of which are more intricate than Task 1. Specifically, the sequencing of "le" bigrams and the word "foglio" demands a heightened motor control effort compared to the signature, a gesture characterized by high automation. The dominance of handcrafted features in yielding better results is unsurprising, given their extensive utilization in literature and their incorporation of dynamic information related to the handwriting process. As mentioned earlier, the well-established significance of these features in supporting the diagnosis of AD is recognized in the field. In contrast, deep features are automatically extracted from the CNN and far outnumber handcrafted features.

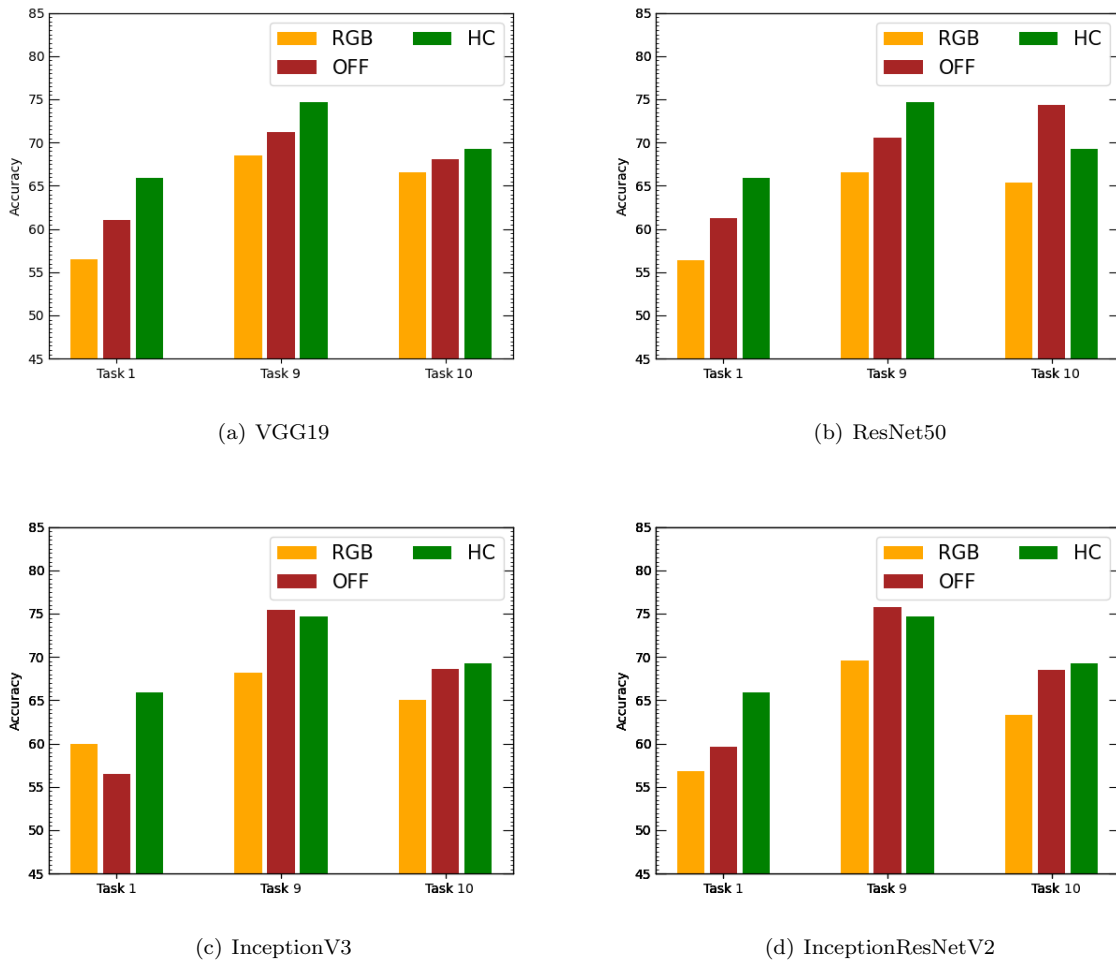


Figure 4.14: Comparison between accuracy achieved by the ML classifiers from RGB, offline and Handcrafted features.

Finally, an additional experiment assessed whether combining classifier responses could enhance overall performance. This experiment involved applying a majority vote rule, and the summarized results can be found in Table 4.14. The table presents accuracy outcomes achieved by combining responses from the considered classifiers using features extracted from both RGB and offline images for each task. The first two rows depict the results of the ML approach applied to RGB and offline features, respectively. Similarly, the third and fourth rows display the results of the FC layer of the CNNs using RGB and offline features, respectively. Additionally, the table includes results obtained by combining responses across all tasks. It is worth noting that the best-combined results were obtained using offline features. While the results are interesting, the observed performance improvement is generally modest. This outcome is likely attributed to the limited number of classifier responses available for combination and the simplicity of the combining rule. Notably, the most favourable outcome was achieved by combining ML classifier responses for offline features, resulting in an accuracy of 75.94%. It is noteworthy that combining responses for all

tasks consistently yields lower results than combining responses for only task 9. This discrepancy is due to the higher accuracy typically obtained for task 9 compared to the other tasks.

Table 4.14: Combining results.

		Task 1	Task 9	Task 10	All Tasks
ML	RGB	59.05	70.25	67.46	69.35
	OFF	61.89	75.94	72.79	74.23
FC	RGB	60	72.34	69.5	68.02
	OFF	66.66	74.63	68.7	73.33

The analysis of the experimental results, focusing on three tasks within the aforementioned protocol, indicates a substantial performance improvement thanks to offline images compared to features extracted from RGB synthetic images. Moreover, the performance appears comparable to that achieved using dynamic features alone.

4.5 Exploiting Lognormal Features to Support the Diagnosis of AD through Handwriting Analysis

The subsequent sections delineate a system to facilitate the diagnosis of AD by analysing handwriting using the Sigma-Lognormal model. This model is specifically employed for characterizing complex movements, and I utilized it to break down each handwriting task into a vector summation of basic time-overlapping movements. This process enabled extracting a set of Sigma-Lognormal parameters from the movements. Subsequently, based on the Lognormal parameters, I derived a set of lognormal features, which were then assessed using ML algorithms [24]. This research aimed to assess whether these simple lognormal features could effectively characterize individuals' handwriting. Specifically, the investigation aimed to determine whether ML classifiers could discern interesting patterns, distinguishing the handwriting of a healthy control from that of an individual affected by AD.

Data

Data involved in this research refer to the first set of lognormal features computed and described in Section 3.3.2. The final step of this experiment compares the outcomes obtained from the lognormal features with those achieved from RGB in-air on-paper images described in Section 3.2.2. Given my uncertainty about which tasks would be best suited for analysis using lognormal features, I opted to include tasks of various types. Six tasks were considered: joining two points with a vertical line continuously four times; tracing a circle ($d = 6cm$) continuously four times; writing continuously four times, in cursive, the bigram 'le'; Copying in reverse order a simple the Italian word "bottiglia"; writing under dictation a telephone number and the Clock Drawing Test (CDT). The initial two tasks fall under the graphic tasks category (tasks 3 and 4); the third and fourth tasks involve copy and reverse copy tasks (tasks 9 and 15); the fifth task is a dictation task (task 23), and the sixth is a graphic task (task 24), assessing cognitive skills. Figure 4.15 shows examples of RGB in-air on-paper images generated from a person executing the tasks involved in this experiment.

Experimental Setting and Results

In this case, the experimental setting resembles the baseline depicted in Figure 4.1, with some differences: lognormal features have been assessed with ML algorithms, while RGB in-air on-paper images were evaluated with CNNs and the FC classifier. No combining rule was applied in this case. Concerning images I trained three CNN models, namely VGG19, ResNet50 and InceptionV3, whose hyperparameters are the same as expressed in Section 4.1. I evaluated lognormal features with standard machine learning algorithms: KNN, RF, MLP, SVM, LR, GB, and XGB. Hyperparameter settings for these algorithms were maintained at default values as provided by scikit-Learn, except for the SVM classifiers, which employed a linear kernel, and the KNN classifier, where the number of neighbours was set to 3. As for other experiments, to ensure statistical significance, I conducted 30 runs for each ML classifier. Performance evaluation of the models considered metrics such as accuracy, Sensitivity, Specificity, Precision, False Negative Rate, and Area Under the ROC Curve. Given the 30 runs for each classifier, the aforementioned metrics were computed for each run, and their average, along with the standard deviation, is presented

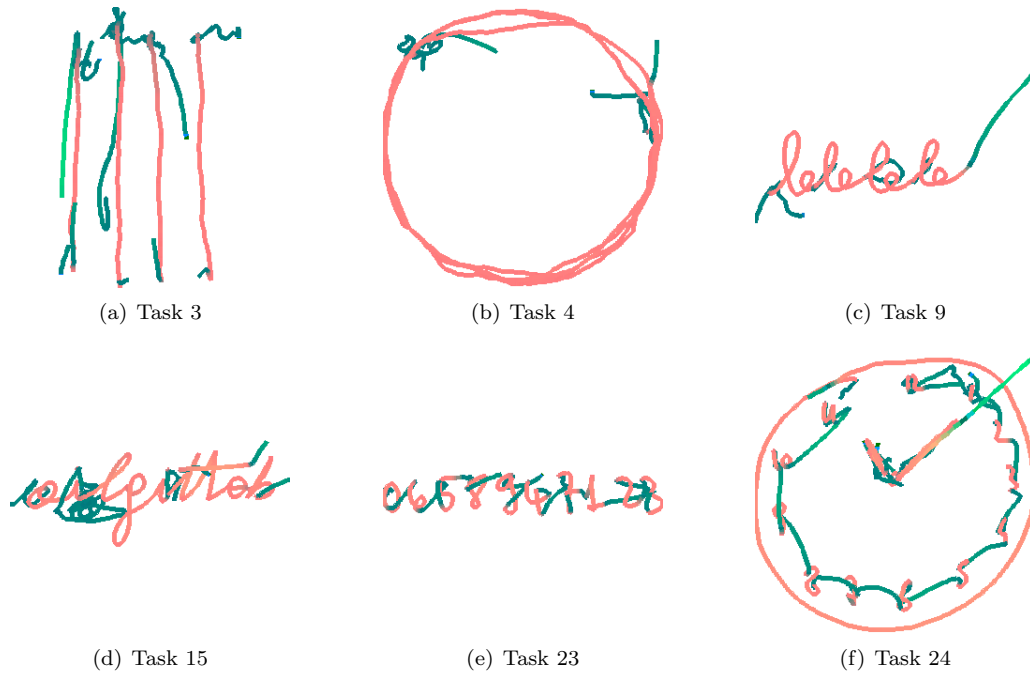


Figure 4.15: Examples RGB in-air on-paper images generated from the execution of tasks by a participant involved in the experiments.

in the subsequent tables. All metrics are expressed in percentages, except for the AUC, with bold values highlighting the best performance achieved.

Examining the accuracies presented in Table 4.15, it is noteworthy that the highest performance, with an accuracy of 74.66% (SVM), was achieved on task 9, while the lowest performance was recorded on task 3 with an accuracy of 58.24% (DT). The SVM algorithm emerged as the best-performing in most cases, except for tasks 4 and 15, where RF yielded higher values. Conversely, the DT classifier consistently delivered the poorest performances across various tasks. Upon closer inspection of the table, a discernible trend emerges: the initial two tasks exhibit inferior performances compared to the others. This trend can be elucidated through an analysis of the tasks in question. They involve graphic elements, assessing the dynamics of simple movements and the individual’s motor control without requiring significant cognitive attention. In contrast, the subsequent tasks involve words, numbers, and the clock drawing test, demanding cognitive attention due to semantic meanings, the inclusion of descending and ascending traits, and the requisite for enhanced coordination, control skills, and the use of working memory. These observations suggest that lognormal features prove more effective in tasks with semantic meaning than in purely graphic tasks, accentuating differences between patients and healthy controls.

T	Accuracy						
	KNN	RF	DT	SVM	LR	GB	XGB
3	64.3 (2.7)	63.8 (2.2)	58.2 (3.1)	66.9 (2.2)	63.9 (1.5)	61.5 (2.3)	61.3 (2.9)
4	62.7 (2.4)	63.9 (2.6)	60.8 (3.2)	59.4 (2.0)	61.3 (2.1)	63.3 (3.7)	62.0 (3.0)
9	62.7 (2.0)	72.9 (2.2)	67.0 (2.7)	74.6 (1.5)	74.2 (1.5)	69.7 (2.7)	71.2 (2.4)
15	64.6 (2.6)	72.0 (2.0)	62.1 (3.5)	68.7 (1.8)	70.4 (1.4)	69.9 (3.5)	70.5 (2.7)
23	66.8 (2.3)	71.8 (1.8)	63.5 (2.8)	73.6 (1.9)	72.6 (1.9)	69.7 (3.2)	69.6 (2.9)
24	67.0 (2.8)	67.1 (3.0)	61.1 (3.9)	72.6 (2.8)	71.5 (2.3)	68.4 (2.9)	70.1 (3.0)

Table 4.15: Average Accuracy achieved on 30 runs for every ML algorithm on lognormal features

Table 4.16 displays the sensitivity values derived from the experimental process. Sensitivity is a crucial metric in medical contexts, offering insights into accurately recognising patients. The highest sensitivity score, 77.47%, is given by RF in task 15, while the lowest, 59.29%, is observed with DT in task 3. This table shows that RF and LR classifiers demonstrate good sensitivity values. However, it is essential to note that superior sensitivity does not necessarily equate to the overall best classifiers, as SVM emerges as the top performer in accuracy. Despite SVM’s supremacy in accuracy, this table highlights that other classifiers exhibit a greater ability to identify patients correctly.

Sensitivity							
T	KNN	RF	DT	SVM	LR	GB	XGB
3	67.2 (2.6)	64.3 (3.0)	59.2 (4.2)	64.5 (2.7)	67.3 (2.1)	62.6 (2.5)	62.5 (3.1)
4	66.6 (3.5)	72.0 (3.9)	63.5 (4.2)	62.5 (3.5)	64.5 (3.2)	70.5 (4.3)	67.9 (3.6)
9	63.5 (2.0)	68.0 (3.7)	66.3 (4.1)	67.3 (1.8)	70.1 (2.0)	66.3 (4.2)	68.6 (3.5)
15	67.7 (3.0)	77.4 (2.1)	66.1 (5.4)	68.8 (3.0)	73.4 (2.7)	75.5 (3.5)	75.9 (3.1)
23	62.4 (3.1)	70.1 (3.0)	62.3 (4.1)	68.6 (2.8)	68.5 (2.1)	67.6 (4.2)	68.2 (4.3)
24	70.5 (3.5)	69.6 (4.6)	62.0 (6.2)	75.1 (3.4)	75.2 (3.6)	71.1 (4.8)	72.5 (5.3)

Table 4.16: Average Sensitivity achieved on 30 runs for every ML algorithm on lognormal features

Table 4.17 illustrates the specificity values obtained in the study. The optimum specificity measure, 82.03%, is reached by SVM for task 9, while the lowest result, 54.06%, is observed with RF in task 4. This measure is closely tied to sensitivity, providing insights into the accurate classification of healthy control participants. Despite SVM being identified as the top classifier based on accuracy, it did not obtain the highest sensitivity values. Consequently, the specificity table reveals elevated values of this metric for the SVM classifier. This implies that while SVM stands out as the best classifier in terms of accuracy, considering sensitivity and specificity considerations suggests its superior ability to recognize healthy controls rather than patients among our study participants.

Specificity							
T	KNN	RF	DT	SVM	LR	GB	XGB
3	61.1 (4.5)	63.2 (4.8)	57.1 (5.9)	69.5 (3.5)	60.3 (2.7)	60.4 (4.3)	59.9 (5.0)
4	58.0 (3.1)	54.0 (3.0)	57.4 (4.2)	55.6 (3.4)	57.3 (2.4)	54.3 (4.7)	54.8 (4.8)
9	61.9 (3.6)	77.8 (2.0)	67.6 (5.5)	82.0 (2.6)	78.4 (2.0)	73.0 (3.2)	73.8 (3.3)
15	60.7 (4.0)	65.0 (3.9)	56.9 (6.4)	68.5 (2.5)	66.5 (2.4)	62.8 (5.1)	63.8 (5.0)
23	71.3 (2.6)	73.4 (2.7)	64.8 (4.7)	78.7 (2.9)	76.9 (3.1)	71.8 (3.5)	71.0 (3.9)
24	63.1 (3.6)	64.0 (4.0)	60.1 (5.0)	69.8 (4.7)	67.1 (3.7)	65.4 (4.7)	67.3 (3.7)

Table 4.17: Average Specificity achieved on 30 runs for every ML algorithm on lognormal features

In Table 4.18, it is evident that the highest precision value, 80.36%, is given by SVM in task 9, while the lowest precision, 59.89%, is observed with DT in task 3. Despite SVM not being the superior classifier in recognizing patients according to sensitivity, this table underscores its precision as the most notable among the classifiers in recognising patients.

Precision							
T	KNN	RF	DT	SVM	LR	GB	XGB
3	65.3 (3.0)	65.7 (2.8)	59.8 (3.4)	69.7 (2.7)	64.9 (1.7)	63.2 (2.7)	62.9 (3.1)
4	66.3 (2.6)	66.2 (2.4)	65.0 (3.3)	63.8 (2.0)	65.2 (2.0)	65.9 (3.3)	65.4 (3.0)
9	63.2 (2.3)	76.4 (1.9)	68.5 (3.3)	80.3 (2.5)	77.6 (2.0)	72.1 (2.8)	73.6 (3.0)
15	69.2 (2.6)	74.4 (2.5)	66.7 (3.7)	74.5 (2.0)	74.3 (1.7)	72.8 (3.5)	73.2 (2.8)
23	69.6 (2.7)	73.8 (2.3)	64.8 (3.8)	77.3 (2.8)	76.0 (2.9)	71.6 (3.6)	71.1 (3.0)
24	69.0 (3.1)	69.6 (3.1)	64.3 (3.9)	74.6 (3.1)	72.8 (2.6)	71.1 (2.9)	72.3 (2.7)

Table 4.18: Average Precision achieved on 30 runs for every ML algorithm on lognormal features

Table 4.19 presents the FNR values computed during the experimental phase. It is evident that the best value, 22.52%, is achieved by RF in task 15, while the worst, 40.71%, is observed with DT in task 3. FNR is closely tied to Sensitivity, as they are complementary metrics. FNR represents the amount of erroneously classified patients, and the lowest, the better. In the medical domain, this information is pivotal because misclassifying a patient is a more critical issue than an error involving a healthy individual.

Table 4.20 presents the AUC values. AUC quantifies the area under the ROC curve, depicting the diagnostic ability of a binary classifier as the discrimination threshold varies, with a higher value indicating superior performance. The table indicates that LR obtained the highest result (0.83) on the ninth task, while DT recorded the lowest outcome on the third task (0.58).

To assess the efficacy of the computed lognormal features, I compared the results presented in the aforementioned tables and outcomes obtained from deep neural networks trained on synthetic RGB images containing both in-air and on-paper traits. Table 4.21 showcases the accuracy performances of

FNR							
T	KNN	RF	DT	SVM	LR	GB	XGB
3	32.7 (2.7)	35.6 (3.0)	40.7 (4.2)	35.4 (2.7)	32.6 (2.1)	37.3 (2.5)	37.4 (3.1)
4	33.3 (3.5)	27.9 (3.9)	36.4 (4.2)	37.4 (3.5)	35.4 (3.2)	29.4 (4.3)	32.0 (3.6)
9	36.4 (2.0)	31.9 (3.7)	33.6 (4.1)	32.6 (1.8)	29.8 (2.0)	33.6 (4.2)	31.3 (3.5)
15	32.2 (3.0)	22.5 (2.1)	33.8 (5.4)	31.1 (3.0)	26.5 (2.7)	24.4 (3.7)	24.0 (3.1)
23	37.5 (3.1)	29.8 (3.0)	37.6 (4.1)	31.3 (2.1)	31.4 (2.1)	32.3 (4.2)	31.7 (4.3)
24	29.4 (3.5)	30.3 (4.6)	37.9 (6.2)	24.8 (3.4)	24.7 (3.6)	28.8 (4.8)	27.4 (5.3)

Table 4.19: Average FNR achieved on 30 runs for every ML algorithm on lognormal features

AUC							
T	KNN	RF	DT	SVM	LR	GB	XGB
3	0.66 (0.02)	0.70 (0.01)	0.58 (0.03)	0.72 (0.02)	0.71 (0.01)	0.66 (0.02)	0.67 (0.02)
4	0.65 (0.02)	0.68 (0.02)	0.60 (0.03)	0.63 (0.04)	0.65 (0.02)	0.67 (0.03)	0.66 (0.03)
9	0.69 (0.01)	0.82 (0.01)	0.66 (0.02)	0.82 (0.01)	0.83 (0.01)	0.79 (0.02)	0.78 (0.02)
15	0.68 (0.02)	0.78 (0.01)	0.61 (0.03)	0.78 (0.01)	0.79 (0.01)	0.76 (0.02)	0.77 (0.02)
23	0.69 (0.01)	0.78 (0.01)	0.63 (0.02)	0.77 (0.01)	0.76 (0.01)	0.75 (0.02)	0.75 (0.02)
24	0.68 (0.02)	0.72 (0.03)	0.74 (0.02)	0.76 (0.02)	0.75 (0.02)	0.73 (0.03)	0.74 (0.02)

Table 4.20: Average AUC achieved on 30 runs for every ML algorithm on lognormal features

the two approaches. In most instances, ML, thus lognormal features, outperformed DL, thus RGB images, particularly with the SVM classifier. DL approaches on images only outperformed the ML approaches on features in the fourth task with the VGG19 network. For a comprehensive comparison, ROC curves for each task were plotted in Figure 4.16 for the classification algorithms/nets that outperformed others in at least one task, specifically LR and SVM among the ML classifiers and VGG19 among the CNNs, referring to Table 4.21.

T	ML							Deep		
	KNN	RF	DT	SVM	LR	GB	XGB	VGG19	ResNet50	Inc.V3
3	64.34	63.83	58.24	66.92	63.98	61.56	61.33	61.62	62.64	62.20
4	62.77	63.97	60.83	59.47	61.31	63.30	62.05	72.19	65.90	71.25
9	62.79	72.97	67.01	74.66	74.25	69.73	71.27	66.83	62.09	70.81
15	64.67	72.03	62.11	68.72	70.43	69.97	70.59	66.82	58.62	63.97
23	66.87	71.80	63.58	73.69	72.68	69.70	69.63	66.01	62.43	70.37
24	67.09	67.10	61.14	72.66	71.50	68.45	70.10	66.48	64.07	65.39

Table 4.21: Comparison results.

These diverse evaluation sources show that the deep approach (utilizing RGB images) surpassed the lognormal-based approach in tasks involving graphic elements (Tasks 3 and 4). Conversely, lognormal features demonstrated their effectiveness in addressing handwriting and cognitive tasks, as evidenced by their superior performance in the remaining tasks.

4.6 A Machine Learning Approach to Analyze the Effects of AD on Handwriting through Lognormal Features

This study aims to create a classification system for AD diagnosis by utilizing handwriting features extracted using the Sigma-Lognormal model [29]. I conducted a thorough analysis of the results to understand the quality of the extracted features and the relationships between these characteristics and the personal information of the examined people. Features involved in this study are described in Section 3.3.2 and represent an expansion of the feature set previously examined. The effectiveness of these features is evaluated through a series of ML experiments, as outlined in the subsequent sections, and the obtained results are discussed.

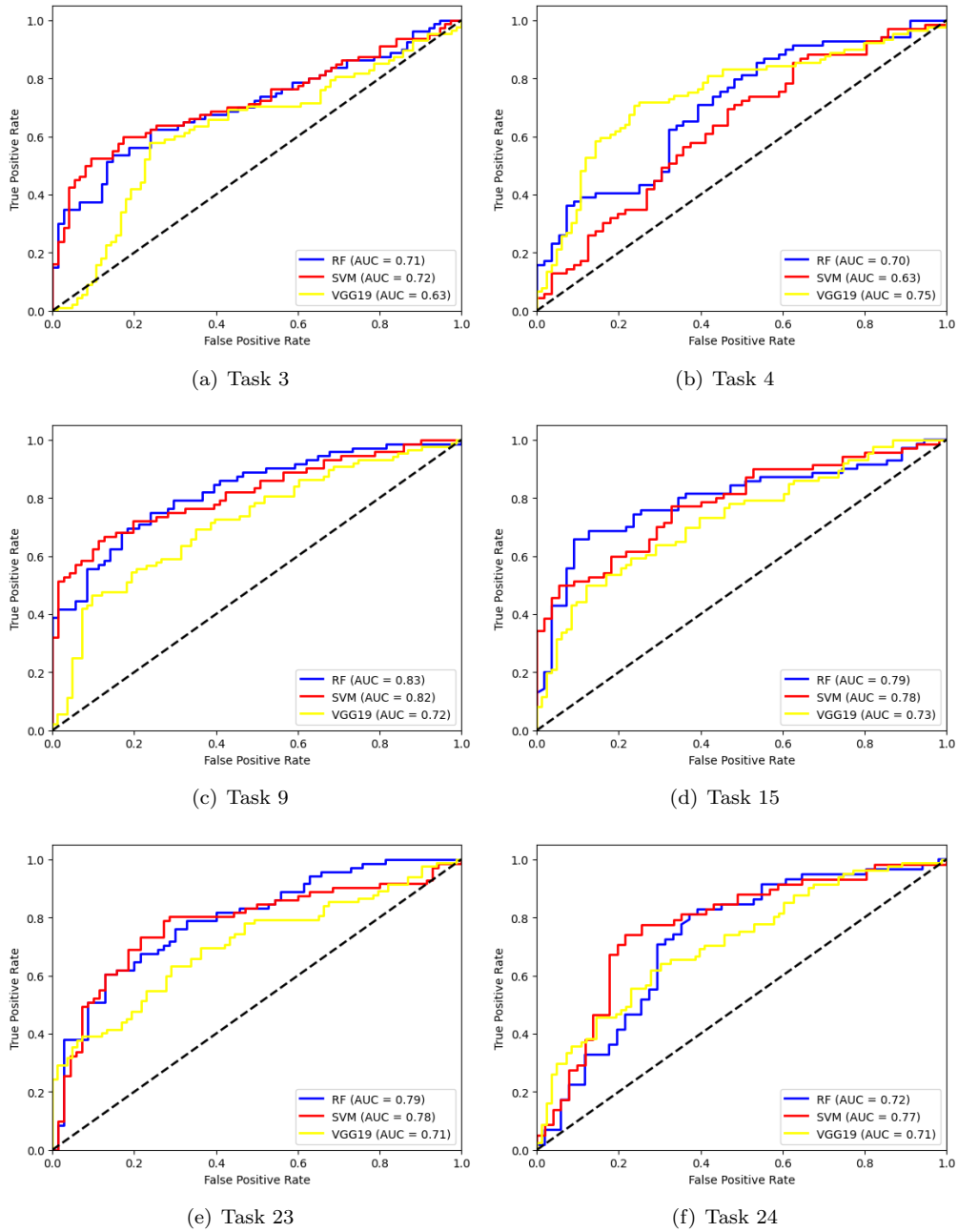


Figure 4.16: Comparison of ROC curves obtained from RF, SVM and VGG19 for every task.

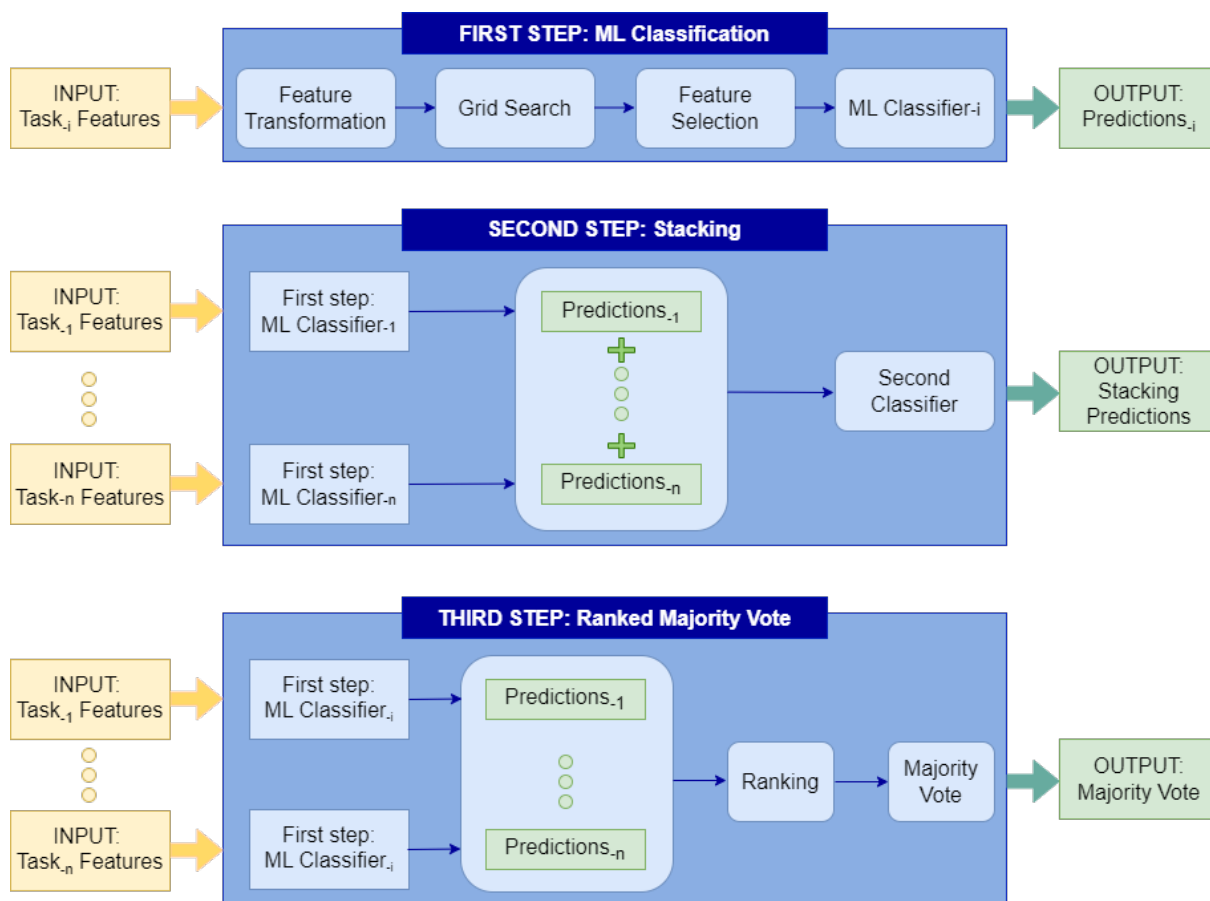


Figure 4.17: Workflow Representation.

Data

The data exploited in this study refer to the lognormal set of features described in Section 3.3.2. This research aims to identify the system’s optimal task performance and evaluate the sigma-lognormal model’s suitability and the extracted features for the specific problem. To achieve this, the proposed experimental setting was employed for every of the 25 tasks of the protocol. The feature computation process produced a collection of features for each task, resulting in individual datasets. In one of the developed experiments I compared the performance achieved on these features with those achieved on handcrafted static and dynamic features, described in Section 3.3.1

Experimental Settings

The workflow employed is detached from the baseline architecture and can be more comprehensively elucidated by considering a three-step approach, as discussed in the following section and depicted in Figure 4.17.

The first step of the workflow is devoted to ML classification. In this case, I opted for seven widely recognized classification algorithms to classify each task dataset: XGB, RF, DT, SVM, MLP, KNN and LR. Before initiating training, I employed a pipeline consisting of three ML techniques to enhance the discriminative capability of the system:

- Feature scaling;
- Grid search;
- Feature Selection: Recursive Feature Elimination (RFE)/SelectKBest

Following the grid search in the initial step of the experimental phase, I implemented a feature selection procedure, enabling the classifier to focus on the most crucial features by eliminating redundant or less informative ones. RFE was chosen as the algorithm for most classifiers, except MLP and KNN, for which SelectKBest was utilized.

Once each algorithm’s optimal set of hyperparameters and features was determined, I performed the training. Specifically, the dataset was randomly partitioned into training and testing sets, allocating 50% of the total samples to each set while maintaining a balance between the two classes: healthy controls and patients. Thirty random runs were executed to ensure reliability and robust performance estimates, and the final results were averaged across these runs. A final comparison was performed with results obtained from handcrafted features (Section 3.3.1. Dynamic features encompass handwriting characteristics such as Start time, Duration, Vertical dimension, Horizontal dimension, Vertical speed peak, Peak of vertical acceleration, Relative initial inclination, Jerk, Pen pressure, etc. Each task performed by each subject generates a feature vector. Notably, these dynamic features were included in this study solely for comparison purposes.

The second phase of the proposed approach involved a stacking technique, a form of ensemble stacking or stacked generalization in ML. This approach combines predictions from multiple models or base learners to generate a more potent and accurate final prediction. By training diverse models on the same dataset and utilizing a ”meta-learner” to learn from their predictions, stacking aims to enhance predictive power and system robustness. The output predictions provided by the classifiers in the first step were utilized. Specifically, the responses obtained for all tasks were merged to form a new feature vector for each sample (person). The final score was determined by averaging the stacking results over the 30 runs.

The third and final step of the proposed experimental approach involved utilizing the outcomes from the initial step outlined above to implement a combining technique based on a ranking. First, I employed a ranking technique for each classifier, ordering the tasks based on their relevance measured by the accuracy metric. Accuracy was chosen for its effectiveness in assessing classifier performance, resulting in a list of tasks sorted in ascending order concerning this metric. Subsequently, I implemented a combination rule known as the majority vote. This rule, applied in problems with multiple classifiers, relies on the majority opinion to determine the final prediction. Each classifier’s prediction is considered a vote, and the class with the most votes is selected as the ultimate prediction. This process involved combining predictions for different sets of tasks, run by run, and the average accuracy was computed over the thirty runs. The list of tasks generated by the ranking process was used to select significant subsets of tasks to apply the majority vote.

Experimental Results

Table 4.22 presents the outcomes derived from the first experimental step, showcasing the average accuracy (in percentage values) for each task and classifier across 30 runs. Notably, the best-performing classifier for each task is highlighted in bold. Average accuracies range from a minimum of 58.31%, achieved by KNN in the first task, to a maximum of 78.41% obtained by RF for task 23. Specifically, RF demonstrated superior performance compared to other classifiers in seven of the 25 tasks, while KNN never achieved the top result. Regardless of the classifier type, the table indicates that the 1st task yielded the worst performance, whereas the 23rd task yielded the highest. Notably, the 1st task involved the execution of a signature, while the 23rd task required writing a telephone number dictated to the participant.

Accuracy								Accuracy							
T	XGB	RF	DT	SVM	MLP	KNN	LR	T	XGB	RF	DT	SVM	MLP	KNN	LR
1	67.6	66.3	60.3	62.7	60.1	58.3	65.0	14	68.3	67.4	61.6	66.2	64.9	65.0	66.7
2	65.3	66.3	60.0	68.6	65.5	61.7	65.1	15	71.0	72.5	67.6	73.2	73.0	69.3	73.3
3	66.9	68.1	64.8	68.5	63.8	62.0	67.6	16	65.8	64.2	59.2	67.4	63.0	61.4	67.4
4	65.0	66.3	58.4	66.4	66.6	62.6	67.7	17	74.6	75.7	71.3	71.8	70.9	65.6	75.0
5	67.2	69.0	62.0	66.8	65.0	61.5	69.9	18	64.5	68.4	62.2	67.2	64.9	62.9	67.7
6	70.6	75.0	67.4	75.6	61.8	64.1	75.0	19	65.0	66.4	61.9	66.2	59.4	66.0	65.1
7	70.5	68.61	68.6	73.7	69.8	66.3	71.4	20	66.2	66.8	64.7	67.6	66.1	66.5	69.9
8	68.3	68.6	65.2	69.1	68.2	64.5	65.6	21	66.3	67.2	58.7	63.8	59.0	61.8	67.0
9	76.5	77.3	70.0	74.6	67.1	74.4	76.0	22	72.9	72.3	68.3	71.6	68.6	66.7	68.8
10	71.2	69.5	62.8	68.9	63.5	65.0	70.2	23	77.4	78.4	70.7	78.3	66.7	75.0	78.0
11	69.0	69.3	63.3	65.5	68.0	64.7	70.1	24	76.4	73.9	65.9	65.3	68.0	62.3	67.5
12	68.3	68.6	62.4	66.1	64.9	60.5	70.6	25	72.8	74.6	63.2	73.1	68.8	68.1	71.3
13	67.3	62.6	58.7	63.7	67.1	62.8	66.7								

Table 4.22: Average Accuracy achieved on 30 runs for every ML algorithm on lognormal features

Since task 23 achieved the highest accuracy, this experiment’s further analysis was performed by computing additional evaluation metrics, including precision, sensitivity, specificity, False Negative Rate, F1 score and Area Under the Curve. In Table 4.23, the first column identifies the classifier used for the 23rd task, while the subsequent columns present the metric values averaged over 30 runs. All metrics are expressed as percentage values except for AUC, which ranges from 0 to 1. The best metric value in each column is highlighted in bold. This table reveals that RF emerged as the top classifier for this task, excelling in accuracy and other metrics, except for precision and specificity. The latter two metrics indicate that RF may not be the optimal classifier for correctly classifying healthy controls. However, in the medical context, prioritizing accurately identifying individuals affected by the illness is crucial, given the more significant consequences of a false prediction. The FNR stands at 22.07%, representing the lowest value among all classifiers, indicating that RF was the best in recognizing patients. Figure 4.18 shows a bar plot of evaluation metrics averaged over 30 runs for every ML classifier tested with the 23rd task.

CLS	ACC	PRE	SEN	SPE	FNR	F1S	AUC
XGB	77.43	77.67	77.03	77.88	22.97	77.18	0.83
RF	78.41	78.72	77.93	78.89	22.07	78.14	0.85
DT	70.78	78.77	58.16	83.25	41.84	65.21	0.70
SVM	78.39	81.53	73.51	83.23	26.49	77.06	0.84
MLP	66.72	90.23	36.38	96.73	63.62	49.67	0.81
KNN	75.06	77.04	71.57	78.54	28.43	74.01	0.80
LR	78.02	79.88	75.37	80.69	24.63	77.24	0.84

Table 4.23: Average results achieved on 30 runs for every ML algorithm on lognormal features, extracted from the execution of task 23, i.e. the one that reached the best performance according to Table 4.22

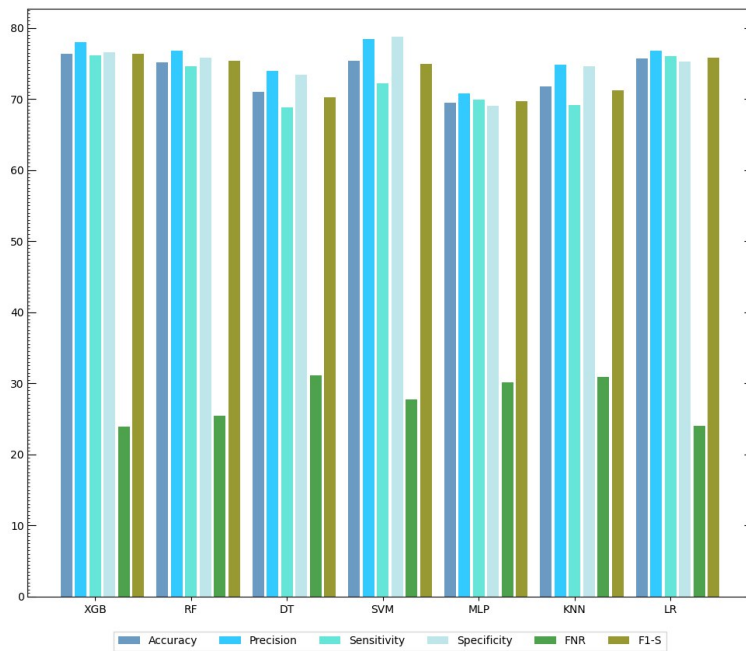


Figure 4.18: Averaged evaluation metrics achieved on 30 runs for every ML algorithm on lognormal features, extracted from the execution of task 23.

Table 4.24 compares the results obtained for each task using handcrafted features (dynamic) and those obtained with the lognormal features computed for this study. The accuracy percentage values in the table are computed by averaging this metric over 30 runs, displaying only the results of the best classifier, with the top performance for each task highlighted in bold. Overall, employing our proposed lognormal features for classification consistently led to better results in most cases, as the bar plot in Figure 4.19 shows at a glance.

The second step of the experimental workflow involved using a stacking approach. Stacking is widely used in ML, particularly when working with multiple classifiers. This is because it can potentially

DYN. FEAT.		LOG. FEAT.		DYN. FEAT.		LOG. FEAT.			
T	CLS	ACC	CLS	ACC	T	CLS	ACC		
1	XGB	64.5	XGB	67.6	14	XGB	64.1	XGB	68.3
2	XGB	63.0	SVM	68.6	15	XGB	64.3	LR	73.3
3	XGB	57.3	SVM	68.5	16	XGB	67.0	SVM	67.4
4	XGB	59.2	LR	67.7	17	XGB	69.3	RF	75.7
5	DT	69.0	LR	69.9	18	XGB	68.9	RF	68.4
6	XGB	57.4	SVM	75.6	19	XGB	68.3	RF	66.4
7	DT	58.4	SVM	73.7	20	XGB	66.1	LR	69.9
8	DT	61.7	SVM	69.1	21	DT	55.1	RF	67.2
9	XGB	64.5	RF	77.3	22	RF	70.0	XGB	72.9
10	XGB	61.6	XGB	71.2	23	XGB	72.3	RF	78.4
11	XGB	66.4	LR	70.1	24	XGB	56.0	XGB	76.4
12	XGB	67.2	LR	70.6	25	DT	59.8	RF	74.6
13	XGB	68.3	XGB	67.3					

Table 4.24: Comparison between average Accuracy achieved on 30 runs for every task with the best-performing ML algorithm for Dynamic and Lognormal features

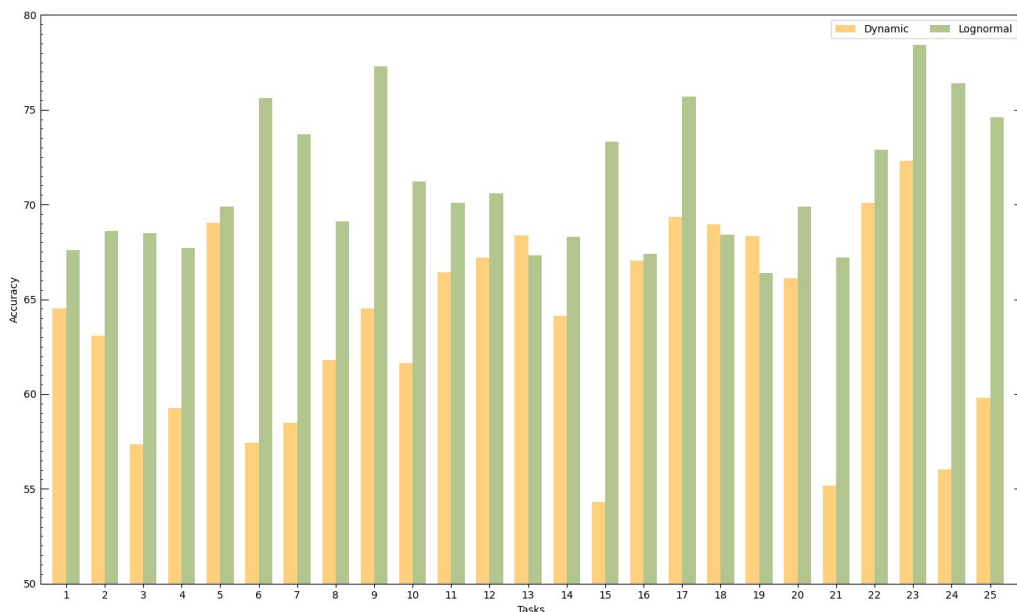


Figure 4.19: Comparison between average Accuracy achieved on 30 runs for every task with the best-performing ML algorithm for Dynamic and Lognormal features.

enhance overall predictive performance compared to utilizing individual models in isolation. The meta-model in stacking can effectively learn to capitalize on various base models' strengths while mitigating their weaknesses. Table 4.25 presents the evaluation of each classifier based on various metrics. These parameters enable a more comprehensive analysis of a classifier's performance. Notably, stacking didn't consistently outperform the average accuracy obtained from the initial classification step for all tasks. Examining all parameters, the best outcome is achieved by applying stacking to the results obtained from XGB as the initial classifier, yielding a final stacking accuracy of 76.29%.

1st CLS	ACC	PRE	SEN	SPE	FNR	F1S	AUC
XGB	76.29	77.99	76.09	76.52	23.91	76.32	0.84
RF	75.15	76.78	74.55	75.75	25.45	75.31	0.83
DT	70.98	73.89	68.85	73.39	31.15	70.2	0.78
SVM	75.38	78.45	72.24	78.7	27.76	74.91	0.83
MLP	69.41	70.78	69.86	69.00	30.14	69.63	0.76
KNN	71.75	74.79	69.1	74.58	30.9	71.16	0.77
LR	75.68	76.78	76.02	75.3	23.98	75.81	0.83

Table 4.25: Stacking results averaged over thirty runs with XGB classifiers, with the output of first-step classifiers

The third phase of the experimental process involved the implementation of two well-established techniques in machine learning: ranking and majority vote. Table 4.26 presents the ranked lists of tasks based on their accuracy for each classifier. Upon inspection of the table, it becomes evident that certain tasks consistently appear in top positions, irrespective of the algorithm used. Specifically, task number 23 is positioned the first five times across seven algorithms, task 9 ranks within the top three positions for six algorithms, and task 17 secures a place within the top four positions for five algorithms. Tasks 6, 7, 15, 22, and 24 also exhibit notable relevance. The 23rd task involves writing a phone number under dictation, the 9th requires continuous writing of the bigram 'le' four times, and the 17th involves writing six words in defined boxes, each with varying levels of complexity. As for the other relevant tasks, the 6th entails writing 'l, m, p'; the 7th involves 'n, l, o, g' in designated spaces; the 15th is a reverse copy of 'bottiglia'; the 22nd involves the direct copy of a phone number, while the 24th is the clock drawing test.

Table 4.27 demonstrates the performance of the majority vote for different sets of tasks, considering the first n tasks from the ranking lists. The initial column, labelled "T set," denotes the number of tasks considered for each set, ranging from a minimum of three to a maximum of 25 (i.e., all tasks). Beyond the fifth set, involving 11 tasks, the accuracy of the majority vote decreases. The highest majority vote accuracy, reaching 82.5%, is obtained by combining predictions given by XGB from the first three tasks of the ranked list.

Considering the intermediate stages of the experimental process, many observations can be derived from the results. After the grid search, a feature selection was performed in the first step. While different sets of features were selected for each classifier and task, common patterns were observed, leading to some notable observations. Certain personal features like age, profession, and education were consistently selected for nearly every task. In the realm of temporal features, f1, f4, f3, f7, and f9 were frequently chosen, while among geometric features, f62, f47, f26, f28, f61, f60, f49, and f64 were prominent selections. As for Signal-to-Noise Ratio (SNR) features, f19, f20, f22, and f23 were frequently chosen. Notably, these features were selected in at least ten tasks out of 25 by the classifier that achieved the best result, XGB.

The selection of temporal features aligns with some general assumptions, indicating that individuals with impairments generally take more time to complete a handwriting task, resulting in more lognormal functions in the velocity profile and segments in the trace acquisition. Among features related to SNR and geometric shapes, the emphasis lies on those describing the variation of a measure sequence. Particularly, relating SNR or geometric features to temporal ones contributes to creating more robust features, enhancing the differentiation between the two classes I aim to distinguish. To better comprehend the significance of each feature for the examined problem, we employed a parametric statistical test, the t-test. This test evaluated whether the difference between the means of the two groups was statistically significant. The test yielded a p-value for each characteristic, indicating the strength of evidence against the null hypothesis. A significance level 0.05 was used as the threshold, below which the null hypothesis was rejected. In the examined case, all the aforementioned features reported a p-value smaller than the threshold, proving that these features significantly distinguish between the two classes. This discussion highlighted valuable features extracted during the initial step. However, further refinement is necessary

XGB	RF	DT	SVM	MLP	KNN	LR
T23	T23	T17	T23	T15	T23	T23
T09	T09	T23	T06	T17	T09	T09
T24	T17	T09	T09	T07	T15	T06
T17	T06	T07	T07	T25	T25	T17
T22	T25	T22	T15	T22	T22	T15
T25	T24	T15	T25	T08	T20	T07
T10	T15	T06	T17	T24	T07	T25
T15	T22	T24	T22	T11	T19	T12
T06	T10	T08	T08	T13	T17	T10
T07	T11	T03	T10	T09	T14	T11
T11	T05	T20	T02	T23	T10	T20
T12	T12	T11	T03	T04	T11	T05
T08	T08	T25	T20	T20	T08	T22
T14	T07	T10	T16	T02	T06	T18
T01	T18	T12	T18	T05	T18	T04
T13	T03	T18	T05	T14	T13	T03
T05	T14	T05	T04	T18	T04	T24
T03	T21	T19	T19	T12	T24	T16
T21	T20	T14	T14	T03	T03	T21
T20	T19	T01	T12	T10	T21	T13
T16	T02	T02	T11	T16	T02	T14
T02	T01	T16	T24	T06	T05	T08
T19	T04	T13	T21	T01	T16	T19
T04	T16	T21	T13	T19	T12	T02
T18	T13	T04	T01	T21	T01	T01

Table 4.26: Tasks ranking for each ML Classifier

T_set	XGB	RF	DT	SVM	MLP	KNN	LR
3	82.5	79.23	76.16	81.22	76.32	77.40	79.35
5	79.81	80.35	76.10	81.88	75.37	75.21	79.95
7	79.46	81.17	77.30	80.26	74.53	75.06	79.72
9	79.67	80.10	77.75	79.99	74.37	74.99	78.10
11	78.82	79.70	77.58	80.32	75.54	74.84	78.02
13	77.94	78.35	77.14	78.64	75.06	74.33	77.67
15	77.86	78.08	76.80	77.76	73.60	72.66	76.42
17	77.60	77.24	76.92	77.20	72.43	72.71	76.69
19	77.19	77.75	76.37	76.83	72.23	72.37	76.36
21	76.74	77.34	75.42	77.24	72.51	72.58	75.56
23	76.54	77.07	75.53	77.49	71.73	72.04	74.58
ALL	76.39	76.64	75.19	76.99	71.93	71.76	74.04

Table 4.27: Majority vote to a different set of ranked tasks.

for a classification system, particularly in the medical domain. To better understand the observed behaviour, I comprehensively analysed our features. First, I examined the individuals' distribution in the dataset based on personal features. It is important to note that, from this point onward, I will refer to educational level as the number of years of school attended by an individual. Given that the dataset encompasses individuals aged between 44 and 88 years with educational levels ranging from 2 to 21 years of school, I investigated whether these personal features could significantly influence handwriting tasks. Table 4.28 illustrates the distribution of individuals in the dataset according to age and education (school years), with these characteristics categorized into two ranges.

Education level Distribution				Age Distribution			
School years	Total	HC	PT	Age intervals	Total	HC	PT
[2, 11]	70	22	48	[44, 66]	69	50	19
[12, 21]	104	63	41	[67, 88]	105	35	70

Table 4.28: Distribution of people according to Education level and Age.

Box plots in Figure 4.20 depict how the contact time feature varies with age, school years, and the presence or absence of AD. These plots refer to the 23rd task, which outperformed others according to Table 4.26. The figures facilitate the comparison of feature distribution between the two education ranges for a specific group. The x-axis represents the education range, the y-axis depicts the contact time feature in seconds, and each plot corresponds to a particular group (All, HC, PT) within a specific age range [44, 66] and [67, 88]. The following trends emerge from this analysis:

- Younger individuals are faster; those with fewer years of school take more time, with an increased deviation (Figure 4.20 (a) and (b)).
- There is no variation among healthy individuals in the first age range, regardless of education (Figure 4.20 (c)).
- Older healthy controls take longer than younger healthy controls, particularly if they have a lower education level. Education years appear to be significant for elderly individuals (Figure 4.20 (c) and (d)).
- The feature shows minimal change for impaired individuals in the first age range, irrespective of education (Figure 4.20 (e)).
- Elderly patients exhibit significantly different behaviours based on their education. Impaired elderly individuals take more time for handwriting tasks with fewer years of school (Figure 4.20 (f)).

These trends extend to other features, especially geometric ones related to contact time and the number of lognormals. This study reveals that a younger patient with Alzheimer's may perform a task similarly or even better than an older healthy person with fewer years of education. These findings offer insights into the outcomes obtained in the first step, highlighting a notable difference between young, healthy controls (c) and older patients (f). The difference diminishes significantly between older healthy controls (d) and younger patients (e) or older individuals in a higher education range (f).

There could be multiple reasons why healthy older individuals are not easily distinguished from younger patients. Older individuals, even without Alzheimer's, may have other impairments affecting their skills, as it is known that abilities deteriorate with age. Regarding younger patients, their performance depends on the disease stage, and they may be more used to compensate for its effects than older individuals who experience a loss of control. It would be intriguing to determine the disease stage of younger patients and understand if they have developed compensation mechanisms that render them similar to older healthy controls.

4.7 Complementary Results

This section describes further examinations involving images presented in Section 3.2 and approaches similar to those employed in the previously detailed works.

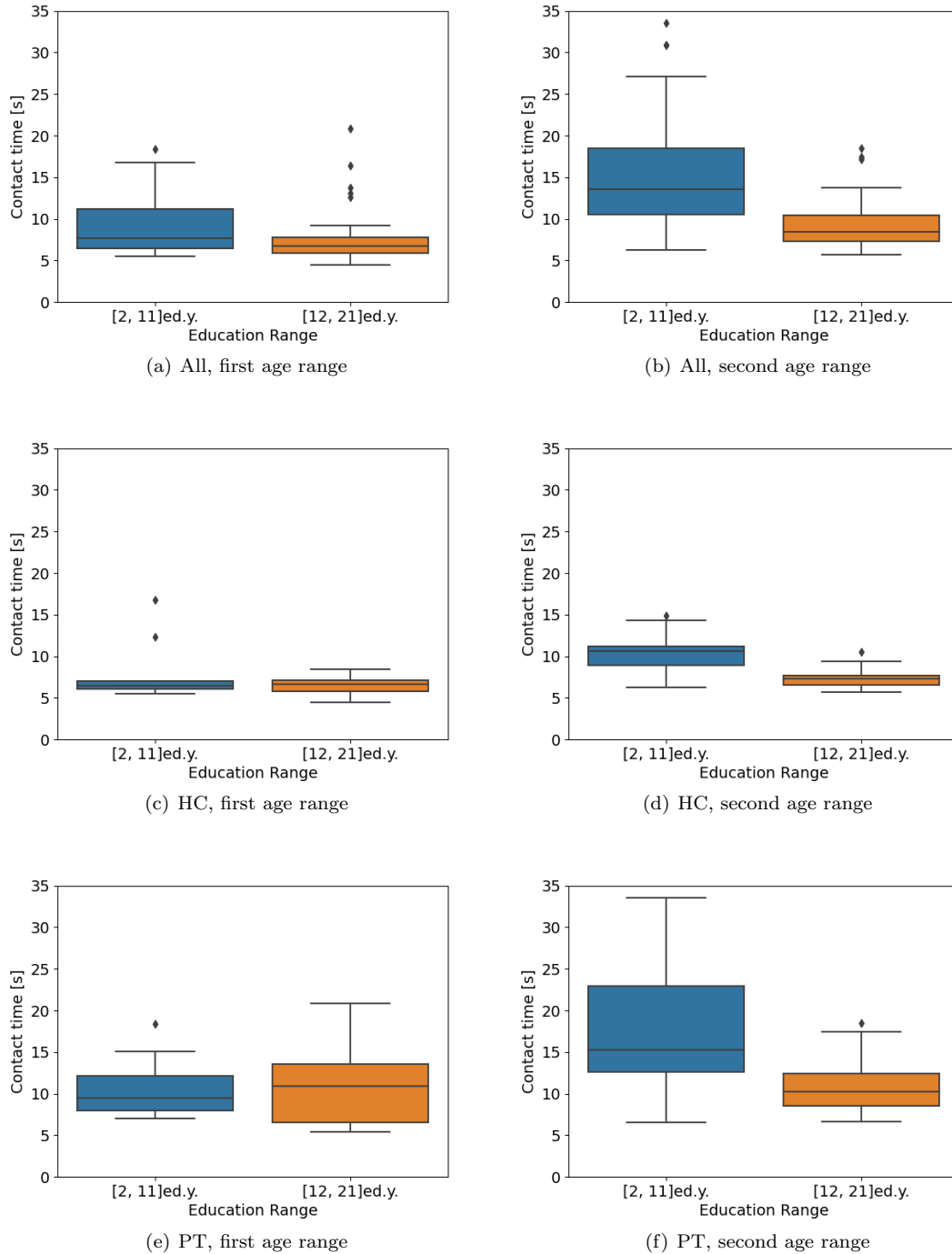


Figure 4.20: Box plots showing how the contact time feature is related to age and education. The first age range is from 44 to 66 years old, while the second is from 67 to 88 years old.

4.7.1 Comparison between Deep results achieved from Binary, RGB and Offline Images on Graphic Tasks

This experiment compares the results obtained by CNNs when employed in the classification through Binary, RGB on paper and Offline images concerning graphic tasks. The CNN models were the same as described in Section 4.1, so VGG19, ResNet50, InceptionV3 and InceptionResNetV2, without changing their hyperparameters setting. These models were pre-trained on Imagenet and trained on the target dataset by applying a 5-fold cross-validation strategy, and every model had the same FC classifier. Figure 4.21 illustrates the simple workflow implemented.

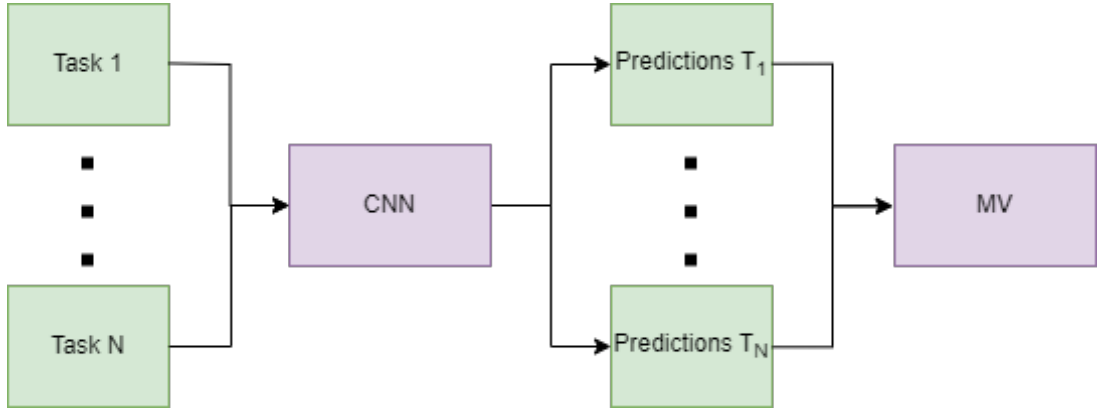


Figure 4.21: Experimental setting.

Table 4.29 shows the results expressed in accuracy (%) for every approach to ease the comparison. InceptionV3 achieves the best result on task 21 with offline images, which reported an accuracy of 74.4%. Moreover, it shows that offline outperforms the other approaches most of the time and that binary never exceeds the others.

Model	Task 2			Task 3			Task 4			Task 5			Task 21			Task 24		
	BIN	RGB	OFF	BIN	RGB	OFF	BIN	RGB	OFF	BIN	RGB	OFF	BIN	RGB	OFF	BIN	RGB	OFF
VGG19	50.1	64.1	64.6	48.1	59.0	65.4	42.6	70.7	67.2	50.6	64.7	69.6	61.2	64.8	61.7	66.2	70.4	69.8
Res.50	45.2	61.1	68.2	50.9	53.5	67.6	47.2	68.8	68.7	47.0	64.4	71.3	59.3	64.3	65.4	65.6	66.8	64.3
Inc.V3	52.4	64.2	68.7	51.9	57.1	65.5	44.4	68.0	65.4	45.2	65.7	70.3	57.5	58.8	74.4	56.2	62.4	60.8
Inc.V2	49.1	60.0	72.4	48.2	62.7	68.1	46.8	68.0	63.0	49.6	65.3	68.3	56.2	62.0	66.7	53.1	55.3	65.8

Table 4.29: Comparison among the evaluated deep approaches, considering binary, RGB and Offline images.

After the classification, I obtained predictions for each individual and task. To consolidate them, I employed a majority vote rule. Since there are six graphic tasks, an even number, I addressed instances of ties as cases to discard. Table 4.30 presents accuracy evaluated solely on non-tie cases and the rejection rate, indicating the percentage of cases where ties occurred. As expected, the best result was given by using offline images, particularly by the ResNet50 model, with a majority vote accuracy of 81.9% and the lowest rejection rate of 13.2%.

	VGG19			ResNet50			InceptionV3			InceptionRNV2		
	BIN	RGB	OFF	BIN	RGB	OFF	BIN	RGB	OFF	BIN	RGB	OFF
MV	68.2	73.6	73.5	69.9	78.2	81.9	71.3	74.4	78.8	71.0	72.8	77.0
REJECTED %	15.4	15.4	18.0	18.2	21.1	13.2	18.2	24.0	14.4	13.1	20.0	13.2

Table 4.30: Results of applying a majority vote rule with reject.

This study further affirms that binary images containing only an approximation of the real handwritten trace are not suitable for this study. On the contrary, if adequately exploited, RGB images can bring interesting outcomes, as they approximate the handwritten trace and encode dynamic information in the three colour channels. Finally, the best data type among the three analyzed seems to be offline images, depicting exactly the real handwritten trace executed by people.

4.7.2 Comparison between Deep results achieved from Multi-Channel Images

In this experiment, I studied the power of CNNs to extract valuable features from Multi-Channel images for detecting people suffering from AD through handwriting. In particular, I considered all the types of MC images generated, i.e. on paper, in air, and in air on paper. I used the same CNN models described for the experiment in Section 4.7.1.

Figure 4.22 shows the workflow implemented, where every task dataset was used to feed convolutional models, exploited in this case for training and testing too. Once the accuracy and the predictions were achieved, tasks were ranked in increasing order of accuracy. The final step involved the application of the majority vote rule on different subsets of tasks, i.e. all, the first seven and the first nine of the ranked list.

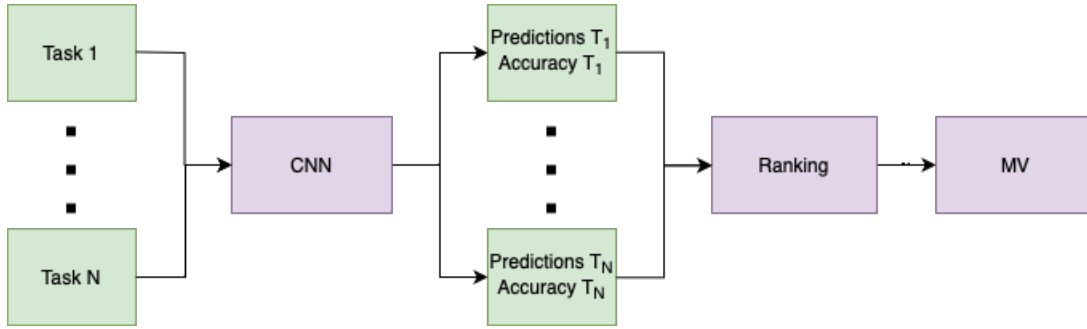


Figure 4.22: Experimental setting.

Table 4.31 shows the results of applying the majority vote rule on all the tasks or subsets. The table doesn't show intermediate results but only the outcome of the last step of the workflow, referring to the combining rule. The subsets are defined considering a list of tasks ranked in increasing order of accuracy. In detail, I considered two subsets of the first seven and the first nine tasks of the ranked list. The evaluation metrics reported are the accuracy and the FNR in percentage. The best results were achieved by combining the first nine tasks of the ranked list, reporting the best accuracy of 82.1% and a good FNR value for in-air on-paper images evaluated by InceptionV3. It is worth noting that the approach considering in-air and on-paper traits outperformed the others.

Data	Model	All task		7-task		9-task	
		Acc	FNR	Acc	FNR	Acc	FNR
On Paper	VGG19	70.6	20.2	77.0	31.0	78.1	27.0
	Res.50	71.7	20.2	71.2	23.6	70.6	25.8
	Inc.V3	75.1	20.2	77.0	20.2	81.0	14.6
In Air	VGG19	72.9	22.4	78.9	21.2	80.7	18.8
	Res.50	70.1	15.3	73.3	13.1	76.3	14.3
	Inc.V3	74.7	12.9	76.0	20.5	79.0	16.9
In Air On Paper	VGG19	76.8	30.3	77.1	34.1	78.1	28.1
	Res.50	71.7	12.4	75.2	21.3	73.5	20.2
	Inc.V3	75.1	9.0	81.6	20.2	82.1	18.0

Table 4.31: Accuracy and FNR results from applying the majority vote rule on all the tasks or subsets of them.

This experiment highlighted the power of MC images, in particular the in air on paper version. It also enhanced the improvements that can be achieved by judiciously applying a combination rule.

4.7.3 Comparison between Deep and Machine Learning results achieved from RGB Images

In Section 3.2.2, I described three types of RGB images. The following work focus on two of them, specifically on paper and in air on paper. In this experiment, I considered a hybrid system with deep networks to extract features evaluated by ML algorithms. Finally, as 34 tasks were considered, a majority vote rule with rejection was applied.

Figure 4.23 shows the experimental workflow. Four CNN models were used as feature extractors, as described in Section 4.1. Once extracted, the features for every task were used for classification with five

well-known ML algorithms. Before the classification, the Recursive Feature Elimination algorithm was applied to select the most valuable features of the dataset. Moreover, ML classifiers used a grid search technique to find the best hyperparameters. After the classification, a majority vote rule with reject was applied to the 34 tasks and used to compare the different approaches.

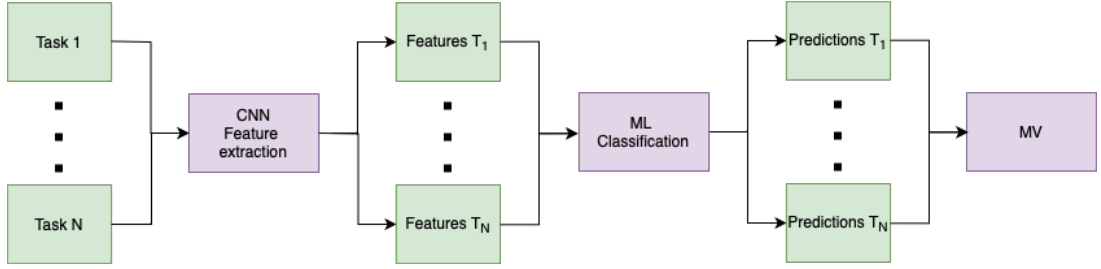


Figure 4.23: Experimental setting.

Table 4.32 shows the results obtained by the experimental setting previously described. No intermediate results are shown in this context. Results are very encouraging, showing the validity of the proposed system. In both the evaluated approaches, the best results were obtained by InceptionV3 with XGB classifier, reporting a majority vote accuracy of 90.2% for on-paper images and an accuracy of 91.1% for in-air on-paper images. It is worth noting that the proposed system shows very low reject rates.

	Model	XGB	RF	DT	SVM	MLP
On Paper	VGG19	81.4 (2.5)	81.6 (3.2)	78.1 (3.8)	73.6 (4.6)	79.3 (3.9)
	Res.50	88.3 (1.9)	85.2 (2.0)	84.3 (4.3)	69.4 (1.7)	84.5 (2.9)
	Inc.V3	90.2 (2.3)	89.5 (2.1)	84.4 (4.0)	77.4 (2.9)	86.1 (2.4)
	Inc.V2	84.6 (2.0)	82.3 (2.3)	79.3 (4.2)	67.3 (1.6)	80.7 (3.7)
In Air On Paper	VGG19	84.0 (3.5)	83.6 (2.2)	81.2 (4.6)	76.9 (3.2)	81.4 (3.5)
	Res.50	90.0 (2.4)	89.3 (1.8)	83.1 (4.9)	74.2 (2.6)	85.3 (3.5)
	Inc.V3	91.1 (2.0)	90.4 (2.0)	84.2 (4.3)	86.6 (2.7)	90.1 (2.2)
	Inc.V2	85.7 (2.7)	83.5 (3.1)	81.4 (4.4)	70.0 (2.3)	82.3 (2.5)

Table 4.32: Results of the majority vote rule with rejection.

This complementary experiment aimed to highlight the power of the various types of RGB images, in particular when a combination rule is applied.

Chapter 5

Evolutionary Algorithms

Evolutionary algorithms comprise optimization techniques inspired by the theory of biological evolution. Their development mimics natural selection to improve solutions for complex problems iteratively. They evolve candidate solutions over successive generations by maintaining a population of potential solutions and applying genetic operators. These versatile algorithms have been successfully applied in various domains, including optimization, machine learning, and neural architecture search. Their adaptive and explorative nature makes them well-suited for solving problems with complex search spaces and unknown or dynamic landscapes.

In my exploration of AD research, I extensively utilized evolutionary algorithms to achieve various objectives. In Section 5.1, I applied a GA to optimize the prediction capabilities of a DL system, enhancing its efficacy in supporting the diagnosis of cognitive impairment. In Section 5.2, I delved into an evolutionary approach for Neural Architecture Search (NAS), aiming to tailor the network architecture to best suit the dataset of handwriting samples for AD diagnosis. Lastly, in Section 5.3, I employed evolutionary algorithms for feature selection on a dataset of handwriting features in the context of AD diagnosis using ML.

5.1 Using Genetic Algorithms to Optimize a DL-based System for the Prediction of Cognitive Impairments

In Chapter 4, I conducted experiments to develop a system to support the diagnosis of AD through handwriting analysis. In particular, in Section 4.3, I made an investigation based on DL and ML techniques applied on synthetic MC on paper images, described in Section 3.2.3. The generated data fed four distinct CNNs. Evolutionary algorithms have demonstrated their efficacy as search tools in addressing numerous real-world challenges characterized by extensive and nonlinear search spaces [27, 34, 33], and they also found widespread application in health-related domains. In this section, I introduce a system employing a GA to enhance the performance of the aforementioned deep architectures [26]. The aim is to identify the optimal subset of tasks that enhances the predictive capability of the networks through the GA. Experimental results affirm the efficacy of the proposed approach.

Experimental Setting and Results

The previously mentioned in-air on-paper MC images, acquired for each task and subject, were organized into a dataset. In this experiment, images from every task and those belonging to the additional tasks were considered, referring to Table 3.2, summing up to 34 tasks. To mitigate overfitting and statistically enhance the network's accuracy, I employed the 5-fold Cross-validation technique. Subsequently, each task dataset fed three distinct CNN models: VGG19, ResNet50, and InceptionV3. These models, pre-trained on the ImageNet dataset, automatically extracted features. The models shared a common classifier with two fully connected layers specifically configured for binary classification, healthy controls and patients, deviating from the ImageNet dataset's thousands of classes. The classification step iterated over the 34 tasks, resulting in 34 predictions and their corresponding confidence levels for each subject after this process. The implemented experimental setting is depicted in Figure 5.1.

The output of each CNN is a binary prediction and its corresponding confidence degree. Once obtained the prediction, I organized two datasets:

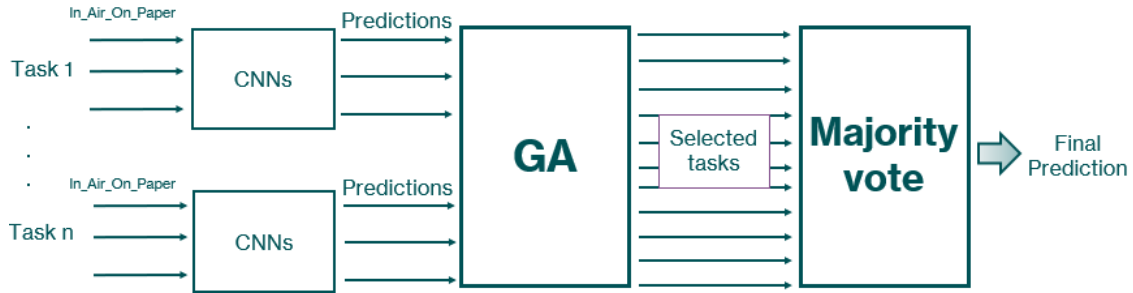


Figure 5.1: The layout of the proposed system. Note that in our case $n = 34$.

- Normal, where every sample referred to a person and consisted of a feature vector containing the 34 binary predictions;
- Weighted, where every sample referred to a person and consisted of a feature vector containing the 34 predictions weighted with the corresponding confidence degree;

The two datasets obtained were divided into two statistically independent sets: a training set (T_r) comprising 80% of the available samples and a test set (T_s) encompassing the remaining samples. I implemented a GA to select the optimal subset of 34 tasks for both approaches to enhance the system’s predictive performance regarding participants’ cognitive state. The chromosomes of this algorithm were binary vectors of 34 elements, where each value is a 0, representing the exclusion or a 1, representing the inclusion of the task corresponding to the position of that value. The evaluation of the i -th individual, representing the tasks subset s_i , consisted in the computation of its fitness, defined as a majority vote rule applied considering only the tasks included in s_i . The GA implementation utilized a generational evolutionary algorithm, initiating with the random generation of a population of P individuals. Subsequently, the fitness of these individuals was assessed by computing the prediction accuracy on the training set. After this preliminary evaluation phase, a new population is generated by selecting $P/2$ pairs of individuals using the tournament selection method of size t . The one-point crossover operator is then applied to each selected pair according to a given probability factor p_c , followed by the mutation operator with a probability of p_m . Finally, these individuals are added to the new population. This process is iterated over a number of N_g generations. The parameters of the GA are detailed in Table 5.1.

Parameter	Symbol	Value
Population size	P	100
Crossover probability	p_c	0.6
Tournament size	t	5
Elitism	e	2
Mutation probability	p_m	0.03
Number of Generations	N_g	1000

Table 5.1: The values of the parameters used for the GA.

T_r was used to evaluate the fitness of individuals generated from the GA, T_s was employed to evaluate the best individual’s performance on unseen data. For every dataset, I performed thirty runs, and after each run, the task subset encoded by the individual with the highest fitness was preserved as the solution for that particular run. The results presented herein are obtained by averaging the outcomes from the thirty best individuals stored.

I conducted three sets of experiments. First, I examined the generalization capability and task reduction effectiveness of the GA-based system for predicting cognitive impairments. This involved plotting the training and test accuracy of the best individual, along with the average number of selected tasks (across the entire population) and the number of selected tasks by the best individual in the thirty runs. Figures 5.2 and 5.3 present the outcomes from the first set of experiments, illustrating the evolution of (i) average training (blue) and test (red) accuracy of the best individual; (ii) average number of selected tasks for the best individual (green) and the entire population (yellow), computed by averaging values from the thirty runs. These plots reveal the impact of overfitting because although the training and the test accuracy increase, the system exhibits superior performance on training samples compared to unseen test samples, indicating a need for enhanced generalization ability. This trend holds across different networks and

approaches. Notably, the best accuracy trend is observed with weighted predictions from InceptionV3, minimizing the gap between train and test accuracy, with the test accuracy reaching 71%. However, analyzing other trends from the plots highlights the system’s effective task-reduction capability. The number of selected tasks, both average and best, decreases with increasing generations, with less than half of the tasks selected. The best result is observed with weighted predictions from InceptionV3, where the lowest number of selected tasks is 14. Additionally, it’s noteworthy that the number of selected tasks by the best individual is consistently lower than the average. This supports the idea that a better performance can be achieved by selecting a subset of available tasks.

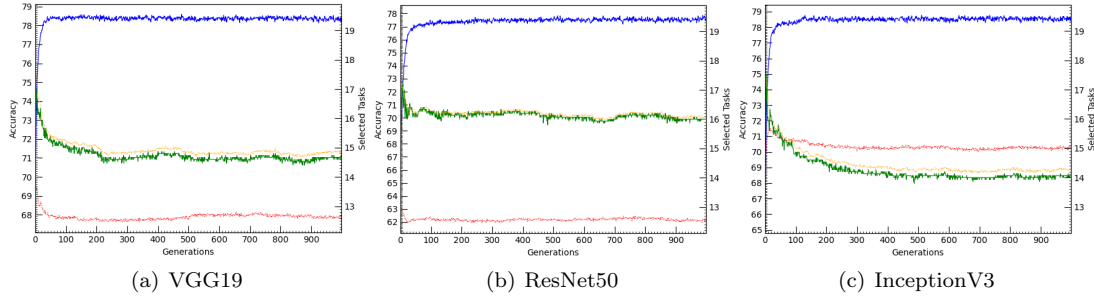


Figure 5.2: Evolution of accuracy and average number of selected tasks for every model of CNN, considering weighted predictions.

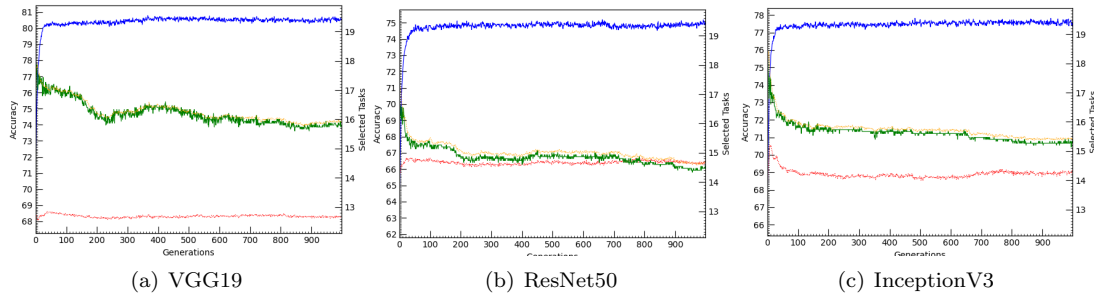


Figure 5.3: Evolution of accuracy and average number of selected tasks for every model of CNN, considering non-weighted predictions.

In the second set, I analyzed the frequency of selected tasks to identify the most relevant ones. Figure 5.4 presents histograms from the second set of experiments, displaying the frequency of selected tasks across the thirty runs. Each of the thirty-four tasks has a corresponding bar, indicating the number of times it was chosen collectively across all CNNs. The first plot refers to normal predictions, while the second refers to weighted predictions. The two plots exhibit a similar trend. It is worth noting that some tasks, such as 1 (signature), 5 (circles), 6 (copy of letters), 22 (copy of phone number), 24 (clock drawing test), and 33 (copy of word), are more frequently selected than others. Many of these tasks involve graphic or copy-related activities, underscoring the effectiveness of our protocol. Graphic tasks assess subjects’ proficiency in writing elementary traits, connecting points, and drawing figures, while copy tasks evaluate their ability to replicate complex graphic gestures or motor planning.

The third experiment involved comparing the results of the proposed approach with those obtained using majority-vote and weighted majority-vote rules applied to all tasks. Table 5.2 illustrates the comparison results, demonstrating that, in most cases, the majority-vote rule applied to GA-selected tasks surpasses the accuracy of the rule applied to the entire task set. Even if the increment of accuracy is not so significant, it is worth noting that this experiment supported the notion that a reduced number of tasks, depending on the selected ones, can perform similarly or even better than considering all the tasks, reducing the computational amount of time and resources needed.

5.2 An evolutionary approach for Neural Architecture Search

DNNs have emerged as a standard tool in various applications, but despite their success, a significant challenge lies in the absence of a universally accepted set of criteria for selecting the architecture that

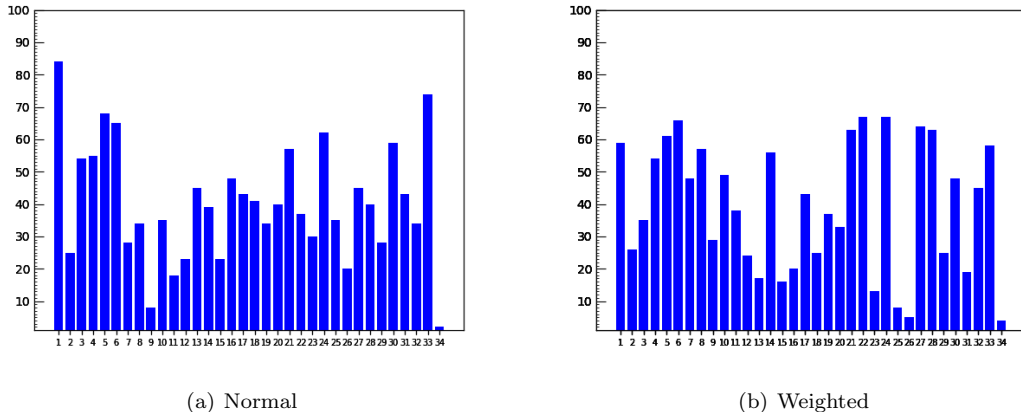


Figure 5.4: Comparison of the number of occurrences of the selected tasks for the three CNN between the normal and the weighted approach.

Normal	GA MV	MV
VGG19	69.30	69.06
ResNet50	68.16	65.74
InceptionV3	70.62	70.16
Weighted	GA MV	MV
VGG19	69.04	72.92
ResNet50	63.99	69.06
InceptionV3	72.27	71.82

Table 5.2: Comparison of results obtained by applying the majority vote rule on a subset of tasks selected by the GA or on all the tasks.

suits a specific problem. This represents a limitation as the chosen architecture and the setting of its hyperparameters significantly influence DNNs’ performance. Typically, experts manually design these architectures, a labour-intensive process requiring a high level of expertise due to its trial-and-error nature. To address this challenge, NAS [148, 113] has gained attention as a technology capable of automatically designing DNN architectures. Evolutionary Computation (EC) approaches have shown notable success among the diverse NAS methods [83]. The following work introduces an approach based on evolutionary computation to optimize CNNs.

Proposed Method

This research focuses on CNNs, consisting of three types of layers: Convolutional layers, Pooling layers, and Fully Connected layers. Table 5.3 outlines each layer type’s variants and associated hyperparameters. Due to the interconnected nature of these hyperparameters, determining the optimal set becomes a challenging task.

NAS emerges as an advanced algorithm in cases involving many hyperparameters or architectural optimisation. Specifically, a variant of NAS, involving an evolutionary algorithm, was chosen for this study. The proposed solution is an evolutionary algorithm designed to optimize a network’s hyperparameters and architecture. Before employing the evolutionary algorithm, certain parameters influencing the operation of evolutionary operators need to be set. These parameters encompass population size, crossover and mutation probabilities, and the number of generations. The evolutionary algorithm is made of multiple steps and is outlined as follows:

1. Generate the initial population of individuals randomly (First generation):
 - Decode the chromosomes and synthesize neural networks;
 - Evaluate the fitness (validation accuracy) of each individual (net) in the population. Each net undergoes training, and the validation accuracy is assessed.
2. Repeat the subsequent generational steps until the termination condition is satisfied:

Layer	Variants	Hyperparameters
Convolutional	Residual block Inception block	Input/Output channels
		Padding
		Stride
		Dilation
		Kernel size
		Activation function
Pooling	Average Max	Input/Output channels
		Pool size
		Padding
		Stride
		Dilation
Fully Connected		Number of layers
		Number of neurons
		Dropout
		Activation function
Model		Optimizer
		Learning rate
		Loss
		Metrics
		Number of epochs
		Batch size

Table 5.3: CNN layers and their hyperparameters.

- Select the fittest individuals for reproduction (Parents).
- Breed new individuals through crossover and mutation operations to generate offspring.
- Evaluate fitness.
- Replace the least-fit individuals of the population with new individuals (Next generation).

Each individual is encoded using a set of hyperparameters and the network architecture. Therefore, the chromosome is divided into two parts. Hyperparameters coding:

- Learning rate: governs the algorithm's update pace or learning of parameter values.
- Max learning rate.
- Learning rate gamma: implements a learning rate decay mechanism during training.
- Gradient clipping: constrains the gradient's range of variation.
- Weight decay: incorporates a weight decay mechanism during training.
- Batch size: denotes the number of samples processed before model update.
- Dropout: addresses the undesired phenomenon of overfitting.

Every hyperparameter gene consists of its actual value; instead, the architecture can't be coded in a unique value. Regarding the architecture, every layer refers to a sequence of genes. In the case of a fully connected layer, a gene is enough to represent the number of neurons; on the contrary, for layers coming from the convolutional part of the architecture, the following genes are needed:

- Type: determines the layer type (convolutional or residual).
- Param: indicates the number of convolutional layers or the type of residual block (basic or bottleneck), depending on the first element's value.
- Double channel: a boolean indicating whether to double the number of output features.

Each individual is initialized with values chosen from a discrete group of possible values. Concerning the architecture, there are minimum and maximum chromosome length constraints. As the population ages, the resolution of the search space increases, allowing individuals to settle in the neighbourhood of a local optimum and refine hyperparameter optimization. Concerning evolutionary operators, selection,

mutation and Crossover were applied in the algorithm, and after that, the offspring was generated. Finally, the evaluation of each individual involves a chromosome decoding, necessary to synthesize the neural network, configuring the architecture and hyperparameters. The obtained model is trained, and its fitness (validation accuracy) is evaluated.

Experimental Results

I conducted three experiments to evaluate the effectiveness of the proposed system, each involving a different input dataset. The goal was to assess how ENAS performs with images of varying sizes, nature, and datasets from diverse tasks with varying sample and class numbers. Initially, I tested the system on well-known benchmark datasets, MNIST [78] and CIFAR-10 [77], followed by an evaluation of a handwriting Alzheimer dataset. Datasets details are shown in Table 5.4. Concerning the handwriting Alzheimer dataset RGB on-paper images (Section 3.2.2) were considered for graphic tasks of the protocol (Section 3.1). The first two tasks consisted of joining two points 5cm apart with a straight continuous horizontal (task 1) or vertical (task 2) line continuously four times. The third and fourth tasks consisted of retracing a 3cm (task 3) or 6cm (task 4) wide circle four times. The fifth task consisted of retracing a complex form (task 5), and the sixth was the well-known clock drawing test (task 6). The results, often reaching state-of-the-art performance, are promising. The experiments were conducted on an Intel Xeon i7-7700 CPU @ 3.60 GHz and Intel Xeon Silver 4110 @ 2.10 GHz, with 377GB of RAM and a Tesla V100 GPU. PyTorch 3.6.9 on Ubuntu 18.04.3 LTS served as the framework.

Dataset	#samples	Size	type	#classes
CIFAR-10	60,000	32×32	RGB	10
MNIST	60,000	28×28	BN	10

Table 5.4: Benchmark datasets.

Regarding experiments on MNIST and CIFAR-10, some hyperparameters and rules were predefined to expedite convergence toward a local optimum. The chosen optimizer is SGD, the scheduler is StepLR, and random rotation is the only applied data augmentation. MNIST experiments achieved performance comparable to state of the art, with a validation accuracy of 99.7% while, as shown by the learning curves of the best individuals obtained at the end of the evolution, in Figure 5.5.

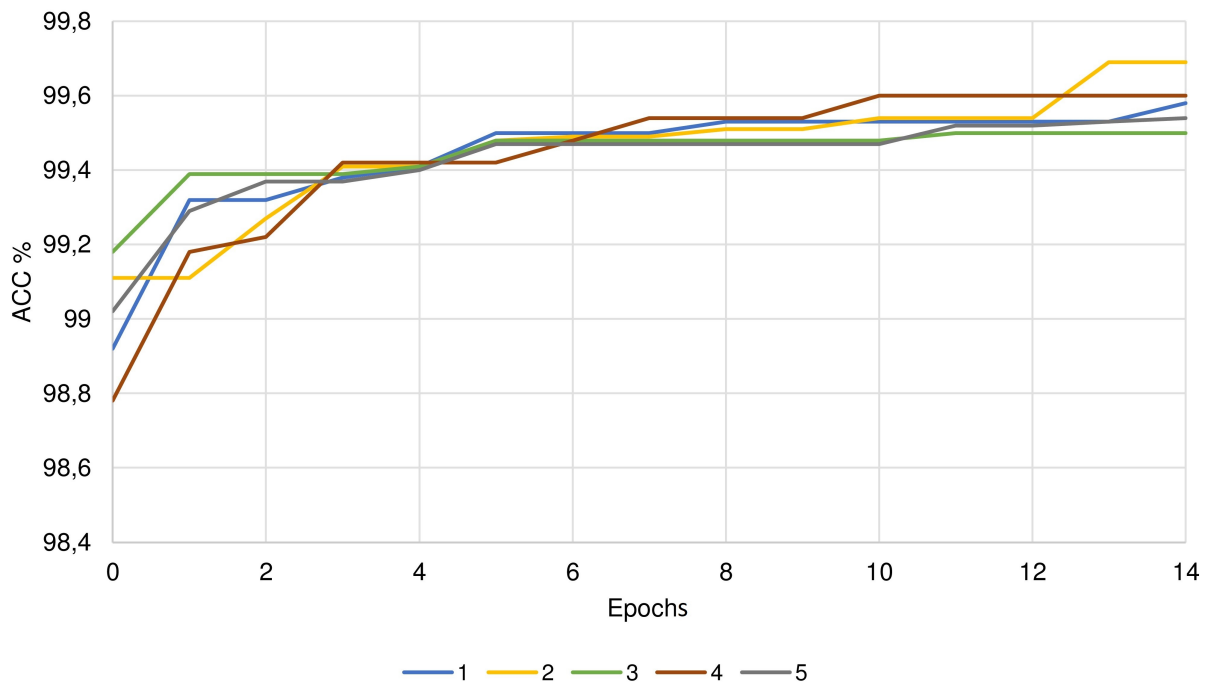


Figure 5.5: Best individuals.

Fitness reached a maximum of 99.69% after the 18th generation, and execution time showed variations

due to algorithmic improvements, as shown in Figure 5.6 (a). Figure 5.6 (b) shows the trend of times over the generations. Here, it is worth noting a detail: for the first part of the evolution, the time curve grows, then it decreases, to keep on growing again in the last five generations. This trend is the result of some improvements in the evolutionary algorithm. During the generations, the fitness value for each individual was stored along its chromosome. This prevents a further evaluation when an already evaluated chromosome shows again.

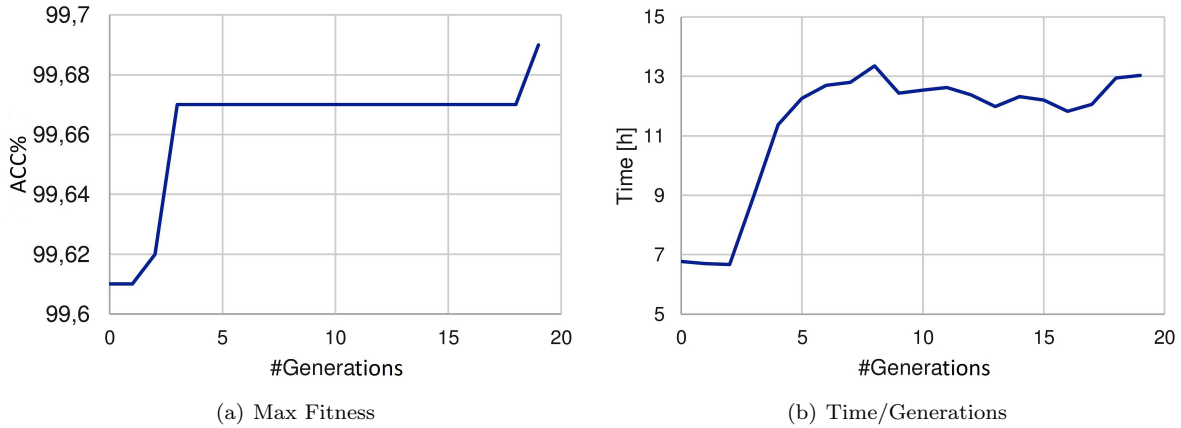


Figure 5.6: MNIST results.

Although CIFAR-10 is not reaching the state-of-the-art accuracy (99.61%), the algorithm exhibited increasing validation accuracy across generations, as shown in Figure 5.7. The execution time initially decreased, indicating evaluations skipped for previously seen chromosomes, but later increased to counter being stuck near a local optimum. To address this, a mechanism was implemented to enhance resolution and generate new chromosomes near the local optimum, causing an upward trend in execution time.

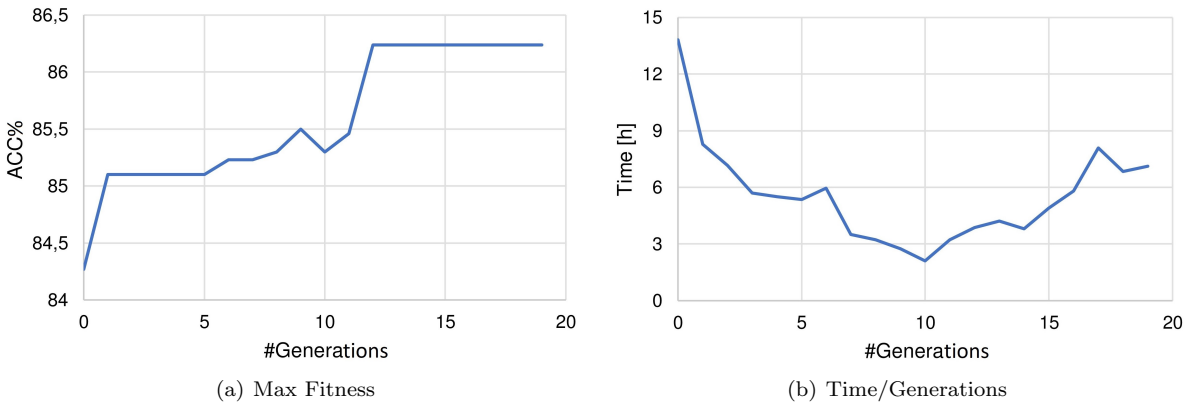


Figure 5.7: CIFAR-10 results.

Before ENAS execution on the Alzheimer dataset, specific hyperparameters and rules were predefined; Adam was chosen as the optimizer, OneCycleLR as the scheduler, batch size set at 4, and no data augmentation was applied. This choice aimed to expedite convergence towards a local optimum. ENAS operated on one task dataset at a time, and after evolution, the best individual’s performance was evaluated. For the classification phase, multiple experiments were conducted using a 5-fold cross-validation strategy, with the dataset split into balanced training, validation, and test sets. Figure 5.8 illustrates the training and validation accuracy trends for the best individual selected by ENAS for each task.

To highlight the ENAS approach, a comparison was conducted with well-known Convolutional Neural Networks (CNNs) such as VGG19, ResNet50, InceptionV3, and InceptionResNetV2. Table 5.5 presents the obtained accuracy values, with the last line representing ENAS, consistently outperforming other

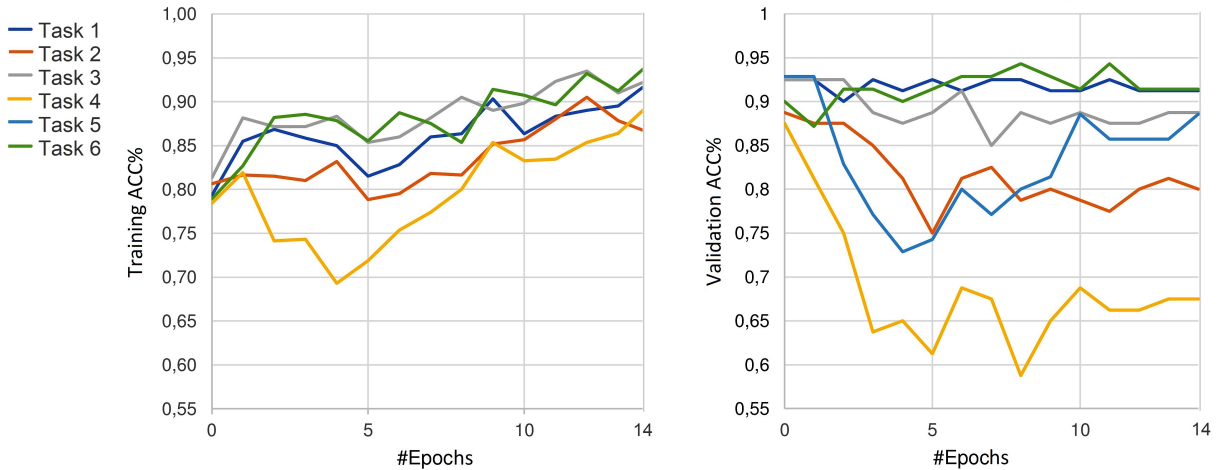


Figure 5.8: Training and validation accuracy of the best individuals for each task

CNNs for every task. ENAS achieved the best overall accuracy (84.57%) on the third task, but for every task, it outperformed the CNNs outcomes.

Model	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
VGG19	67.56	57.64	69.4	67.24	64.59	67.49
ResNet50	59.34	60.5	70.41	64.38	65.69	68.17
Inc.V3	60.98	54.73	72.94	67.71	68.12	61.47
Inc.Res.V2	64.41	61.48	69.97	62.38	67.68	64.53
ENAS	79.93	69.66	84.57	71.62	72.01	83.6

Table 5.5: Accuracy comparison among CNNs and ENAS

The proposed approach couldn't outperform the state of the art on well-known and huge datasets like MNIST or CIFAR. Besides this, it showed outstanding performances on the small dataset of handwriting samples related to Alzheimer's disease compared to results obtained by experts with known architectures.

5.3 Integrating Data Augmentation in Evolutionary Algorithms for Feature Selection

In many ML problems, the presence of hundreds or even thousands of features represents a challenge in identifying the most pertinent subset, as not all available features always introduce valuable information. In this research, I investigate the effect of data augmentation on the performance of evolutionary algorithms when implied in feature selection procedures, particularly regarding GA and Particle Swarm Optimization (PSO). Comparative analyses with two established feature selection algorithms were conducted and tested on several publicly available datasets.

Experimental Settings and Results

This study aims to explore the impact of data augmentation on Evolutionary Computation-based techniques employed in feature selection [59, 19, 141, 27]. Multiple experiments were conducted considering six datasets from various application domains and presenting distinct characteristics regarding sample sizes, features, and classes. Table 5.6 provides a more detailed overview.

Aiming to evaluate the proposed approach combining data augmentation and evolutionary feature selection techniques, I performed three experiments presented in increasing order of complexity. In each experiment conducted for every dataset, I performed twenty runs. Figure 5.9 depicts the overall experiment workflow.

The initial implementation served as a baseline case where datasets were directly employed for classification. In contrast, the second experiment utilized four distinct feature selection methods. The third experiment introduced a data augmentation module. To evaluate the efficacy of GA and PSO, I compared the best results achieved for the three experimental settings presented. Performance assessment was done

Dataset	#Samples	#Features	#Classes
Hand (Sec. 3.3.1)	174	90	2
Isolet [28]	7797	617	26
Mfeat1 [49]	2000	216	10
Mfeat2 [49]	2000	64	10
Ozone [143]	4748	72	2
Toxicity [61]	171	1203	2

Table 5.6: The datasets used in the experiments.

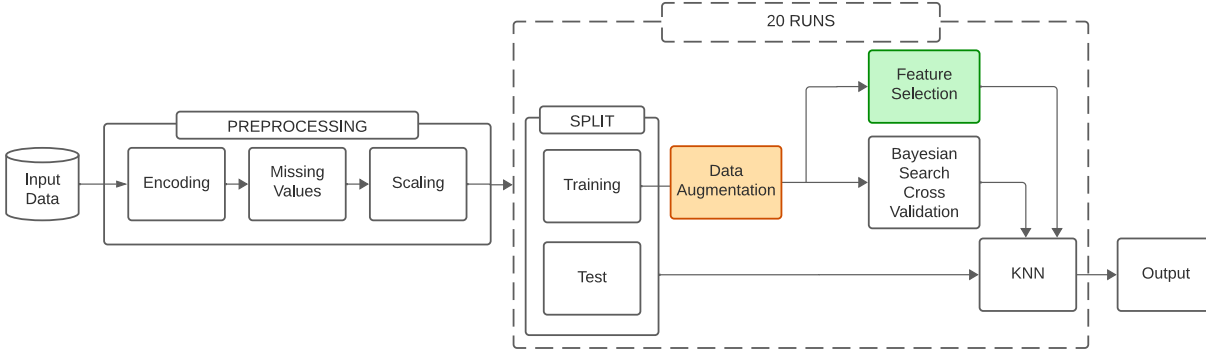


Figure 5.9: Experimental workflow.

considering the average accuracy and its standard deviation over the twenty runs, providing insights into the stability of experimental results. The average number of selected features is also reported, a crucial parameter in machine learning impacting efficiency, interpretability, and model performance.

The first experimental setup provided a fundamental reference point for comparison with other more time-consuming and resource-intensive experiments. Figure 5.9 illustrates the final experimental workflow, with the baseline case represented as the uncoloured section.

First, the input dataset went through a preprocessing phase to ensure data quality and align the dataset with the standards expected by the subsequent classification algorithm. This process involved three operations: encoding categorical features, handling missing values, and scaling all features. The first two operations were selectively applied where necessary. Then the dataset was split into training (80%) and test (20%) sets. Subsequently, the training set underwent Bayesian search [8] to optimize the hyperparameters of the classification algorithm, implemented using a 5-fold cross-validation strategy. The chosen supervised machine learning algorithm was KNN, which, after hyperparameter optimization, was trained on the training set and tested on the test set. Table 5.7 displays the results obtained for each dataset.

Dataset	AVG	STD
Hand	58.8	7.1
Isolet	90.7	0.6
Mfeat1	95.7	1.1
Mfeat2	95.3	1.1
Ozone	94.1	3.8
Toxicity	67.4	5.5

Table 5.7: Baseline experiment results in average accuracy and standard deviation computed over 20 runs.

The second experimental setting is like the baseline with adding a feature selection module, represented in green in Figure 5.9. Before the Bayesian search, the training set was utilized for feature selection (FS), employing various FS techniques to identify the most suitable method, balancing resources and performance. In particular, this experiment compares conventional FS methods with those proposed and based on evolutionary algorithms, specifically PSO and GA.

PSO is an evolutionary computation technique inspired by swarm behaviour applied to feature selection for machine learning. PSO iteratively updates particles' velocities and positions by using a population of particles, each representing a candidate solution with a position vector corresponding to a feature subset. Fitness is determined by a sigmoid function, guiding feature inclusion or exclusion. PSO balances exploration and exploitation through personal and global best positions, converging towards an optimal

feature subset. The algorithm terminates after a set number of iterations, returning the best feature subset for enhanced predictive model performance. Table 5.8 shows the parameters setting for the algorithm.

Parameter	Value
Swarm Size	30
Cognitive Coefficient (ϕ_1)	2.0
Social Coefficient (ϕ_2)	2.0
Number of Iterations	60
Particle Position Limits (p_{min}, p_{max})	$[-1.0, 1.0]$
Velocity Limits (s_{min}, s_{max})	$[-1.0, 1.0]$

Table 5.8: PSO parameters setting.

GA for feature selection efficiently explores the best subset of features for maximum prediction accuracy. Using binary string encoding, the GA evolves solutions over generations, employing selection, crossover, and mutation operations. The process refines feature subsets, with the best-performing subset identified to create an accurate and parsimonious model, preventing overfitting. The GA’s effectiveness is evaluated based on the quality and consistency of the final feature subset across multiple runs, offering flexibility for various data types and models, enabling an intelligent search for optimal feature spaces and revealing hidden interactions. GA’s parameters setting is shown in 5.9.

Parameter	Value
Population Size	50
Crossover Probability	0.6
Mutation Probability	$1/\text{\#features}$
Tournament size	3
Number of Generations	40
Elitism	Keep best

Table 5.9: GA parameters setting.

Two well-known FS algorithms were considered for comparison: RFE and SelectKBest (SKB). Table 5.10 presents the results of the tested feature selection methods, with the best outcomes highlighted in bold. Notably, PSO achieved the best performance, surpassing others in four out of six cases in accuracy and the number of selected features. While PSO demonstrated superiority, the table shows that PSO and GA performance was comparable to the common and FS methods tested.

Dataset(#features)	RFE		SKB		PSO		GA	
	Acc	#feat	Acc	#feat	Acc	#feat	Acc	#feat
Hand (90)	57.7 ± 7.5	45	58.0 ± 6.9	45	58.8 ± 6.8	33.3	58.4 ± 8.0	40.6
Isolet (617)	88.8 ± 1.4	308	88.3 ± 2.1	307	92.1 ± 0.9	291.8	91.5 ± 0.7	302.1
Mfeat1 (216)	95.0 ± 1.7	108	95.4 ± 1.6	109	95.9 ± 0.9	90.9	95.8 ± 1.1	103.3
Mfeat2 (64)	96.0 ± 1.0	32	96.2 ± 0.9	31	95.3 ± 1.0	31.5	95.3 ± 1.2	36.1
Ozone (72)	91.8 ± 3.2	36	92.3 ± 3.9	37	92.4 ± 3.3	29.9	94.2 ± 2.8	32.6
Toxicity (1203)	64.8 ± 5.8	600	64.6 ± 6.6	601	67.1 ± 7.5	509.9	64.4 ± 8.7	597.9

Table 5.10: Feature selection results in average and standard deviation accuracy and average number of selected features computed over 20 runs.

The third experiment, depicted in Figure 5.9, adds a data augmentation module (orange box) to the second experimental setting. After the dataset splitting, the training set underwent augmentation before feature selection, Bayesian search, and training of the KNN algorithm. Three augmentation percentages (10%, 20%, and 30%) were tested. Data augmentation is a common practice in ML to enhance model generalization by addressing imbalanced datasets. The implemented algorithm generates new samples by iteratively perturbing random samples within class-wise constraints, ensuring adherence to the original dataset feature distribution. The process continues until a predetermined augmentation percentage is reached. This approach introduces minimal modifications, ensuring class balance and adjusting sample numbers per class as necessary. Table 5.11 presents the results, highlighting that, in general, increased augmentation percentages correlate with improved performance. Exceptions include the Ozone dataset with the GA method and the Isolet dataset in conjunction with RFE and SKB. However, the highest performance for each dataset was consistently achieved with the maximum augmentation percentage.

Evolutionary feature selection methods outperformed others in four of six datasets, with comparable performance in the remaining cases. Data augmentation did not impact the number of features selected for all algorithms. Comparing PSO and GA, GA tended to select more features than PSO.

Dataset (#features)	DA(%)	RFE		SKB		PSO		GA	
		Acc	#feat	Acc	#feat	Acc	#feat	Acc	#feat
Hand (90)	0	57.7 ± 7.6	45	57.9 ± 6.9	45	58.8 ± 6.8	33.3	58.4 ± 8.1	40.6
	10	61.7 ± 7.0	45	60.1 ± 8.0	46	62.7 ± 7.4	38.0	62.8 ± 8.7	40.7
	20	65.6 ± 6.6	45	64.5 ± 8.5	44	66.3 ± 9.9	33.1	67.6 ± 6.8	40.1
	30	66.7 ± 7.2	44	64.0 ± 9.5	45	68.3 ± 6.2	34.3	69.6 ± 5.7	38.4
Isolet (617)	0	88.8 ± 1.4	308	88.3 ± 2.1	307	92.1 ± 0.9	291.8	91.5 ± 0.8	302.1
	10	82.2 ± 4.6	307	81.7 ± 3.6	308	92.0 ± 0.7	294.9	91.5 ± 1.0	301.5
	20	85.5 ± 3.8	308	85.9 ± 3.4	306	92.6 ± 0.7	295.6	92.4 ± 0.8	304.1
	30	88.9 ± 2.2	308	87.7 ± 2.5	308	92.9 ± 0.9	292.3	92.7 ± 0.7	299.4
Mfeat1 (216)	0	95.0 ± 1.7	108	95.4 ± 1.6	109	95.9 ± 0.9	90.9	95.9 ± 1.1	103.3
	10	95.5 ± 1.4	107	95.6 ± 0.9	109	96.2 ± 1.0	94.5	96.5 ± 0.8	101.2
	20	96.2 ± 0.9	108	96.3 ± 0.9	109	96.5 ± 0.8	89.5	96.5 ± 0.6	105.1
	30	96.5 ± 0.9	106	96.5 ± 0.9	109	96.7 ± 0.6	90.8	96.9 ± 0.9	104.6
Mfeat2 (64)	0	96.0 ± 1.0	32	96.2 ± 0.9	31	95.3 ± 1.1	31.5	95.3 ± 1.2	36.1
	10	96.0 ± 1.1	32	96.1 ± 1.4	33	95.4 ± 1.4	32.0	95.2 ± 1.4	35.6
	20	96.4 ± 0.9	32	95.6 ± 0.9	34	96.1 ± 1.0	32.4	96.2 ± 0.8	36.2
	30	96.1 ± 1.3	32	96.5 ± 0.8	32	95.7 ± 1.3	32.1	96.2 ± 1.0	34.7
Ozone (72)	0	91.8 ± 3.2	36	92.3 ± 3.9	37	92.4 ± 3.4	29.9	94.2 ± 2.9	32.6
	10	92.1 ± 3.1	37	93.3 ± 3.4	35	93.1 ± 2.4	30.2	92.8 ± 2.8	33.7
	20	93.0 ± 2.8	36	93.4 ± 3.2	35	94.1 ± 2.6	29.2	93.4 ± 2.5	33.7
	30	92.9 ± 2.4	36	93.2 ± 2.5	35	94.8 ± 2.1	30.0	94.6 ± 2.2	33.1
Toxicity (1203)	0	64.8 ± 5.8	600	64.6 ± 6.6	601	67.1 ± 7.5	509.9	64.4 ± 8.7	597.9
	10	64.5 ± 6.6	601	67.9 ± 7.5	602	66.1 ± 5.5	511.3	62.2 ± 4.2	593.1
	20	66.7 ± 9.5	601	66.9 ± 7.9	600	65.3 ± 6.3	502.7	63.4 ± 9.1	592.0
	30	70.1 ± 8.9	601	67.1 ± 9.5	600	69.6 ± 8.0	519.8	68.7 ± 8.0	593.3

Table 5.11: Data augmentation and feature selection results in average accuracy and standard deviation computed over 20 runs for every FS technique and DA percentage.

To evaluate GA and PSO effectiveness, I compared the best results with data augmentation to baseline and no feature selection results. The non-parametric Wilcoxon rank-sum test validated the comparisons. Bold values in Table 5.12 represent better results, while starred results indicate no statistically significant difference. GA and PSO achieved the best significant results on three datasets (Hand, Isolet, Mfeat1), notably improving accuracy with data augmentation. Interestingly, for Hand, data augmentation led to a substantial 10% accuracy improvement, whereas Mfeat1 showed a similar trend with a smaller improvement. PSO outperformed GA on all three datasets, but data augmentation allowed GA to excel on two (Hand, Mfeat1), selecting more features and suggesting its advantage in feature selection over PSO.

Dataset (#features)	Baseline	Without DA			30% DA		
	Acc	FS	Acc	#feat	FS	Acc	#feat
Hand (90)	58.8 ± 7.1	PSO	58.8 ± 6.8	33.3	GA	69.6 ± 5.7	38.45
Isolet (617)	90.7 ± 0.6	PSO	92.1 ± 0.9	291.8	PSO	92.9 ± 0.9	292.3
Mfeat1 (216)	95.7 ± 1.1	PSO	95.9 ± 0.9	90.9	GA	96.9 ± 0.9	104.6
Mfeat2 (64)	95.3 ± 1.1	SKB	96.2* ± 0.9	31	SKB	96.5* ± 0.8	32
Ozone (72)	94.1* ± 3.8	GA	94.2* ± 2.8	32.6	PSO	94.8* ± 2.1	30
Toxicity (1203)	67.4* ± 5.5	PSO	67.1* ± 7.5	509.9	RFE	70.1* ± 8.9	601

Table 5.12: Comparison between baseline experiment and best performance achieved with FS and FS combined with DA.

Chapter 6

Conclusions and Future Work

NDs are a group of disorders characterized by the progressive degeneration of the structure and function of the nervous system. Among them, the most common is AD, which predominantly impacts cognitive functions, leading to memory loss, impaired reasoning, and changes in behaviour. Its degeneration is due to the accumulation of abnormal protein aggregates, such as beta-amyloid plaques and tau tangles, in the brain. This condition lacks a cure, and an early diagnosis is crucial as it allows the initiation of timely interventions and treatments, offering individuals a better chance to manage symptoms, maintain quality of life, and potentially slow down the progression of the disease.

AD manifestations extend beyond cognitive decline and may affect motor skills, including handwriting. As the disease progresses, individuals often experience fine motor control and coordination difficulties. This can result in evident changes in their handwriting, including inconsistencies in letter size, spacing, slant, and a general decline in legibility. Understanding these subtle yet significant alterations in handwriting can serve to find potential diagnostic markers and monitor disease progression. Moreover, investigating the impact of neurodegenerative diseases on handwriting may contribute to developing innovative therapeutic interventions to preserve both cognitive and motor abilities in affected individuals. More information about the NDs and how their symptoms affect handwriting can be found in Chapter 2 of this work.

My thesis aims to offer a low-cost, non-invasive, and readily available tool for supporting AD diagnosis by incorporating AI techniques with handwriting analysis. The exploitation of AI techniques strives to leverage distinctive patterns in handwriting to identify early indicators of cognitive decline. The developed system aims to process handwriting samples efficiently, extracting subtle features associated with motor control and cognitive function. The focus on affordability and non-invasiveness ensures greater accessibility, especially for individuals encountering difficulties in performing traditional diagnostic methods. Ongoing research in this field highlights the potential for AI-driven handwriting analysis to transform the diagnostic landscape for neurodegenerative diseases, providing an accessible and scalable solution for widespread adoption and implementation. Section 1.1 describes the objectives of this work, while Section 2.4 comprises a detailed description of the research on NDs involving several AI approaches on different types of data.

During my research years, I worked with data described in Chapter 3 and acquired in 2018, comprising several handwriting samples from a group of 174 people, equally balanced into two classes: healthy controls and AD patients. Participants had to perform 25 handwriting tasks to evaluate different motor and cognitive abilities, usually impaired by the onset of the disease. Starting from the raw data acquired from each handwriting sample, I obtained and generated different data types, i.e. images and features. Chapter 4 shows many experimental settings devoted to understanding which combination of data type, kind of task and AI technique was the most suitable in discriminating people affected by AD from healthy controls. In 4.2, I compared classifiers based on handcrafted and deep features applied to Alzheimer's diagnosis from handwriting on graphic tasks. Deep features were derived by feeding different models of CNNs with binary and RGB on paper images. Every feature set was evaluated with different ML classifiers. This choice allowed me to quickly compare the experimental results relative to the different feature vector representations and, therefore, the role played by the shape and the combined use of both shape and dynamic information.

The outcomes of this experiment show exciting trends. First, deep features exhibit greater promise than handcrafted ones, showcasing superior accuracy, mainly when employed with the RF classifier. On average, the results obtained from handcrafted features perform worse than those from deep features. Across various tasks and classification schemes, CNN extracted features from RGB on paper images consistently

outperform handcrafted features, underscoring the advantage of deep features. Regarding the significance of shape information compared to dynamic information derived from handcrafted features, it becomes apparent that shape information holds relevance, particularly in subject classification. However, there is a noticeable decrease in average performance when considering binary images, which don't comprise dynamic information. This underscores the importance of combining dynamic features, represented by the three RGB channels, with shape features as the most effective approach to address the problem, according to this first experiment.

Section 4.3 shows an experimental setting similar to the previous work, but instead of the binary images, it considers MC images and six graphic tasks. As the previous experiment, also this one highlighted greater promise in using deep extracted features than handcrafted ones. In each task and classification scheme, deep features extracted from CNN models consistently outperformed handcrafted features, except in task 2, where handcrafted features marginally outperformed deep features. In comparing RGB and MC deep features, the analysis reveals that adding an extra channel in generating MC images does not necessarily enhance feature extraction. Generally, classification results obtained with RGB deep features are almost always superior to those obtained with MC deep features. The sole exception is task 5, where the FC classifier, trained with MC deep features from InceptionResNetV2, exhibited slightly better results. This comprehensive approach could enhance the diagnostic system's overall performance by aggregating the classifiers' responses across multiple tasks.

In Section 4.4, I evaluated the proposed system with writing tasks and a new data type: offline images. The study aimed to assess whether extracting features directly from original offline handwriting images, as opposed to synthetic RGB images, considering the authentic shape of the handwritten trace, could yield superior results compared to handcrafted features. The preliminary experimental findings are highly promising and validate the effectiveness of the proposed approach. A primary observation reveals that offline deep features generally outperform RGB deep features. This is particularly noteworthy as RGB images contain dynamic information, though with an approximation of the original shape. In contrast, offline images capture all the actual shape details of handwritten traits, proving crucial for distinguishing patients from healthy controls. This suggests the potential use of offline images in diagnosing AD and raises the possibility of utilizing past examples of a person's handwriting to detect the presence of neurodegenerative disease and estimate its progression.

Another significant finding is that offline deep features perform similarly to handcrafted features. While handcrafted features exhibit higher accuracy in many experiments, this discrepancy is more pronounced for Task 1, a signature involving a highly automated graphic gesture less influenced by the presence of AD. These results align well with other tasks, where shape changes play a more critical role than changes in dynamic features, enabling better detection of AD patients. Notably, the sensitivity is generally higher for offline deep features, a crucial aspect in medical applications where failing to identify a subject with a pathology carries a much higher cost than inaccurately classifying a healthy subject.

Another type of data used during my evaluation is lognormal features. In particular, I computed two different sets. The experiment on the first set is described in 4.5. This study examines the handwriting of individuals, utilizing lognormal features derived from the kinematic theory of rapid movement. I considered various tasks in this work and evaluated different aspects of AD symptoms. Preliminary findings indicate that lognormal features offer better modelling of writing tasks than graphic ones. It's essential to highlight a drawback of this experiment, as no parameter optimization was applied, but despite this, results are interesting, and they outperform results obtained by the deep approach from the convolutional models.

The study in Section 4.6 employs a ML-based classification system, leveraging ensemble techniques and combining rules, to discern between patients and healthy controls using the second set of features derived from the sigma-lognormal model applied to various handwriting tasks. The findings are interesting; the outcomes on lognormal features outperform those on handcrafted features. In particular, this study highlighted the good performance of ML classifiers when dealing with task 23. Contrary to what was expected, the stacking ensemble didn't improve the performance. The majority vote combining rule, instead, allowed an increment in the accuracy by aggregating predictions from the first three tasks in a ranked list based on their predictive ability. The study concludes that the extracted lognormal features prove valuable in exploring handwriting dynamics and fluency. However, the obtained results are not enough to address a medical problem, prompting an investigation into potential explanations:

- Inconsistencies were noted in the dataset, with instances where individuals from the control group took an unexpectedly long time to perform certain tasks or deviated from task requirements. It was observed that the velocity profile generated by the sigma-lognormal model failed to accentuate

differences between healthy controls and patients sufficiently.

- Handwriting and its associated features exhibit correlations with age and education. While distinctions between young HC and elderly PT are discernible, this contrast diminishes notably when comparing elderly HC and young PT.

Finally, I performed complementary experiments in Section 4.7 to complete an evaluation of the systems and approaches presented in the previous works. I improved the deep feature extraction and classification algorithms in this section by implementing grid search and feature selection techniques. In this context, I highlighted the potential of the offline approach with respect to RGB and binary. Moreover, I deployed a majority vote strategy considering subsets of tasks. I compared the performance of MC in air, on paper and in air-on-paper images. This outlined an improvement in the performance considering both air and on-paper traits. The same findings stand for RGB images, showing better performance with in-air on paper images. Finally, these experiments proved the importance of the combination rule used as if applied on a good subset of tasks, it can boost the system's overall performance.

Chapter 5 is devoted to describing a part of my research on evolutionary algorithms. These optimization techniques, inspired by the principles of biological evolution, were strategically applied across various domains such as optimization, machine learning, and neural architecture search. Specifically, I showed how the implementation of a GA could be useful in optimizing the prediction capabilities of a DL system for supporting the diagnosis of cognitive impairment associated with AD. Additionally, I considered an evolutionary approach to diagnosing cognitive impairment by adapting neural network architecture using evolutionary algorithms. The investigation extended to applying evolutionary algorithms for feature selection on a dataset of handwriting features, primarily focusing on enhancing diagnostic precision for cognitive impairment in the context of AD through ML. The collective results underscore the efficacy of evolutionary algorithms in these applications, though a critical need for further experimental phases is acknowledged to refine and optimize the obtained results.

The results achieved throughout my research are interesting, but there is room for improvement and further investigation. In future system development, I mean paying more attention to information related to in-air features, which haven't been sufficiently studied in this context. Since in-air points are captured from the tablet during the execution of writing tasks, assessing their impact on the feature extraction process is a potential avenue for further investigation. Analyzing in-air traits could provide an additional dimension to the study of AD, potentially offering valuable insights into early signs of cognitive decline. Changes in the fluidity, speed, and accuracy of in-air writing movements could indicate neurological changes associated with AD.

More improvements may come from the feature extraction and selection process definition, such as selecting a specific set of features for each task and adopting more powerful combining rules, which could substantially enhance overall classification performance.

Regarding the experiment on the second set of lognormal features, described in Section 4.6, It highlighted the need to improve the system in discriminating between older HC and younger PT. To achieve this, I need to investigate their striking similarities and determine the appropriate techniques or acquisition tasks to consider. A useful direction for future research involves integrating handwriting features with personal attributes to assess if it is feasible to measure or study how individuals cope with AD symptoms while writing and how they compensate. Additionally, future work will extend to applying the system to diverse disease datasets to validate and generalize the findings regarding results and identified relationships. Moving from the analysis of handwriting, Chapter 3 shows that the research community is investing in supporting the AD diagnosis through a large number of approaches relying on different data. In considering potential future advancements, it would be valuable to explore the complementary utilization of multiple data sources to reinforce and validate diagnostic predictions in the context of an AI-driven system for supporting the diagnosis of Alzheimer's disease. A potentially practical approach could involve the integration of new data sources but always guaranteeing some peculiar aspects of the system, which has to be non-invasive and cost-effective. Specifically, the combined analysis of movement, speech, and handwriting could emerge as a versatile and accessible method. Combining these signals might provide a broader and deeper insight into the condition, thereby helping to overcome current challenges in ND diagnosis.

In conclusion, as presented in this thesis, the results obtained in my research have successfully met the objectives I meant to achieve. The conducted experiments have provided insights into identifying tasks and data types, most effectively highlighting differences in handwriting between healthy subjects and those affected by AD. Moreover, through the various experiments, I have recognized the potential of con-

volutional networks in extracting crucial features for this study. In particular, I noticed an enhancement of results achieved on images containing both in-air and on-paper traits and improved performance by applying combination rules to specific tasks rather than all acquired tasks. This is significant as it helps discern which tasks can be excluded from the study and future acquisitions, such as signatures, and which require particular attention, like the dictated telephone number task. Understanding these distinctions aids in optimizing the study design. Despite the interesting outcomes, I have identified several drawbacks that require resolution and further investigation. The work can thus be refined, and additional analyses can be carried out to achieve better performance. This ongoing process of improvement and exploration is crucial for advancing our understanding and refining the methodologies employed in handwriting analysis, contributing to the broader field of research on NDs.

Bibliography

- [1] Agostino Accardo et al. “A device for quantitative kinematic analysis of children’s handwriting movements”. In: *11th Mediterranean Conference on Medical and Biomedical Engineering and Computing 2007*. Ed. by Tomaz Jarm, Peter Kramar, and Anze Zupanic. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 445–448. ISBN: 978-3-540-73044-6.
- [2] M. T. Angelillo et al. “Attentional Pattern Classification for Automatic Dementia Detection”. In: *IEEE Access* 7 (2019), pp. 57706–57716.
- [3] Richard Armstrong. “What causes neurodegenerative disease?” In: *Folia Neuropathologica* 58.2 (2020), pp. 93–112.
- [4] Francesco Ascì et al. “Handwriting Declines With Human Aging: A Machine Learning Study”. In: *Frontiers in Aging Neuroscience* 14 (2022), p. 889930. DOI: 10.3389/fnagi.2022.889930.
- [5] Lerina Aversano et al. “Early Detection of Parkinson’s Disease using Spiral Test and Echo State Networks”. In: *2022 International Joint Conference on Neural Networks (IJCNN)*. 2022, pp. 1–8. DOI: 10.1109/IJCNN55064.2022.9891917.
- [6] “Benefits and Harms of Prescription Drugs and Supplements for Treatment of Clinical Alzheimer-Type Dementia”. In: *Annals of Internal Medicine* 172.10 (2020). PMID: 32340037, pp. 656–668. DOI: 10.7326/M19-3887. eprint: <https://doi.org/10.7326/M19-3887>. URL: <https://doi.org/10.7326/M19-3887>.
- [7] V. Bevilacqua et al. “A Model-Free Computer-Assisted Handwriting Analysis Exploiting Optimal Topology ANNs on Biometric Signals in Parkinson’s Disease Research”. In: *Lecture Notes in Computer Science*. Vol. 10955. 2018, pp. 650–655.
- [8] Xiaojun Bi and Haibo Wang. “Early Alzheimer’s disease diagnosis based on EEG spectral images using deep learning”. In: *Neural Networks*. Vol. 114. Elsevier, 2019, pp. 119–135.
- [9] Luboš Brabenec et al. “Speech disorders in Parkinson’s disease: early diagnostics and effects of medication and brain stimulation”. In: *Journal of neural transmission* 124 (2017), pp. 303–334.
- [10] Zeinab Breijyeh and Rafik Karaman. “Comprehensive Review on Alzheimer’s Disease: Causes and Treatment”. In: *Molecules* 25.24 (2020). ISSN: 1420-3049. DOI: 10.3390/molecules25245789. URL: <https://www.mdpi.com/1420-3049/25/24/5789>.
- [11] Leo Breiman. “Random Forests”. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565.
- [12] Peng Cao et al. “l2,1-l1 regularized nonlinear multi-task representation learning based cognitive performance prediction of Alzheimer’s disease”. In: *Pattern Recognition* 6437 (2018).
- [13] Zhe Cao et al. “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields”. In: *IEEE Trans Pattern Anal Mach Intell* 43.1 (Dec. 2020), pp. 172–186.
- [14] Daniela Carfora et al. “On Extracting Digitized Spiral Dynamics’ Representations: A Study on Transfer Learning for Early Alzheimer’s Detection”. In: *Bioengineering* 9.8 (2022). ISSN: 2306-5354. DOI: 10.3390/bioengineering9080375. URL: <https://www.mdpi.com/2306-5354/9/8/375>.
- [15] Cristina Carmona-Duarte et al. “Sigma-lognormal modeling of speech”. In: *Cognitive computation* 13 (2021), pp. 488–503.
- [16] Cristina Carmona-Duarte et al. “Temporal evolution in synthetic handwriting”. In: *Pattern Recognition* 68 (2017), pp. 233–244. ISSN: 0031-3203. DOI: <https://doi.org/10.1016/j.patcog.2017.03.019>. URL: <https://www.sciencedirect.com/science/article/pii/S0031320317301280>.
- [17] Chih-Chung Chang and Chih-Jen Lin. “LIBSVM: A Library for Support Vector Machines”. In: *ACM Transactions on Intelligent Systems and Technology* 2 (3 2011), 27:1–27:27.
- [18] Lin Chen, Hezhe Qiao, and Fan Zhu. “Alzheimer’s Disease Diagnosis With Brain Structural MRI Using Multiview-Slice Attention and 3D Convolution Neural Network”. In: *Frontiers in Aging Neuroscience* 14 (2022), p. 871706.

- [19] N. D. Cilia et al. “A ranking-based feature selection approach for handwritten character recognition”. In: *Pattern Recognition Letters* (2018).
- [20] Nicole D Cilia et al. “Deep transfer learning algorithms applied to synthetic drawing images as a tool for supporting Alzheimer’s disease prediction”. In: *Machine Vision and Applications* 33.3 (2022), p. 49.
- [21] Nicole D. Cilia et al. “Diagnosing Alzheimer’s disease from on-line handwriting: A novel dataset and performance benchmarking”. In: *Engineering Applications of Artificial Intelligence* 111 (2022), p. 104822. ISSN: 0952-1976. DOI: <https://doi.org/10.1016/j.engappai.2022.104822>. URL: <https://www.sciencedirect.com/science/article/pii/S0952197622000902>.
- [22] Nicole D. Cilia et al. “From Online Handwriting to Synthetic Images for Alzheimer’s Disease Detection Using a Deep Transfer Learning Approach”. In: *IEEE Journal of Biomedical and Health Informatics* 25.12 (2021), pp. 4243–4254. DOI: 10.1109/JBHI.2021.3101982.
- [23] Nicole Dalia Cilia et al. “An Experimental Protocol to Support Cognitive Impairment Diagnosis by using Handwriting Analysis”. In: *Procedia Computer Science* 141 (2018). The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2018) / The 8th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2018) / Affiliated Workshops, pp. 466–471. ISSN: 1877-0509. DOI: <https://doi.org/10.1016/j.procs.2018.10.141>. URL: <https://www.sciencedirect.com/science/article/pii/S1877050918317903>.
- [24] Nicole Dalia Cilia et al. “Lognormal Features for Early Diagnosis of Alzheimer’s Disease Through Handwriting Analysis”. In: *International Graphonomics Conference*. Springer. 2022, pp. 322–335.
- [25] Nicole Dalia Cilia et al. “Offline handwriting image analysis to predict Alzheimer’s disease via deep learning”. In: *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE. 2022, pp. 2807–2813.
- [26] Nicole Dalia Cilia et al. “Using Genetic Algorithms to Optimize a Deep Learning Based System for the Prediction of Cognitive Impairments”. In: *Italian Workshop on Artificial Life and Evolutionary Computation*. Springer. 2021, pp. 139–150.
- [27] Nicole Dalia Cilia et al. “Variable-length representation for EC-based feature selection in high-dimensional data”. In: *Applications of Evolutionary Computation: 22nd International Conference, EvoApplications 2019, Held as Part of EvoStar 2019, Leipzig, Germany, April 24–26, 2019, Proceedings 22*. Springer. 2019, pp. 325–340.
- [28] Ron Cole and Mark Fanty. *ISOLET*. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C51G69.1994>.
- [29] Tiziana D’Alessandro et al. “A Machine Learning Approach to Analyze the Effects of Alzheimer’s Disease on Handwriting Through Lognormal Features”. In: *International Graphonomics Conference*. Springer. 2023, pp. 103–121.
- [30] Erika D’Antonio et al. “A markerless system for gait analysis based on OpenPose library”. In: *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. 2020, pp. 1–6. DOI: 10.1109/I2MTC43012.2020.9128918.
- [31] Quang Dao, Mounim A. El-Yacoubi, and Anne-Sophie Rigaud. “Detection of Alzheimer Disease on Online Handwriting Using 1D Convolutional Neural Network”. In: *IEEE Access* 11 (2023), pp. 2148–2155. DOI: 10.1109/ACCESS.2022.3232396.
- [32] Giuseppe De Gregorio et al. “A Multi Classifier Approach for Supporting Alzheimer’s Diagnosis Based on Handwriting Analysis”. In: *Pattern Recognition. ICPR International Workshops and Challenges*. Ed. by Alberto Del Bimbo et al. Cham: Springer International Publishing, 2021, pp. 559–574. ISBN: 978-3-030-68763-2.
- [33] C. De Stefano, F. Fontanella, and C. Marrocco. “A GA-Based Feature Selection Algorithm for Remote Sensing Images”. In: *Applications of Evolutionary Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 285–294.
- [34] C. De Stefano et al. “A Bayesian Approach for Combining Ensembles of GP Classifiers”. In: *Lecture Notes in Computer Science. Multiple Classifier Systems. MCS 2011* 6713 (2011), pp. 26–35.
- [35] Claudio De Stefano et al. “Handwriting analysis to support neurodegenerative diseases diagnosis: A review”. In: *Pattern Recognition Letters* 121 (2019). Graphonomics for e-citizens: e-health, e-society, e-education, pp. 37–45. ISSN: 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2018.05.013>. URL: <https://www.sciencedirect.com/science/article/pii/S0167865518301880>.

- [36] J. Deng et al. “ImageNet: A large-scale hierarchical image database.” In: *CVPR*. IEEE Computer Society, 2009, pp. 248–255.
- [37] Günther Deuschl et al. “The burden of neurological diseases in Europe: an analysis for the Global Burden of Disease Study 2017”. In: *The Lancet Public Health* 5.10 (2020), e551–e567. ISSN: 2468-2667. DOI: [https://doi.org/10.1016/S2468-2667\(20\)30190-0](https://doi.org/10.1016/S2468-2667(20)30190-0). URL: <https://www.sciencedirect.com/science/article/pii/S2468266720301900>.
- [38] Moises Diaz et al. “Dynamically enhanced static handwriting representation for Parkinson’s disease detection”. In: *Pattern Recognition Letters* 128 (2019), pp. 204–210. ISSN: 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2019.08.018>. URL: <https://www.sciencedirect.com/science/article/pii/S0167865518307013>.
- [39] Moises Diaz et al. “Sequence-based dynamic handwriting analysis for Parkinson’s disease detection with one-dimensional convolutions and BiGRUs”. In: *Expert Systems with Applications* 168 (2021), p. 114405.
- [40] Moisés Díaz et al. “Graphomotor Evolution in the Handwriting of Bengali Children Through Sigma-Lognormal Based-Parameters: A Preliminary Study”. In: 2019.
- [41] M. Djioua and R. Plamondon. “Studying the variability of handwriting patterns using the Kinematic Theory”. In: *Human Movement Science* 28.5 (2009). Disruptions of Handwriting, pp. 588–601. ISSN: 0167-9457. DOI: <https://doi.org/10.1016/j.humov.2009.01.005>. URL: <https://www.sciencedirect.com/science/article/pii/S0167945709000190>.
- [42] Moussa Djioua and Rejean Plamondon. “A New Algorithm and System for the Characterization of Handwriting Strokes with Delta-Lognormal Parameters”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.11 (2009), pp. 2060–2072.
- [43] Peter Drotár et al. “A new modality for quantitative evaluation of Parkinson’s disease: In-air movement”. In: *13th IEEE International Conference on BioInformatics and BioEngineering*. 2013, pp. 1–4. DOI: 10.1109/BIBE.2013.6701692.
- [44] Peter Drotár et al. “Analysis of in-air movement in handwriting: A novel marker for Parkinson’s disease”. In: *Computer Methods and Programs in Biomedicine* 117.3 (2014), pp. 405–411. ISSN: 0169-2607. DOI: <https://doi.org/10.1016/j.cmpb.2014.08.007>. URL: <https://www.sciencedirect.com/science/article/pii/S0169260714003204>.
- [45] Peter Drotár et al. “Contribution of different handwriting modalities to differential diagnosis of Parkinson’s Disease”. In: *2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA) Proceedings*. 2015, pp. 344–348. DOI: 10.1109/MeMeA.2015.7145225.
- [46] Peter Drotár et al. “Decision Support Framework for Parkinson’s Disease Based on Novel Handwriting Markers”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 23.3 (2015), pp. 508–516. DOI: 10.1109/TNSRE.2014.2359997.
- [47] Peter Drotár et al. “Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson’s disease”. In: *Artificial Intelligence in Medicine* 67 (2016), pp. 39–46. ISSN: 0933-3657. DOI: <https://doi.org/10.1016/j.artmed.2016.01.004>. URL: <https://www.sciencedirect.com/science/article/pii/S0933365716000063>.
- [48] Peter Drotár et al. “Prediction potential of different handwriting tasks for diagnosis of Parkinson’s”. In: *2013 E-Health and Bioengineering Conference (EHB)*. 2013, pp. 1–4. DOI: 10.1109/EHB.2013.6707378.
- [49] Robert Duin. *Multiple Features*. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5HC70>.
- [50] Pakize Erdogmus and Abdullah Talha Kabakus. “The promise of convolutional neural networks for the early diagnosis of the Alzheimer’s disease”. In: *Engineering Applications of Artificial Intelligence* 123 (2023), p. 106254. ISSN: 0952-1976. DOI: <https://doi.org/10.1016/j.engappai.2023.106254>. URL: <https://www.sciencedirect.com/science/article/pii/S0952197623004384>.
- [51] Valery L Feigin et al. “The global burden of neurological disorders: translating evidence into policy”. In: *The Lancet Neurology* 19.3 (2020), pp. 255–265. ISSN: 1474-4422. DOI: [https://doi.org/10.1016/S1474-4422\(19\)30411-9](https://doi.org/10.1016/S1474-4422(19)30411-9). URL: <https://www.sciencedirect.com/science/article/pii/S1474442219304119>.
- [52] Miguel A. Ferrer et al. “IDeLog: Iterative Dual Spatial and Kinematic Extraction of Sigma-Lognormal Parameters”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.1 (2020), pp. 114–125.
- [53] G. Fisco et al. “Combining EEG signal processing with supervised methods for Alzheimer’s patients classification”. In: *BMC Medical Informatics and Decision Making Neural Networks*. Vol. 18:35. 2018.

- [54] M. F. Folstein, S. E. Folstein, and P. R. McHugh. “‘Mini-mental state’: A practical method for grading the cognitive state of patients for the clinician”. In: *J. Psychiatric Res.* 12.3 (1975), pp. 189–198.
- [55] Zoltan Galaz et al. “Comparison of CNN-Learned vs. Handcrafted Features for Detection of Parkinson’s Disease Dysgraphia in a Multilingual Dataset”. In: *Frontiers in Neuroinformatics* 16 (2022). ISSN: 1662-5196. DOI: 10.3389/fninf.2022.877139. URL: <https://www.frontiersin.org/articles/10.3389/fninf.2022.877139>.
- [56] J Garre-Olmo et al. “Kinematic and Pressure Features of Handwriting and Drawing: Preliminary Results Between Patients with Mild Cognitive Impairment, Alzheimer Disease and Healthy Controls”. In: *Curr Alzheimer Res* 14 (2017), pp. 1–9.
- [57] Matej Gazda, Máté Hireš, and Peter Drotár. “Multiple-Fine-Tuned Convolutional Neural Networks for Parkinson’s Disease Diagnosis From Offline Handwriting”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 52.1 (2022), pp. 78–89. DOI: 10.1109/TSMC.2020.3048892.
- [58] Peyvand Ghaderyan, Ataollah Abbasi, and Sajad Saber. “A new algorithm for kinematic analysis of Handwriting data; towards a reliable handwriting-based tool for early detection of Alzheimer’s disease”. In: *Expert Systems With Applications* 12106 (2018).
- [59] David E. Goldberg. *Genetic Algorithms in Search Optimization and Machine Learning*. Addison-Wesley, 1989.
- [60] Luis C. Guayacán and Fabio Martínez. “Visualising and quantifying relevant parkinsonian gait patterns using 3D convolutional network”. In: *Journal of Biomedical Informatics* 123 (2021), p. 103935. ISSN: 1532-0464. DOI: <https://doi.org/10.1016/j.jbi.2021.103935>. URL: <https://www.sciencedirect.com/science/article/pii/S1532046421002641>.
- [61] Seref Gul et al. “Structure-based design and classifications of small molecules regulating the circadian rhythm period”. In: *Scientific Reports* 11 (2021). URL: <https://api.semanticscholar.org/CorpusID:237546851>.
- [62] Yao Guo et al. “Detection and assessment of Parkinson’s disease based on gait analysis: A survey”. In: *Frontiers in Aging Neuroscience* 14 (2022). ISSN: 1663-4365. DOI: 10.3389/fnagi.2022.916971. URL: <https://www.frontiersin.org/articles/10.3389/fnagi.2022.916971>.
- [63] Ujjwal Gupta, Hritik Bansal, and Deepak Joshi. “An improved sex-specific and age-dependent classification model for Parkinson’s diagnosis using handwriting measurement”. In: *Computer Methods and Programs in Biomedicine* 189 (2020), p. 105305. ISSN: 0169-2607. DOI: <https://doi.org/10.1016/j.cmpb.2019.105305>. URL: <https://www.sciencedirect.com/science/article/pii/S0169260719315159>.
- [64] Youssef El-Hayek et al. “Tip of the Iceberg: Assessing the Global Socioeconomic Costs of Alzheimer’s Disease and Related Dementias and Strategic Implications for Stakeholders”. In: *Journal of Alzheimer’s Disease* 70 (June 2019), pp. 1–19. DOI: 10.3233/JAD-190426.
- [65] K. He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.
- [66] Patricia Henríquez et al. “Characterization of healthy and pathological voice through measures based on nonlinear dynamics”. In: *IEEE transactions on audio, speech, and language processing* 17.6 (2009), pp. 1186–1195.
- [67] Máté Hireš et al. “Convolutional neural network ensemble for Parkinson’s disease detection from voice recordings”. In: *Computers in Biology and Medicine* 141 (2022), p. 105021. ISSN: 0010-4825. DOI: <https://doi.org/10.1016/j.combiomed.2021.105021>. URL: <https://www.sciencedirect.com/science/article/pii/S0010482521008155>.
- [68] Zhentao Hu et al. “VGG-TSwinformer: Transformer-based deep learning model for early Alzheimer’s disease prediction”. In: *Computer Methods and Programs in Biomedicine* 229 (2023), p. 107291.
- [69] Meiyang Huang et al. “Deep-gated recurrent unit and diet network-based genome-wide association analysis for detecting the biomarkers of Alzheimer’s disease”. In: *Medical Image Analysis* 73 (2021), p. 102189.
- [70] A. Iavarone et al. “The frontal assessment battery (FAB): Normative data from an Italian sample and performances of patients with Alzheimer’s disease and frontotemporal dementia”. In: *Funct Neurol.* 19 (July 2004), pp. 191–195.
- [71] D. Impedovo et al. “Writing Generation Model for Health Care Neuromuscular System Investigation”. In: *Proceedings of CIBB 2013*. Springer, 2014, pp. 137–148.

- [72] Jyoti Islam and Yanqing Zhang. “Brain MRI analysis for Alzheimer’s disease diagnosis using an ensemble system of deep convolutional neural networks”. In: *Brain informatics* 5 (2018), pp. 1–14.
- [73] D. D. Kairamkonda et al. “Analysis of Interpretable Handwriting Features to Evaluate Motoric Patterns in Different Neurodegenerative Diseases”. In: *2022 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. 2022, pp. 1–6. DOI: 10.1109/SPMB55497.2022.10014966.
- [74] Nagaendran Kandiah et al. “Treatment of dementia and mild cognitive impairment with or without cerebrovascular disease: Expert consensus on the use of Ginkgo biloba extract, EGb 761®”. In: *CNS Neuroscience & Therapeutics* 25.2 (2019), pp. 288–298. DOI: <https://doi.org/10.1111/cns.13095>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cns.13095>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cns.13095>.
- [75] Hajer Khachnaoui, Rostom Mabrouk, and Nawres Khlifa. “Machine learning and deep learning for clinical data and PET/SPECT imaging in Parkinson’s disease: a review”. In: *IET Image Processing* 14.16 (2020), pp. 4013–4026.
- [76] Niklas König et al. “Can Gait Signatures Provide Quantitative Measures for Aiding Clinical Decision-Making? A Systematic Meta-Analysis of Gait Variability Behavior in Patients with Parkinson’s Disease”. In: *Frontiers in Human Neuroscience* 10 (2016). ISSN: 1662-5161. DOI: 10.3389/fnhum.2016.00319. URL: <https://www.frontiersin.org/articles/10.3389/fnhum.2016.00319>.
- [77] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. 2009. URL: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- [78] Y. Lecun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
- [79] Baiying Lei et al. “Deep and Joint Learning of Longitudinal Data for Alzheimer’s Disease Prediction”. In: *Pattern Recognition* 107247 (2020).
- [80] Lanlan Li et al. “Use of deep-learning genomics to discriminate healthy individuals from those with Alzheimer’s disease or mild cognitive impairment”. In: *Behavioural Neurology* 2021 (2021).
- [81] Zhu Li et al. “Early diagnosis of Parkinson’s disease using Continuous Convolution Network: Handwriting recognition based on off-line hand drawing without template”. In: *Journal of Biomedical Informatics* 130 (2022). ISSN: 1532-0464. DOI: <https://doi.org/10.1016/j.jbi.2022.104085>.
- [82] Laurence Likforman-Sulem et al. “EMOTHAW: A Novel Database for Emotional State Recognition From Handwriting and Drawing”. In: *IEEE Transactions on Human-Machine Systems* 47.2 (Apr. 2017), pp. 273–284. ISSN: 2168-2305. DOI: 10.1109/thms.2016.2635441. URL: <http://dx.doi.org/10.1109/THMS.2016.2635441>.
- [83] Yuqiao Liu et al. “A Survey on Evolutionary Neural Architecture Search”. In: *IEEE Transactions on Neural Networks and Learning Systems* 34.2 (2023), pp. 550–570.
- [84] Claudio Loconsole et al. “A model-free technique based on computer vision and sEMG for classification in Parkinson’s disease by using computer-assisted handwriting analysis”. In: *Pattern Recognition Letters* 121 (2019), pp. 28–36.
- [85] Justin M. Long and David M. Holtzman. “Alzheimer Disease: An Update on Pathobiology and Treatment Strategies”. In: *Cell* 179.2 (2019), pp. 312–339. ISSN: 0092-8674. DOI: <https://doi.org/10.1016/j.cell.2019.09.001>. URL: <https://www.sciencedirect.com/science/article/pii/S0092867419310074>.
- [86] Karmele Lopez-de-Ipina et al. “Advances on automatic speech analysis for early detection of Alzheimer disease: a non-linear multi-task approach”. In: *Current Alzheimer Research* 15.2 (2018), pp. 139–148.
- [87] Angelo Marcelli, Antonio Parziale, and Rosa Senatore. “Some Observations on Handwriting from a Motor Learning Perspective”. In: vol. 1022. Aug. 2013.
- [88] Delazer Margarete, Laura Zamarian, and Atbin Djamshidian. “Handwriting in Alzheimer’s Disease”. In: *Journal of Alzheimer’s Disease* 82 (May 2021), pp. 1–9. DOI: 10.3233/JAD-210279.
- [89] Jiao Meng et al. “Image-based Handwriting Analysis for Disease Diagnosis”. In: *2022 41st Chinese Control Conference (CCC)*. 2022, pp. 4058–4062. DOI: 10.23919/CCC55666.2022.9902136.
- [90] Momina Moetesum et al. “A Survey of Visual and Procedural Handwriting Analysis for Neuropsychological Assessment”. In: *Neural Comput. Appl.* 34.12 (June 2022), pp. 9561–9578. ISSN: 0941-0643. DOI: 10.1007/s00521-022-07185-6. URL: <https://doi.org/10.1007/s00521-022-07185-6>.

- [91] Steven T Moore et al. “Long-term monitoring of gait in Parkinson’s disease”. In: *Gait & posture* 26.2 (2007), pp. 200–207.
- [92] Jan Mucha et al. “Identification and Monitoring of Parkinson’s Disease Dysgraphia Based on Fractional-Order Derivatives of Online Handwriting”. In: *Applied Sciences* 8.12 (2018). ISSN: 2076-3417. DOI: 10.3390/app8122566. URL: <https://www.mdpi.com/2076-3417/8/12/2566>.
- [93] Nickson Mwamsojo et al. “Reservoir Computing for Early Stage Alzheimer’s Disease Detection”. In: *IEEE Access* 10 (2022), pp. 59821–59831. DOI: 10.1109/ACCESS.2022.3180045.
- [94] Ron Nachum et al. “A Novel Computer Vision Approach to Kinematic Analysis of Handwriting with Implications for Assessing Neurodegenerative Diseases”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021. DOI: 10.1109/EMBC46164.2021.9630492.
- [95] Z. S. Nasreddine et al. “The Montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment”. In: 53.4 (Apr. 2005), pp. 695–699.
- [96] Mícheál Ó Breasail et al. “Wearable GPS and accelerometer technologies for monitoring mobility and physical activity in neurodegenerative disorders: A systematic review”. In: *Sensors* 21.24 (2021), p. 8261.
- [97] Christian O’Reilly and Réjean Plamondon. “Development of a Sigma–Lognormal representation for on-line signatures”. In: *Pattern Recognition* 42.12 (2009). New Frontiers in Handwriting Recognition, pp. 3324–3337. ISSN: 0031-3203. DOI: <https://doi.org/10.1016/j.patcog.2008.10.017>. URL: <https://www.sciencedirect.com/science/article/pii/S0031320308004470>.
- [98] Christian O’Reilly and Réjean Plamondon. “Development of a Sigma–Lognormal representation for on-line signatures”. In: *Pattern Recognition* 42.12 (2009). New Frontiers in Handwriting Recognition, pp. 3324–3337. DOI: <https://doi.org/10.1016/j.patcog.2008.10.017>.
- [99] Juan R Orozco et al. “Voice pathology detection in continuous speech using nonlinear dynamics”. In: *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*. IEEE, 2012, pp. 1030–1033.
- [100] Hyung-Soon Park et al. “Development of a VR-based treadmill control interface for gait assessment of patients with Parkinson’s disease”. In: *2011 IEEE International Conference on Rehabilitation Robotics*. 2011, pp. 1–5. DOI: 10.1109/ICORR.2011.5975463.
- [101] A. Parziale et al. “Cartesian genetic programming for diagnosis of Parkinson disease through handwriting analysis: Performance vs. interpretability issues”. In: *Artificial Intelligence in Medicine* 111 (2021), p. 101984. ISSN: 0933-3657. DOI: <https://doi.org/10.1016/j.artmed.2020.101984>. URL: <https://www.sciencedirect.com/science/article/pii/S0933365720312495>.
- [102] Antonio Parziale, Antonio Della Cioppa, and Angelo Marcelli. “Investigating One-Class Classifiers to Diagnose Alzheimer’s Disease from Handwriting”. In: *Image Analysis and Processing – ICIAP 2022*. Ed. by Stan Sclaroff et al. Cham: Springer International Publishing, 2022, pp. 111–123. ISBN: 978-3-031-06427-2.
- [103] Antonio Parziale et al. “A Decision Tree for Automatic Diagnosis of Parkinson’s Disease from Offline Drawing Samples: Experiments and Findings”. In: *Image Analysis and Processing – ICIAP 2019*. Ed. by Elisa Ricci et al. Cham: Springer International Publishing, 2019, pp. 196–206. ISBN: 978-3-030-30642-7.
- [104] Jonas J. de Paula et al. “Impairment of fine motor dexterity in mild cognitive impairment and Alzheimer’s disease dementia: association with activities of daily living”. In: *Brazilian Journal of Psychiatry* 38.3 (July 2016), pp. 235–238. ISSN: 1516-4446. DOI: 10.1590/1516-4446-2015-1874. URL: <https://doi.org/10.1590/1516-4446-2015-1874>.
- [105] Clayton R. Pereira et al. “Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson’s disease identification”. In: *Artificial Intelligence in Medicine* 87 (2018), pp. 67–77. ISSN: 0933-3657. DOI: <https://doi.org/10.1016/j.artmed.2018.04.001>. URL: <https://www.sciencedirect.com/science/article/pii/S093336571730369X>.
- [106] Giuseppe Pirlo et al. “Early Diagnosis of Neurodegenerative Diseases by Handwritten Signature Analysis”. In: *ICIAP Workshops*. 2015, pp. 290–297.
- [107] Réjean Plamondon et al. “The lognormal handwriter: learning, performing, and declining”. In: *Frontiers in Psychology* 4 (2013). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2013.00945. URL: <https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00945>.
- [108] Réjean Plamondon et al. “The lognormal handwriter: learning, performing, and declining”. In: *Frontiers in Psychology* 4 (2013). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2013.00945. URL: <https://www.frontiersin.org/article/10.3389/fpsyg.2013.00945>.

- [109] “Practice guideline update summary: Mild cognitive impairment report of the guideline development, dissemination, and implementation”. English. In: *Neurology* 90.3 (Jan. 2018), pp. 126–135. ISSN: 0028-3878. DOI: 10.1212/WNL.0000000000004826.
- [110] Joan Prats-Climent et al. “Artificial Intelligence on FDG PET Images Identifies Mild Cognitive Impairment Patients with Neurodegenerative Disease”. In: *Journal of Medical Systems* 46.8 (2022), p. 52.
- [111] María Luisa Barragán Pulido et al. “Alzheimer’s disease and automatic speech analysis: A review”. In: *Expert Systems with Applications* 150 (2020), p. 113213. ISSN: 0957-4174. DOI: <https://doi.org/10.1016/j.eswa.2020.113213>. URL: <https://www.sciencedirect.com/science/article/pii/S0957417420300397>.
- [112] Nihar M. Ranjan, Gitanjali Mate, and Maya Bembde. “Detection of Parkinson’s Disease using Machine Learning Algorithms and Handwriting Analysis”. In: *Journal of Data Mining and Management* 8.1 (2023), pp. 21–29.
- [113] Pengzhen Ren et al. “A comprehensive survey of neural architecture search: Challenges and solutions”. In: *ACM Computing Surveys (CSUR)* 54.4 (2021), pp. 1–34.
- [114] Sara Rosenblum, Batya Engel-Yeger, and Yael Fogel. “Age-related changes in executive control and their relationships with activity performance in handwriting”. In: *Human Movement Science* 32.2 (2013), pp. 363–376. ISSN: 0167-9457. DOI: <https://doi.org/10.1016/j.humov.2012.12.008>. URL: <https://www.sciencedirect.com/science/article/pii/S0167945712001649>.
- [115] Sara Rosenblum et al. “Handwriting as an objective tool for Parkinson’s disease diagnosis”. In: *Journal of neurology* 260 (2013), pp. 2357–2361.
- [116] Andrea Sabo et al. “Assessment of Parkinsonian gait in older adults with dementia via human pose tracking in video data”. In: *Journal of NeuroEngineering and Rehabilitation* 17.1 (July 2020), p. 97. ISSN: 1743-0003. DOI: 10.1186/s12984-020-00728-9. URL: <https://doi.org/10.1186/s12984-020-00728-9>.
- [117] Andrea Sabo et al. “Estimating Parkinsonism Severity in Natural Gait Videos of Older Adults With Dementia”. In: *IEEE Journal of Biomedical and Health Informatics* 26.5 (2022), pp. 2288–2298. DOI: 10.1109/JBHI.2022.3144917.
- [118] Cristina L Saratzaga et al. “MRI deep learning-based solution for Alzheimer’s disease prediction”. In: *Journal of personalized medicine* 11.9 (2021), p. 902.
- [119] Konstantin Sarin et al. “A three-stage fuzzy classifier method for Parkinson’s disease diagnosis using dynamic handwriting analysis”. In: *Decision Analytics Journal* 8 (2023). DOI: <https://doi.org/10.1016/j.dajour.2023.100274>.
- [120] Rosa Senatore, Antonio Della Cioppa, and Angelo Marcelli. “Automatic Diagnosis of Parkinson Disease through Handwriting Analysis: A Cartesian Genetic Programming Approach”. In: *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. 2019, pp. 312–317. DOI: 10.1109/CBMS.2019.00071.
- [121] Rosa Senatore et al. “Distinctive Handwriting Signs in Early Parkinson’s Disease”. In: *Applied Sciences* 12.23 (2022). ISSN: 2076-3417. DOI: 10.3390/app122312338. URL: <https://www.mdpi.com/2076-3417/12/23/12338>.
- [122] Jamie Shotton et al. “Real-time human pose recognition in parts from single depth images”. In: *CVPR 2011*. 2011, pp. 1297–1304. DOI: 10.1109/CVPR.2011.5995316.
- [123] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: <http://arxiv.org/abs/1409.1556>.
- [124] Anju Singh et al. “Oxidative Stress: A Key Modulator in Neurodegenerative Diseases”. In: *Molecules* 24.8 (2019). ISSN: 1420-3049. DOI: 10.3390/molecules24081583. URL: <https://www.mdpi.com/1420-3049/24/8/1583>.
- [125] Alastair Smith. “On the Use of Drawing Tasks in Neuropsychological Assessment”. In: *Neuropsychology* 23 (Mar. 2009), pp. 231–9. DOI: 10.1037/a0014184.
- [126] Aimee Spector et al. “Efficacy of an evidence-based cognitive stimulation therapy programme for people with dementia: Randomised controlled trial”. In: *The British Journal of Psychiatry* 183.3 (2003), pp. 248–254. DOI: 10.1192/bjp.183.3.248.
- [127] Stephen Stahl. “The New Cholinesterase Inhibitors for Alzheimer’s Disease, Part 2: Illustrating Their Mechanisms of Action”. In: *The Journal of clinical psychiatry* 61 (Dec. 2000), pp. 813–4. DOI: 10.4088/JCP.v61n1101.

- [128] Jodie Stephenson et al. “Inflammation in CNS neurodegenerative diseases”. In: *Immunology* 154.2 (2018), pp. 204–219. DOI: <https://doi.org/10.1111/imm.12922>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/imm.12922>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/imm.12922>.
- [129] Martin Strassnig and Mary Ganguli. “About a peculiar disease of the cerebral cortex: Alzheimer’s original case revisited”. In: *Psychiatry (Edgmont (Pa. : Township))* 2 (Sept. 2005), pp. 30–3.
- [130] C. Szegedy, S. Ioffe, and V. Vanhoucke. “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning”. In: *AAAI*. 2016.
- [131] C. Szegedy et al. “Rethinking the Inception Architecture for Computer Vision”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 2818–2826.
- [132] Christian Tackenberg, Luka Kulic, and Roger M. Nitsch. “Familial Alzheimer’s disease mutations at position 22 of the amyloid β -peptide sequence differentially affect synaptic loss, tau phosphorylation and neuronal cell death in an ex vivo system”. In: *PLOS ONE* 15.9 (Sept. 2020), pp. 1–14. DOI: [10.1371/journal.pone.0239584](https://doi.org/10.1371/journal.pone.0239584). URL: <https://doi.org/10.1371/journal.pone.0239584>.
- [133] Catherine Taleb et al. “Detection of Parkinson’s disease from handwriting using deep learning: a comparative study”. In: *Evolutionary Intelligence* 16 (Sept. 2020). DOI: [10.1007/s12065-020-00470-0](https://doi.org/10.1007/s12065-020-00470-0).
- [134] Aleksandr Talitckii et al. “Comparative Study of Wearable Sensors, Video, and Handwriting to Detect Parkinson’s Disease”. In: *IEEE Transactions on Instrumentation and Measurement* 71 (2022), pp. 1–10. DOI: [10.1109/TIM.2022.3176898](https://doi.org/10.1109/TIM.2022.3176898).
- [135] Janani Venugopalan et al. “Multimodal deep learning models for early detection of Alzheimer’s disease stage”. In: *Scientific reports* 11.1 (2021), p. 3254.
- [136] Dimitrios Vlachakis et al. “Improving the utility of polygenic risk scores as a biomarker for alzheimer’s disease”. In: *Cells* 10.7 (2021), p. 1627.
- [137] P. Werner et al. “Handwriting Process Variables Discriminating Mild Alzheimer’s Disease and Mild Cognitive Impairment”. In: *Journal of Gerontology: PSYCHOLOGICAL SCIENCES* 61.4 (2006), pp. 228–36.
- [138] Perla Werner et al. “Handwriting Process Variables Discriminating Mild Alzheimer’s Disease and Mild Cognitive Impairment”. In: *The Journals of Gerontology: Series B* 61.4 (July 2006), P228–P236. ISSN: 1079-5014. DOI: [10.1093/geronb/61.4.P228](https://doi.org/10.1093/geronb/61.4.P228). eprint: <https://academic.oup.com/psychsocgerontology/article-pdf/61/4/P228/9909050/P228.pdf>. URL: <https://doi.org/10.1093/geronb/61.4.P228>.
- [139] Peter West. *Handwriting*. Random House, 2012.
- [140] David M. Wilson et al. “Hallmarks of neurodegenerative diseases”. In: *Cell* 186.4 (2023), pp. 693–714. ISSN: 0092-8674. DOI: <https://doi.org/10.1016/j.cell.2022.12.032>. URL: <https://www.sciencedirect.com/science/article/pii/S0092867422015756>.
- [141] B. Xue et al. “A Survey on Evolutionary Computation Approaches to Feature Selection”. In: *IEEE Transactions on Evolutionary Computation* 20.4 (Aug. 2016), pp. 606–626.
- [142] Mounim A. El-Yacoubi et al. “From aging to early-stage Alzheimer’s: Uncovering handwriting multimodal behaviors by semi-supervised learning and sequential representation learning”. In: *Pattern Recognition* 86 (2018).
- [143] Kun Zhang, Wei Fan, and XiaoJing Yuan. *Ozone Level Detection*. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5NG6W>. 2008.
- [144] Yifan Zhang et al. “Deep Long-Tailed Learning: A Survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023), pp. 1–20. DOI: [10.1109/TPAMI.2023.3268118](https://doi.org/10.1109/TPAMI.2023.3268118).
- [145] Yu Zhang et al. “Strength and Similarity Guided Group-level Brain Functional Network Construction for MCI Diagnosis”. In: *Pattern recognition* 6731 (2018).
- [146] Zigeng Zhang, Christian O’Reilly, and Réjean Plamondon. “Comparing Symbolic and Connectionist Algorithms for Correlating the Age of Healthy Children with Sigma-Lognormal Neuromuscular Parameters”. In: *2022 26th International Conference on Pattern Recognition (ICPR)*. 2022, pp. 4385–4391. DOI: [10.1109/ICPR56361.2022.9956651](https://doi.org/10.1109/ICPR56361.2022.9956651).
- [147] Jenny Ziviani and Margaret Wallen. “Chapter 11 - The Development of Graphomotor Skills”. In: *Hand Function in the Child (Second Edition)*. Ed. by Anne Henderson and Charlane Pehoski. Second Edition. Saint Louis: Mosby, 2006, pp. 217–236. ISBN: 978-0-323-03186-8. DOI: <https://doi.org/10.1016/B978-032303186-8.50014-9>. URL: <https://www.sciencedirect.com/science/article/pii/B9780323031868500149>.

- [148] B. Zoph et al. “Learning Transferable Architectures for Scalable Image Recognition”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2018, pp. 8697–8710.