PhD- FSTM-2025-045
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 04/04/2025 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN *Informatique*

by

Nina HOSSEINI KIVANANI
Born on 22 September 1985 in Kermanshah, Iran

ADVANCING DEMENTIA SCREENING THROUGH HANDWRITING ANALYSIS AND DATA AUGMENTATION

Dissertation defence committee

Prof. Dr. Christoph SCHOMMER, dissertation supervisor
Professor, Université du Luxembourg

Prof. Dr Sanaz FALLAHKHAIR
Professor, University of Brighton

Prof. Dr Luis A. LEIVA, Chairman
Professor, Université du Luxembourg

Prof. Dr Miguel Ángel FERRER,
Professor, Universidad de Las Palmas de Gran Canaria

Prof. Dr Elena SALOBRAR-GARCÍA, Vice Chairman
Professor, Universidad Complutense de Madrid

Dissertation

Department of Computer Science
Faculty of Science, Technology, and Medicine

Advancing Dementia Screening Through Handwriting Analysis and Data Augmentation

Nina Hosseini Kivanani

Acknowledgements

They say a PhD is a marathon, not a sprint. Well, if that's true, then I must have taken a few scenic detours, tripped over my own hypotheses, and possibly run in circles a few times—but I made it! And I certainly didn't do it alone.

First and foremost, I am grateful to Prof. Christoph Schommer for providing me with the opportunity to embark on this PhD journey. It has been a long and winding road, filled with challenges, discoveries, and a few questionable research rabbit holes. While navigating these, I was fortunate to have Prof. Luis Leiva, whose support, and commitment to helping me through all the papers kept this thesis on track. Without his insight and encouragement, this work might have taken different, perhaps more chaotic, direction. I also extend my sincere thanks to Prof. Elena Salobrar-García, whose invaluable help with data collection made a significant difference in this research.

I would like to sincerely thank my esteemed examiners, Prof. Sanaz Fallahkhair, Prof. Miguel Ferrer, and Prof. Jochen Klucken, for their time and effort in reviewing my thesis. I greatly appreciate their willingness to evaluate my work and provide valuable feedback.

To my incredible colleagues and friends, thank you for your support, camaraderie, and for making academic life so much more bearable (and sometimes even fun!). Whether it was deep discussions about research, moral support through tight deadlines, or much-needed coffee breaks filled with laughter, I couldn't have asked for a better PhD squad.

To my parents, there are no words to fully express my gratitude. Your endless love, sacrifices, and unwavering belief in me have been my anchor throughout this journey. And to my brother, Soheil, thank you for always being there, for your encouragement, and for tolerating my academic rants even when they made absolutely no sense.

A special and heartfelt thank you to my partner, who has been my rock through all the ups and downs, reminding me that there is life beyond research papers, debugging nightmares, and conference deadlines. Your patience, love, and support have meant the world to me (especially during the times when I questioned all my life choices).

And finally, to coffee, deadlines, and sheer stubbornness—you played a vital role in getting me here.

To everyone who has been part of this journey, whether through academic guidance, emotional support, or simply by listening to my endless ramblings about research, I am profoundly grateful.

Nina Hosseini Kivanani — Luxembourg, April 2025

Dedicated to my Behi ...

Abstract

Early detection of dementia, particularly Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI), remains a significant challenge in healthcare. This dissertation investigates handwriting-based cognitive assessments as a viable alternative, leveraging off-line (scanned images) and on-line (digitally captured) drawing tasks to enhance classification accuracy through deep learning (DL), data augmentation, and transfer learning.

One of the central contributions of this work is the study of dataset size requirements for AI-based dementia screening (*The Magic Number*), demonstrating that EfficientNet achieves reliable performance with only half of the available data. These findings challenge the assumption that large-scale datasets are indispensable for robust screening. Another key aspect involves a deep feature concatenation (DFC) framework (*Better Together*), which integrates multiple handwriting sources—pentagon drawings, sentences, and signatures—leading to a classification improvement from 60% (single-source) to 80% (multi-source with augmentation).

A comparative assessment of drawing modalities (*Blueprint of Tomorrow*) establishes that off-line handwriting analysis provides better results than on-line methods, with EfficientNet achieving 90% accuracy in binary classification. This investigation extends to two studies focused on the House Drawing Test (HDT), where one examines off-line and on-line representations independently, and the other introduces an approach that converts and refines on-line data into off-line format. This method yields 82% accuracy (86% AUC), reinforcing the suitability of static image-based models for automated analysis.

This investigation extends to two studies focused on the House Drawing Test (HDT), where one examines off-line and on-line representations independently (*Predicting Alzheimer's Disease and Mild Cognitive Impairment*), and the other introduces an approach that converts and refines on-line data into off-line format (*Screening of Alzheimer's Disease*). The latter achieves 82% accuracy (86% AUC), reinforcing the suitability of static image-based models for automated analysis.

This dissertation also presents the first evaluation of data augmentation techniques for handwriting-based dementia detection (*Ink of Insight*). The effects of augmentation strategies are analyzed across classical machine learning (SVM, RF, k-NN) and DL architectures, with EfficientNet achieving 87% accuracy (91% AUC). Further investigation (*Efficient Automatic Data Augmentation*) explores automated augmentation methods, demonstrating that non-learnable techniques (TrivialAugment, UniformAugment) improve generalization by up to 15% with minimal computational demands.

These studies establish handwriting analysis as a scalable, non-invasive, and cost-effective tool for dementia screening. By refining dataset requirements, exploring augmentation strategies, and integrating multiple handwriting sources, this research advances AI applications in neurodegenerative disease diagnostics and lays the groundwork for further developments in computational healthcare.

List of Publications

This dissertation is mainly based on the investigations and studies presented in the following publications:

1. **N. Hosseini-Kivanani**, E. Salobrar-García, L. Elvira-Hurtado, I. López-Cuenca, R. de Hoz, J. M. Ramírez, P. Gil, M. Salas, C. Schommer, and L. A. Leiva, "Better together: Combining different handwriting input sources improves dementia screening," in *Proceedings of the 2023 IEEE 19th International Conference on e-Science*, 2023, pp. 1–7. DOI: 10.1109/e-Science58273.2023.10254799

2. **N. Hosseini-Kivanani**, C. Schommer, and L. A. Leiva, "The Magic Number: Impact of sample size for dementia screening using transfer learning and data augmentation of clock drawing test images," in *Proceedings of the 2023 IEEE International Conference on E-health Networking, Applications and Services (Healthcom)*, 2023, pp. 101–106. DOI: 10.1109/Healthcom56612.2023.10472399

3. **N. Hosseini-Kivanani**, E. Salobrar-García, L. Elvira-Hurtado, I. López-Cuenca, R. de Hoz, J. M. Ramírez, P. Gil, M. Salas-Carrillo, C. Schommer, and L. A. Leiva, "Ink of insight: Data augmentation for dementia screening through handwriting analysis," in *Proceedings of the 2024 8th International Conference on Medical and Health Informatics (ICMHI)*, 2024, pp. 224–229. DOI: 10.1145/3673971.3673992

4. **N. Hosseini-Kivanani**, E. Salobrar-García, L. Elvira-Hurtado, M. Salas, C. Schommer, and L. A. Leiva, "Predicting Alzheimer's disease and mild cognitive impairment with off-line and on-line house drawing tests," in *Proceedings of the 2024 IEEE 20th International Conference on e-Science*, 2024, pp. 1–10. DOI: 10.1109/e-Science62913.2024.10678661

5. **N. Hosseini-Kivanani**, E. Salobrar-García, L. Elvira-Hurtado, M. Salas, C. Schommer, and L. A. Leiva, "Screening of Alzheimer's disease and mild cognitive impairment through integrated on-line and off-line house drawing tests," *IADIS International Journal on Computer Science and Information Systems*, pp. 37–53, 2024.

6. **N. Hosseini-Kivanani**, E. Salobrar-García, L. Elvira-Hurtado, M. Salas, C. Schommer, and L. A. Leiva, "Blueprint of Tomorrow: Contrasting Off-Line and On-Line Drawing Tasks for Alzheimer's Disease Screening," in *Proceedings of Intelligent Data Engineering and Automated Learning – IDEAL 2024, Lecture Notes in Computer Science*, vol. 15346, V. Julian *et al.*, Eds. Cham: Springer, 2025, pp. 422–433. DOI: 10.1007/978-3-031-77731-8_38

7. **N. Hosseini-Kivanani**, I. Oliveira, S. Kilinç, and L. A. Leiva, "Efficient Automatic Data Augmentation of CDT Images to Support Cognitive Screening," in *Proceedings of the 17th International Conference on Agents and Artificial Intelligence- ICAART 2025*, vol. 3. pp. 600–607. DOI: 10.5220/0000196100003890

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AA	AutoAugment
Acc	accuracy
AD	Alzheimer's disease
ADRDA	Alzheimer's Disease and Related Disorders Association
AI	Artificial Intelligence
AVC	All Variability Chain
AUC	area under the receiver operating characteristic curve
AutoDA	automatic data augmentation
AWS	Augmentation-Wise Weight Sharing
BiGRU	Bidirectional Gated Recurrent Unit
BiLSTM	Bidirectional Long Short-Term Memory
BiRNN	Bidirectional vanilla RNN
BO	Bayesian Optimization
CDR	Clinical Dementia Rating
CDT	Clock Drawing Test
cCDT	conventional CDT
CNNs	Convolutional Neural Networks
CONV	Convolutional
CV	Cross-validation

DA	data augmentation
DFC	deep feature concatenation
DL	Deep Learning
dCDT	digital Clock Drawing Test
DNN	deep neural network
DSM V	Statistical Manual of Mental Disorders V
dTDT	digital Tree Drawing Test
DTW	Dynamic Time Warping
EHR	Electronic Health Record
Fast AA	Fast AutoAugment
FC	Fully Connected
fMRI	Functional magnetic resonance imaging
GRUs	Gated Recurrent Units
HCSC	Hospital Clinico San Carlos
HCs	healthy controls
HDT	House Drawing Test
***k*-NN**	*k*-nearest neighbors
LC	learning curve
LSTMs	Long Short-Term Memory networks
ML	Machine Learning
MCI	Mild Cognitive Impairment
MMSE	Mini-Mental State Examination
MoCA-K	Korean version of the Montreal Cognitive Assessment
MRI	magnetic resonance imaging

NAS	Neural Architecture Search
NINCDS	National Institute of Neurological and Communicative Disorders and Stroke
PBA	Population-Based Augmentation
PD	Parkinson’s disease
PDT	Pentagon Drawing Test
PET	positron emission tomography
POOL	Pooling
RA	RandAugment
RF	random forest
RCFT	Rey–Osterrieth Complex Figure Test-copy
ReLU	Rectified Linear Unit
RNNs	Recurrent Neural Networks
SSIM	Structural Similarity Index Measure
SVM	Support Vector Machine
TA	TrivialAugment
TDT	Tree Drawing Test
TMT	Trail Making Test
UA	UniformAugment
VGG	Visual Geometry Group

1 Introduction and Related Work

1.1 Introduction

Dementia is a complex clinical syndrome marked by a progressive decline in cognitive functions, including memory, problem-solving, and language. This deterioration ultimately affects an individual's ability to carry out daily tasks and maintain independence [49, 12]. Among the various forms of dementia, Alzheimer's disease (AD) is the most prevalent, presenting significant medical, social, and economic challenges on a global scale [16]. As life expectancy continues to increase, the number of people affected by AD is predicted to grow exponentially, placing additional strain on healthcare systems and caregivers [27].

Rising life expectancy in the global population above 60 years of age, which continues climbing without a clear limit, has introduced additional health challenges [24]. The growing number of older adults correlates with an increased incidence of neurodegenerative conditions, including AD and Parkinson's disease (PD) [21], originally identified in 1907 by Alois Alzheimer [20]. These conditions involve amyloid plaques and tau-related neurofibrillary tangles, as well as synaptic loss in the neocortex and limbic system. These changes often begin before noticeable symptoms appear, affecting memory, visuospatial skills, and other cognitive processes [47].

A related condition, Mild Cognitive Impairment (MCI), represents an intermediate stage between normal aging and dementia [5]. Individuals diagnosed with MCI exhibit cognitive deficits that exceed what is expected of their age and educational background, but do not yet interfere significantly with their daily activities or autonomy [2]. Although some individuals with MCI remain stable or regain normal cognitive function, a considerable proportion eventually develop AD [41]. Therefore, identification and monitoring of MCI are essential, as early intervention can help slow disease progression or allow patients and families to prepare for future care needs [18].

Despite extensive research efforts, there is currently no definitive cure for AD, and available pharmacological treatments primarily focus on symptom management [6, 4]. As a result, early detection has emerged as a crucial strategy, offering the potential to extend functional independence, improve treatment efficacy, and enhance overall quality of life [13]. From a public health perspective, early diagnosis can also reduce long-term care costs, optimize healthcare resource allocation, and mitigate the broader societal impact of dementia care [49, 29].

The effectiveness of early detection depends on the availability of reliable, accessible, and cost-effective methods, particularly in primary care settings and underserved communities [34]. Current diagnostic approaches typically involve clinical assessments, standardized cognitive tests, neuroimaging, and biomarker analysis. While techniques such as magnetic resonance imaging (MRI) and positron emission tomography (PET) provide valuable diagnostic insights, they are often expensive, invasive, and resource-intensive [7]. This highlights the need for alternative screening strategies that are noninvasive, easily deployable, and suitable for integration into routine clinical practice or community settings [10].

Handwriting-based assessments, including drawing tasks, have gained increasing attention in recent years due to their ability to assess multiple cognitive functions, such as visuospatial processing, motor coordination, and executive function [8]. These assessments are simple to administer and interpret, making them practical for both clinicians and patients. Advances in technology now enable automatic analysis of handwriting data, providing objective and quantifiable metrics that can support clinical decision-making [35]. Detecting cognitive impairments at an early stage creates opportunities for timely intervention, participation in therapeutic trials, and improved long-term care planning, particularly in the context of an aging population [52].

This dissertation investigates the potential of handwriting-based diagnostics for AD and MCI. Through a series of studies, it examines advanced methodologies, including deep feature concatenation, data augmentation, and dual-modal data integration. These approaches address key challenges in the field, such as small dataset sizes, variability in handwriting features, and the need for robust diagnostic models. By incorporating both offline and online handwriting data, this research enhances the diagnostic framework, facilitating more comprehensive feature extraction, and improving classification accuracy.

The findings presented in this dissertation demonstrate that handwriting analysis, supported by Artificial Intelligence (AI), offers a promising approach to dementia diagnosis. By providing a cost-effective, accessible, and non-invasive alternative to traditional assessment methods, these advancements contribute to the development of early detection tools that can be used in clinical and home-based environments. Beyond advancing scientific understanding, this work presents practical applications for improving cognitive health assessment on a global scale.

1.2 Related Work

Recent advancements in AI-driven tools for dementia diagnosis have prioritized the development of non-invasive, scalable, and cost-effective cognitive assessment methods [1]. The Clock Drawing Test (CDT), Trail Making Test (TMT), and Pentagon Drawing Test (PDT), long-standing neuropsychological assessments, have been central to automated cognitive impairment detection due to their ability to evaluate multiple cognitive domains with minimal patient burden [31]. AI-based tools analyzing CDT, PDT, and TMT have demonstrated high accuracy in detecting cognitive impairments and offer a promising alternative to traditional paper-based tests [30].

The digitization of CDT tools has progressed significantly, with an emphasis on integrating real-time feedback and deep learning for automated scoring [43]. Their study introduced mCDT, a mobile application that leverages deep learning models such as U-Net and CNN for qualitative scoring of CDT drawings. By utilizing mobile sensors, mCDT enables real-time data capture and automated analysis of hand-drawn clock images, assessing key parameters such as contour accuracy, number placement, clock hand positioning, and spatial alignment. The system was validated using 159 CDT hand-drawn images and tested on 219 subjects, achieving high sensitivity (89.33%), specificity (92.68%), and precision (98.15%) in differentiating normal and impaired cognitive performance. Their findings highlight the potential of mobile-based CDT assessments in reducing reliance on subjective manual scoring while improving accessibility to cognitive screening in clinical and community settings. The integration of mobile sensors and AI-driven analysis marks a significant step toward scalable, cost-effective, and real-time cognitive assessment tools that can be deployed beyond traditional clinical environments.

Kuok et al. (2024) developed a mobile application that digitizes the CDT for automated data capture and classification of cognitive impairment. The system employs CoreML and Vision Frameworks to process drawings on iPads, addressing errors related to conceptual deficits, spatial planning, and stimulus-bound responses. By automating CDT scoring, the application reduces subjectivity and improves accessibility for clinical use. The study demonstrates the potential of digital CDT assessments but identifies dataset size as a limitation. The manually curated dataset consists of 52 normal and 63 abnormal clock drawings, categorized into "Pass" and "Fail" classes. Data augmentation techniques, including noise addition, blurring, cropping, and rotation, were applied to mitigate overfitting. Despite achieving a validation accuracy of 86.7% and test accuracy of 81%, broader dataset diversity remains necessary for improving model performance. The model was tested on clinically validated drawings and classified common error types with confidence scores ranging from 75% to 100%. While the findings indicate promise for cognitive impairment screening, larger datasets and further evaluation of real-world usability are needed. The study contributes to the development of digital tools for early dementia detection, highlighting challenges related to model generalization and clinical integration [33].

Similarly, Prange et al. (2021) introduced an automated cognitive assessment tool for the TMT, utilizing digital pen input to capture dynamic pen features [45]. Their system measured completion time and generated structured reports detailing errors, ensuring transparent and interpretable results. Evaluated with 40 elderly participants, the system demonstrated high concordance with manual assessments, reinforcing the role of automated methods in enhancing reliability and transparency in cognitive test evaluations. The use of digital pens further enhances objectivity and standardization, addressing common errors associated with human scoring biases [23]. Unlike traditional paper-based assessments, digital pen technology allows real-time analysis of task execution, providing clinicians with quantifiable metrics such as stroke velocity, pen pressure, hesitation time, and corrective movements. These features improve diagnostic accuracy and facilitate the early detection of cognitive decline by revealing subtle impairments that may be overlooked in manual scoring. Furthermore, automated scoring systems integrated with digital pens ensure greater reproducibility and transparency, allowing for remote assessments and integration into large-scale screening programs [38].

Handwriting analysis has emerged as a promising method for early AD detection, leveraging the interconnection between cognitive and motor functions [50]. Mitra and Rehman (2024) introduced an ensemble Machine Learning (ML) model for AD detection using handwriting kinetics. Their approach, applied to the DARWIN dataset containing handwriting data from 174 individuals (89 AD patients, 85 healthy participants), employed stacking techniques to integrate multiple classifiers. The model achieved 97.14% accuracy, 95% sensitivity, and 97.5% area under the receiver operating characteristic curve (AUC), surpassing existing methods. This study reinforces the clinical potential of machine learning-based handwriting analysis as a non-invasive and cost-effective diagnostic tool [40].

Deep Learning (DL) has further advanced CDT analysis. Chen et al. (2020) [11] developed a CNN-based CDT scoring system, achieving high diagnostic accuracy while addressing variability in patient drawings. Chan et al. (2021) emphasized the benefits of combining digital and traditional CDT approaches, demonstrating that CNN-based analyses can match or exceed traditional clinical assessments in sensitivity and specificity [9].

Park et al. (2024) explored CNNs for distinguishing MCI from healthy controls using CDT images. Their study, involving 177 CDT images (103 healthy controls, 74 a-MCI patients), applied preprocessing and data augmentation techniques. Their CNN model achieved an accuracy of 88.7% and an AUC of 0.886, outperforming the Korean version of the Korean version of the Montreal Cognitive Assessment (MoCA-K). Grad-CAM visualizations indicated that the model focused on key features such as clock letter placement and hand positions, validating its interpretability and clinical relevance. This research demonstrates the feasibility of CNN-based CDT analysis as a reliable MCI screening tool [44].

Beyond CDT and TMT, DL has benefited other drawing-based cognitive assessments. Li et al. (2021) developed a deep learning framework for PDT scoring, assessing visuospatial function. Their study utilized 823 PDT images, implementing two CNN-based approaches: supervised transfer learning with Inception-V3 and an object detection method using YOLOv5m. The object detection approach significantly outperformed classification-based methods, achieving 93.8% accuracy and an AUROC of 0.954. By improving scoring efficiency and reliability, this work demonstrates the applicability of CNN architectures in the automation of cognitive tests [36].

The digitization of handwriting assessments has enhanced diagnostic precision by capturing temporal and kinematic data. Studies by Souillard-Mandar et al. (2015)[48] and Davoudi et al. (2021)[14] demonstrated the effectiveness of digital CDT systems in differentiating AD from MCI using stroke dynamics and pen pressure. However, these systems often require specialized hardware, limiting accessibility. Kuok et al. (2024) [33] addressed this challenge by developing a CDT system for consumer-grade tablets, reducing dependency on proprietary medical devices. Despite this advancement, dataset limitations and concerns over model generalizability persist.

To improve scalability, Impedovo et al. (2019) proposed a framework that integrates digital and pen-and-paper methods while maintaining diagnostic reliability [26]. Their approach combines handwriting tasks with digitized cognitive assessments, facilitating broader clinical adoption. By analyzing kinematic features such as movement speed, pen pressure, and in-air time, their study involving over 100 participants demonstrated that handwriting-based screening could effectively distinguish between healthy individuals and those with MCI. The results support the development of scalable handwriting protocols that incorporate both conventional and digital methods, ensuring accessibility while preserving continuity with established neurological evaluation standards.

Prange et al. (2021) advanced this field by developing an automatic scoring system for the TMT that applies digital pen technology to cognitive assessment [45]. The TMT, widely used to assess executive functioning and processing speed, is typically scored manually with a stopwatch, introducing variability. Their system not only measures total completion time but also captures pen movement patterns, including spatial accuracy, pen pressure, and in-air time. By incorporating a digital pen with the standard paper-based TMT, the system aligns with existing clinical protocols while allowing for automated feature extraction. The analysis detects errors such as missed connections, incorrect sequences, and prolonged pauses, generating structured reports that provide clinicians with both quantitative metrics and visual representations of patient performance. In an evaluation involving 40 elderly participants, the system showed strong agreement with manual scoring while also identifying patterns of cognitive decline through error analysis. The study confirmed the relevance of in-air movement analysis, aligning with prior handwriting-based screening research. The inclusion of explainability features also mitigates concerns associated with black-box machine learning models, offering clinicians greater transparency in decision-making. The ability to provide real-time feedback and structured digital records enhances the practicality of this approach in clinical settings, supporting more standardized and widely accessible cognitive assessments.

Multi-task learning methods have been applied to cognitive screening to improve diagnostic accuracy. Qi et al. (2024) examined the combination of handwriting-based stroke features with gait analysis for AD detection, reporting a classification accuracy of 96.17%. Their results surpassed models that relied solely on handwriting or gait assessments, suggesting that integrating fine motor and motor coordination features provides a more comprehensive representation of cognitive decline. The study identified handwriting pressure, stroke kinematics, and gait stability as predictive markers for early-stage AD [46].

Fan et al. (2024) expanded on this approach by introducing a multi-target, multi-task learning model aimed at predicting AD progression through correlated cognitive scores. Their method incorporated assessments such as the Clinical Dementia Rating (CDR) and Mini-Mental State Examination (MMSE) to capture the interrelations between cognitive domains. Unlike machine learning models designed for a single cognitive metric, this framework considered latent connections among neuropsychological scores, which led to improved predictive performance [17].

Liu et al. (2024) further explored multi-task learning by developing a feature-sharing deep learning model for AD classification and MMSE score estimation. Their study, published in Big Data Mining and Analytics, demonstrated that jointly modeling diagnostic classification and cognitive score prediction improves accuracy and clinical relevance [37]. By incorporating structured feature-sharing modules, their model effectively captured cross-task dependencies, achieving leading performance across multiple AD-related datasets. These findings highlight the role of deep learning in refining neurodegenerative disease assessment by leveraging shared representations across related tasks.

1.2.1 Off-line vs. On-line Data

A fundamental distinction in handwriting-based data collection concerns whether the drawn content is captured as a static image (off-line) or recorded in real time (on-line), preserving each stroke's dynamic properties. These two modalities offer distinct advantages and limitations in clinical applications. Off-line handwriting analysis, typically based on scanned images, provides a stable format that is easily stored and processed. However, it does not include temporal and kinematic details, which are essential for assessing fine motor function [15]. In contrast, on-line handwriting captures stroke timing, velocity, and pen pressure, allowing for a more detailed evaluation of cognitive and neuromotor function. Efforts to bridge the gap between these approaches have led to the development of techniques to reconstruct on-line handwriting trajectories from off-line images, expanding the diagnostic potential of handwriting-based assessments [51].

Off-line data refers to static images of handwriting tasks, typically obtained from paper-and-pencil assessments that are later digitized by scanning or photographing [15]. This approach is widely used in clinical settings due to its simplicity and accessibility. Traditional pen-and-paper tests do not require specialized equipment, making them easy to administer and familiar to both clinicians and patients [39]. Furthermore, off-line handwriting analysis is cost-effective as it does not depend on digital devices or real-time tracking sensors, making it particularly suitable for resource-constrained environments. Scanned or photographed images can also be processed using standard image analysis techniques, enabling the application of deep learning models such as Convolutional Neural Networks (CNNs) to identify spatial patterns associated with cognitive decline [42].

Despite these advantages, offline handwriting analysis has notable limitations. The absence of temporal and kinematic data prevents the capture of stroke order, writing speed, and pen pressure—key indicators of fine motor control and cognitive function [15]. In addition, scanned images can introduce noise and artifacts, including shadows, distortions, and extraneous markings, which can hinder the accurate extraction of features and reduce classification accuracy [22]. Recent research has attempted to recover stroke dynamics from offline handwriting images, but these methods remain limited in capturing real-time neuromotor features [3].

On-line data collection, in contrast, captures the dynamic aspects of handwriting in real-time using digital devices equipped with sensors [32]. This modality provides richer information by tracking stroke velocity, hesitations, pressure, and movement sequences, offering deeper insights into neuromotor function. Studies have demonstrated that on-line handwriting metrics are valuable in distinguishing cognitive impairments, as patients with MCI and AD often exhibit slower writing speeds and increased hesitation compared to healthy individuals [19, 32]. Furthermore, on-line handwriting analysis supports real-time feedback, allowing clinicians to observe and adjust assessments during data collection [52].

Despite its advantages, on-line handwriting analysis presents challenges in clinical implementation. The requirement for specialized hardware, such as digitizing tablets or sensor-equipped pens, increases both cost and logistical complexity [25]. Moreover, older individuals or those with limited experience using digital devices may require training before they can effectively complete handwriting tasks in a digital format [28].

Díaz et al. (2019) [15] introduced a method for embedding temporal and velocity-based features into static handwriting representations, demonstrating that "dynamically enhanced" off-line images can improve classification performance in PD detection. Their approach retained critical dynamic features by plotting discrete points instead of linking strokes, allowing speed variations to be visualized within a single static representation. The inclusion of pen-up movements further contributed to identifying neuromotor control deficits in PD patients. Their findings indicated that this hybrid representation performed better than models relying solely on either static or dynamic handwriting features, reinforcing the value of incorporating kinematic information into off-line handwriting analysis. Building on this approach, Xiong et al. (2023) [51] developed a deep learning framework for reconstructing handwriting trajectories from static character images, addressing the inherent limitations of off-line handwriting analysis. Their model employed an encoder-decoder structure with two prediction streams: one dedicated to reconstructing stroke sequences as discrete point trajectories and another generating frame-based motion representations. A sequence consistency module was introduced to synchronize both streams during training, allowing the model to capture complementary spatial and temporal characteristics. Experimental evaluations demonstrated that this method produced handwriting trajectories closely resembling real on-line data. These results suggest that recovering dynamic features from static handwriting could extend the scope of handwriting-based cognitive and motor assessments. This advancement holds significance for neurodegenerative disease diagnosis, particularly in settings where real-time handwriting data collection may not be feasible, yet kinematic features remain essential for early-stage detection.

1.3 Key Scientific Contributions

This section summarizes the key publications included in the thesis, highlighting the goals, methods, results, and overall relevance of each article to the field.

1. **Paper 1**: “Better Together: Combining Different Handwriting Input Sources Improves Dementia Screening.”
   - Aim & Significance: This study examines the impact of integrating multiple handwriting tasks—including pentagon drawings, sentences, and signatures—on the early detection of AD. The research aims to assess whether diverse input sources enhance diagnostic accuracy and mitigate data limitations.
   - Methods: A deep feature concatenation (DFC) approach is introduced that merges feature vectors from CNN-based submodels. The models include a custom CNN alongside state-of-the-art architectures (e.g., EfficientNet, ResNet, DenseNet).

- Results: The model achieved an 80% classification accuracy, revealing a 20-point improvement after augmenting smaller subsets of scanned handwriting data.
- Impact: The findings underscore the potential of multi-source handwriting analysis in dementia screening, providing a cost-effective, non-invasive diagnostic tool. This approach reduces the reliance on single-task assessments, offering a more holistic evaluation of cognitive decline

2. **Paper 2**: "The Magic Number: Impact of Sample Size for Dementia Screening Using Transfer Learning and Data Augmentation of Clock Drawing Test Images."
    - Aim & Significance: This study systematically investigates the impact of varying sample sizes on classification performance in AD detection using CDT images.
    - Methods: A learning curve analysis was conducted on pre-trained CNN models, including VGG-16, ResNet-152, and DenseNet-121. The study explores the effects of training with different proportions of the dataset, both with and without data augmentation (DA), to evaluate the feasibility of using smaller datasets without significantly compromising model performance.
    - Results: Findings indicate that approximately 50–75% of the dataset is sufficient to achieve near-peak accuracy for AD classification tasks. The DenseNet-121 model, in particular, demonstrated robust performance, reaching competitive accuracy levels with as little as 50% of the data. While binary classification (healthy vs. patient) reaches optimal accuracy with 50% of the dataset, multi-class classification (six AD severity stages) requires around 75% for comparable results.
    - Impact: This study provides empirical guidelines on the minimum data requirements for effective AD screening using DL models, making it particularly valuable for clinical settings where data collection is challenging. The work underscores the feasibility of using smaller datasets for dementia screening while maintaining high classification accuracy, thus facilitating broader adoption of AI-driven diagnostic tools in healthcare.
3. **Paper 3**: "Ink of Insight: Data Augmentation for Dementia Screening Through Handwriting Analysis."
    - Aim & Significance: This research focuses exclusively on the design and effect of augmentation techniques for analyzing a small set of PDT with CNN.

   - Methods: A benchmark analysis was conducted using standard geometric transformations (flipping, rotation, and translation) with different intensity levels. The study compares the effects of augmentation on both classic ML models (Support Vector Machine (SVM), random forest (RF), k-nearest neighbors (k-NN)) and state-of-the-art CNN architectures (VGG-16, ResNet-152, DenseNet-121, EfficientNet, and a custom CNN).
   - Results: The application of validated augmentation techniques led to a classification accuracy improvement of approximately 10–30% in distinguishing mild AD from healthy controls. Among the tested models, EfficientNet achieved the highest accuracy of 87% and an AUC of 91% in binary classification, while multiclass classification (healthy, mild AD, moderate AD) reached 76% accuracy and 77% AUC.
   - Impact: The findings highlight that selective transformations (e.g., ignoring color-based manipulations) can enhance performance in clinically relevant tasks.

4. **Paper 4**: “Predicting Alzheimer’s Disease and Mild Cognitive Impairment with Off-line and On-line House Drawing Tests.”
   - Aim & Significance: This study introduces our new dataset capturing House Drawing Test (HDT) in both scanned images (off-line) and digitized pen strokes (on-line) for AD screening.
   - Methods: Compares CNNs models for images vs. Recurrent Neural Networks (RNNs) models for sequences, with different augmentation strategies.
   - Results: The findings show that RNNs outperform CNNs in multi-class classification (distinguishing AD, MCI, and healthy participants), whereas CNNs perform better in binary classification tasks.
   - Impact: The study demonstrates the complementary nature of off-line and on-line handwriting data and provides insights into optimal classification strategies for different diagnostic settings.

5. **Paper 5**: “Screening of Alzheimer’s Disease and Mild Cognitive Impairment Through Integrated On-line and Off-line House Drawing Tests.”
   - Aim & Significance: Continues exploring house drawing tasks, but investigates a pipeline that merges the dynamic and static data representations for improved detection.

- Methods: The approach uses CNNs for off-line images and RNNs for on-line stroke data. Data augmentation techniques, including standard geometric transformations and the All Variability Chain (AVC), were applied.
- Results: The combination of on-line and off-line handwriting data improved classification performance compared to using either data type alone, with a 10% increase in multi-class classification accuracy. The BiGRU model achieved the best results for on-line data, while EfficientNet showed the highest accuracy for off-line images.
- Impact: The findings suggest that using both handwriting data formats together enhances cognitive impairment detection. This study supports the development of clinical tools that capture both real-time stroke patterns and static handwriting details to improve early diagnosis of AD and MCI.

6. **Paper 6**: "Blueprint of Tomorrow: Contrasting Off-line and On-line Drawing Tasks for Alzheimer's Disease Screening."
    - Aim & Significance: More in-depth analysis of the advantages and disadvantages of on-line vs. off-line data collection.
    - Methods: Studies both clock drawing and pentagon drawing tasks, each captured off-line or on-line, and compares classification performance under identical conditions.
    - Results: The analysis indicates that off-line images of clock and pentagon drawings outperform their on-line counterparts, particularly in tasks requiring the differentiation of mild and moderate AD cases. EfficientNet achieved the best results, with 90% accuracy in binary classification and 82% accuracy in multi-class classification. The on-line models, particularly RNNs, struggled with the small dataset and the inherent variability of stroke patterns.
    - Impact: The results emphasize the potential of simple, image-based models for clinical applications, especially when datasets are small and variability in stroke dynamics is high.
7. **Paper 7**: "Efficient Automatic Data Augmentation of CDT Images to Support Cognitive Screening."
    - Aim & Significance: Explores new "automatic" data augmentation (AutoDA) algorithms (e.g., RandAugment (RA), TrivialAugment (TA), UniformAugment (UA)), comparing them to classical/manual augmentation on multiple CDT benchmarks.

- Methods: Evaluates the computational overhead, classification accuracy, and AUC under different training conditions using state-of-the-art CNN architectures.
- Results: The findings show that non-learnable augmentation methods (TA and UA) perform as well as or better than RA in many cases, particularly on smaller datasets. The best-performing setup, TA with EfficientNet, improved classification accuracy by up to 15% compared to the baseline (no augmentation). TA and UA consistently boosted model performance across all datasets, while RA demonstrated inconsistent gains and higher computational costs.
- Impact: The study highlights that simple, non-learnable AutoDA methods can effectively improve the accuracy and generalization of clinical models without computationally intensive policy learning.

1.4 Summary and Thesis Organization

This introduction has provided a rationale for focusing on handwriting-based AD screening, highlighted the gaps in the current literature, and explained how each paper in this cumulative thesis addresses these critical issues. The broad approach applies advanced DL methods and curated data augmentation, with an emphasis on small-scale clinically relevant data.

The subsequent chapters delve into each research direction in detail. Figure 1.1 presents an overview of the thesis structure, highlighting the key topics explored:

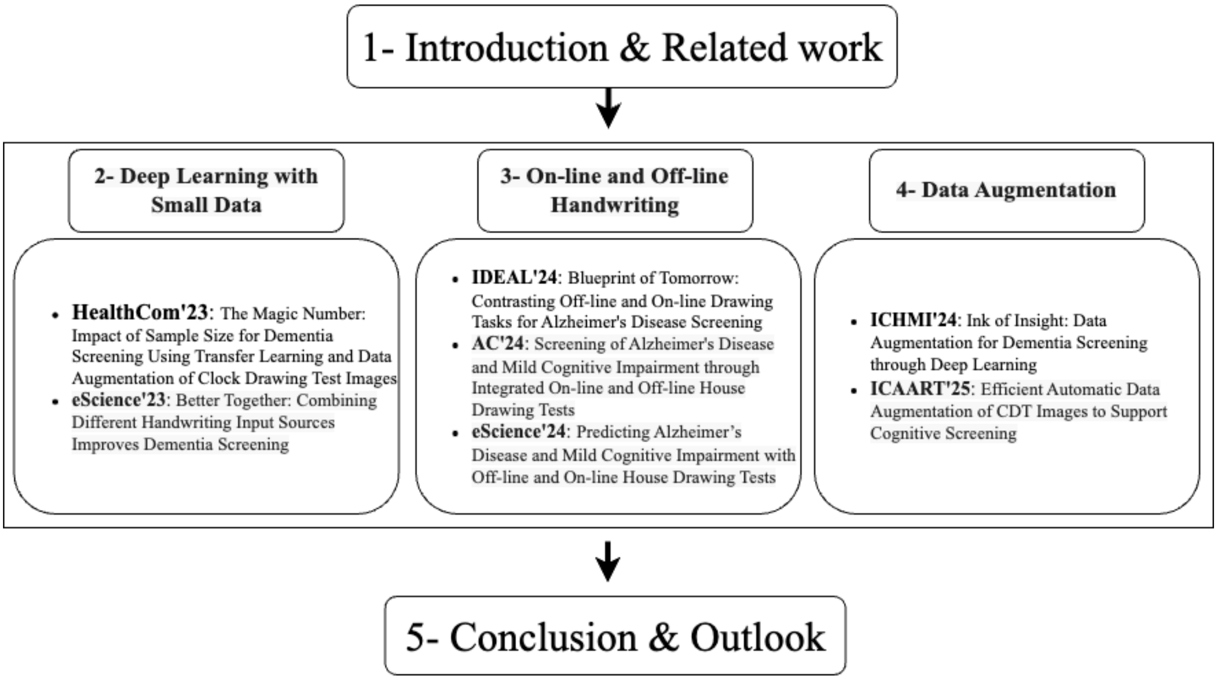
- Chapter 2 (Deep Learning with Small Data): This chapter analyzes the impact of sample size on CNN models using the CDT dataset, emphasizing that larger datasets lead to better model performance. Given the limitations of small datasets, the chapter explores the combination of different handwriting input sources to enhance classification accuracy. Papers:
  1. HealthCom'23: The Magic Number: Impact of Sample Size for Dementia Screening Using Transfer Learning and Data Augmentation of Clock Drawing Test Images
  2. eScience'23: Better Together: Combining Different Handwriting Input Sources Improves Dementia Screening
- Chapter 3 (On-line and Off-line Handwriting): This chapter investigates the use of both off-line and on-line handwriting data for AD screening. It compares the effec-

tiveness of these modalities and reveals that off-line data consistently outperforms on-line data for small-scale datasets. Papers:

1. IDEAL'24: Blueprint of Tomorrow: Contrasting Off-line and On-line Drawing Tasks for Alzheimer's Disease Screening
2. AC'24: Screening of Alzheimer's Disease and Mild Cognitive Impairment through Integrated On-line and Off-line House Drawing Tests
3. eScience'24: Predicting Alzheimer's Disease and Mild Cognitive Impairment with Off-line and On-line House Drawing Tests

- Chapter 4 (Data Augmentation): This chapter focuses on practical and affordable data augmentation techniques to improve model performance in clinical settings. Non-learnable methods like TA and UA are highlighted for their effectiveness in addressing small, imbalanced datasets. Papers:

1. ICHMI'24: Ink of Insight: Data Augmentation for Dementia Screening through Deep Learning
2. ICAART'25: Efficient Automatic Data Augmentation of CDT Images to Support Cognitive Screening


1- Introduction & Related work

2- Deep Learning with Small Data

- **HealthCom'23**: The Magic Number: Impact of Sample Size for Dementia Screening Using Transfer Learning and Data Augmentation of Clock Drawing Test Images
- **eScience'23**: Better Together: Combining Different Handwriting Input Sources Improves Dementia Screening

3- On-line and Off-line Handwriting

- **IDEAL'24**: Blueprint of Tomorrow: Contrasting Off-line and On-line Drawing Tasks for Alzheimer's Disease Screening
- **AC'24**: Screening of Alzheimer's Disease and Mild Cognitive Impairment through Integrated On-line and Off-line House Drawing Tests
- **eScience'24**: Predicting Alzheimer's Disease and Mild Cognitive Impairment with Off-line and On-line House Drawing Tests

4- Data Augmentation

- **ICHMI'24**: Ink of Insight: Data Augmentation for Dementia Screening through Deep Learning
- **ICAART'25**: Efficient Automatic Data Augmentation of CDT Images to Support Cognitive Screening

5- Conclusion & Outlook

Figure 1.1: Structural outline of the thesis, highlighting the key topics and corresponding published works.

The concluding chapter (Chapter 5) offer cross-paper reflections, unify the distinct findings into a coherent perspective, and propose an outlook on the future of digital and AI-assisted medicine.

2 Deep Learning with Small Data

Dementia screening, particularly in the early identification of ADs, faces two significant challenges. The first challenge is the reliance of DL models on extensive data collections, a demand that is rarely met in clinical settings due to the logistical and financial constraints involved in acquiring and annotating drawing tasks. The second challenge concerns the predominant focus on single-source handwriting assessments, leaving unexplored the potential benefits of incorporating a variety of handwriting samples.

The *HealthCom'23* study examines the interplay between dataset size and model performance in the context of dementia screening using CDT images. A systematic evaluation of convolutional neural networks on this task reveals that model performance attains a plateau beyond a certain data threshold. This result indicates that, when paired with appropriate image modification techniques and precise parameter tuning, DL models can yield dependable outcomes even in data-scarce conditions.

In contrast, the *eScience'23* study addresses the scarcity of data by merging multiple handwriting sources—including pentagon drawings, personal sentences, and signatures—through a method that concatenates features extracted from different neural network architectures. This approach aims to capture a more complete profile of cognitive function by integrating heterogeneous handwriting inputs. Empirical results demonstrate that the fusion of these varied sources, in combination with specialized image processing procedures, enhances classification performance and alleviates the limitations inherent in small sample sizes.

Collectively, these investigations reveal an important gap in current research on computer-aided dementia detection. They illustrate that DL models in this domain may operate effectively without vast data repositories and that a diversified approach to handwriting analysis can improve the discriminative capacity of diagnostic systems. This work establishes a basis for developing clinically pertinent screening tools, thereby refining early intervention strategies in dementia care.

2.1 The Magic Number: Impact of Sample Size for Dementia Screening Using Transfer Learning and Data Augmentation of Clock Drawing Test Images

Abstract†

Dementia is a disease characterized by memory impairment and a gradual disability in performing daily activities. Automated screening for early detection of dementia can lead to more adequate and timely treatment. Our work focuses on predicting various stages of dementia severity using pre-trained Deep Learning (DL) models and a public Clock Drawing Test (CDT) dataset. However, the relationship between sample size and model performance is not yet well understood. This may lead to an overreliance on a large number of samples for model training, which may eventually deter reliable outcomes. We found that the classification performance of DL models tends to plateau once a certain number of samples is reached, therefore, it is possible to work on a small data regime with DL models in this task. This research not only advances the field of medical image analysis for dementia screening but also offers broader implications for DL applications in healthcare. Ultimately, the understanding of how sample size affects model performance can guide future research and support more intelligent and efficient utilization of DL models in addressing complex health-related challenges.

Intex Terms

Sample size estimation; Clock Drawing Test; Deep Learning; Alzheimer's disease

2.2 Introduction

Dementia is a progressive neurological disorder that causes memory loss and cognitive decline, critically impairing an individual's ability to perform daily activities. The prevalence of dementia is increasing, with Alzheimer's disease (AD) being identified as the most common form, accounting for the majority of cases [27]. Despite substantial research progress, a cure for dementia continues to remain out of reach, highlighting the need for research and development of innovative interventions [27]. The impact of this

†**N. Hosseini-Kivanani**, C. Schommer, and L. A. Leiva, "The Magic Number: Impact of Sample Size for Dementia Screening Using Transfer Learning and Data Augmentation of Clock Drawing Test Images," in *Proceedings of the 2023 IEEE International Conference on E-health Networking, Applications and Services (Healthcom'23)*, 2023, pp. 101–106. DOI: 10.1109/Healthcom56612.2023.10472399

disease extends beyond the affected individuals, including their families, caregivers, and healthcare systems. According to current estimates, the global cost of dementia reaches one trillion dollars annually and is expected to increase in the future [2].

Detecting dementia typically requires a variety of cognitive tests for neuropsychological impairment [21], among which the Clock Drawing Test (CDT) is a widely used tool. Specifically for the study of AD, the CDT has demonstrated a high diagnostic efficiency [23, 26], especially among the elderly [7, 22]. It assesses the cognitive health of patients through a simple yet revealing task: drawing a clock set to a specific time (usually ten minutes after eleven). The simplicity, non-invasiveness, and intuitiveness of this test make it an accessible tool for assessing cognitive health across diverse populations, including those with limited literacy or physical disabilities.

The evaluation of CDT images involves analyzing the quality of the drawings, specifically the positioning of the numbers and the clock hands. Clinicians use different scoring systems for their assessments, which help determine the severity or progression of dementia. In this paper, we rely on the well-established Shulman six-point scale [32] to classify CDT images with Deep Learning (DL) models.

Recently, significant advancements have been made with regard to the analysis and interpretation of CDT images. Jimenez-Mesa et al. [20] introduced a computer-aided diagnosis system based on DL for automated diagnosis. Similarly, [30] developed a deep neural network (DNN) model using 40000 CDT drawings. Their model achieved an accuracy of 90% in binary classification (impaired vs. control participants), and up to 77% accuracy in identifying individuals with probable dementia. Nevertheless, their research primarily centered on the binary classification problem. In another study by [39], a DNN-based prediction model was designed to detect cognitive decline, effectively distinguishing between cognitively impaired and control participants. All these previous works aimed at automating the scoring process for CDT images, specifically targeting their use for screening purposes, and mostly focusing on binary classification tasks.

A current limitation of the state of the art, as discussed in the next section, is the lack of systematic and comprehensive understanding of the optimal sample size required for DL models, particularly in the context of dementia diagnosis, and the absence of focused studies on the nuanced effects of varying sample sizes on model performance. This is important because it is often assumed that DL requires a large number of samples for model training. At the same time, collecting and labeling CDT images is time-consuming. Therefore, it would be a quite feat if DL models could provide reliable outcomes with small data.

This work concentrates on optimizing the sample size used for DL model training with CDT images. By exploring various sample sizes and their effects on model performance, we aim to improve the efficiency and effectiveness of diagnostic processes for dementia. Our findings shed light on the important topic of sample size in DL model performance, providing a practical roadmap for researchers, clinicians, and practitioners dealing with limited data availability.

2.3 Related Work

The growing body of research supports the potential of DL models for early and accurate dementia detection, facilitating more timely and effective treatment for patients [4]. For example, previous work has shown promise in the diagnosis of neurological disorders [24] and the prediction of early-stage dementia [5].

Recent research proposed the use of automatic scoring systems as alternatives to traditional manual evaluation techniques [3, 12]. Notably, Chen et al. [12] aimed to automate the CDT scoring process using DL models, reporting an accuracy of 96.65% for binary classification and 72.2% for multi-class classification based on the six-point Shulman scale [32], which categorizes drawings from perfect clock representations to those that are severely disorganized and unidentifiable as clocks (see Figure 2.2).

However, the process of collecting a large quantity of high-quality data for DL models is both time-consuming and expensive. This represents a significant challenge, especially in medical research where data collection involves strict ethical regulations and privacy concerns [35]. Nonetheless, the relationship between sample size and DL performance is poorly understood. While larger datasets can potentially lead to improved model performance, the results are not guaranteed. Therefore, determining an appropriate sample size is vital, as it significantly affects the robustness, reliability, and generalizability of a model's predictions.

Knowing the adequate amount of training data is essential [35], but few studies have systematically evaluated the impact of sample size on model accuracy [6]. Althnian et al. [1] showed that smaller sample sizes when combined with careful feature selection, can enhance the performance of machine learning (ML) classifiers. This lack of a clear rule emphasizes the need for further research on how the quantity of data influences the performance of ML models in the medical field. Our study aims to contribute to this ongoing topic, examining the influence of data size on the effectiveness of DL models in dementia diagnosis.

Several studies have investigated the impacts of dataset size on the classification performance in the medical domain [6, 34]. For example, Varma and Simon [36] used a dataset comprising only 40 samples and examined the performance of models using two different

Cross-validation (CV) approaches for data selection. Their study primarily focused on the choice of validation method. In contrast, Combrisson et al. [14] varied the sample size and used K-fold CV exclusively. Their study reported that with smaller sample sizes, classification accuracy was above the chance level 62.5% with the p-value $< .05$ (in a 2-class or 4-class classification problem). In another study in the medical domain, Althnian et al. [1] prepared three subsets of different sizes and employed a range of metrics to compare the performance of six classic ML models. They concluded that a set of 10 features and a smaller amount of data could enhance the performance of classifiers. Meanwhile, Han et al. [16] took a slightly different approach and investigated the optimal number of feature sets using a random forest classifier. They suggested that optimal data can vary from one dataset to another if no specific pattern is defined. Hence, to address this, they proposed using an out-of-bag error and 'SearchSize' exploration, leading to an improvement in accuracy.

Finally, various studies have tackled the challenge of small datasets by augmenting training sets; see e.g. [11]. However, most of this research has predominantly focused on increasing the data size, with little attention given to examining the impact of the sample size on performance. Mostly, existing research has concentrated on the extent to which the dataset size can affect the classification performance in different domains (e.g., [8, 40]).

In sum, this paper addresses a gap in the current research on using DL models in the clinical domain, in particular for dementia diagnosis. The paper investigates the optimal sample size required for DL models, exploring its impact on model performance. By clarifying the optimal data requirements, this study paves the way for more efficient dementia diagnostic processes, even in settings with limited data resources.

2.4 Materials and Methods

Our focus is to investigate whether using an optimal dataset size for DL model training can achieve similar, or even improved, classification performance levels for dementia screening, as compared to using the full dataset. Therefore, we sought to develop DL models capable of predicting different stages of dementia severity using an optimal sample size. To achieve this, our approach builds on and replicates the study conducted by Chen et al. [12], who previously demonstrated the efficacy of training DL model on a public CDT dataset.

Given the characteristics of the CDT dataset (small size and imbalanced classes), we used data augmentation techniques, which have been proven effective in generating new training samples, by applying various transformations to the original images [37] to see how the model performance with and without augmentation in different sample sizes will change. This approach is two-fold. Firstly, it helps balance the dataset by generating a range of inputs from which the model can learn. Secondly, it acts as a robust regularization technique, ensuring that the model generalizes well to new data and preventing overfitting [25, 31].

2.4.1 Dataset

The CDT dataset we used in this work has 1375 images. The participants' age varied from 18 to 98 years, with an average age of 69.8 years±14.7 years. Based on the Shulman scoring system, the images have been classified into six categories, each representing different stages of dementia severity. The scores ranged from 1 to 6, each indicative of increasing dementia severity, with 1 indicating low severity and 6 indicating high severity. A score of 0 represents healthy subjects (Figure 2.1).

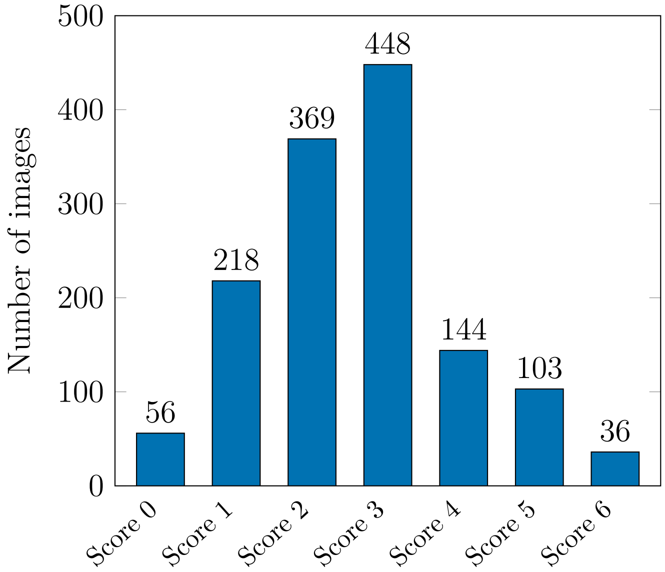
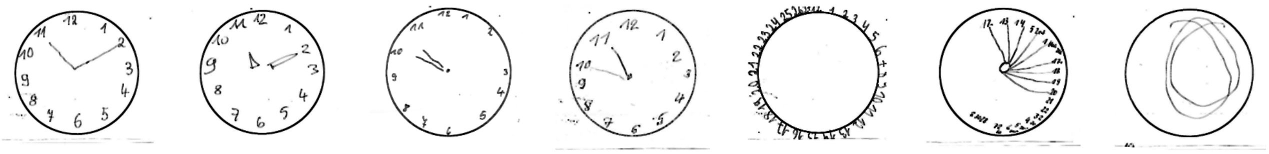


Figure 2.1: Total number of drawings for screening and scoring. Score 0 refers to healthy subjects. Score 6 refers to subjects who are unable to draw anything related to a clock.

During data acquisition, participants were presented with a piece of paper containing a pre-printed circle. They were then instructed to draw the clock numbers from 1 to 12 and to set the clock hands to point to "11:10 o'clock." As illustrated in Figure 2.2 (with a Shulman Score 0), the clock hands should be positioned on '11' and '2' to accurately represent this time.



(a) Score 0 (b) Score 1 (c) Score 2 (d) Score 3 (e) Score 4 (f) Score 5 (g) Score 6

Figure 2.2: Sample drawings of the CDT dataset that was used in this work, scored according to Shulman's scale.

2.4.1.1 Data pre-processing

The paper-and-pencil drawings of both patients and healthy participants were scanned in 256-bit grayscale PNG format at 849×1168 px, which were resized to 224×224 px, according to the expected input size of our pre-trained DL models, as explained in the next section. We manually revised all images and removed 20 low-quality images, mostly due to bad scanning and noisy images.

As mentioned before, the dataset is highly imbalanced and not very large for today's standards in DL, so we use data augmentation to address this issue in our study. However, note that not all data augmentations apply in our case. For example, mirroring CDT images would destroy the semantics of the drawings. Similarly, changing the hue or saturation has no effect on those images since they are grayscaled. Therefore, we applied the following operations: scaling, rotation, and translation. These operations produce new, transformed images that help to increase the size and diversity of the training data without compromising its clinical validity [18]. The resulting dataset was perfectly balanced, comprising 448 CDT images per class, or 3136 images overall.

Further, to quantify the quality of the augmented data, we computed the structural similarity index measure (SSIM) [38] of all augmented images against the original images. As shown in Figure 2.3, the SSIM values are overall between 0.65 to 0.85 with the highest frequency range from 8 to 10, which indicates that the augmented images are not near-duplicates of the original data. Rather, they are new images that, as shown later, eventually helped to improve model performance.

2.4.2 Deep learning models

We employed transfer learning to leverage the power of pre-trained DL models in our experiments. Transfer learning is an approach that enables the use of neural network models that have been previously trained on a representative dataset, such as ImageNet [15], to be used as a starting point for solving a related problem by fine-tuning the model on a new dataset [28]. This method helps reduce the need for large computational resources

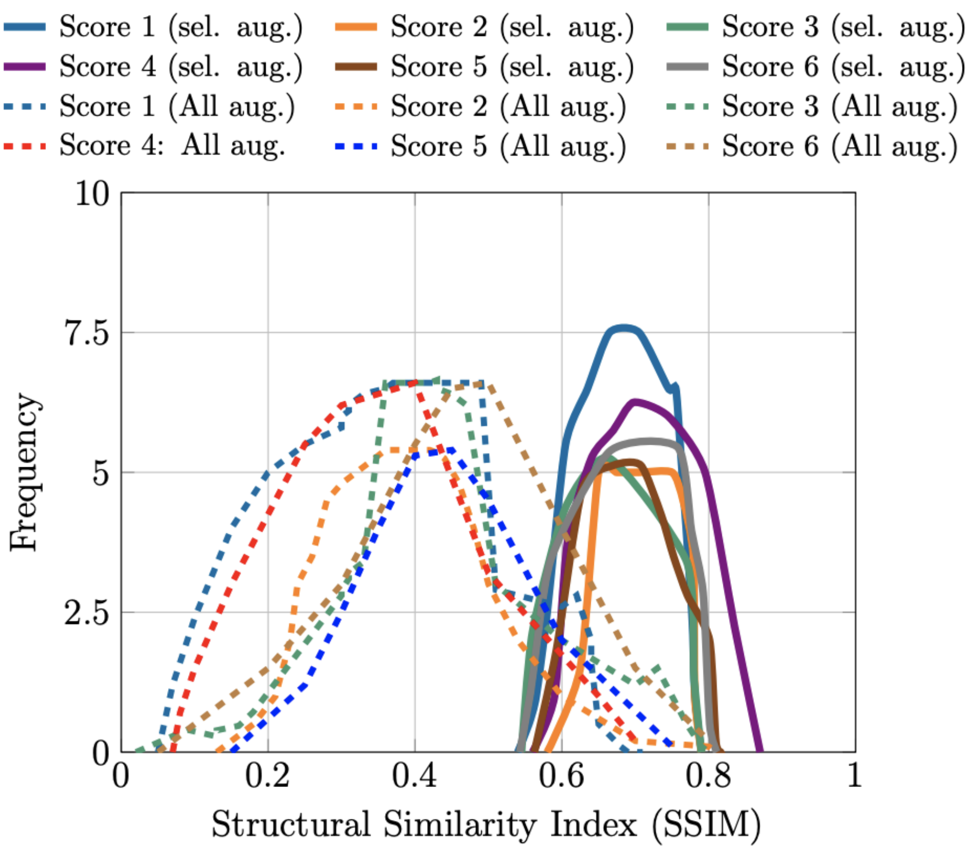


Figure 2.3: SSIM distributions. Dashed plots correspond to the results considering all the augmentation (All aug.) techniques collectively. The selected augmentation techniques are Rotation, Scaling, and Translation offset.

and extensive labeled data, yet still allows us to achieve high performance on the target task.

We fine-tuned three Convolutional Neural Networks (CNNs) for binary (healthy vs. non-healthy) and multi-class (six Shulman scores) classification tasks. Binary classification provides a fundamental understanding of the model's capability to differentiate between normal and abnormal cognitive functioning. Conversely, the multi-class classification task aims to distinguish among the six stages of dementia severity according to Shulman's scale, which enables a more detailed understanding of dementia progression, as reflected in the CDT drawings. In the following, we delve into the specifics of the CNN architectures used:

- VGG-16 [33]: It comprises 16 CNN layers with a 3x3 kernel and three subsequent Fully Connected (FC) layers, was designed by the Visual Geometry Group (VGG) at Oxford University. It has garnered recognition due to its straightforward architecture and efficiency in extracting features.

- ResNet-152 [17]: It comprises 152 CNN layers, providing flexibility and a smaller parameter count compared to models like VGG. It effectively minimizes the error rate to 3.5% and owes its remarkable performance to the use of skip connections.

- DenseNet-121 [19]: It comprises 121 CNN layers, ensuring that every layer has a direct connection to the outputs of all the layers preceding it. It also comprises DenseBlocks interconnected by transition layers.

2.5 Sample size analysis

The size of a dataset plays a central role in ML, enabling the effective training and testing of predictive models. A frequent question that often arises is how much data is sufficient or required for model training, and this remains an open challenge [9]. The answer to this question is not straightforward, as it involves finding a balance influenced by various factors. These include the complexity of the task, the diversity present within the data, and the sophistication of the chosen model. Further, small-scale studies carry a higher likelihood of committing either type I or II errors, thereby reducing the probability of identifying true effects [10].

Therefore, it is critical to identify and apply strategies that can reduce data requirements without excessively compromising the model performance. How to effectively use smaller datasets is especially beneficial for researchers with restricted access to large datasets. We aim to demonstrate how these strategies can be effectively applied to deal with the limitations imposed by smaller datasets while ensuring robust model performance. We will report the usual model performance metrics to assess the efficacy of our approach: classification accuracy and area under the Receiver Operating Characteristic curve AUC. These metrics will provide a comprehensive understanding of the model's capacity to correctly classify and distinguish between different stages of dementia severity.

2.5.1 Learning Curves

A learning curve (LC) is a graphical representation that illustrates the performance of a model over time [29]. The LC serves beyond merely visualizing the current performance of the model. It can also be used as a predictive tool. Once the LC has been established, we can extrapolate to estimate the accuracy of the model if it is to be trained on the entirety of the available training data. This allows us to predict the potential performance using additional data.

In this work, we generated LCs for accuracy and AUC to visually assess the model's improvement and to identify any potential areas where performance may decline. This can help researchers to use data wisely, e.g. to decide whether to stop or continue model training based on the observed performance over different data splits.

2.5.2 Model Training

We selected training subsets corresponding to splits of 10%, 25%, 50%, 75%, and 95% of the entire dataset for our study. A training split of 10% was fixed in each case, except for the 95% training split where the test split was set to 5%. For each of the selected subsample splits, the model was trained from scratch. We repeat this procedure five times, using different initialization seeds for each split to verify the validity of our results and consider any variability that may occur during individual training iterations.

To ensure consistency, we used the experimental setup outlined in the study conducted by Chen et al. [13]. Our CNN models were trained using the Adam optimizer algorithm, with learning rate values ranging from 0.0001 to 0.1, while maintaining a fixed batch size of 16. Additionally, the cross-entropy loss function was used in the training process to measure the performance of a classification model.

Additionally, the recorded performance metrics facilitated the plotting of the LC in the subsequent stages of our experiment. This analysis enables us to effectively analyze, interpret, and optimize the performance of our models. This also helps us understand the trade-off between computational cost and model performance, thereby allowing us to maximize the efficiency of our model given the available data.

2.6 Results and Discussion

Table 2.1: Best and second-best Accuracy and AUC results overall, relative to the total size of the CDT dataset.

	Binary Classification		**Multi-class Classification**	
Classifier	Accuracy @ Sample size		Accuracy @ Sample size	
VGG-16	0.97 @ 100%	0.95 @ 95%	0.68 @ 100%	0.68 @ 95%
ResNet-152	0.97 @ 100%	0.88 @ 95%	0.71 @ 100%	0.71 @ 95%
DenseNet-121	0.98 @ 100%	0.98 @ 95%	0.77 @ 100%	0.77 @ 95%

The results for both binary and multi-class classification can be found in Table 2.1. DenseNet-121 performed the best, with almost 98% accuracy for binary classification, even with the small sample size of 1585 images (50% of the data, after data augmentation). For multi-class classification experiments, ResNet-152 achieved an accuracy of 88%

for binary classification using 95% of the data. These results suggest that these models can provide competitive performance when fine-tuned on small data samples.

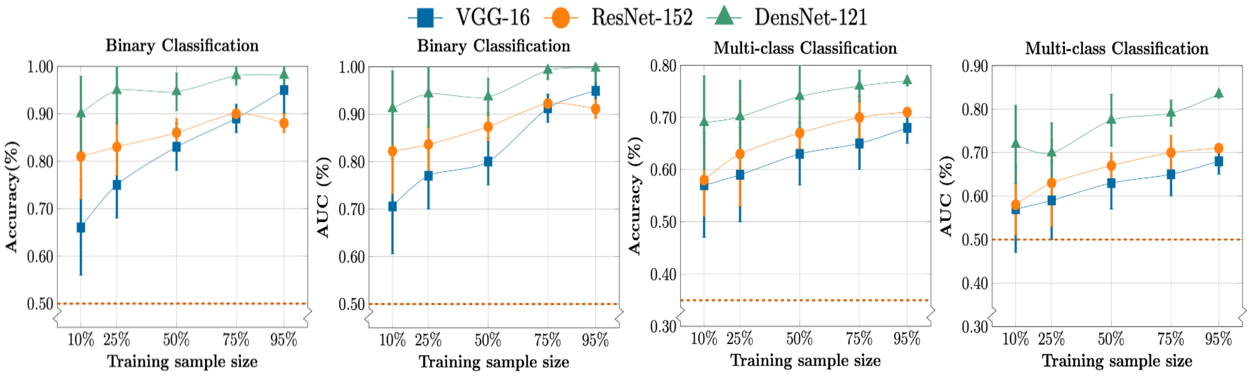


Figure 2.4: LC for binary classification (two leftmost plots), and multi-class classification (two rightmost plots), showing model accuracy and AUC results vs. training sample size. The dashed orange lines represent the performance of a random classifier.

The LCs in Figure 2.4 are aligned with common trends observed in previous studies (e.g., [14]), where a swift performance enhancement was noticeable with the expansion of the training set size. For sample sizes larger than 50%, the performance of the DenseNet model stabilized and reached the second-highest accuracy when using the whole dataset (Table 2.1).

In terms of binary classification accuracy, our analysis did not reveal any statistically significant differences when using DenseNet and 25% of the dataset compared to using more data. This was confirmed by a chi-squared test of proportions ($\chi^2(4, N = 417) = 13.13, p > .05$). Similarly, for multi-class classification accuracy, no significant differences were noted when employing 50% of the dataset compared to higher data splits ($\chi^2(4, N = 834) = 2.28, p > .05$). These findings suggest that using all the available data may not always be crucial to obtain the best performance results. For architectures like VGG and ResNet, a minimum of 75% of the dataset seems necessary to attain peak performance.

In contrast to previous studies, that employed ML models using all the available data, our experiments examined the role of sample size and model performance across varying sample sizes. In a nutshell, if the performance of a classifier is good enough with only a subset of the dataset, then such a classifier can be used in conditions where limited data is present, which is quite frequent in clinical domains.

Although the emphasis in previous research has been on how increasing data size improves the ability of CNNs (e.g., [9]), there is a big gap in studies regarding the optimized sample sizes. Our results showed that with 75% of the data we reached to the highest accuracy in all CNN models but DenseNet required only 50% of the data to achieve statistically similar results as when using all the data. Until now, it was expected that adding more labeled data would improve model performance, but we have shown that we do not need as much data.

The advantage of applying DL on a small dataset is quite apparent, given that CNNs are highly data-dependent and usually necessitate more computational costs than traditional ML models. Although our findings are specific to a public CDT dataset and a handful of state-of-the-art CNN models used in a previous study (i.e. [13]), we believe that our results provide valuable insights into the understanding of selecting the optimum sample size for the development of improved DL models.

2.7 Conclusion and future work

We have investigated the impact of sample size and the performance of DL models (state-of-the-art CNN classifiers). Our findings indicate that classification accuracy and AUC tend to plateau once a certain number of samples are reached. Therefore, we do not really need to use all the available data for training. For multi-class classification, results suggest we may need a larger portion of the dataset, even though the results shown in Tables 2.1 report very similar figures. Future research should explore other datasets together with other DL architectures. This way, further improvements could be achieved in terms of predictive capabilities for the early detection of dementia.

References for Paper 1

[1] Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou Elwafa, and Heba Kurdi. 2021. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Applied Sciences*, 11, 2, 796.

[2] Alzheimer's Association. 2022. 2022 Alzheimer's disease facts and figures. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 18, 4, (Apr. 2022), 700–789. DOI: `10.1002/alz.12638`.

[3] Samad Amini et al. 2021. An Artificial Intelligence-Assisted Method for Dementia Detection Using Images from the Clock Drawing Test. *Journal of Alzheimer's Disease*, 83, 2, 581–589. Publisher: IOS Press BV. DOI: `10.3233/JAD-210299`.

[4] Liana G Apostolova and Paul M Thompson. 2008. Mapping progressive brain structural changes in early alzheimer's disease and mild cognitive impairment. *Neuropsychologia*, 46, 6, 1597–1612.

[5] Damaris Aschwanden, Stephen Aichele, Paolo Ghisletta, Antonio Terracciano, Matthias Kliegel, Angelina R Sutin, Justin Brown, and Mathias Allemand. 2020. Predicting cognitive impairment and dementia: a machine learning approach. *Journal of Alzheimer's Disease*, 75, 3, 717–728.

[6] Indranil Balki et al. 2019. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Canadian Association of Radiologists Journal*, 70, 4, 344–353.

[7] Sabyasachi Bandyopadhyay, Jack Wittmayer, David J. Libon, Patrick Tighe, Catherine Price, and Parisa Rashidi. 2023. Explainable semi-supervised deep learning shows that dementia is associated with small, avocado-shaped clocks with irregularly placed hands. en. *Scientific Reports*, 13, 1, (May 2023), 7384. DOI: `10.1038/s41598-023-34518-9`.

[8] Jayme Garcia Arnal Barbedo. 2018. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Computers and electronics in agriculture*, 153, 46–53.

[9] Saul Calderon-Ramirez, Diego Murillo-Hernandez, Kevin Rojas-Salazar, David Elizondo, Shengxiang Yang, Armaghan Moemeni, and Miguel Molina-Cabello. 2022. A real use case of semi-supervised learning for mammogram classification in a local clinic of costa rica. *Medical & biological engineering & computing*, 60, 4, 1159–1175.

[10] Evan C Carter, Felix D Schönbrodt, Will M Gervais, and Joseph Hilgard. 2019. Correcting for bias in psychology: a comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2, 2, 115–144.

[11] Hung-Yu Chen, Der-Chiang Li, and Liang-Sian Lin. 2016. Extending sample information for small data set prediction. In *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 710–714.

[12] Shuqing Chen, Daniel Stromer, Harb Alnasser Alabdalrahim, Stefan Schwab, Markus Weih, and Andreas Maier. 2020. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports 2020 10:1*, 10, 1, (Nov. 2020), 1–11. Publisher: Nature Publishing Group.

[13] Shuqing Chen, Daniel Stromer, Harb Alnasser Alabdalrahim, Stefan Schwab, Markus Weih, and Andreas Maier. 2020. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports*, 10, 1, 1–11.

[14] Etienne Combrisson and Karim Jerbi. 2015. Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of neuroscience methods*, 250, 126–136.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: a large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[16] Sunwoo Han and Hyunjoong Kim. 2019. On the optimal size of candidate feature set in random forest. *Applied Sciences*, 9, 5, 898.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[18] Nina Hosseini-Kivanani et al. 2023. Better together: combining different handwriting input sources improves dementia screening. In *Proc. eScience: AI4Health*. IEEE.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

[20] Carmen Jimenez-Mesa et al. 2022. Automatic classification system for diagnosis of cognitive impairment based on the clock-drawing test. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 34–42.

[21] Edith Kaplan. 1990. The process approach to neuropsychological assessment of psychiatric patients. *The Journal of Neuropsychiatry and Clinical Neurosciences*.

[22] Sangsoon Kim, Seungmin Jahng, Kyung-Ho Yu, Byung-Chul Lee, and Yeonwook Kang. 2018. Usefulness of the Clock Drawing Test as a Cognitive Screening Instrument for Mild Cognitive Impairment and Mild Dementia: an Evaluation Using Three Scoring Systems. *Dementia and Neurocognitive Disorders*, 17, 3, (Dec. 2018), 100–109. DOI: 10.12779/dnd.2018.17.3.100.

[23] Kang Soo Lee, Eun A Kim, Chang Hyung Hong, Dong-Woo Lee, Byoung Hoon Oh, and Hae-Kwan Cheong. 2008. Clock drawing test in mild cognitive impairment: quantitative analysis of four scoring methods and qualitative analysis. *Dementia and Geriatric Cognitive Disorders*, 26, 6, 483–489.

[24] Alexander Selvikvåg Lundervold and Arvid Lundervold. 2019. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29, 2, 102–127.

[25] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. 2022. A review: data pre-processing and data augmentation techniques. *Global Transitions Proceedings*.

[26] Brian J Mainland and Kenneth I Shulman. 2017. Clock drawing test. *Cognitive screening instruments: A practical approach*, 67–108.

[27] M Monica Moore, Mirella Díaz-Santos, and Keith Vossel. 2021. Alzheimer's association 2021 facts and figures report. *Alzheimer's Association*.

[28] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22, 10, 1345–1359.

[29] Claudia Perlich. 2010. Learning curves in machine learning. In *Encyclopedia of Machine Learning*.

[30] Kenichiro Sato, Yoshiki Niimi, Tatsuo Mano, Atsushi Iwata, and Takeshi Iwatsubo. 2022. Automated evaluation of conventional clock-drawing test using deep neural network: potential as a mass screening tool to detect individuals with cognitive decline. *Frontiers in neurology*, 13, 896403.

[31] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6, 1, 1–48.

[32] Kenneth I Shulman, Dolores Pushkar Gold, Carole A Cohen, and Carla A Zucchero. 1993. Clock-drawing and dementia in the community: a longitudinal study. *International journal of geriatric psychiatry*, 8, 6, 487–496.

[33] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[34] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J Casson. 2019. Machine learning algorithm validation with a limited sample size. *PloS one*, 14, 11, e0224365.

[35] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J. Casson. 2019. Machine learning algorithm validation with a limited sample size. en. *PLOS ONE*, 14, 11, (Nov. 2019), e0224365. DOI: 10.1371/journal.pone.0224365.

[36] Sudhir Varma and Richard Simon. 2006. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7, 1, 1–8.

[37] Jason Wang, Luis Perez, et al. 2017. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11, 2017, 1–8.

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13, 4, 600–612.

[39] Young Chul Youn et al. 2021. Use of the clock drawing test and the rey–osterrieth complex figure test-copy with convolutional neural networks to predict cognitive impairment. *Alzheimer's Research & Therapy*, 13, 1, 1–7.

[40] Xiangxin Zhu, Carl Vondrick, Charless C Fowlkes, and Deva Ramanan. 2016. Do we need more training data? *International Journal of Computer Vision*, 119, 1, 76–92.

2.8 Better Together: Combining Different Handwriting Input Sources Improves Dementia Screening

Abstract†

Alzheimer's disease (AD) is a cognitive disorder, marked by memory loss and impaired reasoning, that requires early detection methods to better manage and potentially slow down the disease's progression. Recent advances in machine learning have offered new possibilities for AD detection using handwriting analysis, however previous work has considered only one type of input source, e.g. clock or pentagon drawings. Here we propose to develop an efficient method for detecting AD's early symptoms using Deep Feature Concatenation (DFC) models considering multiple handwriting sources: pentagon drawings, self-reported sentences, and signatures. Substantial performance improvements were observed when considering all input sources together with data augmentation techniques. For example, classification accuracy increased from 60% (best model, without data augmentation) to 80% (DFC and data augmentation). Our findings show that the use of diverse input sources can lead to an efficient and cost-effective method for early AD detection. Looking forward into the future, our study highlights the potential of DFC in supporting home-based healthcare diagnoses which is a crucial step in integrating artificial intelligence into healthcare practices.

Intex Terms

Deep Learning; Convolutional Neural Networks; deep feature concatenation; dementia; image processing.

2.9 Introduction and Related work

Dementia is a condition primarily characterized by a gradual and irreversible decline in cognitive ability, which results in memory loss and language impairment that negatively affect the daily lives of the elderly [8]. The loss of neurons in various regions of the nervous system causes this neurodegenerative disorder, among which Alzheimer's disease (AD) is the most prevalent form.

† **N. Hosseini-Kivanani**, E. Salobrar-García, L. Elvira-Hurtado, I. López-Cuenca, R. de Hoz, J. M. Ramírez, P. Gil, M. Salas, C. Schommer, and L. A. Leiva, "Better together: Combining different handwriting input sources improves dementia screening," in *Proceedings of the 2023 IEEE 19th International Conference on e-Science (eScience'23)*, 2023, pp. 1–7. DOI: 10.1109/eScience58273.2023.10254799

Since currently there is no known cure for AD, the importance of an accurate and timely diagnosis of dementia cannot be overstated, as it is the cornerstone for implementing effective treatment and providing necessary support for patients and their families. Traditional diagnostic methods, on the other hand, can be subjective and time-consuming, since they require multiple tests, including expensive imaging techniques such as magnetic resonance imaging, invasive tests such as serological or cephalorachidian fluid analysis, and neuropsychological tests by a highly trained professional.

In light of these challenges, the development of automated, objective screening methods for dementia is a priority in the quest for efficient and precise diagnosis of dementia. Notably, the most beneficial methods would be non-invasive and user-friendly in order to minimize any additional burden on the individuals undergoing the diagnosis.

In digital medicine, there is growing interest in using DL models, in particular Convolutional Neural Networks (CNNs), to automatically score cognitive impairment tests instead of traditional manual methods. More concretely, transfer learning has been shown to enhance the performance of pre-trained CNNs in image classification tasks [21, 18], thus suggesting its potential for clinical applications.

Recent studies [12, 11, 26] have demonstrated the effectiveness of CNNs in analyzing patients' cognitive function assessment over various popular tests, including e.g. the PDT [13], the Clock Drawing Test (CDT) [1, 5, 3], and the Rey–Osterrieth Complex Figure Test-copy Rey–Osterrieth Complex Figure Test-copy (RCFT) [26]. Recent work proposed a CNN-based method for the automatic diagnosis of cognitive impairment based on CDT drawings [11] that were classified into healthy and non-healthy categories, demonstrating its potential for implementation in hospitals and clinics, particularly in resource-limited settings.

The ability of CNNs to learn complex image features and patterns associated with cognitive functioning makes these algorithms promising tools for developing objective and efficient diagnostic methods for dementia screening. Despite the drawing tests' proven ability to accurately predict cognitive impairment, they are not sufficient on their own to provide a comprehensive assessment of the user's overall cognitive state.

Despite the substantial progress made in this field on single input sources (e.g. CDT [11, 5] and PDT [12, 13]), to our knowledge, concatenation-based models (i.e. models considering multiple input sources) have not been considered so far in AD screening contexts. Concatenation-based models have primarily been used in clinical work that involves the analysis of medical images such as Functional magnetic resonance imaging (fMRI) [16]. However, handling multiple input sources present unique challenges due to their variability in terms of styles, shapes, etc. Additionally, they may contain irrelevant or noisy information that could affect classification accuracy. It is therefore essential to develop

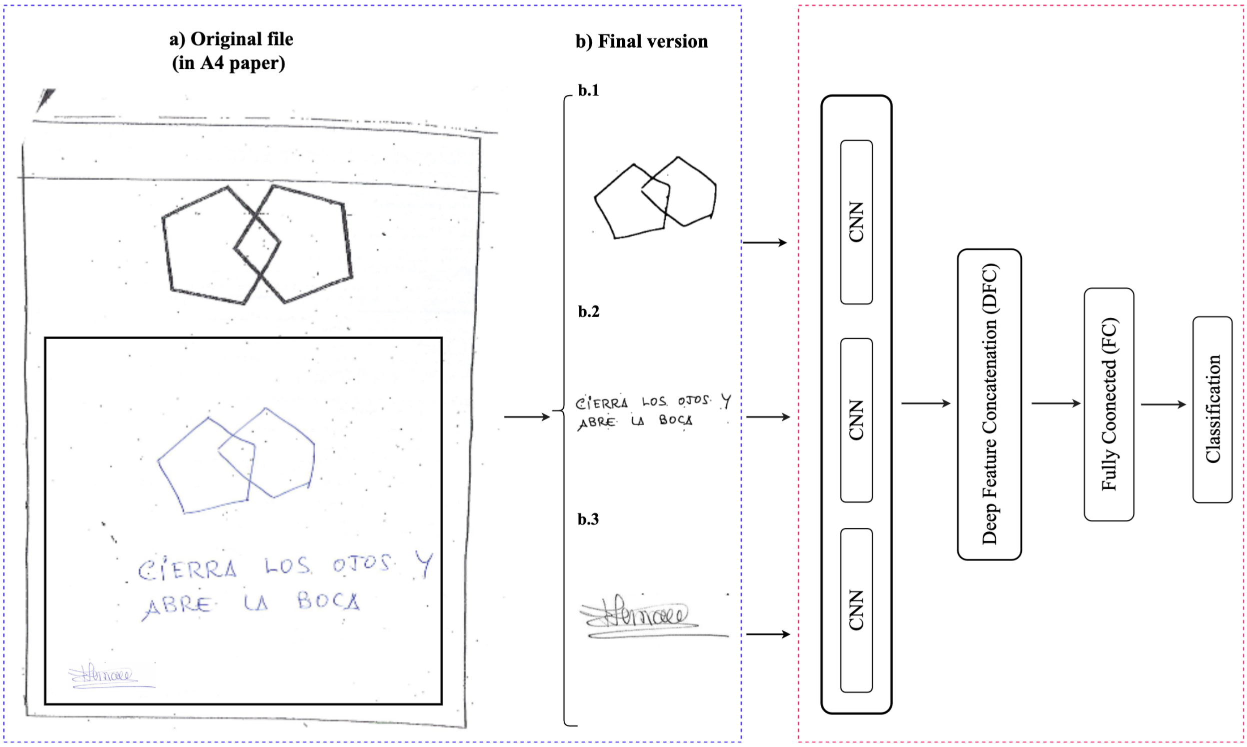


Figure 2.5: Final output of image preprocessing which shows in the blue box (left side): a) prompting pentagon on top with participants' drawings at the bottom; b) final images after preprocessing. Structure of the proposed method on the red box (right side).

methods for effectively extracting and combining relevant features from these images in order to ensure accurate classification.

In response to this need, we propose a deep feature concatenation (DFC) method which enables the efficient combination of features derived from various CNN models. Additionally, since clinical datasets are typically very small for DL standards, we propose simple data augmentation techniques to handle various handwriting data sources, and the use of the Structural Similarity Index Measure (SSIM) [25] to quantify the quality of the augmented data. Our experiments show significant improvements in classification performance in several scenarios, as presented and discussed later. Taken together, our contributions represent a potential pathway to assist practitioners in better detecting early symptoms of AD using handwriting data, potentially reducing the variability linked to human subjectivity when interpreting clinical data, and ultimately supporting home-based healthcare diagnosis.

2.10 Method

We studied state-of-the-art DL models (pre-trained CNNs and a custom CNN) for classifying AD disease according to various handwriting tasks. Building on previous work that showed the effectiveness of DFC in other medical domains [24, 15], we studied this approach in the context of handwriting. Figure 2.5 summarizes our method. In a nutshell, we use various CNN models to automatically extract features from handwritten images and then concatenate those features via DFC. The concatenated feature vectors are finally classified via a softmax function that predicts a probability distribution, followed by the argmax operation that selects the class with the highest probability.

2.10.1 CNN architectures

CNNs have been remarkably successful across a range of image-based classification tasks, particularly in the healthcare domain(e.g., [6]). Transfer learning, particularly *feature transfer* wherein features are extracted from pre-trained models, is a common approach for adopting these models for specific tasks [27]. For our study, we selected four pre-trained models that have proven their efficacy in the medical domain [17]. Further, to broaden the versatility of our approach, we also added a custom CNN model that we describe below.

- **VGG-16** [20]: Developed by the Visual Geometry Group (VGG) in Oxford, VGG-16 is a deep neural network comprising 16 CNN layers with a 3x3 kernel, followed by three fully connected (FC) layers. Noted for its simplicity and effectiveness in feature extraction.

- **ResNet-152** [9]: This is a deep architecture, with 152 CNN layers. It achieves an error rate of 3.5% and uses skip connections between CNN layers to achieve its excellent performance [9].

- **DenseNet-121** [10]: This is a 121-layer deep CNN model that uses a "dense" connectivity pattern between layers. This configuration allows each layer to have direct access to the output of all preceding layers. The architecture includes CNN layers with 7x7 kernels and DenseBlocks containing interconnected CNN layers with 1x1 and 3x3 kernels.

- **EfficientNet** [22]: This is a CNN architecture that employs a scaling technique to harmonize depth, width, and resolution, leading to improved performance and efficiency. With its ability to outperform with fewer parameters, EfficientNet is frequently chosen for applications with limited computational resources [22].

- **Custom CNN**: It comprises five CNN layers, each with 32 filters of size 5, followed by a pooling layer, an FC layer, and Rectified Linear Unit (ReLU) layers for all layers except the output layer, which uses linear activation. To avoid overfitting, a dropout rate of 0.1 was implemented after each CNN layer. We explored various model configurations during our research, however, most of them did not show satisfactory results. For example, removing one or two layers did not result in better performance, it even led to overfitting issues, which we resolved by introducing the Dropout layers.

2.11 Experiments

2.11.1 Participants

We recruited 85 participants from the Memory Unit of the Hospital Clinico San Carlos (HCSC) in Madrid. The participants' ages ranged from 61 to 88 years, with a mean of 73.92 ± 6.78 years. Statistical analysis revealed no significant differences in age between the healthy group, the mild AD group, and the moderate AD group ($p > .05$). The study included 35 female and 50 male participants.

Table 2.2: Demographics of Study Participants (mean±SD).

Attr.	**Healthy**	**Mild AD**	**Moderate AD**
Pentagon	30	3	3
Sentence	30	3	3
Signature	29	22	15
Total	**89**	**28**	**21**
Age	72.4 ± 6.07	81.33 ± 4.62	81.66 ± 2.32
MMSE	28.17 ± 2.15	23 ± 2.65	19 ± 2.01
Gender	Female = 35, Male = 50		

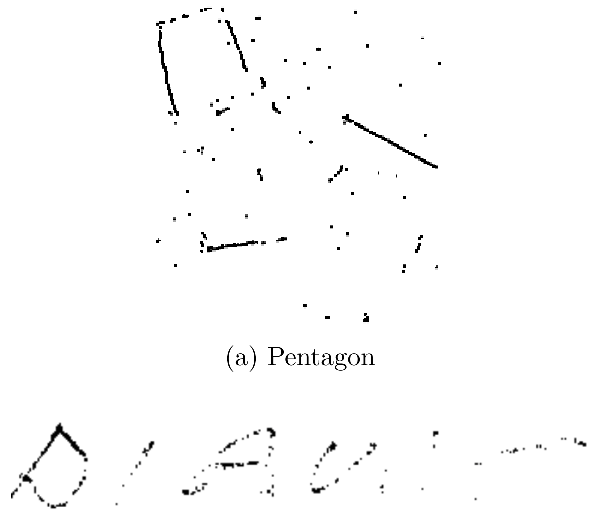
Following the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS), the Alzheimer's Disease and Related Disorders Association (ADRDA) workgroup [14], and the Statistical Manual of Mental Disorders V (DSM V) guidelines [7], cognitively impaired patients were divided into two groups: mild AD and moderate AD.

The demographic information, which included age and gender, along with the clinical information for each participant and Mini-Mental State Examination (MMSE) [23] scores, is presented in Table 2.2.

Participants with MMSE scores above 26 were considered cognitively healthy. The MMSE scores for participants diagnosed with cognitive impairment ranged from 25 to 17. Following standard practice [19], we excluded participants with a medical history of neurological or psychiatric disorders, serious medical conditions, or systemic disorders affecting vision. Additionally, we excluded those with ophthalmological conditions such as glaucoma or suspected glaucoma, media opacity, or retinal diseases to avoid biases caused by the vision problems of the participants.

2.11.2 Datasets

During the cognitive assessment tests, neuropsychologists used various tools to evaluate the cognitive performance of the participants. Among these, we rely on the PDT subtest of the MMSE results, which assessed visuoconstructional skills and cognitive impairment. The total number of collected images is 138, of which 30 images from the pentagon and sentence parts were removed from the healthy group because of the very low quality of the scanned images (see Figure 2.6).



(a) Pentagon

(b) Sentence

Figure 2.6: Example of badly scanned images.

Following the PDT protocol, participants were instructed to copy two overlapping pentagons with interlocking shapes to form a rhombus. In order to gain a more holistic understanding of the participants' cognitive abilities, the cognitive assessment incorporated further handwriting data (e.g., sentence and signature) from the same participant. Below each pentagon drawing, participants were requested to write a sentence of their choice and provide a signature on an A4-size blank paper (see Figure 2.5a).

The paper-and-pencil drawings of both healthy subjects and patients were scanned in PDF format and saved as PNG files. Subsequently, we converted the PNG images to grayscale format, resized them to the standard dimensions of 224×224 pixels, and removed noise using the "pad" (to get them in the same shape) and "canny" (for edge detection of images) operations as provided by the OpenCV library (see Figure 2.5b). We then removed low-quality images by manual inspection and the remaining images were labeled as either 'healthy', 'mild AD', or 'moderate AD'. The last two categories correspond to AD patients, which allows us to conduct two kinds of experiments: *binary* (patient vs healthy) and *multiclass* (healthy vs mild AD vs moderate AD) classification tasks.

One of the challenges in clinical settings is the lack of large datasets, which may potentially hinder the success of DL models. To address this limitation, we used the Albumentations open-source toolkit [4] to augment the handwritten images for model training and evaluation. We applied the following data augmentation techniques: elastic transformations, grid distortions, horizontal flipping, translation, and rotations to the image (see Figure 2.7). Note that not all data augmentations make sense in our data, given the nature of our grayscale handwritten images. For example, changing hue or inverting colors would do more harm than good.

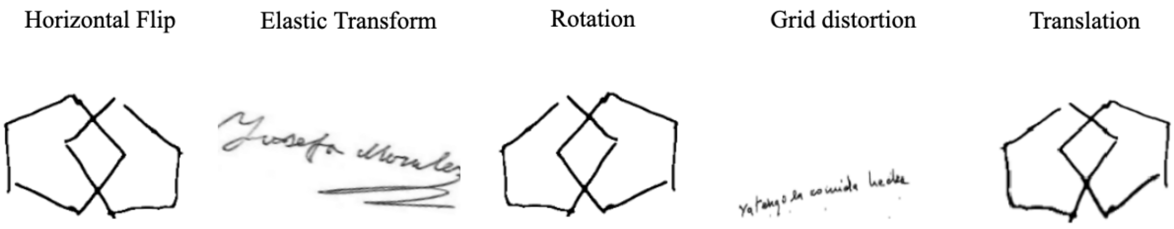


Figure 2.7: Examples of applying selected data augmentation techniques on different input sources. From left to right: Horizontal Flip, Elastic transform, Rotation, Grid distortion, and Translation offset.

The resulting dataset, after data augmentation, was partitioned into the train, validation, and test splits using stratified sampling, to ensure that each partition reflects the same data distribution as in the original dataset. The total size of our dataset increased to 240 images: 120 corresponding to healthy individuals and 120 to patients, among whom 66 represented mild AD and 54 moderate AD.

We used SSIM [25] to measure the quality between original and augmented images. This method uses sliding windows to compare structural distortions between two images. An SSIM value of 1 indicates that the two images are identical, whereas a value of 0 indicates that the two images are completely dissimilar. In our case, a high SSIM value between the original and augmented images implies that the augmentation process has successfully preserved the original images' structural information and visual quality. In contrast, a low SSIM value would suggest that the augmented image differs significantly from the original image and may not be of sufficient quality for the intended use.

As shown in Figure 2.8, our results reported SSIM values between 0.6 and 0.8, suggesting that the augmented images were reasonable variants (not near-duplicates) of the original data. In contrast, when using all available augmentation techniques provided by the Albumentations toolkit, the distribution of SSIM values is comprised of values between 0.1 and 0.7, indicating that the augmented images are much more different than the original images, which is not desirable in our research.

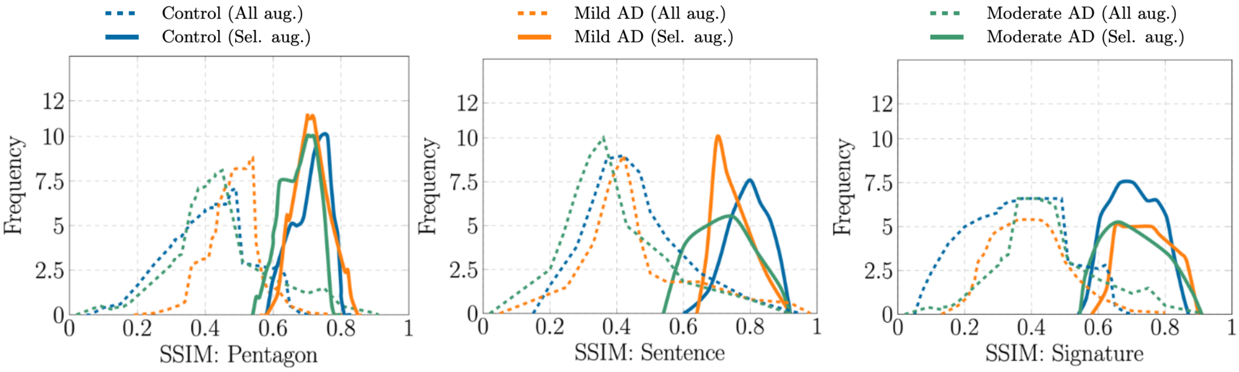


Figure 2.8: SSIM distributions. Dashed plots correspond to the results considering all the augmentation techniques collectively (All aug.). The selected augmentation (Sel. aug.) techniques are Elastic transform (`e`), Rotation (`r`), Grid distortion (`g`), Horizontal flip (`f`), and Translation offset (`t`). We used `e+g+r+f+t` for Pentagon, `e+g+t` for Sentence, and `e+t` for Signature.

As hinted previously, it is important to note that not all data augmentation techniques will provide advantages. Each technique has its own strengths and weaknesses, and its utility often depends on the specific dataset and the task at hand. As illustrated in Figure 2.8 some augmentation methods produce images that correlate better with the original images. This usually enhances the learning process and potentially leads to more accurate and generalizable models [2]. Conversely, other methods might introduce noise or misleading patterns into the data, confounding the learning process and potentially leading to poorer performance. This shows that the chosen augmentation positively influences model learning and performance, while simultaneously reducing the likelihood of adverse or neutral impacts.

2.11.3 Model training

All models were trained on the training set for 100 epochs, using early stopping with 10 epochs as a form of regularization. Early stopping prevents overfitting, maintaining the optimal model weights before the model starts memorizing the training data. We used the popular Adam optimizer with variable learning rates (see next section) and categorical cross-entropy as a loss function in multiclass classification experiments. For binary classification experiments, we used the binary cross-entropy loss function.

2.11.4 Evaluation

The performance of our models was measured using two metrics: Accuracy and area under the receiver operating characteristic curve (AUC). On the one hand, Accuracy refers to the ratio of correctly classified patients to the overall number of participants. It serves as a straightforward measure of the overall performance of the classification models. On the other hand, AUC illustrates the relationship between sensitivity (the true positive rate) and specificity (the true negative rate) for any given classification model. High AUC values indicate that the model possesses a strong discriminatory capacity between classes (e.g., healthy subjects vs patients).

Continuing our discussion about the selected metrics, it is essential to note that each of them brings forth unique insights into the performance of our models. While Accuracy offers a general view of the model's capability to differentiate between classes, AUC provides a deeper understanding of the model's reliability across various classification thresholds, proving invaluable in scenarios where the cost of misclassification can be substantial. Hence, the integration of these two metrics serves as a comprehensive approach to evaluating and interpreting the performance of our models in a robust manner.

2.12 Results and Discussion

We discuss the experimental results for both binary and multiclass classification settings. The results are based on CNN models that were trained with the following hyperparameters: number of epochs (100), batch size (16), dropout (0.1), weight decay (0.01), and learning rate (0.0001 to 0.1).

Figure 2.9 and Table 2.3 show the results for different model combinations. We use the following nomenclature: "model A + model B" to indicate that "model A" was used to

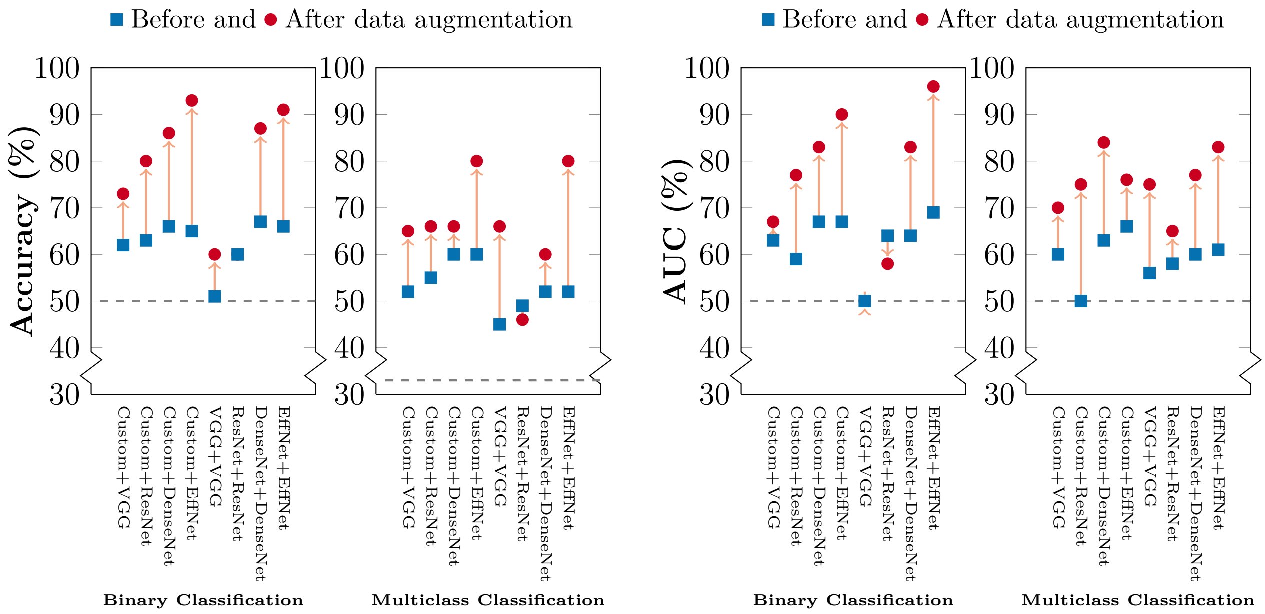


Figure 2.9: Accuracy and AUC results before and after data augmentation considering all input sources (pentagon, sentence, and signature), both for binary (2 classes, leftmost plot) and multiclass (3 classes, rightmost plot) classification experiments. Dashed lines denote the performance of a random classifier, as a way to illustrate the empirical lower bound in classification performance.

process Pentagon images and that "model B" was used to process both sentences and signatures. As can be observed, it is clear that our DFC approach of concatenating all input sources is significantly superior to any model that considers fewer input sources; see Table 2.3.

Overall, most of our DFC models are considered to improve substantially after data augmentation, except VGG-16 for both binary and multiclass classification, as observed in Figure 2.9. On the other hand, we can see that the combination of our custom CNN (for processing Pentagon images) and EfficientNet (for processing sentences and signatures) is the one that achieves the highest accuracy: 93% and 80% for binary and multiclass classification, respectively. In terms of AUC, the combination of EfficientNet models (EffNet+EffNet) is the best performer, although it is only 6 percentual points higher than our Custom+EffNet model.

Our results show that concatenated-based CNN models with augmented data outperform previous studies that did not use concatenated-based models with augmentation (e.g., [13]). These findings suggest that the combination of custom CNN and EfficientNet (Custom + EffNet) is a promising option for automatically evaluating handwriting tasks.

Overall, our findings have the potential to improve the accuracy of AD detection and treatment outcomes. Importantly, our DFC model does not need a large number of data, which means that we can reach a robust classifier for detecting AD patients. These findings underline the substantial role data augmentation plays in boosting model performance and further demonstrate the benefits of using diverse input sources in the performance of the models.

Table 2.3: Performance results of other input source combinations before and after data augmentation: Pentagon+Sentence (abbreviated as Pen+Sen), Pentagon+Signature (Pen+Sig), and Sentence+Signature (Sen+Sig). We use the following nomenclature: “model A + model B” to indicate that “model A” was used to process the first input source (e.g. Pentagon) and “model B” was used to process the second input source (e.g. Sentence).

			Custom + VGG	**Custom + ResNet**	**Custom + DenseNet**	**Custom + EffNet**	**VGG + VGG**	**ResNet + ResNet**	**DenseNet + DenseNet**	**EffNet + EffNet**
			Acc. | AUC	Acc. | AUC	Acc. | AUC	Acc. | AUC	Acc. | AUC	Acc. | AUC	Acc. | AUC	Acc. | AUC
Binary	**Before**	**Pen+Sen**	0.67 | 0.66	0.6 | 0.6	0.4 | 0.4	0.66 | 0.67	0.6 | 0.61	0.6 | 0.6	0.8 | 0.89	0.67 | 0.76
		Pen+Sig	0.6 | 0.83	0.6 | 0.6	0.8 | 0.83	0.8 | 0.79	0.6 | 0.59	0.6 | 0.6	0.73 | 0.73	0.93 | 0.92
		Sen+Sig	0.6 | 0.78	0.6 | 0.6	0.4 | 0.4	0.87 | 0.85	0.6 | 0.64	0.6 | 0.6	0.4 | 0.4	0.8 | 0.84
	After	**Pen+Sen**	0.8 | 0.79	0.6 | 0.64	0.4 | 0.4	0.87 | 0.93	0.4 | 0.4	0.93 | 0.93	0.8 | 0.8	0.93 | 0.93
		Pen+Sig	0.1 | 0.1	0.93 | 0.93	0.87 | 0.86	0.93 | 0.92	0.6 | 0.6	0.6 | 0.6	0.93 | 0.92	0.4 | 0.4
		Sen+Sig	0.4 | 0.43	0.53 | 0.51	0.93 | 0.93	0.93 | 0.92	0.6 | 0.6	0.47 | 0.42	0.6 | 0.6	0.93 | 0.93
Multiclass	**Before**	**Pen+Sen**	0.2 | 0.52	0.6 | 0.7	0.2 | 0.40	0.6 | 0.74	0.4 | 0.58	0.2 | 0.4	0.6 | 0.7	0.8 | 0.89
		Pen+Sig	0.27 | 0.48	0.6 | 0.7	0.2 | 0.59	0.73 | 0.78	0.53 | 0.63	0.2 | 0.4	0.6 | 0.72	0.87 | 0.9
		Sen+Sig	0.2 | 0.4	0.6 | 0.7	0.6 | 0.69	0.93 | 0.94	0.6 | 0.7	0.6 | 0.7	0.6 | 0.72	0.73 | 0.83
	After	**Pen+Sen**	0.6 | 0.78	0.73 | 0.81	0.87 | 0.91	0.87 | 0.92	0.8 | 0.89	0.8 | 0.88	0.93 | 0.95	0.93 | 0.96
		Pen+Sig	0.4 | 0.57	0.8 | 0.87	0.93 | 0.95	0.93 | 0.96	0.8 | 0.89	0.8 | 0.88	0.93 | 0.96	0.93 | 0.96
		Sen+Sig	0.4 | 0.58	0.8 | 0.87	0.8 | 0.9	0.93 | 0.95	0.8 | 0.89	0.8 | 0.88	0.93 | 0.95	0.93 | 0.95

2.13 Conclusion and future work

We have successfully developed a robust and efficient model, capable of accurately classifying handwritten images into healthy individuals and AD patients ranging from mild to moderate severity. Our research primarily focused on evaluating the potential improvement in classification performance through the incorporation of handwriting data from diverse sources and feature concatenation. Our findings put forward the efficacy of the combination of DFC and data augmentation techniques in developing more holistic and precise models for AD screening.

The potential implications of our study are manifold, with particularly important implications within clinical settings. The developed DFC model enhances healthcare providers' decision-making capabilities, especially for untrained professionals, fostering improved patient care and mitigating the likelihood of unnecessary procedures or subjective diagnoses. As a screening method, it can be used anywhere from primary care settings to daycare facilities. Looking forward, we suggest future research should focus on creating a smartphone app, grounded in the established framework, that can collect and analyze handwritten data on the go. This app could potentially integrate multiple models (e.g. binary and multiclass classifiers) in order to account for different practitioners' needs. Furthermore, our methodology exhibits promising potential for the classification of other handwriting tasks, such as CDT or RCFT drawings. Hence, future work should focus on exploring and validating the model's proficiency in these tasks.

Acknowledgment

Work supported by the UCM research group (Grupo de Investigación básica en Ciencias de la Visión del IIORC, UCM-GR17-920105), the Horizon 2020 FET program of the European Union (grant CHIST-ERA-20-BCI-001), and the European Innovation Council Pathfinder program (grant 101071147).

References for Paper 2

[1] Samad Amini et al. 2021. An Artificial Intelligence-Assisted Method for Dementia Detection Using Images from the Clock Drawing Test. *Journal of Alzheimer's Disease*, 83, 2, 581–589. Publisher: IOS Press BV. DOI: `10.3233/JAD-210299`.

[2] Prashant Bhardwaj and Amanpreet Kaur. 2021. A novel and efficient deep learning approach for COVID-19 detection using X-ray imaging modality. en. *International Journal of Imaging Systems and Technology*, 31, 4, 1775–1791. DOI: `10.1002/ima.22627`.

[3] Russell Binaco, Nicholas Calzaretto, Jacob Epifano, Sean McGuire, Muhammad Umer, Sheina Emrani, Victor Wasserman, David J. Libon, and Robi Polikar. 2020. Machine Learning Analysis of Digital Clock Drawing Test Performance for Differential Classification of Mild Cognitive Impairment Subtypes Versus Alzheimer's Disease. *Journal of the International Neuropsychological Society*, 26, 7, (Aug. 2020), 690–700.

[4] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. 2020. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11, 2, (Feb. 2020), 125.

[5] Shuqing Chen, Daniel Stromer, Harb Alnasser Alabdalrahim, Stefan Schwab, Markus Weih, and Andreas Maier. 2020. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports 2020 10:1*, 10, 1, (Nov. 2020), 1–11. Publisher: Nature Publishing Group.

[6] Xuxin Chen et al. 2022. Recent advances and clinical applications of deep learning in medical image analysis. en. *Medical Image Analysis*, 79, (July 2022), 102444. DOI: `10.1016/j.media.2022.102444`.

[7] Martin Guha. 2014. Diagnostic and Statistical Manual of Mental Disorders: DSM-5 (5th edition). *Reference Reviews*, 28, 3, (Jan. 2014), 36–37. DOI: `10.1108/RR-10-2013-0256`.

[8] Harald Hampel et al. 2011. The future of Alzheimer's disease: The next 10 years. *Progress in Neurobiology*, 95, 4, (Dec. 2011), 718–728.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[10] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

[11] C. Jiménez-Mesa et al. 2022. Automatic Classification System for Diagnosis of Cognitive Impairment Based on the Clock-Drawing Test. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13258 LNCS, 34–42.

[12] Yike Li, Jiajie Guo, and Peikai Yang. 2022. Developing an Image-Based Deep Learning Framework for Automatic Scoring of the Pentagon Drawing Test. *Journal of Alzheimer's disease: JAD*, 85, 1, 129–139.

[13] Jumpei Maruta, Kentaro Uchida, Hideo Kurozumi, Satoshi Nogi, Satoshi Akada, Aki Nakanishi, Miki Shinoda, Masatsugu Shiba, and Koki Inoue. 2022. Deep convolutional neural networks for automated scoring of pentagon copying test results. *Scientific Reports*, 12, 1, (Dec. 2022), 9881.

[14] Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. 1984. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. en. *Neurology*, 34, 7, (July 1984), 939–939. DOI: `10.1212/WNL.34.7.939`.

[15] Long D. Nguyen, Dongyun Lin, Zhiping Lin, and Jiuwen Cao. 2018. Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. ISSN: 2379-447X. (May 2018), 1–5. DOI: `10.1109/ISCAS.2018.8351550`.

[16] Neelum Noreen, Sellappan Palaniappan, Abdul Qayyum, Iftikhar Ahmad, Muhammad Imran, and Muhammad Shoaib. 2020. A Deep Learning Model Based on Concatenation Approach for the Diagnosis of Brain Tumor. *IEEE Access*, 8, 55135–55144.

[17] Mohammad Rahimzadeh and Abolfazl Attar. 2020. A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2. en. *Informatics in Medicine Unlocked*, 19, (Jan. 2020), 100360.

[18] Tawsifur Rahman, Muhammad E. H. Chowdhury, Amith Khandakar, Khandaker R. Islam, Khandaker F. Islam, Zaid B. Mahbub, Muhammad A. Kadir, and Saad Kashem. 2020. Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray. en. *Applied Sciences*, 10, 9, (Jan. 2020), 3233. DOI: `10.3390/app10093233`.

[19] Elena Salobrar-García et al. 2019. Changes in visual function and retinal structure in the progression of Alzheimer's disease. *PLOS ONE*, 14, 8, (Aug. 2019), e0220535. DOI: `10.1371/journal.pone.0220535`.

[20] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. (2014).

[21] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. 2016. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, 35, 5, (May 2016), 1299–1312. DOI: `10.1109/TMI.2016.2535302`.

[22] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, (May 2019), 6105–6114.

[23] Tom Tombaugh and Nancy J McIntyre. 1992. The Mini-Mental State Examination: A Comprehensive Review - Tombaugh - 1992 - Journal of the American Geriatrics Society - Wiley Online Library. *Journal of the American Geriatrics Society*, 40, 9, 922–935.

[24] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D. Wang. 2021. Multimodal deep learning models for early detection of Alzheimer's disease stage. en. *Scientific Reports*, 11, 1, (Feb. 2021), 3254. DOI: `10.1038/s41598-020-74399-w`.

[25] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13, 4, (Apr. 2004), 600–612.

[26] Young Chul Youn et al. 2021. Use of the Clock Drawing Test and the Rey–Osterrieth Complex Figure Test-copy with convolutional neural networks to predict cognitive impairment. *Alzheimer's Research & Therapy*, 13, 1, (Apr. 2021), 85.

[27] Wenbo Zhu, Jinhong Zhang, and Jose Romagnoli. 2022. General Feature Extraction for Process Monitoring Using Transfer Learning Approaches. *Industrial & Engineering Chemistry Research*, 61, 15, (Apr. 2022), 5202–5214. DOI: 10.1021/acs.iecr.1c04565.

3 On-line and Off-line Handwriting

Handwriting and drawing assessments are being scrutinized as non-invasive methods for diagnosing neurodegenerative conditions such as ADs and MCI. These tasks require individuals to conceive, coordinate, and execute fine motor actions, thereby exposing subtle cognitive and motor deficiencies at early stages. A central issue under investigation is whether static imagery or dynamic trajectory data convey superior diagnostic insight. Three recent studies—*IDEAL'24, AC'24, and eScience'24*—have examined this question by contrasting static representations (scanned images) with dynamic recordings (discrete point sequences), both in isolation and in combination, to screen for AD and MCI.

IDEAL'24 examines two drawing tasks (pentagon and clock) using both static (scanned images) and dynamic (discrete point sequences) modalities. This study evaluates binary classification—differentiating Alzheimer's disease from healthy individuals—and multi-class classification—delineating mild Alzheimer's disease, moderate Alzheimer's disease, and healthy subjects—using CNNs for static data and RNNs for dynamic data. The findings reveal that, in contexts of limited data, static imagery tends to yield superior accuracy compared to dynamic trajectories, with binary classification accuracies approaching 90% versus 60% and multi-class accuracies of approximately 82% versus 53%.

AC'24 centers on the Hand-Drawing Test (HDT) and involves 58 participants categorized into AD, MCI, and healthy control groups. Each drawing is recorded in two formats: as a scanned static image and as a digitally captured dynamic trace. Comparative analyses between models using a single modality and those that amalgamate spatial and temporal features reveal that a multimodal approach enhances classification performance. Data expansion strategies are used to mitigate the limitations of small sample sizes, and the outcomes indicate that the combined features distinguish MCI from normal aging more reliably than either modality alone.

eScience'24 builds upon the focus on the HDT by introducing an "OnOff-line" approach in which dynamic recordings are supplemented and converted into static images. This method takes advantage of the intricate movements inherent in dynamic data while capitalizing on CNN architectures designed for image analysis. The results indicate that this hybrid method frequently outperforms both conventional static and dynamic pipelines, particularly in binary classification scenarios (distinguishing healthy individuals from patients). Such findings suggest that converting and integrating data modalities may counteract the shortcomings of direct dynamic modeling, especially in studies constrained by limited participant numbers.

Collectively, these investigations articulate two pivotal themes. The appropriateness of static versus dynamic data is contingent upon factors such as sample magnitude, task complexity, and model architecture. Furthermore, the conversion or fusion of data types may counterbalance the inherent limitations of any single modality. The remainder of this chapter elucidates the methodological underpinnings of each study, detailing data preprocessing, model architecture selection, and classification strategies, while also addressing approaches to overcome small sample challenges and the implications for diagnostic applications in clinical and community settings.

3.1 Blueprint of Tomorrow: Contrasting Off-line and On-line Drawing Tasks for Alzheimer's Disease Screening

Abstract[†]

Alzheimer's disease (AD) is the leading cause of dementia. Although there is currently no cure for AD, early detection of cognitive decline can help clinicians mitigate its impact. Recently, Machine Learning (ML) approaches have been developed to automatically analyze handwriting and hand-drawing tasks to support the early diagnosis of AD. In this paper, we study pentagon and clock drawing tests using both off-line (scanned image pixels) and on-line (discrete point sequences) data as input to several ML models (i.e., DensNet, ResNet, EfficientNet, RNN, LSTM, and GRU). Our study is the first to determine the most effective modality (on-line vs. off-line) and drawing tasks to distinguish healthy controls from AD patients (binary classification) as well as two stages of AD severity (multi-class classification). Our results suggest that, contrary to other domains, the off-line modality outperforms the on-line one, sometimes by a large margin: 90% vs. 60% accuracy in binary classification and 53% vs. 82% accuracy in multi-class classification. This suggests that, for drawing tasks and small-scale datasets, image-based representations may be more effective in predicting AD than those relying on more complex data representations.

Keywords

Alzheimer disease; Off-line handwriting; On-line handwriting; Deep learning; Data augmentation; Classification.

3.2 Introduction

Neurodegenerative disorders are among the primary causes of disability worldwide, marked by an irreversible loss of neurons that culminates in progressive neurological decline, manifesting as motor and cognitive impairments. Alzheimer's disease (AD) and Parkinson's

[†] **N. Hosseini-Kivanani**, E. Salobrar-García, L. Elvira-Hurtado, M. Salas, C. Schommer, and L. A. Leiva, "Blueprint of Tomorrow: Contrasting Off-Line and On-Line Drawing Tasks for Alzheimer's Disease Screening," in *Intelligent Data Engineering and Automated Learning – IDEAL 2024, Lecture Notes in Computer Science*, vol. 15346, V. Julian *et al.*, Eds. Cham: Springer, 2025, pp. 422–433. DOI: 10.1007/978-3-031-77731-8_38

disease (PD) are particularly notable for their widespread prevalence, which affects approximately 50 million and 10 million individuals, respectively. AD, in particular, is closely linked with cognitive deficits, affecting memory, attention, language comprehension, and spatial awareness [32]. The global demographic trend toward an older population underscores the critical need for early AD detection and intervention as the most prevalent form of dementia. However, the present diagnostic landscape reveals a concerning trend, with an estimated 75% of dementia cases worldwide going undiagnosed and rates of early-stage detection being considerably lower [10]. Enhancing screening methodologies in accessible settings, particularly in primary healthcare, emerges as a strategic response to improve diagnostic rates [8]. Research indicates that primary healthcare practitioners face substantial challenges in the early detection of dementia and in the timely referral of patients to specialized care [10]. Thus, the development and implementation of accessible and efficient screening tools for use in primary healthcare or by individuals at home are crucial steps toward closing the diagnostic gap, potentially leading to improved detection rates of AD.

Handwriting and hand-drawing tasks[‡] entail the coordination of fine motor movements and cognitive processes, making them popular as a psychometric tool to evaluate and diagnose AD [35], leveraging the correlation between declining drawing abilities and the onset of AD. Deterioration in handwriting skills, characterized by inconsistencies in size, spacing, and letter formation, indicates a progression of the disease [7]. Recent studies (e.g., [19, 23]) have shown the potential of handwriting-related tasks to reveal specific cognitive deficits indicative of AD. Researchers have explored various automated methods, including drawing tasks [19], neuroimaging [31], and gait assessments [11], to capture cognitive impairments across multiple domains. However, the current need for healthcare professionals' reliance on manual analysis highlights a significant bottleneck. This puts forward the importance of developing automated tools to make the AD screening process easier, quicker, and more affordable, particularly in non-specialist settings.

Our contributions are straightforward yet significant. Firstly, we gathered handwriting data from both AD patients and healthy individuals. The data includes two types of drawings—pentagons and clocks—captured *simultaneously* in two formats: off-line, as scanned images, and on-line, as sequences of discrete points. This dual-method collection allows us to: (i) Compare how different hand-drawing tasks perform in classification tests, (ii) Assess various machine learning models to see which best classifies the data, and (iii) Examine the differences in using static images versus dynamic, point-by-point data in model performance.

[‡] We consider 'handwriting' and 'hand-drawing' synonymous because both tasks involve the same neurophysiological and peripheral processes involved in motor control.

Secondly, our experiments focus on distinguishing AD (mild AD and moderate AD) patients from healthy controls using state-of-the-art neural network models. We used pre-trained convolutional neural networks (CNNs) for analyzing the off-line data and RNNs for the on-line data. Additionally, we used data augmentation techniques to enhance the models' ability to generalize. Interestingly, our findings reveal that the off-line modality consistently outperformed the on-line one, achieving higher accuracy in both binary (90% versus 60%) and multi-class (53% versus 82%) classification tasks. This suggests that simpler, static image-based approaches may be more effective for tasks like drawing analysis in AD research than those relying on more complex, temporal data, at least when working with small-scale datasets.

3.3 Related Work

Technological advancements in ML and computer vision have significantly enhanced the efficiency and objectivity of remote patient monitoring systems by providing real-time data for improved care beyond conventional healthcare settings [6]. Handwriting analysis has been shown to be effective in detecting cognitive decline and changes in motor skills in AD, thus serving as an effective diagnostic tool [18]. However, despite these advances, there is no research aimed at comparing model performance using *both* off-line and on-line data from *identical* patient/healthy cohorts [5, 29].

Recent studies have used CNNs (e.g., [17, 18, 3]) and RNNs (e.g., [17, 2, 6]) for early detection of AD, showing that Deep Learning (DL) models can significantly enhance the accuracy of diagnosing AD in its early stages. We can find notable works that studied each modality separately (cf. off-line [19, 15, 3] and on-line [6, 23, 21]), suggesting that on-line data are preferred over off-line data, given that on-line handwriting provides a feature-rich representation, including, e.g., temporal and spatial sequences of discrete points that are not available in the off-line representation. Collectively, these studies highlight the potential of ML technologies not only to revolutionize AD diagnosis but also to enable more personalized and timely therapeutic interventions, ultimately improving patient outcomes. Most of these studies have relied on some form of data augmentation, given the limited number of samples in clinical datasets. In this regard, Dao et al. [6] used Generative Adversarial Networks (GANs) as an alternative to data augmentation, and trained AD classifiers with RNNs that achieved 89% accuracy. However, GANs require a significant amount of data to begin with, which is often not available in most cases.

Finally, we should mention relevant studies that have compared different drawing symbols for AD screening, such as clocks drawngs [1] using CNN models, which achieved an AUC score of 81%. By combining clock drawing with age and education using logistic regression, their model improved to 91%. Pentagon drawings [26] reached an accuracy of 93% using GoogLeNet for binary classification, distinguishing between correct and incorrect pentagon drawings from patients only. Another study focused on letters [6], and obtained high accuracy by using DL models for detecting and classifying early-stage of ADs patients based on on-line handwriting loop patterns. In sum, it remains unclear which is the most adequate input modality for AD screening and also what the most adequate drawing symbols are to achieve competitive performance.

By systematically addressing this gap in the research literature, our study paves the way for a more holistic understanding of AD classification models, opening promising directions to more accurate AD screening approaches in the future. For example, [23] reported that combining multiple drawing tasks improves detection accuracy by capturing different cognitive impairments, achieving a classification accuracy of 75.2%.

3.4 Materials and Methods

3.4.1 Participants

Thirty-three individuals were recruited from the Memory Unit of the Hospital Clinico San Carlos (HCSC) in Madrid between January 2023 and January 2024. The group consisted of 22 patients and 11 healthy controls (HCs), all aged between 70 and 89. Participants were asked to both clocks and pentagons, which are well-established symbols in cognitive assessment tasks [9]. All participants had normal vision and hearing. They underwent a neuropsychological assessment of their drawing tasks in a clinical setting to minimize distractions and reduce background noise. Each participant was individually assessed, beginning with an informed consent form. Cognitive status was evaluated using the Mini-Mental State Examination (MMSE) [39]. Patients with AD were classified into mild AD and moderate AD, based on guidelines from the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS), the Alzheimer's Disease and Related Disorders Association (ADRDA) workgroup [28], and the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [13]. Statistical analysis (ANOVA test, after verification of normality and homoscedasticity) confirmed that there were no significant age differences between the healthy, mild AD, and moderate AD groups ($F(2, 55) = 2.04, p > .139$).

Table 3.1: Demographics and the Number of Drawing Tasks of this Study.

Drawing Task	**Num. of drawing**	**HC**	**Mild AD**	**Moderate AD**
Pentagon	33	11	8	14
Clock	33	11	8	14
Total	**66**	**22**	**16**	**28**
Gender (F & M)		8F, 3M	5F, 3M	12F, 4M
Age (Mean ± SD)		82.64±2.46	76.5±5.75	78.94±4.78
MMSE (Mean ± SD)		29.9±0.83	25.33±1.21	22.37±3.58

3.4.2 Drawing tasks and preprocessing

Participants were instructed to draw the pentagons and clocks using a Repaper tablet (size: 10.9-inch) [§] with a blank sheet attached and a regular pen that had an accelerometer connected to the Repaper app via Bluetooth for data capturing (see Figure 3.1). This setup was designed to provide a familiar pen-and-paper experience to participants while being able to capture on-line and off-line data simultaneously. The participants were asked to draw each symbol from memory. We collected 66 drawings in total.

The on-line data (discrete point sequences) were stored as SVG files (as per the Repaper app) and then converted to JSON format, comprising sequences of `{x,y,t}` points. The off-line data (image pixels) were stored as PDF files (scanned with the HP Color LaserJet Pro scanner) and then converted to PNG images and resized to square size (224×224 px) as this is standard for CNN models. We applied the canny edge detector to enhance the quality of the scanned images.

3.4.3 Data augmentation

We created synthetic samples to make the models more robust and generalizable [12]. For off-line data, we applied the usual geometric transformations, where suitable:[¶] For on-line data, jittering, scaling, and warping have been proposed [22]; however, more recently, Maslych et al. [27] found that an “All Variability Chain” (AVC) of transformations (gaussian, frame-skip, spatial, perspective, rotate, scale) provided a significant boost in classification performance with RNNs, achieving state-of-the-art accuracy in gesture recognition. Therefore, we adopt the AVC approach to augment our on-line data; see Figure 3.1. After data augmentation, we concluded to a dataset consisting of 300 images (off-line data) and 300 point sequences (on-line data), where all groups were balanced up

[§] `https://www.iskn.co/eu`

[¶] For example, pentagons can be flipped in horizontal or vertical axes, whereas clocks cannot be flipped because it would destroy their semantics.

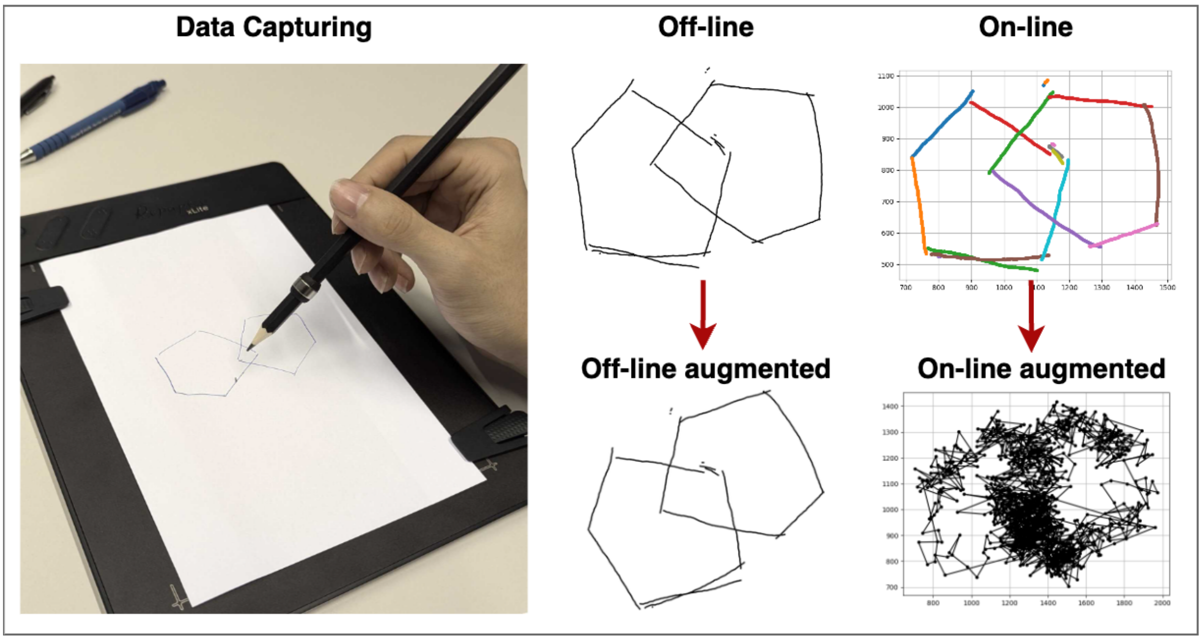


Figure 3.1: Example of pentagon drawing on a tablet (left) and drawing samples in off-line and on-line version (right), before and after applying data augmentation.

to 100 observations; e.g., 92 variations of the 8 pentagons from the Mild AD group were created; see Figure 3.2.

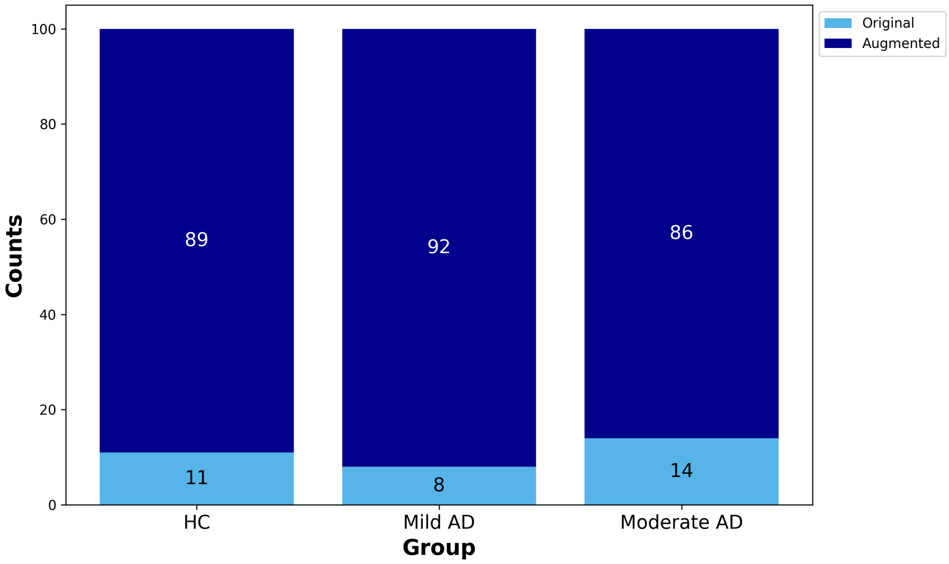


Figure 3.2: Class distribution of Pentagon drawings before and after data augmentation.

To evaluate the quality of the augmented data, we used the Structural Similarity Index Measure (SSIM) [41] for off-line data and Dynamic Time Warping (DTW) [34] for on-line data. We use SSIM to compare augmented data against original images to ensure that key structural details are preserved even as variations are introduced. This balance is critical since, while the augmented data are inherently different, maintaining structural similarity ensures that the variations remain realistic and relevant for training robust models. By using SSIM, we can confirm that the augmentation process does not distort the data to the extent that it loses its representative characteristics. In this sense, structural similarity is beneficial, as it ensures that the augmented data faithfully represents the original data. SSIM values range from 0.7 to 0.8 (M=0.75, SD=0.03), whereas DTW values range between 123 and 5678 (M=2000, SD=850), indicating that augmented images are not near-duplicates of the original data but rather new images that eventually should help improving model performance.

3.4.4 Models

3.4.4.1 Convolutional Neural Nets

We use three state-of-the-art pre-trained CNNs for analyzing the off-line data: ResNet50 [14], DenseNet121 [20], and EfficienNet [37]. ResNet and DenseNet use residual connections, which are instrumental to train very deep models. While ResNet performs an element-wise addition to pass the output to the next layer, DenseNet connects all layers directly to each other through concatenation. However, EfficienNet uses a uniform compound scaling technique that achieves the same performance as state-of-the-art CNNs but with much better efficiency. These CNNs were trained on the large ImageNet dataset, and we fine-tuned them to our AD dataset by transfer learning [42].

3.4.4.2 Recurrent Neural Nets

Since transfer learning for on-line data is not currently possible, as there are no public pre-trained models available, we train three RNNs from scratch: Vanilla RNN, LSTM [16], and GRU [4]. LSTM is an improvement over vanilla RNNs by adding long-term memory, making them ideal for complex sequences. GRU is a simplification of LSTM while retaining the same performance, making them ideal for cases where computational efficiency is crucial. These RNN models include a hidden layer of 100 units with hyperbolic tangent activation and 0.1 dropout, followed by a softmax output layer. We experimented with other combinations of layers and different hidden units, but we did not observe improvements with regard to this configuration.

3.4.4.3 Training and evaluation

All CNNs and RNNs are trained with the popular Adam optimizer, with a learning rate of 0.001 and decay rates $\beta_1 = \beta_2 = 0.99$. The loss function is categorical cross-entropy, consistent with our binary and multi-class classification tasks. All models use a batch size of 32 (images or sequences) and use up to 50 epochs for training with early stopping (patience of 10 epochs) to prevent overfitting. We train each model on 80% of the data and test on the remaining 20% of both on-line and off-line data. We then used stratified 5-fold cross-validation on the training set only, ensuring that each fold was representative of the whole by maintaining approximately the same percentage of samples of each class as in the training subset. We computed the classification accuracy (Acc) and Area Under the ROC curve (AUC) to assess model performance.

3.5 Results and discussion

Figure 3.3 summarizes the results of our experiments, depicting the differences between CNN classifiers (top row) for off-line data (ResNet, DenseNet, and EfficienNet) and RNN classifiers (bottom row) for on-line data (vanilla RNN, LSTM, and GRU). These results are instrumental for assessing the efficacy of binary (distinguishing AD patients from HC) and multi-class classification (distinguishing HC, mild AD, and moderate AD) tasks in the context of AD screening using hand-writing tasks (i.e., pentagon and clock).

In binary classification of pentagon drawings, the before-augmentation scenario showed modest performance across models. After-augmentation, however, there was an increase in classifier accuracy, particularly with EfficienNet, which improved substantially, from 60% to 90%. A similar trend was observed in the classification of clock drawings, where the after-augmentation results underscored the effective impact of our data augmentation strategies on model performance.

For multi-class classification, the differential impact of data augmentation was again significant. EfficientNet was the best performer, especially for pentagon drawings, confirming the model's ability to differentiate between various stages of AD severity. The trend of improved after-augmentation accuracy was consistent across other models, although the degree of improvement varied.

Our results indicate that off-line data, when enhanced through strategic data augmentation, provides a more stable and consistent basis for AD classification compared to on-line data. This stability can be attributed to the static nature of off-line data, which, unlike on-line data, is less affected by the variabilities introduced by the temporal and dynamic components of on-line drawing. In non-clinical domains, researchers have shown

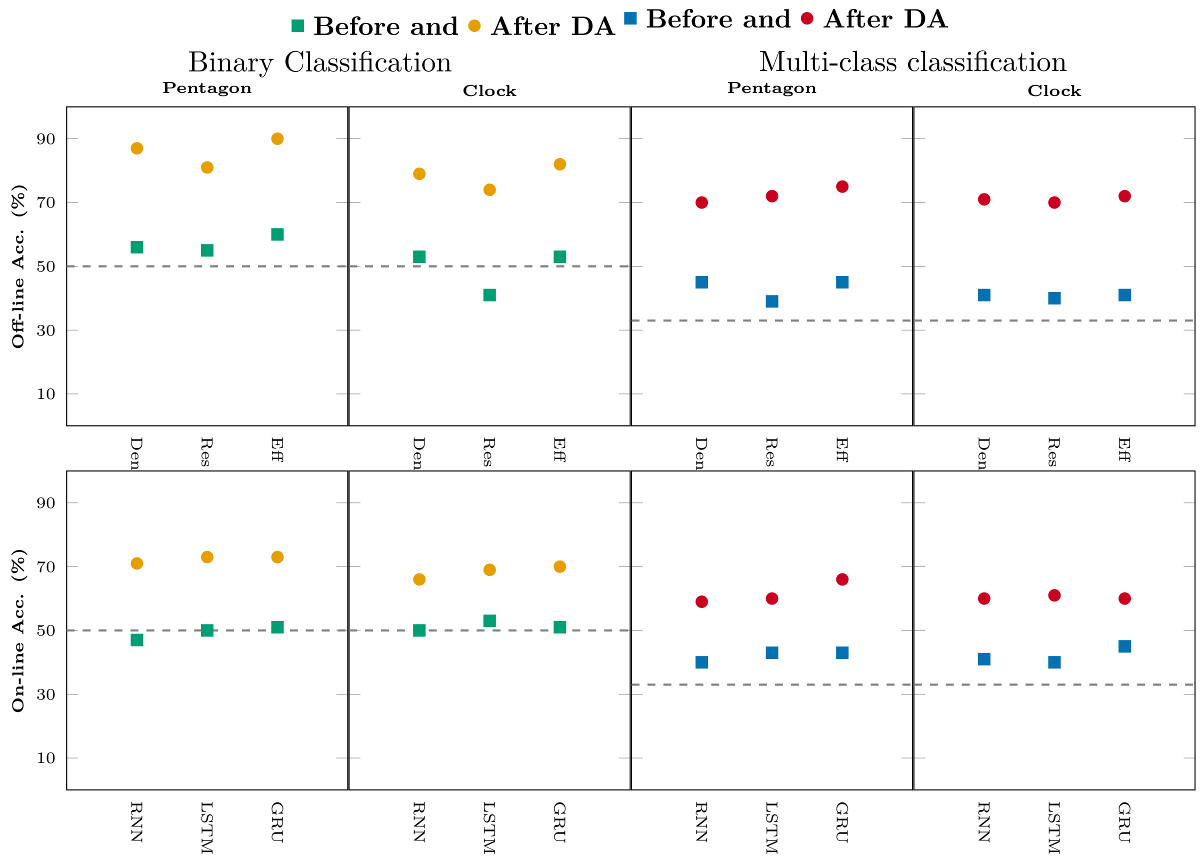


Figure 3.3: Experiment results. CNN classifiers for off-line data (Res: ResNet, Den: DenseNet, Eff: EfficienNet) are depicted in the top row, whereas RNN classifiers for on-line data (vanilla RNN, LSTM, GRU) are depicted in the bottom row. The dashed lines represent the performance of a random classifier, illustrating the empirical lower bound.

that on-line data is preferred over off-line data (e.g., [38, 24]), given the rich patterns and movement dynamics involved [25]. In our experiments, however, we observed that pre-trained CNNs outperformed RNNs trained from scratch due to the size and variability in online data that our RNNs were not able to capture as effectively as the CNNs. This variability also seems to affect tasks differently; for example, more cognitively demanding drawing tasks allowed for an easier distinction of AD patient handwriting from HCs; see e.g., Figure 3.3.

Our study builds upon previous research that focused on one drawing type for AD screening [18, 3]. However, our findings suggest that both pentagon and clock drawings are suitable for AD screening, with EfficientNet achieving the highest performance in binary classification (90% accuracy, 92% AUC) and in multi-class classification settings (75% accuracy, 79% AUC) for using pentagons, followed by clocks (Binary classification: 82% accuracy and 79% AUC, multi-class classification: 70% accuracy and 75% AUC), highlighting the effectiveness of these simple tasks.

Our results also indicate that the performance gap before and after data augmentation differs across tasks, with larger differences in off-line data. Previous results by Maslych et al. [27] reported improved performance of an RNN model in several handwriting tasks using AVC for data augmentation, although all the tasks were focused on gesture recognition and a specific dataset. In any case, it seems clear that data augmentation is necessary for both off-line and on-line drawing tasks.

Without data augmentation, most models behaved like a random classifier; see the dashed lines in Figure 3.3. After data augmentation, however, we observed a significant improvement in model performance across all tasks for both binary and multi-class classification scenarios. These improvements were more apparent for CNN models, which somehow disagrees with previous findings in AD screening that reported similar performance for RNNs [2, 30].

Interestingly, Souillard-Mandar et al. [36] found that the digital Clock Drawing Test had superior diagnostic performance compared to traditional paper-and-pencil methods for differentiating healthy individuals from cognitive impairment subjects (only binary classification), using traditional ML models without any data augmentation. Previous studies have demonstrated the effectiveness of various augmentation strategies in similar contexts for off-line data [19, 33]. In our work, we designed and optimized our augmentation strategy to ensure its suitability for our specific datasets. This involved iterative testing and refinement of different augmentation techniques, such as geometric transformations.

3.6 Limitations and future work

The main limitation of our study is the small sample size of the original dataset, which had to be augmented with suitable variations in order to fine-tune the CNN models and, more importantly, to train the RNN models from scratch. However, we should remark that dealing with small sample sizes is a well-known and pervasive issue among clinical studies [21, 40]. Recruiting participants with AD is very challenging due to strict criteria and ethical concerns. Despite our best efforts, it took us one year to recruit 33 suitable participants. On the other hand, future work should consider different data augmentation approaches for on-line trajectories, since we have observed that the AVC method [27] is suboptimal for AD screening. Additionally, an interesting direction for future research could be assessing whether simpler, custom-built CNN models can achieve comparable or superior results with reduced complexity.Ultimately, despite these shortcomings, our findings show promise and could lead to practical clinical applications.

3.7 Conclusion

We have analyzed the impact of off-line vs. on-line handwriting data for AD screening using pentagon and clock drawings with suitable data augmentation techniques. We trained several CNNs and RNNs for binary and multi-class classification settings. Our results show that data augmentation is always beneficial and that pentagons have better discriminative power than clocks. Our results also show that CNNs outperform RNNs in all settings, contradicting what was previously known in non-clinical work.

Our observed improvements in performance suggest that while our current strategy enhances model robustness, further optimizations could indeed yield even better results, a possibility that we aim to explore in future work. We acknowledge that there are numerous opportunities for additional refinement, and future work will continue to explore and optimize these techniques to maximize their efficacy in enhancing model robustness and accuracy. Our code and models are available upon reasonable request.

Acknowlegments

Work supported by the UCM research group (Grupo de investigacion básica en Ciencias de la Visión del IIORC, UCM-GR17-920105), the Horizon 2020 FET program of the European Union (grant CHIST-ERA-20-BCI-001), and the European Innovation Council Pathfinder program (grant 101071147).

References for Paper 3

[1] Samad Amini et al. 2021. An Artificial Intelligence-Assisted Method for Dementia Detection Using Images from the Clock Drawing Test. *J. Alzheimers Dis.*, 83, 2.

[2] Asma Bensalah, Antonio Parziale, Giuseppe De Gregorio, Angelo Marcelli, Alicia Fornés, and Josep Lladós. 2023. I can't believe it's not better: in-air movement for alzheimer handwriting synthetic generation. In *Proc. IGS*.

[3] Shuqing Chen, Daniel Stromer, Harb Alnasser Alabdalrahim, Stefan Schwab, Markus Weih, and Andreas Maier. 2020. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports*, 10, 1.

[4] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proc. EMNLP*.

[5] Nicole D. Cilia, Giuseppe De Gregorio, Claudio De Stefano, Francesco Fontanella, Angelo Marcelli, and Antonio Parziale. 2022. Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking. *Engineering Applications of Artificial Intelligence*, 111, (May 2022), 104822. Retrieved Mar. 14, 2024 from.

[6] Quang Dao, Mounîm A. El-Yacoubi, and Anne-Sophie Rigaud. 2023. Detection of Alzheimer Disease on Online Handwriting Using 1D Convolutional Neural Network. *IEEE Access*, 11. Retrieved Nov. 24, 2023 from.

[7] Margarete Delazer, Laura Zamarian, and Atbin Djamshidian. 2021. Handwriting in Alzheimer's Disease. *Journal of Alzheimer's Disease*, 82, 2, 727–735. Publisher: IOS Press BV. DOI: 10.3233/JAD-210279.

[8] Tilly Eichler, Jochen René Thyrian, Johannes Hertel, Bernhard Michalowsky, Diana Wucherer, Adina Dreier, Ingo Kilimann, Stefan Teipel, and Wolfgang Hoffmann. 2015. Rates of formal diagnosis of dementia in primary care: The effect of screening. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1, 1, (Mar. 2015), 87–93.

[9] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. 1975. "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.*, 12, 3.

[10] Serge Gauthier, Pedro Rosa-Neto, Jos{\'e} A Morais, and Claire Webster. 2021. World Alzheimer Report 2021: Journey through the diagnosis of dementia. *Alzheimer's Disease International*, 2022, 30.

[11] Behnaz Ghoraani, Lillian N Boettcher, Murtadha D Hssayeni, Amie Rosenfeld, Magdalena I Tolea, and James E Galvin. 2021. Detection of mild cognitive impairment and alzheimer's disease using dual-task gait assessments and machine learning. *Biomed. Signal Process. Control*, 64.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, (Nov. 2016). ISBN: 9780262337373.

[13] Martin Guha. 2014. Diagnostic and Statistical Manual of Mental Disorders: DSM-5 (5th edition). *Reference Reviews*, 28, 3.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. CVPR*.

[15] Yuji Higaki. 2023. Clock-Drawing Test and Cube-Copying Test to Quickly Screen Dementia: In Combination with the Mini-Mental State Examination Scores. *Internal Medicine*. The Japanese Society of Internal Medicine, 2579–23.

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.*, 9, 8.

[17] Nina Hosseini-Kivanani, Elena Salobrar-García, Lorena Elvira-Hurtado, Mario López-Cuenca, Christoph Schommer, and Luis A. Leiva. 2024. Predicting alzheimer's disease and mild cognitive impairment with off-line and on-line house drawing tests. In *Proc. e-Science*. IEEE.

[18] Nina Hosseini-Kivanani, Christoph Schommer, and Luis. A Leiva. 2023. The Magic Number: Impact of Sample Size for Dementia Screening Using Transfer Learning and Data Augmentation of Clock Drawing Test Images. In *Proc. Healthcom*.

[19] Nina Hosseini-Kivanani et al. 2024. Ink of insight: data augmentation for dementia screening through deep learning. In *Proc. ICMHI*.

[20] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Proc. CVPR*.

[21] Donato Impedovo and Giuseppe Pirlo. 2018. Dynamic Handwriting Analysis for the Assessment of Neurodegenerative Diseases: A Pattern Recognition Perspective. *IEEE Rev. Biomed. Eng.*, 12.

[22] Brian Kenji Iwana and Seiichi Uchida. 2021. An empirical survey of data augmentation for time series classification with neural networks. en. *PLOS ONE*, 16, 7, (July 2021), e0254841. DOI: `10.1371/journal.pone.0254841`.

[23] Masatomo Kobayashi, Yasunori Yamada, Kaoru Shinkawa, Miyuki Nemoto, Kiyotaka Nemoto, and Tetsuaki Arai. 2022. Automated Early Detection of Alzheimer's Disease by Capturing Impairments in Multiple Cognitive Domains with Multiple Drawing Tasks. eng. *Journal of Alzheimer's disease: JAD*, 88, 3, 1075–1089.

[24] Luis A. Leiva, Vicent Alabau, Verónica Romero, Alejandro H. Toselli, and Enrique Vidal. 2015. Context-aware gestures for mixed-initiative text editing UIs. *Interact. Comput.*, 27, 6.

[25] Luis A. Leiva, Moises Diaz, Miguel A. Ferrer, and Réjean Plamondon. 2021. Human or Machine? It Is Not What You Write, But How You Write It. In *Proc. ICPR*.

[26] Jumpei Maruta, Kentaro Uchida, Hideo Kurozumi, Satoshi Nogi, Satoshi Akada, Aki Nakanishi, Miki Shinoda, Masatsugu Shiba, and Koki Inoue. 2022. Deep convolutional neural networks for automated scoring of pentagon copying test results. *Scientific Reports*, 12, 1.

[27] Mykola Maslych, Eugene Matthew Taranta, Mostafa Aldilati, and Joseph J. Laviola. 2023. Effective 2D Stroke-based Gesture Augmentation for RNNs. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23). Association for Computing Machinery, New York, NY, USA, (Apr. 2023), 1–13. ISBN: 9781450394215. DOI: `10.1145/3544548.3581358`.

[28] Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. 1984. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. en. *Neurology*, 34, 7.

[29] Stephan Müller, Oliver Preische, Petra Heymann, Ulrich Elbing, Christoph Laske, Aurel Popa-Wagner, and Junhong Yu. 2017. Increased Diagnostic Accuracy of Digital vs. Conventional Clock Drawing Test for Discrimination of Patients in the Early Course of Alzheimer's Disease from Cognitively Healthy Individuals. *Front. Aging Neurosci.*, 11.

[30] Nickson Mwamsojo, Frederic Lehmann, Mounim A. El-Yacoubi, Kamel Merghem, Yann Frignac, Badr-Eddine Benkelfat, and Anne-Sophie Rigaud. 2022. Reservoir Computing for Early Stage Alzheimer's Disease Detection. *IEEE Access*, 10.

[31] M. Odusami, R. Maskeliūnas, and R. Damaševičius. 2022. An intelligent system for early recognition of alzheimer's disease using neuroimaging. *Sensors*, 22.

[32] Richard J. Perry and John R. Hodges. 1999. Attention and executive deficits in Alzheimer's disease: A critical review. *Brain*, 122, 3, (Mar. 1999), 383–404.

[33] Raksit Raksasat, Surat Teerapittayanon, Sirawaj Itthipuripat, Kearkiat Praditpornsilpa, Aisawan Petchlorlian, Thiparat Chotibut, Chaipat Chunharas, and Itthi Chatnuntawech. 2023. Attentive Pairwise Interaction Network for AI-Assisted Clock Drawing Test Assessment of Early Visuospatial Deficits. en. SSRN Scholarly Paper. Rochester, NY, (Jan. 2023). DOI: 10.2139/ssrn.4327538.

[34] Pavel Senin. 2008. Dynamic time warping algorithm review. Tech. rep. 855. Information and Computer Science Department University of Hawaii, USA.

[35] Alastair D. Smith. 2009. On the use of drawing tasks in neuropsychological assessment. *Neuropsychology*, 23, 2, 231–239.

[36] William Souillard-Mandar, Randall Davis, C. Rudin, R. Au, D. Libon, R. Swenson, C. Price, M. Lamar, and Dana L. Penney. 2015. Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. *Mach. Learn.*, 102.

[37] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proc. ICML*.

[38] Charles C. Tappert and Pauline H. Mosley. 2001. Recent Advances in Pen Computing. Tech. rep. 166. Pace University.

[39] Tom Tombaugh and Nancy J McIntyre. 1992. The Mini-Mental State Examination: A Comprehensive Review. *J. Am. Geriatr. Soc.*, 40, 9.

[40] Gennaro Vessio. 2019. Dynamic Handwriting Analysis for Neurodegenerative Disease Assessment: A Literary Review. *Applied Sciences 2019, Vol. 9, Page 4666*, 9, 21, (Nov. 2019), 4666–4666. Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/APP9214666.

[41] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13, 4.

[42] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *J. Big Data*, 3, 1.

3.8 Screening of Alzheimer's disease and mild cognitive impairment through integrated on-line and off-line house drawing tests

Abstract[†]

Objective: Evaluate the effectiveness of machine learning (ML) algorithms in classifying mild cognitive impairment (MCI) and Alzheimer's disease (AD) using features derived from the House Drawing Test (HDT). Methods: The HDT was administered to 58 participants, categorized into AD (n = 22), MCI (n= 25), and Healthy Controls (HC, n = 11). Drawings were simultaneously captured using an electronic pen (on-line format) and scanned (off-line format). Results: The models achieved high classification accuracy across all groups: HC vs. MCI (67%), MCI vs. AD (70%), HC vs. AD (76%). Our results showcase the potential of ML for early screening of neurodegenerative diseases.

Keywords

On-Line, Off-Line, Deep Learning, Alzheimer's Disease, Mild Cognitive Impairment, House Drawing.

3.9 Introduction

Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) present significant challenges in an aging society, necessitating early diagnosis for effective disease management. MCI, often a precursor to AD, is characterized by a cognitive decline that, while noticeable, does not yet significantly impair daily activities. Identifying MCI as early as possible is essential, as it allows for interventions that may delay the onset of AD. There is thus an urgent demand for diagnostic tools and strategies to facilitate early detection and intervention [31, 20].

Cognitive assessments have become a focus for early detection of MCI and AD, particularly relying on drawing tasks that assess constructional abilities [18, 29], such as the Rey-Osterrieth Complex Figure [3], the Clock Drawing Test (CDT) [7], and the House

[†] **N. Hosseini-Kivanani**, E. Salobrar-García, L. Elvira-Hurtado, M. Salas, C. Schommer, and L. A. Leiva, "Screening of Alzheimer's disease and mild cognitive impairment through integrated on-line and off-line house drawing tests," *IADIS International Journal on Computer Science and Information Systems*, pp. 37–53, 2024.

Drawing Test (HDT) [33]. However, traditional cognitive assessments, based on pen and paper, are often time-consuming, prompting the development of quicker, semi-quantitative alternatives [19, 2]. Drawing tests based on electronic pens provide more quantifiable metrics (e.g., drawing latency or visual quality) to differentiate between individuals with and without neurodegenerative diseases [22, 32]. Despite these advancements, little research has compared how traditional (off-line, scanned images) and digital (on-line, time series) drawings perform in practice. While previous work noted that on-line representations offer richer features than off-line data [1, 8], a systematic comparison between these input types for diagnostic accuracy is currently lacking. To bridge this gap, we investigate the impact of data augmentation (DA) on both off-line and on-line representations for MCI and AD screening.

We focus on HDT drawings given the test's complexity and its demand for visuospatial and cognitive planning abilities, as previous research has shown that more complex tasks are more sensitive to early cognitive impairments [28]. Further, the HDT's requirement for participants to draw from memory, as opposed to copying, places a higher cognitive load, making it a robust tool for early detection of conditions like MCI and AD [25].

3.10 Related work

Handwriting analysis [‡] has emerged as a cost-effective and reliable method for early detection of AD and MCI. Various studies (e.g., [11]) have used handwriting-based features to differentiate between AD, MCI, and Healthy Controls (HC). However, task effectiveness can vary significantly; for example, symbols like the spiral may not fully capture the fine-grained details of spatial awareness, planning, and memory, which are particularly affected in MCI and AD patients.

[10] analyzed kinematic and pressure features of handwriting in 52 participants. The tasks included drawing of crossed pentagons, spirals, 3D houses, and the CDT. Their study highlighted the potential of on-line features in distinguishing between healthy subjects and those with cognitive impairments. Supporting these findings, [30] reported significant differences in temporal measures and pressure among AD, MCI, and HC groups.

Traditional off-line cognitive assessments have primarily focused on identifying outlines and details using scoring systems, often overlooking the sequence of drawing actions. A limited number of studies have employed digital tools, such as pens or tablets, to record the drawing process (see, e.g., [7]). [24] used a digital pen to analyze continuity and symmetry variables, offering insights into cognitive functions beyond traditional methods.

[‡] In this paper, we consider *handwriting* and *hand-drawing* interchangeably, as both involve similar neurophysiological and peripheral processes involved in motor control.

Similarly, [17] used a tablet to automatically extract stroke parameters and spatial information. They found that AD patients produced more fragmented drawings, took longer pauses, and demonstrated lower accuracy than individuals with normal cognition.

Our study contributes to the research literature by evaluating the effectiveness of various computational models for detecting AD and MCI (e.g., [14, 5]). For example, digital parameters of the CDT have effectively demonstrated cognitive processes and distinguished between patients with amnestic MCI, mild AD, and those with normal cognition [34]. However, specific drawing behaviors in MCI patients remain underexplored.

3.11 Methodology

We sought to explore cognitive and motor functions through a drawing task designed to assess creativity and precision across different cognitive stages (HC vs. MCI, MCI vs. AD, HC vs. AD). We recruited 58 participants from the Memory Unit of the Hospital Clinico San Carlos (HCSC) in Madrid between January 2023 and January 2024, including 11 HC, 25 with MCI, and 22 AD. Cognitive status was assessed using the Mini-Mental State Examination (MMSE). A Chi-square test showed no significant association between sex distribution and diagnosis group, $\chi^2(1, N = 47) = 2.09$, $p = 0.148$, and a t-test indicated no significant age differences between HC and MCI groups ($t(36) = 0.65, p > .05$) (Table 3.2). However, a significant difference in MMSE scores was found, indicating lower cognitive function in individuals with MCI compared to healthy controls ($t(36) = 3.38, p < .05$) (Table 3.2).

Table 3.2: Summary of user demographics (Mean ± SD) and Age-MMSE correlations.

Characteristic	**HC** (n=11)	**MCI** (n=25)	**AD** (n=22)	**Total** (n=58)
Age (years)	82.6 ± 2.5	81.4 ± 5.9	79.4 ± 4.1	80.9 ± 5.0
MMSE Score	29.9 ± 0.8	26.0 ± 2.1	23.5 ± 3.6	26.6 ± 3.3
Gender (F/M)	8/3	15/10	17/5	40/18
Age-MMSE Corr.	-0.16	0.28	-0.23	0.15
***p*-value**	.170	.030	.310	.200

3.11.1 Data Collection and Preprocessing

Participants were instructed to draw a house symbol on a Repaper tablet (dimensions: 10.9 inches)[§] with a blank sheet of paper affixed and using a standard pen equipped with an accelerometer. This setup replicated a typical pen-and-paper drawing experience while capturing digital data via Bluetooth to the Repaper app. A total of 58 drawings were collected. The **on-line** data, representing discrete point sequences, were initially saved as SVG files and then converted to JSON format, containing multivariate sequences of `(x,y,t)` points. The **off-line** data, captured as high-resolution images using an HP Color LaserJet Pro scanner, were stored as PDFs, converted to PNG format, and resized to 224x224 pixels to standardize inputs for deep learning (DL) models. This resizing aligns with common computer vision practices for compatibility with pretrained DL models. To further enhance image quality, the Canny edge detector was applied to highlight edges in the scanned images.

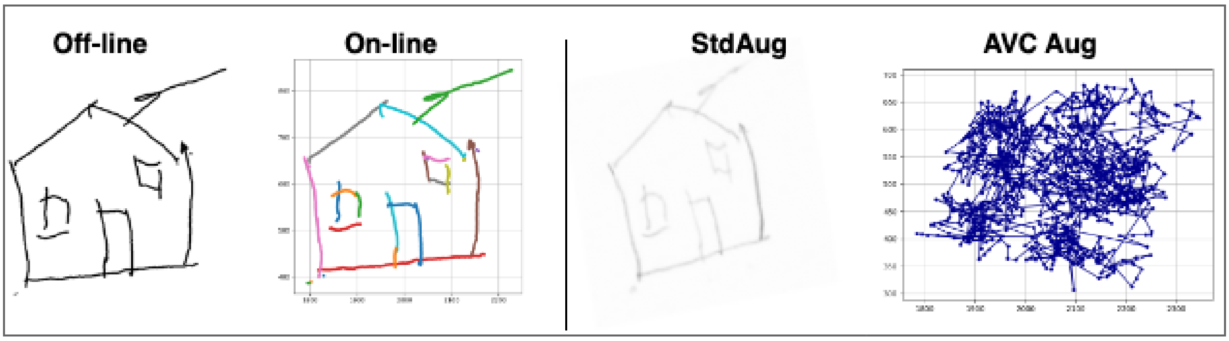


Figure 3.4: Sample of off-line and on-line drawing with standard and AVC augmentation.

Data Augmentation: To improve model robustness and generalizability, we applied DA techniques to both on-line and off-line versions. For off-line version, we used geometric transformations such as rotation, translation, scaling, and flipping to increase variability and reduce overfitting. For on-line version, we employed standard techniques such as jittering and, based on recent findings by the AVC technique proposed by [21]. The AVC included Gaussian noise addition, frame-skipping, spatial modifications, perspective adjustments, and scaling. After DA, the dataset included 300 images (off-line version) and 300 point sequences (on-line version), evenly distributed across 100 observations per group.

[§] `https://www.iskn.co/eu`

3.11.2 Experimental Setup

Our study employs Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to analyze handwriting for cognitive impairment assessment. CNNs are used to analyze pixel-based images, detecting spatial patterns and textures that are indicative of subtle cognitive changes. We use three CNN architectures: ResNet50 [12], which employs skip connections to maintain information across deeper layers; DenseNet121 [15], known for efficient feature propagation through densely connected layers; and EfficientNet [27], which optimizes the network's architecture to handle diverse handwriting styles effectively.

In addition to image-based analysis, RNNs are applied to interpret stroke sequences. This approach captures temporal dynamics and sequential nuances in handwriting, providing insights into the cognitive processes underlying stroke patterns. We implement three types of RNNs: the Bidirectional Vanilla RNN (BiRNN) for straightforward sequential tasks; Bidirectional Long Short-Term Memory (BiLSTM) [13] for retaining information across longer sequences; and Bidirectional Gated Recurrent Unit (BiGRU) [6], which balances computational efficiency with performance.

Training Details: All models were trained using the Adam optimizer with a learning rate of $\eta = 0.001$ and decay rates of $\beta_1 = 0.99$ and $\beta_2 = 0.999$. We used binary cross-entropy as the loss function for all binary classification tasks (HC vs. MCI, MCI vs. AD, and HC vs. AD). We used a batch size of 32 and up to 100 training epochs, with early stopping (patience of 40 epochs) to avoid overfitting. The augmented dataset was split into 80% for training and 20% for testing, ensuring the test set represented unseen data. Stratified 5-fold cross-validation was conducted on the training set to maintain class proportions across folds. Model performance was evaluated using classification accuracy (Acc.) and the Area Under the ROC Curve (AUC).

3.12 Results and Discussion

Our experiments are crucial for understanding the progression of cognitive decline and distinguishing between HC, individuals with MCI, and those with AD. As highlighted in previous research [9, 30], distinguishing MCI from HC and AD can be particularly challenging due to overlapping characteristics.

Table 3.3 summarizes the Accuracy and AUC results for both on-line and off-line datasets, comparing models with and without DA. The data reveal that applying DA, particularly standard DA (StdAug), consistently improves performance across all models and settings.

Off-line Data: EfficientNet demonstrated significant performance gains across all binary classification tasks when standard DA was applied. Specifically, for "HC vs. MCI," accuracy increased from 50%|52% to 65%|66%, for "MCI vs. AD" from 49%|49% to 69%|70%, and for "HC vs. AD" from 53%|55% to 76%|77%. This indicates that off-line data setups benefit substantially from standard DA.

On-line Data: BiGRU was the top performer for on-line data, showing marked improvements post-DA. In the "HC vs. MCI" task, performance increased from 51%|54% to 67%|69%, in "MCI vs. AD" from 47%|45% to 70%|72%, and in "HC vs. AD" from 45%|47% to 75%|76%. While on-line data also benefitted from DA, results varied more between models.

Comparison of DA Techniques: The comparison between standard DA and AVC DA shows that standard DA generally yields higher performance gains. For example, in the "HC vs. AD" group, GRU achieved similar results with standard DA (75%|76%) and AVC DA (68%|70%), but overall, standard DA consistently outperformed AVC across different settings.

Table 3.3: Binary classification results achieved before and after DA (Standard & AVC) for different DL Models.

<table>
<tr><th rowspan="4"></th><th colspan="5">HC vs MCI</th><th colspan="5">MCI vs AD</th><th colspan="5">HC vs AD</th></tr>
<tr><th colspan="2">Off-line</th><th colspan="3">On-line</th><th colspan="2">Off-line</th><th colspan="3">On-line</th><th colspan="2">Off-line</th><th colspan="3">On-line</th></tr>
<tr><th>Before</th><th>StdAug</th><th>Before</th><th>StdAug</th><th>AVCaug</th><th>Before</th><th>StdAug</th><th>Before</th><th>StdAug</th><th>AVCaug</th><th>Before</th><th>StdAug</th><th>Before</th><th>StdAug</th><th>AVCaug</th></tr>
<tr><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th><th>(Acc.|AUC)</th></tr>
<tr><td><b>Res</b></td><td>42 | 48</td><td>57 | 49</td><td>N/A</td><td>N/A</td><td>N/A</td><td>46 | 47</td><td>63 | 65</td><td>N/A</td><td>N/A</td><td>N/A</td><td>51 | 50</td><td>65 | 64</td><td>N/A</td><td>N/A</td><td>N/A</td></tr>
<tr><td><b>Den</b></td><td>49 | 51</td><td>61 | 63</td><td>N/A</td><td>N/A</td><td>N/A</td><td>49 | 50</td><td>60 | 62</td><td>N/A</td><td>N/A</td><td>N/A</td><td>51 | 50</td><td>73 | 77</td><td>N/A</td><td>N/A</td><td>N/A</td></tr>
<tr><td><b>Eff</b></td><td>50 | 52</td><td><b>65 | 66</b></td><td>N/A</td><td>N/A</td><td>N/A</td><td>49 | 49</td><td><b>69 | 70</b></td><td>N/A</td><td>N/A</td><td>N/A</td><td>53 | 55</td><td><b>76 | 77</b></td><td>N/A</td><td>N/A</td><td>N/A</td></tr>
<tr><td><b>RNN</b></td><td>N/A</td><td>N/A</td><td>50 | 51</td><td>61 | 60</td><td>56 | 59</td><td>N/A</td><td>N/A</td><td>50 | 53</td><td>61 | 65</td><td>55 | 59</td><td>N/A</td><td>N/A</td><td>48 | 52</td><td>72 | 75</td><td>60 | 58</td></tr>
<tr><td><b>LSTM</b></td><td>N/A</td><td>N/A</td><td>50 | 55</td><td>65 | 66</td><td>59 | 59</td><td>N/A</td><td>N/A</td><td>47 | 49</td><td>65 | 66</td><td>60 | 57</td><td>N/A</td><td>N/A</td><td>46 | 52</td><td>75 | 75</td><td>59 | 59</td></tr>
<tr><td><b>GRU</b></td><td>N/A</td><td>N/A</td><td>51 | 54</td><td><b>67 | 69</b></td><td>59 | 60</td><td>N/A</td><td>N/A</td><td>47 | 45</td><td><b>70 | 72</b></td><td>61 | 64</td><td>N/A</td><td>N/A</td><td>45 | 47</td><td><b>75 | 76</b></td><td>68 | 70</td></tr>
</table>

The improvements observed in our study align with existing literature that suggests data augmentation can enhance ML model performance by providing more diverse training data, thereby improving generalization [26]. Specifically, our findings underscore that standard DA outperforms more complex techniques like AVC DA, particularly in tasks that require distinguishing subtle cognitive differences, such as between HC and MCI. This suggests that simpler, well-tuned DA methods might be more beneficial for certain medical datasets, where the quality and interpretability of data are paramount [23].

3.12.1 Limitations and Future Work

Our study has some limitations worth of mentioning. Mainly, the small sample size, which is a pervasive problem in medical studies [4, 16, 14], and the focus on a single type of drawing task may limit the generalizability of our findings. Additionally, the study's reliance on a specific neuropsychological test (the HDT) may not fully capture the diversity of cognitive impairments across different populations and tasks. Future research should explore other cognitive assessment tasks to validate further our findings. Despite these limitations, our results hold promise and could pave the way for future clinical applications using a simple handwriting test as a non-invasive, low-cost method.

Acknowlegments

Work supported by the UCM research group (Grupo de Investigación básica en Ciencias de la Visión del IIORC, UCM-GR17-920105), the Horizon 2020 FET program of the European Union (grant CHIST-ERA-20-BCI-001), and the European Innovation Council Pathfinder program (grant 101071147).

References for Paper 4

[1] Asma Bensalah, Antonio Parziale, Giuseppe De Gregorio, Angelo Marcelli, Alicia Fornés, and Josep Lladós. 2023. I Can't Believe It's Not Better: In-air Movement for Alzheimer Handwriting Synthetic Generation. In *Proc. IGS*.

[2] Janice Y. Chan, Benny K. Bat, Anson Wong, Tak-Kit Chan, Zhixuan Huo, Benjamin H. Yip, and Kenneth K. Tsoi. 2021. Evaluation of digital drawing tests and paper-and-pencil drawing tests for the screening of mild cognitive impairment and dementia: a systematic review and meta-analysis of diagnostic studies. *Neuropsychol. Rev.*, 31, 4.

[3] Wen-Ting Cheah, Wei-Der Chang, Jwu-Jia Hwang, Sheng-Yi Hong, Li-Chen Fu, and Yu-Ling Chang. 2019. A screening system for mild cognitive impairment based on neuropsychological drawing test and neural network. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 3543–3548.

[4] Hung-Yu Chen, Der-Chiang Li, and Liang-Sian Lin. 2016. Extending sample information for small data set prediction. In *Proc. IIAI-AAI*.

[5] Shuqing Chen, Daniel Stromer, Harb Alnasser Alabdalrahim, Stefan Schwab, Markus Weih, and Andreas Maier. 2020. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Sci. Rep.*, 10, 1.

[6] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proc. EMNLP*.

[7] Nicole Cilia, Tiziana D'Alessandro, Claudio De Stefano, and Francesco Fontanella. 2022. Deep transfer learning algorithms applied to synthetic drawing images as a tool for supporting Alzheimer's disease prediction. *Mach. Vis. Appl.*, 33.

[8] Nicole D Cilia, Tiziana D'Alessandro, Claudio De Stefano, Francesco Fontanella, and Mario Molinara. 2021. From online handwriting to synthetic images for alzheimer's disease detection using a deep transfer learning approach. *IEEE Journal of Biomedical and Health Informatics*, 25, 12, 4243–4254.

[9] Zihan Ding, T. Lee, and A. Chan. 2022. Digital cognitive biomarker for mild cognitive impairments and dementia: a systematic review. *J. Clin. Med.*, 11.

[10] Josep Garre-Olmo, Marcos Faúndez-Zanuy, Karmele López-de-Ipiña, Laia Calvó-Perxas, and Oriol Turró-Garriga. 2017. Kinematic and pressure features of handwriting and drawing: preliminary results between patients with mild cognitive impairment, alzheimer disease and healthy controls. *Curr. Alzheimer Res.*, 14, 9.

[11] Peyvand Ghaderyan, A. Abbasi, and Sajad Saber. 2018. A new algorithm for kinematic analysis of handwriting data; towards a reliable handwriting-based tool for early detection of alzheimer's disease. *Expert Syst. Appl.*, 114.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.*, 9, 8.

[14] Nina Hosseini-Kivanani et al. 2024. Ink of insight: data augmentation for dementia screening through deep learning. In *Proc. ICMHI.* Japan, Yokohama.

[15] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Proc. CVPR.*

[16] Donato Impedovo and Giuseppe Pirlo. 2018. Dynamic handwriting analysis for the assessment of neurodegenerative diseases: a pattern recognition perspective. *IEEE Rev. Biomed. Eng.*, 12.

[17] Ko Woon Kim, Sung Yun Lee, Jongdoo Choi, Juhee Chin, Byung Hwa Lee, Duk L Na, and Jee Hyun Choi. 2020. A comprehensive evaluation of the process of copying a complex figure in early- and late-onset alzheimer disease: a quantitative analysis of digital pen data. *J. of medical internet research*, 22, 8, e18136.

[18] Patricia Knechtl and Johann Lehrner. 2023. Visuoconstructional abilities of patients with subjective cognitive decline, mild cognitive impairment and alzheimer's disease. *J. Geriatr. Psychiatry Neurol.*

[19] Masahiro Kobayashi, Yuya Yamada, Koji Shinkawa, Mitsuhiro Nemoto, Kiyoshi Nemoto, and Takeo Arai. 2022. Automated early detection of alzheimer's disease by capturing impairments in multiple cognitive domains with multiple drawing tasks. *J. Alzheimer's Dis.*, 88, 3.

[20] Jeffrey L. Liss, Sara Seleri Assunção, Jeffrey Cummings, Alireza Atri, David S. Geldmacher, Sheila F. Candela, and Marwan N. Sabbagh. 2021. Practical recommendations for timely, accurate diagnosis of symptomatic alzheimer's disease (mci and dementia) in primary care: a review and synthesis. *J. Intern. Med.*, 290, 2.

[21] Mykola Maslych, Eugene Matthew Taranta, Mostafa Aldilati, and Joseph J. Laviola. 2023. Effective 2D Stroke-based Gesture Augmentation for RNNs. In *Proc. the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23).

[22] Fredrik Öhman, Jens Hassenstab, Daniel Berron, Michael Schöll, and Katherine V. Papp. 2021. Current advances in digital cognitive assessment for preclinical alzheimer's disease. *DADM*, 13, 1.

[23] Luis Perez and Jason Wang. 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv:1712.04621.*

[24] Amir Poreh, Jennifer B Levin, and Max Teaford. 2020. Geriatric complex figure test: a test for the assessment of planning, visual spatial ability, and memory in older adults. *Applied Neuropsychology: Adult*, 27, 2, 101–107.

[25] Isabelle Rouleau, David P Salmon, and Nelson Butters. 1996. Longitudinal analysis of clock drawing in alzheimer's disease patients. *Brain Cognit.*, 31, 1.

[26] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *J. Big Data*, 6, 1.

[27] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proc. ICML.*

[28] Luigi Trojano and Guido Gainotti. 2016. Drawing disorders in alzheimer's disease and other forms of dementia. *J. Alzheimer's Dis.*, 53, 1.

[29] Maria Tsatali, Konstantina Avdikou, Menelaos Gialaouzidis, Despina Minopoulou, Alexandra Emmanouel, Eleni Kouroundi, and Magda Tsolaki. 2022. The discriminant validity of rey complex figure test (rcft) in subjective cognitive decline, mild cognitive impairment and alzheimer's disease dementia in greek older adults. *Appl. Neuropsychol. Adult.*

[30] P. Werner, S. Rosenblum, G. Baron, J. Heinik, and A. Korczyn. 2006. Handwriting process variables discriminating mild alzheimer's disease and mild cognitive impairment. *J. Gerontol. Series B*, 61, 4.

[31] Anders Wimo, Katrin Seeher, Renata Cataldi, Elena Cyhlarova, Joseph L. Dielemann, Ola Frisell, and Tarun Dua. 2023. The worldwide costs of dementia in 2019. *Alzheimer's Dement.*, 19, 7.

[32] Fang Xu, Yue Ding, Zongheng Ling, Xinyu Li, Yixue Li, and Shibiao Wang. 2020. DCDT: a digital clock drawing test system for cognitive impairment screening. In *Proc. ICDE.*

[33] Young Chul Youn et al. 2021. Use of the clock drawing test and the rey–osterrieth complex figure test-copy with convolutional neural networks to predict cognitive impairment. *Alzheimer's Res. Therapy.*

[34] X Zhang, Y Zhao, L Lv, G Min, Q Wang, and Y Li. 2021. A study on the performance characteristics and diagnostic efficacy of digital clock drawing test in patients with amnesic mild cognitive impairmen. *Chin. J. Behav. Med. Brain Sci*, 30.

3.13 Predicting Alzheimer's disease and mild cognitive impairment with off-line and on-line house drawing tests

Abstract[†]

There is growing interest in developing reliable, non-invasive, and cost-effective methods for early diagnosis of neurodegenerative diseases such as Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD). In this regard, handwriting-based tasks have shown potential in differentiating MCI and AD patients from healthy controls (HCs). However, previous work has reported mixed results when using different symbols and data representations. We address this research gap by developing computational models (convolutional and recurrent neural networks) to differentiate MCI and AD from HCs with off-line (scanned images) and on-line (discrete time series) house drawings. Notably, we observed that augmenting on-line data and then converting it to off-line format, a method we refer to as "OnOff-line", yielded the best performance results in binary classification tasks. These findings highlight the effectiveness of on-line representations in capturing handwriting dynamics more accurately. Ultimately, our work opens new avenues for future research to enhance automated diagnostic of MCI and AD from handwriting analysis.

Keywords

Mild Cognitive Impairment, Alzheimer's Disease, Handwriting, Drawing, Deep Learning, Classification.

3.14 Introduction

Alzheimer's Disease (AD) is the primary cause of dementia, leading to substantial cognitive and behavioral decline [1], and is expected to impact up to 152 million individuals worldwide by 2050 [5]. The early detection and understanding of Mild Cognitive Impairment (MCI), which often precedes dementia, has garnered considerable attention from both researchers and healthcare professionals, as it represents a pivotal phase between healthy aging and AD. MCI is characterized by mild cognitive symptoms (e.g., memory

[†] **N. Hosseini-Kivanani**, E. Salobrar-García, L. Elvira-Hurtado, M. Salas, C. Schommer, and L. A. Leiva, "Predicting Alzheimer's disease and mild cognitive impairment with off-line and on-line house drawing tests," in *Proceedings of the 2024 IEEE 20th International Conference on e-Science*, 2024, pp. 1–10. DOI: 10.1109/e-Science62913.2024.10678661

and thinking skills) that do not significantly impede daily life activities. Although not all MCI cases progress to dementia, there is an elevated risk of evolving into AD [19], particularly for people who experience memory deficits [38, 53]. Furthermore, the availability of specific diagnostic tests can be constrained by factors such as economic limitations, healthcare infrastructure, or geographical location [28]. These restrictions can adversely affect the prompt diagnosis and intervention for individuals with cognitive impairments. This limitation underscores the need for alternative, patient-friendly diagnostic approaches.

Current research and technological advancements are focused on improving the accessibility and accuracy of diagnostic tools. Given the challenges and inaccuracies associated with traditional (manual) diagnosis methods for MCI and AD, there is growing interest in using Machine Learning (ML) to enhance these processes. For example, ML can identify biomarkers such as neurofibrillary tangles and senile plaques, which correlate highly with specific structural changes in the brain caused by AD [48]. This shift towards technology-driven diagnostics aims to streamline the early detection of AD and MCI, facilitating timely interventions that could potentially slow the progression of these neurodegenerative diseases.

Handwriting analysis has seen substantial growth recently, following recent advancements in neuroscience, and in particular, it has proven useful in diagnosing AD due to its ability to detect changes in cognitive and motor skills [20, 15, 69]. The community has acknowledged handwriting-based tests[‡] as a viable, non-invasive method for early detection of cognitive decline [16, 13, 17], a necessity in light of increasing AD prevalence and an aging population.

In the medical field, it is widely recognized that handwriting deterioration (e.g., irregular size, spacing, and letter formation) is one of the earlier indicators of cognitive disorders. This deterioration stems from the impact of cognitive diseases on motor skills, which involve cognitive, kinesthetic, and perceptual-motor abilities. Therefore, observing changes in handwriting can serve as a critical early sign of cognitive decline, highlighting the need for early diagnosis and interventions to mitigate the severity of these conditions [35].

The choice of symbols for handwriting/drawing tests significantly impacts the accuracy and utility of the results. The Clock Drawing Test (CDT), widely used for cognitive evaluation, has been shown to require large sample sizes to yield reliable results [31]. The Pentagon drawing test (PDT), which measures visuospatial abilities through a copy task, is not well-suited for detecting broader cognitive impairments beyond spatial skills [33]. Sentence-writing tasks [21] are language-dependent, which poses a challenge in multilingual settings or among individuals with language deficits, and handwritten signatures do

[‡] In this paper, we consider *handwriting* and *hand-drawing* synonymous because both tasks involve the same neurophysiological and peripheral processes involved in motor control.

not perform well for AD screening [32, 23].
Both the Tree Drawing Test (TDT) [59, 61] and the HDT [22] are increasingly being used by clinicians nowadays. These tests evaluate a range of cognitive functions, including planning, organization, spatial awareness, and motor control [22]. Despite their potential, computational modeling using these symbols remains unexplored. The TDT, though useful, presents certain challenges. Trees can vary greatly in their structure and complexity, which may introduce variability in the interpretation of results. In contrast, the HDT offers a more standardized and relatable symbol. Drawing a house typically involves a combination of straight lines, angles, and geometric shapes, which can effectively assess visuoconstructional abilities and cognitive function without the variability introduced by more complex symbols. It is simple to administer, language independent and captures a wide range of cognitive abilities. Given these advantages, we develop computational models using HDT data in this paper.
Another area that remains underexplored is the use of different input representations that can improve model performance. Namely, the vast majority of previous work uses scanned images (off-line representations) for AD classification [36, 31, 32, 44, 10]. Only a handful of papers have explored the use of time series data (on-line representations) in this regard [2, 16]. We argue that on-line data can capture better handwriting dynamics; however, in the medical domain, these kinds of datasets are really small, so ML models tend to overfit [25]. This is arguably the main reason why previous work has only focused on off-line data. In this paper, we explore a novel approach: convert on-line data to off-line data. This enables the combination of fine-grained handwriting dynamics with static image data, for which computer vision DL models have proved useful for AD classification.
In sum, this paper makes the following contributions:

- Comparison of Deep Learning (DL) models for classification of AD, MCI, and HCs using both On-line and Off-line HDTs.
- On-line to off-line conversion: a novel approach where time series data are encoded into (pixel-based) intensity values as images.

3.15 Related Work

ML in digital medicine has recently shown considerable promise in enhancing healthcare outcomes and improving the efficiency of the diagnostic process [2, 7, 12, 42]. ML algorithms have been used to classify MCI, AD, and cognitive normal groups using neurocognitive tests, with high area under the ROC curve (AUC) values, indicating strong predictive performance [50]. Handwriting analysis, particularly in the context of AD and

MCI, offers valuable insights for screening and diagnosis (e.g., [39]). Recent studies have identified distinct patterns in the early stages of AD through handwriting, demonstrating the efficacy of drawing skills as indicators of cognitive decline [56, 18, 69, 24].

Handwriting data can be captured either as scanned images, also known as *off-line* data, or as discrete sequences of $\{x, y\}$ (sometimes $\{x, y, t\}$) points, also known as *on-line* data. The former has been the focus of most of the previous work in AD screening [36, 12]. The latter has gained attention only recently [2, 16], which is surprising because, in other domains, it has been shown that on-line data representations provide richer movement dynamics, including, for example, detailed timing and (sometimes) pressure information [41, 49, 70, 35], which are features not available in off-line data representations.

When combined with cognitive functioning assessments, handwriting kinematic measures can differentiate between MCI, AD, and healthy controls (HCs). The reported classification accuracy ranges from 69% to 72% in differentiating participants, although the classification accuracy for the MCI group alone is relatively poor [69]. Müller et al. [46] reported that the digital Clock Drawing Test (dCDT) has a higher diagnostic accuracy for discriminating MCI patients from HCs compared to the conventional CDT (cCDT) (81.3% vs. 62.5%). Robens et al. [58] used the digital Tree Drawing Test (dTDT), achieving 77% AUC when discriminating MCI from HCs and 90% AUC when discriminating AD from HCs. Faundex et al. [22] used the House Drawing Test to analyze (not classify) handwriting movements in AD patients and HCs. They noted that on-line data revealed subtler motor impairments that traditional off-line methods would miss.

A study by Garre et al. [25] investigated handwriting and drawing copy-tasks involving several symbols (e.g., spiral, house, or crossed pentagons) to differentiate between AD, MCI, and HC. They found that kinematic features such as pen tip velocity and pressure could classify participants based on their cognitive status with varying accuracy, from 63.5% (drawing of a 3D house and CDTs) to 100% (drawing of a spiral). Unfortunately, the spiral symbol does not include considerations of fine-grained details and facets of spatial awareness, planning, and memory, all of which are particularly affected in MCI and AD patients.

El-Yacoubi et al. [70] tried to identify cognitive states based on handwriting characteristics (e.g., velocity, acceleration, or stroke length) when copying predefined sentences. Their study revealed distinct clusters that corresponded to different cognitive profiles. For example, one cluster was dominated by HC and MCI patients, while another was dominated by MCI and early-stage AD patients. This clustering highlighted that MCI patients exhibited handwriting behaviors that were intermediate between HCs and early-stage AD patients. Another study by Raksasat et al. [57] developed an Attentive Pairwise Interaction Network (API-Net) aimed at enhancing the automatic scoring of the CDT. They achieved 79% F1-score, slightly outperforming a convolutional neural network (ResNet-152 by 3%) for multi-class classification.

As shown before, there is no consensus about which performance metric should be reported and which handwriting task should be administered. Further, very few datasets are available for AD and MCI screening, mostly in off-line form. To address these issues, we collected a comprehensive dataset [§] featuring HDTs, drawn by MCI and AD patients as well as HC, using *both* on-line and off-line data. Our dataset is, therefore, unique in the sense that it encompassed off-line and on-line representations of the same drawing from the same subject. Accordingly, we compare and contrast the classification performance of various computational models (neural networks) using either off-line or on-line handwriting to detect MCI and AD. We also investigate a novel approach, which consists of converting on-line data to off-line, which turned out to outperform models trained from scratch on either on-line or off-line data. Taken together, our results provide new insights into the performance of ML models when classifying MCI and AD patients using different handwriting data representations.

3.16 Experimental Setup

Digital pens bring new challenges, including increased costs and a potential compromise in handwriting input due to variations in pen grip and user familiarity with digital interfaces [43]. To address these concerns, we used a Repaper tablet,[¶] featuring conventional pencils equipped with a small accelerometer, which allowed us to replicate the natural paper-based handwriting experience while at the same time capturing discrete time series data of $\{x, y\}$ points. The on-line data capture of these $\{x, y\}$ points is crucial as it provides a continuous, real-time digital record of the writing process, enabling detailed analysis of handwriting trajectories. These trajectories are particularly meaningful in our study as they offer insights into neuromotor control in elderly populations, helping to differentiate between normal aging processes and specific impairments associated with MCI and AD.

In addition to on-line data, off-line data captures static end results of the handwriting, such as the final drawing. The key difference between off-line and on-line data is that while off-line data provides a snapshot of the final product, on-line data captures the process. By comparing both off-line and on-line versions from the same patient, we can gain a comprehensive understanding of both the result and the process of handwriting, which is essential for diagnosing and differentiating between MCI and AD.

[§] Available upon reasonable request.

[¶] `https://www.iskn.co/`

3.16.1 Participants

Table 3.4: Demographic information of our dataset.

	HC	MCI	AD
	(n = 11)	(n = 25)	(n = 22)
Gender (male + female)	3 + 8	15 + 10	5 + 17
Age (M ± SD)	82.63 ± 2.46	81.44 ± 5.89	79.36 ± 4.09
MMSE (M ± SD)	29.91 ± 0.83	27.60 ± 2.18	23.45 ± 3.57

We recruited 58 individuals aged between 70 and 89 years at the Memory Unit of the Hospital Clinico San Carlos (HCSC) in Madrid, including 25 patients with MCI, 22 patients with AD, and 11 HC; see Table 3.4 and Figure 3.5. All individuals had normal vision and no hearing problems. Cognitive capabilities were assessed in a clinical setting optimized to minimize distractions and background noise. This assessment used the Mini-Mental State Examination (MMSE) [63] to evaluate cognitive status. Patients were classified into MCI or AD based on criteria from the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS), Alzheimer's Disease and Related Disorders Association (ADRDA) [45], and the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [27]. Statistical analysis with one-way ANOVA tests revealed no significant age differences between the three groups ($F(2, 55) = 2.04, p > .139$), indicating a balanced distribution across participants. Conversely, significant differences in MMSE scores were found among the groups ($F(2, 55) = 25.64, p < .0001$), with the HC group scoring higher than the MCI and AD groups ($p < .0001$), as expected.

Figure 3.5 presents the distribution of (a) age and (b) MMSE scores for the three groups. The HC group displays a broader age range with a slightly higher median age compared to the MCI and AD groups. Notably, the statistical test indicates no significant difference ($p > .05$) in age distribution among the groups, suggesting that age alone does not allow to differentiate between these groups. The boxplots reveal a clear descending trend in median MMSE scores from HC to AD, indicating a progressive cognitive decline. The MCI scores lie between those of the HC and AD groups, consistent with their intermediary diagnostic status. The difference across groups was statistically significant ($p < .001$).

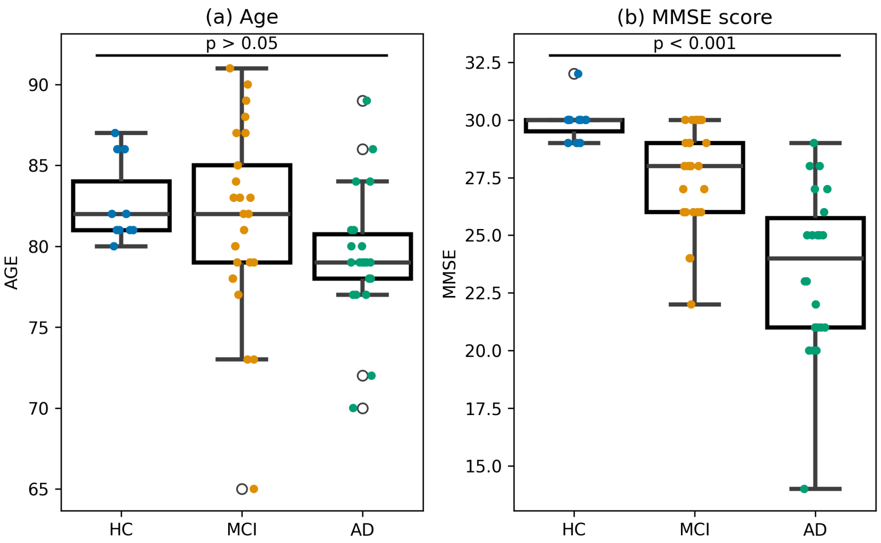


Figure 3.5: Data distribution of (a) age and (b) MMSE score of HC, MCI, and AD groups.

3.16.2 Procedure

Participants were tasked to draw a house with the instrumented pencil (1 drawing per participant, totaling 58 drawings). The Repaper tablet simultaneously recorded off-line and on-line data. The on-line data were stored as SVG files (default format in Repaper) and then converted to JSON files for later post-processing. The off-line data were digitized with an HP Color LaserJet Pro scanner, stored as PDF files (default format in LaserJet), and then converted into PNG files for later post-processing. The PNG files were enhanced with the Canny edge detector [4] and resized to a standard resolution of 224×224 px.

3.17 Modeling Methodology

Figure 3.6 shows the full system pipeline. As explained later, we use Convolutional Neural Networks (CNNs) to classify off-line data and Recurrent Neural Networks (RNNs) to classify on-line data. We chose CNNs and RNNs for our tasks because of their effectiveness in analyzing complex drawing tasks such as the CDT [31, 37], HDT [22] and PDT tasks [33], which are commonly used to assess cognitive impairment [8] for both off-line and on-line data. CNNs are particularly effective for processing off-line data, such as pixel-based images, where they can detect spatial patterns and intricate details within the static drawings, essential for identifying signs of cognitive decline. On the other hand, RNNs, especially Long Short-Term Memory networks (LSTMs) and Gated Recurrent Units (GRUs), excel in analyzing on-line data by capturing the sequential nature of stroke movements, which is critical for understanding the temporal progression in tasks like drawing tests. This approach allows us to capitalize on the specific strengths of CNNs for off-line image analysis and RNNs for on-line sequential data, providing a comprehensive assessment of cognitive function.

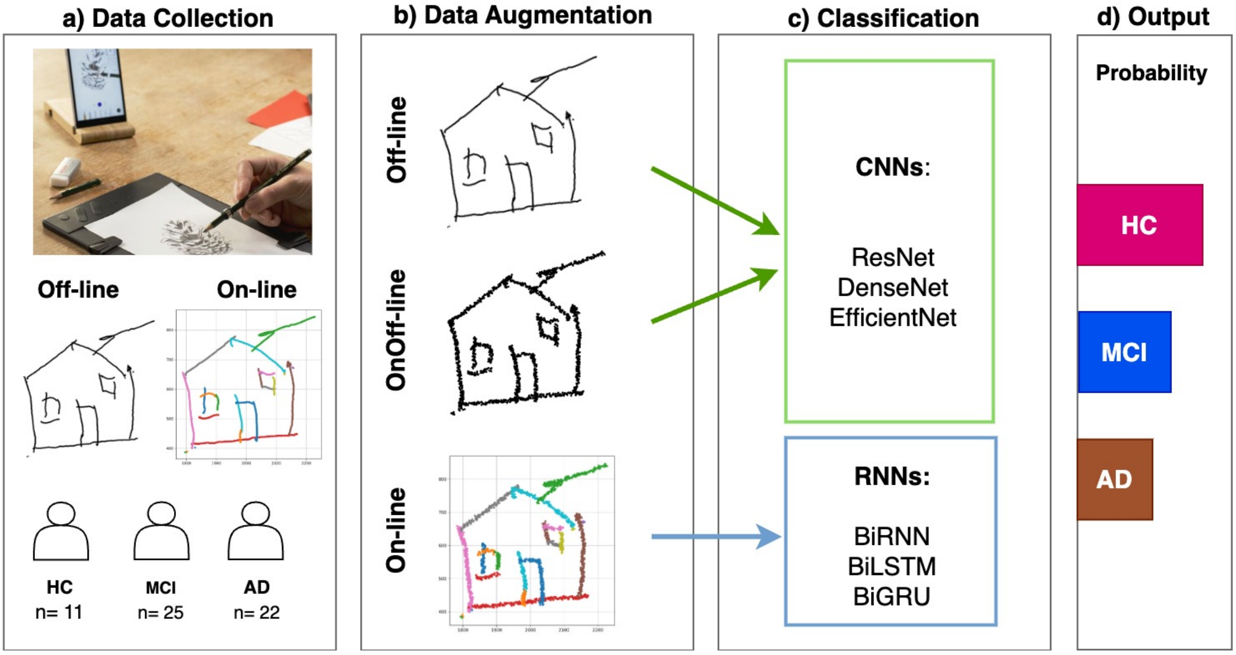


Figure 3.6: Full system pipeline, from data collection and preprocessing to model training and classification.

3.17.1 Data Augmentation

Data augmentation is essential to increase the robustness and generalizability of ML models, especially in digital medicine, where sample sizes are quite often too small for today's ML standards. We considered three versions of data augmentation (Figure 3.6b), as follows.

Off-line version: We applied a series of standalone geometric transformations to the images, such as scaling and small rotations, in a selective manner to preserve the image semantics (e.g., no vertical flipping). For this, we used the Albumentations library.[‖]

On-line version: The discrete point sequences were modified according to commonly used transformation methods for time series, such as jittering and scaling [64].

OnOff-line version: The augmented on-line data were converted to off-line data (PNG format), following the same steps indicated in Section 3.16.2. This variant is a compromise solution between on-line and off-line approaches; the idea is to leverage the potentially

[‖] https://albumentations.ai/

larger variability produced by time series transformations and combine it with the high performance of pre-trained CNNs.

Thanks to data augmentation, we expanded our dataset to include 300 images (off-line representations) and 300 sequences (on-line representations), ensuring a balanced representation across categories. As an example, for the MCI category, we generated 75 unique variations from the initial set of 25 house drawings.

While augmented data should introduce variability, it should still retain the essential features of the original data to be useful for training. To ensure this, the quality and variability of the augmented dataset were rigorously assessed using the Structural Similarity Index Measure (SSIM) [67] for off-line data and the Dynamic Time Warping (DTW) [60] metric for on-line data. SSIM is a similarity metric where 0 indicates no similarity, and 1 means full similarity. SSIM is commonly used to measure the visual similarity between images by considering changes in structural information, luminance, and contrast. DTW is a distance metric where values greater than 0 indicate deviation from full similarity. DTW, on the other hand, compares time-series data by aligning sequences in a way that minimizes the cumulative distance.

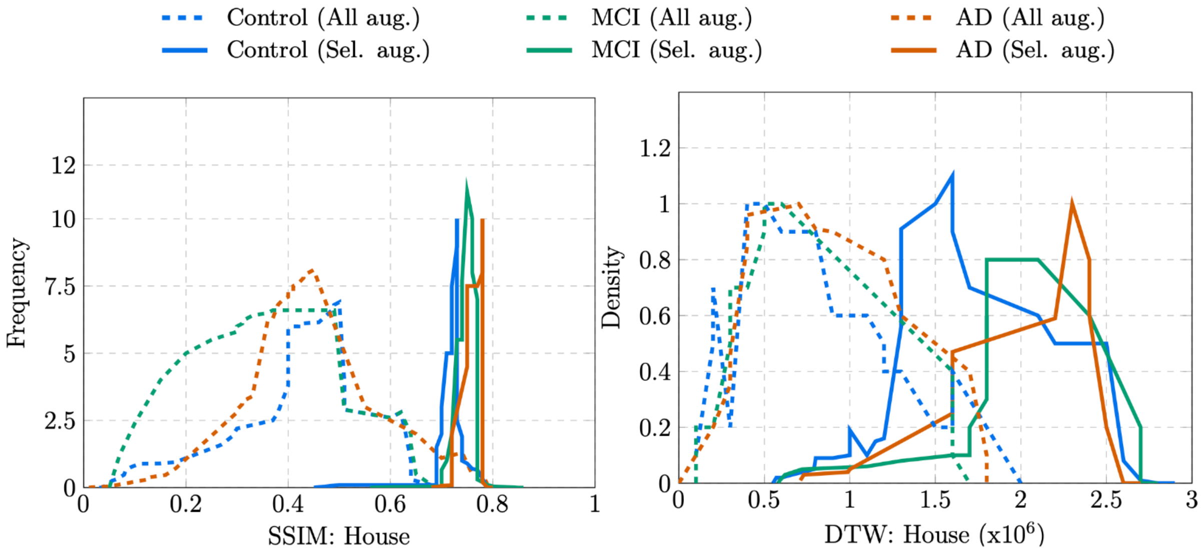


Figure 3.7: Distributions of SSIM and DTW distributions across HC, MCI, and AD. Dashed plots correspond to the results considering all the augmentation techniques collectively (labeled as “All aug.”). The solid plots show the results from the selected augmentation (labeled as “Sel. aug.”) techniques.

Note that, in this context, “similarity” refers to maintaining the structural integrity of the original data while introducing the desired variations. The goal of using similarity measures in the context of data augmentation is not to create augmented data identical

to the originals but rather to ensure that the transformations do not lead to the loss of important features or introduce unrealistic characteristics. Too high SSIM or DTW values might indicate insufficient augmentation, whereas too high SSIM values might indicate that the augmented data is too similar to the original, suggesting insufficient variability and, therefore, potentially less effective augmentation. Conversely, SSIM values that are too low might indicate excessive distortion, where the structural integrity of the data is compromised. Similarly, higher values suggest greater deviation from the original time-series patterns for DTW, while very low values could imply that the augmentation did not introduce meaningful variations. Thus, both SSIM and DTW provide critical insights into the balance between maintaining essential features and introducing sufficient variability, ensuring that the augmented data remains valid for model training.
Figure 3.7 represents the plots for SSIM and DTW. With SSIM scores ranging between 0.7 and 0.75 (see Figure 3.7 b) and DTW ranging from 180 to 4987 (see Figure 3.7b), we confirm that the augmented data comprise novel variations rather than mere replicas of the originals.

3.17.2 Convolutional Neural Networks

CNNs are inspired by the hierarchical structure of the human visual cortex [40] and are widely used for image classification tasks in digital medicine (e.g., [9]). We selected three state-of-the-art CNNs pre-trained on the popular ImageNet dataset, which offers a vast range of images across multiple categories:[**] ResNet50 [29], DenseNet121 [34], and EfficientNet [62]. ResNet uses skip connections to allow gradients to flow through the network directly, preventing the vanishing gradient problem and enabling the training of very deep networks. DenseNet features a unique architecture where each layer is connected to every other layer in a feed-forward fashion, significantly reducing the number of parameters and enhancing feature propagation. EfficientNet scales up CNNs in a more structured manner using a compound coefficient to ensure that depth, width, and resolution grow uniformly. These models were fine-tuned to our dataset using transfer learning [68].

3.17.3 Recurrent Neural Networks

RNNs are preferred to handle sequential data, where understanding spatiotemporal dynamics is important [26]. Since no pre-trained RNN models for sequence classification are currently available as open source, we designed three models from scratch, each based on one type of RNN memory cell.

[**] https://www.image-net.org/

Bidirectional vanilla RNN (BiRNN), (with no memory), which operates without the use of gating mechanisms, makes it simpler and faster for tasks where long-term dependencies are less critical. This model analyzes the sequence dynamics in a straightforward manner, though it may struggle with longer sequences due to the vanishing gradient problem.

Bidirectional Long Short-Term Memory (BiLSTM) [30] uses LSTM units to capture long-range dependencies within the data effectively. This model is particularly adept at handling the challenges of sequence classification, where understanding across large time lags is crucial. The bidirectional architecture enhances its capability to integrate context from both past and future inputs, providing a robust analysis of the sequence's temporal features.

Bidirectional Gated Recurrent Unit (BiGRU) [11] uses GRU cells that streamline the architecture of LSTMs while retaining their ability to manage long-term dependencies. GRUs simplify the gating mechanism found in LSTMs, leading to faster training times without a significant trade-off in performance. Like BiLSTM, the bidirectional approach allows the BiGRU to glean comprehensive insights from both directions of the sequence, enhancing its predictive accuracy in complex scenarios.

Generally, bi-directionality allows the models to analyze an input sequence in the forward and backward directions, offering a more comprehensive understanding of the sequence's temporal features. Our three RNNs have one hidden layer comprising 128 neurons selected through Bayesian optimization with the *Keras Tuner* library [51], hyperbolic tangent as an activation function, and a dropout rate of 0.1 to prevent overfitting. This setup precedes a softmax output layer, ensuring a probabilistic distribution over the classification labels as model output (Figure 3.6d).

3.17.4 Training Procedure

We split our augmented dataset as disjoint partitions of: 80% training, and 20% testing. The test partition simulates unseen data, as it is only used for final model evaluation. We train our CNN and RNN models using the Adam optimizer with a learning rate of $\eta = 0.001$ and momentums $\beta_1 = \beta_2 = 0.99$. The loss function is binary cross-entropy for two-class classification tasks (in our case, two categories: HC and patient) and categorical cross-entropy for multi-class classification tasks (in our case, three categories: HC, MCI, and AD). We use a batch size of 32 and train each model for up to 100 epochs with early stopping (patience of 40 epochs, meaning that if validation loss does not improve over 40 consecutive epochs, training stops, preserving the optimal model weights).

To ensure a consistent proportion of samples across different classes, we used a stratified 5-Fold Cross-Validation, a type of K-Fold Cross-Validation that splits the entire set into

k number of folds. The number of folds, k, was set to 5. Consequently, the training set–which represents 80% of the whole dataset–was split into 5 folds. The first fold was used as the validation set, while the remaining 4 folds served as the training set. This process was repeated 5 times to guarantee that the entire set was used for both training and validation purposes.

3.17.5 Evaluation Metrics

To assess the performance of our models, we compute classification accuracy (Acc) and Area Under the Receiver Operating Characteristic curve (AUC). Together, these two metrics serve as fundamental tools in the assessment of predictive models, particularly in applications where the balance between sensitivity (true positive rate) and specificity (true negative rate) is crucial.

Accuracy is defined as the percentage of true cases (true positives and true negatives) that are correctly identified relative to the total number of cases. This metric offers a straightforward measure of a model's overall performance in correctly predicting outcomes. The AUC, on the other hand, provides insight into the discriminative power of any classifier. It is calculated by plotting the true positive rate against the false positive rate and measuring the area under the resulting curve. The true positive rate quantifies the model's ability to correctly identify actual positives, while the false positive rate measures how often the model incorrectly classifies negatives as positives. If AUC = 50%, the classifier is no better than random guessing, highlighting its ineffectiveness.

3.18 Results and Discussion

We first evaluated the performance of the proposed CNN and RNN models, with and without data augmentation, for the three data representations considered (Off-line, OnOff-line, and On-line). Figure 3.8 shows that data augmentation yields notable improvements in all cases. Without it, most models behave like a random classifier, especially when considering binary classification tasks; see the leftmost plots in Figure 3.8. This emphasizes that data augmentation is essential to train competent computational models.

Subsequently, we found that both jittering and scaling were particularly effective in improving classification accuracy for models that rely on on-line data representations. This contradicts previous findings in PD screening, where jittering failed to improve performance due to the introduction of noise that mimicked dyskinesia charateristics [64]. It is important to note that time series data should not be excessively modified to avoid significant distortions after data augmentation. Contrary to studies in PD (e.g., [64]), our methods proved effective, suggesting that the type of data augmentation and the nature of the disease significantly influence model outcomes.

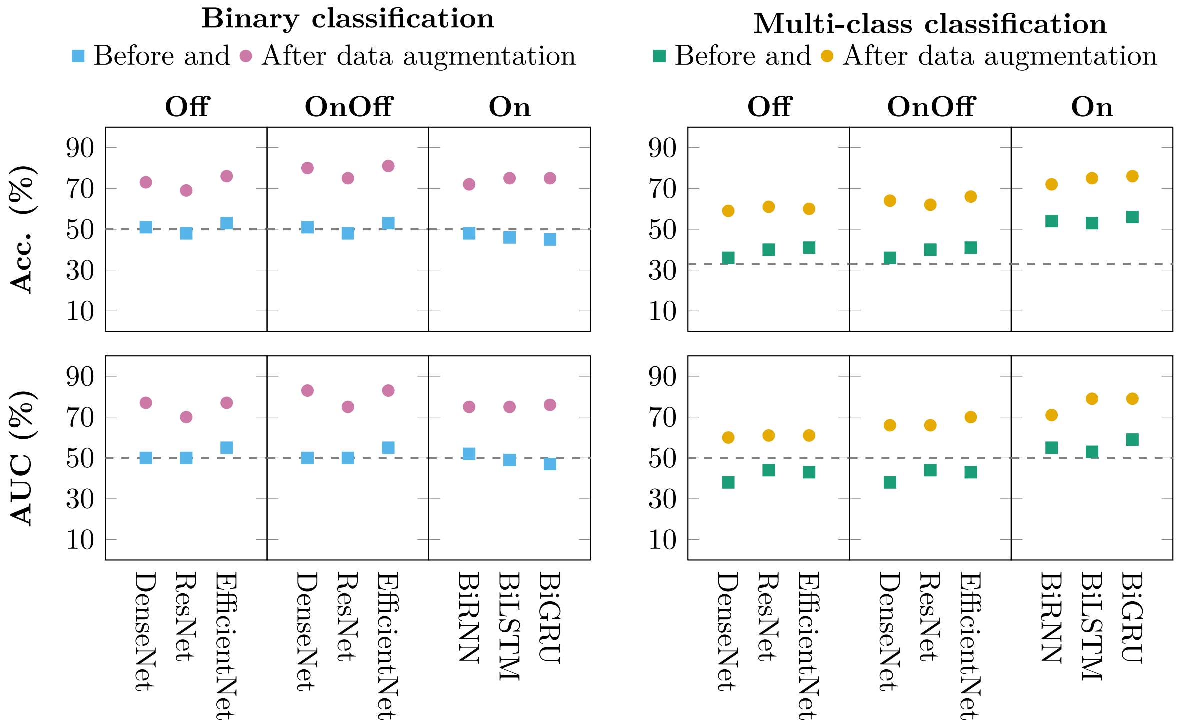


Figure 3.8: Classification accuracy (top row) and AUC scores (bottom row) using different models. Dashed lines indicate the performance of a random classifier, serving as an empirical lower bound for comparison.

Our study aims to go beyond merely achieving accurate and high-performance classification results; it seeks to understand the effectiveness of different input data formats as novel diagnostic methods. Overall, our CNN and RNN models succeeded in distinguishing AD patients from HC or MCI, outperforming random guessing by a large margin. Interestingly, our experiments achieved the best results in binary classification tasks when augmenting on-line (time series) data and then converting them to off-line (images) data, referred to as OnOff-line in our experiments. To the best of our knowledge, this is the first work to report this finding.

Conversely, RNNs, specifically BiGRU, achieved the best performance for multi-class classification tasks. In binary classification tasks, MCI and AD patients are considered in the same group, which may introduce some noise and ambiguities in the multi-class case, affecting the BiGRU model's performance. However, for multi-class classification tasks, where HC, MCI, and AD are treated as separate classes, BiGRU models can effectively use their architecture to distinguish between these groups more accurately. BiGRU worked best for multi-class classifications and performed as well as EfficientNet for binary classifications, making it the most versatile model for our dataset.

CNNs have previously been recognized for their high accuracy in predicting the conversion from MCI to AD [56, 71]. ML models have also been used to detect the progression of AD stages. For example, Bucholc et al.[6] introduced a hybrid Random Forest model that achieved 87.5% accuracy; however, the reported performance varied across different measures (e.g., MRI, age, and cognitive measures). In contrast, our models deliver higher accuracy across classification tasks and are robust across different data formats. Another study by Piers et al.[54] used on-line data to examine neurocognitive behavior over time, demonstrated the utility of digital pen technology in cognitive evaluation. In line with these observations, our study shows that on-line data representations are preferred, even when converted to off-line data.

Table 3.5 presents the classification results from various studies involving handwriting analysis to diagnose AD, using both off-line and on-line representations and across different datasets. Where applicable, each model's performance is evaluated under binary and multi-class classification scenarios. When using off-line data representations, EfficientNet models on pentagon drawings achieved the highest binary classification accuracy. This is followed closely by the InceptionResNetV2 models on letter drawings, which also show robust performance. In multi-class classification, where the complexity of class differentiation increases, all models generally show reduced performance, exemplifying the challenges posed by more complex classification tasks. On-line data representations performed best in this case.

Classification accuracy varies notably across the different drawings and formats, underscoring the influence of drawing complexity and data format on model performance. Our findings suggest that simpler shapes may facilitate higher accuracy in binary classification due to fewer complexities distinguishing ADs from HCs. However, as the task complexity increases in multi-class scenarios, where the model must differentiate between multiple stages or types of cognitive impairment, model performance generally declines. This is evident in the more complex "Clock" and "Letter" drawings, which involve more intricate details and potentially more variation in individual execution.

The MCI group showed characteristics that were intermediate between HCs and ADs, which made it challenging to distinguish them clearly, as it was also stated in Werner et al. work [69]. However, our study demonstrated better performance in distinguishing MCI from HC and AD with higher accuracy in our multi-class classification task, with an accuracy of (See Table 3.5). Compared with previous work, several factors of our study stand out:

1. **Data augmentation**: Our results demonstrate that data augmentation signifi-

Table 3.5: Classification results and comparison to the state of the art.

Drawing		Models	Binary	Multi-class
Off-line data	Clock [31]	DenseNet-121	74%	61%
	Letter [14]	InceptionResNetV2	74.6%	N/A
	Pentagon [33]	EfficientNet	87%	76%
	House (Off), ours	EfficientNet	76%	60%
	House (OnOff), ours	EfficientNet	82%	66%
On-line data	Clock [3]	DL	83.44%	N/A
	Signature [66]	SVM	75.71%	N/A
	Letter [16]	1D-CNN	89%	N/A
	House (On), ours	BiGRU	72%	61%
	House (OnOff), ours	EfficientNet	82%	66%

Note: All these studies used data augmentation to improve their model performance.

cantly enhances model performance, which aligns with findings by Dao et al.[16] and Bensalah et al.[2] (see Table 3.5). Unlike their studies, which focused on GANs, our work shows that traditional augmentation techniques like jittering and scaling can also yield high accuracy. This highlights the versatility of these simple augmentation techniques across different datasets and model architectures.

2. **Model performance**: Similar to Bucholc et al. [6], who reported high accuracy with a hybrid Random Forest model, our study finds that BiGRU performs exceptionally well in multi-class classification tasks. This suggests that RNNs, particularly BiGRU, are highly effective in capturing the temporal dynamics of handwriting data, which is critical for distinguishing between HC, MCI, and AD.

3. **Data Representations**: Our findings that on-line data representations are preferred, even when converted to off-line data (here OnOff-line format), are consistent with Piers et al. [13], who suggested that using both on-line and off-line handwriting analysis with deep transfer learning and GANs can improve early-stage Alzheimer's disease detection. This underscores the importance of preserving temporal information in handwriting data for accurate classification.

4. **Digital Drawing Tools**: The use of digital drawing tools, as highlighted in previous studies [58, 47, 54]), supports our approach of using on-line pen stroke data to analyze drawing characteristics indicative of cognitive impairment. All those previous studies emphasized the potential of these tools for early dementia detection but only a handful of them actually did some computational modeling tasks.

3.19 Limitations and Future Work

We acknowledge that relying solely on a single type of drawing task (in our case, a house) may limit the generalizability of our findings across different clinical settings where other forms of cognitive assessments are used. Although, as discussed in the Introduction section, the HDT evaluates a range of cognitive functions [22], therefore, in line with recent work, it should become the standard task to screen MCI and AD patients. We also should note that previous work has considered pentagons [52], clocks [31], and signatures [32, 55], achieving similar results (sometimes lower) to ours.

On the other hand, our experimental findings are drawn from a relatively small sample size, which might limit their applicability to a broader population of patients. However, the challenge of recruiting a large and diverse cohort is a common and pervasive issue in digital medicine [35, 65]. Despite these limitations, our results hold promise and could pave the way for future clinical applications using a simple handwriting test as a non-invasive, low-cost, and accurate method.

3.20 Conclusion

Handwriting analysis has been used by neurologists assessing suspected dementia patients in conjunction with a range of other measurements and tests. We have investigated how neural networks can use off-line and on-line house drawings to classify AD and MCI patients. We have observed that on-line data converted to off-line data is the most efficient approach to distinguish patients from HC (binary classification), whereas on-line data representations are preferred over off-line data for distinguishing AD, MCI, and HC (multi-class classification). Taken together, our results show the potential to enhance home-based healthcare services, using non-invasive, low-cost handwriting tests like the one we have investigated in this paper.

Acknowlegments

Work supported by the UCM research group (Grupo de Investigación básica en Ciencias de la Visión del IIORC, UCM-GR17-920105), the Horizon 2020 FET program of the European Union (grant CHIST-ERA-20-BCI-001), and the European Innovation Council Pathfinder program (grant 101071147).

References for Paper 5

[1] Alzheimer's Association. 2022. 2022 Alzheimer's disease facts and figures. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 18, 4, (Apr. 2022), 700–789.

[2] Asma Bensalah, Antonio Parziale, Giuseppe De Gregorio, Angelo Marcelli, Alicia Fornés, and Josep Lladós. 2023. I Can't Believe It's Not Better: In-air Movement for Alzheimer Handwriting Synthetic Generation. In *Graphonomics in Human Body Movement. Bridging Research and Practice from Motor Control to Handwriting Analysis and Recognition*. Antonio Parziale, Moises Diaz, and Filipe Melo, editors. Cham, 136–148.

[3] Russell Binaco, Nicholas Calzaretto, Jacob Epifano, Sean McGuire, Muhammad Umer, Sheina Emrani, Victor Wasserman, David J. Libon, and Robi Polikar. 2020. Machine Learning Analysis of Digital Clock Drawing Test Performance for Differential Classification of Mild Cognitive Impairment Subtypes Versus Alzheimer's Disease. *Journal of the International Neuropsychological Society*, 26, 7, (Aug. 2020), 690–700.

[4] Gary Bradski. 2000. The openCV library. *Miller Freeman Inc*, 25, 11, 120–123.

[5] Zeinab Breijyeh and Rafik Karaman. 2020. Comprehensive review on alzheimer's disease: causes and treatment. *Molecules*, 25, 24, 5789.

[6] Magda Bucholc, Sofya Titarenko, Xuemei Ding, Callum Canavan, and Tianhua Chen. 2023. A hybrid machine learning approach for prediction of conversion from mild cognitive impairment to dementia. *Expert Systems with Applications*, 217, (May 2023), 119541.

[7] Joyce Y. C. Chan, Baker K. K. Bat, Adrian Wong, Tak Kit Chan, Zhaohua Huo, Benjamin H. K. Yip, Timothy C. Y. Kowk, and Kelvin K. F. Tsoi. 2022. Evaluation of Digital Drawing Tests and Paper-and-Pencil Drawing Tests for the Screening of Mild Cognitive Impairment and Dementia: A Systematic Review and Meta-analysis of Diagnostic Studies. *Neuropsychology Review*, 32, 3, 566–576. Retrieved Jan. 24, 2023 from.

[8] T. Charernboon. 2017. Diagnostic accuracy of the overlapping infinity loops, wire cube, and clock drawing tests for cognitive impairment in mild cognitive impairment and dementia. *International Journal of Alzheimer's Disease*, 2017. DOI: 10.1155/2017/5289239.

[9] Bo-Lin Chen, Kuan-Ting Hu, Kuo-Sheng Cheng, and Chien-Yu Chen. 2024. Automatic CDT Scoring Using Machine Learning with Interpretable Feature. In *Proc. ICBBB' 24*. (Mar. 2024), 55–59. Retrieved Mar. 14, 2024 from.

[10] Shuqing Chen, Daniel Stromer, Harb Alnasser Alabdalrahim, Stefan Schwab, Markus Weih, and Andreas Maier. 2020. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports 2020 10:1*, 10, 1, (Nov. 2020), 1–11.

[11] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*.

[12] Nicole Cilia, Tiziana D'Alessandro, Claudio De Stefano, and Francesco Fontanella. 2022. Deep transfer learning algorithms applied to synthetic drawing images as a tool for supporting Alzheimer's disease prediction. *Machine Vision and Applications*, 33, (May 2022).

[13] Nicole Dalia Cilia, Tiziana D'Alessandro, Claudio De Stefano, Francesco Fontanella, and Mario Molinara. 2021. From online handwriting to synthetic images for Alzheimer's disease detection using a deep transfer learning approach. *IEEE Journal of Biomedical and Health Informatics.*

[14] Nicole Dalia Cilia, Tiziana D'Alessandro, Claudio De Stefano, and Francesco Fontanella. 2022. Offline handwriting image analysis to predict alzheimer's disease via deep learning. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2807–2813.

[15] Nicole Dalia Cilia, Claudio De Stefano, Francesco Fontanella, Mario Molinara, and Alessandra Scotto Di Freca. 2019. Handwriting analysis to support alzheimer's disease diagnosis: a preliminary study. In *Computer Analysis of Images and Patterns: 18th International Conference, CAIP 2019, Salerno, Italy, September 3–5, 2019, Proceedings, Part II 18*. Springer, 143–151.

[16] Quang Dao, Mounîm A. El-Yacoubi, and Anne-Sophie Rigaud. 2023. Detection of Alzheimer Disease on Online Handwriting Using 1D Convolutional Neural Network. *IEEE Access*, 11, 2148–2155.

[17] Claudio De Stefano, Francesco Fontanella, Donato Impedovo, Giuseppe Pirlo, and Alessandra Scotto di Freca. 2019. Handwriting analysis to support neurodegenerative diseases diagnosis: a review. *Pattern Recognition Letters*, 121, 37–45.

[18] Claudio De Stefano, Francesco Fontanella, Donato Impedovo, Giuseppe Pirlo, and Alessandra Scotto di Freca. 2019. Handwriting analysis to support neurodegenerative diseases diagnosis: A review. *Pattern Recognition Letters*. Graphonomics for e-citizens: e-health, e-society, e-education 121, (Apr. 2019), 37–45.

[19] Charles DeCarli. 2003. Mild cognitive impairment: prevalence, prognosis, aetiology, and treatment. English. *The Lancet Neurology*, 2, 1, (Jan. 2003), 15–21. Retrieved Mar. 19, 2024 from.

[20] Margarete Delazer, Laura Zamarian, and Atbin Djamshidian. 2021. Handwriting in Alzheimer's Disease. en. *Journal of Alzheimer's Disease*, 82, 2, (Jan. 2021), 727–735.

[21] Marcos Faundez-Zanuy, Jiri Mekyska, and Donato Impedovo. 2021. Online Handwriting, Signature and Touch Dynamics: Tasks and Potential Applications in the Field of Security and Health. en. *Cognitive Computation*, 13, 5, (Sept. 2021), 1406–1421. Retrieved Mar. 14, 2024 from.

[22] Marcos Faundez-Zanuy, Enric Sesa-Nogueras, Josep Roure-Alcobé, Josep Garré-Olmo, Karmele Lopez-de-Ipiña, and Jordi Solé-Casals. 2014. Online drawings for dementia diagnose: in-air and pressure information analysis. In *XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013: MEDICON 2013, 25-28 September 2013, Seville, Spain*. Springer, 567–570.

[23] Marcos Faundez-Zanuy et al. 2013. Biometric applications related to human beings: there is life beyond security. *Cognitive Computation*, 5, 136–151.

[24] Katrina E. Forbes, Michael F. Shanks, and Annalena Venneri. 2004. The evolution of dysgraphia in Alzheimer's disease. *Brain Research Bulletin*, 63, 1, (Mar. 2004), 19–24.

[25] Josep Garre-Olmo, Marcos Faúndez-Zanuy, Karmele López-de-Ipiña, Laia Calvó-Perxas, and Oriol Turró-Garriga. 2017. Kinematic and pressure features of handwriting and drawing: preliminary results between patients with mild cognitive impairment, alzheimer disease and healthy controls. *Current Alzheimer Research*, 14, 9, 960–968.

[26] Rajib Ghosh. 2021. A Recurrent Neural Network based deep learning model for offline signature verification and recognition system. *Expert Systems with Applications*, 168, (Apr. 2021), 114249. Retrieved Apr. 2, 2024 from.

[27] Martin Guha. 2014. Diagnostic and Statistical Manual of Mental Disorders: DSM-5 (5th edition). *Reference Reviews*, 28, 3, (Jan. 2014), 36–37.

[28] Youssef H El-Hayek et al. 2019. Tip of the iceberg: assessing the global socioeconomic costs of alzheimer's disease and related dementias and strategic implications for stakeholders. *Journal of Alzheimer's Disease*, 70, 2, 323–341.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. the IEEE CVPR*, 770–778.

[30] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9, 8, (Nov. 1997), 1735–1780.

[31] Nina Hosseini-Kivanani, Christoph Schommer, and Luis. A Leiva. 2023. The Magic Number: Impact of Sample Size for Dementia Screening Using Transfer Learning and Data Augmentation of Clock Drawing Test Images. In *Healthcom*. China, 23–28.

[32] Nina Hosseini-Kivanani et al. 2023. Better Together: Combining Different Handwriting Input Sources Improves Dementia Screening. In *IEEE e-Science*. Cyprus, 1–7.

[33] Nina Hosseini-Kivanani et al. 2024. Ink of insight: data augmentation for dementia screening through handwriting analysis. In *Proceedings of the 2024 8th International Conference on Medical and Health Informatics*, 224–229.

[34] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Proc. the IEEE CVPR*, 4700–4708.

[35] Donato Impedovo and Giuseppe Pirlo. 2018. Dynamic Handwriting Analysis for the Assessment of Neurodegenerative Diseases: A Pattern Recognition Perspective. *IEEE reviews in biomedical engineering*, 12, 209–220.

[36] Carmen Jimenez-Mesa, Ignacio Alvarez Illan, Alberto Martin-Martin, Diego Castillo-Barnes, Francisco Jesus Martinez-Murcia, Javier Ramirez, and Juan M. Gorriz. 2020. Optimized one vs one approach in multiclass classification for early alzheimer's disease and mild cognitive impairment diagnosis. *IEEE Access*, 8, 96981–96993. Publisher: Institute of Electrical and Electronics Engineers Inc. DOI: `10.1109/ACCESS.2020.2997736`.

[37] C. Jiménez-Mesa et al. 2022. Automatic Classification System for Diagnosis of Cognitive Impairment Based on the Clock-Drawing Test. en. In *Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications* (Lecture Notes in Computer Science). José Manuel Ferrández Vicente, José Ramón Álvarez-Sánchez, Félix de la Paz López, and Hojjat Adeli, editors. Springer International Publishing, 34–42.

[38] K. Kantarci et al. 2009. Risk of dementia in MCI: combined effect of cerebrovascular disease, volumetric MRI, and 1H MRS. eng. *Neurology*, 72, 17, (Apr. 2009), 1519–1525.

[39] Jacek Kawa, Adam Bednorz, Paula Stępień, Jarosław Derejczyk, and Monika Bugdol. 2017. Spatial and dynamical handwriting analysis in mild cognitive impairment. *Computers in Biology and Medicine*, 82, 21–28.

[40] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521, 7553, (May 2015), 436–444.

[41] Luis A. Leiva, Moises Diaz, Miguel A. Ferrer, and Réjean Plamondon. 2021. Human or Machine? It Is Not What You Write, But How You Write It. In *ICPR*. (Jan. 2021), 2612–2619.

[42] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. A survey on deep learning in medical image analysis. eng. *Medical Image Analysis*, 42, (Dec. 2017), 60–88.

[43] Sara Marullo, Maria Pozzi, Monica Malvezzi, and Domenico Prattichizzo. 2022. Analysis of postures for handwriting on touch screens without using tools. *Scientific Reports*, 12, 1, (Jan. 2022), 296.

[44] Jumpei Maruta, Kentaro Uchida, Hideo Kurozumi, Satoshi Nogi, Satoshi Akada, Aki Nakanishi, Miki Shinoda, Masatsugu Shiba, and Koki Inoue. 2022. Deep convolutional neural networks for automated scoring of pentagon copying test results. *Scientific Reports*, 12, 1, (Dec. 2022), 9881.

[45] Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. 1984. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. en. *Neurology*, 34, 7, (July 1984), 939–939.

[46] S. Müller, O. Preische, Petra Heymann, U. Elbing, and C. Laske. 2017. Increased diagnostic accuracy of digital vs. conventional clock drawing test for discrimination of patients in the early course of alzheimer's disease from cognitively healthy individuals. *Frontiers in Aging Neuroscience*, 9.

[47] Stephan Müller, Laura Herde, Oliver Preische, Anja Zeller, Petra Heymann, Sibylle Robens, Ulrich Elbing, and Christoph Laske. 2019. Diagnostic value of digital clock drawing test in comparison with CERAD neuropsychological battery total score for discrimination of patients in the early course of Alzheimer's disease from healthy individuals. en. *Scientific Reports*, 9, 1, (Mar. 2019), 3543. Retrieved Mar. 14, 2024 from.

[48] David G Munoz and Howard Feldman. 2000. Causes of Alzheimer's disease. *Cmaj*, 162, 1, 65–72.

[49] Nickson Mwamsojo, Frederic Lehmann, Mounim A. El-Yacoubi, Kamel Merghem, Yann Frignac, Badr-Eddine Benkelfat, and Anne-Sophie Rigaud. 2022. Reservoir Computing for Early Stage Alzheimer's Disease Detection. *IEEE*, 10, 59821–59831.

[50] Sanjay Nagaraj and Tim Q Duong. 2021. Deep learning and risk score classification of mild cognitive impairment and alzheimer's disease. *Journal of Alzheimer's Disease*, 80, 3, 1079–1090.

[51] Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. 2019. Kerastuner. `https://github.com/keras-team/keras-tuner`. (2019).

[52] Ingyu Park, Yun Joong Kim, Yeo Jin Kim, and Unjoo Lee. 2020. Automatic, Qualitative Scoring of the Interlocking Pentagon Drawing Test (PDT) Based on U-Net and Mobile Sensor Data. en. *Sensors*, 20, 5, (Jan. 2020), 1283.

[53] R. C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, and L. Fratiglioni. 2014. Mild cognitive impairment: a concept in evolution. en. *Journal of Internal Medicine*, 275, 3, 214–228.

[54] Ryan J. Piers et al. 2017. Age and Graphomotor Decision Making Assessed with the Digital Clock Drawing Test: The Framingham Heart Study. *Journal of Alzheimer's Disease*, 60, 4, (Nov. 2017), 1611–1620. Retrieved May 23, 2023 from.

[55] Giuseppe Pirlo, Moises Diaz, Miguel Angel Ferrer, Donato Impedovo, Fabrizio Occhionero, and Urbano Zurlo. 2015. Early diagnosis of neurodegenerative diseases by handwritten signature analysis. *Lecture Notes in Computer Science*, 9281, 290–297.

[56] Gabriel Poirier, Alice Ohayon, Adrien Juranville, France Mourey, and Jeremie Gaveau. 2021. Deterioration, Compensation and Motor Control Processes in Healthy Aging, Mild Cognitive Impairment and Alzheimer's Disease. *Geriatrics*, 6, 1, (Mar. 2021), 33.

[57] Raksit Raksasat, Surat Teerapittayanon, Sirawaj Itthipuripat, Kearkiat Praditpornsilpa, Aisawan Petchlorlian, Thiparat Chotibut, Chaipat Chunharas, and Itthi Chatnuntawech. 2023. Attentive pairwise interaction network for ai-assisted clock drawing test assessment of early visuospatial deficits. *Scientific Reports*, 13, 1, 18113.

[58] Sibylle Robens, Petra Heymann, Regine Gienger, Andreas Hett, S. Müller, C. Laske, Roland Loy, T. Ostermann, and U. Elbing. 2019. The digital tree drawing test for screening of early dementia: an explorative study comparing healthy controls, patients with mild cognitive impairment, and patients with early dementia of the alzheimer type. *Journal of Alzheimer's disease : JAD*, 68 4, 1561–1574.

[59] Sibylle Robens, Petra Heymann, Regine Gienger, Andreas Hett, Stephan Müller, Christoph Laske, Roland Loy, Thomas Ostermann, and Ulrich Elbing. 2019. The digital tree drawing test for screening of early dementia: an explorative study comparing healthy controls, patients with mild cognitive impairment, and patients with early dementia of the alzheimer type. *Journal of Alzheimer's Disease*, 68, 4, 1561–1574.

[60] Pavel Senin. 2008. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii, USA*, 855, 1-23, 40.

[61] Michelangelo Stanzani Maserati, Corrado Matacena, Luisa Sambati, Federico Oppi, Roberto Poda, Maddalena De Matteis, Roberto Gallassi, et al. 2015. The tree-drawing test (koch's baum test): a useful aid to diagnose cognitive impairment. *Behavioural neurology*, 2015.

[62] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proc. the 36th International Conference on Machine Learning.* PMLR, (May 2019), 6105–6114.

[63] Tom Tombaugh and Nancy J McIntyre. 1992. The Mini-Mental State Examination: A Comprehensive Review. *Journal of the American Geriatrics Society*, 40, 9, 922–935.

[64] Terry T. Um, Franz M. J. Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proc. ICMI '17.* (Nov. 2017), 216–220.

[65] Gennaro Vessio. 2019. Dynamic Handwriting Analysis for Neurodegenerative Disease Assessment: A Literary Review. *Applied Sciences 2019, Vol. 9, Page 4666*, 9, 21, (Nov. 2019), 4666–4666.

[66] Zelong Wang, Majd Abazid, Nesma Houmani, Sonia Garcia-Salicetti, and Anne-Sophie Rigaud. 2019. Online signature analysis for characterizing early stage alzheimer's disease: a feasibility study. *Entropy*, 21, 10, 956.

[67] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13, 4, (Apr. 2004), 600–612.

[68] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big Data*, 3, 1, (May 2016), 9. Retrieved Feb. 17, 2024 from.

[69] Perla Werner, Sara Rosenblum, Gady Bar-On, Jeremia Heinik, and Amos Korczyn. 2006. Handwriting Process Variables Discriminating Mild Alzheimer's Disease and Mild Cognitive Impairment. *The Journals of Gerontology: Series B*, 61, 4, (July 2006), P228–P236.

[70] Mounim A El-Yacoubi, Sonia Garcia-Salicetti, Christian Kahindo, Anne-Sophie Rigaud, and Victoria Cristancho-Lacroix. 2019. From aging to early-stage alzheimer's: uncovering handwriting multimodal behaviors by semi-supervised learning and sequential representation learning. *Pattern Recognition*, 86, 112–133.

[71] Young Chul Youn et al. 2021. Use of the Clock Drawing Test and the Rey–Osterrieth Complex Figure Test-copy with convolutional neural networks to predict cognitive impairment. *Alzheimer's Research & Therapy*, 13, 1, (Apr. 2021), 85.

4 Data Augmentation

Data augmentation techniques are critical in ML and DL, particularly within clinical domains where labeled samples are scarce and class distributions are uneven. In dementia screening, handwriting and drawing assessments serve as non-invasive diagnostic tools for early detection; yet, the limited number of annotated datasets restricts the scope of DL models. This chapter focuses on straightforward and rapid augmentation methods that address these constraints without imposing extra demands on data collection.

Recent research indicates that applying augmentation methods can mitigate model overfitting, expand the variety of training samples, and enhance predictive performance in cognitive screening. Traditional approaches, such as geometric modifications, are widely used in natural image processing. Their adaptation to clinical handwriting and drawing tasks, however, requires careful calibration to preserve diagnostically pertinent features. Here, we examine strategies that range from basic geometric adjustments to advanced automatic augmentation techniques specifically tailored for cognitive assessment images. Two investigations form the basis of our discussion. The first study, *ICMHI'24*, examines the effect of applying geometric and spatial modifications on limited handwriting datasets for dementia classification. In this investigation, both conventional and deep learning methods are compared to determine the improvement derived from augmentation. The second study, *ICAART'25*, assesses fixed and adjustable automatic augmentation approaches developed for CDT images. This work contrasts methods that adjust augmentation policies with those that rely on predetermined procedures, thereby evaluating which strategies best preserve clinical content while enhancing model performance. This chapter contributes in three significant ways. First, it provides a thorough analysis of diverse augmentation techniques and investigates their influence on models used for dementia screening, distinguishing between conventional and automatic approaches. Second, it considers the broader ramifications for medical image classification and outlines promising directions for the incorporation of augmentation strategies into artificial intelligence–based diagnostic tools.

The ensuing sections detail the methodologies, experimental outcomes, and primary observations from these two investigations, offering a comprehensive perspective on the role of data augmentation in refining cognitive screening models. The findings herein advance the research on artificial intelligence in healthcare and emphasize the importance of accessible and scalable augmentation methods in clinical machine learning applications.

4.1 Ink of insight: Data augmentation for dementia screening through handwriting analysis

Abstract[†]

We investigate the use of handwriting data as a means of predicting early symptoms of Alzheimer's disease (AD). Thirty-six subjects were classified based on the standardized pentagon drawing test (PDT) using deep learning (DL) models. We also compare and contrast classic machine learning (ML) models with DL by employing different data augmentation (DA) techniques. Our findings indicate that DA greatly improves the performance of all models, but the DL-based ones are the ones that achieve the best and highest results. The best model (EfficientNet) achieved a classification accuracy of 87% and an area under the receiver operating characteristic curve (AUC) of 91% for binary classification (healthy or AD patients), whereas for multiclass classification (healthy, mild AD, or moderate AD) accuracy was 76% and AUC was 77%. These results underscore the potential of DA as a simple, cost-effective approach to aid practitioners in screening AD in larger populations, suggesting DL models are capable of analyzing handwriting data with a high degree of accuracy, which may lead to better and earlier detection of AD.

Keywords

Alzheimer's Disease; Screening; Pentagon Drawing Test; Data Augmentation; Image Classification; Machine Learning; Deep Learning

4.2 Introduction and Related Work

Alzheimer's disease refers to a dementia syndrome characterized by primary impairments of cortical cognitive functions, including memory, language, and praxis, that gradually progress over time [15]. These impairments have a high functional impact and are often accompanied by various neuropsychiatric symptoms [7]. As the disease progresses, the number of damaged neurons and the extent of affected brain regions increases, leading to a greater need for assistance from family members, friends, and professional caregivers for daily tasks [1]. The early stages of AD are characterized by memory loss, recognition

[†] **N. Hosseini-Kivanani**, E. Salobrar-García, L. Elvira-Hurtado, I. López-Cuenca, R. de Hoz, J. M. Ramírez, P. Gil, M. Salas-Carrillo, C. Schommer, and L. A. Leiva, "Ink of insight: Data augmentation for dementia screening through handwriting analysis," in *Proceedings of the 2024 8th International Conference on Medical and Health Informatics (ICMHI)*, 2024, pp. 224–229. DOI: 10.1145/3673971.3673992

problems (such as an object or face recognition [8]), visual impairments, and deficits in spatial perception [21], despite relatively normal visual acuity values and intact visual fields [23]. Recent research has shown that assessing visuospatial function, in addition to brain scanning, can aid in the early detection of impairments. Effective screening tests can identify visuospatial dysfunction, which may manifest years before the onset of clinical symptoms [18]. However, existing screening measures for cognitive changes face challenges, particularly with regard to their limited intra-individual reliability, which hinders accurate tracking of cognitive changes over time.

Drawing tests, frequently used in dementia screening, can reflect the presence of the condition through changes in a person's drawing ability [22]. However, the subjectivity in scoring systems used in these tests and their limited scope in capturing a range of drawing attributes often result in missing subtle yet clinically significant indicators of cognitive decline. This means that no single scoring system is reported as the most effective and reliable evaluation method (e.g., [11]). This highlights the need for more comprehensive and objective screening methods. There is a growing interest in exploring more advanced analytical approaches, such as the integration of machine learning (ML) techniques, to augment the diagnostic effectiveness of cognitive screening tools.

Recent advancements in artificial intelligence (AI), particularly in deep learning (DL), have significantly impacted healthcare, especially when it comes to diagnosing neurodegenerative diseases like AD (e.g., [11, 24, 6]). DL models have played an instrumental role in the analysis of neuroimaging, detecting complex patterns in brain scans that are imperceptible to the human eye. Our study focuses on refining DL models for dementia screening and emphasizing the importance of DA techniques in contexts with limited high-quality and diverse data. This approach is vital for improving model robustness, especially in applications like automated analysis of scanned paper-based handwriting and drawings, which are crucial in AD screening. Recent research has highlighted DL's transformative role in healthcare, particularly in the early detection and management of cognitive impairments [3, 12, 6, 14].

Relevant studies (e.g., [12, 6]) have highlighted the precision of DL models, particularly convolutional neural networks (CNNs). However, the effectiveness of these models is often limited by the small size of available datasets. Maruta et al. [19] demonstrated that the fine-tuned GoogleNet CNN outperforms other CNN models like VGG-16, ResNet-50, and Inception-v3 in automatically evaluating the pentagon drawings for constructional apraxia. Additionally, Tasaki et al. [28] conducted a study on the usage of a DL model called PentaMind. which analyzes hand-drawn images of intersecting pentagons to extract cognition-related features. The model was trained on 13,777 images and successfully extracted features such as line waviness, which shows an improvement over conventional visual assessment methods. Jiménez-Mesa et al. [16] proposed a CNN-based method for diagnosing cognitive impairment through the Clock Drawing Test (CDT), effectively classifying drawings as healthy or patient, indicating its potential for hospital and clinic use, particularly in resource-limited areas. The use of DL in cognitive impairment tests is not without limitations, primarily due to the limited dataset sizes and variability. DA emerges as a pivotal solution to enhance model robustness and accuracy. It involves generating additional training data from existing datasets, increasing size, diversity, and quality. However, challenges exist in preserving clinical relevance and avoiding artificial biases.

Summary of Contributions

Our research builds upon significant advancements in ML for cognitive impairment screening, aiming to tackle the existing challenges. This brings us to the core objectives of our research. Firstly, we aim to develop robust DL Models for AD screening to refine and enhance the existing models. Secondly, our study focuses on the importance of DA in clinical settings, emphasizing the preservation of data integrity and reliability. Thirdly, we explore the comparative advantages of DL over classic ML in the context of AD screening, providing a comprehensive insight into the future of digital screening in AD.

4.3 Materials and Methods

Our goal is to improve the performance of ML models in classifying handwriting data by implementing suitable DA techniques. Although DA has shown advantages in other scientific domains, its application to handwriting data in clinical contexts has received little attention. This is primarily because the augmented data is often either too similar to the original data or too distorted for the models to learn effectively from it (e.g., [25]). This study compares classic ML models (SVMs, RFs, k-NNs) and DL models (CNNs) in the context of classifying binary (healthy vs. patient) and multiclass (healthy, mild AD, and moderate AD) classification tasks, both with and without applying DA.

4.3.1 Data Collection and Tasks

The study recruited 36 subjects (13 female and 23 male) from the Memory Unit of the (Hospital Clinico San Carlos (HCSC) for a study on cognitive and neurophysiological characteristics of individuals at high risk of dementia. Subjects were categorized according to the guidelines of the National Institute of Neurological and Communicative Disorders and

Stroke (NINCDS), Alzheimer's Disease and Related Disorders Association (ADRDA) [20] and the Statistical Manual of Mental Disorders V (DSM V) [9]. Based on these guidelines, the subjects were classified into two groups of patients (mild AD, n=3, and moderate AD, n=3) and one group of healthy subjects (control, n=30). All the subjects provided informed consent prior to participation. The participants' ages ranged from 61 to 88 years old, with a mean age of 73.92 ± 6,78 years old. No significant differences in age were observed among the healthy group, mild AD, or moderate AD based on p-value $> .05$. The study included 30 individuals aged between 61 to 84 years who were cognitively healthy with no evidence of brain injury and had MMSE score above 26. Non-healthy participants had MMSE scores between 25 and 17.

4.3.2 Image Preprocessing and Data Augmentation

Participants were given a blank A4-sized paper and asked to copy a figure of two overlapping pentagons with an interlocking shape (as shown in Figure 4.1a). The paper-and-pencil drawings were converted from PDF files to image format (PNG format) (Figure 4.1a and b) to be processed with our classic ML and DL models. The PNG images were then converted from color images (three channels) to grayscale (one channel) (Figure 4.1c). The resulting images were resized to standard dimensions (224×224 px). Any nonrelevant information, such as the original printed images from the clinicians, that appeared on the top side of the original file (Figure 4.1a), was removed during the preprocessing pipeline. Finally, images were padded to remove noise from the image and make them in the same shape, and the canny edge detector from OpenCV library [4] was used to improve the resulting image (Figure 4.1d). Low-quality and noisy images (in total, 14 images from the healthy group) were manually filtered out.

ML (and, particularly, DL) models perform better when trained on large datasets; however, obtaining such large-scale datasets is really challenging in clinical fields. To address this, DA techniques can be used to artificially increase the size of the dataset. By generating additional images from the input images, these techniques can help reduce the risk of overfitting and increase the model's generalizability, leading to better overall performance. These techniques include applying geometric transformations (such as flipping, cropping, rotating, and translating), changing the color space of the images, mixing images, or even using generative adversarial networks [25]. In this study, we only applied geometric transformations to images for DA, carefully avoiding transformations that would potentially destroy the semantics of the original image and are not suitable for our grayscale handwritten images. Therefore, techniques commonly used in broader computer vision applications, such as hue adjustments or color inversion, were deliberately excluded from

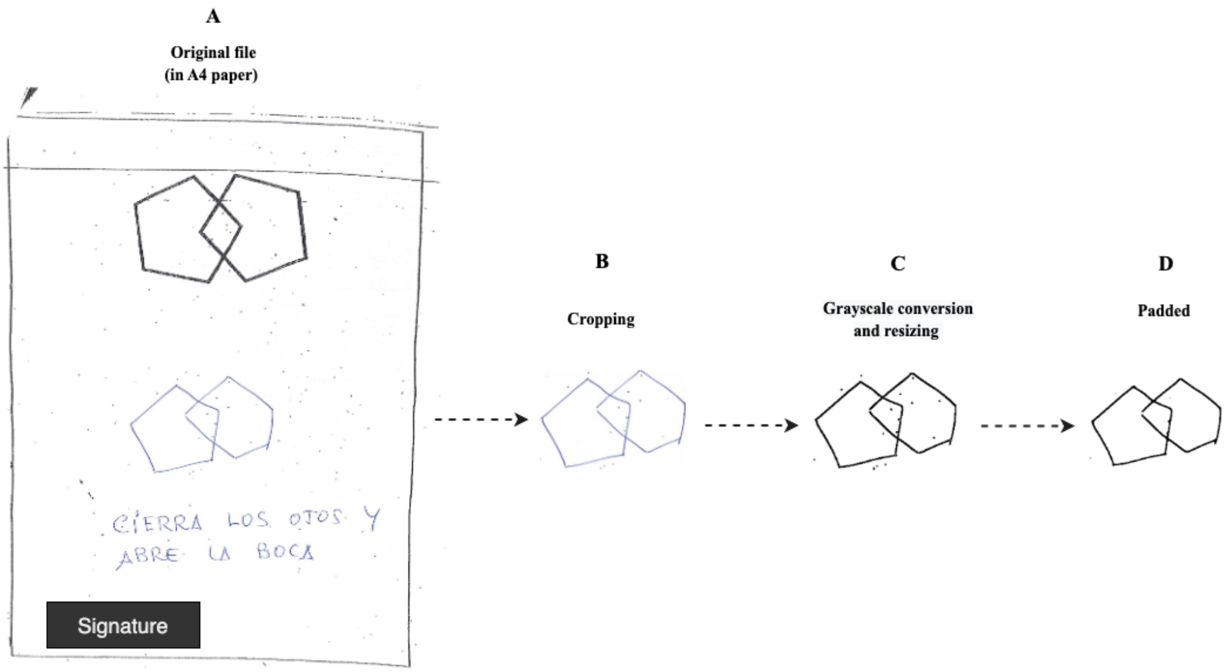


Figure 4.1: Example of the preprocessing steps for Pentagon Drawing Test (PDT) images: prompting pentagon (A) on top with participants' drawings at the bottom; image processing (B, C); and final image (D) for model input.

our process. Our approach was to maintain the integrity of the original handwritten samples, ensuring that the essential characteristics of these images were preserved.

To determine the quality of the augmented data, we used the structural similarity index (SSIM) [29]. SSIM measures the similarity between two images by considering the human visual perception of differences in terms of luminance, contrast, and structure. SSIM is a widely used measurement tool because of its low computational complexity and ability to compare synthetic and original images. The SSIM method uses a sliding window to analyze the structural distortion between two similar images. The SSIM score ranges from 0 to 1, with a score of 1 indicating that the images are the same and a score of 0 indicating that the images are totally different. For applying DA techniques such as elastic transformation, grid distortion, and rotation to the images in our training set, we used the Albumentations open-source toolkit [5]. These DA techniques were applied to the images in our training set, which resulted in an increased sample size. Crucially, we allocated all original images from patient subjects exclusively to the test set to ensure a robust testing protocol. The training set consisted of 60 images for healthy and 60 for patient classes. The test set comprised 6 images for healthy and 6 images for patient classes.

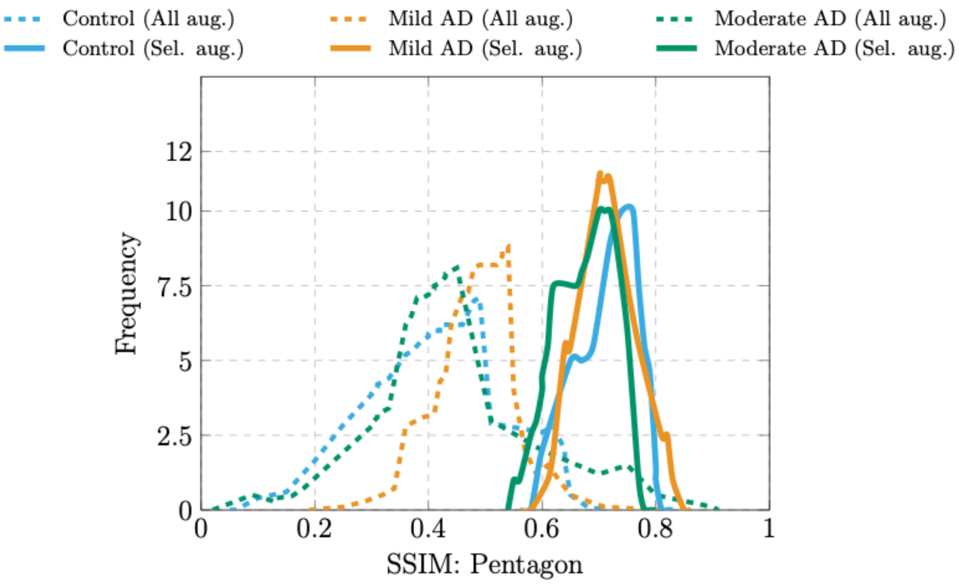


Figure 4.2: SSIM distributions for Pentagon drawings. Dashed lines represent the SSIM results of all augmentation techniques (All aug.), while solid lines correspond to the selected augmentation techniques (Sel. aug).

After DA, as shown in Figure 4.2, the SSIM values ranged from 0.6 to 0.7, indicating that the augmented images are not near-duplicates of the original data but are rather new images. However, when all DA techniques from the Albumentations toolkit were applied, the distribution of SSIM values was from 0.1 to 0.7, indicating that the augmented images are much more different than the original images, which is not desirable in our research.

4.3.3 Classic ML and DL models for AD screening

We selected classic ML and DL models based on their proven strengths and applicability to medical image analysis. Classic ML models were SVM, RF, and k-NN. They require manual feature extraction, whereas DL models automatically identify and optimize relevant features from data.

Among DL models, CNNs are the most popular and widely used in image-related tasks [31], due to their ability to automatically detect features by using a composition of the different types of layers: (i) Convolutional (CONV) are the primary building blocks of a CNN model for extracting features such as colors, edges, and corners from the input by applying the convolution operation through a sliding kernel, (ii) Pooling (POOL) are used to reduce the dimensionality of the feature maps computed by the CONV layers, and (iii) Fully-connected layers (FC) are placed at the end of the model's architecture to flatten the output of the previous layer and to add non-linearities to the model.

We evaluated various state-of-the-art CNN architectures for AD screening. All models have the same input layer (224x224 px grayscale images) and the same output layer (with Softmax activation function):

VGG-16 [26] features 16 CONV layers with 3×3 kernels, followed by 3 FC layers before the output layer.

ResNet-152 [10] is a deep residual network architecture with 152 CONV layers. It uses skip connections between CONV layers, a kernel size of 3×3, and batch normalization. The model has two FC layers before the output layer.

DenseNet-121 [13] is a deep CNN composed of 121 layers, including CONV layers with 7×7 kernels, and DenseBlocks, which are groups of CONV layers with 1×1 and 3×3 kernels interconnected through transition layers, and finally an FC layer followed by the output layer.

EfficientNet [27] has multiple CONV layers with a mix of different kernels, followed by corresponding POOL layers and a single FC layer before the output layer.

Custom CNN that we designed with five CONV layers with 3×3 kernels, followed by two POOL layers and one FC layer before the output layer.

Except for our proposed Custom CNN model, the other CNNs are pre-trained on the ImageNet dataset, which contains 1M images distributed over 1000 classes. Therefore, we used transfer learning to finetune those architectures on our dataset. Accordingly, the dimensionality of the output layer is reduced from 1000 classes to 2 or 3, depending on the classification experiments. We used 2 classes in binary classification experiments and 3 classes in multiclass classification experiments.

4.3.4 Model training

To train the classic ML models (SVM, k-NN, RF), we used 5-fold cross-validation, which involves randomly dividing the dataset into 5 groups or folds. For the SVM classifier, a “C” value of 0.1, a “gamma” of 0.0001, and a “linear” kernel were determined to be best. For the k-NN classifier, the “manhattan” metric with “n_neighbors” set to 3 and “weights” configured as “distance” was used. The RF classifier, on the other hand, used a “max_depth” of 15, “max_features” of 9, a “min_impurity_decrease” of 1e-05, and “n_estimators” set at 70.

To train the DL models (CNNs), we used grid search to find the optimal parameters for each model. The learning rate varied between 0.0001 and 0.1, weight decay was fixed at

0.01, and the Adam optimization was employed. The models were trained over 50 epochs, using a batch size of 16, and the Cross-Entropy loss function was applied to optimize classification performance.

4.4 Results and Discussion

The efficiency of ML and DL models was evaluated using accuracy and area under the receiver operating characteristic curve (AUC). Accuracy represents the ratio of correct classifications to the total number of samples. AUC reports the performance of a classifier as a trade-off between the True Positive Rate and False Positive Rate, ranging from 0.5 (indicating random performance) to 1 (indicating perfect performance).

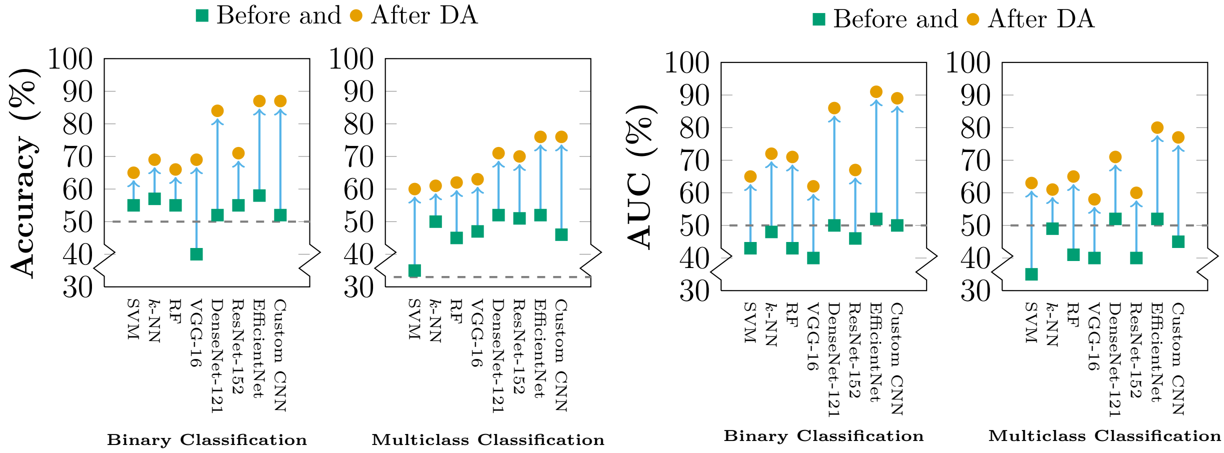


Figure 4.3: Accuracy and AUC values of classic ML and DL models, both before and after DA and for both binary class and multiclass classification experiments. Dashed lines represent the performance of a random classifier, illustrating the empirical lower bound in classification performance.

We have explored various classic ML and DL models for binary (healthy and patients) and multiclass (healthy, mild AD, and moderate AD) classification. The results are presented in Figure 4.3. Both classic ML and DL models showed an increase in accuracy after DA. This improvement was significant when compared to the baseline model (without DA). The results obtained from the classifiers that employed EfficientNet and our custom CNN outperformed all the other models, with 0.87 accuracy and 0.91 AUC scores for binary classification and 0.76 accuracy and 0.8 AUC for multi-class classification. In sum, DA led to a 10 to 30% increase in binary classification experiments and to a 10 to 20% increase in multi-class classification experiments.

Our work showcases the ability of classic ML and DL models to accurately classify AD patients, with a particular emphasis on integrating DA techniques. These DA methods were carefully selected based on their suitability in analyzing cognitive assessment tests used in AD diagnosis, addressing the limitations of current approaches in the existing literature (e.g., [30]). According to the SSIM results (Figure 4.2), the most appropriate DA techniques for PDT images include elastic transformation, grid distortion, horizontal flipping, translation offset, and rotations. Hosseini-Kivanani et al. [11] highlighted the importance of accurately choosing the DA techniques, showing that flipping and rotation can destroy the semantics of a CDT image. In contrast, in this work, flipping and rotation are appropriate DA techniques for PDT images, given their symmetry.

DL models have been used in various research for different types of cognitive assessments such as the paper-and-pencil CDT or cube drawing (e.g., [6, 2, 16, 24]). However, none of these studies have specifically focused on the use of DA. Furthermore, while there have been a few efforts to apply DL to automatically screen PDT images, these have not included the use of DA, as seen in [17, 19]. Our Custom CNN model, enhanced with DA, demonstrated exceptional proficiency in evaluating PDT images and outperformed previous studies' results with fewer data used in their studies.

After benchmarking our custom CNN against other state-of-the-art models, we found that it performs better in many cases, particularly when the data has a simple underlying pattern. The simpler structure of our Custom model allows it to learn and generalize these patterns more effectively, leading to higher performance. This suggests that our Custom CNN model with well-designed augmented images excels at certain tasks, such as simple drawings by patients, and is valuable for detecting AD patients from healthy individuals. Furthermore, it outperforms recent work that used pre-trained CNN models in similar task [19].

Our findings underscore the transformative potential of DA in enhancing the DL model's performance. By artificially increasing the dataset's size and diversity, both ML and DL models can be trained to be more robust and accurate, ultimately leading to improved patient outcomes in clinical settings. This research lays the groundwork for future advancements in AD treatment and care, aiming to ultimately improve the quality of life for those affected by AD.

4.5 Conclusion

This work provides valuable insights into the effectiveness of using DA on small clinical datasets for AD screening through handwriting analysis. Both classic ML and DL models were able to achieve better performance than they could without DA. Our method, which is practical for clinical use, offers a cost-effective solution to assist healthcare professionals in patient screening and minimizes subjectivity in interpreting clinical data, particularly in resource-limited settings. It can have a significant impact by helping doctors make more informed decisions and eventually provide better treatment options for patients.

Acknowledgments

Work supported by the UCM research group (Grupo de Investigación básica en Ciencias de la Visión del IIORC, UCM-GR17-920105), the Horizon 2020 FET program of the European Union (grant CHIST-ERA-20-BCI-001), and the European Innovation Council Pathfinder program (grant 101071147).

References for Paper 6

[1] Alzheimer's Association. 2022. 2022 Alzheimer's disease facts and figures. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 18, 4, (Apr. 2022), 700–789.

[2] Samad Amini et al. 2021. An Artificial Intelligence-Assisted Method for Dementia Detection Using Images from the Clock Drawing Test. *Journal of Alzheimer's Disease*, 83, 2, 581–589. Publisher: IOS Press BV. DOI: `10.3233/JAD-210299`.

[3] Sabyasachi Bandyopadhyay, Jack Wittmayer, David J. Libon, Patrick Tighe, Catherine Price, and Parisa Rashidi. 2023. Explainable semi-supervised deep learning shows that dementia is associated with small, avocado-shaped clocks with irregularly placed hands. en. *Scientific Reports*, 13, 1, (May 2023), 7384. DOI: `10.1038/s41598-023-34518-9`.

[4] Gary Bradski. 2000. The openCV library. *Miller Freeman Inc*, 25, 11, 120–123.

[5] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. 2020. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11, 2, (Feb. 2020), 125.

[6] Shuqing Chen, Daniel Stromer, Harb Alnasser Alabdalrahim, Stefan Schwab, Markus Weih, and Andreas Maier. 2020. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports 2020 10:1*, 10, 1, (Nov. 2020), 1–11. Publisher: Nature Publishing.

[7] Jeffrey L. Cummings. 2004. Alzheimer's disease. *The New England Journal of Medicine*, 351, 1, (July 2004), 56–67. DOI: `10.1056/NEJMra040223`.

[8] John D. W. Greene and John R. Hodges. 1996. Identification of famous faces and famous names in early Alzheimer's disease: Relationship to anterograde episodic and general semantic memory. *Brain*, 119, 1, (Feb. 1996), 111–128. DOI: `10.1093/brain/119.1.111`.

[9] Martin Guha. 2014. Diagnostic and Statistical Manual of Mental Disorders: DSM-5 (5th edition). *Reference Reviews*, 28, 3, (Jan. 2014), 36–37.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[11] Nina Hosseini-Kivanani, Christoph Schommer, and Luis. A Leiva. 2023. The Magic Number: Impact of Sample Size for Dementia Screening Using Transfer Learning and Data Augmentation of Clock Drawing Test Images. In *International Conference on E-health Networking, Application & Services (Healthcom)*. IEEE, China, 23–28.

[12] Nina Hosseini-Kivanani et al. 2023. Better Together: Combining Different Handwriting Input Sources Improves Dementia Screening. In *IEEE 19th International Conference on e-Science (e-Science)*. IEEE, Cyprus, 1–7.

[13] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

[14] Donato Impedovo and Giuseppe Pirlo. 2018. Dynamic Handwriting Analysis for the Assessment of Neurodegenerative Diseases: A Pattern Recognition Perspective. *IEEE reviews in biomedical engineering*, 12, 209–220. Publisher: IEEE Rev Biomed Eng.

[15] Jessica J Jalbert, Lori A Daiello, and Kate L Lapane. 2008. Dementia of the Alzheimer Type | Epidemiologic Reviews | Oxford Academic. *Epidemiologic reviews*, 30, 1, 15–34. Retrieved Nov. 16, 2022 from `https://academic.oup.com/epirev/article/30/1/15/623289`.

[16] C. Jiménez-Mesa et al. 2022. Automatic Classification System for Diagnosis of Cognitive Impairment Based on the Clock-Drawing Test. *Lecture Notes in Computer Science*, 13258 LNCS, 34–42.

[17] Yike Li, Jiajie Guo, and Peikai Yang. 2022. Developing an Image-Based Deep Learning Framework for Automatic Scoring of the Pentagon Drawing Test. *Journal of Alzheimer's disease: JAD*, 85, 1, 129–139.

[18] José Eduardo Martinelli, Juliana Francisca Cecato, Marcos Oliveira Martinelli, Brian Alvarez Ribeiro de Melo, and Ivan Aprahamian. 2018. Performance of the Pentagon Drawing test for the screening of older adults with Alzheimer's dementia. *Dementia & Neuropsychologia*, 12, 1, (Jan. 2018), 54–60. Publisher: Academia Brasileira de Neurologia, Departamento de Neurologia Cognitiva e Envelhecimento.

[19] Jumpei Maruta, Kentaro Uchida, Hideo Kurozumi, Satoshi Nogi, Satoshi Akada, Aki Nakanishi, Miki Shinoda, Masatsugu Shiba, and Koki Inoue. 2022. Deep convolutional neural networks for automated scoring of pentagon copying test results. *Scientific Reports*, 12, 1, (Dec. 2022), 9881.

[20] Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. 1984. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. en. *Neurology*, 34, 7, (July 1984), 939–939. DOI: `10.1212/WNL.34.7.939`.

[21] Mario F Mendez, Monique M Cherrier, and Robert S Meadows. 1996. Depth Perception in Alzheimer's Disease. *Perceptual and motor skills*, 83, 3, 987–995. Retrieved Nov. 17, 2022 from.

[22] Gabriel Poirier, Alice Ohayon, Adrien Juranville, France Mourey, and Jeremie Gaveau. 2021. Deterioration, Compensation and Motor Control Processes in Healthy Aging, Mild Cognitive Impairment and Alzheimer's Disease. *Geriatrics*, 6, 1, 33. Publisher: Geriatrics (Basel).

[23] Elena Salobrar-García et al. 2019. Changes in visual function and retinal structure in the progression of Alzheimer's disease. *PLOS ONE*, 14, 8, (Aug. 2019), e0220535. DOI: `10.1371/journal.pone.0220535`.

[24] Kenichiro Sato, Yoshiki Niimi, Tatsuo Mano, Atsushi Iwata, and Takeshi Iwatsubo. 2022. Automated Evaluation of Conventional Clock-Drawing Test Using Deep Neural Network: Potential as a Mass Screening Tool to Detect Individuals With Cognitive Decline. *Frontiers in Neurology*, 13, 831–831. Retrieved Jan. 18, 2023 from `https://www.frontiersin.org/articles/10.3389/fneur.2022.896403`.

[25] Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6, 1, (July 2019).

[26] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. (2014).

[27] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, (May 2019), 6105–6114.

[28] Shinya Tasaki, Namhee Kim, Tim Truty, Ada Zhang, Aron S. Buchman, Melissa Lamar, and David A. Bennett. 2023. Interpretable deep learning approach for extracting cognitive features from hand-drawn images of intersecting pentagons in older adults. en. (Apr. 2023).

[29] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13, 4, (Apr. 2004), 600–612.

[30] Victor Wasserman et al. 2020. Visuospatial performance in patients with statistically-defined mild cognitive impairment. *Journal of Clinical and Experimental Neuropsychology*, 42, 3, (Apr. 2020), 319–328. DOI: 10.1080/13803395.2020.1714550.

[31] Guangle Yao, Tao Lei, and Jiandan Zhong. 2019. A review of Convolutional-Neural-Network-based action recognition. *Pattern Recognition Letters*. Cooperative and Social Robots: Understanding Human Activities and Intentions 118, (Feb. 2019), 14–22. DOI: 10.1016/j.patrec.2018.05.018.

4.6 Efficient Automatic Data Augmentation of CDT Images to Support Cognitive Screening

Abstract[†]

We investigate the effectiveness of learnable and non-learnable automatic data augmentation (AutoDA) techniques in enhancing Deep Learning (DL) models for classifying Clock Drawing Test (CDT) images used in cognitive dysfunction screening. Specifically, we evaluate TrivialAugment (TA) and UniformAugment (UA), adapted for clinical image classification to address data scarcity and class imbalance. Our experiments across three public datasets demonstrate significant improvements in model performance and generalization. Notably, TA increased classification accuracy by up to 15%, while UA achieved a 12% improvement. These techniques offer a computationally efficient alternative to learnable methods like RandAugment (RA), which we also compare against, delivering comparable (and sometimes better) results with a much lower computational overhead. Our findings indicate that AutoDA techniques, particularly TA and UA, can be effectively applied in clinical settings, providing robust tools for early detection of cognitive disorders, including Alzheimer's disease and dementia.

Keywords

Drawing; Handwriting; Cognitive impairments; Data augmentation; Neural networks.

4.7 Introduction

Data augmentation (DA) is crucial for Deep Learning (DL) models in clinical settings, where acquiring large, labeled datasets is often challenging. By applying transformations like rotation, scaling, and cropping, DA creates diverse training samples that reduce overfitting and enhance model generalization [8, 30]. This is particularly vital in medical applications where data is scarce and imbalanced, as seen in radiology and Alzheimer's disease (AD) screening [11, 19, 27], where DA can significantly improve accuracy. Despite these advantages, the success of DA relies on preserving the clinical relevance of images. In some tasks, such as object detection for medical image analysis, traditional DA techniques have shown limitations [17]. Indeed, improper augmentation can introduce noise that

[†] **N. Hosseini-Kivanani**, I. Oliveira, S. Kilinç, and L. A. Leiva, "Efficient Automatic Data Augmentation of CDT Images to Support Cognitive Screening," in *Intelligent Data Engineering and Automated Learning – ICAART 2025, Lecture Notes in Computer Science*, vol. 3, 2025, pp. 600–607. DOI: 10.5220/0000196100003890

disrupts the learning process [18]. Therefore, while DA has demonstrated its value in healthcare [3, 25], developing more sophisticated augmentation strategies tailored to the unique challenges of medical data remains a priority.
Several studies have explored the use of drawing tasks to improve the detection of AD. These tasks capture different, complementary aspects of cognitive impairments, enhancing the automated detection of AD and mild cognitive impairment (MCI) [11, 19]. However, there remains a gap in research focusing on customizing automatic data augmentation (AutoDA) techniques for cognitive assessment tools like the Clock Drawing Test (CDT), widely used in cognitive dysfunction screening. In this paper, we address this gap by evaluating and adapting state-of-the-art AutoDA techniques for CDT images. Our aim is to maintain clinical relevance while improving model robustness. Our key contributions are:

- We adapt AutoDA techniques to the specific clinical requirements of CDT images, preserving diagnostic relevance while achieving significant improvements in detection accuracy and model generalization across three public datasets.
- By comparing learnable and non-learnable augmentation methods, we provide practical insights and guidelines for applying data augmentation effectively in cognitive dysfunction screening.

Our experimental results demonstrate that AutoDA methods achieve up to a 15% improvement in accuracy compared to models without data augmentation, depending on the dataset. The results highlight the effectiveness of applying tailored AutoDA techniques for improving the early diagnosis of cognitive impairments, such as AD and dementia. This work supports enhanced clinical decision-making and lays the foundation for more advanced diagnostic technologies in healthcare.

4.8 Related work

Traditional DA methods for images, such as random cropping, flipping, and color jittering, require manual design and domain expertise to be effective. While these basic transformations are straightforward to implement, they may not capture the complex variations needed for specialized tasks or datasets. Specialized methods, including Cutout [6], Mixup [38], and CutMix [36], have been proposed to enhance model performance by introducing more sophisticated augmentation techniques. Cutout randomly masks out square regions of input images, Mixup generates new training examples by linearly interpolating between pairs of examples, and CutMix replaces regions of an image with patches from another image along with their labels. While effective for specific tasks, transferring these methods to other tasks or datasets often requires extensive manual effort and tuning.

Recent advances have shifted towards AutoDA strategies to alleviate the manual effort involved in designing and tuning augmentation policies. Table 4.1 summarizes the state of the art. AutoAugment (AA) [4] uses reinforcement learning to search for optimal policies, yielding significant performance improvements at the cost of heavy computational resources. Fast AutoAugment (Fast AutoAugment (Fast AA)) [20] reduces this computational burden by leveraging Bayesian Optimization (BO), while Population-Based Augmentation (PBA) [10] introduces an evolutionary algorithm to explore augmentation schedules. Faster AutoAugment [9] further accelerates the process by employing differentiable policy search, but this comes with some performance degradation. This enables the selection of operations in a more efficient manner, though it comes with some degradation in performance compared to the original AutoAugment. Despite its speed improvements, Faster AA is less commonly used when performance is the top priority but is valuable for environments where efficiency is critical.

RA [5], inspired by the findings of Fast AA and PBA, simplifies automated DA by removing the need for an extensive search phase. It introduces only two hyperparameters: the number of operations (N) and their magnitude (M). RA employs random sampling from a predefined set of augmentation operations, applying them with fixed magnitudes determined through a simple grid search. This approach reduces computational overhead and allows for straightforward optimization. However, RA still requires a computationally intensive offline grid search to find optimal hyperparameters.

Similar to RA, UA [21] and TA [24] also make use of randomness to simplify the augmentation process. UA eliminates the search phase entirely by uniformly sampling augmentation operations from a predefined set and applying them with equal probability. This method hypothesizes that uniform sampling within an approximately invariant augmentation space can achieve effective results. TA applies a single random augmentation to each image, sampling the augmentation strength anew for each instance. Both methods avoid the computational complexity of search-based techniques while still benefiting from the diversity introduced by random augmentations.

The previously mentioned methods rely on randomness to enhance DA, but Augmentation-Wise Weight Sharing (AWS) [34] takes a different approach. AWS uses Neural Architecture Search (NAS) [39] for automatic augmentation search, reducing computational costs while maintaining performance with a dynamic augmentation policy that adapts during training. By focusing on augmentations in later training stages, AWS boosts model performance. However, it still demands significant computation in the initial and fine-tuning phases, which can be a challenge for those with limited resources. Another method, Model-

Adaptive Data Augmentation (MADAug) [14], adjusts augmentation policies dynamically based on model performance. It employs a policy network to select augmentations and uses a bi-level optimization scheme to adapt augmentations throughout different stages of training. Our work similarly explores when and what augmentations should be applied during training to optimize performance.

BO-Aug [37] proposes a new method for automated DA by utilizing a continuous policy search space and evaluating policy groups rather than individual policies. By using Bayesian Optimization as the search algorithm, BO-Aug achieves state-of-the-art or comparable performance with relatively low computational costs compared to AA and RA.

AutoDA	**Error (%)**	**Non-learnable**
RandAugment (RA)	15.0	No
AutoAugment (AA)	16.5	No
AWS	18.5	No
Fast AA	19.4	No
UniformAugment (UA)	19.6	Yes
MADAug	21.5	No
TrivialAugment (TA)	21.9	Yes
Faster AA	23.5	No
BO-Aug†	36.8	No

Table 4.1: Overview of AutoDA techniques for DL models, tested on ImageNet, sorted by error rate (lower is better).

†BO-Aug used Tiny ImageNet, a subset of 100k ImageNet images.

Despite the significant amount of research focused on AutoDA strategies, there is limited work specifically targeting medical images. MedAugment [22] is one of the few methods designed for medical imaging. It employs two distinct augmentation spaces: pixel-level (photometric) and spatial (geometric) transformations. MedAugment introduces a sophisticated sampling strategy that constrains the number of operations applied sequentially, ensuring that the integrity of medical image features is preserved. Unfortunately, MedAugment focuses on X-ray data, which differs significantly from hand-drawn data, such as the CDT images we are studying. Additionally, MedAugment relies on ground-truth segmentations, which are not applicable to handwriting images and require learning a DA policy, rendering it unsuitable for real-time application in DL training pipelines.

Building on these insights, our work aims to examine DA techniques for Computer Vision models applied to drawing tasks for cognitive impairment assessment, specifically AD. Our approach not only addresses the limitations of existing methods but also explores a novel domain in medical image augmentation. We focus on creating augmentation strategies that preserve the semantic content of hand-drawn elements while introducing sufficient variability to enhance model performance. By avoiding extensive computational requirements and reliance on specialized datasets, our method is suitable for real-time use and contributes to the advancement of DL applications in medical imaging.

4.9 Methodology

Our task consists of spotting early signs of cognitive decline via hand-drawn clock images. This is framed as a binary classification problem between healthy controls (HCs) and mild cognitive impairment (MCI) patients. This is a really challenging and appealing task for several reasons. First, MCIs are at high risk of progressing to dementia, although their impairments do not severely impact daily or social functioning. In fact, MCIs might remain stable or reverse to healthy cognition [1]. Second, the drawing abilities of HCs and MCIs are often on par, which makes it difficult to differentiate both groups with DL models. Third, being able to tell HCs and MCIs apart means that practitioners could start treating the patients as soon as possible, as once they are diagnosed with AD, it is irreversible.

4.9.1 Materials

The CDT is a paper-and-pencil cognitive screening tool that is quick to apply, well accepted by patients, easy to score, and independent of language, education, and culture. It also has good inter-rater and test-retest reliability, high levels of sensitivity and specificity, concurrent validity, and predictive validity [31]. In the CDT, subjects must draw a clock, including the numbers 1 to 12, as well as the clock hands, usually pointing to “10:00”, “11:10”, or similar. The drawing is then scored according to a normalized system, among which the Shulman [32] and MoCA [26] scoring systems are the most popular ones.

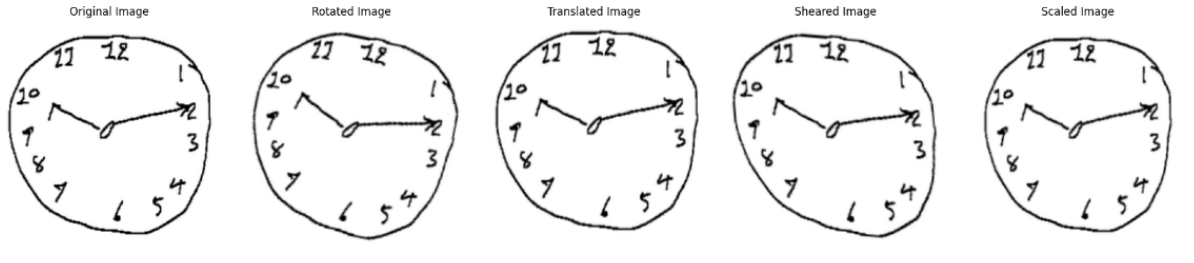


Figure 4.4: Examples of CDT images from Ruengchaijatuporn dataset before (original image) and after augmentation.

We used three publicly available CDT datasets for this study, each containing images from both HCs and MCI patients. These datasets provide a rich variety of clock images, enabling the exploration of different augmentation strategies and deep-learning models.

1. Dataset Chen [2] 2020 dataset. It contains 1,021 images categorized as HCs (n=50) and six patient subgroups. Images in subgroups 1 (n=164) and 2 (n=233) correspond to MCIs. The average age in both HCs and MCIs is 69.8 years. There are 58% females and 42% males.

2. Ruengchaijatuporn dataset [29] 2022 dataset. It contains 918 images labeled according to the MoCA score. We selected those of HCs (score of 26 or higher, n=550) and MCIs (scores between 18 and 25, n=322). The median age in both groups is 67 years. There are 77% females and 23% males.

3. Raksasat dataset [28] 2023 dataset. It contains 3,108 images categorized as six user groups. We consider group 5 ("perfect clock", n=1623) as HCs and group 4 ("minor visuospatial deficits", n=1047) as MCIs. The median age in both groups is 67 years. There are 66% females and 33% males.

To maintain consistency across all datasets, we ensured that all images had a square aspect ratio by cropping each image to its shortest dimension. This step was essential because DL models such as EfficientNet require square inputs to avoid distortion and ensure optimal performance. After cropping, the images were resized to 224×224 px, matching the input size required by pre-trained models. No additional preprocessing, such as color normalization or denoising, was applied, as the clock images are relatively clean.

4.9.2 AutoDA methods

We systematically evaluate two non-learnable AutoDA methods, TA and UA, which have demonstrated state-of-the-art performance in various computer vision tasks [24, 21]. These methods are particularly appealing for real-time applications because they do not require learning augmentation policies during training, thus reducing computational overhead. The augmentation process in both methods follows three main steps:

- Random Sampling: A set of augmentations is randomly chosen from a predefined list of operations, such as cropping, rotation, brightness adjustment, etc.

- Magnitude Randomization: The intensity of each selected augmentation is randomized within a specified range.

- Application of Augmentation: The selected augmentations are sequentially applied to the image, resulting in a modified version of the original input.

In TA, a single transformation is applied per augmented image with a randomly chosen strength. In UA, k transformations are selected, each being applied with a probability of 0.5, with a randomly picked magnitude. Following the original paper [21], we set $k = 2$. For comparison, we also evaluate RA [5], a state-of-the-art and widely used learnable AutoDA method that dynamically optimizes augmentation strategies during training. Unlike TA and UA, which rely on fixed augmentations, RA introduces two key hyperparameters: the number of augmentation operations and the magnitude, which are optimized during the training process. This learnable approach allows RA to adapt the augmentation policies based on the dataset's characteristics, making it particularly useful in domains such as medical imaging, where data scarcity and class imbalance are common challenges. In our implementation, we search the RA hyperparameters N and M over discrete sets, with N values ranging from 2 to 3 and M values ranging from 4 to 5, as part of the optimization process to find the best-performing augmentation combination. While learnable methods like RA can potentially improve model performance by adjusting augmentations to the data, non-learnable methods such as TA and UA provide a computationally efficient alternative by avoiding the complexity and overhead associated with policy optimization.

4.9.2.1 Transformation operations

A key detail in AutoDA methods is the "augmentation pool," i.e., the set of available transformation operations and their ranges. Table 4.2 details the transformations considered in the study. Only geometric transformations were applied in carefully curated ranges so as not to destroy image semantics and thus ensure clinical relevance. Transformations were applied using the Albumentations library[‡].

Transformation	**Range**	**Description**
Rotation	[-10, 10]	degrees
Shear	[0.2, 10]	degrees
Scale	[-0.05, 0.05]	% of original size
Translation	[-0.02, 0.02]	% of bounding box

Table 4.2: Overview of considered augmentation operations and transformation ranges.

4.9.3 DL models

We provide classification results according to EfficientNet [33] and DenseNet [15] as a way of establishing a common reference model for all the benchmarked AutoDA methods. Effi-

[‡] https://albumentations.ai/

cientNet, a lightweight deep learning model with 5 million parameters, has demonstrated state-of-the-art performance in various medical imaging applications. Its efficiency and scalability make it an ideal choice for this study, particularly given the relatively small size of the datasets involved. EfficientNet uses a compound scaling method to systematically balance the network's depth, width, and resolution, ensuring optimal performance across different scales.

In contrast, DenseNet features a densely connected architecture where each layer is directly connected to every other layer, promoting efficient feature reuse and enhancing gradient flow. This structure not only reduces the number of parameters but also improves the model's learning capacity by mitigating vanishing gradients and enabling the extraction of richer, more detailed feature representations. DenseNet's design is particularly advantageous for complex tasks like medical image classification, where capturing intricate patterns in the data is critical for accurate diagnosis.

By benchmarking these two architectures, we provide a comprehensive evaluation of the impact of non-learnable AutoDA methods on model performance, ensuring a robust comparison across different augmentation strategies.

The models are trained using the Adam optimizer with a learning rate of $\eta = 0.0005$. We use a batch size of 32 images, and training is carried out for up to 100 epochs. Early stopping is employed to prevent overfitting, with a patience threshold of 10 epochs. This approach ensures that training halts if validation accuracy does not improve over 10 consecutive epochs, while retaining the best-performing model weights. Balanced classification accuracy is used as the monitoring metric. Additionally, the Area Under the Receiver Operating Characteristic (AUC) curve is used to evaluate the discriminative power of the classifier, providing further insight into its performance.

		Chen dataset						**Ruengchaijatuporn dataset**						**Raksasat dataset**					
		TA		**UA**		**RA**		**TA**		**UA**		**RA**		**TA**		**UA**		**RA**	
		Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
EfficientNet	**DA train only**	85	85	80	80	84	84	58	58	56	56	59	59	77	77	78	78	78	78
	DA train + val.	**95**	**95**	90	90	80	80	58	58	58	58	60	60	**80**	**80**	77	77	79	79
	DA val. only	85	85	85	85	85	85	62	62	60	60	57	57	76	76	77	77	77	77
	DA all splits	90	90	91	91	90	90	62	62	60	60	**64**	**64**	78	78	76	76	78	78
	No DA	80 Acc. 80 AUC						56 Acc. 56 AUC						77 Acc. 77 AUC					
DenseNet	**DA train only**	90	90	89	89	89	89	53	53	59	59	50	50	71	71	68	68	67	67
	DA train + val.	90	90	90	90	89	89	54	54	49	49	56	56	71	71	69	69	64	64
	DA val. only	88	88	78	78	83	83	57	57	54	54	61	61	69	69	72	72	72	72
	DA all splits	92	92	90	90	**93**	**93**	**69**	**69**	68	68	67	67	**75**	**75**	74	74	72	72
	No DA	65 Acc. 65 AUC						55 Acc. 55 AUC						67 Acc. 67 AUC					

Table 4.3: Performance results on three public datasets. For each dataset, the best result is highlighted in boldface.

4.9.4 Procedure

We split each dataset into three randomly disjoint sets: 70% training, 20% validation, and 10% testing. The testing set is reserved as a held-out partition that is used only after a model is trained since it simulates unseen data. The splits are also stratified to ensure that HC and MCI images are evenly allocated to the training/validation/testing sets.

In this work, we investigate five different DA conditions for the training and evaluation of our models. The baseline condition, **No DA**, involves no DA at all, where the model is trained, validated, and tested on the original, non-augmented data. The first augmentation condition, **DA train only**, applies DA solely to the training set, leaving the validation and test sets unmodified. This allows the model to benefit from augmented samples during training while preserving the original, unaltered validation and test sets for unbiased evaluation. The second condition, **DA train + val.**, applies DA to both the training and validation sets, enabling the model to generalize better by encountering augmented samples in both phases while still maintaining a pristine test set. The third condition, **DA val. only**, applies augmentation solely to the validation set, allowing the original training and test sets to remain unaltered. Finally, in **DA all splits**, DA is applied to all three partitions—training, validation, and test—offering the most challenging scenario where the model is trained, validated, and evaluated with real and augmented data.

In each condition, DA is applied by ensuring that the majority class has 10% more instances than in the original dataset and matching the number of instances in the minority class. This way, we address both class imbalance and data scarcity issues during model training.

4.10 Results

Table 4.3 compares the performance of various DA strategies across two deep learning architectures (EfficientNet and DenseNet) and three benchmark datasets: Chen dataset, Ruengchaijatuporn dataset, and Raksasat dataset. We evaluate the effects of TA, UA, and RA under multiple augmentation regimes.

EfficientNet shows the best performance in the Chen dataset, achieving the highest accuracy and AUC of 95% under the DA train + val strategy. This indicates that EfficientNet is particularly effective on smaller datasets when both training and validation data are augmented. However, DenseNet performs better in the Ruengchaijatuporn and Raksasat datasets, especially under the DA all splits strategy, where it achieves 69% accuracy and AUC (Ruengchaijatuporn) and 75% accuracy and AUC (Raksasat), outperforming

EfficientNet on these more complex datasets.

- **Performance on Chen dataset**: EfficientNet: The highest performance was achieved with DA train + val, reaching 95% accuracy and 95% AUC, a significant improvement over the baseline (No DA) of 80% accuracy and 80% AUC. Augmenting only the training set yielded an accuracy of 85%, demonstrating that augmenting the validation set can help mitigate overfitting and improve generalization. DenseNet: The best performance was observed with DA all splits, reaching 92% accuracy and 93% AUC. While DenseNet performed well, EfficientNet had slightly better results on this dataset.

- **Performance on Ruengchaijatuporn dataset**: EfficientNet: The most significant improvement occurred with TA, where accuracy improved by 12% from 56% (No DA) to 62%. However, the model struggled compared to DenseNet under most DA strategies. DenseNet: DenseNet outperformed EfficientNet across all DA regimes, particularly under DA all splits, where it reached 69% accuracy and AUC. UA also led to strong results, with DenseNet achieving 68% accuracy, demonstrating its robustness in handling this highly imbalanced dataset.

- **Performance on Raksasat dataset**: EfficientNet: EfficientNet: The best results were observed with DA train + val, achieving 80% accuracy and AUC. RA, applied to the training set only, yielded 78% accuracy, but its performance was inconsistent across other strategies. DenseNet achieved the best performance under the DA all splits condition with 75% accuracy and 75% AUC.

Overall, the results show that EfficientNet performs well on datasets like Chen, achieving the highest accuracy and AUC, especially with DA train + val. However, DenseNet performs better on more complex datasets like Ruengchaijatuporn, consistently achieving higher accuracy and AUC, particularly when using DA all splits, as seen in the Ruengchaijatuporn dataset (69% accuracy, 69% AUC).

Overall, the results show that EfficientNet performs well on datasets like Chen, achieving the highest accuracy and AUC, especially with DA train + val. While DenseNet performs better on more complex datasets like Ruengchaijatuporn, consistently achieving higher accuracy and AUC (69% accuracy, 69% AUC), EfficientNet outperforms DenseNet on the Raksasat dataset, with its best performance in the DA train + val. condition (80% accuracy, 79% AUC), compared to DenseNet's best performance of 75% accuracy, 72% AUC under the DA all splits condition.

The improved generalization, particularly with TA and UA on imbalanced datasets like Ruengchaijatuporn dataset, highlights the potential for these techniques to be applied in real-world clinical environments. Simpler augmentation strategies (like TA) performed comparably or better than more complex approaches (like RA), emphasizing the importance of efficiency and resource constraints in clinical decision-making.

4.11 Discussion

Our results show that applying non-learnable data augmentation techniques, particularly TA and UA, significantly boosts the performance of DL models for CDT image classification in cognitive dysfunction screening. These findings are evident across three public datasets—Chen, Ruengchaijatuporn, and Raksasat.

On the Chen dataset, EfficientNet demonstrated superior performance, particularly when both training and validation splits were augmented, achieving an accuracy of 95% and AUC of 95%. This suggests that EfficientNet is highly effective in simpler dataset structures, leveraging its architecture to maximize the benefits of DA. Conversely, DenseNet consistently outperforms EfficientNet in handling more complex datasets such as Ruengchaijatuporn, where it shows up to a 14% increase in accuracy and 14% improvement in AUC compared to EfficientNet. This superior performance can be attributed to DenseNet's capacity to reuse features more effectively across layers, which enhances generalization in complex clinical datasets characterized by limited data and inherent variability. However, on the Raksasat dataset, the results slightly diverge. While DenseNet achieved its best performance under the DA all splits condition with 75% accuracy and 72% AUC, EfficientNet slightly outperformed DenseNet under the DA train + val. condition, achieving 80% accuracy and 80% AUC.

Our findings are consistent with prior work in medical imaging, where augmentation strategies have been shown to enhance model performance by diversifying training data. Dutta et al. [7] reported similar performance improvements in radiological classification tasks using data augmentation, while Tufail et al. [35] demonstrated the role of augmentation in enhancing Alzheimer's disease detection. These results confirm the broad applicability of TA and UA beyond CDT screening, indicating their potential utility across clinical domains reliant on image-based diagnostics.

Moreover, the results of the Ruengchaijatuporn dataset highlight the importance of selecting appropriate augmentation strategies for imbalanced datasets. TA led to a 12% improvement in accuracy, showcasing its capability to handle dataset imbalance effectively. UA, while achieving robust performance with a 68% accuracy, further demonstrates that simpler augmentation strategies can be highly effective in clinical applications where data is limited and heavily skewed. This finding echoes prior research by Shorten and Khoshgoftaar [30], who stressed the importance of augmentation in handling class imbalances. Although RandomAugment (RA) provided some gains, especially in the Ruengchaijatuporn dataset (64% AUC for EfficientNet), its improvements were less consistent compared to TA and UA. This reinforces the practical benefits of non-learnable methods, which offer a better balance between computational efficiency and performance gains in clinical applications. Lim et al. [20] demonstrated that simpler augmentation methods, such as Fast AutoAugment, can match or exceed the performance of more complex, learned strategies while requiring significantly fewer computational resources. This aligns with our findings, where non-learnable methods provided comparable performance to RA, but with much lower complexity and computational costs.

Another key takeaway from our results is the effectiveness of selective augmentation strategies. Applying augmentation to both training and validation sets (DA train + val.) consistently yielded the best performance across all datasets for EfficientNet, while DenseNet excelled with DA all splits in more complex datasets. Conversely, augmenting only the training set (DA train only) delivered strong results on the Raksasat dataset for EfficientNet, with 80% accuracy and AUC, underscoring the efficiency of targeted augmentation. These results suggest that over-augmenting validation and test sets can introduce noise, as observed in Ruengchaijatuporn, where DA all splits resulted in only marginal improvements (69% accuracy and AUC for DenseNet), consistent with Chlap et al. [3], who cautioned against over-augmentation in medical imaging due to potential overfitting and biased model evaluations.

Overall, this study presents strong evidence that non-learnable augmentation methods, such as TA and UA, are not only computationally efficient but also highly effective in improving model performance for medical image classification tasks. By enhancing model generalization across various datasets, these techniques hold significant promise for real-time healthcare applications where accurate and timely decision-making is critical.

4.11.1 Limitations and Future work

One limitation is that the AutoDA techniques we selected for analysis are those suitable for real-time (TA and UA) or near real-time (RA) processing. There are many other approaches that are learnable and have achieved slightly better performance on common benchmarks, such as ImageNet (Table 4.1), but unfortunately, they are too slow to be usable in practice. Also, it remains unclear if the results achieved on ImageNet would transfer to the medical domain. The research literature suggests otherwise [16, 23, 32].

Another limitation of our work is that we have considered only one type of drawing to support cognitive dysfunction screening, albeit the most popular one. Future work should

go beyond CDTs to better assess the generalizability of AutoDA methods. For example, some drawings, like Pentagon Drawing Test (PDT) images, allow for other DA operations such as vertical and horizontal flipping [13, 12].
Furthermore, investigating the effectiveness of AutoDA techniques across multiple domains can reveal further insights into their potential for improving model performance in other computer vision applications. Future research could explore the integration of non-learnable methods with semi-supervised learning approaches to further improve performance, particularly in scenarios where labeled data is scarce. Expanding the application of these augmentation strategies to other diagnostic fields, such as neuroimaging and pathology, could unlock further potential and lead to advancements in clinical diagnostics.

4.12 Conclusion

This study highlights the value of non-learnable AutoDA methods in improving the performance and generalization of DL models for cognitive dysfunction screening using CDT images. TA and UA, in particular, demonstrated significant improvements in model accuracy while maintaining computational efficiency, making them well-suited for medical image analysis. Our results indicate that augmentation strategies must be carefully tailored to the input data and task at hand, particularly in the medical domain, where preserving the integrity of diagnostic features is paramount. By addressing these challenges, our work contributes to the advancement of DL-based diagnostic tools and sets the stage for further exploration of augmentation strategies in medical imaging.

Acknowledgments

Research supported by the Horizon 2020 FET program of the European Union through the ERA-NET Cofund funding (grant CHIST-ERA-20-BCI-001) and the European Innovation Council Pathfinder program (SYMBIOTIK project, grant 101071147).

References for Paper 7

[1] Emilie Blair, Darin Zahuranec, Kenneth M. Langa, Jane Forman, Bailey K. Reale, Colleen Kollman, Bruno Giordani, and Deborah A. Levine. 2022. Impact of patient mild cognitive impairment on physician decision-making for treatment. *Journal of Alzheimer's Disease*, 78, 4.

[2] Shuqing Chen, Daniel Stromer, Harb Alnasser Alabdalrahim, Stefan Schwab, Markus Weih, and Andreas Maier. 2020. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports*, 10, 1.

[3] P. Chlap, Hang Min, Nym Vandenberg, J. Dowling, L. Holloway, and A. Haworth. 2021. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65.

[4] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. 2019. Autoaugment: learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

[5] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. 2020. Randaugment: practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*.

[6] Terrance Devries and Graham W. Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *CoRR*. arXiv:1708.04552.

[7] S. Dutta, P. Prakash, and C. Matthews. 2020. Impact of data augmentation techniques on a deep learning based medical imaging task. 11318, 113180M - 113180M–10. DOI: 10.1117/12.2549806.

[8] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321.

[9] Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. 2020. Faster autoaugment: learning augmentation strategies using backpropagation. In *Computer Vision–ECCV 2020: 16th European Conference*.

[10] Daniel Ho, Eric Liang, Ion Stoica, P. Abbeel, and Xi Chen. 2019. Population based augmentation: efficient learning of augmentation policy schedules. *CoRR*. arXiv:1905.05393.

[11] Nina Hosseini-Kivanani, Elena Salobrar-García, Lorena Elvira-Hurtado, Mario Salas, Christoph Schommer, and Luis A. Leiva. 2024. Predicting alzheimer's disease and mild cognitive impairment with off-line and on-line house drawing tests. In *2024 IEEE 20th International Conference on e-Science (e-Science)*.

[12] Nina Hosseini-Kivanani et al. 2023. Better Together: Combining Different Handwriting Input Sources Improves Dementia Screening. In *2023 IEEE 19th International Conference on e-Science (e-Science)*.

[13] Nina Hosseini-Kivanani et al. 2024. Ink of insight: data augmentation for dementia screening through handwriting analysis. In *Proceedings of the 2024 8th International Conference on Medical and Health Informatics*, 224–229.

[14] Chengkai Hou, Jieyu Zhang, and Tianyi Zhou. 2023. When to learn what: model-adaptive data augmentation curriculum. *CoRR*. arXiv:2309.04747.

[15] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[16] F. Jonske et al. 2023. Why does my medical ai look at pictures of birds? exploring the efficacy of transfer learning across domain boundaries. *ArXiv*, abs/2306.17555. DOI: `10.48550/arXiv.2306.17555`.

[17] Aghiles Kebaili, J. Lapuyade-Lahorgue, and S. Ruan. 2023. Deep learning approaches for data augmentation in medical imaging: a review. *Journal of Imaging*, 9.

[18] Byungchan Ko and Jungseul Ok. 2021. Time matters in using data augmentation for vision-based deep reinforcement learning. *CoRR*. arXiv:2102.08581.

[19] Masatomo Kobayashi, Yasunori Yamada, Kaoru Shinkawa, M. Nemoto, K. Nemoto, and T. Arai. 2022. Automated early detection of alzheimer's disease by capturing impairments in multiple cognitive domains with multiple drawing tasks. *Journal of Alzheimer's Disease*, 88.

[20] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. 2019. Fast autoaugment. In *Advances in neural information processing systems*.

[21] Tom Ching LingChen, Ava Khonsari, Amirreza Lashkari, Mina Rafi Nazari, Jaspreet Singh Sambee, and Mario A. Nascimento. 2020. Uniformaugment: a search-free probabilistic data augmentation approach. *CoRR*. arXiv:2003.14348.

[22] Zhaoshan Liu, Qiujie Lv, Yifan Li, Ziduo Yang, and Leizhao Shen. 2023. Medaugment: universal automatic data augmentation plug-in for medical image analysis. *CoRR*. arXiv:2306.17466.

[23] Mohammad Amin Morid, Alireza Borjali, and Guilherme Del Fiol. 2021. A scoping review of transfer learning research on medical image analysis using imagenet. *Computers in biology and medicine*. DOI: `10.1016/J.COMPBIOMED.2020.104115`.

[24] S. G. Muller and F. Hutter. 2021. Trivialaugment: tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*.

[25] J. Nalepa, Michal Marcinkiewicz, and M. Kawulok. 2019. Data augmentation for brain-tumor segmentation: a review. *Frontiers in computational neuroscience*, 13.

[26] Ziad S. Nasreddine, Natalie A. Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L. Cummings, and Howard Chertkow. 2019. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 11, 1.

[27] R. Ogawa, T. Kido, and T. Mochizuki. 2019. Effect of augmented datasets on deep convolutional neural networks applied to chest radiographs. *Clinical radiology*.

[28] Raksit Raksasat, Surat Teerapittayanon, Sirawaj Itthipuripat, Kearkiat Praditpornsilpa, Aisawan Petchlorlian, Thiparat Chotibut, Chaipat Chunharas, and Itthi Chatnuntawech. 2023. Attentive pairwise interaction network for ai-assisted clock drawing test assessment of early visuospatial deficits. *Scientific Reports*, 13, 1.

[29] Natthanan Ruengchaijatuporn et al. 2022. An explainable self-attention deep neural network for detecting mild cognitive impairment using multi-input digital drawing tasks. *Alzheimer's Research & Therapy*, 78, 14.

[30] Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6.

[32] Kenneth I Shulman, Dolores Pushkar Gold, Carole A Cohen, and Carla A Zucchero. 1993. Clock-drawing and dementia in the community: a longitudinal study. *International journal of geriatric psychiatry*, 8, 6, 487–496.

[31] Bárbara Spenciere, Heloisa Alves, and Helenice Charchat-Fichman. 2017. Scoring systems for the clock drawing test: a historical review. *Dementia & neuropsychologia*, 11, 1.

[32] M. Taher, Fatemeh Haghighi, Ruibin Feng, M. Gotway, and Jianming Liang. 2021. A systematic benchmarking analysis of transfer learning for medical image analysis. *Domain adaptation and representation transfer, and affordable healthcare and AI for resource diverse global health : third MICCAI workshop, DART 2021 and first MICCAI workshop, FAIR 2021 : held in conjunction with MICCAI 2021 : Strasbou...*, 12968, 3–13. DOI: 10.1007/978-3-030-87722-4_1.

[33] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International conference on machine learning*.

[34] Keyu Tian, Chen Lin, Ming Sun, Luping Zhou, Junjie Yan, and Wanli Ouyang. 2020. Improving auto-augment via augmentation-wise weight sharing. *CoRR*. arXiv:2009.14737.

[35] Ahsan Bin Tufail, Kalim Ullah, Rehan Ali Khan, M. Shakir, Muhammad Abbas Khan, Inam Ullah, Yong-Kui Ma, and Md. Sadek Ali. 2022. On improved 3d-cnn-based binary and multiclass classification of alzheimer's disease using neuroimaging modalities and data augmentation methods. *Journal of Healthcare Engineering*, 2022. DOI: 10.1155/2022/1302170.

[36] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. 2019. Cutmix: regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*.

[37] Chunxu Zhang, Ximing Li, Zijian Zhang, Jiaxu Cui, and Bo Yang. 2022. Bo-aug: learning data augmentation policies via bayesian optimization. *Applied Intelligence*, 53.

[38] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. 2017. Mixup: beyond empirical risk minimization. *CoRR*. arXiv:1710.09412.

[39] Barret Zoph and Quoc V. Le. 2017. Neural architecture search with reinforcement learning. *CoRR*. arXiv:1611.01578.

5 Conclusion and Outlook

5.1 Summary of Main Findings

This dissertation evaluates the performance of ML and DL models in differentiating AD and MCI from healthy individuals through handwriting-based cognitive assessments. Both static (scanned images) and dynamic (discrete point sequences) drawing tasks were investigated, encompassing multiple input sources such as the CDT, PDT, HDT, signature, and sentence-writing.

A principal observation is that off-line handwriting data surpasses on-line data, with binary classification reaching 90% accuracy and multi-class classification 82%. This result contradicts earlier expectations that the temporal nuances of dynamic stroke data would yield superior predictive value. Instead, image-based representations exhibit greater stability, clarity, and discriminative capability, especially in contexts where data are limited.

To address the challenge of data scarcity, the study investigates augmentation methodologies. Non-learnable DA techniques, such as TA and UA, enhance classification performance by as much as 15%, with the magnitude of improvement varying in relation to dataset intricacy and task demands. Within the DL models evaluated, EfficientNet achieves an accuracy of 95% on well-structured datasets, while DenseNet demonstrates superior generalization on imbalanced and more heterogeneous datasets.

The comparative analysis between static and dynamic modalities reaffirms the clinical dependability of scanned handwriting samples. Although dynamic data encapsulates richer temporal information, the resulting features tend to be noisier and less effective for classification. These outcomes highlight the necessity for feature extraction, preprocessing, and the development of integrated architectures that combine both spatial and temporal information.

Overall, the findings advance the field of AI-assisted cognitive screening by guiding the selection of appropriate models, data augmentation strategies, and preprocessing protocols for neurocognitive assessments.

5.1.1 Drawnstic: AI-Assisted Cognitive Screening through Digital Handwriting

This study introduces Drawnstic, an AI-powered cognitive screening tool that leverages handwriting analysis for the early detection of AD and MCI. By integrating ML models with digital handwriting processing, Drawnstic offers a quantitative, objective framework for assessing cognitive deterioration, supplementing traditional screening protocols with enhanced precision and facilitating timely clinical intervention.

The system supports both off-line and on-line handwriting input, allowing users to either draw on a digital canvas or upload scanned paper-based tests. Deep learning models extract critical handwriting features, including stroke pressure, velocity, and spatial patterns, enabling early detection of subtle motor and cognitive impairments.

The system accommodates both off-line and on-line handwriting inputs, allowing users to either upload paper-based cognitive tests or draw directly on a digital canvas. DL models extract and analyze key handwriting characteristics—including spatial trajectory, and —to identify subtle neurocognitive and motor impairments indicative of early-stage dementia.

Designed for accessibility and scalability, Drawnstic is optimized for tablet-based touchscreen interfaces to ensure ease of use across diverse populations. Its telehealth integration enables remote monitoring, while seamless Electronic Health Record (EHR) interoperability allows for real-time clinical reporting and data-driven decision-making.

5.1.1.1 System Infrastructure and Implementation

Drawnstic is implemented as a web-based framework with a Vue 3–driven front-end and a Flask-powered back-end, ensuring scalability, modularity and cross-platform adaptability. The front-end, constructed with JavaScript, HTML, and CSS, incorporates Pinia for state management and is configured for larger tablet-based interfaces, touch-based interfaces (>7 inches) to facilitate touch-based handwriting input. Accessibility features—including screen-reader compatibility, adjustable typography, high-contrast visual modes, and semantic markup—enhance usability for individuals with visual or motor impairments.

The back-end governs user authentication, session oversight, and data storage, using Argon2 encryption to fortify credential security and utilizing a SQL-based relational database (SQLite for development, PostgreSQL for deployment). Image metadata is systematically purged to safeguard user confidentiality, and uploaded files undergo standardization to maintain consistency across varying hardware. A Flask-based API mediates interactions between the user interface and the processing engine, managing image submissions, stroke analysis, and diagnostic evaluations.

For handwriting examination, Drawnstic integrates DL architectures: EfficientNet for static handwriting assessments, and a bidirectional GRU network for temporal data interpretation. The system accommodates diverse input mechanisms—including digital canvases, optical capture, file uploads, and drag-and-drop functionality—while a hybrid ensemble of convolutional and recurrent networks generates classification outputs, confidence estimations, and task completion metrics.

5.1.1.2 Authentication and Patient Management

Ensuring data security and user authentication is a core aspect of Drawnstic's infrastructure. The system implements secure login mechanisms that enable users to register, log in, log out, or delete accounts. Authentication is session-based, using encrypted session tokens, while credentials are securely stored using Argon2 password hashing. Patients maintain control over their personal data, with options to inspect, modify, or expunge records in compliance with data protection mandates.

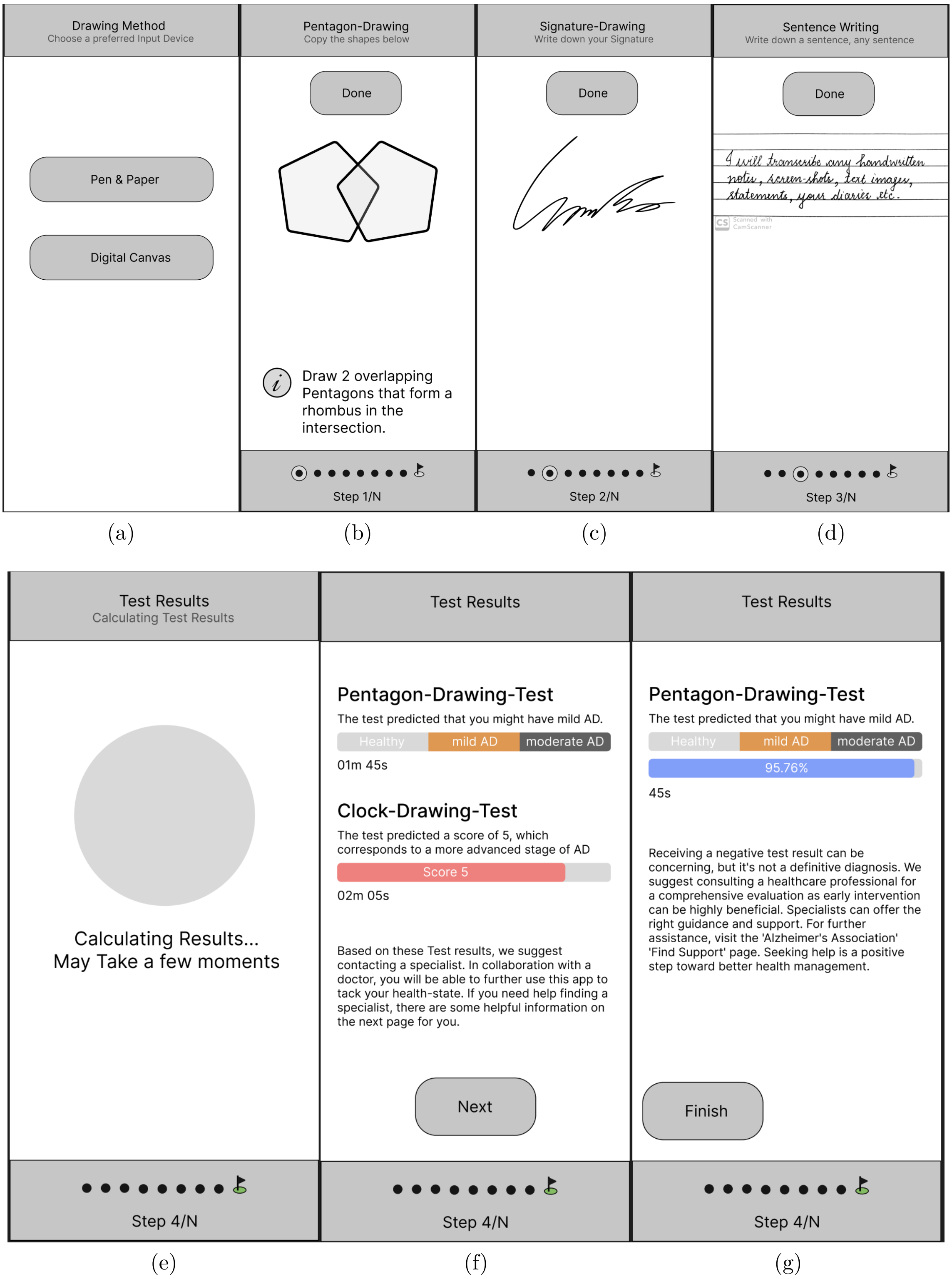
To improve usability for elderly individuals and those with cognitive impairments, Drawnstic incorporates alternative authentication mechanisms, including OAuth2 third-party authorization and biometric verification. Robust password recovery procedures ensure continuous accessibility without compromising security. Furthermore, temporal handwriting data is anonymized—retaining only essential stroke coordinates—to uphold privacy, while detailed access logs and security protocols mitigate unauthorized breaches.

Role-based access control allows caregivers and clinicians to manage multiple patient records. A hierarchical authorization model ensures that only authorized users (e.g., doctors, nurses, or designated caregivers) can access sensitive patient data. Patients maintain control over their personal information with options to review, update, or delete records in accordance with GDPR regulations.

5.1.1.3 Clinical Integration and Future Enhancements

Building upon prior AI-assisted cognitive assessment platforms such as NeuroQWERTY, DCTclock, and Visuospatial Clock Drawing Test systems, Drawnstic differentiates itself by incorporating unconstrained handwriting analysis, capturing a broader array of cognitive and motor deficits. While dynamic handwriting evaluation presents inherent challenges—primarily due to sensor limitations restricting data capture to spatial coordinates—forthcoming enhancements aim to incorporate richer temporal markers, refine augmentation methodologies, and develop hybrid architectures that integrate spatial and temporal features more effectively.

Beyond AD screening, expanding Drawnstic's diagnostic capabilities to PD and vascular dementia will increase its clinical utility. Multilingual handwriting analysis, cross-cultural dataset expansion, and adaptive AI models will ensure robustness across diverse populations. Additionally, gamification elements and cognitive engagement tasks may improve user compliance and data quality.



Draw 2 overlapping Pentagons that form a rhombus in the intersection.

I will transcribe any handwritten notes, screen-shots, text images, statements, your diaries etc.

(a) (b) (c) (d)

Calculating Results...
May Take a few moments

Pentagon-Drawing-Test

The test predicted that you might have mild AD.

01m 45s

Clock-Drawing-Test

The test predicted a score of 5, which corresponds to a more advanced stage of AD

02m 05s

Based on these Test results, we suggest contacting a specialist. In collaboration with a doctor, you will be able to further use this app to tack your health-state. If you need help finding a specialist, there are some helpful information on the next page for you.

Pentagon-Drawing-Test

The test predicted that you might have mild AD.

45s

Receiving a negative test result can be concerning, but it's not a definitive diagnosis. We suggest consulting a healthcare professional for a comprehensive evaluation as early intervention can be highly beneficial. Specialists can offer the right guidance and support. For further assistance, visit the 'Alzheimer's Association' 'Find Support' page. Seeking help is a positive step toward better health management.

(e) (f) (g)

Figure 5.1: Schematic of a sample drawing task in the Drawnstic application.

5.2 Outlook

This dissertation establishes off-line and on-line handwriting analysis, particularly in different drawing tasks, as a robust and accessible mechanism for the early detection of ADs. Its inherent simplicity renders it deployable in primary healthcare settings and non-specialist environments, offering an economically viable alternative to invasive diagnostic procedures.

A key insight from this work is the differential diagnostic potential of off-line and on-line handwriting data. Off-line drawing tasks, processed as static images, yield higher classification accuracy due to their stability and structured spatial features. In contrast, on-line handwriting captures temporal stroke dynamics, which, while rich in neuromotor information, often suffer from noise and variability due to sensor limitations. Beyond integrating these modalities, multimodal fusion with other diagnostic signals (e.g., speech analysis, gaze tracking, neuroimaging, and motor coordination data) presents an exciting direction for enhancing the specificity and sensitivity of AI-driven cognitive screening.

Another significant research direction lies in personalized AI-driven screening, as handwriting variability is influenced by demographic and cognitive factors such as age, education level, linguistic background, and motor function. The development of adaptive diagnostic models that adjust to individual cognitive baselines could mitigate biases and enhance generalization across diverse patient populations. Additionally, refining preprocessing techniques and feature extraction strategies for both static and dynamic handwriting data remains an important challenge for improving model robustness.

The integration of real-time handwriting assessment tools into routine cognitive health checkups presents an exciting avenue for early disease detection and clinical decision-making. By combining AI-driven handwriting analysis with medical expertise, these advancements will optimize neurodegenerative disease screening, improving early detection, patient outcomes, and healthcare resource allocation.

While this dissertation primarily focuses on AD and MCI screening, handwriting-based AI models hold broader potential for detecting other neurological and psychiatric conditions, such as:

- PD (micrographia detection through signature and letter-writing analysis).
- Stroke rehabilitation monitoring (tracking motor recovery via drawing tasks).
- Dysgraphia and neurodevelopmental disorders (evaluating fine motor impairments in children and adults).

- Cognitive decline prediction (longitudinal handwriting changes as early biomarkers of neurodegeneration).

In conclusion, this dissertation marks a critical step toward AI-assisted cognitive diagnostics, yet significant work remains to bridge the gap between algorithmic innovation and clinical utility. The next phase of research must focus on multimodal integration, adaptive modeling, regulatory readiness, and real-world deployment strategies. By addressing these challenges, AI-driven handwriting analysis could emerge as a cornerstone of scalable, cost-effective, and non-invasive cognitive health screening—revolutionizing early detection and disease management in aging populations.

Bibliography

[1] Samad Amini, Lifu Zhang, Boran Hao, Aman Gupta, Mengting Song, Cody Karjadi, Rhoda Au, and Ioannis Paschalidis. 2021. Remote diagnosis of dementia using ai methods on clock drawing images. *Alzheimer's & Dementia*, 17, e051911.

[2] Joao Apostolo et al. 2016. Mild cognitive decline. a position statement of the cognitive decline group of the european innovation partnership for active and healthy ageing (eipaha). *Maturitas*, 83, 83–93.

[3] Taylor Archibald, Mason Poggemann, Aaron Chan, and Tony Martinez. 2021. Trace: a differentiable approach to line-level stroke recovery for offline handwritten text. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part III 16*. Springer, 414–429.

[4] Alireza Atri. 2019. Current and future treatments in alzheimer's disease. In *Seminars in neurology* number 02. Vol. 39. Thieme Medical Publishers, 227–240.

[5] Henry Brodaty et al. 2017. Operationalizing the diagnostic criteria for mild cognitive impairment: the salience of objective measures in predicting incident dementia. *The American Journal of Geriatric Psychiatry*, 25, 5, 485–497.

[6] Francesca R Buccellato, Marianna D'Anca, Gianluca Martino Tartaglia, Massimo Del Fabbro, Elio Scarpini, and Daniela Galimberti. 2023. Treatment of alzheimer's disease: beyond symptomatic therapies. *International Journal of Molecular Sciences*, 24, 18, 13900.

[7] Adrià Casamitjana, Paula Petrone, Alan Tucholka, Carles Falcon, Stavros Skouras, José Luis Molinuevo, Verónica Vilaplana, Juan Domingo Gispert, Alzheimer's Disease Neuroimaging Initiative, et al. 2018. Mri-based screening of preclinical alzheimer's disease for prevention clinical trials. *Journal of Alzheimer's Disease*, 64, 4, 1099–1112.

[8] Joyce YC Chan, Tak Kit Chan, and Kelvin KF Tsoi. 2020. Comparison of the different approaches of drawing test for the detection of dementia and mild cognitive impairment: a systematic review and meta-analysis: dementia care research (research projects; nonpharmacological)/assessment and care planning. *Alzheimer's & Dementia*, 16, e040852.

[9] Joyce YC Chan, Sarah TY Yau, Timothy CY Kwok, and Kelvin KF Tsoi. 2021. Diagnostic performance of digital cognitive tests for the identification of mci and dementia: a systematic review. *Ageing Research Reviews*, 72, 101506.

[10] Wen-Ting Cheah, Jwu-Jia Hwang, Sheng-Yi Hong, Li-Chen Fu, Yu-Ling Chang, Ta-Fu Chen, I-An Chen, Chun-Chen Chou, et al. 2022. A digital screening system for alzheimer disease based on a neuropsychological test and a convolutional neural network: system development and validation. *JMIR Medical Informatics*, 10, 3, e31106.

[11] Shuqing Chen, Daniel Stromer, Harb Alnasser Alabdalrahim, Stefan Schwab, Markus Weih, and Andreas Maier. 2020. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports*, 10, 1, 20854.

[12] Gabriele Cipriani, Sabrina Danti, Lucia Picchi, Angelo Nuti, and Mario Di Fiorino. 2020. Daily functioning and dementia. *Dementia & neuropsychologia*, 14, 2, 93–102.

[13] Xinyi Cui, Hai Lin, et al. 2024. Recent advances on early diagnosis and immunotherapy drug for alzheimer's disease. *MEDS Basic Medicine*, 2, 1, 131–138.

[14] Anis Davoudi, Catherine Dion, Shawna Amini, Patrick J Tighe, Catherine C Price, David J Libon, and Parisa Rashidi. 2021. Classifying non-dementia and alzheimer's disease/vascular dementia patients using kinematic, time-based, and visuospatial parameters: the digital clock drawing test. *Journal of Alzheimer's Disease*, 82, 1, 47–57.

[15] Moises Diaz, Miguel Angel Ferrer, Donato Impedovo, Giuseppe Pirlo, and Gennaro Vessio. 2019. Dynamically enhanced static handwriting representation for parkinson's disease detection. *Pattern Recognition Letters*, 128, 204–210.

[16] Liyuan Fan et al. 2020. New insights into the pathogenesis of alzheimer's disease. *Frontiers in neurology*, 10, 1312.

[17] Xuanhan Fan, Menghui Zhou, Jun Qi, Yun Yang, and Po Yang. 2024. A multi-target multi-task approach based on correlated multiple cognitive scores for ad progression prediction. In *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[18] Houqian Gao, Mingzhu Liu, Zijian Zhao, Caixia Yang, Ling Zhu, Yanning Cai, Yanlian Yang, and Zhiyuan Hu. 2020. Diagnosis of mild cognitive impairment and alzheimer's disease by the plasma and serum amyloid-beta 42 assay through highly sensitive peptoid nanosheet sensor. *ACS applied materials & interfaces*, 12, 8, 9693–9700.

[19] Peyvand Ghaderyan, Ataollah Abbasi, and Sajad Saber. 2018. A new algorithm for kinematic analysis of handwriting data; towards a reliable handwriting-based tool for early detection of alzheimer's disease. *Expert Systems with Applications*, 114, 428–440.

[20] Michel Goedert and Maria Grazia Spillantini. 2006. A century of alzheimer's disease. *science*, 314, 5800, 777–781.

[21] Veer B Gupta and Vivek K Gupta. 2016. Impaired energy metabolism: involvement in neurodegenerative processes and cns ageing.

[22] Yahia Hamdi, Hanen Akouaydi, Houcine Boubaker, and Adel M Alimi. 2022. Handwriting quality analysis using online-offline models. *Multimedia Tools and Applications*, 81, 30, 43411–43439.

[23] Anika Heimann-Steinert, Antje Latendorf, Alexander Prange, Daniel Sonntag, and Ursula Müller-Werdan. 2021. Digital pen technology for conducting cognitive assessments: a cross-over study with older adults. *Psychological Research*, 85, 3075–3083.

[24] Yujun Hou, Xiuli Dan, Mansi Babbar, Yong Wei, Steen G Hasselbalch, Deborah L Croteau, and Vilhelm A Bohr. 2019. Ageing as a risk factor for neurodegenerative disease. *Nature Reviews Neurology*, 15, 10, 565–581.

[25] Donato Impedovo and Giuseppe Pirlo. 2018. Dynamic handwriting analysis for the assessment of neurodegenerative diseases: a pattern recognition perspective. *IEEE reviews in biomedical engineering*, 12, 209–220.

[26] Donato Impedovo, Giuseppe Pirlo, Gennaro Vessio, and Maria Teresa Angelillo. 2019. A handwriting-based protocol for assessing neurodegenerative dementia. *Cognitive computation*, 11, 576–586.

[27] Syed Fahad Javaid, Clarissa Giebel, Moien AB Khan, and Muhammad Jawad Hashim. 2021. Epidemiology of alzheimer's disease and other dementias: rising global burden and forecasted trends. *F1000Research*, 10, 425.

[28] Amy Jenkins, Stephen Lindsay, Parisa Eslambolchilar, Ian M Thornton, and Andrea Tales. 2016. Administering cognitive tests through touch screen tablet devices: potential issues. *Journal of Alzheimer's disease*, 54, 3, 1169–1182.

[29] Danko Jeremic, Lydia Jiménez-Díaz, and Juan D Navarro-López. 2021. Past, present and future of therapeutic strategies against amyloid-β peptides in alzheimer's disease: a systematic review. *Ageing research reviews*, 72, 101496.

[30] Carmen Jiménez-Mesa et al. 2023. Using explainable artificial intelligence in the clock drawing test to reveal the cognitive impairment pattern. *International Journal of Neural Systems*, 33, 04, 2350015.

[31] Shridevi Karande and Vrushali Kulkarni. 2024. Prognosis of dementia using early fusion approach with digital clock drawing and trail-making tests.

[32] Jacek Kawa, Adam Bednorz, Paula Stępień, Jarosław Derejczyk, and Monika Bugdol. 2017. Spatial and dynamical handwriting analysis in mild cognitive impairment. *Computers in Biology and Medicine*, 82, 21–28.

[33] Bryan Zi Wei Kuok, Malcolm HS Koh, and Kenneth YT Lim. 2024. An exploratory study integrating deep learning in digital clock drawing test on consumer platforms for enhanced detection of mild cognitive impairment. In *International Conference on Human-Computer Interaction*. Springer, 175–181.

[34] Christoph Laske et al. 2015. Innovative diagnostic tools for early detection of alzheimer's disease. *Alzheimer's & Dementia*, 11, 5, 561–578.

[35] Aoyu Li, Jingwen Li, Jiali Chai, Wei Wu, Suamn Chaudhary, Juanjuan Zhao, and Yan Qiang. 2024. Detection of mild cognitive impairment through hand motor function under digital cognitive test: mixed methods study. *JMIR mHealth and uHealth*, 12, e48777.

[36] Yike Li, Jiajie Guo, and Peikai Yang. 2022. Developing an image-based deep learning framework for automatic scoring of the pentagon drawing test. *Journal of Alzheimer's Disease*, 85, 1, 129–139.

[37] Jin Liu, Xu Tian, Hanhe Lin, Hong-Dong Li, and Yi Pan. 2024. Multi-task learning for alzheimer's disease diagnosis and mini-mental state examination score prediction. *Big Data Mining and Analytics*, 7, 3, 828–842.

[38] Anna MacKay-Brandt, Nadine Schwab, Irene Piryatinksy, Maxine Krengel, Malvina Pietrzykowski, Dave Gansler, Andrea Suazo Rivas, Alyssa DiFalco, and Stan Colcombe. 2023. 93 digitized trail making test in the nki-rockland sample normative lifespan neuroimaging study. *Journal of the International Neuropsychological Society*, 29, s1, 768–769.

[39] Aathira Manoj, Priyanka Borate, Pankaj Jain, Vidya Sanas, and Rupali Pashte. 2016. Offline handwriting recognition system using convolutional network.

[40] Uddalak Mitra and Shafiq Ul Rehman. 2024. Ml-powered handwriting analysis for early detection of alzheimer's disease. *IEEE Access*.

[41] Rose Emily Nina-Estrella. 2019. Mild cognitive impairment: diagnosis and treatment. *Psychiatry and Neuroscience Update: From Translational Research to a Humanistic Approach-Volume III*, 323–331.

[42] Olusanya Olamide Omolara, O. Oyedepo, Elegbede Adedayo Wasiat, Adeola Adetola Olufunmilayo, and Ojo Olufemi Samue. 2023. An offline handwriting age range prediction system using an optimized deep learning technique. *International Journal of Emerging Technology and Advanced Engineering*. DOI: `10.46338/ijetae0423_12`.

[43] Ingyu Park and Unjoo Lee. 2021. Automatic, qualitative scoring of the clock drawing test (cdt) based on u-net, cnn and mobile sensor data. *Sensors*, 21, 15, 5239.

[44] Jin-Hyuck Park. 2024. Clock drawing test with convolutional neural networks to discriminate mild cognitive impairment. *The European Journal of Psychiatry*, 38, 3, 100256.

[45] Alexander Prange, Michael Barz, Anika Heimann-Steinert, and Daniel Sonntag. 2021. Explainable automatic evaluation of the trail making test for dementia screening. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–9.

[46] Hengnian Qi, Xiaorong Zhu, Yinxia Ren, Xiaoya Zhang, Qizhe Tang, Chu Zhang, Qing Lang, and Lina Wang. 2024. A study of assisted screening for alzheimer's disease based on handwriting and gait analysis. *Journal of Alzheimer's Disease*, 101, 1, 75–89.

[47] Nadia Alejandra Rivero-Segura, AA Guerrero-Cruz, and OS Barrera-Vázquez. 2020. Age-related neurodegenerative diseases: an update. *Clinical Genetics and Genomics of Aging*, 27–41.

[48] William Souillard-Mandar, Randall Davis, Cynthia Rudin, Rhoda Au, David J Libon, Rodney Swenson, Catherine C Price, Melissa Lamar, and Dana L Penney. 2016. Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. *Machine learning*, 102, 393–441.

[49] Fadi Thabtah, Swan Ong, and David Peebles. 2022. Detection of dementia progression from functional activities data using machine learning techniques. *Intelligent Decision Technologies*, 16, 3, 615–630.

[50] M Vimaladevi, R Thangamani, S Priyadharshini, B Varshini, and A Tarun. 2024. Prediction of alzheimer's disease by analyzing handwriting dynamics using machine learning algorithms. In *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 1298–1304.

[51] Yu-Jie Xiong, Yu-Fan Dai, and Dan Meng. 2023. Deep frame-point sequence consistent network for handwriting trajectory recovery. In *2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2151–2158.

[52] Nan-Ying Yu and Shao-Hsia Chang. 2019. Characterization of the fine motor problems in patients with cognitive dysfunction–a computerized handwriting analysis. *Human movement science*, 65, 71–79.