



NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

**SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS**

**INTERDISCIPLINARY MSc PROGRAM
"DATA SCIENCE AND INFORMATION TECHNOLOGIES"
SPECIALIZATION
"BIOINFORMATICS – BIOMEDICAL DATA SCIENCE"**

MSc THESIS

**Interpretable Machine Learning for the classification of
Mild Cognitive Impairment patients using actigraphy
data**

Marios A. Gavrielatos

Supervisor:

Elias S. Manolakos, PhD, Professor, Department of
Informatics and Telecommunication, National and Kapodistrian
University of Athens

ATHENS

JUNE 2024



ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

**ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΑΣ"
ΕΙΔΙΚΕΥΣΗ
"ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ – ΕΠΙΣΤΗΜΗ ΒΙΟΙΑΤΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ"**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ ΜΕΤΑΠΤΥΧΙΑΚΟΥ

**Ερμηνεύσιμη μηχανική μάθηση για την ταξινόμηση
ασθενών με ήπια γνωσιακή διαταραχή
χρησιμοποιώντας δεδομένα ακτιγραφίας**

Μάριος Α. Γαβριελάτος

Επιβλέπων:

**Ηλίας Σ. Μανωλάκος, Καθηγητής, Τμήμα Πληροφορικής και
Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο
Αθηνών**

ΑΘΗΝΑ

ΙΟΥΝΙΟΣ 2024

MSc THESIS

Interpretable Machine Learning for the classification of Mild Cognitive Impairment patients using actigraphy data

Marios A. Gavrielatos
SRN: 7115152100023

Supervisor: **Elias S. Manolakos, Professor**, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens

EXAMINATION COMMITTEE:

Elias S. Manolakos	Professor, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens
Nikolaos Skarmas	Professor of Neurology, Department of Medicine, National and Kapodistrian University of Athens
Stavros Perantonis	Research Director, National Center for Scientific Research "Demokritos"

ATHENS

JUNE 2024

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ ΜΕΤΑΠΤΥΧΙΑΚΟΥ

Ερμηνεύσιμη μηχανική μάθηση για την ταξινόμηση ασθενών με ήπια γνωσιακή διαταραχή χρησιμοποιώντας δεδομένα ακτιγραφίας

Μάριος Α. Γαβριελάτος
A.M.: 7115152100023

ΕΠΙΒΛΕΠΩΝ: **Ηλίας Σ. Μανωλάκος, Καθηγητής,** Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Ηλίας Σ. Μανωλάκος	Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Νικόλαος Σκαρμέας	Καθηγητής Νευρολογίας, Ιατρική Σχολή, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Σταύρος Περαντώνης	Διευθυντής Ερευνών, ΕΚΕΦΕ, Δημόκριτος

ΑΘΗΝΑ

ΙΟΥΝΙΟΣ 2024

ABSTRACT

Early and accurate diagnosis of Mild Cognitive Impairment (MCI) is crucial as it is often a precursor stage to Alzheimer's disease, allowing for timely intervention and treatment that can potentially slow the progression of cognitive decline. A non-invasive diagnostic approach would be highly beneficial, as it would be more comfortable and accessible for patients, increasing the likelihood of early detection and enabling prompt action to preserve cognitive function and quality of life. Developing easy and non-invasive digital medicine methods for diagnosing MCI could lead to earlier intervention and better management of this condition, which is a major step towards preventing or delaying the onset of Alzheimer's disease and its devastating effects on individuals and their families.

Actigraphy, which involves the continuous monitoring of physical activity and rest-activity cycles using wearable devices, can offer valuable insights into the presence of MCI. This graduate thesis project was conducted in collaboration with the Aeginition Longitudinal Biomarker Investigation of Neurodegeneration (ALBION) study at the University of Athens, Greece. 7-day actigraphy time series data from individuals with normal cognitive function and those with MCI were collected via an actigraph device. We extracted parametric features, including Mesor (average activity), Amplitude (highest magnitude of activity), and Acrophase (the timing of the largest peak), via a multi-cosinor analysis. We then developed a machine-learning pipeline that uses exclusively these parametric features of recent actigraphy to classify MCI vs. Normal samples, trying to predict the "Diagnosis" endpoint, a binary label assigned by expert clinicians after a patient's biannual evaluation. Our workflow performs a wide range of signal preprocessing steps, examining a variety of machine learning setups, and thus explores the full range of the parametric feature set's class separation capabilities. The selected models are analyzed via a thorough SHAP (SHapley Additive exPlanations) visualization pipeline to distinguish the most informative features and their effect on the classification results, providing valuable insights to clinicians.

We have established that for the ALBION dataset in its current state, lower Mesor, Acrophase, and Amplitude feature values are associated with the MCI class. Specifically, lower Mesor, Acrophase, and Amplitude values have a strong correlation with the MCI samples, while higher Acrophase is associated with the Normal class. Our results are in general agreement with the available recent literature. Moreover, we identified a collection of machine learning models reaching expected specificity 0.88 at recall (sensitivity) level of 0.5 on unseen data. Notably, the developed computational pipeline allows us to fully explore the very large space of predictive model solutions in terms of their sensitivity vs. specificity tradeoff. As ALVION is an ongoing longitudinal study, we expect the generalization abilities of interpretable machine learning models to get better as the dataset size increases and the class balance improves over the years.

SUBJECT AREA: Interpretable Machine Learning, time series data analysis, digital health

KEYWORDS: Alzheimer's disease, Minor Cognitive Impairment, MCI, actigraphy, time series, machine learning, classification, data science, cosinor analysis, Shapley Values.

ΠΕΡΙΛΗΨΗ

Η έγκαιρη και ακριβής διάγνωση της Ήπιας Γνωσιακής Διαταραχής (ΗΓΔ) αποτελεί σημαντικό πρόβλημα, καθώς η ΗΓΔ συχνά είναι πρόδρομο στάδιο της νόσου του Αλτσχάιμερ. Μια μη επεμβατική διαγνωστική προσέγγιση ευρέως προσβάσιμη στο πληθυσμό θα ήταν εξαιρετικά σημαντική, αυξάνοντας την πιθανότητα έγκαιρης ανίχνευσης και επιτρέποντας την άμεση αντίδραση για τη διατήρηση της γνωσιακής λειτουργίας και της ποιότητας ζωής των ατόμων με ΗΓΔ. Η ανάπτυξη μη επεμβατικών μεθόδων ψηφιακής ιατρικής χαμηλού κόστους για την έγκαιρη διάγνωση της ΗΓΔ μπορεί να οδηγήσει σε καλύτερη διαχείριση αυτής της πάθησης και αποτελεί σημαντικό βήμα προς την πρόληψη ή την καθυστέρηση της εμφάνισης της νόσου του Αλτσχάιμερ και των καταστροφικών επιπτώσεων της στους πάσχοντες και τις οικογένειές τους.

Η ακτιγραφία, η οποία περιλαμβάνει τη συνεχή παρακολούθηση των κύκλων σωματικής δραστηριότητας και ανάπαυσης με χρήση φορητών συσκευών, μπορεί να προσφέρει πολύτιμες πληροφορίες για την παρουσία της ΗΓΔ σε υποψήφιους ασθενείς. Η παρούσα διπλωματική εργασία πραγματοποιήθηκε σε συνεργασία με την ομάδα της μελέτης ALBION, του Αιγινήτειου Νοσοκομείου του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών. Τα δεδομένα ακτιγραφίας (χρονοσειρές επτά (7) ημερών) από άτομα με φυσιολογική γνωσιακή λειτουργία και άτομα με ΗΓΔ συλλέχθηκαν μέσω συσκευής ακτιγράφου πριν την επίσκεψη στο Νοσοκομείο. Μέσω της ανάλυση συνημίτονου, υπολογίσαμε παραμετρικά χαρακτηριστικά του σήματος, όπως η μέση δραστηριότητα, (Mesor), το Εύρος (μεγαλύτερο μέγεθος του σήματος, Amplitude) και η Ακρόφαση (ο χρόνος της μεγαλύτερης κορυφής του σήματος, Acrophase). Στη συνέχεια, αναπτύξαμε μια ροή εργασίας μηχανικής μάθησης που χρησιμοποιεί αυτά τα χαρακτηριστικά του σήματος για την ταξινόμηση δειγμάτων ΗΓΔ έναντι φυσιολογικών, προσπαθώντας να προβλέψουμε την ανεξάρτητη μεταβλητή (endpoint) «Διάγνωση», μια δυαδική ταξινόμηση των δειγμάτων που πραγματοποιείται από ειδικούς ιατρούς μετά από αξιολόγηση του κάθε ασθενούς στη κλινική χωρίς τη χρήση των δεδομένων ακτιγραφίας. Η ροή εργασιών που αναπτύξαμε εκτελεί ένα ευρύ φάσμα βημάτων προεπεξεργασίας του σήματος δραστηριότητας, εξετάζοντας μια ποικιλία ρυθμίσεων, και με αυτόν τον τρόπο διερευνάται το πλήρες εύρος των δυνατοτήτων διαχωρισμού των δύο κλάσεων μέσω των παραμετρικών χαρακτηριστικών. Τα επιλεγμένα καλύτερα μοντέλα πρόβλεψης ερμηνεύονται μέσω SHAP (SHapley Additive Explanations) ανάλυσης για να εντοπιστούν τα πιο σημαντικά χαρακτηριστικά και η επίδρασή τους στα αποτελέσματα ταξινόμησης, παρέχοντας πολύτιμες πληροφορίες στους κλινικούς γιατρούς για κάθε δείγμα.

Συμπεράναμε ότι για το σύνολο δειγμάτων της μελέτης ALBION, στην τρέχουσα κατάστασή του, οι χαμηλότερες τιμές Μέσορ, Ακροφάσης και Εύρους συσχετίζονται με την κλάση ΗΓΔ. Συγκεκριμένα, οι χαμηλότερες τιμές Μέσορ, Ακροφάσης και Εύρους έχουν ισχυρή συσχέτιση με τα δείγματα ΗΓΔ, ενώ υψηλότερη Ακροφάση σχετίζεται με την κλάση των φυσιολογικών δειγμάτων. Τα αποτελέσματά μας συμφωνούν με τη διαθέσιμη πρόσφατη βιβλιογραφία. Επιπλέον, παρουσιάζουμε μια συλλογή μοντέλων μηχανικής μάθησης που φτάνουν τιμές για την αναμενόμενη ειδικότητα (specificity) 0,88 και ευαισθησία (sensitivity) 0,5 σε νέα δεδομένα για το μοντέλο. Συγκεκριμένα, η ροή εργασίας αυτή, μας επιτρέπει να εξερευνήσουμε πλήρως τον μεγάλο χώρο των πιθανών λύσεων που προσφέρουν τα ανεπτυγμένα μοντέλα μηχανικής μάθησης όσον αφορά την ευαισθησία και την ειδικότητα των προβλέψεών τους. Καθώς το ALBION είναι μια συνεχιζόμενη μελέτη, αναμένουμε ότι οι ικανότητες γενίκευσης των ερμηνεύσιμων μοντέλων μηχανικής εκμάθησης που μπορούν να αναπτυχθούν θα

βελτιώνονται καθώς αυξάνεται το πλήθος των δειγμάτων τη μελέτης και η ισορροπία των κλάσεων σε αυτά.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Ερμηνεύσιμη μηχανική μάθηση, ανάλυση δεδομένων χρονοσειρών, ψηφιακή υγεία

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Νόσος του Alzheimer, Ήπια Γνωστική Διαταραχή, ΗΓΔ , ακτιγραφία, χρονοσειρές, μηχανική μάθηση, ταξινόμηση, επιστήμη δεδομένων, ανάλυση συνημίτονου, τιμές Shapley.

ACKNOWLEDGMENTS

The current MSc thesis was conducted in the Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens (NKUA) under the supervision of Professor Elias S. Manolakos. I would like to express my sincere gratitude to Professor Manolakos, for his invaluable guidance, and support throughout our cooperation. He was the one who introduced me to core and advanced Machine Learning concepts which I was able to study in depth throughout this project.

I am deeply grateful to the Aiginition Longitudinal Biomarker Investigation of Neurodegeneration (ALBION) team led by Prof. Skarmas for their cooperation and for providing the dataset for this research. Their expertise and insights have been critical in ensuring the practical applicability and relevance of this work. I would also like to extend my appreciation to the members of the examination committee, Prof. Nikolaos Skarmas of the Department of Neurology at NKUA, and Dr. Stavros Perantonis, President of the Scientific Council, and lead Researcher at the "Demokritos" National Research Center for their guidance and helpful comments. Furthermore, I would like to thank the administration of the HYPATIA cloud of the Greek node of ELIXIR for providing valuable computational resources, which were crucial for the realization of this project.

Finally, I would like to express my heartfelt gratitude to my family for their unconditional love, encouragement, and unwavering support throughout my academic journey. Their belief in me has been a constant source of motivation, encouraging me to pursue my dreams and overcome challenges along the way.

CONTENTS

PREFACE	13
1. INTRODUCTION	14
1.1 Alzheimer's disease demographics	14
1.2 Alzheimer's disease history	14
1.3 Alzheimer's distinctions in early and late-onset	15
1.4 Alzheimer's neurobiological base.....	15
1.5 Mild Cognitive Impairment and Alzheimer's disease progression	16
1.6 The ALBION study	16
1.7 Related work	16
1.8 Goals and importance of this study.....	18
1.9 Thesis organization	19
2. METHODS	20
2.1 Dataset description	20
2.1.1 Data collection.....	20
2.1.2 Endpoints	22
2.2 Dataset analysis and preprocessing	23
2.2.1 Datasets generation for analysis.....	26
2.2.2 Feature Extraction.....	29
2.2.3 Feature Selection and Processing	31
2.3 Machine Learning pipeline.....	33
2.3.1 Concepts	33
2.4 Visualization and further exploration.....	39
2.4.1 Nested Cross-Validation results exploration.....	39
2.4.2 Interpretation of model predictions.....	40
3. RESULTS & DISCUSSION.....	44
3.1 Summary of results	44
3.2 Interpretation of model predictions	49

3.3 Feature Kolmogorov-Smirnov tests.....53

3.4 Spatial allocation of samples 54

4. CONCLUSIONS 55

ABBREVIATIONS – ACRONYMS 57

ANNEX I.....59

REFERENCES.....63

LIST OF FIGURES

Figure 1: Projected number of individuals aged 65 and older with Alzheimer's Dementia from 2020 to 2060 in the US.	14
Figure 2: Participants in the ALBION study perform five annual assessments (A-E).....	22
Figure 3: Histogram of zero values percentage per actigraphy.....	24
Figure 4: Venn diagram showing the relationship between the Diagnosis (reduced dataset) and CSF endpoints.	25
Figure 5: Actigraphy length histogram.....	25
Figure 6: 5-day actigraphies of the MCI class (red) and the Normal class samples (blue) overlaid.....	26
Figure 7: Average 1-day actigraphies of the MCI class (red) and the Normal Controls class (blue).	27
Figure 8: Euclidean distance of 1-day time series (MCI vs. Normal) for different 8-hour time intervals (frames) of the day.	28
Figure 9: Histogram of 1-day actigraphy values (A) before normalization and (B) after square root normalization.....	29
Figure 10: Graphical representation of Mesor, Amplitude, and Acrophase.	30
Figure 11: The nested cross-validation (nCV) pipeline comprises an outer and inner loop, each implementing a stratified 5-fold CV.....	36
Figure 12: Example of hyperparameter set selection (model selection) using the one-SEM rule.	37
Figure 13: Final model building and SHAP values generation.....	41
Figure 14: Results summary Recall vs Specificity plots.....	44
Figure 15: Each point in the plot defines an experiment, and the color represents the classifier family.	45
Figure 16: Each point in the plot defines a set of experiments, and the color represents the actigraphy used.....	46
Figure 17: SHAP analysis of 1-day STL decomposed actigraphy dataset.....	50
Figure 18: SHAP-based analysis of 1-day actigraphy dataset.....	51
Figure 19: SHAP analysis of Frame 2 actigraphy dataset.....	52
Figure 20: Dimensionality reduction using UMAP of the selected features for datasets 1 (A), 2 (B), and 4 (C) from Table 5.....	54

Figure A1: Nested cross-validation (nCV) pipeline with MRMR feature selection in the outer loop.	60
Figure A2: Nested cross-validation (nCV) pipeline with SFM feature selection in the inner loop.	62

LIST OF TABLES

Table 1: Positive (MCI) and Negative (Normal controls) sample distribution based on different endpoints.....	24
Table 2: Datasets used in the study, their length, and the number of extracted features....	31
Table 3: Applied rules for the one-SEM hyperparameter selection approach.	38
Table 4: Summary of model selection approaches.	39
Table 5: Best performing experiments per actigraphy and recall levels – nCV median metrics.	47
Table 6: Final models estimated generalization performance.	48
Table 7: Kolmogorov-Smirnov two-sample test results.	53

PREFACE

This project was conducted using a data set provided by ALBION, an ongoing longitudinal study at the Aiginition Hospital, University of Athens, Greece, supervised by Prof Skarmas. The study aims to decipher the preclinical and prodromal stages of Alzheimer's disease (AD). It took place in Athens, Greece, as a thesis project to fulfill the requirements of the international Master's program "Data Science and Information Technologies (DSIT)," Specialization: "Bioinformatics and Biomedical Data Science" administered at the Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens, under the supervision of Professor Elias S. Manolakos.

The primary objective of this study was the development of interpretable machine learning models adept at distinguishing between normal control and Mild Cognitive Impairment individuals using exclusively features extracted from 7-day actigraphy time series data of the patients while also ensuring the interpretability of the model predictions for clinical relevance. As ALBION is an ongoing effort, 121 actigraphy samples were available for the data analysis at the time of this project.

1. INTRODUCTION

1.1 Alzheimer's disease demographics

Dementia is an umbrella term to describe various neurodegenerative disorders with symptoms that are related to memory loss, language problems, difficulties in problem-solving scenarios, and other thinking-related skills. These symptoms are related to specific changes in the brain. Dementia has a broad spectrum of severity from the mildest to the most severe stage. Alzheimer's disease (AD) is a type of dementia caused by damage to nerve cells, known as neurons. AD is the most common type of dementia, accounting for 60% to 80% of cases. It is also one of the most expensive medical conditions, as it costs \$321 billion and is expected to reach \$1 trillion by 2050 in the US [1] while worldwide, the annual cost is over \$1.3 trillion and is forecasted to rise to \$2.8 trillion by 2030 [1]. The disease affects an estimated 6.7 million Americans aged 65 and older (1 in 9 people) in 2023, while 73% are age 75 or older. As the population ages, these numbers are likely to grow continuously [2], and by 2060, 13.9 million Americans are estimated to have Alzheimer's disease [3], [4] (Figure 1). The same is true for the European population and also for the Greek population, as in 2018, 1.99% of the population suffered from dementia, and by 2050 this number is projected to reach 3.95% [5].

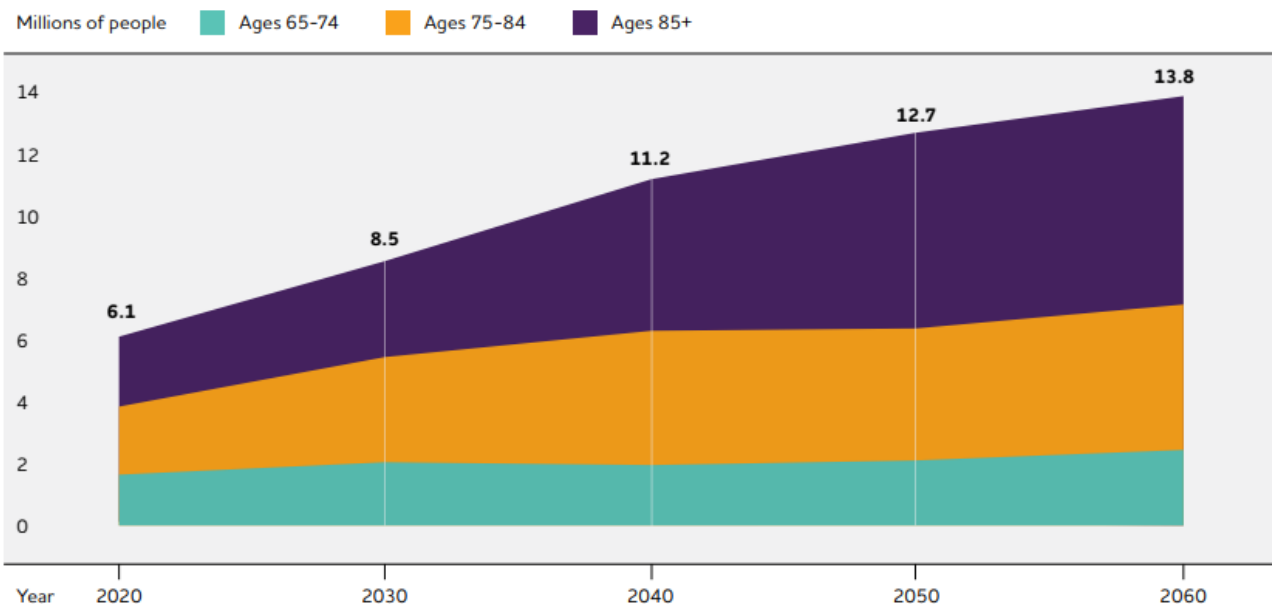


Figure 1: Projected number of individuals aged 65 and older with Alzheimer's Dementia from 2020 to 2060 in the US. Figure adapted from the Alzheimer's Association. [2]

1.2 Alzheimer's disease history

In modern medicine, the term dementia was accepted as a medical term in 1797 by Philippe Pinel [6]. Jean Etienne Dominique Esquirol described it as the "disabilities are shown in discernment, intellectual ability and will due to brain diseases, and is to lose joyfulness enjoyed and is that the rich become poor" [7]. In 1906, Dr. Alois Alzheimer conducted a biopsy of the brain to examine the relationship between medical records and the clinical and neurobiological symptoms of a 51-year-old patient. He noticed changes in the brain tissue that were not previously observed in people at this age, like abnormal clumps, now known as amyloid plaques, and tangled bundles of fibers, now called tau tangles, while regions that are associated with memory, language, judgment, and thinking in the brain were impaired

and the cerebral cortex was thinned [8]. During the early days of AD research, the disease was viewed as a clinicopathological entity. If the patient exhibited the clinical symptoms of amnesic dementia and other causes were ruled out, then the patient was assumed to be suffering from AD [9]. Advancements in the area and the availability of biomarkers such as cerebrospinal fluid (CSF) and PET markers have opened the road to understanding the neurobiological base of the disease [10].

1.3 Alzheimer's distinctions in early and late-onset

AD cases could be either inherited or sporadic. Only a very small percentage of AD cases (less than 10%) is inherited via an autosomal dominant way, and they usually showcase a very early onset and fast progression of the disease. The Amyloid-beta Precursor Protein (APP) and the presenilins, Presenilin 1 (*PSEN1*) and Presenilin 2 (*PSEN2*), are the three primary genes that are responsible for early onset familial AD. If an individual carries a genetic variant of these three genes, s/he is likely to develop AD early in life as abnormal proteins are produced. These abnormal proteins, which are amyloid fragments and are named amyloid beta ($A\beta$) peptides, are produced as the result of the breakdown of APP. These peptides cluster together to create plaques. These cause the deterioration of neurons by disrupting normal cell communication, leading to neural dysfunction and, eventually, cell death. The extracellular $A\beta$ plaques and tau tangle promote neuroinflammation in which immune cells of the brain tissue known as microglia release pro-inflammatory molecules, forming a chronic inflammatory condition that further degenerates the brain neurons.

In contrast to early-onset AD, in late-onset AD, which develops at a later age, environmental and genetic causes act as risk factors contributing to the disease's development. The most important and impactful gene that influences the risk of Alzheimer's is the apolipoprotein E (*APOE*) gene and its variants. The seriousness of the risk is isoform-dependent (apoE4 > apoE3 > apoE2) [11]. ApoE co-deposited with $A\beta$ in amyloid plaques, while there is also a lot of evidence that supports the claim that apoE plays an essential role in $A\beta$ fibrilization and amyloid deposition [12]. Regarding tau, apoE4 carriers show an increased tau deposition regardless of the presence of $A\beta$ plaques [13], while it has been shown to promote tau phosphorylation and cell death [14].

1.4 Alzheimer's neurobiological base

Biomarkers are molecules essential for diagnosing and monitoring disease and, therefore, are directly associated with treating a disease as the doctors can act precisely and determine the individual disease risk. By detecting the presence and measuring the concentration and activity of these molecules, researchers can identify the likelihood of a clinical event. The cerebrospinal fluid (CSF) is an important source of AD biomarkers as it is directly associated with the brain's extracellular space. As already described, the neuropathological hallmarks of Alzheimer's are the amyloid plaques and the neurofibrillary tangles [15]. According to the "amyloid hypothesis", the accumulation of the amyloid plaques with the concatenation of $A\beta$ peptides is the primary cause of AD pathology. According to the "tau hypothesis," the loss of function of the microtubule-stabilizing tau proteins via a mechanism of hyper-phosphorylation of the tau proteins results in the degradation of the cytoskeleton and, therefore, cell death [16]. Based on the aforementioned hypothesis, amyloid pathology can be diagnosed by an increase in CSF β -amyloid ($A\beta_{42}$) concentration or by an increase in amyloid-specific radiotracer uptake, as measured by PET CT, while tau pathology can be determined by the rise in CSF phosphorylated tau species (pTau) but an increase in CSF tau proteins (tTau) is not an indication of neurodegenerations [17].

1.5 Mild Cognitive Impairment and Alzheimer's disease progression

The progression of AD follows a distinct pattern, with symptoms worsening over time. In the early stages of the disease, patients exhibit mild memory loss. At the same time, skills such as reasoning and conduct are also affected to a certain degree, as abstract concepts such as numbers may appear challenging. This minor neurocognitive disorder is referred to as Mild Cognitive Impairment (MCI), which describes a noticeable decline in cognitive abilities greater than what would be expected for normal aging but does not meet the criteria for dementia. Individuals with MCI may experience difficulties with memory, thinking, language, or other cognitive functions. However, these impairments are not severe enough to significantly interfere with daily activities or independence. MCI is considered a transitional stage between normal cognitive aging and more serious cognitive decline. This cognitive decline may extend to other areas of the patient's life and manifest as difficulties in planning, decreased problem-solving skills, and other challenges in everyday life. During these early stages of the disease, the patient may be aware of this cognitive decline. These individuals have a higher risk of developing dementia and AD as they tend to progress to probable dementia at a rate of approximately 10% to 15% per year. As the disease progresses, memory loss will become more severe, and long-term memories will be affected; individuals will become confused, disoriented, and struggle to complete routine chores. Behavioral alterations, such as agitation, despair, and social disengagement, are also distinct symptoms of Dementia and AD [18].

1.6 The ALBION study

This thesis project was conducted in collaboration with the Aiginition Longitudinal Biomarker Investigation Of Neurodegenerations (ALBION) study. ALBION is a longitudinal ongoing study initiated in 2018 in the Cognitive Disorders Clinic of Aiginition Hospital of the National and Kapodistrian University of Athens. The Aiginition research team, led by Prof. Scarmeas, investigates the preclinical and prodromal stages of Alzheimer's disease to find insights about potential markers for early detection and prediction of dementia and mild cognitive impairment, the earlier stages of AD. This study's participants are individuals over 40 who believe to have experienced symptoms of cognitive decline, but they are cognitively normal or have mild cognitive deficits. Each participant undergoes extensive tests to collect information regarding demographic, medical, clinical, nutritional, and neuropsychological determinants, and information from lifestyle activities. Apart from these, biological samples (blood, stools, CSF) are collected from each individual. Finally, data are collected via neuroimaging techniques, and each participant is assigned a wearable actigraphy device. The data used in this project come from the wrist actigraphy devices as well as further information regarding the diagnosis, whether an individual is of normal or impaired cognition, assigned by a team of medical doctors. It is important to note that the ALBION study is ongoing, and the dataset is under development, with new samples added every four months.

1.7 Related work

Several studies have utilized actigraphy data to investigate the association between activity, sleep patterns, circadian rhythms, and cognitive impairment in older adults. *Roger-Soeder et al.* [19] found evidence that lower average activity levels (Mesor), lower magnitude of activity (Amplitude), and earlier activity peaks (Acrophase) are associated with a higher possibility of cognitive decline. *Tranah et al.* [20] examined circadian activity patterns in older women and their association with MCI and dementia. Using logistic regression and cosine analysis, they inferred odds ratios and 95% confidence intervals. They found that reduced activity levels and later peak activity times (Acrophase) were linked to a higher risk of

cognitive decline. Similarly, *Li et al.* [21] also associated a lower Amplitude with MCI and AD dementia development, as well as higher intraday variability and lower intraday stability, without finding associations between cognitive decline and Acrophase. *Milman et al.* [22] demonstrated that individuals with cognitive impairments experienced less efficient sleep and increased awakenings compared to those without impairment, with a trend toward decreased relative Amplitude (RA) observed in the cognitively impaired group.

Most studies limited their investigation to sleep actigraphy to infer correlations between sleep-wake patterns and MCI. *Diem et al.* [23] explored sleep-wake patterns measured by actigraphy to assess the risk of incident MCI and dementia. Their statistical analysis revealed that lower average sleep efficiency, longer sleep latency, and increased variability in total sleep time correlated with a heightened risk of cognitive impairment. *Basta et al.* [24] investigated the association between objective sleep patterns and cognitive performance in elderly individuals diagnosed with AD, MCI, and cognitively unimpaired controls. The study revealed that while AD patients exhibited longer total sleep time compared to MCI and normal controls, elderly patients with MCI have similar objective sleep duration to normal controls. It appears that within the MCI group, those who sleep longer have more severe cognitive impairment. *Posner et al.* [25] discovered that earlier sleep/wake times were associated with a higher risk of dementia, but not MCI, and smaller day/night activity differentials were noted in samples with increased dementia risk. *Alfini et al.* [26] discovered that those with MCI displayed greater nighttime activity and reduced morning activity by analyzing sleep patterns and circadian rest/activity rhythms (RARs).

Cai et al. [27] conducted a systematic review and meta-analysis on sleep parameters obtained through polysomnography or actigraphy in patients with amnesic mild cognitive impairment (aMCI). Their findings suggested that individuals with aMCI experience more severe sleep disturbances compared to cognitively normal elderly individuals. *Winer et al.* [28] discovered associations between tau burden and wristwatch actigraphy-measured sleep quality, as well as β -amyloid deposition and self-reported sleep quality, underscoring the complex interplay between sleep disturbances and cognitive function. *Lauriola et al.* [29] compared individuals experiencing subjective cognitive decline (SCD) with cognitively normal individuals. Their study suggested that subjects with SCD exhibited reduced sleep efficiency and increased overnight wakefulness, indicative of objective sleep disruption. *Scarlett et al.* [30] identified a correlation between sleep disturbances and cognitive performance, particularly in cognitive response time (CRT) and overall response time.

Apart from cognitive decline in older adults, actigraphy watches have been used to find correlations between AD, other types of dementia, and activity patterns. *Lysen et al.* [31] linked poor sleep, as measured by longer sleep latency, waking after sleep onset, time in bed, lower sleep efficiency, and an earlier "lights out" time, to an increased risk of dementia, particularly AD. Nevertheless, no correlation was discovered between the incidence of dementia and 24-hour activity cycles. *Dowling et al.* [32] investigated the effect of light treatment alone and the incorporation of melatonin into bright-light therapy for the treatment of rest-activity disruption in institutionalized patients with AD. According to the study's findings, light therapy alone did not enhance rest-activity rhythm, wakefulness during the day, or sleep at night, while on the other hand, light treatment plus melatonin increased daytime wake time and activity levels and strengthened the rest-activity rhythm. *Targa et al.* [33] proposed the potential role of the circadian rest-activity pattern in predicting cognitive decline in patients with mild-moderate Alzheimer's disease. *Bringas et al.* [34] successfully implemented a deep learning analysis, utilizing Convolutional Neural Networks to classify

three stages of Alzheimer's disease progression—Early-Stage, Middle-Stage, and Late-Stage—based on activity data from a mobile device. Alphen et al. [35] found that dementia patients who live in institutions or the community spend most of their days inactive and have less intense physical activity levels than healthy individuals. *Anderson et al.* [36] performed an investigation on sleep/wake disturbance in individuals with frontotemporal dementia (FTD). They revealed notable disruptions during the illness, which were different from those observed in patients with AD. Patients with FTD had reduced morning activity and increased nighttime activity, which may indicate a phase delay. All FTD patients had lower sleep efficiency and overall sleep.

Parametric cosine and non-parametric analysis are not the only feature engineering types implemented in the study of MCI and AD dementia. Wavelet-based analysis, which is a type of parametric analysis, has not been applied to human actigraphy data [37]. However, Hayes et al. [38] performed a wavelet-based analysis using an unobtrusive sensor system to monitor walking speed and home activity. They revealed the variation in median walking speed was twice as high in the MCI class, and the inter-daily variability was also elevated.

Activity patterns are not random but comprise small patterns that construct larger curve formations. These patterns are also called fractal temporal patterns and can be extracted by performing fractal analysis. The fractal patterns contain information about the circadian rhythms. Several studies have used fractal patterns to separate cognitively declined samples from normal controls [39]–[41].

Empirical model decomposition (EMD) is a data-adaptive approach, meaning the user does not need to make any assumptions regarding the nature of the data, as in the parametric cosinor and wavelet analysis, as the model considers the irregularity of the activity curves. The activity signal will get split into parts called intrinsic mode functions (IMFs) until it cannot be decomposed any further [42], [43].

As we will see in the following chapters, our findings using the ALVION dataset align in general terms with the outcomes in the abovementioned studies regarding the association between parametric features extracted from actigraphy time series and MCI. In addition, the proposed machine learning pipeline entails a comprehensive analysis of actigraphy time series to classify MCI patients and individuals with normal cognitive function based solely on activity observed by a wearable device within a 24-hour cycle. Moreover, we focus on outcomes of our machine learning pipeline that yield interpretable results, thereby providing valuable insights to healthcare professionals engaged in the diagnostic process of MCI down to the individual sample level. We present a collection of classification models capable of analyzing a wide range of activity patterns, offering the basis for developing a tool for clinical decision support, using non-intrusively collected data as input.

1.8 Goals and importance of this study

The primary objective of our study is to develop interpretable machine-learning methods and an associated pipeline that, using the actigraphy time series of individuals alone, could classify them as normal controls (NC) vs. MCI. The present study focuses exclusively on the parametric features extracted using cosinor analysis [44] because they are easily interpretable by medical doctors. We are also using interpretable machine learning methods aiming to provide clinicians with valuable insights about a patient's current status based on the actigraphy profile a few days before s/he is examined at the clinic. Our focus is on developing a pipeline that not only identifies the most influential features but also elucidates the direction and magnitude of their impact on the model predictions. By emphasizing

interpretability in a machine learning framework, we aim to empower healthcare professionals involved in the diagnosis process, providing them with the basis for developing a useful AI tool for clinical decision support.

Our approach seeks to bridge the gap between advanced machine learning techniques and the practical needs of clinicians, ultimately improving the quality of the accuracy and efficiency of MCI diagnosis. Distinguishing the two groups is challenging as the actigraphy patterns of the samples in the positive class (MCI) and the negative class (NC) are very difficult to discern. Despite the challenging classification problem and the limited number of available samples, we managed to identify the key features and how they contribute to distinguishing the two classes, demonstrating the reduced activity levels of the MCI samples compared to the NC cases. We also propose a machine-learning workflow applicable in biomedical studies to analyze and interpret actigraphy datasets.

1.9 Thesis organization

This thesis is organized into four chapters to provide the reader with a comprehensive and easy-to-follow exposition of an interpretable machine learning-based classification approach to diagnose individuals with Mild Cognitive Impairment vs. Normal Control individuals based on actigraphy time series data alone. In the current chapter, we present the background information of our analysis and the key characteristics of the ALBION study, providing the data for this project. It establishes the context by delineating the research problem and articulating the main goals and significance of the study. Moreover, it reviews the state of the art and discusses how this study attempts to extend it. In the next chapter, Methods, we describe the characteristics of our dataset, the research methodology, and the key tools and concepts implemented to develop our computational pipeline and estimate our models' performance, as well as the exploratory analysis that makes our results interpretable with clinical relevance. In Chapter 3, Results and Discussion, we present the main findings of the ALVION dataset analysis and explore the space of the possible solutions produced by our pipeline while presenting key experiments and their interpretation to demonstrate the flexibility and power of the analysis approach. Finally, in chapter 4, we present the concluding remarks, compare our findings with the literature to demonstrate the novelties of this study as well as its limitations, and present ideas for further research.

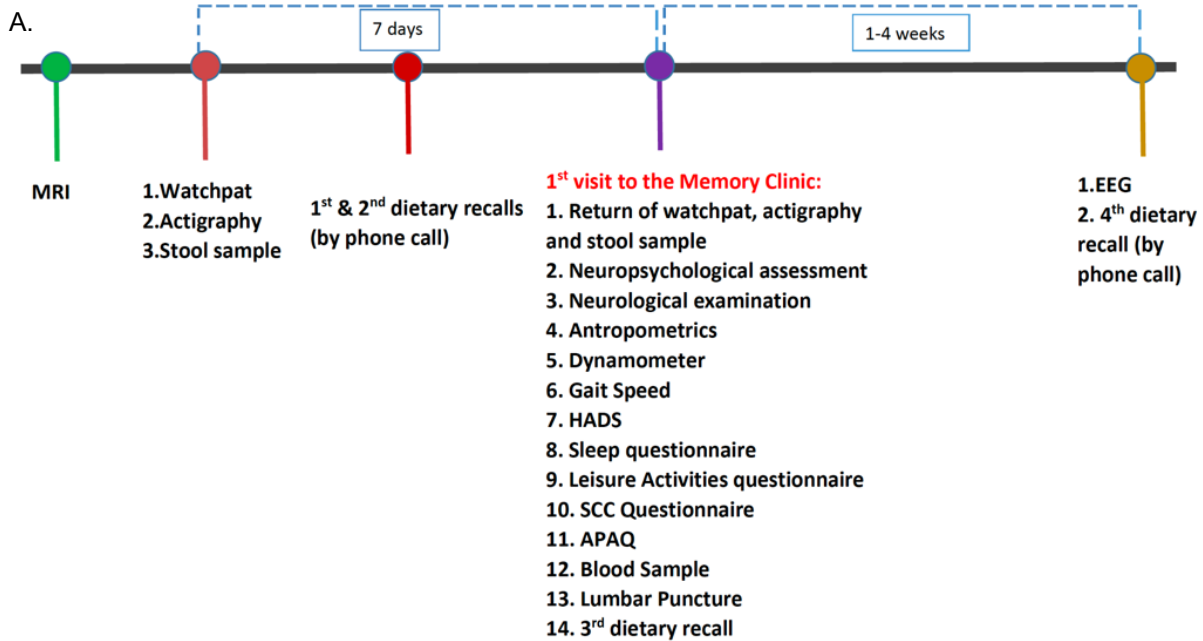
2. METHODS

2.1 Dataset description

2.1.1 Data collection

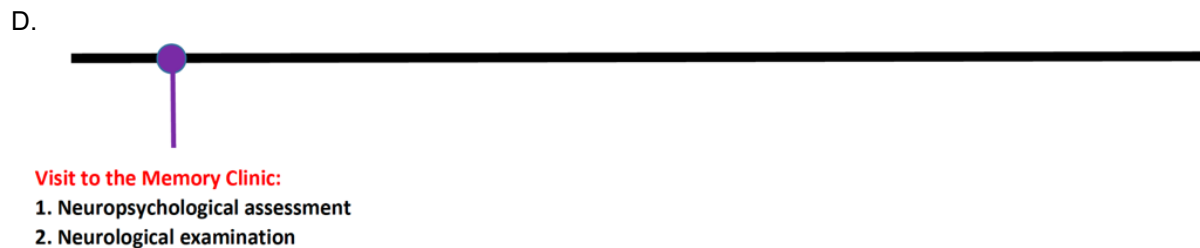
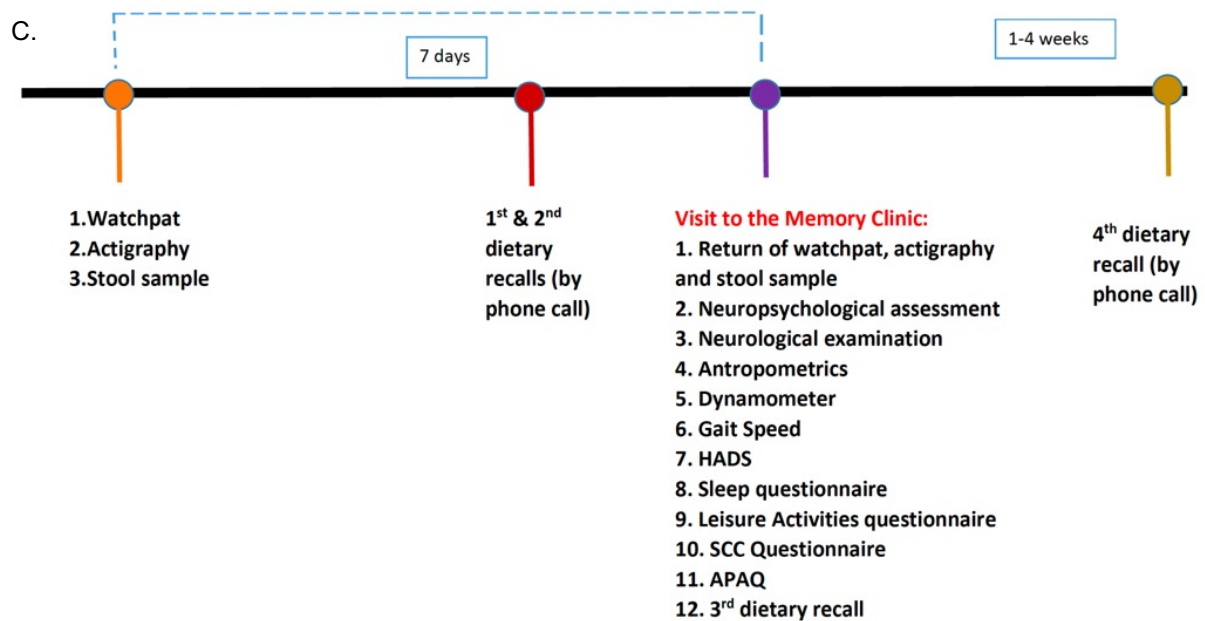
Subjects participating in the ALBION study undergo five assessments, as depicted in Figure 2. During the first and third assessments, subjects are given an actigraphy watch (Actiwatch 2 Phillips-Respironics) [45] and are advised to wear it in the non-dominant hand for seven days and nights, except for bath time. After exporting the data, a table with the activity parameter is produced as a CSV file. Using a piezoelectric sensor, the Actiwatch device records subject activity counts every 15 seconds (0.0667 Hz). Activity counts reflect the peak acceleration over a vertical axis detected over each epoch and are traditionally used in determining sleep and wake intervals. The date and time of the measurements (timestamp), a marker used to identify a particular event, white light detection from the environment, and the sleep/wake status determined by a Sleep Interval Detection Algorithm are additional variables included in the raw data. In the present study, we use only the Activity counts and the date-time measurements (5760 time points per day).

Since ALBION is an ongoing study, new subjects are introduced, and their time series data are incorporated into the dataset every four months. At the time of this thesis, there were 121 recorded actigraphies, of which 111 were first-assessment and only 10 were second-assessment actigraphies.





* These procedures were performed till August 2021. After that, visit 2 will include only neurological and neuropsychological assessment. Watchpat, Actigraphy and stool sample could be administered in V2, if not administered in V1.



Watchpat, Actigraphy and stool sample could be administered in V5, if not administered in previous visits (baseline or repeated measure).

E.

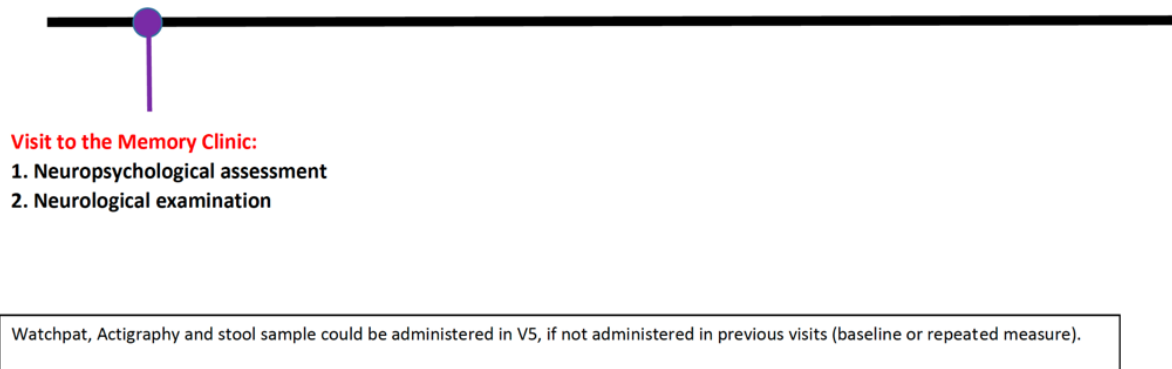


Figure 2: Participants in the ALBION study perform five annual assessments (A-E). Actigraphies are recorded on the 1st and 3rd visits. Participants are followed up longitudinally for at least ten years to be able to observe potential cognitive.

2.1.2 Endpoints

An SVA file with 544 subject features is acquired along with the actigraphy device export file. We can separate these features into the following categories: clinical neurology scores, laboratory and imaging tests, demographic, medical history, and family history. In this thesis, we are mainly concerned with the "Diagnosis" endpoint. After a Diagnostic consensus meeting, all of ALBION's principal investigators and researchers, including neurologists and neuropsychologists, assign a Diagnosis category to all subjects. For that endpoint, subjects are characterized as having "Normal Cognition," "Impairment but not MCI," "MCI" (reduced performance in one or more neuropsychiatric tests but not gravely affecting independence in everyday routine), or "Dementia" (same as "MCI" but with affected independence). It is important to note that the class "impairment but not MCI" is not a valid clinical category, but it is an extra category to define healthy individuals with a normal cognitive decline, and thus we should treat them as healthy.

The DSM-IV-TR criteria [46] are utilized to diagnose "Dementia" and its subcategories in greater depth. The National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer Disease and Related Disorders Association (NINCDS/ADRDA) criteria [47] are used to determine whether an AD diagnosis is likely or feasible. When determining if a subject has vascular dementia, it is necessary to consider several variables, including the history of stroke or any related clinical signs, the time passed since the stroke when dementia first appeared, and the Hachinski Ischemia Scale score [48]. For the diagnosis of "MCI" and its subtypes, the Petersen criteria [49] are utilized. Based on the categories mentioned above, we can categorize the subjects into two classes:

1. Class 1, which includes subjects belonging to the "MCI" and the "Dementia" category (what we call positive samples),
2. Class 0, which includes subjects belonging to the "Normal" and the "Impairment but not MCI" categories (what we call negative samples).

Based on the rules above, the negative and positive groups concerning the Diagnosis endpoint are populated as follows:

1. Class 1: 32 samples (with actigraphy time series)

2. Class 0: 89 samples (with actigraphies)

The number of actigraphy samples that are positive for the CSF biomarkers are the following:

1. Amyloid B: 46 positive samples
2. Tau protein: 14 positive samples
3. PhosphoTau protein: 15 positive samples

Based on the ALBION protocol and the ROCHE brochure criteria, which are robust and clinically validated, cut-off values that are valid worldwide [50], [51] are the following:

- If pTau/Abeta42 ratio >0.024 , the test result is positive (If pTau/Abeta42 ratio ≤ 0.024 , the test result is negative).
- If tTau/Abeta42 ratio >0.28 , the test result is positive (If tTau/Abeta42 ratio ≤ 0.28 , the test result is negative).

The sample is considered positive if one of the aforementioned rules evaluates to “true”. Based on this, the negative and positive groups of our actigraphy samples are populated as follows according to the CSF endpoint:

1. Class 1: 17 samples
2. Class 0: 102 samples

This work focuses only on the "Diagnosis" endpoint as the target (binary) variable. We decided to include only samples for which we also had information about their CSF biomarkers in the analysis. Consequently, we excluded two samples, leaving us with a total of 119 samples (30 positive, 89 negative).

2.2 Dataset analysis and preprocessing

Each subject initiated its actigraphy session at a different time of the day, as observed by aligning the different samples based on the daytime (timestamps). Zero values in our dataset represent a lack of activity, either a sleep or resting state. By plotting the distribution of the zero-value frequency of the samples, we expect to see a close-to-normal curve (Figure 3). However, one sample, “6661”, stands out with an unusually high zero count, with over 95% of its activity points being zeros, meaning a lack of activity for 95% of the actigraphy session. We considered this sample as an outlier and removed it from the analysis.

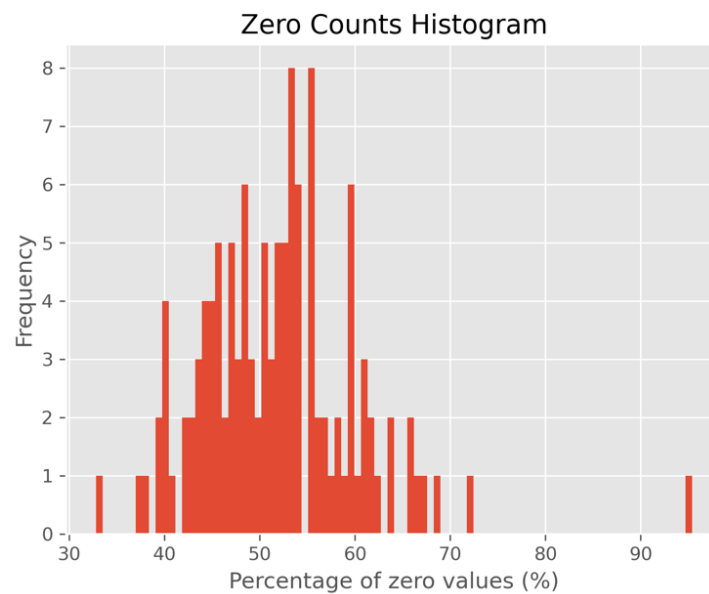


Figure 3: Histogram of zero values percentage per actigraphy. The outlier actigraphy on the far right (sample “6661”) has 95.4% zero values, and thus it is removed from the dataset.

A summary of the considered dataset's categories is provided in Table 1 below (Figure 4).

Table 1: Positive (MCI) and Negative (Normal controls) sample distribution based on different endpoints. Our dataset is imbalanced as the Normal sample size is close to three times or five times the size of the Positive samples for the Diagnosis or the CSF endpoint respectively. It is worth noting that the intersection of those two endpoints reveals that only a few samples are positive in both two categorizations.

	Diagnosis	Amyloid B	Tau	PhosphoTau	CSF	Intersection of Diagnosis and CSF
Positive samples	30	46	14	15	17	8
Negative samples	88	72	104	103	101	79
Total samples	118	118	118	118	118	87

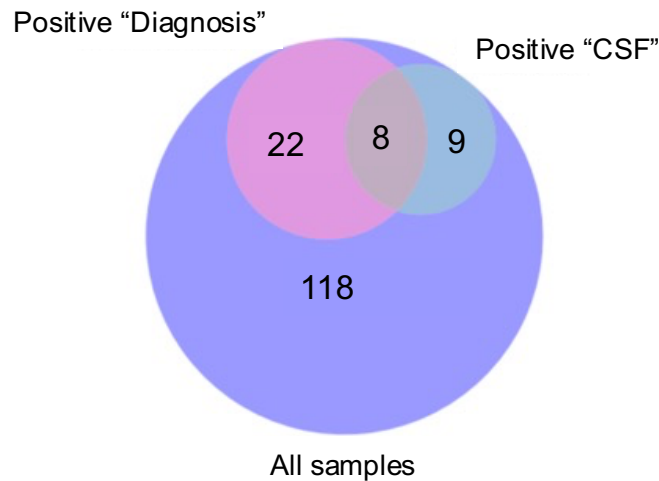


Figure 4: Venn diagram showing the relationship between the Diagnosis (reduced dataset) and CSF endpoints. Positive Diagnosis (30) and positive CSF (17) have only eight (8) common samples.

The actigraphy time series of different samples is not of the same length. Figure 5 shows a histogram of the time series lengths of our dataset. It illustrates that most time series have similar lengths, but some outlier samples exist. The longest actigraphy record, sample "6385", has 43,682 data points (7 days, 14 hours, 1 minute), while the shortest, sample "7075", has only 27,673 data points (4 days, 19 hours, 18 minutes, 15 seconds).

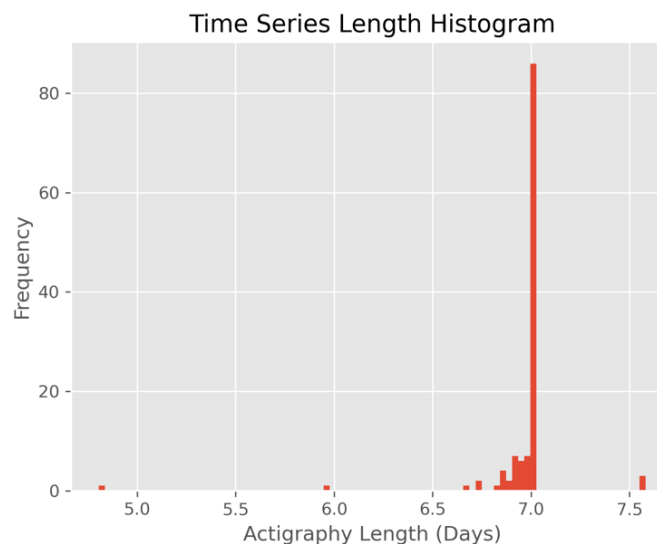


Figure 5: Actigraphy length histogram. Most of the samples have a length of close to 7 days.

To tackle the problem of unequal lengths, we decided to investigate the different approaches discussed below and create additional time series datasets from which we would extract features for the downstream machine learning predictive modeling steps.

2.2.1 Datasets generation for analysis

2.2.1.1 The 5-day actigraphy time series

The first approach uses 5-day actigraphies (28,800 data points, i.e., 5 days x 5,760 time-points per day) (Figure 6). By trimming the time series down to 5 days, we eliminated many front and rear ends missing (NaN) values. After the time-series trimming, two samples still exhibited NaN values at the front and/or rear ends, requiring restoration. To tackle this problem, we used a data imputation method that fills the missing time points using the mean daytime activity values of the same subject and time points from other days of the actigraphy.

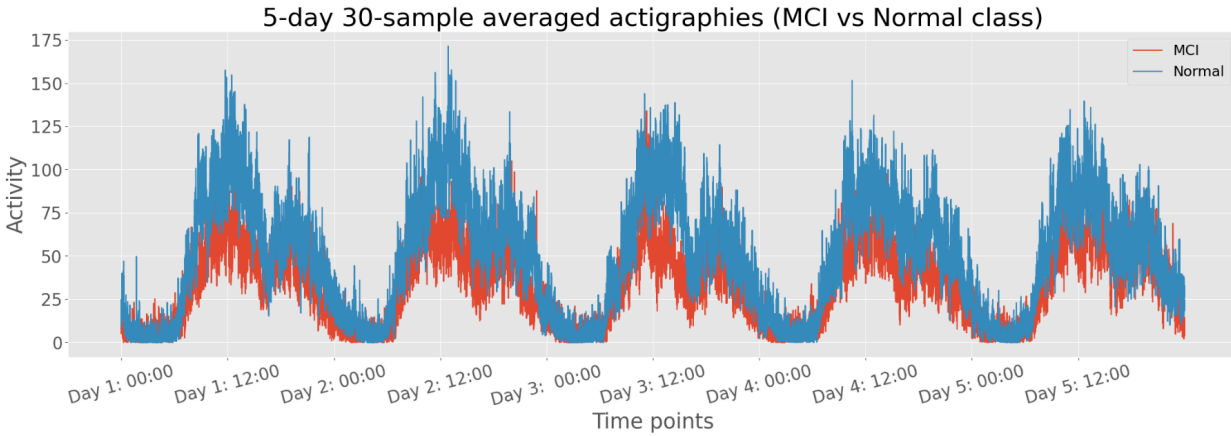


Figure 6: 5-day actigraphies of the MCI class (red) and the Normal class samples (blue) overlaid. The two classes intertwine in some areas of the time series, making their distinction a challenging task.

2.2.1.2 The 1-day averaged actigraphy time series

While exploring the 5-day actigraphies we observed substantial inter-daily variability. To address this issue, we decided to calculate the 1-day average of the time series, encompassing not just the five days used in the previous approach but the entire length 7 days of each time series. By doing so, we avoid imputing missing values and utilize more activity points to estimate the true mean activity value for each time point. This way, we created a 1-day averaged actigraphy for each sample, effectively mitigating the problem of inter-daily variability. This approach allowed us to perform our analysis based on the mean activity time series representing an individual's average activity level at any time point of the day. From now on, we will refer to the 1-day average dataset simply as the 1-day dataset.

Figure 6 depicts the MCI and Normal Controls class samples of 5-day actigraphies. As we observe, the two class samples overlap in a non-consistent manner on different days. This happens because the activity of each subject may vary from day to day due to its inherent inter-daily activity variability. For instance, the activity pattern of Day 1 of an MCI subject may resemble a Normal Control subject closer than another MCI subject. This inconsistency may interfere with the effectiveness of downstream feature extraction methods employed for classification. In contrast, when averaging the actigraphies to produce a 1-day time series, we observed that the mean actigraphies, over the samples of the two classes, exhibit improved separation (Figure 7).

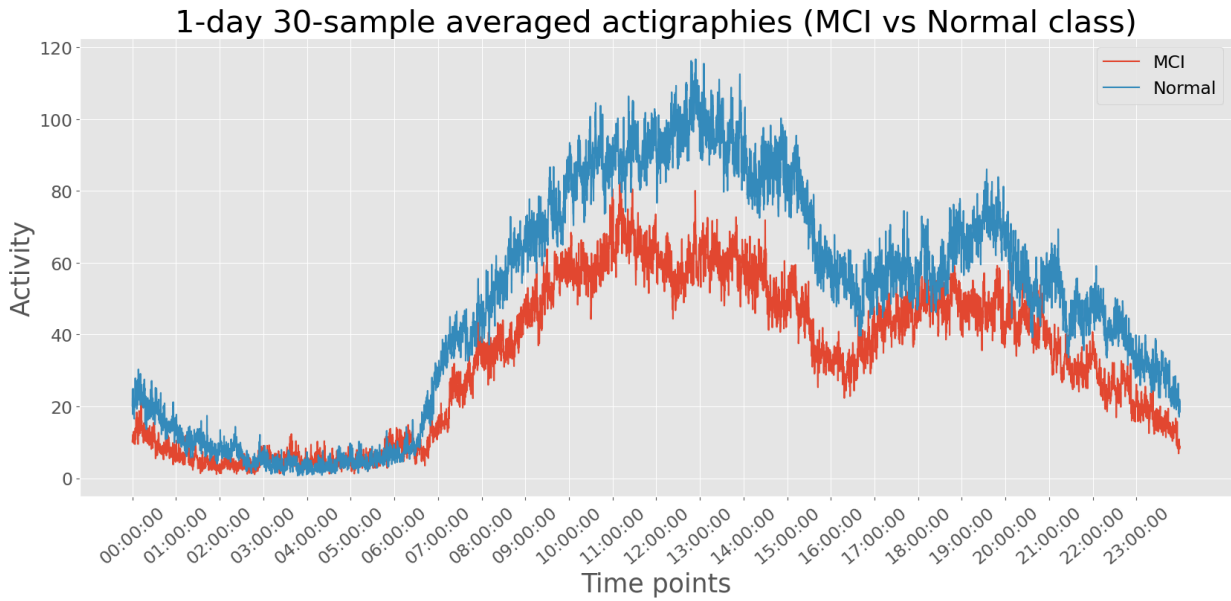


Figure 7: Average 1-day actigraphies of the MCI class (red) and the Normal Controls class (blue).
The 1-day averaged time series of the two classes are better separated than the 5-day original actigraphies, facilitating the classification of the samples of the two classes.

2.2.1.3 Actigraphy time frames

Upon creating the 1-day time series for each subject, we observed that the Euclidean distances between the mean actigraphies of the two classes vary for different time intervals within the day. The bar plot in Figure 8 effectively shows that the highest Euclidean distance is observed in frames (8-hour time intervals) 8:00-16:00 and 10:00-18:00. Consequently, we decided to analyze further the time frame from 8:00 to 16:00, the second 8-hour period of the day, which we will refer to as the Frame 2 dataset henceforth.

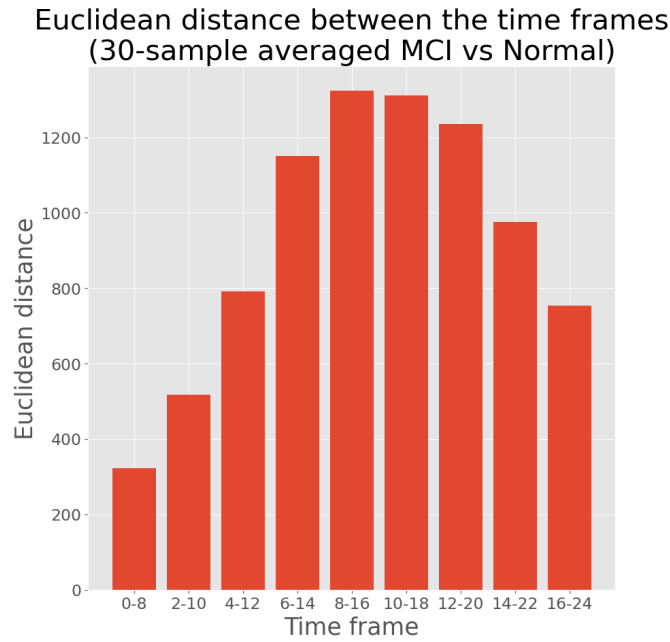


Figure 8: Euclidean distance of 1-day time series (MCI vs. Normal) for different 8-hour time intervals (frames) of the day. The Euclidean distance is calculated between the averaged 1-day actigraphies of the two classes, calculated by 30 randomly selected samples from each class. We observe that the Euclidean distance between the two classes is larger in time frames 8:00-16:00 and 10:00-18:00. This motivated us to investigate classification using only frame 2 (8:00-16:00).

2.2.1.4 Time series – STL decomposition

STL refers to "Seasonal and Trend decomposition using Loess," It is a statistical method introduced in [52] for analyzing a time series into three components: trend, seasonality, and residuals. The trend curve captures the underlying pattern and the progression or direction of the time series. Seasonality, or seasonal variation, captures the patterns or fluctuations that occur in regular periods over the time series, and lastly, the residuals contain fluctuations and noise that cannot be attributed to the seasonal or trend components and represent the random variation and irregularities in the data. By performing STL decomposition on an actigraphy sample, we obtain three new time series from which we can extract features for classification. In our project, we applied STL decomposition on the 1-day and Frame 2 datasets.

2.2.1.5 Time series - Resampling and Normalization

Resampling means representing the original time series with fewer data points; thus, it is a compression action. We decided to explore different resampling frequencies to inspect how our data set behaves under different resolutions. By smoothing the time series, we might be able to reduce the noise of our data. We aggregate the data based on a reduction function when we resample based on a specified frequency using low-pass filtering. After extensive experimentation, we decided to reduce the original time series to one- time point per 30 seconds or 1 minute and use the mean as a resampling (smoothing) function. We determined these resampling periods based on preliminary knowledge from experiments already performed.

Figure 9A demonstrates the skewness of all activity values in the 1-day actigraphy dataset. We aim to make the data distribution more uniform and explore if this transformation helps create features that can better separate the two classes. By observing the plot, it is clear that

the zero activity values dominate the rest. We want to preserve the zero values since they represent a lack of activity, mainly related to sleep and/or rest periods during the day, which we do not want to filter out. The two most widely used transformation techniques to reduce the skewness of the data are the log and square root transformation techniques. We choose to utilize the square root transformation since the skewness is moderate, and we want to preserve the zero values of our data, something that is not possible with log transformation since we need to add a pseudo count to the entire data, to make up for the inability of the log transformation to deal with zero values, which will disproportionately impact the low values. To apply the transformation, we calculate the square root of each activity point.

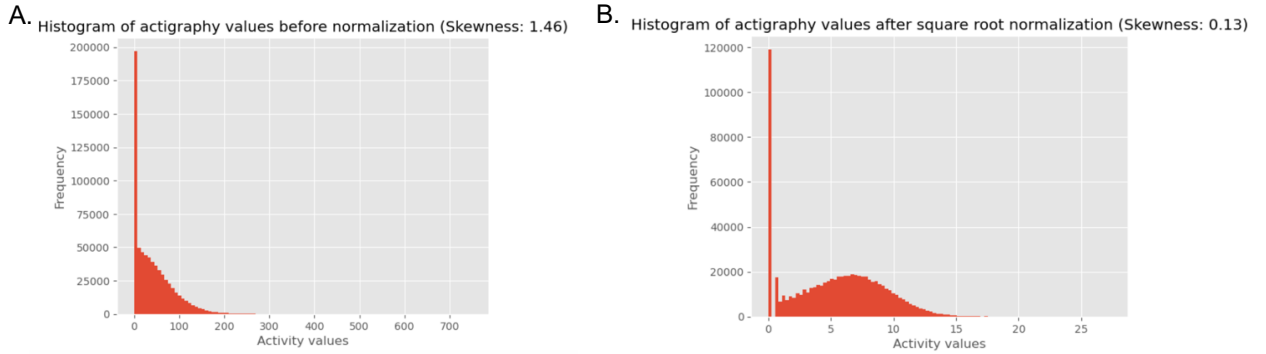


Figure 9: Histogram of 1-day actigraphy values (A) before normalization and (B) after square root normalization. We observe that the skewness of our data is close to zero, meaning that the activity points of our data have a distribution closer to the normal. Furthermore, we observe that the zero values, which are important in our analysis, are preserved.

2.2.2 Feature Extraction

We investigated four families of features, categorizing them into interpretable and non-interpretable. Doctors can easily understand interpretable features that provide valuable insights into how machine learning models work. On the other hand, non-interpretable features hold less clinical significance but could lead to higher classification performance and be useful as potential digital biomarkers for detecting MCI. The interpretable group comprises parametric [44] and non-parametric [25] features extracted from the time series. In contrast, the non-interpretable features we considered are Fourier or Wavelet transform-based features [53]. The primary focus of this thesis was to use machine learning methods to derive insights that can be of valuable help to physicians. Thus, the results ought to be interpretable. Based on preliminary results, which show that parametric features outperform non-parametric features (not included in this thesis project), we decided to delve deeper into the parametric features-based analysis of the different time series constructed based on the original actigraphy measurements.

2.2.2.1 Parametric features

The statistical analysis employed for extracting parametric features is widely known as "cosinor analysis," and it is used to analyze and model regular patterns in time series data [44], such as biological rhythms. By fitting curves of one or more cosines to the data, cosinor analysis enables the characterization of rhythmic patterns. It offers valuable information about the trends and timing of the observed periodic phenomena and their peaks. The

features of interest here generated by the cosinor analysis are the *Mesor*, *Acrophase*, and *Amplitude* (Figure 10).

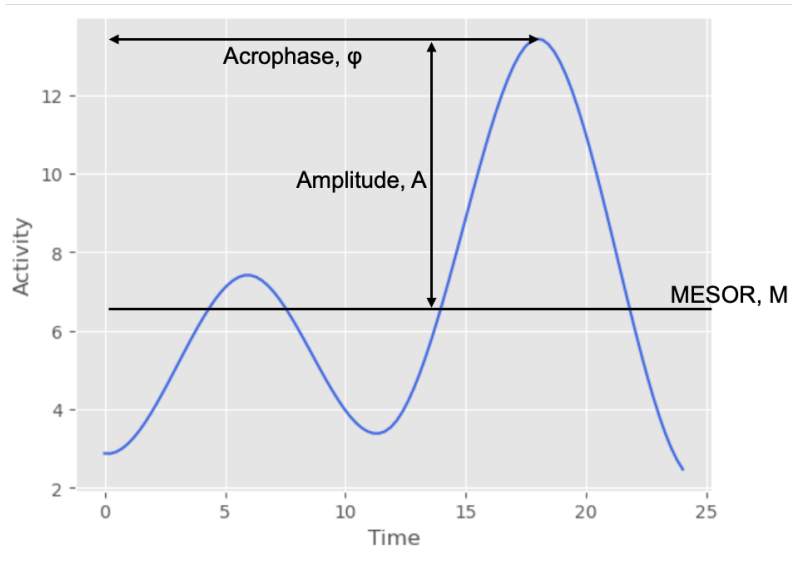


Figure 10: Graphical representation of Mesor, Amplitude, and Acrophase.

In our experiments, we experimented with fitting the time series data with functions being a mixture of one to four cosine curves to extract features. The *Mesor* represents the midpoint or average value of the fitted curve, indicating the baseline level around which the curve's oscillations occur. The *Acrophase* refers to the time of the day of the maximum peak point of the fitted curve, indicating when the activity reaches its highest value. The *Amplitude* represents the maximum deviation, or magnitude, of the fitted curve from its baseline, the *Mesor*, reflecting the range or intensity of the oscillations in the activity pattern. Additionally, cosinor analysis produces two more features. The *R²*, known as the coefficient of determination, measures the curve's goodness of fit to the actual measurements, with higher values indicating a better fitting. The *Marginal Error* (ME), also referred to as the residual error or residual sum of squares, quantifies the discrepancy, or difference, between the observed data points and the values predicted by the fitted curve, representing the residual or unexplained variation in the data set. Considering that we fit each time series with four different parametric model curves, including 1, 2, 3, or 4 cosines, 20 features (4×5) are extracted. Moreover, since STL decomposed time series are characterized by three different components (trend, seasonality, and residuals), we will end up with 60 features in total in the STL case. However, *R²* and ME features are not valuable to physicians and have been excluded from the study, thus reducing the total number of interpretable features to 12 (4×3) for non-decomposed and 36 (3×12) for STL decomposed time series, respectively. This is a reasonable number of features for the analysis to be robust, considering that the ALBION dataset is currently small, and the MCI class (minority) is underrepresented. Following the parsimonious modeling (Occam's razor) principle, we aim to identify the simplest machine learning models that use only interpretable features as input and can generalize well on unseen data.

Since we aim to fit one to four cosinor curves, a multi-component cosinor analysis becomes necessary. This analysis extends the standard cosinor analysis and allows for identifying

multiple underlying rhythmic patterns or oscillations present in the data. The regression model for a multi-component cosinor analysis with $p=1,2,3,4$, cosines can be expressed as:

$$Y(t) = M + \sum_j A_j \cos\left(\frac{2\pi t}{\tau_j} + \varphi_j\right) + e(t), j = 1, 2, \dots \quad (1)$$

In (1), M represents the Mesor, A the magnitude, φ the phase, and τ the period (duration of one cycle) of each cosine; $e(t)$ represents the error term. We set $\tau = 1$ hour [44].

The different time series analyzed, and the number of features extracted for machine learning purposes are summarized in Table 2. The dataset names used in the rest of this report will be the ones in the table.

Table 2: Datasets used in the study, their length, and the number of extracted features. We extract one Amplitude, Mesor, and Acrophase feature for each fitted cosine curve on the time series. We fit four cosine curves.

Resampling period	Time series dataset	Length (in time points)	Number of features
30 seconds	1-day	2880	12
	1-day STL	2880 x 3 (trend, residuals, seasonality)	36
	1-day Trend	2880	12
	Frame 2	960	12
	Frame 2 STL	960 x 3 (trend, residuals, seasonality)	36
	Frame 2 Trend	960	12
1 minute	1-day	1440	12
	1-day STL	1440 x 3 (trend, residuals, seasonality)	36
	1-day Trend	1440	12
	Frame 2	480	12
	Frame 2 STL	480 x 3 (trend, residuals, seasonality)	36
	Frame 2 Trend	480	12

2.2.3 Feature Selection and Processing

In the following paragraph, we will describe specific techniques we used to select and preprocess features before we provide them as inputs to the machine learning pipeline.

2.2.3.1 Feature Selection methods

Features selection, or variable selection, is the task that identifies the most informative features from a larger pool and has a vital role in machine learning and data science in general. Eliminating less informative features can reduce noise, which may prove harmful to simple classification models such as Logistic Regression, Gaussian Naive Bayes, etc. By lowering the number of input variables, we also reduce the complexity and cost of computations associated with modeling and, quite often, raise a trained model's

generalization performance on unseen data. This study investigates the effectiveness of two orthogonal feature selection approaches discussed below.

Maximum Relevance - Minimum Redundancy (MRMR)

Maximum Relevance Minimum Redundancy (MRMR) was first introduced in 2005 [54], [55], and ever since, it has been one of the most widely used methods for automated feature selection, surpassing other popular algorithms such as Boruta [56]. MRMR, in contrast to Boruta, is characterized as a minimal-optimal method since it tries to find the smallest set of features that can achieve optimal predictive results. Therefore, if two features are relevant to the target variable but have the same predictive power, MRMR will only select one and discard the other. Consequently, MRMR selects features that have both predictive power and tend to be unrelated to each other.

MRMR [55] is a model-agnostic method; its only user-defined parameter is the number of features to use. We decided to test our machine learning pipeline with 10, 20, and 30 (whenever applicable) features. Whenever the maximum number of features is close to these three options, we use the maximum number instead. The features are selected iteratively. In each iteration, the algorithm picks the feature with the maximum relevance concerning the target variable and minimum redundancy concerning the features already selected at previous iterations. Specifically, the scoring function for feature f used at iteration i is defined as:

$$score_i(f) = \frac{relevance(f \mid target)}{redundancy(f \mid features \text{ selected until } i - 1)} \quad (2)$$

Select From Model (SFM)

The second feature selection approach we tried is the *SelectFromModel* from the scikit-learn python package [57] (sklearn v1.2.2). Unlike MRMR, this meta-transformer approach picks features according to their importance weights assigned by a specific classifier after training. The user can again define the maximum number of features to be selected by the classifier, set to 10, 20, or 30 in our case. A limitation of this approach is that it applies only to specific model families that support it. Among the models we used in this study, we can apply SFM with Logistic Regression [58], Linear Discriminant Analysis [59], Random Forest [60], and XGBoost [61]. A big advantage of the SFM approach is its small computational time compared to MRMR.

Variance Filter

Before applying any feature selection algorithm, we use a Variance Filter (sklearn v1.2.2) to identify and eliminate features with minimal variability across the dataset. This filtering helps remove quasi-constant features that do not contribute helpful information to the downstream analysis. In our case, we applied a variance filter with a threshold of 0.01; thus, features with a variance less than 0.01 are considered low variability and subsequently removed.

2.2.3.2 Feature Processing

Standardization and Normalization

Normalization and Standardization are two techniques used in data preprocessing to transform the features of a dataset before using them as input to train a machine learning algorithm. Normalization is essential for distance-based classifiers, such as Logistic

Regression (LR) [58], Support Vector Machines (SVM) [62], and k-Nearest Neighbors (KNN) [63]. It ensures that the features of a sample are transformed to a similar range, thus preventing any single feature from dominating the distance calculations. In our case, we utilize the *MinMaxScaler()* from the *scikit-learn* package (v1.2.2) for normalization to scale the features of a sample to a specific range, typically between 0 and 1, while preserving the relative relationships among the feature values.

On the other hand, standardization is not required for distance-based classifiers, but it can still benefit specific algorithms and situations. It transforms the features into zero mean and unit variance across the samples of the whole dataset, allowing for a more straightforward interpretation and comparison of feature importance. We utilize the *StandardScaler()* of the *scikit-learn* package (v1.2.2) for Standardization, which scales the features using their mean and standard deviation.

In our experiments, we perform classification tests twice, with and without standardization. This allows us to compare performance and determine the impact of standardization on the classification results. When standardization and normalization are combined, we typically perform standardization first to center the data and then proceed with normalization to ensure a consistent scaling across all sample features.

Synthetic Minority Over-sampling (SMOTE)

The Synthetic Minority Over-sampling (SMOTE) technique [64] was introduced to address the issue of imbalanced datasets. SMOTE helps to balance a dataset by oversampling the minority class. The idea behind SMOTE is to create synthetic samples for the minority class by generating new data points. Oversampling is done by finding the nearest neighbors of each minority class instance and creating new instances along the line connecting them. Using SMOTE can increase the number of positive (MCI) class instances, making the dataset more balanced. This is important because it allows machine learning algorithms to be trained with a presumably more representative dataset, improving their performance in accurately classifying the minority class. However, while SMOTE evens out the dataset's classes, it does so without significantly increasing the imbalanced class's variance since it creates samples that are derived from the same distribution of the minority class samples.

2.3 Machine Learning pipeline

2.3.1 Concepts

This section will analyze the basic tools and classification model families used in this study before presenting the final pipeline.

2.3.1.1 Supervised Learning and Classifiers

Supervised is a type of machine learning where a model learns from labeled training data to make predictions or classifications. It involves training a model using input features and corresponding target labels to learn their relationship. The trained model aims to generalize the learned patterns from the training data to predict or classify future unseen samples accurately.

Our experiments employed a wide variety of classification model families to explore and compare their performance. They include simple baseline models such as Logistic Regression (LR) [58] (sklearn v1.2.2), Linear Discriminant Analysis (LDA) [59] (sklearn v1.2.2), Gaussian Naive Bayes (GNB) [65] (sklearn v1.2.2), k-Nearest Neighbors (KNN) [63] (sklearn v1.2.2), more elaborate models such as Support Vector Machines (SVM) [62]

(sklearn v1.2.2), and ensemble tree-based models such as Random Forests (RF) [60] (sklearn v1.2.2), and XGBoost (XGB) [61] (xgboost v2.0.0).

By using a diverse range of classification model families, we aimed to assess their strengths and weaknesses. Each model family follows a different fitting approach, which can result in variations in performance depending on the nature of the data, and through that, we sought to gain a comprehensive understanding of how different algorithms address the classification task and identify suitable combinations of (actigraphy-based time series, interpretable features set, and model family) “triplets” for our specific classification task.

2.3.1.2 Hyperparameters tuning - Optuna

Hyperparameters are settings or configurations we choose before training a model, such as its learning rate, the number of hidden layers in a neural network, or the regularization strength in linear models. Tuning these hyperparameters is crucial to improve the performance and generalization of a model family. Identifying the best set of hyperparameters for a learning algorithm is known as hyperparameter optimization or tuning in machine learning. This results in selecting an optimal model instance (algorithm with fixed hyperparameters).

Optuna (optuna v3.3.0) is a popular hyperparameter tuning package [66] that automates the search process by intelligently exploring the hyperparameter space. In our implementation, Optuna uses an optimization algorithm to efficiently search for the best set of hyperparameters by evaluating the model's performance on a validation dataset via a 5-fold stratified Cross-Validation (CV). It iteratively suggests new hyperparameter combinations based on the results of previous evaluations, gradually converging toward the optimal configuration in a Bayesian manner. With Optuna, we can save time and effort by automating the hyperparameter tuning process and finding optimal settings for our model.

We use the Optuna package with specific settings tailored to our needs. In each hyperparameter tuning session, we perform 100 tuning “trials” to explore a wide space of hyperparameter combinations and allow our models to reach or get close to the global optimum test set performance. For the sampling algorithm, we chose the *TPESampler* based on a Tree-structured Parzen Estimator [67]. It can efficiently explore the hyperparameters space by adapting the combinations sampling distribution to focus on promising areas, leading to faster convergence and more effective search. To optimize the tuning process, we incorporated the *SuccessiveHalvingPruner* algorithm, which terminates unpromising trials early during the optimization, allowing us to allocate computational resources to more promising ones.

2.3.1.3 Performance Metrics

Our experiments use different metrics to establish a holistic view of the different models' performance. The metrics used are Balanced Accuracy, F1 score, F2 score, Matthew's correlation coefficient (MCC) for both the test and the training sets, Precision, Recall, and Specificity. The different metrics provide different insights about the model's performance while allowing the mitigation of the limitations and biases of individual metrics. Finally, using multiple metrics enhances our confidence in the results. Since our dataset is unbalanced, we have chosen MCC as the performance metric of the training procedure [68].

2.3.1.4 Nested Cross-Validation

Our machine learning pipeline implements Nested Cross-Validation (nCV) [69] to select the best classifier model for a given experiment (a time series, preprocessing, feature set

combination). nCV ensures that we correctly estimate the expected model's performance with unknown data and mitigate the risk of overfitting, which arises when the model closely fits the training data but performs poorly on new, unseen data, allowing us to assess its generalization ability effectively. nCV is a method of choice to select estimators expected to make accurate predictions in real-world scenarios, primarily when training relies on small-size and imbalanced datasets, as in our case [70].

Our nCV pipeline (Figure 11) consists of two main components: the outer and inner loops, with the latter employed for hyperparameter tuning. The available data is divided into training and test sets in the outer loop using stratified cross-validation with five folds (CV (5)). The test fold of the split remains untouched, and classifiers are not exposed to this data to prevent any information leakage. However, the training set (remaining four folds) is passed through the inner loop CV, where Optuna is used for hyperparameter tuning. In that loop, the performance of each hyperparameter combination is evaluated using yet another 5-fold stratified cross-validation. Subsequently, different approaches are employed to select the optimal trial (hyperparameter combination) from the 100 trial sets constructed by Optuna. Finally, the classifier model is trained using the selected optimal hyperparameter combination (best model instance for the current iteration of the outer loop) and the training data. This entire process is repeated for each outer loop iteration, corresponding to a different train-test data split, to provide a reliable estimate of the model's expected performance statistics (mean and variance) on unseen data. During one nCV repetition (round), each dataset sample will appear once in the test set. To increase this number and be able to assess the per-sample classification performance of different algorithms, we implemented multiple nCV rounds with different random seeds controlling the outer and inner loop train-test and train-val splits, respectively. So, at the end of the nCV process (completion of all rounds), we can estimate the probability of misclassification of every sample for every algorithm. This can help us identify “stubborn samples” that tend to be misclassified with high probability by all ML algorithms using a specific feature set family (e.g., parametric features).

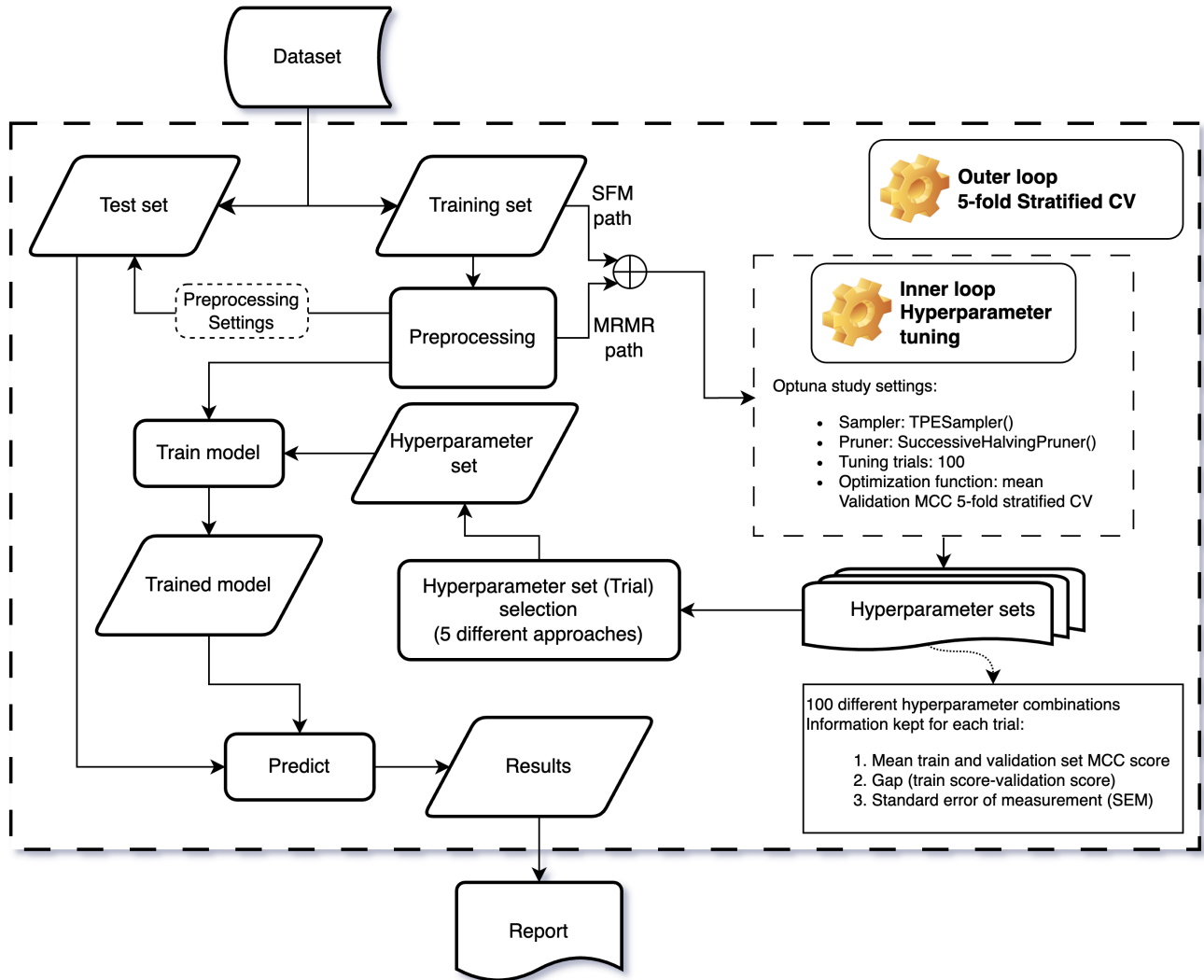


Figure 11: The nested cross-validation (nCV) pipeline comprises an outer and inner loop, each implementing a stratified 5-fold CV. Initially, the dataset is split into training and test sets. Preprocessing occurs once (before the inner CV begins) in the MRMR feature selection path or in each inner loop in the SFM features election path. The training set is then used for the hyperparameter tuning in the inner CV (5) loop using Optuna, and the optimal hyperparameter set is selected after applying four different model selection methods. The final model instance is trained using the selected optimal hyperparameters combination and the training set data. The trained final model instance is used to make predictions using the test set, and the results of the outer loop are stored before proceeding to the next train-test split iteration of the current nCV round. The whole process is repeated in each stratified nCV round.

2.3.1.5 Hyperparameters selection methods

Upon the end of each inner split/hyperparameter tuning run, we acquire a table with 100 hyperparameter combinations, each corresponding to an Optuna trial. Our goal in the hyperparameter set selection stage is to pick an optimal combination in some sense to reduce the potential overfitting so the model selected can generalize well on unseen data. For this purpose, we employ the following five different model selection approaches.

Optimum Validation score

This is the most common and widely used approach. The pipeline selects the hyperparameter combination that achieved the highest validation score, in our case, the mean validation MCC value of the inner 5-fold CV (one Optuna trial). We will refer to this approach as “Best.”

Standard Error of the Mean (SEM)

The One Standard Error of the Mean (One-SEM) technique aims to identify the simplest model instance whose mean validation score is within one standard error of the mean of the “best” model's validation score. Its motivation is to minimize the probability of overfitting by using a less complex model than the “Best” following Occam's razor principle [71]. Figure 12 shows an example illustrating the concept, where we observe how the one-SEM rule is applied. We find the “best” trial with the highest mean validation score and then select the model instance with the simplest hyperparameter combination and a mean MCC score in the band defined by the “one-SEM” rule (see the two horizontal dashed lines), aiming to reduce the chances of overfitting.

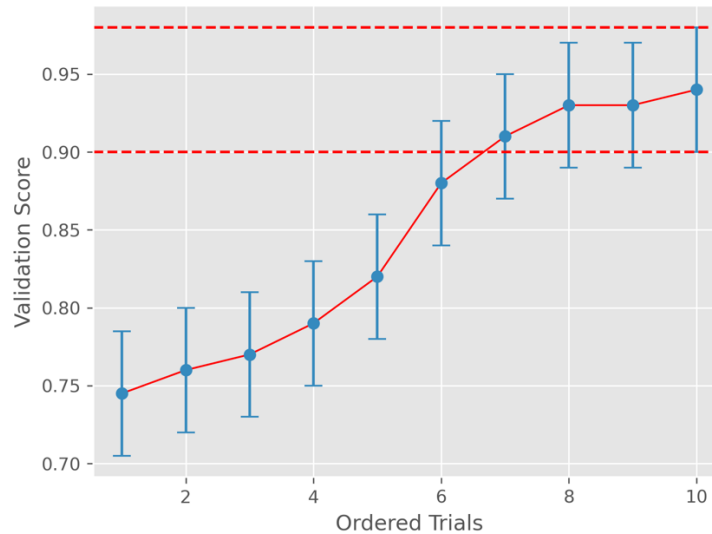


Figure 12: Example of hyperparameter set selection (model selection) using the one-SEM rule. The trials are ordered based on validation score in ascending order. The trial with the best mean validation score is first identified along with its standard error of the mean (SEM) considering all train-test splits of the inner nCV loop. This is trial 10 in the Figure. Then all trials that have a lower mean validation score but no less more than one SEM relatively to the best trial's mean score are identified. Between these trials we select the one with the simplest hyperparameters combination. The trials considered are 7, 8, 9, and 10 in the Figure.

In more detail, after acquiring a table with 100 hyperparameter combinations upon the end of the Optuna tuning, we sort the table in ascending order based on the MCC validation score. We identify the trial with the best mean validation score and filter out all trials that fall outside the boundaries defined by the best MCC score and the SEM score of that trial using the one-SEM rule. The resulting table is then reordered according to the values of each hyperparameter, based on each hyperparameter's importance. For instance, in the SVM classification model, which has two hyperparameters, 'C' and 'degree,' the table is first ordered by 'C' which we have chosen as more important than the other hyperparameter. If multiple trials have the same 'C' value, they are ordered based on the 'degree'

hyperparameter. This approach is similarly applied to the rest of the classification model families. The applied rules are summarized in Table 3. We select the first row of the table which defines the trial with the simplest hyperparameters combination.

Table 3: Applied rules for the one-SEM hyperparameter selection approach. This table outlines the prioritization of hyperparameters for various classification model families using the one-SEM rule. Each row represents a classification model family and lists the hyperparameters in the order of their importance, from left to right, with the simplest values (low or high) indicated in parentheses.

Classification model family	Hyperparameter (simpler values)			
LR	C (low)			
GNB	var_smoothing (low)			
LDA	shrinkage (high)	solver (low)		
KNN	n_neighbors (low)			
SVM	C (low)	degree (low)		
RF	n_estimators (low)	max_depth (low)	min_samples_split (low)	min_samples_leaf (low)
XGB	n_estimators (low)	max_depth (low)	reg_lambda (high)	reg_alpha (high)

Gap Heuristic Techniques

The Gap heuristic approaches discussed below are similar. However, they differ in certain aspects. We will analyze the first in detail and then note the main differences to the other methods.

The first approach focuses on optimizing the training score and will be called "Gap score optimization 1 (GSO1)". The procedure starts with the acquisition of the hyperparameter trials' data frame, which includes for each trial information about the mean MCC scores of the validation and the training set of the inner loop CV (5) and the "Gap score," meaning the absolute difference between the mean training and mean validation scores. In the GSO1 heuristic, we first find the best training MCC score and calculate a threshold by multiplying it with a number K , which, in our case, is set to $K=0.85$. Then, we find the subset of trials with a training MCC score larger than $K \times (\text{Best MMC Score})$. Finally, we select the trial with the smallest Gap score among them. Using this approach, we aim to minimize the score difference between the training and the validation set to reduce overfitting and thus boost the generalization ability of the selected trial. The main drawback of this heuristic is that we arbitrarily choose the number K to calculate the threshold.

The second heuristic approach, called "Gap Score optimization 2 (GSO2)", is similar to the previous one. However, the first sorting of the data frame is done based on the mean Validation MCC score instead of the mean Training MCC score of the inner loop CV. Finally, for the third Gap approach, which is a variation of the first, we first find the top $K=20$ trials

with the highest mean training MCC score instead of setting a threshold. Again, the user should decide on the value of K trials considered.

Table 4: Summary of model selection approaches. We apply all five approaches to the same validation trial.

Model Selection Approach	Selection details
Best validation score	Select the hyperparameters set that produced the highest mean validation score
Standard Error of the Mean (SEM)	Select the hyperparameter set that is the simplest one with a validation score within one standard error of the best validation model's score.
Gap Score Optimization 1 (GSO1)	Find the trials with Training score $> K * \text{Best training score}$. Select among them the hyperparameters set with the smallest Gap (absolute difference between Training and Validation scores).
Gap Score Optimization 2 (GSO2)	Find the trials with Validation score $> K * \text{Best validation score}$. Select among them the hyperparameters set with the smallest Gap score (absolute difference between Training and Validation scores) Default value $K=0.85$.
Gap score Optimization 3 (GSO3)	Find the K trials with the highest mean training score. Select among them the hyperparameter set with the smallest Gap score (absolute difference between Training and Validation scores). Default value $K=20$.

Depending on which feature selection approach we used in each experiment, we implement two alternative nCV pipelines. In experiments that employ MRMR feature selection (Figure A1), a computation heavier and significantly slower than SFM, the feature selection step is executed only once per nCV outer loop iteration, right after each train-test outer split.

On the contrary, SFM's reduced computational time allows us to perform the feature selection step in the inner loop after each inner training and validation set split (Figure A2). Upon the end of the hyperparameter tuning and the optimal hyperparameter set selection, we perform an SFM feature selection in the outer loop using the selected model instance with the optimal hyperparameters. The SFM allows us to create pipelines that theoretically can generalize better than the MRMR ones because by performing feature selection using only the inner loop's training set, SFM does not "touch" the validation set for feature selection. We perform the final feature selection in the outer loop, using the optimal model instance selected during the hyperparameters tuning process.

2.4 Visualization and further exploration

2.4.1 Nested Cross-Validation results exploration

A crucial part of this study is using visualization to understand better, explore, and evaluate the obtained results. Getting into more detail, for each machine learning experiment setting, we create a boxplot per performance metric based on the 50 values obtained from the 50 total outer train-test splits of the nCV experiment (10 rounds of 5-fold outer-CV splits). These boxplots help us to visually compare experiment statistics by inspecting their expected median and inter-quartile range for individual performance metrics across many experiment settings. We also use two dimensionality reduction methods to explore the feature space of our datasets: Principal Component Analysis (PCA) [72] and Uniform Manifold Approximation and Projection (UMAP) [73].

As the pipeline can be used to define and implement a plethora of predictive modeling experiments, we must first explore the solution space in terms of the Recall and Specificity tradeoff and identify the experiments of most interest that meet the clinical criteria of the ALBION study. To that end, we create Recall vs. Specificity scatter plots and tables that allow us to explore the different experiments, sort them in different ways (e.g., by achieved median MCC per actigraphy and classifier algorithm), and then apply certain constraints (e.g., a minimum recall and/or specificity threshold) to identify experiments of interest. The scatter plots are colored in terms of the classifier model and actigraphy dataset to facilitate the analysis. Using these plots, we can explore the solutions space and select experiments of interest to focus on. To do that, we apply a set of rules formed after discussion with the medical experts. Positive MCI class predictions of this actigraphy-based approach may advise the doctors to perform additional tests on the patients, such as the invasive Lumbar puncture (LP) procedure, which can cause distress to the study participants. As such, our primary goal is to minimize the number of false positives by maximizing the number of true negatives (Specificity) and then maximize the number of true positives (Recall). Therefore, we primarily focus our attention on experiments with specificity ≥ 0.7 and recall ≥ 0.5 .

We carefully consider each actigraphy dataset, starting from the highest achieved recall and moving down to a recall equal to 0.5 when we try to build the list of top models at the end of the nCV process. Our approach systematically tries to identify the top models per actigraphy at each recall level. If just one model stands out, achieving the highest specificity level, it gets added to the list. When two experiments exhibit the highest specificity level using the same classifier, we include both of them in the list. However, suppose there are more than two experiments achieving the highest specificity level but employing two or more different classifier families. In this case, we expanded our selection from that actigraphy type and picked up to three models to include in the top models' list.

2.4.2 Interpretation of model predictions

2.4.2.1 SHAP-based analysis

Upon selection of the top experiments of interest, the next step of our study is the development of an interpretation pipeline that will help generate valuable insights crucial for the diagnosis process. To interpret the selected experiment results, we use Shapley Additive exPlanations (SHAP) package [74]. This package uses the classical Shapley values from game theory [75] and their related extensions to produce insights on how and to what extent each feature affects the class predictions made by a model instance. However, it is essential to note that SHAPs can identify correlations and not causal relations between features and the target variable.

The SHAP analysis is executed through the following structured workflow. For every chosen experiment, we must apply its specific experiment settings. These settings include the combination of machine learning algorithm, feature selection method, the number of features used, optimal hyperparameter selection method, and pre-processing steps. The first step is to select the features used to build the final model instance for evaluation and interpretation. To do that, we examine the report from the corresponding nCV experiment, extract the selected feature set in each outer CV (5) loop iteration, and create a dictionary with keys the feature names and values of their frequencies (number of appearances in the nCV outer loop iterations). When using SFM for feature selection, we calculate the average number of selected features, K , across all outer loops. When using MRMR, where the user defines the number of desired features K to be used, the average number equals K . We select the top

K features (with the highest frequencies) from the created dictionary in both cases. If K is a real number (SFM) we round it to the next integer.

Following feature selection, we use the whole dataset to execute a CV (stratified 5-fold) Optuna-based hyperparameters tuning search using the same range of hyperparameters as in the nCV process, and the best trial is selected using the hyperparameters selection heuristic defined in the experiment's settings. Finally, we execute the same ten rounds of the nCV pipeline outer loop (using the same seeds) but only for the final model. The train and test sets are pre-processed based on the experiment's settings. Each iteration uses the training set to train the final model instance (with the selected features and optimal hyperparameters). The test set is used to make model predictions, calculate the performance metrics, and compute SHAP values. At the end of all iterations, performance metrics statistics and average SHAP values are calculated. Figure 13 summarizes this process.

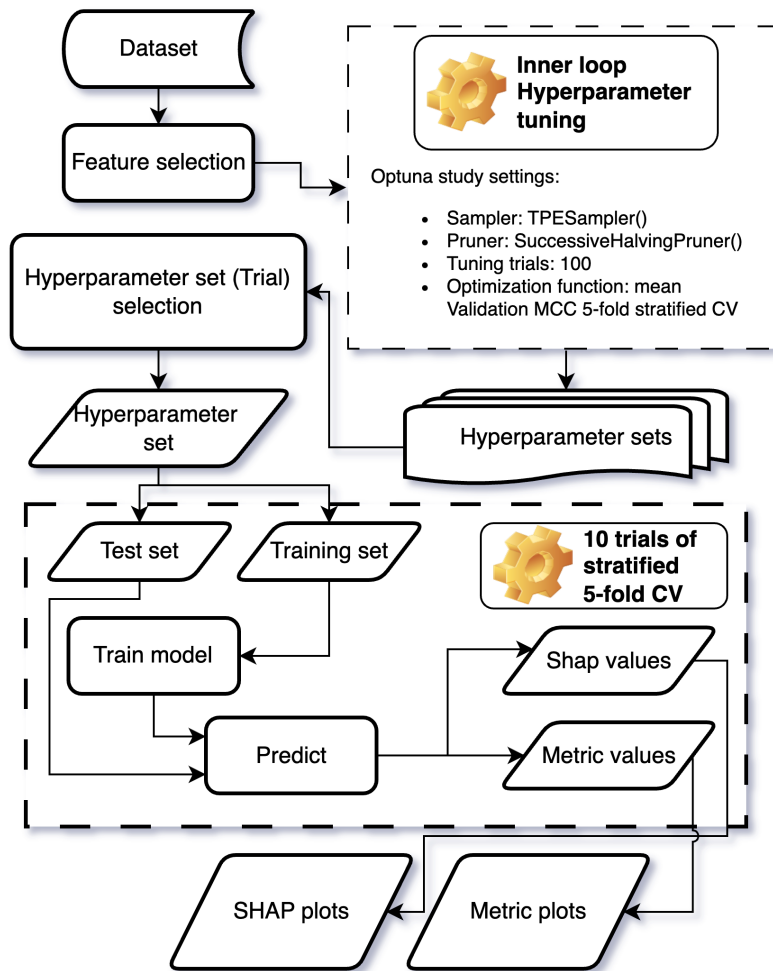


Figure 13: Final model building and SHAP values generation. Once we have selected an experiment of interest, we perform feature selection based on the experiment's settings. Once the dataset has been reduced to the features of interest, we perform hyperparameter tuning in the same way as the main pipeline. Using the selected hyperparameter set, we perform the same 10 trials of a stratified 5-fold CV, with the same seeds as the initial nCV experiments, to produce SHAP and Metric values for each experiment. Using these values, we produce SHAP and Metric plots.

We create bar plots and beeswarm plots using the average SHAP values we obtained. The Bar plot is a global feature importance plot, and the overall significance of each feature is determined by calculating the average absolute value of that feature's SHAP values across all samples. The beeswarm plot illustrates which features influence the model's predictions and to what extent. It represents each sample's explanation as a single dot on the respective feature's row, where that feature's SHAP value determines the dot's horizontal position. Dots accumulate vertically along each feature row to illustrate their density, and the color is employed to indicate the feature values for each sample.

To delve deeper into the individual sample's classification, we assess each sample's probability of correct classification. To do so, we compute the percentage of correct classification per sample (sample accuracy) for the chosen nCV experiments and related experiments using the same settings but different ML model families. For instance, if the best model for an actigraphy type utilizes GNB, we first calculate the percentage of correct classifications for every sample, considering all rounds of the nCV outer loop when using GNB. In addition, we calculate the average percentage of correct classifications per sample when using all model families and the same experiment settings as for the GNB. Finally, we calculate the correct classification percentage per sample using the same nCV outer loops but the final GNB model instance, built as shown in Figure 13.

Using this per-sample information for each experiment (three classification expected accuracy scores per sample), we can create two types of insightful plots: a dot plot and two heatmaps, one for each class. The dot plot can be used to compare the per-sample accuracy of the final model instance against the corresponding nCV experiment using the same ML algorithm, as well as the average correct classification percentage across all ML model families used in nCV for this experiment. Regarding the heatmaps, the rows correspond to the dataset samples, and the columns correspond to the per-sample accuracy of the different model families used in nCV, their average classification score, and the final best model instance score. The heatmaps can help us identify “stubborn samples” (and “easy samples”) that get misclassified (that get correctly classified) by all ML algorithms used in the experiment. We can analyze these samples more carefully using SHAP values to identify the features that best explain the ML models' extreme success (or extreme failure) in classifying these samples. This can provide insights as to which features best explain misclassifications and which samples are extremely difficult to classify, potentially due to noise, outliers, poor feature set selection, or even mislabeling of samples.

Apart from the global SHAP analysis, which gives us insights into the aggregated way our model makes predictions based on the whole dataset, we produce plots that focus on individual predictions and samples, thus performing a local SHAP analysis. These local explanations are crucial for understanding the model's behavior in specific instances, especially when predictions seem counterintuitive or erroneous. For that purpose, we create three types of plots: decision plots, heatmap plots, and waterfall plots.

Decision plots visualize SHAP values for each feature and each instance (sample) in the dataset, allowing us to decipher how the model makes predictions. The x-axis represents the model's output, and the y-axis lists the features, ordered by descending importance. SHAP values are added to each instance, moving from the bottom of the plots towards the top. This visualization allows us to determine how each feature impacts the model prediction for each test sample represented by a line. We highlight the misclassified samples by changing the continuous line into a dashed one, making it easier to identify these instances and understand the reasoning behind their misclassification. Decision plots are particularly useful

for detecting patterns, and interactions between features, and identifying outliers or unusual instances the model may be struggling with.

By combining these different types of SHAP plots, we can comprehensively understand our model's behavior globally across the dataset and locally for individual instances or subsets of them. This interpretation is crucial for building trust in the model, identifying potential biases or areas for improvement, and aligning the model's reasoning with domain knowledge and expectations.

The results of the SHAP analysis are validated through a 2-sample Kolmogorov-Smirnov (KS) test, which is a nonparametric statistical test used to determine if two samples come from the same distribution. This test is particularly useful when comparing two independent samples without assuming any specific distribution. The method involves computing the empirical cumulative distribution function (ECDFs) for each sample and then calculating the maximum vertical distance between these ECDFs, the so-called “D” statistic. This statistic, along with the associated p-value, helps to evaluate the significance of the observed differences. The test is especially useful when the underlying distribution is unknown like in our case.

3. RESULTS & DISCUSSION

3.1 Summary of results

This section summarizes the results of the plethora of experiments performed using the developed ML pipeline. The reader can find the complete set of results in the supplementary material. We first visually summarize the results of non-SMOTE experiments obtained using an actigraphy time series (TS) smoothed down to 1 minute period. Figure 14A provides a visual performance summary of these experiments. Each dot in the recall-specificity plot represents a set of experiments. The median MCC of these experiments (each calculated over the 50 values produced by the 10 nCV rounds) ranged from -0.12 to 0.41. By focusing on the most interesting experiments that achieved a recall ≥ 0.5 and a specificity ≥ 0.70 , we obtain Figure 14B, summarizing 240 experiments.

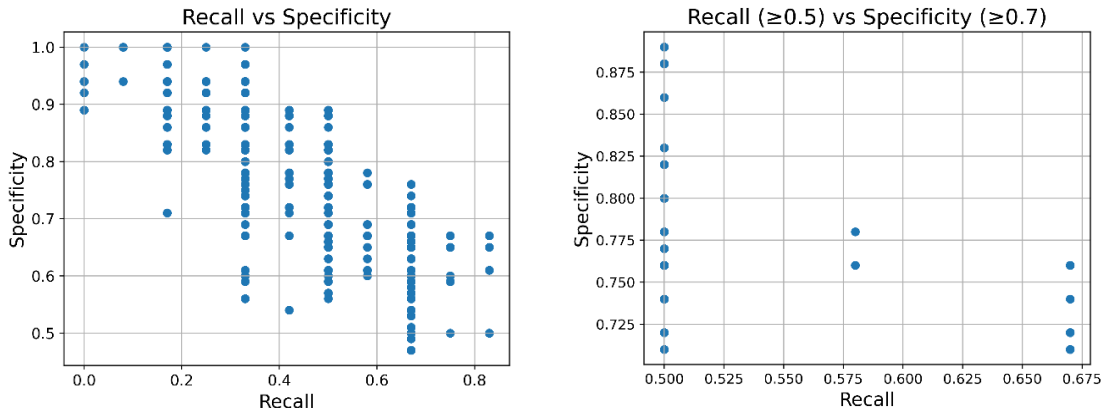


Figure 14: Results summary Recall vs Specificity plots. Each dot in the plots corresponds to a set of experiments achieving a specific value in the recall-specificity space. A. All experiments (2498) are represented. B. The experiments surpassing the Recall ≥ 0.5 and Specificity ≥ 0.7 cut-offs (240 experiments). All experiments in this plot are non-SMOTE, using 1-min smoothed actigraphy time series).

Figure 15 shows the same information as Figure 14B but adds information about the classifier model family used. Experiments with the same recall value have been spread out on the recall axis per family model for visualization purposes. These experiments are grouped in the same arrowbox that points to their recall value. The classifier model family that offers the best scores in terms of recall and specificity is LDA since LDA experiments (red dots) occupy the top position (highest specificity) at every recall level. Some LDA experiments also offer a more balanced performance. XGB models, on the other hand, offer solutions with high specificity, and GNB models offer solutions with high Recall. SVM experiments do not seem to provide solutions matching the other classifiers.

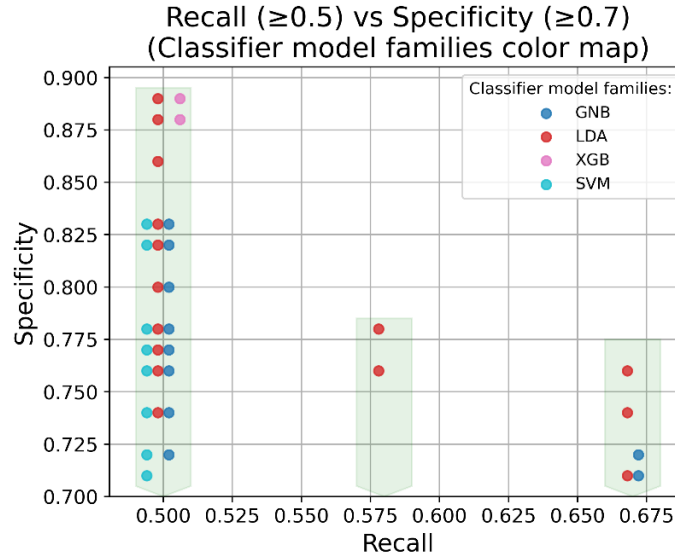


Figure 15: Each point in the plot defines an experiment, and the color represents the classifier family. The plot presents the solution space of experiments that surpass the Recall (≥ 0.5) and Specificity (≥ 0.7) cut-offs. The points have been jittered to showcase the different families at each coordinate. The number of non-SMOTE, 1-minute resampled actigraphy experiments that surpass the cut-offs are 98 LDA, 73 SVM, 65 GNB, and 4 XGB experiments.

Figure 16 also depicts the same information as Figure 14B but now adds the information of the actigraphy dataset used. Experiments that have the same recall value have been spread out on the recall axis per actigraphy used for visualization purposes. These experiments are again grouped in the same arrowbox that points to their recall value. By following the same visualization approach as for Figure 15, we observe that the actigraphy dataset that offers the best specificity or recall, and more balanced results between the two is the decomposed 1-day STL dataset. Moreover, 1-day, 1-day Trend, and Frame 2 STL datasets also offer experiments with high specificity. In addition, Frame 2 and Frame 2 Trend datasets offer experiments with high recall but without reaching the top specificity scores.

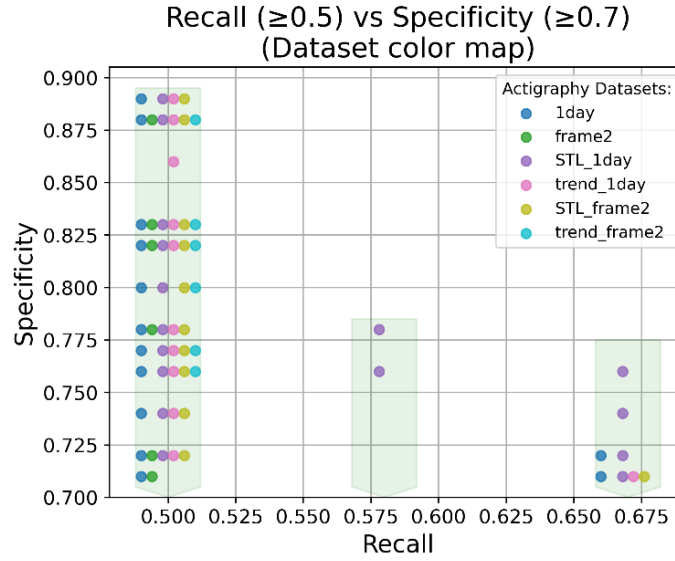


Figure 16: Each point in the plot defines a set of experiments, and the color represents the actigraphy used. Only experiments that surpass the Recall (≥ 0.5) and Specificity (≥ 0.7) cut-offs are represented. The number of non-SMOTE, 1-minute resampled actigraphy experiments that surpass the cut-offs are: 97 1-day STL, 43 1-day, 40 1-day Trend, 34 Frame 2 STL, 18 Frame 2, and 8 Frame 2 Trend type experiments.

For each recall level ≥ 0.5 (Figure 14B) and each type of actigraphy time series reaching that level, we show in Table 5 the best two experiments (or the best three if they employ different classifiers) in terms of achieving the highest specificity. We consider these to be the top 20 setups in terms of expected generalization performance identified using our comprehensive data analysis. The table reports their median test set metrics estimated using the 50 outer loop train-test splits of the repeated nCV process. Please note that the experiments (setups) included in the table for each actigraphy type may differ in terms of preprocessing steps, classifier algorithms, feature selection methods, number of features, and model selection methods. The LDA is the most used algorithm in these experiments (appears in 14 out of 20 rows of the Table) and is among the top choices for all actigraphy-based time series analyzed. The GNB and XGB classifiers also appear 3 times each in the table. GNB achieves the highest recall level at the expense of lower specificity for different actigraphies. XGB achieves very high specificity but only with STL decomposed actigraphies, which result in larger feature sets. The rest of the algorithms considered in the study are not represented in the set of top-performing experiments per actigraphy summarized in Table 5.

For the top 20 experiments summarized in Table 5, a final hyperparameter tuning step was performed, using the entire dataset and 5-fold CV, as described in section 2.3.1.4, in order to build a “final model” per experiment. The outer loop train-test splits of repeated nCV were then used to estimate the final models’ expected generalization performance. Final models constructed this way that reached the specified recall and specificity levels are listed in Table 6. For the models bolded in this table, we performed SHAPs-based analysis to assess how the parametric features correlate with the models’ classification decisions, as discussed in section 2.4.2. We selected for SHAPs-based interpretation one representative model per actigraphy type among those listed in Table 6. In making this selection, we tried to balance recall and specificity, prioritizing using the original undecomposed datasets over their trend components resulting after STL decomposition.

Table 5: Best performing experiments per actigraphy and recall levels – nCV median metrics. We split the different actigraphy datasets with a bolder line. The experiments that use SFM choose the optimum number of features and thus we report the mean number of features +/- one SD.

No.	Classifier	FS method	Number features	TS normalized	Features scaled	Hyperparameters Selection method	Actigraphy dataset	Balanced Accuracy	MCC	Recall	Specificity
1	GNB	MRMR	12	FALSE	FALSE	GSO2	1-day	0.6389	0.2658	0.6667	0.7222
2	LDA	MRMR	12	TRUE	TRUE	SEM	1-day	0.6389	0.3589	0.5	0.8856
3	LDA	SFM	5.56 (+/-0.83)	TRUE	TRUE	GSO1	1-day	0.6381	0.3131	0.5	0.8824
4	LDA	MRMR	12	FALSE	TRUE	Best	Frame 2	0.6348	0.2646	0.5	0.8824
5	LDA	SFM	5.14 (+/-0.8)	FALSE	TRUE	GSO1	Frame 2	0.6217	0.2941	0.5	0.8333
6	GNB	MRMR	12	FALSE	FALSE	GSO2	1-day Trend	0.6389	0.2658	0.6667	0.7059
7	LDA	SFM	5.46 (+/-0.78)	TRUE	TRUE	GSO1	1-day Trend	0.6389	0.3612	0.5	0.8889
8	LDA	SFM	5.82 (+/-0.93)	TRUE	TRUE	Best	1-day Trend	0.6389	0.3529	0.5	0.8824
9	GNB	MRMR	20	FALSE	FALSE	GSO2	1-day STL	0.6618	0.3235	0.6667	0.7222
10	LDA	MRMR	10	FALSE	TRUE	GSO3	1-day STL	0.6667	0.3608	0.6667	0.7647
11	LDA	MRMR	20	FALSE	TRUE	GSO3	1-day STL	0.6879	0.3534	0.6667	0.7647
12	LDA	SFM	10.0 (+/-0.0)	FALSE	TRUE	GSO2	1-day STL	0.6667	0.3273	0.5833	0.7778
13	XGB	MRMR	20	FALSE	TRUE	GSO3	1-day STL	0.6389	0.3466	0.5	0.8889
14	XGB	SFM	9.46 (+/-1.14)	FALSE	FALSE	GSO3	1-day STL	0.6389	0.3284	0.5	0.8824
15	LDA	MRMR	12	FALSE	TRUE	Best	Frame 2 Trend	0.6389	0.2646	0.5	0.8333
16	LDA	SFM	5.4 (+/-0.69)	FALSE	TRUE	GSO3	Frame 2 Trend	0.6381	0.3235	0.5	0.8824
17	LDA	MRMR	30	FALSE	TRUE	GSO3	Frame 2 STL	0.6389	0.247	0.6667	0.7059
18	LDA	MRMR	20	FALSE	TRUE	Best	Frame 2 STL	0.6389	0.3333	0.5	0.8889
19	LDA	MRMR	30	FALSE	TRUE	GSO1	Frame 2 STL	0.6389	0.3284	0.5	0.8889
20	XGB	MRMR	10	FALSE	FALSE	SEM	Frame 2 STL	0.6389	0.3456	0.5	0.8824

Table 6: Final models estimated generalization performance. With bold are noted the experiments we are analyzing further as they have the optimum combination of Recall and Specificity.

No.	Model name	FS method	Number of features	TS normalized	Features scaled	Hyperparameter Selection approach	Actigraphy dataset	Balanced Accuracy	MCC	Recall	Specificity
1	GNB	MRMR	12	FALSE	FALSE	GSO2	1-day	0.6356	0.2512	0.5000	0.7435
2	LDA	SFM	5	FALSE	TRUE	GSO1	Frame 2	0.6618	0.3333	0.5000	0.8578
3	GNB	MRMR	12	FALSE	FALSE	GSO2	1-day Trend	0.6356	0.2526	0.5000	0.7222
4	LDA	MRMR	20	FALSE	TRUE	GSO3	1-day STL	0.6944	0.3429	0.6667	0.7647
5	GNB	MRMR	20	FALSE	FALSE	GSO2	1-day STL	0.6618	0.3357	0.5000	0.8235
6	XGB	MRMR	20	FALSE	TRUE	GSO3	1-day STL	0.6593	0.3608	0.5000	0.8824
7	LDA	SFM	5	FALSE	TRUE	GSO3	Frame 2 Trend	0.6503	0.3333	0.5000	0.8333

3.2 Interpretation of model predictions

The first experiment we considered for SHAPs-based interpretation uses the 1-day STL decomposed actigraphy time series. It therefore uses a relatively large number of features (20) resulting from the multi-cosinor analysis of its three components (trend, seasonality, and residuals). Using the LDA classifier, it achieved the highest level of recall among all experiments (0.6667) at a specificity level that exceeded the 0.7 threshold (0.7647). The SHAPs bar plot of the features (Figure 17A) shows that the Mesors of the seasonality component, particularly the "season_1day_2_mesor" and "season_1day_1_mesor" features, have the highest impact on the model's output, indicating their highest importance as influential features for classification. We remark that the numbers 1_ or 2_ preceding the "mesor" feature names indicate the number of cosine curves used when performing multi-cosinor fitting to the decomposed actigraphy time series data. The beeswarm plot (Figure 17B) reveals the intensity of this influence, indicating that lower Mesor values are associated with the positive (MCI) class. In comparison, the association of higher Mesor values is not as strong with the NC class. After the Mesors, the second most important set of features seem to be the Amplitudes, specifically "season_1day_4_amplitude" and "season_1day_3_amplitude", which have a similar effect on model decisions as the Mesors, but with lower intensity. Finally, "trend_1day_3_acrophase" is the only Acrophase feature selected by MRMR in this experiment. The beeswarm plot suggests a very strong association of high "trend_1day_3_acrophase" values with the NC class, which is not observed for the Mesor and Amplitude features.

Keeping everything discussed in mind and examining the decision plots, we can reach similar conclusions regarding the features and their effect on the model's decisions. In Figure 17C and 17D, we observe how each sample (corresponding to a line trajectory) is guided toward the edges of the plot and what features contribute most to the path toward its final destination. We observe that the "trend_1day_3_acrophase" feature appears at a higher (3rd) position in the list of features in Figure 17D (mispredicted samples) relative to Figure 17C (all samples), suggesting that, in addition to the Mesors, "trend_1day_3_acrophase" plays a significant role in mispredictions. We also observe in Figure 17D that trajectories for some red samples, although misclassified (FPs) experience a shift towards the negative correct class, which can be explained considering the strong association of high "trend_1day_3_acrophase" values with the normal class.

Overall, the global (dataset level) and local (sample level) SHAP analysis reveals the relative importance of characteristics of the seasonality and trend components of the STL decomposed actigraphies, predominantly their Mesors and Amplitudes, in the model's predictions. The analysis also identifies Acrophase as a feature requiring further investigation.

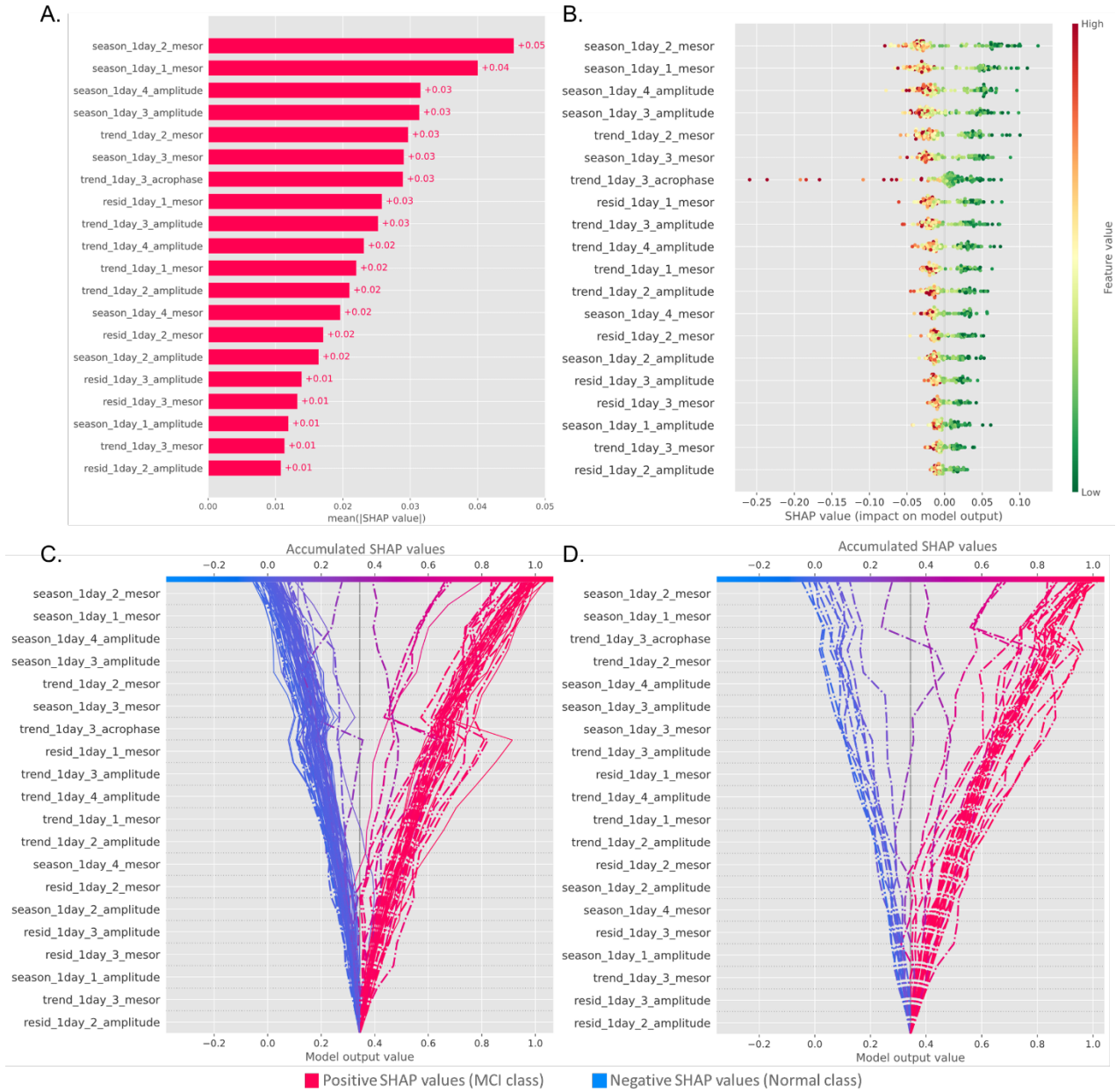


Figure 17: SHAP analysis of 1-day STL decomposed actigraphy dataset. Bar plot (A), beeswarm plot (B), decision plot of all samples (C), and decision plot of only the misclassified samples (D). Experiment details: Classifier: LDA, Feature Selection: MRMR, Number of features: 20, Time Series Normalized: FALSE, Features Scaled: TRUE, Hyperparameter Selection method: GSO3, Resampling period: 1min, Actigraphy Dataset: 1-day STL decomposed.

Next, we examined the final model in Table 6 which uses the non-decomposed 1-day average actigraphy time series (Figure 18) and exceeds the specified recall and specificity thresholds using the GNB algorithm and 12 features extracted with the MRMR feature selection method. The SHAP visualizations show in that case that the "1day_3_acrophase" feature is the most influential in determining the model's predictions, reinforcing what we have previously established using the STL decomposed time series' trend component. The bar plot (Figure 18A) indicates its substantial impact on the model's output. The decision plot reveals that this feature guides NC samples towards the negative correct class as high

“3_acrophase” values are associated with the negative class, in contrast to other features that generally only move MCI samples towards the positive class as the intensity of the associations of these features with the negative class is very low. The Amplitude and then the Mesor features follow the “3_acrophase” feature in terms of importance. The beeswarm plot (Figure 18B) also suggests that the other three Acrophase features, resulting from the multi-cosinor analysis of the time series, do not have a consistent association with the direction of the model's predictions.

A close inspection of the decision plots reveals that a lot of NC samples start moving towards the wrong positive MCI class end, but finally the top 2 features, “3_acrophase” and “3_amplitude”, guide them to the correct negative class end, something which agrees with the observation that only the “3_acrophase” values have a strong association with the negative class. Overall, the visualizations consistently highlight the critical role of “3_acrophase” in explaining the model's predictions, making it the most important feature of the average 1-day actigraphy dataset.

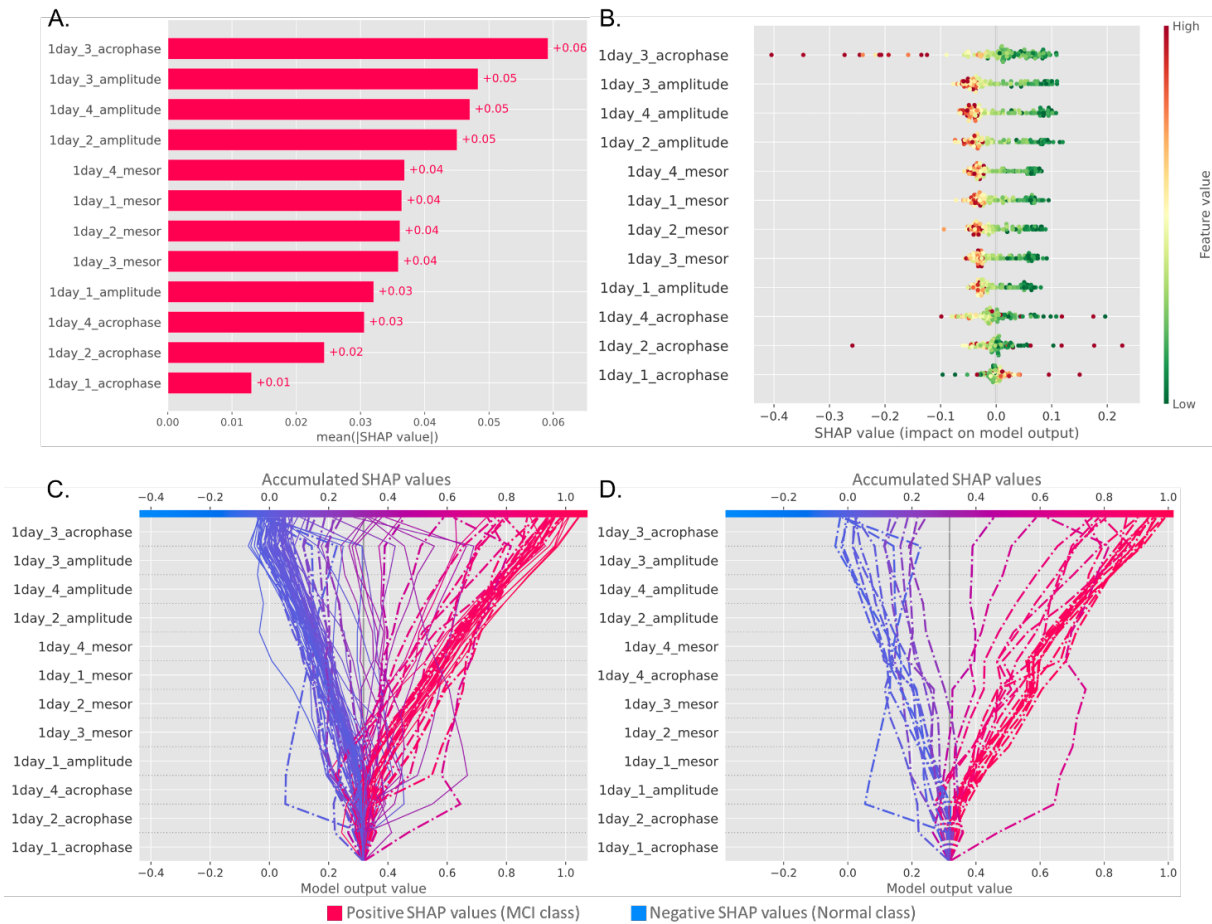


Figure 18: SHAP-based analysis of 1-day actigraphy dataset. Features bar plot (A), beeswarm plot (B), decision plot of all samples (C) and decision plot of only the misclassified samples (D). Experiment details: Model: GNB, Feature Selection: MRMR, Number of features: 12, Time Series normalized: FALSE, Features Scaled: FALSE, Hyperparameter Selection method: GSO2, Resampling period: 1min, Dataset: 1-day average actigraphy.

As already described in section 2.2.1.3, the highest distance between mean actigraphies of the two classes in the ALVION dataset is exhibited in the second 8-hour time window of the day, which we call “Frame 2” (8:00-16:00). We performed SHAPS-based analysis using the final model for Frame 2 in Table 6 that uses LDA and only 5 features selected using the SFM method while exhibiting recall and specificity levels above the set thresholds for this study. SHAPS-based analysis of this dataset (see Figure 19) reveals the Mesor as the most important feature, with low Mesor values associated strongly with positive class (MCI) model predictions. The only non-Mesor selected feature is “1_acrophase” in which low (high) values are associated with the positive (negative) class. The decision plots reinforce the observation, as the “Mesor” features mainly explain the model’s behavior, while the “frame2_1_acrophase” appears to have a relatively lesser influence.

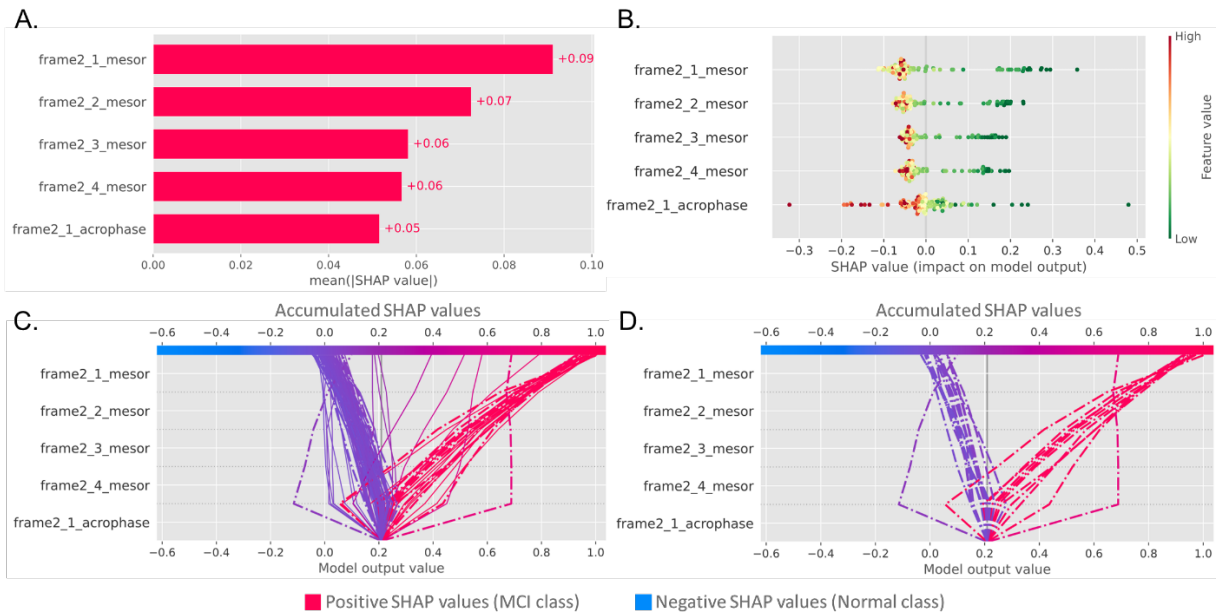


Figure 19: SHAP analysis of Frame 2 actigraphy dataset. Feature bar plots (A), beeswarm plot (B), decision plot of all samples (C), and decision plots of only the misclassified samples (D). Experiment details: Classifier: LDA, Feature Selection method: MRMR, Number of features: 5, Time Series normalized: FALSE, Features Scaled: TRUE, Hyperparameter Selection method: GSO1, Resampling period: 1min, Actigraphy Dataset: Frame 2.

Considering the important features in the experiments above, we observe that the number of cosine curves fitted in the multi-cosinor modeling of the 1-day actigraphy time series does not play an important role, as the true Mesors and Amplitude values can be approximated correctly even if the fitted curve does not superimpose the optimal number of cosine components.

The differences between the SFM and the MRMR-based experiments are minimal. SFM-based experiments can match the classification performance of MRMR-based experiments which use more features, as the SFM method can discard features that are not helpful for the model. Moreover, the user does not to specify the number of top features to be selected (as for the MRMR algorithm). The biggest advantage of SFM is its computational speed compared to MRMR. However, while MRMR is a general model-agnostic method, SFM is a model-specific method and is not applicable to all classifiers. Due to the slower

computational time of MRMR, we are only able to use MRMR in the outer loop, and thus have some data leakage on the hyperparameter tuning process, something that is avoided using the SFM algorithm. As mentioned earlier, the classifiers used in this study that can be applied with SFM are Logistic Regression [58], Linear Discriminant Analysis [59], Random Forest [60], and XGB [61]. By incorporating feature selection within the inner cross-validation loop of nCV, SFM effectively mitigates the risk of data leakage during the selected model training and evaluation process in the outer loop of nCV.

3.3 Feature Kolmogorov-Smirnov tests

We employed the KS two-sample test to compare the ECDFs of the features of our dataset between the ground truth classes for the 1-day and the Frame 2 dataset-based experiments. The results of the test are presented in Table 7 where the rows are sorted based on average absolute SHAP importance values.

Table 7: Kolmogorov-Smirnov two-sample test results. The table displays the p-values from the Kolmogorov-Smirnov (KS) test, comparing the ECDFs of important features between the two classes of samples. The features are sorted in descending order of absolute SHAP importance values from the evaluation experiments. A low p-value (below 0.05) suggests a significant difference between the two classes. We present the KS analysis results comparing the two classes for the 1-day (left) and Frame 2 (right) based experiments.

1-day based experiment			Frame 2 based experiment		
Feature	KS test p-value	Mean absolute SHAP values	Feature	KS test p-value	Mean absolute SHAP values
3_acrophase	0.1069	0.06	1_mesor	0.0019	0.009
3_amplitude	0.0004	0.05	2_mesor	0.0110	0.007
4_amplitude	0.0001	0.05	3_mesor	0.0110	0.006
2_amplitude	0.0002	0.05	4_mesor	0.0039	0.006
4_mesor	0.0068	0.05	1_acrophase	0.2196	0.005
1_mesor	0.0001	0.04			
2_mesor	0.0009	0.04			
3_mesor	0.0021	0.04			
1_amplitude	0.0053	0.03			
4_acrophase	0.3266	0.02			
2_acrophase	0.3367	0.02			
1_acrophase	0.9734	0.01			

The null hypothesis of the KS two-sample test is that the two ECDFs are identical ($F(x) = G(x)$ for all x), with the alternative hypothesis being that they are not identical. A p-value below 0.05 indicates that we reject the null hypothesis, suggesting a significant difference between the ECDFs of the two samples. The p-values for Acrophase features across both classes of samples (subjects) are consistently high, indicating no significant differences between the two classes for this family of features. However, it is worth noting that for “3_acrophase”, the only acrophase feature SHAP analysis indicates as important, its p-value gets closer to the threshold (0.05) compared to the rest of the acrophase features. In

contrast, Amplitude and Mesor features exhibit very small p-values that indicate important ECDF differences between the MCI and Normal classes for these features.

3.4 Spatial allocation of samples

Figure 20 depicts the three-dimensionally reduced selected datasets after using UMAP. This allows us to visualize the spatial distribution of the samples in two dimensions and gives us insights into possible difficulties and limitations regarding our feature set. As discussed, each sample is found in the test set 10 times during the final model's evaluation. We consider a sample as misclassified when the final model made a wrong prediction for its label more than 50% of those times. Misclassified samples are represented by an "X" mark and correctly classified samples are denoted by a square. The blue or red color is used to denote a Normal or MCI sample respectively based on the ground truth. In Figure 20, we observe that most misclassified MCI samples (FN) are in areas dense with Normal samples (TN), and the opposite is also true but not to the same extent. We also notice that there are distant MCI samples scattered in clusters with mostly Normal samples. These MCI samples are almost always misclassified (FN), reducing the model's recall performance and pointing to possible limitations of the parametric feature set in correctly detecting them.

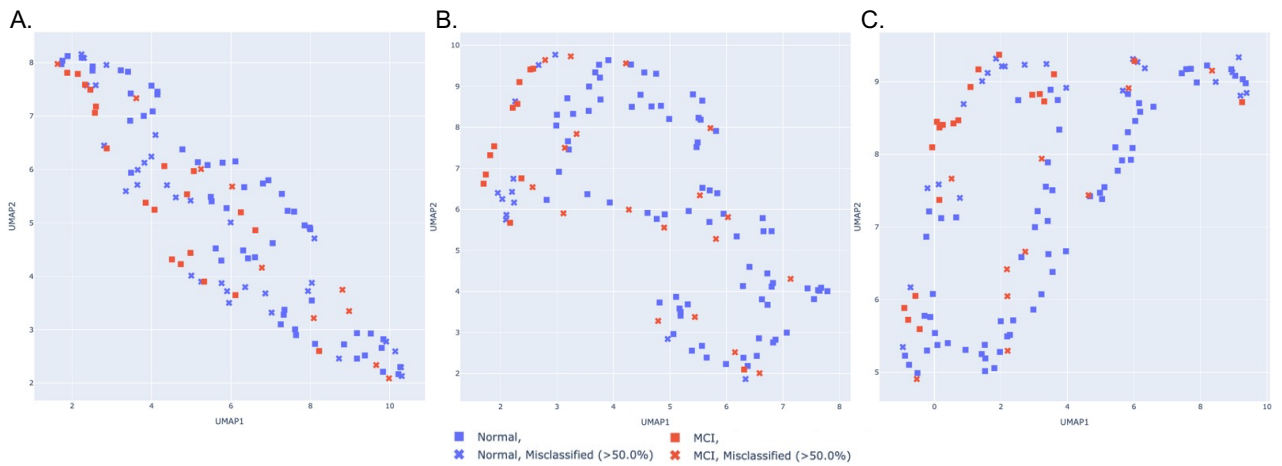


Figure 20: Dimensionality reduction using UMAP of the selected features for datasets 1 (A), 2 (B), and 4 (C) from Table 5. MCI samples are shown in red color while the Normal Control samples in blue color. Samples that are mostly correctly classified or misclassified are represented with a filled square or X shape respectively.

4. CONCLUSIONS

People diagnosed with MCI are characterized by cognitive difficulties that are more profound than in normal aging, however, these individuals are still able to live and act independently. As such, their activity levels are not clearly distinguished from the normal samples, while those diagnosed with AD present a significant decline in their motor functions. We attempted to classify MCI and Normal samples by developing interpretable machine-learning methods that utilize parametric features extracted from a well-established multi-cosinor analysis of actigraphy data sets trying to predict the “Diagnosis” endpoint assigned by clinicians at the end of a patient's visit. Even though the two classes overlap considerably, using simple parametric features we show that we can separate them, by distinguishing the most predominant MCI cases while at the same time minimizing the false positive calls, which was an important consideration of our analysis.

We were able to perform a plethora of experiments, testing the generalization abilities of various classifiers with diverse characteristics, transformed actigraphy datasets, and preprocessing techniques. This approach allowed us to fully explore the discriminative capabilities of the well-established parametric features family. Our findings revealed that lower values of Mesor, Amplitude, and Acrophase are associated with the MCI samples in the ALVION dataset. However, only higher levels of Acrophase exhibited a strong association with the Normal class in our classification models. We also identified a collection of models that harness the full potential of the parametric features derived from the multi-cosinor analysis to deliver an expected specificity reaching 0.88 on unseen data with recall (sensitivity) at 0.5 for the “Diagnosis” endpoint.

In the literature, the parametric set of features is predominantly used in studies that employ actigraphy time series to study neurodegenerative diseases mainly MCI, dementia, and transition to AD [76]. Specifically, in the study by Roeger-Soeder et al. [19], the authors found that lower Mesors, Amplitude, and Acrophase were associated with a higher likelihood of cognitive decline, in a cohort of older men living in the community. Another study by Tranah et al. [20], in a cohort of community-dwelling women, found that lower Amplitude is associated with an increased risk of cognitive decline as well, however, higher Acrophase was associated with the same risk as opposed to the previous study. Posner et al. [25], utilizing the same cohort as Tranah et al., investigated non-parametric features and discovered that both lower values of Acrophase (contradicting Tranah's findings on the same cohort) and higher values for L5 midpoint (timing of low activity) were associated with an increased risk of developing dementia, although the results did not reach statistical significance. Finally, Li et al. [21] associated lower Amplitude with the development of MCI and AD. They observed that during normal aging circadian Amplitude, Acrophase, and interdaily stability progressively decreased over time, while intradaily variability progressively increased. The rate of progression of these effects doubles during MCI and further doubles once more after the diagnosis of AD. Specifically for Acrophase, the transition from normal state to MCI did not accelerate this advance, whereas the transition from MCI to AD did.

The findings of our investigative research align with the core assertions put forth in the aforementioned literature. Specifically, our results corroborate those studies that suggest

lower values for Mesors, Amplitude, or Acrophase—or a combination of them—are indicative characteristics of an actigraphy sample comprising individuals with MCI. However, it is worth noting that one particular study deviates from this consensus, claiming that higher Acrophase values are more characteristic of women with MCI. However, another study involving the same cohort contradicts this viewpoint, lending support to the opposite stance. It is crucial to emphasize that the scope of our research did not extend to investigating the progression of parametric metrics in patients transitioning from a cognitively normal state to MCI and, eventually, AD.

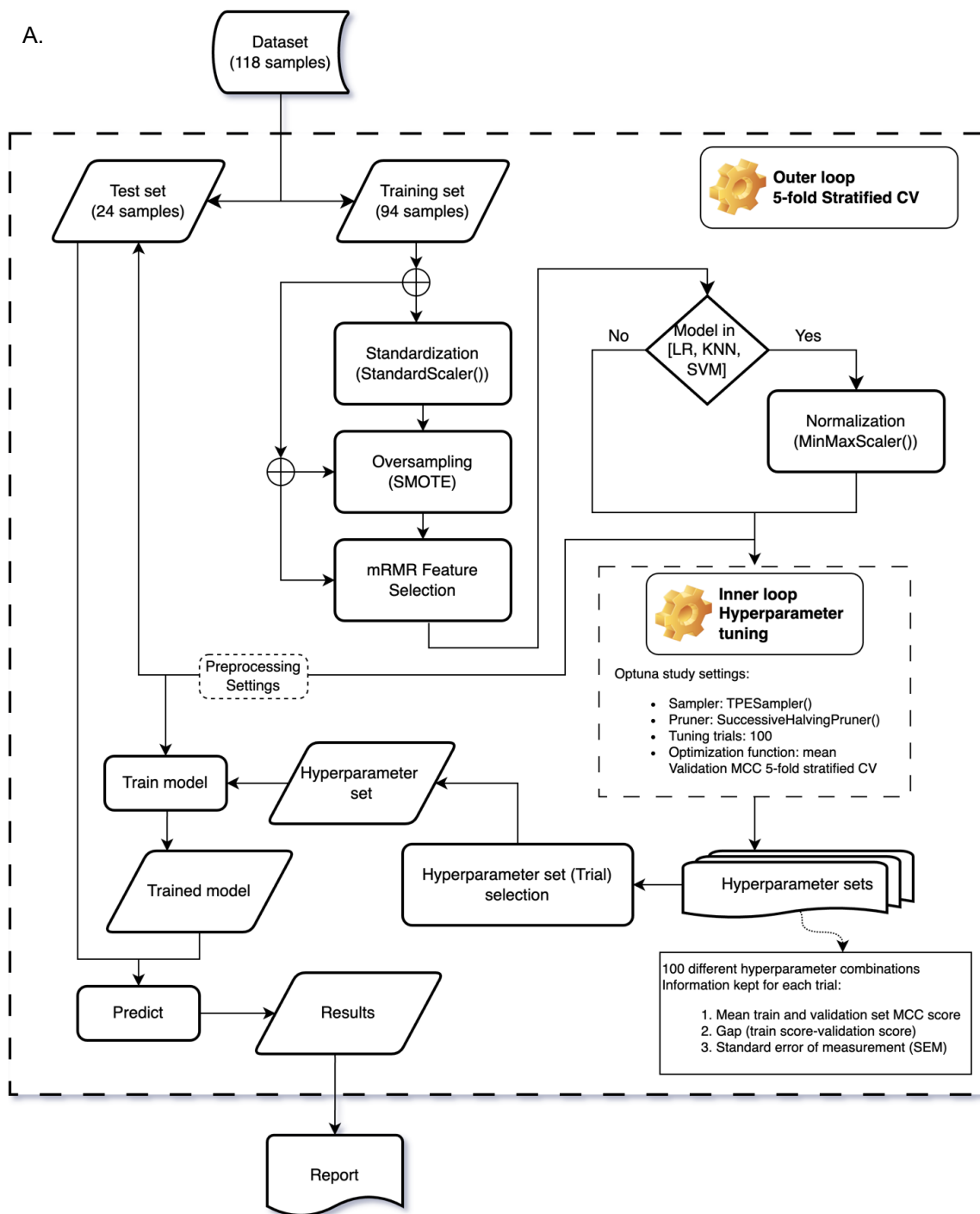
As the ALBION study is an ongoing project, the future incorporation of new actigraphy samples will have a great impact on the development of more robust machine-learning models. A bigger dataset is expected to provide a larger diversity, a better representation of the population, and can improve the learning of complex patterns. This will also allow us to explore more complex feature sets, like those derived from a Wavelet analysis, or even experiment with Deep Neural Networks which can create more complex features that may be able to separate our actigraphy dataset more efficiently than the simple, yet interpretable, parametric features. Finally, it could become possible to associate actigraphy-extracted features with objectively measured molecular targets, such as the CSF biomarkers, which were not available at the time of this analysis. In any case, there will always be a need to strike a balance between generalization ability and interpretability of developed “open box” AI models especially when it comes to clinical decision support or deciphering the underlying complex biological mechanisms. Therefore, this research provides useful insights on how to develop predictive modeling approaches that can achieve such a balance, how to properly assess their expected generalization performance on unseen data, and how to make sense of their predictions even down to the level of each individual sample.

ABBREVIATIONS – ACRONYMS

ALBION	Aiginition Longitudinal Biomarker Investigation Of Neurodegenerations
MCI	Mild Cognitive Impairment
AD	Alzheimer's disease
CSF	Cerebrospinal fluid
PET	Positron Emission Tomography
<i>APP</i>	Amyloid-beta Precursor Protein
<i>PSEN1</i>	Presenilin 1
<i>PSEN2</i>	Presenilin 2
A β	Amyloid beta
<i>APOE</i>	Apolipoprotein E
CSF	CerebroSpinal Fluid
pTau	Phosphorylated tau species
tTau	tau proteins
NINCDS/ADR DA	National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer Disease and Related Disorders Association
DSM-IV-TR	Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision
NC	Normal controls
STL	Seasonal and Trend decomposition using Loess
ME	Marginal Error
MRMR	Maximum Relevance Minimum Redundancy
SFM	Select From Model
LR	Logistic Regression
SVM	Support Vector Machine
KNN	k-Nearest Neighbor
SMOTE	Synthetic Minority Over-sampling
LDA	Linear Discriminant Analysis
GNB	Gaussian Naive Bayes
RF	Random Forest
XGB	XGBoost

MCC	Matthew's Correlation Coefficient
nCV	Nested Cross Validation
CV	Cross Validation
SEM	Standard Error of the Mean
GSO1	Gap Score optimization 1
GSO2	Gap Score optimization 2
GSO3	Gap Score optimization 3
PCA	Principal Component Analysis
UMAP	Uniform Manifold Approximation and Projection
LP	Lumbar puncture
SHAP	SHapley Additive exPlanations

ANNEX I



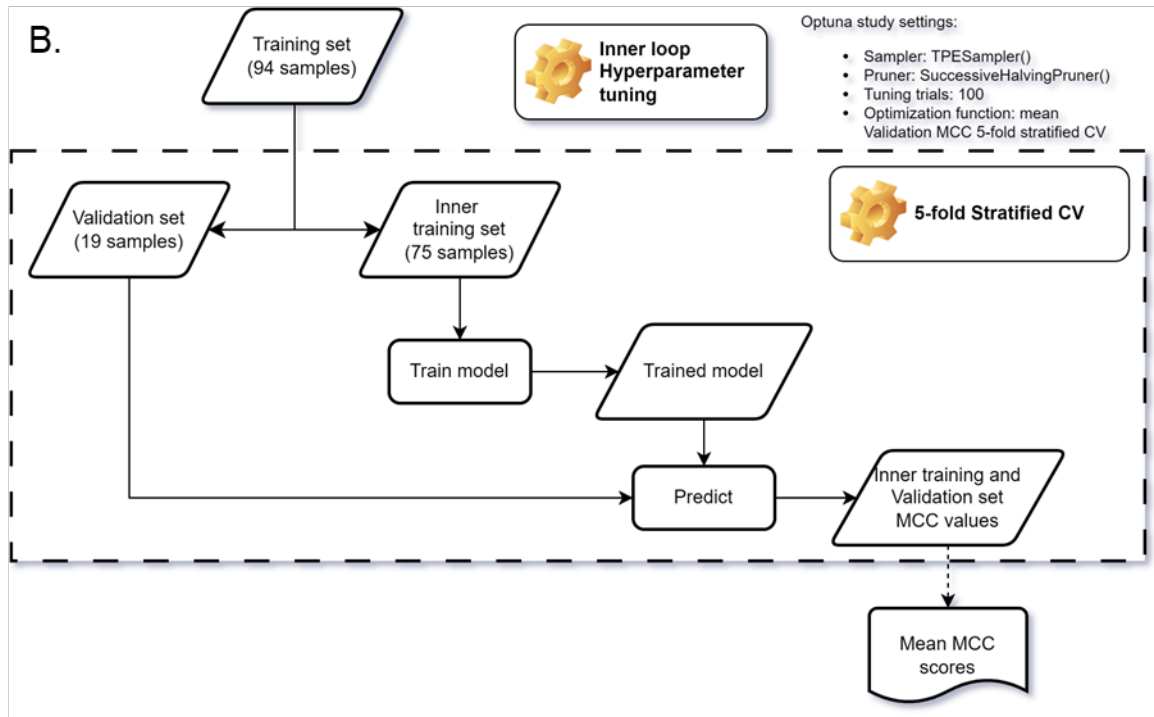
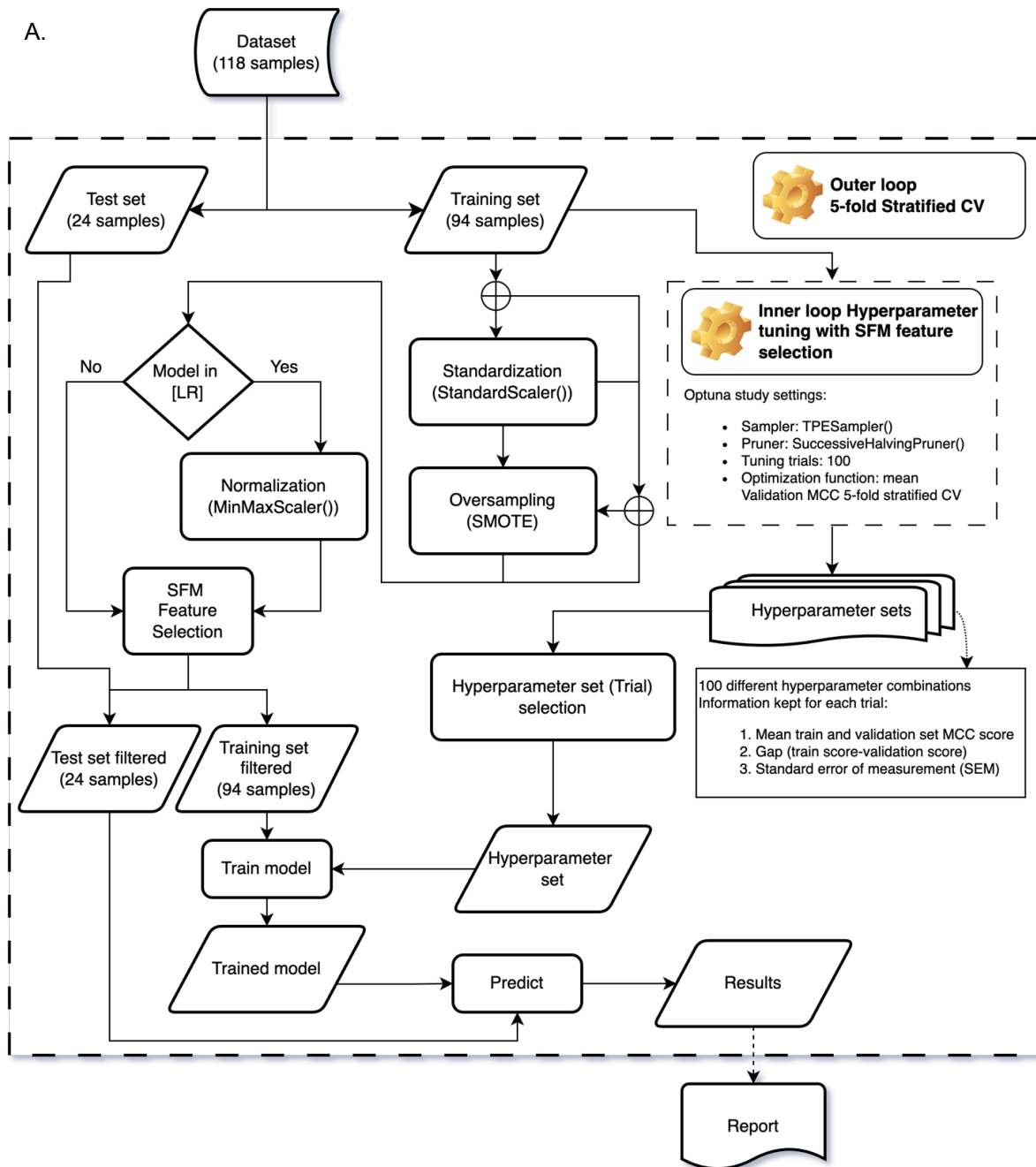


Figure A1: Nested cross-validation (nCV) pipeline with MRMR feature selection in the outer loop. The nCV consists of an outer and inner stratified 5-fold CV. The full dataset will first be split into the test and the training set (A). The latter will be used to perform the preprocessing steps, which will be implemented later on the test set using the same settings. Using MRMR we perform feature selection in the outer loop. The training set will be used for the hyperparameter tuning in the inner loop (B) using Optuna. Specifically, in the inner CV (5) the training set is split into the inner training and validation sets in each tuning trial (100 in total). This split is performed 100 times, and the training sets are used to train the new training set will be processed and will be used for feature selection using SFM. In each tuning trial we perform a 5-fold stratified CV, and we acquire a mean MCC score for each CV. Therefore, we produce 100 mean MCC scores, and the hyperparameter set will be selected using 5 different approaches. The final model of the outer split will be trained using the selected hyperparameters. The trained model will make predictions using the test set and the results of the outer loop will be stored.

A.



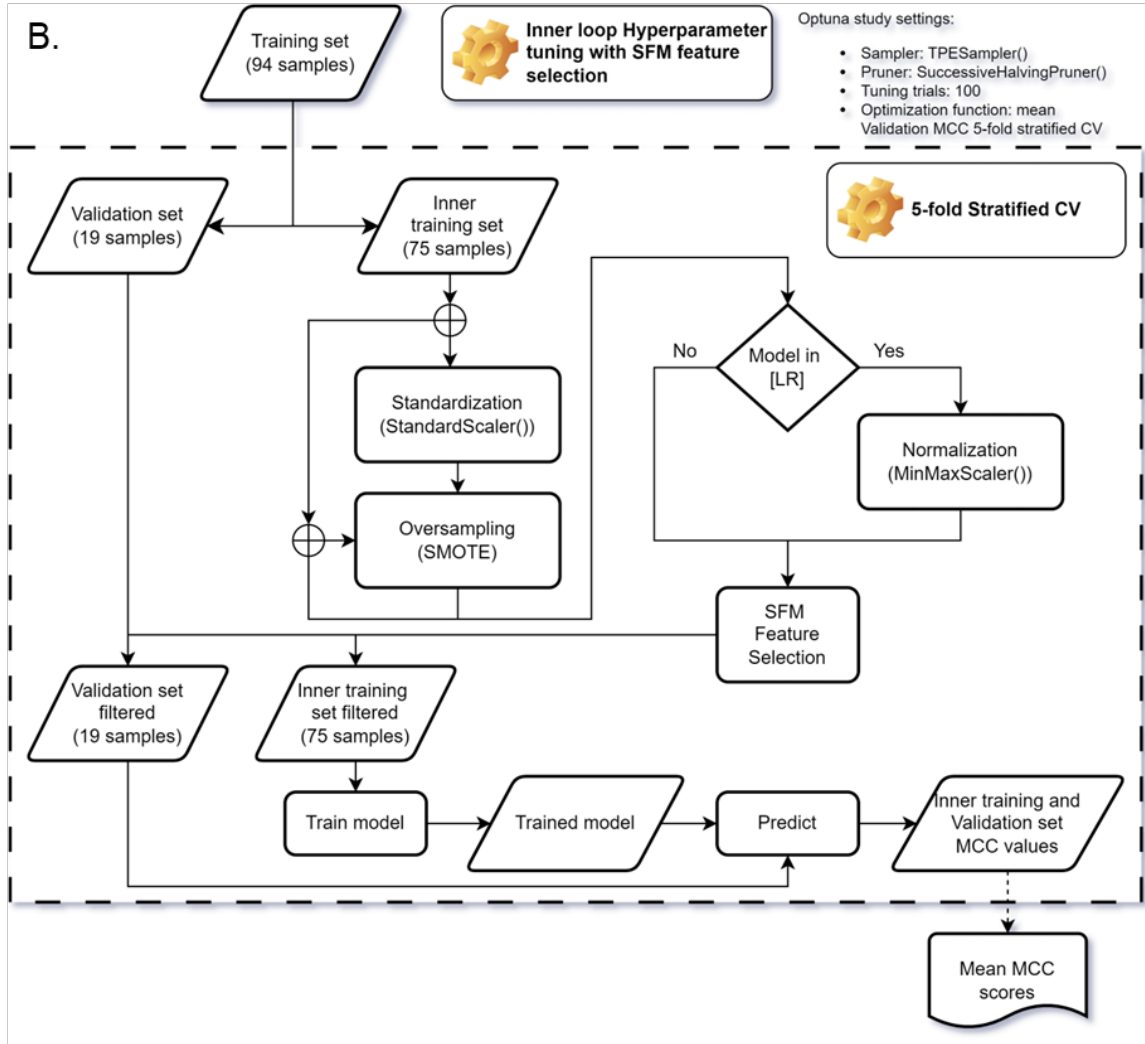


Figure A2: Nested cross-validation (nCV) pipeline with SFM feature selection in the inner loop. The nCV is consisted of an outer and inner stratified 5-fold CV. The full dataset will first be split into the test and the training set (A). The latter will be used to perform the preprocessing steps which will be implemented on the test set using the same setting. The unprocessed training set will be used for the hyperparameter tuning in the inner loop (B) using Optuna. The outer loop training set will be split into an inner training and validation set in each tuning trial (100 in total). The new training set will be processed and will be used for feature selection using SFM. In each tuning trial we perform a 5-fold stratified CV and we acquire a mean MCC score for each CV. Therefore, we produce 100 mean MCC scores and the hyperparameter set will be selected using 5 different approaches. The final model of the outer split will be trained using the selected hyperparameters. The trained model will make predictions using the test set and the results of the outer loop will be stored.

REFERENCES

- [1] World Health Organization, *Global status report on the public health response to dementia*. 2021.
- [2] “2021 Alzheimer’s disease facts and figures,” *Alzheimers. Dement.*, vol. 17, no. 3, pp. 327–406, Mar. 2021.
- [3] K. B. Rajan, J. Weuve, L. L. Barnes, E. A. McAninch, R. S. Wilson, and D. A. Evans, “Population estimate of people with clinical Alzheimer’s disease and mild cognitive impairment in the United States (2020-2060),” *Alzheimers. Dement.*, vol. 17, no. 12, pp. 1966–1975, Dec. 2021.
- [4] K. A. Matthews *et al.*, “Racial and ethnic estimates of Alzheimer’s disease and related dementias in the United States (2015-2060) in adults aged ≥ 65 years,” *Alzheimers. Dement.*, vol. 15, no. 1, pp. 17–24, Jan. 2019.
- [5] Alzheimer Europe, *Prevalence of dementia in Europe*. 2019.
- [6] R. M. Torack, “The early history of senile dementia,” *Alzheimer’s disease*. Free Press, New York, 1983.
- [7] E. Esquirol, “Des maladies mentales considérées sous les rapports médical, hygiénique et médico-legal.” J. B. Baillière, Paris, 1838.
- [8] H. D. Yang, D. H. Kim, S. B. Lee, and L. D. Young, “History of Alzheimer’s Disease,” *Dement Neurocogn Disord*, vol. 15, no. 4, pp. 115–121, Dec. 2016.
- [9] R. C. Petersen, “How early can we diagnose Alzheimer disease (and is it sufficient)? The 2017 Wartenberg lecture,” *Neurology*, vol. 91, no. 9, pp. 395–402, Aug. 2018.
- [10] D. S. Knopman *et al.*, “Alzheimer disease,” *Nat Rev Dis Primers*, vol. 7, no. 1, p. 33, May 2021.
- [11] Y. Namba, M. Tomonaga, H. Kawasaki, E. Otomo, and K. Ikeda, “Apolipoprotein E immunoreactivity in cerebral amyloid deposits and neurofibrillary tangles in Alzheimer’s disease and kuru plaque amyloid in Creutzfeldt-Jakob disease,” *Brain Res.*, vol. 541, no. 1, pp. 163–166, Feb. 1991.
- [12] A.-C. Raulin, S. V. Doss, Z. A. Trottier, T. C. Ikezu, G. Bu, and C.-C. Liu, “ApoE in Alzheimer’s disease: pathophysiology and therapeutic strategies,” *Mol. Neurodegener.*, vol. 17, no. 1, p. 72, Nov. 2022.
- [13] J. Therriault *et al.*, “Association of Apolipoprotein E $\epsilon 4$ With Medial Temporal Tau Independent of Amyloid- β ,” *JAMA Neurol.*, vol. 77, no. 4, pp. 470–479, Apr. 2020.
- [14] A. R. Wadhvani, A. Affaneh, S. Van Gulden, and J. A. Kessler, “Neuronal apolipoprotein E4 increases cell death and phosphorylated tau release in alzheimer disease,” *Ann. Neurol.*, vol. 85, no. 5, pp. 726–739, May 2019.
- [15] A. Anoop, P. K. Singh, R. S. Jacob, and S. K. Maji, “CSF Biomarkers for Alzheimer’s Disease Diagnosis,” *Int. J. Alzheimers. Dis.*, vol. 2010, Jun. 2010.
- [16] D. R. Williams, “Tauopathies: classification and clinical update on neurodegenerative diseases associated with microtubule-associated protein tau,” *Intern. Med. J.*, vol. 36, no. 10, pp. 652–660, Oct. 2006.
- [17] N. Scarmeas *et al.*, “Initial Data and a Clinical Diagnosis Transition for the Aiginition Longitudinal Biomarker Investigation of Neurodegeneration (ALBION) Study,” *Medicina*, vol. 58, no. 9, p. 1179, Aug. 2022.
- [18] P. B. Rosenberg and C. Lyketsos, “Mild cognitive impairment: searching for the prodrome of Alzheimer’s disease,” *World Psychiatry*, vol. 7, no. 2, pp. 72–78, 2008.
- [19] T. S. Rogers-Soeder *et al.*, “Rest-Activity Rhythms and Cognitive Decline in Older Men: The Osteoporotic Fractures in Men Sleep Study,” *J. Am. Geriatr. Soc.*, vol. 66, no. 11, pp. 2136–2143, Nov. 2018.
- [20] G. J. Tranah *et al.*, “Circadian activity rhythms and risk of incident dementia and mild cognitive impairment in older women,” *Ann. Neurol.*, vol. 70, no. 5, pp. 722–732, Nov. 2011.
- [21] P. Li *et al.*, “Circadian disturbances in Alzheimer’s disease progression: a prospective observational cohort study of community-based older adults,” *Lancet Healthy Longev*, vol. 1, no. 3, pp. e96–e105, Dec. 2020.
- [22] N. Milman *et al.*, “0271 Actigraphy Based Measures of Sleep Disruption and Circadian Rhythms in the WashU Biomarkers of Alzheimer’s Disease in Sleep and EEG (BASE) Cohort,” *Sleep*, vol. 45, no. Supplement_1, pp. A122–A122, May 2022.
- [23] S. J. Diem *et al.*, “Measures of Sleep-Wake Patterns and Risk of Mild Cognitive Impairment or Dementia in Older Women,” *Am. J. Geriatr. Psychiatry*, vol. 24, no. 3, pp. 248–258, Mar. 2016.
- [24] M. Basta *et al.*, “Associations between sleep duration and cognitive impairment in mild cognitive impairment,” *J. Sleep Res.*, vol. 28, no. 6, p. e12864, Dec. 2019.

- [25] A. B. Posner *et al.*, “Predicting incident dementia and mild cognitive impairment in older women with nonparametric analysis of circadian activity rhythms in the Study of Osteoporotic Fractures,” *Sleep*, vol. 44, no. 10, Oct. 2021.
- [26] A. Alfini *et al.*, “Associations of actigraphic sleep and circadian rest/activity rhythms with cognition in the early phase of Alzheimer’s disease,” *Sleep Adv.*, vol. 2, no. 1, p. zpab007, Apr. 2021.
- [27] S. Cai *et al.*, “Characteristics of Sleep Structure Assessed by Objective Measurements in Patients With Amnesic Mild Cognitive Impairment: A Meta-Analysis,” *Front. Neurol.*, vol. 11, p. 577126, Nov. 2020.
- [28] J. R. Winer *et al.*, “Tau and β -Amyloid Burden Predict Actigraphy-Measured and Self-Reported Impairment and Misperception of Human Sleep,” *J. Neurosci.*, vol. 41, no. 36, pp. 7687–7696, Sep. 2021.
- [29] M. Lauriola *et al.*, “Sleep changes without medial temporal lobe or brain cortical changes in community-dwelling individuals with subjective cognitive decline,” *Alzheimers. Dement.*, vol. 13, no. 7, pp. 783–791, Jul. 2017.
- [30] S. Scarlett, R. A. Kenny, M. D. O’Connell, H. Nolan, and C. de Looze, “Associations between cognitive function, actigraphy-based and self-reported sleep in older community-dwelling adults: Findings from the Irish Longitudinal Study on Ageing,” *Int. J. Geriatr. Psychiatry*, vol. 36, no. 5, pp. 731–742, May 2021.
- [31] T. S. Lysen, A. I. Luik, M. K. Ikram, H. Tiemeier, and M. A. Ikram, “Actigraphy-estimated sleep and 24-hour activity rhythms and the risk of dementia,” *Alzheimers. Dement.*, vol. 16, no. 9, pp. 1259–1267, Sep. 2020.
- [32] G. A. Dowling *et al.*, “Melatonin and bright-light treatment for rest-activity disruption in institutionalized patients with Alzheimer’s disease,” *J. Am. Geriatr. Soc.*, vol. 56, no. 2, pp. 239–246, Feb. 2008.
- [33] A. D. S. Targa *et al.*, “The circadian rest-activity pattern predicts cognitive decline among mild-moderate Alzheimer’s disease patients,” *Alzheimers. Res. Ther.*, vol. 13, no. 1, pp. 1–10, Sep. 2021.
- [34] S. Bringas, S. Salomón, R. Duque, J. L. Montaña, and C. Lage, “A Convolutional Neural Network-Based Method for Human Movement Patterns Classification in Alzheimer’s Disease,” *Proc. AMIA Annu. Fall Symp.*, vol. 31, no. 1, p. 72, Nov. 2019.
- [35] H. J. M. van Alphen, K. M. Volkers, C. G. Blankevoort, E. J. A. Scherder, T. Hortobágyi, and M. J. G. van Heuvelen, “Older Adults with Dementia Are Sedentary for Most of the Day,” *PLoS One*, vol. 11, no. 3, p. e0152457, Mar. 2016.
- [36] K. N. Anderson, C. Hatfield, C. Kipps, M. Hastings, and J. R. Hodges, “Disrupted sleep and circadian patterns in frontotemporal dementia,” *Eur. J. Neurol.*, vol. 16, no. 3, pp. 317–323, Mar. 2009.
- [37] F. H. Chan, B. M. Wu, F. K. Lam, P. W. Poon, and A. M. Poon, “Multiscale characterization of chronobiological signals based on the discrete wavelet transform,” *IEEE Trans. Biomed. Eng.*, vol. 47, no. 1, pp. 88–95, Jan. 2000.
- [38] T. L. Hayes, F. Abendroth, A. Adami, M. Pavel, T. A. Zitzelberger, and J. A. Kaye, “Unobtrusive assessment of activity patterns associated with mild cognitive impairment,” *Alzheimers. Dement.*, vol. 4, no. 6, pp. 395–405, Nov. 2008.
- [39] K. Hu, E. J. W. Van Someren, S. A. Shea, and F. A. J. L. Scheer, “Reduction of scale invariance of activity fluctuations with aging and Alzheimer’s disease: Involvement of the circadian pacemaker,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 8, pp. 2490–2494, Feb. 2009.
- [40] K. Hu *et al.*, “Progression of Dementia Assessed by Temporal Correlations of Physical Activity: Results From a 3.5-Year, Longitudinal Randomized Controlled Trial,” *Sci. Rep.*, vol. 6, p. 27742, Jun. 2016.
- [41] P. Li *et al.*, “Fractal regulation and incident Alzheimer’s disease in elderly individuals,” *Alzheimers. Dement.*, vol. 14, no. 9, pp. 1114–1125, Sep. 2018.
- [42] Y.-H. Wang, H.-W. V. Young, and M.-T. Lo, “The inner structure of empirical mode decomposition,” *Physica A: Statistical Mechanics and its Applications*, vol. 462, no. C, pp. 1003–1017, 2016.
- [43] N. E. Huang *et al.*, “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,” *Proc. R. Soc. Lond. A Math. Phys. Sci.*, vol. 454, pp. 903–998, Mar. 1998.
- [44] G. Cornelissen, “Cosinor-based rhythmometry,” *Theor. Biol. Med. Model.*, vol. 11, p. 16, Apr. 2014.
- [45] Philips-Respironics (<https://www.philips.com.au/healthcare/product/HC1044809/actiwatch-2-activity-monitor>), “Actiwatch 2 - Activity Monitor.” 2018.
- [46] D. L. Segal, “Diagnostic and statistical manual of mental disorders (DSM-IV-TR),” *The Corsini Encyclopedia of Psychology*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 30-Jan-2010.

- [47] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "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," *Neurology*, vol. 34, no. 7, pp. 939–944, Jul. 1984.
- [48] V. C. Hachinski *et al.*, "Cerebral blood flow in dementia," *Arch. Neurol.*, vol. 32, no. 9, pp. 632–637, Sep. 1975.
- [49] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. Kokmen, and E. G. Tangelos, "Aging, memory, and mild cognitive impairment," *Int. Psychogeriatr.*, vol. 9 Suppl 1, pp. 65–69, 1997.
- [50] O. Hansson *et al.*, "CSF biomarkers of Alzheimer's disease concord with amyloid- β PET and predict clinical progression: A study of fully automated immunoassays in BioFINDER and ADNI cohorts," *Alzheimers. Dement.*, vol. 14, no. 11, pp. 1470–1481, Nov. 2018.
- [51] T. Bittner *et al.*, "Technical performance of a novel, fully automated electrochemiluminescence immunoassay for the quantitation of β -amyloid (1-42) in human cerebrospinal fluid," *Alzheimers. Dement.*, vol. 12, no. 5, pp. 517–526, May 2016.
- [52] R. B. Cleveland, W. S. Cleveland, and I. Terpenning, "STL: A Seasonal-Trend Decomposition Procedure Based on Loess," vol. 6, no. 1, p. 3, Mar. 1990.
- [53] I. P. Karpouzis, "Signal processing and machine learning pipeline to diagnose early cognitive decline," MSc, National and Kapodistrian University of Athens, 2024.
- [54] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, "Minimum redundancy maximum relevance feature selection approach for temporal gene expression data," *BMC Bioinformatics*, vol. 18, no. 1, p. 9, Jan. 2017.
- [55] Z. Zhao, R. Anand, and M. Wang, "Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform," in *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2019, pp. 442–452.
- [56] M. B. Kursu and W. R. Rudnicki, "Feature Selection with the Boruta Package," *J. Stat. Softw.*, vol. 36, pp. 1–13, Sep. 2010.
- [57] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *arXiv [cs.LG]*, 02-Jan-2012.
- [58] D. R. Cox, "The Regression Analysis of Binary Sequences," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 20, no. 2, pp. 215–242, 1958.
- [59] A. J. Izenman, "Linear Discriminant Analysis," in *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, A. J. Izenman, Ed. New York, NY: Springer New York, 2008, pp. 237–280.
- [60] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [61] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *arXiv [cs.LG]*, 09-Mar-2016.
- [62] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
- [63] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, "k-Nearest Neighbor Classification," in *Data Mining in Agriculture*, A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, Eds. New York, NY: Springer New York, 2009, pp. 83–106.
- [64] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, Jun. 2002.
- [65] H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," in *2019 International Engineering Conference (IEC)*, 2019, pp. 165–170.
- [66] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019, pp. 2623–2631.
- [67] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for Hyper-Parameter Optimization," in *Advances in Neural Information Processing Systems*, 2011, vol. 24.
- [68] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Jan. 2020.
- [69] S. Parvande, H.-W. Yeh, M. P. Paulus, and B. A. McKinney, "Consensus features nested cross-validation," *Bioinformatics*, vol. 36, no. 10, pp. 3093–3098, May 2020.
- [70] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS One*, vol. 14, no. 11, p. e0224365, Nov. 2019.
- [71] F. J. Bargagli Stof, G. Cevalani, and G. Gnecco, "Simple Models in Complex Worlds: Occam's Razor and Statistical Learning Theory," *Minds Mach.*, vol. 32, no. 1, pp. 13–42, Mar. 2022.

- [72] I. Jolliffe, "Principal Component Analysis," *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, Chichester, UK, 29-Sep-2014.
- [73] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, Sep. 2018.
- [74] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *arXiv [cs.AI]*, 22-May-2017.
- [75] A. Roth and L. Shapley, *The Shapley value: Essays in honor of Lloyd S. shapley*. Cambridge, England: Cambridge University Press, 2005.
- [76] C. Gao *et al.*, "Approaches for Assessing Circadian Rest-Activity Patterns Using Actigraphy in Cohort and Population-Based Studies," *Current Sleep Medicine Reports*, vol. 9, no. 4, pp. 247–256, Dec. 2023.