An improved understanding of Paradoxical Insomnia : A knowledge-based approach using machine learning tools

Olivier Pallanca

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAX071

Thèse de doctorat

ÉCOLE POLYTECHNIQUE

IP PARIS

An improved understanding of Paradoxical Insomnia: a knowledge-based approach using machine learning tools

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'Ecole polytechnique

École doctorale n°626
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Palaiseau, le 29/08/2023, par

MR OLIVIER PALLANCA

Composition du Jury :

Isabelle ARNULF Professeure des universités - praticienne hospitalière Sorbonne université	Présidente
Céline VENS Professeure associée KU Leuven	Rapporteuse
André ANJOS Directeur d'études Idiap Research Institute	Rapporteur
Vincent NAVARRO Professeur des universités - praticien hospitalier Institut du cerveau et de la moelle	Examinateur
Jesse READ Professeur École polytechnique	Directeur de thèse
Jean-Marc STEYAERT Professeur émérite École polytechnique	Co-directeur de thèse

Acknowledgements

I want to thank Jean-Marc Steyaert for trusting me and allowing me to do this thesis work despite the headwinds. I would also like to thank Michalis Vazirgiannis for accepting me into his team and allowing me to investigate the field of Machine Learning. I thank Jesse for accepting an old MD as his first thesis student. I want to thank Adriana Tapus for getting me through all the administrative and methodological stages. I thank Sammy Khalife, my faithful thesis companion turned friend, for his open-mindedness, great knowledge, and wise advice throughout this Ph.D. I thank Khalida Douibi and Celiane Ayad for their huge work, contribution, and support. I thank Paul Boniol for his insight into Hankel's matrices. I thank Alexandre Gramfort, Agathe De Vulpian, Maylis Solal, and Ying Yu Yang for their work and contributions to EEG signal processing. I want to thank the entire Dascim team, especially Johannes Lutzeyer, for all the interesting conferences that enabled me to discover 1001 data processing methods. I thank Andre Anjos for his careful review of my work.

I want to thank the hospital team, especially the EEG technician, Vincent Navarro, for his support. I thank Ana Gales, and Aurore Besnard for their work and support.

I thank the patients and participants who supported me in this journey, especially Sandrine, Clara, Laure, Mina, and Safa. Their insights, encouragement, and criticism were invaluable. I thank my family for believing in me. Finally, I thank the APHP Foundation, especially Rodolphe Gouin and Sophie Lemaire, for their facilitating role in funding this Ph.D. and a special thank to Marie-Laure Traux Binsse for giving me the resources to carry out this project, thanks to his unfailing support and trust.

Preface

At staring this thesis, I was the archetypal doctor-turned-geek, eager to delve into the trove of data at the Pitié-Salpêtrière Hospital. The journey began with establishing a multimodal data collection unit, but it was soon evident that data collection was only the first step. Aggregating and processing these data presented unforeseen challenges, leading to a realization: data in healthcare requires meticulous preparation before it can be meaningfully analyzed.

Embarking on this doctoral journey at the intersection of medicine and machine learning, I have been continually reminded of the transient nature of knowledge, as eloquently captured by Nikola Tesla: "The history of science shows that theories are perishable." This thesis represents a foray into the dynamic, ever-evolving landscape of artificial intelligence, with a particular focus on its application in healthcare.

As a physician, my initial foray into the world of machine learning was met with the challenge of unlearning and relearning. My medical training had instilled a patient-centric approach, contrasting starkly with the data-driven, algorithmic focus of machine learning. Yet, it was this very difference that fueled my motivation: the transformative potential of AI in healthcare, specifically in sleep medicine, a field burdened by chronic insomnia yet rich in data waiting to be deciphered.

This thesis represents a journey of balancing two worlds. On one hand, the empathetic, nuanced practice of medicine; on the other, the precise, data-centric realm of machine learning. The aim was to harness machine learning to not only predict and personalize treatment for chronic insomnia but also to illuminate the debated clinical concept of Paradoxical Insomnia.

Throughout, I have strived to merge the analytical prowess of AI with the compassionate approach of medicine. This interdisciplinary endeavor has underscored the importance of collaboration in advancing healthcare, requiring a continuous process of learning and adapting.

Under the tutelage of Jesse Read, the journey took a turn towards not just achieving accurate predictions but understanding and evaluating the reliability of these machine learning models. This paradigm shift—from seeking good prediction scores to ensuring the reliability of predictions in a medical context—has been the crux of my doctoral research.

In conclusion, this thesis is more than an academic pursuit; it is a narrative of bridging two distinct yet complementary worlds and a testament to the potential of interdisciplinary collaboration in revolutionizing healthcare.

Contents

Acronyms for Sleep Medicine

AG Actigraphy. 118

BDI Beck Depression Inventory. 27

CBT-I Cognitive and Behavioral Therapy for Insomnia. 11, 23, 146

CID Chronic Insomnia Disorder. 6, 7, 11, 20, 22, 26, 31, 46, 49, 50, 52, 118, 119, 187, 188

DBAS Dysfunctional Beliefs and Attitudes about Sleep Scale. 27

EEG Electroencephalogram. 11–13, 16, 22–24, 29–32, 50, 118

EMG Electromyogram. 16, 17

ESS Epworth Sleepiness Scale. 22, 23, 26, 50, 81

ICD International statistical classification of Diseases. 19, 187, 188

ICSD International Classification of Sleep Disorders. 11, 19–21, 23, 28, 187, 188

ISI Insomnia Severity Index. 11, 22, 23, 26, 27, 50, 81

MMPI Minnesota Multiphasic Personality Inventory. 11, 49, 50, 52, 54, 62, 76–78, 81, 118

MSLT Multiple Sleep Latency Test. 21

N-REM Non rapid eye movement. 12, 15, 16, 18, 19, 29–31

nSSM negative Sleep State Misperception. 11, 12, 21, 28, 31, 32, 187

OSA Obstructive Sleep Apnea. 49

ParI Paradoxical insomnia. 6–8, 11, 19–21, 23, 27–32, 39, 40, 46, 49, 118, 119, 187

PLM Periodic Limb Movements. 49

PSG Polysomnography. 24, 28, 31, 40, 187

PsyI Psychophysiological Insomnia. 11, 29, 31, 118, 187

REM Rapid Eye Movement. 11, 12, 15–19, 24, 30, 49

RLS Restless Legs Syndrome. 49

SE Sleep Efficiency. 11, 30, 128

SOL Sleep Onset Latency. 11, 22, 24, 30, 32, 118, 128

STAI State-Trait Anxiety Inventory. 27

TST Total Sleep Time. 11, 22, 24, 28–32

WASO Wake After Sleep Onset. 11, 22, 24, 30, 31

y.o. years old. 18, 49

Acronyms for Machine Learning

AdaBoost Adaptive Boosting. 146

CA Classification Accuracy. 146

CV Cross Validation. 118

DT Decision Trees. 33, 43–45

EDA Exploratory Data Analysis. 33, 41, 53

ETMPE Ensemble Type Model for Prediction Explanation. 156

FI Feature Importances. 118

IMF Intrinsic Mode Functions. 81

KMEANS K-Means Clustering. 33, 41, 42

KNN K-Nearest Neighbors. 33, 39, 43, 45

LASSO Least Absolute Shrinkage and Selection Operator. 23, 33, 38, 39, 46, 118

LR Logistic Regression. 33, 43, 44

ML Machine Learning. 6, 7, 31, 32, 34, 35, 37, 40, 44, 46, 47, 159

MLP Multilayer Perceptron. 146

MoSA Morris Sensitivity Analysis. 118

OLS Ordinary Least Squares. 33, 38

PCA Principal Component Analysis. 33, 42

RF Random Forest. 33, 39, 43–45

ROC Receiver Operating Characteristic. 146

SHAP SHapley Additive exPlanations. 46, 118

SVM Support Vector Machine. 33, 39, 43, 45

t-SNE t-Distributed Stochastic Neighbor Embedding. 33, 42

XAI Explainable AI. 118

XGB Extreme Gradient Boosting. 43, 146

Chapter 1

Introduction

This dissertation is at an intersection of sleep medicine driven by expert knowledge and methods driven by data (Machine Learning). The modern understanding of sleep science has several open questions; we launch this thesis with the intent to provide answers to some of these questions. It is necessary to overcome the inherent cross-disciplinary difficulties (different views, traditions, and terminology used in the different fields) to do so. The general hypothesis is that we can produce new knowledge from "old" data by harnessing Machine Learning and then pass this knowledge back to the medical community.

1.1 Historical Perspective: Medical Practice and the Rise of Data

Since the beginning of human civilization, medical knowledge has primarily relied on clinical and anatomical observations. Medicine was initially seen as an art passed down through mentorship. However, advancements in tools and methodologies, from early mechanisms for observing living organisms and blood cells in the 17th century to the digitization of medical data in the 21st century, have significantly influenced medical practices and research.

A major medical revolution occurred in the late 19th century with the introduction of descriptions of pathogenic bacteria and the emergence of medical imaging departments and electrocardiograms. In the 20th century, we witnessed numerous medical breakthroughs, including the invention of the Electroencephalogram (EEG) in 1929, the development of X-ray scanners and high-resolution ultrasounds in the 1970s, and the application of Magnetic Resonance Imaging (MRI) to humans in 1977. With the exponential growth in computing power, the digital transformation of medical examinations has accelerated, particularly in the 21st century.

The dynamics and evolution of medical knowledge and publications have reached unprecedented levels. In 1950, it took 50 years to double the number of medical publications; today, it takes only a few months. However, as medical knowledge has expanded, integrating and assimilating this vast amount of data has become increasingly challenging for the human brain. Compounding this problem is the cumulative decline in the number of doctors graduating in France and practicing in the private sector, which has put additional pressure on those who remain. Doctors are now expected to keep up with demanding and well-informed patients, handle administrative tasks, publish academic papers, and cope with the ever-increasing number of publications to read.

The overwhelming amount of medical publications, especially in the last decade (see Figure 1.1), coupled with controversies and limitations, poses a significant challenge for doctors who must navigate this sea of information to provide evidence-based medical care. Though impressive, this explosion in medical activity raises concerns about the reliability of published medical studies. Artificial Intelligence (AI) systems designed to synthesize vast amounts of data, like IBM's Watson, have faced challenges [215] in clinical decision-making. These systems often recommend unsafe or incorrect treatments due to limitations in the data used [96] to train them. The reliance on single studies with statistical significance, measured by a p-value of less than 0.05, has also contributed to a high rate of non-replication in research findings[91].

The question arises: How can research and the effectiveness of treatment and medical knowledge be improved amidst the influx of often unreliable data?

1.2 Motivation and Aim: Contributions to Sleep Medicine

To explore potential solutions in the sleep medicine area, this thesis draws inspiration from dermatology, specifically the advancements made in melanoma care. Over time, dermatologists have refined their understanding of melanoma from observations and inference studies on small datasets to predictive models based on genetic associations and clinical characteristics. Today, in that field, the accumulated knowledge allows a better understanding of genetic mutations, the metastatic process,

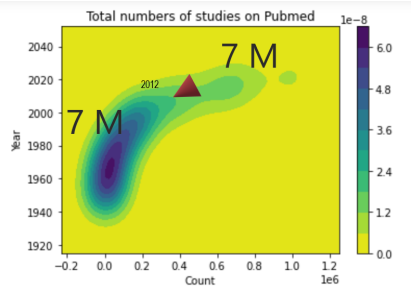


Figure 1.1: Divariate evolution of the number of Pubmed publications corresponding to the entry "Study" from 1935 to 2022

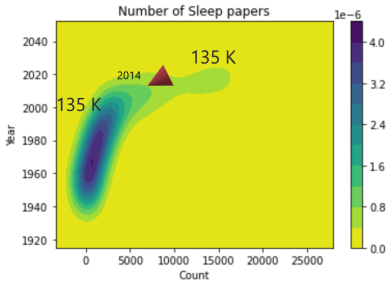


Figure 1.2: Divariate evolution of the number of Pubmed publications corresponding to the entry "Sleep" from 1935 to 2022

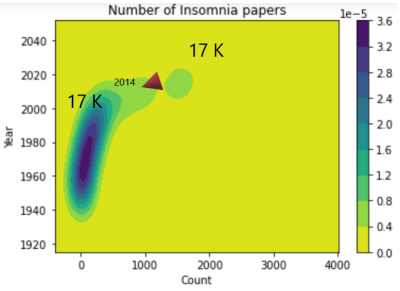


Figure 1.3: Divariate evolution of the number of Pubmed publications corresponding to the entry "Insomnia" from 1935 to 2022

and the different stages and responses to treatment. In the end, a dozen clinical types are described, and physicians can subgroup melanoma patients and match genotypes with therapies. These discoveries have recently benefited from predictive models that combined immune-related genes with clinical and morphologic characteristics to estimate melanoma patient survival and improve the decision-making in the treatment [125]. We are far from that in the sleep research area, but by applying a similar approach to sleep medicine, which emerged as an interdisciplinary field in the second half of the 20th century, we hope that using predictive models could help to gain insights into Chronic Insomnia Disorder (CID), particularly Paradoxical insomnia (ParI).

Although the severity of insomnia is infinitely less conducive to active research on an individual level, insomnia affects a significant portion of the population (10% to 30%) and remains poorly understood and challenging to categorize. However, its proper management is crucial due to its significant social, economic, and medical consequences, which are often underestimated. Despite ongoing research, the diagnosis of insomnia lacks clear objective criteria or biomarkers, and the treatment is moderately effective, with a 50% relapse rate at three years [145]. The International Classification of Sleep Disorders (ICSD-3) published in 2014 [183] consolidated various subtypes of CID into a single category, eliminating previously recognized subtypes due to poor convincing evidence. This simplification carries with it the risk of the clinician and the future sleep specialist losing some of the subtleties of the various manifestations of this disorder. It is also interesting to note that since these last classification, the number of publications has been the same as the total amount from 1945 to 2014 (See Figures 1.2 and 1.3) where the red cone represents the initiation point of the doubling of the publication count up until 12-31-2022). Perhaps this observation should spur us to find a new way of exploiting this gigantic mass of available data through tools that enable us to extract knowledge differently. At least, that's what we've tried to do in this work, based on classic data from a sleep assessment centre.

Indeed, a systematic evaluation and treatment unit for CID was established at the Pitié-Salpêtrière Hospital in 2011 to address these challenges. This thesis work builds upon the data collected through this unit to understand treatment-resistant CID better and contribute to the validation of insomnia subtypes observed in clinical practice, especially ParI.

The decision to employ Machine Learning (ML) in this research is motivated by the belief in ML potential to uncover novel patterns and relationships between variables using predictive models. Recent studies have emphasized the advantages of this approach, demonstrating that traditional linear models used in classical statistical inference may not capture the complexity of the data and may yield divergent results when applied to the same dataset [30]. In contrast to classical inference, ML techniques are specifically designed to extract knowledge directly from the data, making them well-suited for discovering intricate patterns that may have scientific implications in understanding biological, physiological, or clinical processes. By bridging the theoretical foundation provided by classical inference with the predictive capabilities of ML on real-world data, our working hypothesis is that leveraging the strengths of both approaches to identify patterns will bring meaningful insight into insomnia research, not only scientifically but also reproducible on unseen data. This last point aims to be a first step in the personalization and real-time use of data assisted by ML tools to improve the diagnosis and treatment of CID.

1.3 Research Questions

Thanks to many detailed publications on ParI described in datasets comparable to ours, but generally with far fewer features, we could use the learning capabilities of predictive tools used in Machine Learning to verify what the main and most accurate models among most efficient and popular binary classifier for each type of definitions published on our dataset. Once the prediction is achieved, we expect to explain each of the predictions with the assistance of different feature selection tools and compare them.

Our primary hypothesis is that we will find homogeneity in the different predictions to define ParI from the data perspective without preconception.

Q1. Can we provide an improved definition of ParI using a data-driven approach with Machine Learning tools? We will need to investigate the following points (sub-hypotheses):

- gather, study, and validate the definitions provided hitherto in the literature regarding ParI
- show that our dataset is a sufficient representation of the general insomniac population
- provide a robust and explainable definition with a minimum number of features, for reasons of interpretability and reliability

Our second aim is to understand better the treatment outcome, especially the resistance factor and relapses in CID. We aim not to evaluate a treatment's efficacy but to understand the factors determining its negative evolution.

Our secondary hypothesis is that we will find new insight into treatment resistance in a sample of CID patients using binary classification between favorable evolution (i.e., significant improvement) or unfavorable evaluation (no improvement of the problem or even worsening).

Q2. What are the main factors (features) driving negative evolution in treatment outcomes of CID?

We will need to investigate the following points (sub-hypotheses):

- find high-performing models
- identify the important features
- extract insight from the results

Our third interest in this work is times series analysis, specifically in the form of EEG brain recording signals. Indeed as we use brain recording every day, we wanted to know if the ML approach could bring some new insight into the two first hypotheses from a brain analysis perspective. But as we know, extracting useful features from a brain recording could be hard, especially when the recording is ambulatory like ours. Our third hypothesis is that we can find a reliable ML algorithm to extract meaningful data from non-controlled clinical recording with two sub-objectives: - Standardize the prediction of sleep stages and spindles from brain sleep recording to reproduce our research on other datasets with the same feature extraction protocol. - Depending on the results from above, use reliable brain-derived features extraction to predict ParI profiles and treatment outcomes in CID.

Q3. Can we use ML tools on brain recordings to accurately classify ParI profiles and treatment outcomes of CID?

We will need to investigate the following points (sub-hypotheses):

- automate the classification of sleep stages (as reliably as a human expert – or as close to this as possible)
- turn this information into features
- build predictive models for classification of ParI
- build predictive models for prediction of treatment outcomes

1.4 Thesis Organisation

This thesis will be elaborated as outlined below. Given this thesis's 'interface' (cross-disciplinary) aspect, we have made particular efforts to provide the key references to each field; throughout the manuscript. In the same spirit, two glossaries specific to each research area are present at the beginning of the thesis, and two sub-sections containing the main definitions for each research area make up the first part of the Appendix (B.1.1 for Sleep Medicine and B.1.2 for Machine Learning).

- Chapter 2 will summarize the main concepts of sleep medicine; and ParI in particular. We will also explain why we chose Machine Learning tools to tackle these problems instead of traditional statistics and methods.
- Chapter 3 will describe the data collection methodology, the difficulties encountered, the design and construction of the datasets (needed to overcome such difficulties), and their main characteristics; essentially to justify our final datasets as a representative insomniac population.
- Chapter 4, is devoted to testing ML and Deep Learning tools in different predictive tasks concerning the Macro and Micro sleep analysis. The first two experiments will concern sleep stages prediction; the following two will predict specific sleep biomarkers like spindles, and the last will be a cluster-based analysis of insomniac patients from ambulatory EEG recordings. The main objective is to see to what extent those methods could improve the standardization of sleep pattern analyses, enhance the understanding of insomnia, and participate in predicting treatment outcomes.
- Chapter 5 is devoted to the study of ParI in our clinical sample through the definitions provided by the scientific literature on this topic; in particular 20 proposed 'formula' (definitions) of Paradoxical insomnia (ParI). After evaluating the prevalence of each formula and their correlation in our sample, we evaluated the predictability of each formula (i.e., accuracy); and we proposed a new definition. A secondary outcome of this chapter is that we propose a protocol suited to medical doctors to explain the predictions of ML tools.
- Chapter 6 is devoted to predicting therapeutic response. We included our new definition of paradoxical insomnia in the data set. We used the same protocol as before, i.e., we selected the most appropriate binary classifier after hyperparameter adjustment and evaluated the accuracy using different metrics after cross-validation. We then evaluate the explainability of the prediction using the same protocol.
- Finally, in Chapter 7, we synthesize and discuss the general results and outcome of the thesis research, its contributions, and potential impact; we mention limitations and elaborate on some future perspectives.

Introduction La thèse présentée vise à approfondir la compréhension de l'insomnie paradoxale (IP), un des sous types de l'insomnie chronique qui affectent 10 à 20 Contexte et Problématique L'IP se manifeste par une discordance entre la perception subjective du sommeil et les mesures objectives, telles que celles obtenues par l'enregistrement d'un électroencéphalogramme (EEG). Les patients souffrant d'IP ont l'impression de ne pas dormir une grande partie de leur nuit de sommeil quand un enregistrement de leur sommeil montre une quantité de sommeil en général normale. Ainsi malgré des traitements adaptés, cette perception peut persister et conduire à un sentiment d'échec et d'impuissance de la part du thérapeute et du patient, entrainer une anxiété accrue et à une surenchère de traitements parfois iatrogènes entrainant des risques psycho-sociaux accrus. A ce jour, il existe encore une compréhension incomplète de ce trouble et une controverse concernant son existence propre. En effet il n'est pas encore tranché si l'IP est un sous-type de l'insomnie chronique ou un simple symptôme commun à tous les patients insomniaque chronique étiqueté mauvaise perception du sommeil . Cette controverse pose la question encore en suspens concernant l'existence même d'une entité propre nommée IP qui nécessiterait donc une définition claire et un traitement spécifique et la notion de mauvaise perception du sommeil qui serait un continuum intrinsèque à l'insomnie chronique. Dans les deux cas, il existe un problème commun non résolu qui est la définition d'un seuil de perception du sommeil considéré comme normal. Mais, au-delà de ce seuil à définir, il reste à trancher si l'IP

correspond à un sous type distinct de l'insomnie chronique, qui serait donc défini par d'autres caractéristiques cliniques, physiologiques ou psychologiques. Pour essayer de répondre à ces questions, nous avons décidé d'utiliser des outils d'AA pour utiliser sans à priori toutes les données disponibles concernant un groupe d'insomniaques et prédire le degré de perception du sommeil et les définitions publiées jusqu'ici de l'IP. Nous voulons également étudier l'impact de ces problématiques sur la réponse thérapeutique. Enfin, nous avons voulu savoir si les outils d'IA pourraient nous permettre d'exploiter de manière plus fiables et reproductibles les données complexes utilisées en neurophysiologie. Ces problématiques correspondent aux 3 hypothèses de recherche décrites ci-après. La première hypothèse testée dans cette thèse est qu'il est possible d'améliorer la définition d'un seuil de perception anormal du sommeil utilisable en clinique pour définir l'IP à l'aide d'une approche fondée sur les données et l'apprentissage automatique. Cette première hypothèse inclue de tester l'hétérogénéité des définitions déjà publiées sur un dataset représentatif d'insomniaques chroniques, et la proposition d'une unification de la définition basée sur une analyse de sommeil sur sept nuits consécutives au lieu d'une nuit habituellement. La deuxième hypothèse de recherche est que nous pouvons obtenir une meilleure compréhension des facteurs responsables de l'efficacité ou de la résistance à un traitement classique de l'insomnie chronique à l'aide d'une approche fondée sur les données et l'apprentissage automatique. Cette deuxième hypothèse inclut la possibilité d'une prédiction fiable du succès ou de l'échec thérapeutique sur des nouveaux patients. La troisième hypothèse générale est que l'on peut utiliser un algorithme de ML fiable pour extraire des caractéristiques significatives à partir de données EEG brutes et automatiser les interprétations et les prédictions pour pouvoir uniformiser la recherche sur le sommeil sans dépendre de la variabilité inter-experts. Matériels et Méthodes La première hypothèse est testé sur une base de données multimodale de 335 patients souffrant d'insomnie chronique (IC) constituée dans un centre spécialisé dans le diagnostic et la prise en charge de l'insomnie. Cette base inclut des données cliniques, psychométriques, actimétriques et polysomnographiques, comme l'EEG. Chaque patient inclus a été suivi pendant au moins six mois, permettant une évaluation précise du diagnostic et de la réponse au traitement standard. En utilisant des outils d'AA, l'étude a cherché à identifier des sous-groupes de patients et à tester des hypothèses existantes concernant les profils d'IP à travers l'analyse de l'EEG et psychométriques. La deuxième hypothèse concernant la réponse au traitement et l'implication des différents sous-types de l'insomnie chronique a été conduite sur une base plus élargie de 423 patients mais des données actimétriques moins exhaustives. La troisième hypothèse concerne différentes sous hypothèses dédiées à l'utilisation des tracés EEG pour prédire les stades de sommeil, l'extraction et la prédiction des fuseaux de sommeil, et la prédiction de l'intensité de la somnolence subjective. Résultats obtenus pour la première hypothèse : Les résultats de l'étude sur l'insomnie paradoxale (IP) révèlent une grande hétérogénéité dans les définitions existantes de cette condition. Plusieurs formules utilisées pour diagnostiquer l'IP ont montré que la majorité des patients étudiés étaient classifiés comme souffrant d'IP selon au moins une définition, mais il n'y avait pas de consensus général. La recherche a également indiqué qu'un groupe homogène de patients atteints d'insomnie chronique (IC) n'étaient jamais classés comme souffrant d'IP, quelles que soient les formules utilisées. Cette observation suggère que la perception erronée du sommeil n'est pas pathognomonique de l'IC. En utilisant l'apprentissage automatique, notre travail à permis de proposer une nouvelle définition de l'IP, basée sur une analyse temporelle plus longue et moins sujette aux aléas d'une seule nuit d'enregistrement, qui semble mieux refléter la réalité des patients et permet de distinguer plus clairement l'IP des autres formes d'IC. Résultats obtenus pour la deuxième hypothèse : L'hypothèse 2 de l'étude visait à améliorer la compréhension des résultats du traitement de l'insomnie chronique, en se concentrant sur les facteurs de résistance et de rechute. L'étude a utilisé des modèles prédictifs reconnus, notamment le Random Forest, l'Extreme Gradient Boosting et le Support Vector Machine, pour prédire l'issue du traitement chez 423 patients, atteignant une précision supérieure à 0,8. Les résultats ont souligné l'importance de la nouvelle définition de l'IP définie dans la première partie comme prédicteur majeur de la réponse au traitement. Cette découverte ouvre la voie à une approche plus personnalisée dans le traitement de l'insomnie chronique, bien que des études supplémentaires soient nécessaires pour une compréhension plus approfondie et pour valider ces résultats. Résultats obtenus pour la troisième hypothèse : A partir de l'EEG brut, ni un algorithme combinant la densité spectrale de puissance et la décomposition en mode empirique pour l'EEG, ni l'utilisation de réseaux de neurones (Convolutional Neural Network) n'ont montré une performance suffisante pour la prédiction fiable des états de sommeil, avec une précision inférieure à l'objectif fixé correspondant à l'agreement interscoreur (précision de 0,8). De plus le test de l'algorithme CNN pour l'évaluation

automatisée du sommeil a montré qu'il ne reproduit pas les résultats précédents sur d'autres ensembles de données, soulignant les limites de la transférabilité des algorithmes. L'évaluation de différents algorithmes pour la prédiction des fuseaux de sommeil a révélé une performance variable et souvent insuffisante, nécessitant la vérification par des experts. Enfin l'application de méthodes de clustering pour analyser les données EEG n'a pas permis de différencier significativement les groupes de patients basés sur des caractéristiques du sommeil. Ainsi globalement, cette hypothèse met en lumière les défis et les limites des techniques d'apprentissage automatique et d'analyse de données EEG dans le contexte de l'insomnie chronique.

Implications Cliniques et Perspectives Futures Les résultats de cette recherche ont des implications cliniques importantes. Ils suggèrent que l'approche actuelle de traitement de l'IP pourrait nécessiter une révision, en mettant davantage l'accent sur la perception subjective du sommeil sur plusieurs nuits. Cette approche pourrait aider à identifier plus précisément les patients souffrant réellement d'IP et à leur fournir des traitements plus ciblés et efficaces. En outre, cette recherche ouvre la voie à de futures études utilisant l'AA pour mieux comprendre et traiter d'autres troubles du sommeil. La capacité de l'AA à analyser de grandes quantités de données et à identifier des modèles complexes peut révolutionner la manière dont nous abordons les troubles du sommeil, conduisant à des diagnostics plus précis et à des traitements plus personnalisés. Conclusion Cette thèse représente une avancée significative dans la compréhension et le traitement de l'insomnie paradoxale. En utilisant des outils d'AA pour analyser des ensembles de données complexes, cette recherche contribue à une meilleure caractérisation de l'IP et à une prédiction plus précise de la réponse au traitement de l'IC. Les découvertes faites dans ce cadre pourraient transformer la pratique clinique et offrir de nouvelles perspectives pour les patients souffrant de troubles du sommeil.

Chapter 2

Background

Part I: Sleep and Insomnia

Chapter Highlights (PART I: Sleep and Insomnia)

1. Sleep Physiology and Sleep Medicine The classification of sleep into stages is based on the Electroencephalogram (EEG). The stages have identifiable brainwave patterns, and the scoring rules are well-established.

2. Sleep Physiology and Sleep States Sleep is characterized by three distinct levels of regulation: Wake (W), Non rapid eye movement (N-REM), and Rapid Eye Movement (REM) sleep. The sleep cycle comprises four stages: N1, N2, N3, and REM.

3. Sleep: From Normal Sleep to Disorders (a Categorization) Description of the different sleep disorders published in the third edition of the International Classification of Sleep Disorders (ICSD). Chronic Insomnia Disorder (CID) Classification of Insomnia has been harmonized with other international classifications and removed the different subtypes of CID like ParI, Psychophysiological Insomnia and Idiopathic Insomnia. These subtypes were removed for the lack of consensual scientific agreement, especially ParI.

4. Insomnia or Paradoxical Insomnia – Is an Objective Definition of Insomnia Possible? This thesis focuses largely on ParI; we discuss why there are so many definitions for this categorization and what are the difficulties in obtaining one. Indeed, despite more than 20 definitions published and attempts to find biomarkers, there is no one agreed-upon definition of ParI, and the scientific community remains divided on whether it is a disorder in its own right.

5. Launching this Thesis Our global thesis hypothesis is that ParI is a complex subtype of CID needing a more data-driven approach to be understood. We will take a Machine Learning approach.

Key Terms and concepts

Acronym/term	Definition	Ref.
CBT-I	Cognitive and Behavioral Therapy for Insomnia	p. 169 (B.1.1)
CID	Chronic Insomnia Disorder	p. 192 (B.9)
EEG	Electroencephalogram	p. 169 (B.1.1)
ICSD	International Classification of Sleep Disorders	p. 25 (2.4.1)
ISI	Insomnia Severity Index	p. 31 (2.4.1)
MMPI	Minnesota Multiphasic Personality Inventory	p. 32 (2.4.1)
ParI	Paradoxical Insomnia	p. 32 (2.4.2)
PsyI	Psychophysiological Insomnia	p. 170 (B.1.1)
nSSM	negative Sleep State Misperception	p. 170 (B.1.1)
SE	Sleep Efficiency	p. 170 (B.1.1)
SOL	Sleep Onset Latency	p. 170 (B.1.1)
TST	Total Sleep Time	p. 170 (B.1.1)
WASO	Wake After Sleep Onset	p. 170 (B.1.1)

This chapter covers the theoretical background needed to understand the value of using Machine Learning (ML) tools to study chronic Insomnia. First, we'll look at the basics of sleep physiology, the clinical assessment of sleep, and the main disorders encountered. We will then review the particularity of Paradoxical insomnia (ParI) or negative nSSM in the somnological clinic and the difficulty of defining it according to homogeneous criteria. Finally, we will see how ML tools could help us better understand this problem compared to the more conventional statistical tools.

2.1 Sleep Physiology and Sleep Medicine

It is well known that sleep can be classified into stages [7]; and that there are deep sleep and REM sleep. To a large extent, we owe these discoveries to physiological researchers and the contribution of EEG, a real-time recording of brain waves obtained by attaching flat metal discs (electrodes) to the scalp and reflecting the summation of the activity of millions of neurons close to the electrode. Figure 2.1 shows characteristic changes in brain wave amplitude and frequency during wakefulness.

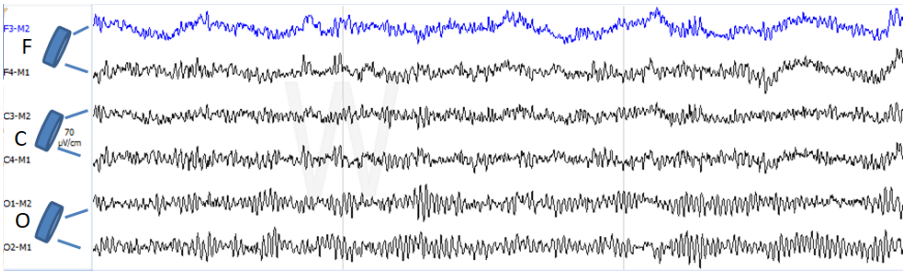


Figure 2.1: Typical EEG of relaxed waking state with alpha rhythm - increased amplitude in the occipital region [158]

Two additional physiological signals for identifying sleep exist: Electrooculogram and the Electromyogram.

The transformation by a sleep expert of the visual analysis of these different signals per 30-second period (epochs) into a staging code is called manual sleep scoring. The scoring rules are well-established and internationally recognized [191].

The five main stages of sleep are Wakefulness (W), N1 (Stage 1), N2 (Stage 2), N3 (Stage 3-4), and REM; where N1 to N3 are often referred to as N-REM sleep (as opposed to REM sleep). We can see in Figure 2.2 the characteristic aspect of the main stages. We describe each in more detail in Section 2.2.

In Figure 2.3 are represented EEG patterns caused by artifacts that have nothing to do with the brain signal. To avoid misinterpretations, the expert's task is to recognize all possible EEG aspects.

Standardized rules were edited to harmonize sleep scoring and the minimum quality standards to limit this possible misinterpretation. So, after applying these scoring rules, when all the epochs of each physiological signal are scored, we can visualize a chronological juxtaposition of all epochs scores called hypnogram (See Figure 2.4).

The EEG patterns of interest for scoring sleep stages are alpha, theta, delta waves, sleep spindles, and K-complexes, We as exemplified in Figure 2.5).

What's important to remember is that a scorer epoch corresponds to a kind of generalization of the EEG aspect observed over 30 seconds; an epoch is not an exact reflection of the frequency bands observed. To be scored in a certain stage, the EEG frequency corresponding to that 30-sec epoch must be present for at least 15 seconds. There is an exception for the W state; when the arousal duration is between three and 15 seconds of the total epochs, it is scored as a **micro-arousal**.

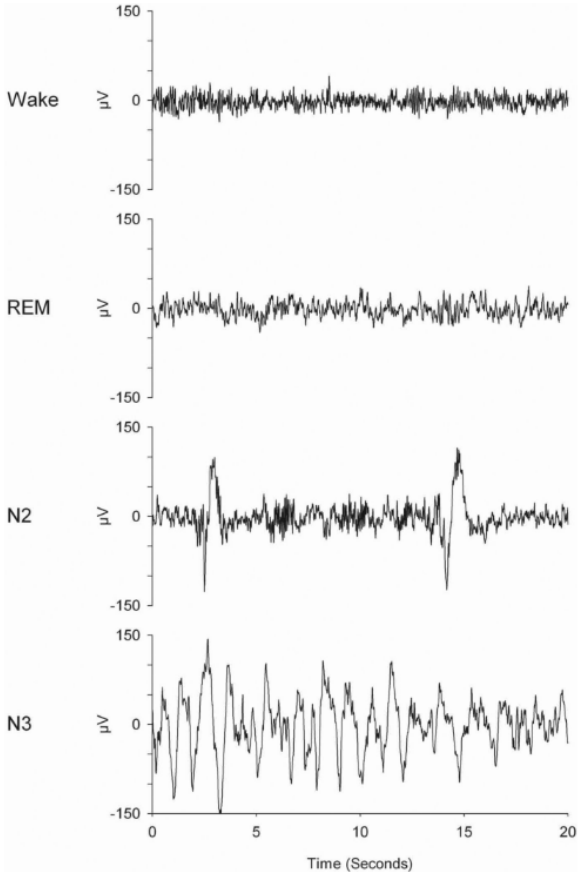


Figure 2.2: EEG visualization for the main sleep stages. N1 is a transitional state (see Figure 2.7 for EEG aspects) [31]

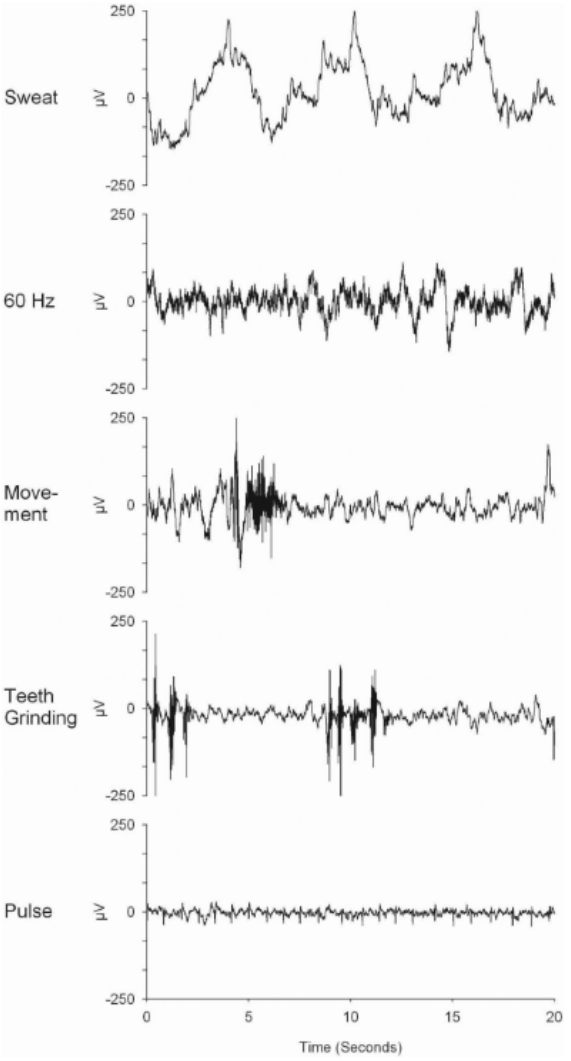


Figure 2.3: Classical artefacts seen in EEG [31]

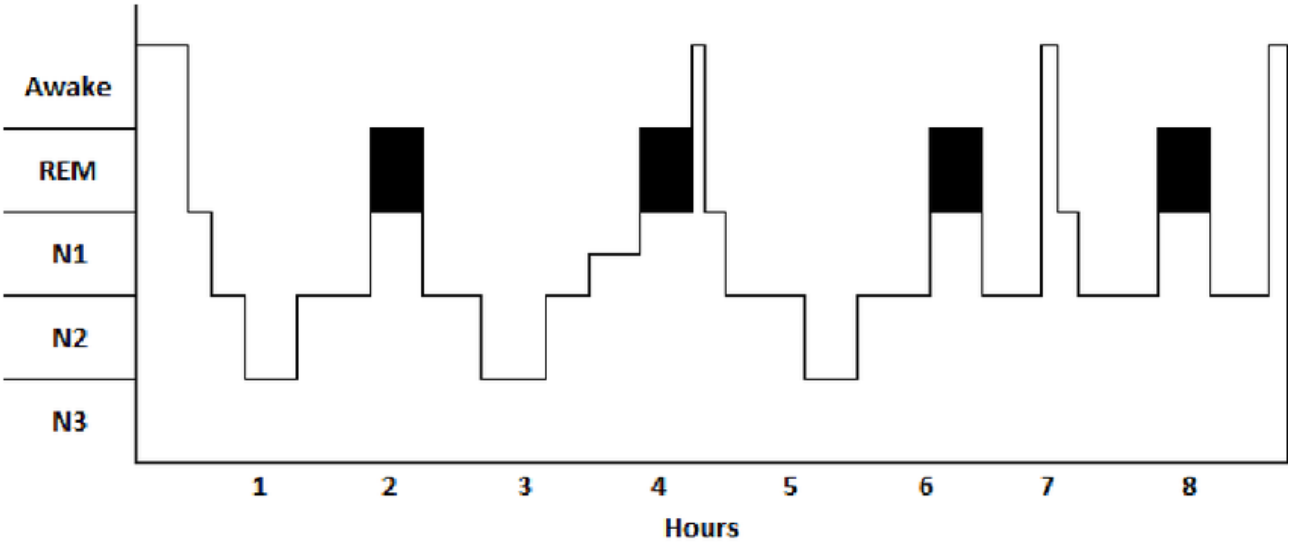


Figure 2.4: Typical normal hypnogram of scored human sleep staging in young-middle age adult [140]

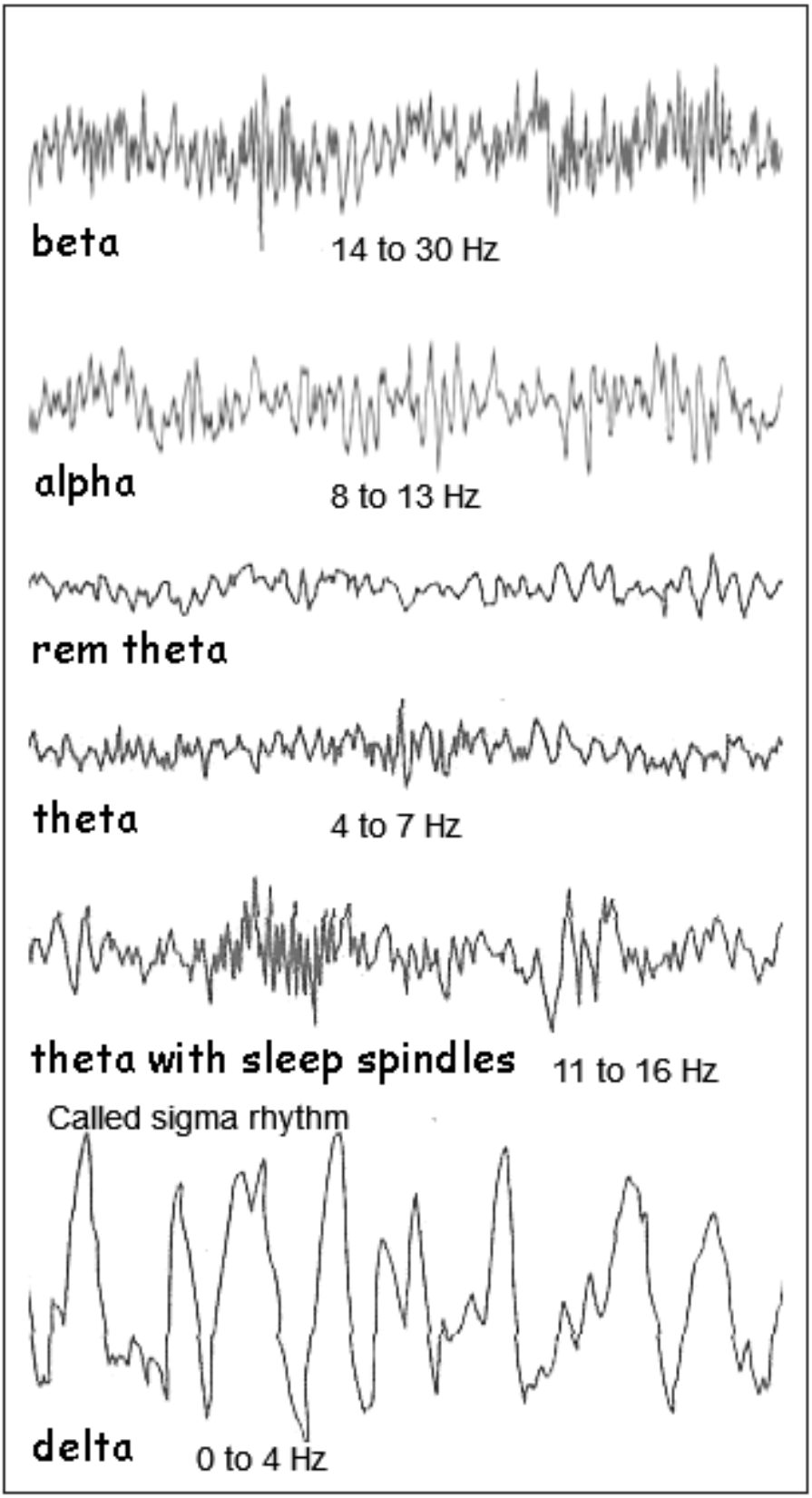


Figure 2.5: Main EEG frequencies and bandwidth seen in sleep scoring Adapted from G.Boeree 2009

2.2 Sleep Physiology and Sleep States

We will now briefly review the brain systems behind the signals recorded and their link with each stage.

Sleep

Sleep is a partial, periodic, immediately reversible suspension of consciousness under sufficient stimulation of the organism's sensory-motor and sensory relations with the environment. It is an active physiological process with several distinct behavioral stages and characteristic physiological states such as Heart Rate (HR), respiratory pattern, and cerebral rhythms. Sleep is different from a state of coma or anesthesia.

As we saw, sleep is characterized by three distinct levels of regulation: wakefulness (W), N-REM sleep, and REM sleep. The sleep state is a complex and dynamic process regulated by different brain and brainstem areas (mainly the Reticular formation). The alternation of these three states is done according to precise physiological rules under the dependence of homeostatic regulation processes, the circadian clock, and dedicated intra-cerebral nuclei.

Let's review each system.

Wakefulness (being awake)

It's a complex set of interdependent and competing systems involving different neurotransmitters that control the waking state or arousal. These wake circuits are all involved and interdependent to keep the brain awake no matter what, for obvious survival reasons. So, the waking state is volitional up to a point. These circuits are represented in Figure 2.6. In Figure 2.1, we can observe the digitization of the electrical signal when the brain is awake; these wake circuits are all involved and interdependent, aiming to keep but in a relaxed or drowsy state with a typical alpha rhythm.

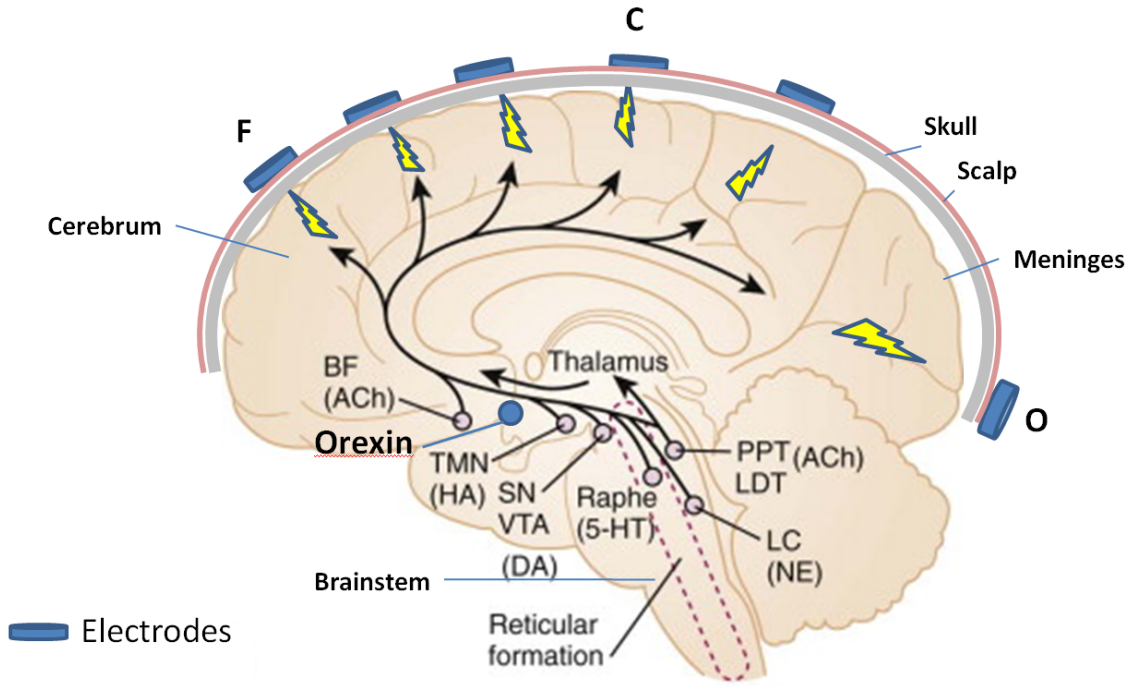


Figure 2.6: Main neural groups facilitating EEG desynchronization (wake) and awake motor behaviors such as walking. These neuronal groups are identified and differentiated by the neurotransmitters produced like Histamine (HA) in the Tubero Mamilar Nucleus (TMN), Acetylcholine (ACh) in the Basal Forebrain (BF), and pons (PPT/LDT), Serotonin (5-HT) in the Raphe nucleus, Noradrenaline (NA) in the Locus Coeruleus (LC) and Dopamine (DA) in the substantia nigra (SN). These systems depend on a neuropeptide synthesized in the lateral and dorsomedial hypothalamus, Orexin (hypocretin)). The electrodes mentioned, F (Frontal), C (Central), and O (Occipital), correspond to the electrodes usually used to record sleep. Adapted from [182]

The activity of the arousal circuits prevents the onset of sleep. The neuronal circuits of wakefulness constitute a permissive system that inhibits sleep. The N-REM and REM sleep dynamic could be triggered when this inhibition is lifted. This is when the different sleep stages can occur in a determined order. Above all, it's important to understand that **the deregulation of these neurotransmitters can be at the root of many sleep disorders, including Insomnia**, when hyper-activated. Indeed, Insomnia, especially in healthy young subjects, is more often a disorder of wakefulness regulation than sleep network functioning. In that case, thanks to sleep homeostasis, sleep always prevails in the absence of neurological disease, and the first sleep stage coming first physiologically is N-REM Sleep.

N-REM sleep

Here are the main sleep stages in detail.

N-REM sleep is divided into three stages of increasing depth. The ventrolateral preoptic nucleus controls the system involved in N-REM sleep. This nucleus inhibits the arousal system and thus promotes sleep. N1, N2, and N3 stages represent different intensities in the hyperpolarization of the thalamocortical neurons. The EEG pattern reflects these intensities, with a specific pattern recognizable in stage N2, the sleep spindles (See Figure 2.5).

Stage N1: This is the transition stage between wakefulness and sleep, during which a person may experience light sleep and muscle relaxation. The brain frequency slows down on the EEG signal, passing from alpha or beta to theta rhythm without spindles or K-complex (Figure 2.7).

Stage N2: This is the stage of light sleep during which a person's heart rate and body temperature decrease, and their brain waves become slower. On the EEG signal, the brain frequency keeps going in theta rhythm but with spindles and K-complex. Also, the percentage of delta rhythm is less than 20% (Figure 2.8).

Stage N3: This is the stage of deep sleep or Slow Wave Sleep during which a person's brain waves become even slower, and their body undergoes restorative processes such as tissue repair and growth hormone release. On the EEG signal, the brain frequency slows down in the delta rhythm; spindles can be seen in (Figure 2.9).

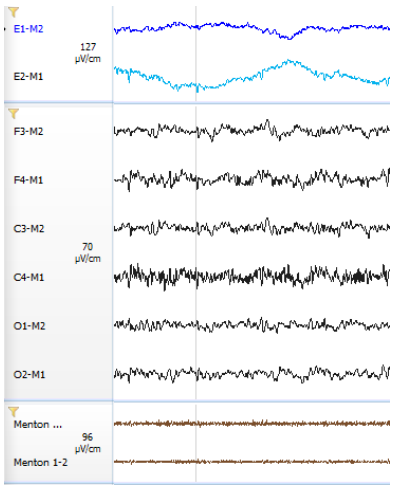


Figure 2.7: Typical N1 stage with dominant theta rhythm (from personal collection)

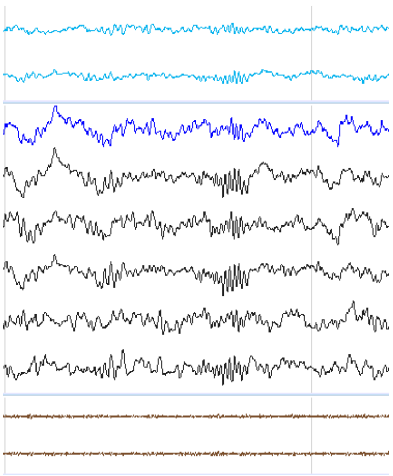


Figure 2.8: Typical N2 stage with K-complex and spindles

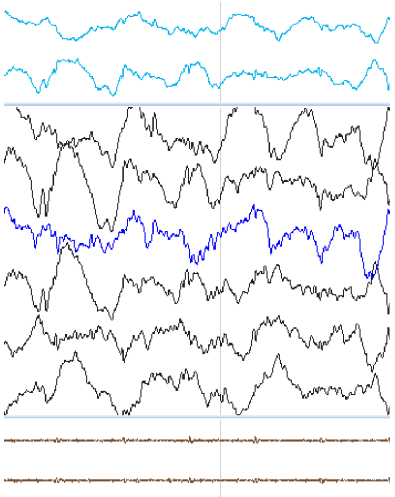


Figure 2.9: Typical N3 stage with dominant delta rhythm

REM sleep (paradoxical sleep)

This sleep state is characterized by a brain activity close to EEG de-synchronization observed in the W state but mixed with slower theta rhythms and specific waves called "sawtooth waves" (Figure 2.2). Although REM sleep could be identified by an experienced sleep expert only with EEG, the task is made easier using EMG and EOG to characterize better rapid eye movements and muscle atonia generally associated. During REM sleep, the brain is highly active (vivid dreams), and the body undergoes various physiological changes. Indeed, heart rate and breathing become faster and more irregular, and blood pressure and body temperature could change.

REM sleep consists of two distinct periods: phasic REM and tonic REM. Phasic REM is characterized by bursts of rapid eye movements recognizable by their phase inversion (Figure 2.10). Tonic REM sleep comprises the same background activity without eye movements (Figure 2.11). REM sleep is usually accompanied by muscular atonia with brief contractions visible on the EMG sensors, but sometimes it could be missing.

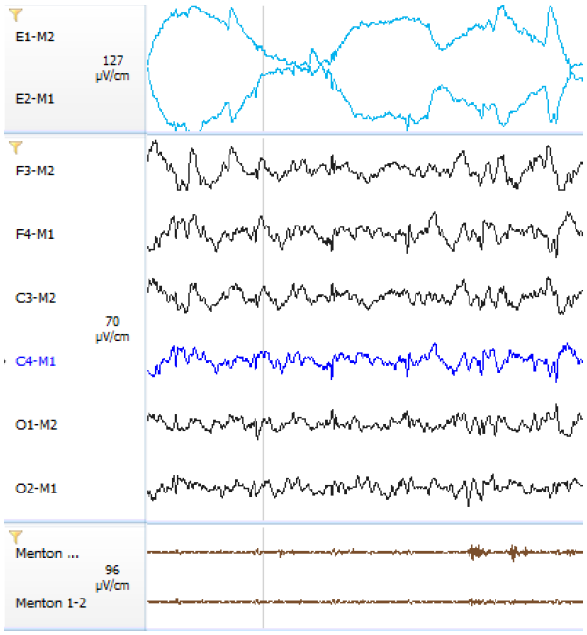


Figure 2.10: Typical Phasic REM sleep

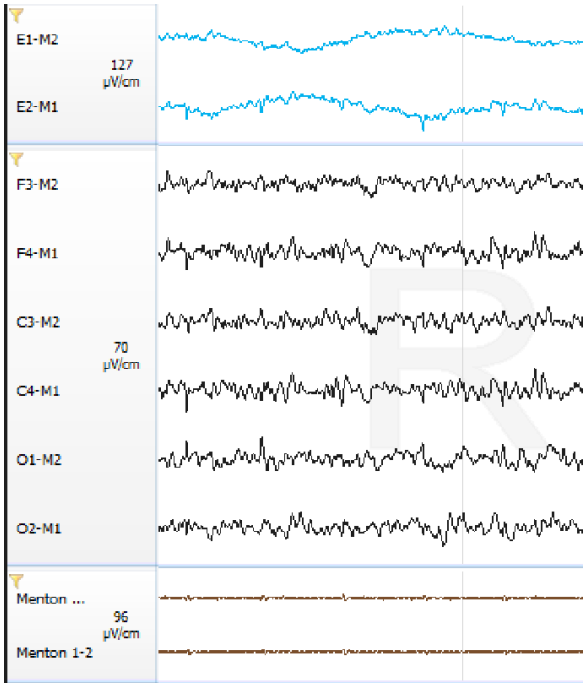


Figure 2.11: Typical tonic REM sleep

Now that we've seen the physiology of sleep circuits, we need to understand the role of the biological clock.

Biological clock

To briefly describe this other sleep-regulating system, we'll say that the biological clock in the suprachiasmatic nucleus is the conductor of the variation of biological rhythms. Under its influence, the body's various functions fluctuate over 24 hours, like temperature. This is the "circadian rhythm", partly defined by the alternation of day and night. Every 24 hours or so, these activities peak and trough. These peaks respond to a temporal structure genetically programmed by our organism. Over time, natural selection has favored an endogenous rhythm. It synchronizes with a time close to the Earth's rotation, even without environmental signals (daylight, social activity, etc). This endogenous rhythmicity is regulated by genes that underpin the functioning of the biological clock.

This clock regulates our sleep rhythm and is located in the suprachiasmatic nucleus of the hypothalamus. Each person has their clock rhythmicity, and the importance of the biological clock in sleep is linked to its favorable or unfavorable action on the conditions that promote the activation of sleep circuits and the deactivation of wakefulness circuits. The role of the biological clock is, so to speak, to determine the "sleep gates" by creating the optimal conditions to fall asleep (e.g., by causing a drop in temperature). Thus, the biological clock plays a central role in the sleep-wake balance, particularly its close link with melatonin secretion. Melatonin is an hypnogen. Its intracerebral secretion is regulated by the biological clock and light intensity, especially blue light. All this regulation, which can vary from one individual to another depending on the length of the clock period, is at the origin of different chronotypes in the population, resulting in those who go to bed early, those who go to bed late, and the "normals".

2.3 Sleep: From Normal Sleep to Disorders (a Categorization)

Normal sleep for adults

This Section aims to describe normal sleep in healthy adults. We won't be talking about children's sleep, which has physiological peculiarities parallel with brain development.

A normal adult's average sleep time is between 7h30 before working days and 8h30 on weekends. The physiological variation in duration is between a minimum of 5-6h and a maximum of 9-10h. But what makes sleep normal is also, and above all, the distribution and duration of sleep stages. The sleep distribution or structure shown in Figure 2.4 underpin the importance ofN-REM sleep in the first part of the night. REM sleep normally appears quantitatively more at the end. The appearance of these sleep stages obeys an arrangement in the form of cycles, averaging 90 minutes each and bringing together a different proportion of each stage, starting withN-REM and ending with REM (influence of the circadian clock). On a normal night, we generally count 4 to 6 sleep cycles. At the end of a standard 8-hour night's sleep, its restorative effect is assured if the following quantitative parameters are met:

1. Wakefulness during sleep generally for less than 5% of the night.
2. N1 Sleep as a transitional stage accounts for 2-5% of sleep.
3. N2 sleep represents between 45% and 55% of sleep.
4. N3 sleep accounts for 13% to 23% of sleep.
5. REM sleep accounts for 20-25% of sleep
6. Micro-arousals can occur during sleep and do not indicate an abnormality if less than 15-20 per hour.

These values are slightly modified with age but remain stable up to 60 yo (See Figure 2.12. From age 65-70, we could observe a progressively increasing reduction in stage N3 and a physiological increase in stages N1 and W symbolized by the increasing time spent in Wake After Sleep Onset, often with polyphasic sleep. Sleep in older people is, therefore, a subject in itself. All of the above illustrate the importance of age groups in sleep studies.

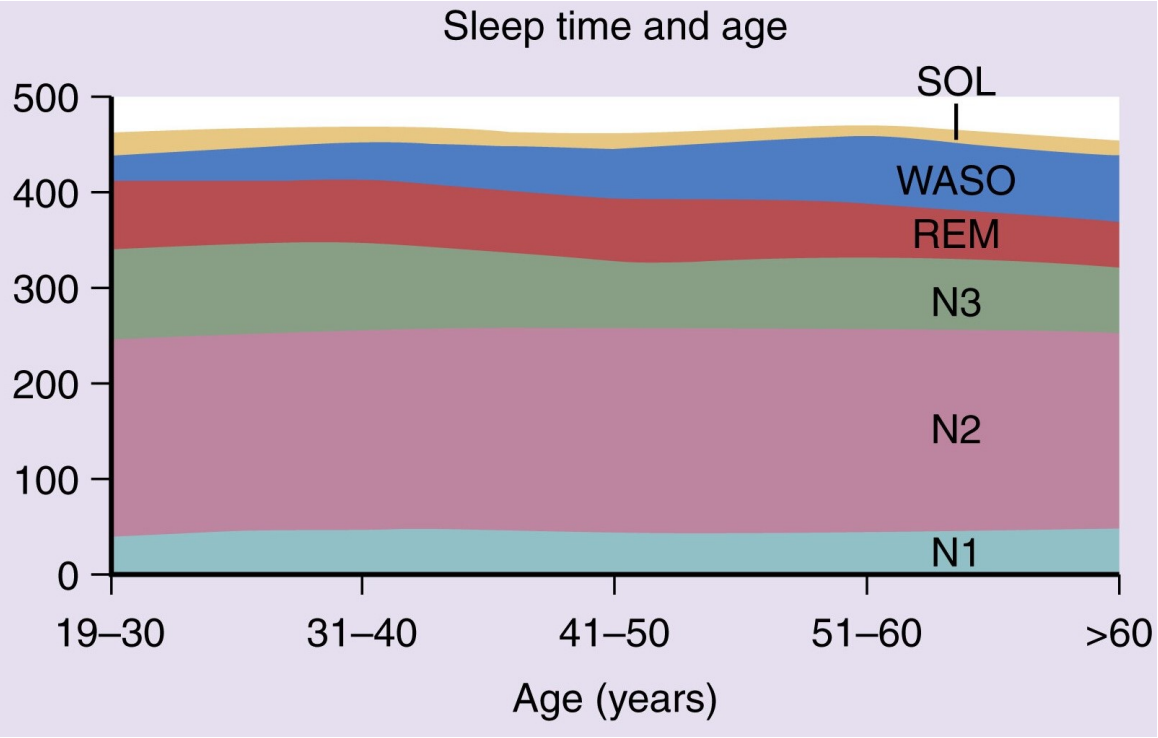


Figure 2.12: Graphic evolution of sleep stages proportion for a night duration 7h45 (465 minutes). W = SOL + WASO [107]

Sleep disorders

Sleep disorders are categorized in a standard way in the third edition of ICSD-3 [183]. The main sleep disorders listed are:

1. **Sleep-related breathing disorders**: Breathing disorders linked to sleep primarily affect normal respiration during nighttime hours. The principal issue is the recurrent interruption of breathing, which can be either partial (hypopnea) or total (apnea), causing sleep fragmentation and resulting in daytime consequences like heightened drowsiness and diminished cognitive functions. These disorders are chiefly categorized into obstructive and central sleep apnea, based on whether the disruption originates from obstructions in the upper airways or malfunctions within the respiratory control center.

2. **Central disorders of hypersomnolence**: The symptoms are excessive daytime sleepiness despite adequate sleep at night, often accompanied by automatic behavior and sleep attacks. Specific questionnaires can assess the complaint, but the diagnosis must be confirmed by multiple naps recorded during the day showing a short sleep latency (less than 8 min). This test is the Multiple Sleep Latency Test. The emblematic disorder is narcolepsy.

3. **Circadian rhythm sleep-wake disorders**: It's a mismatch between the timing of the sleep-wake cycle and the external environment, leading to impaired daytime functioning. It is a disorder of the biological clock, of which the most frequent are the delayed sleep-wake phase disorder and shift work disorder.

4. **Parasomnias**: They correspond to behaviors or experiences during sleep. Depending on the stage of sleep in which they appear, we will have the disorders of N-REM sleep like sleepwalking or sleep terrors, or those appearing during REM sleep like REM sleep behavior disorder or Nightmare disorder. Other parasomnias are described as night eating disorders or hallucinations.

5. **Sleep-related movement disorders**: They are characterized by abnormal movements during sleep, such as periodic limb movement disorder and restless legs syndrome.

6. **Other sleep disorders**: a group of sleep disorders that do not fit into the other categories.

As this is the subject of our study, we will describe Insomnia in detail in the next Section.

2.4 Insomnia or Paradoxical Insomnia – Is an Objective Definition of Insomnia Possible?

In this Section, we place ParI in the global context of Insomnia and the evolution of diagnostic criteria since the 1970s. At the same time, the tools, especially questionnaires or objective measures, evolved. We think it is important first to develop the general Insomnia diagnostic context and the assessment tools before describing ParI diagnostic and definitions. In the Subsection 2.4.1, we will evaluate the main tools used to assess Insomnia, some of which were used in our sample are marked with a † and will benefit from an extensive description.

2.4.1 A historical perspective

Unlike many other medical disorders, Insomnia has been described and classified very differently across time; i.e., the existing definition of Insomnia can already be considered unstable. Indeed, in addition to the ICSD-3, Insomnia is also described in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5 [8]) and the International Statistical Classification of Diseases and related health problems (ICD-11 [156]). These last two classifications are aligned with the diagnostic criteria proposed in the ICSD-3 that we will retain as the primary reference. However, this has not always been the case. This harmonization of classifications is the result of a long process that began with the publication of the first classification in 1979, the Diagnostic Classification of Sleep and Arousal Disorders (DCSAD),

which was created in the United States under the auspices of the Association of Sleep Disorders Centers and the Association for the Psychophysiological Study of Sleep. Remembering the history of the various classifications is crucial, as it aids in grasping the intricate nature of Insomnia and the breadth of what it encompasses..

Although we have had many definitions, one can argue that now we have too few (from 16 subtypes to only 2), as we still have not arrived at an objective definition of Insomnia or its sub-types, in particular, ParI (our main point of study in this work), whereas in everyday clinical practice, we can continue to observe clear evidence for subtypes. This is a key motivation for our work. See, Table 2.1.

ICSD-2
1. Adjustment sleep disorders*
2. Psycho-physiological Insomnia
3. Paradoxical Insomnia (ParI) **
4. Idiopathic Insomnia
5. Insomnia due to mental disorder
6. Behavioral Insomnia of Childhood
7. Insomnia due to a medical condition
8. Insomnia due to a drug or substance
9. Non Organic Insomnia
10. Organic Insomnia

Table 2.1: ICSD-2:Different subtypes of Insomnia described in the second edition of the ICSD (2005) *acute Insomnia **formerly nSSM

ICSD-3
1. Chronic Insomnia disorder
2. Short-term Insomnia disorder
3. Other Insomnia disorder
4. Isolated symptoms-normal variants
5. Excessive time in bed
6. Short sleeper

Table 2.2: ICSD-3: Different subtypes of Insomnia described in the third edition of the ICSD (2014)

This simple description is a sufficient reference now for the diagnosis (of CID). Indeed, it is even mentioned in ICSD-3 that *"the degree of sleep disturbance required to assign a chronic Insomnia disorder diagnosis is somewhat arbitrary."*

One may even wonder **how 35 years after the first classification of sleep disorders [190] in which numerous Insomnia sub-types are mentioned, there are only these minimalist diagnostic criteria to explain this disorder**. Indeed, with the constant progress of medicine for a disorder affecting between 10 and 30 percent of the population depending on its intensity, how is it possible not to be able to define clear subgroups of Insomniacs that could lead to a better understanding and, therefore, better management? This is all the more difficult to understand given that this simplification of Insomnia categorization has occurred simultaneously as an almost exponential increase in publications on the subject (See Figure 1.3).

Supporting our point, the authors of ICSD-3 stated that distinguishing these subtypes is challenging. This is the case because the present definitions fall short of providing the scientific community with a solid basis for decision-making, and the existing identification methods are not consistently reliable.[183].

However, this claim seems to be more of a concession of the scientific community's inability to decode the intricate nature of studies on this topic, primarily due to the absence of clear diagnostic criteria. This lack of precision since the earliest classifications has allowed for including Insomnia patients under varying criteria in different studies. Consequently, the surge in publications since the 1970s has muddled the understanding of this ubiquitous disorder rather than clarifying it. The increasing amount of data on the subject results in more confusion, creating a cyclic dilemma akin to a snake biting its tail.

We argue that there is an objective truth (definition) for ParI, and we set out to gather evidence towards defining it.

Drawing Parallels with Hypersomnia: The Complexity of Classifying Insomnia Subtypes

Our thesis choice, the study of subtypes of a disorder that lacks an official classification, demands a robust defense. To highlight our perspective, we draw an analogy with Hypersomnia, a disorder at the opposite spectrum, also mentioned in ICSD-3 [183].

Unlike Insomnia, Hypersomnia, particularly type 1 Narcolepsy, has a distinct biomarker – the generalized absence of hypocretin peptides in human narcoleptic brains [166]. This condition serves as

a model of the interplay between the immune system, the nervous system, and the sleep-wake system [115]. As a result, it presents a model with specific biological and somnological criteria, marking it as a "primary disorder" or "intrinsic disorder" as outlined in the ICSD-3's Central Disorders of Hypersomnolence chapter. The diagnosis is precise, with objective sleep measures and clear thresholds for determining the presence of the disorder. However, upon closer examination and comparing the broader classification of "Central Disorders" (disorders of neurological origin), we could find individualized subtypes of Hypersomnia like Idiopathic Hypersomnia (IH), but also Hypersomnia "secondary to", for example, to mental disorders. This up-to-date classification can be seen in Table2.4. But the point here is that this Hypersomnia subtyping mirrors the last classification of Insomnia in the ICSD-2, now removed (see Table 2.3). So, all these subtypes distinctions still present in Hypersomnia would have disappeared for Insomnia for insufficient specificity in the diagnostic criteria previously used, unlike the ones used in Hypersomnia.

ICSD-2: Insomnia categorization
Psycho-physiological Insomnia
Paradoxical Insomnia (ParI) *
Idiopathic Insomnia
Insomnia due to mental disorder
Insomnia due to a medical condition
Insomnia due to a drug or substance
Non Organic Insomnia
Organic Insomnia
Isolated Symptoms and Normal Variants
Short Sleeper

Table 2.3: ICSD-2:Differents Subtypes of Insomnia described in the second edition of the ICSD (2005) *formerly nSSM

ICSD-3: Hypersomnia categorization
1.Narcolepsy Type 1
2.Narcolepsy Type 2
3.Idiopathic Hypersomnia
4.Kleine-Levin Syndrome
Hypersomnia Associated with PD
Hypersomnia Due to an MD
Hypersomnia Due to molecules
Insufficient Sleep Syndrome
Isolated Symptoms and Normal Variants
Long Sleeper

Table 2.4: ICSD-3:Different Subtypes of Hypersomnia described in the third edition of the ICSD (2014), MD = Medical disorder, PD = Psychiatric Disorder

However, when we look closely at the diagnostic criteria for Idiopathic Hypersomnia, for example, it is a clear individualized subtype, exactly as Paradoxical Insomnia in the previous classification). So, the reason for keeping Idiopathic Hypersomnia is the presence of objective criteria clearly defined, especially the results obtained in the Multiple Sleep Latency Test (MSLT). This test is used as the main physiological proof of sleepiness [34]. This criteria to be fulfilled require the presence of the two following objective sub-criteria:

1. The MSLT shows a mean sleep latency of $\leq$ 8 minutes.

2. Total 24-hour sleep time is$\geq$ 660 minutes (typically 12–14 hours) on 24-hour polysomnographic monitoring (performed after correction of chronic sleep deprivation) or by wrist actigraphy in association with a sleep log (averaged over at least seven days with unrestricted sleep.

This criterion gives the feeling of a clear cut-off to validate or not the subjective complaint of sleepiness described by the patient. But, it is mentioned in note 4 under the diagnostic criteria: *Occasionally, patients fulfilling other criteria may have an MSLT mean sleep latency longer than 8 minutes and total 24-hour sleep time shorter than 660 minutes.* ***Clinical judgment should be used in deciding if these patients should be considered to have idiopathic Hypersomnia (IH).*** *Great caution should be exercised to exclude other conditions that might mimic the disorder. A repeat MSLT at a later date is advisable if the clinical suspicion for IH remains high*[183]. In the same classification, it's mentioned that no consistent precipitating factor has been identified, that the prevalence is unknown, and that the pathophysiology of IH is unknown.

Put together; we have a disorder (Idiopathic Hypersomnia) whose origin we do not know, which has precise objective diagnostic criteria but **can be modified according to clinical judgment**. This is exactly the criticism leveled at Insomnia subtypes, especially Paradoxical Insomnia. Therefore, we argue **we should not withdraw a diagnostic categorization (of Insomnia) that has been clinically observed for decades because we still do not have the explanation**. This is the argument that is central to the subject of this thesis.

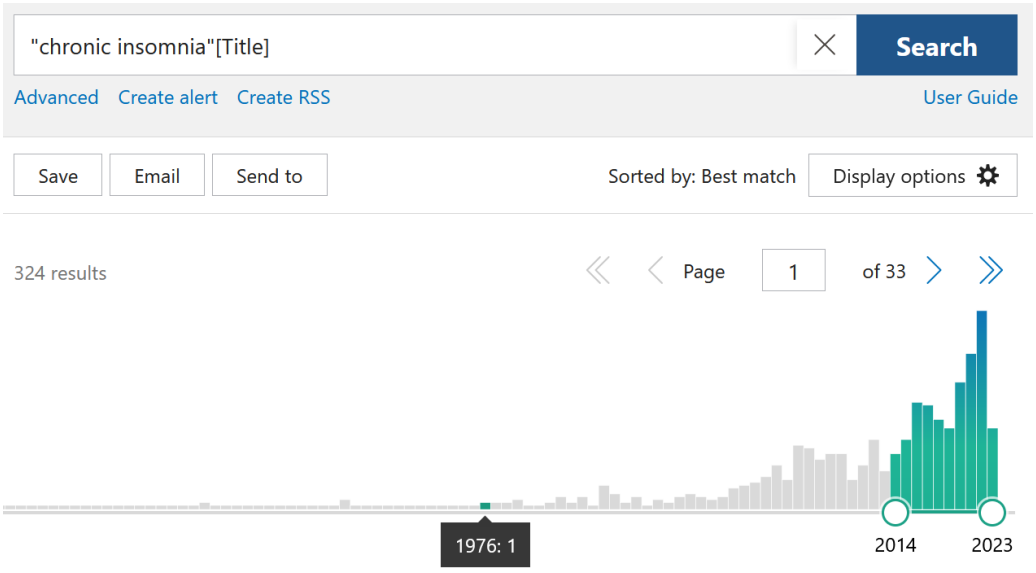


Figure 2.13: Pubmed research for Chronic Insomnia in April 2023

Related work trying to find biomarker of Insomnia subtypes

The main cause of seeing Insomnia removed from the official classification is arguably due to lack of objective biomarkers. Now, we look at other work that has tried identifying such biomarkers with modern data-driven approaches.

As we saw, half of the publications on Insomnia have occurred since 2014, with many proposing either new diagnostic criteria or new approaches, making it difficult to compare different studies. One of the difficulties, in particular, concerns the definition of Insomnia diagnoses, the comparability of groups, and associated or non-associated comorbidities. For Machine-Learning and Big Data purposes, the biggest database to date with physiological and psycho-social is the UK Biobank with 500000 subjects [200]. However, concerning full physiological sleep data crossed with a psychological evaluation, the biggest databases are less than 1000 subjects, even to find subgroups of CID[19]. The biggest sample using several thousand subjects with online questionnaires didn't show sufficiently high sensitivity and specificity, especially without physiological data [168].

One study empirically derives and evaluates potential subtypes of CID through cluster analysis from Polysomnography recording [141]. They performed a cluster analysis using Euclidean distance and Ward's method on a population. But they chose priory distinct clustering variables on theoretical grounds and previous research on the objective sleep parameters, including TST, SOL, and WASO. Other possible cluster solutions were examined against external variables associated with CID for validity, like neurocognitive performance, sleep-onset measures of quantitative EEG, and heart rate variability. The studies concerned 100 volunteers (61 females, mean age 41.4 *SD* 11.8) with the CID diagnostic. They found two distinct clusters: Insomnia with Normal Sleep Duration (I-NSD) and Insomnia with Short Sleep Duration (I-SSD). Surprisingly, they found no differences in subjective sleepiness between the two groups measured by the Epworth Sleepiness Scale as a possible effect of sleep deprivation. The study by [48] examined empirically derived symptom cluster profiles in 175 individuals (63% female) with CID based on scores on validated questionnaires (Insomnia Severity Index (ISI)), Glasgow Sleep Effort Scale, Fatigue Severity Scale, Beliefs and Attitudes about Sleep, ESS and Pre-Sleep Arousal Scale). They found three symptom cluster profiles: "High Subjective Wakefulness" (HSW), "Mild Insomnia" (MI), and "Insomnia-Related Distress" (IRD). However the population is a mix of psychophysiological Insomnia and Insomnia disorder comorbid with Obstructive Sleep Apnea. Thus, we could observe a bias through the mean and Apnea-hypopnea index, respectively 29.2 ($\pm$ 8.8) and 15.2 ($\pm$ 21.6). This means a weight close to obesity and moderate sleep apnea syndrome. They showed that the MI type corresponds to the patients having more apnea; the first would be related to a wrong perception of sleep, and the 3rd the most related to the worries before sleeping.

Other authors have attempted to work on much larger patient cohorts using an online platform and linked database that extensively surveys sleep, personality, and affect traits, life events, and health conditions. A study [19] involving 4322 subjects and the inclusion criteria were the completion

at their convenience of at least one of 34 available questionnaires assessing six dimensions with 523 items in total (Sleep(5), Life history(2), Fatigue and arousal(7), Personality traits (9), Mood (8) and Happiness(2)), a demographic questionnaire, and an assessment of their Insomnia Severity Index (ISI). On the only basis of ISI score > 10, 2224 (51%) participants fulfilled the probable Insomnia disorder criterion. The authors mentioned that they confirmed the validity of this threshold in a subsample of 244 subjects. They used a model-based unsupervised (see definition in B.1.2) clustering technique named latent class analysis and identified five subtypes (see Figure B.9 in Appendix). The subtypes are mainly driven by the personality, mood, and happiness questionnaires with no influence on ESS, the chronotype, or the family history of Insomnia. The researchers validated their five-subtype model in a nonoverlapping sample and found exceptional subtype stability after a mean follow-up of 4.8 years, especially for subtype 1, highly influenced by negative affect, childhood trauma, and life events (80% of women). For the follow-up, they used 207 items only (after colinearity exclusion and LASSO regularization (see the definition in B.1.2). They also investigated the clinical relevance of these subtypes for the developmental trajectories of sleep complaints, current comorbidities, depression risk, and response to benzodiazepine intake, as well as an EEG biomarker and the effectiveness of cognitive-behavioral therapy for Insomnia for two of the subtypes. For the response to benzodiazepine intake, they found a clinical relevance of using those subtypes by showing differential subjectively experienced effects (112 subjects). The EEG biomarker was an Auditory event-related potential for classic tones and deviant tones recorded during an oddball task, but they did not find any significant control subtype difference (n=16 for subtype 2, n=13 for subtype 4, n=31 for the control group). For CBT-I, again insufficient data were available for subtypes 1, 3, and 5; only 43 subjects subtype 2 (mean age 50·8±12·9 years, 88% females) and 25 subtypes 4 (mean age 53·2±9·8 years, 96% females) could participate. They couldn't show a significant difference but a slight difference in the treatment response between these two subtypes regarding the decrease in ISI score. **In the end**, although this study has the merit of crossing multiple data on Insomniac profiles with many subjects, it seems there are several limitations to be accepted by the community. First, the subjects were recruited online, presenting a basic bias regarding profiles. Secondly, there is no standardization of the questionnaires since only one out of 34 is required to be included in the study. There is no objective assessment of sleep. The EEG assessment was performed on a small fraction of the patients and only on two subgroups of five, as well as the online CBT-I. Thus, this study confirms the need for standardization and objectification of subgroups of Insomniacs, showing that this could allow the evaluation of the response to treatment, but this remains to be determined. It opens the way to publications using ML tools like LASSO to select the most significant parameters. Finally, this study also shows that it can be easy to set up an online questionnaire but is much more difficult to follow and evaluate the subjects objectively since only 0.01% of the sample could have an EEG and 0.03% an CBT-I.

The purpose of this long summary of the history of Insomnia classification from 1979 to 2019 is to show the complexity of the concept of chronic Insomnia, the complexity of its classification, and the difficulty of harmonizing the results of the various studies published over the last 40 years.

Before digging into the ParI concept and the main studies that have attempted to determine what ParI is, it is worth describing the tools used in most studies to understand better their scope and the features used.

Assessment of Insomnia

Apart from research, in clinical practice, the most important aspect of sleep assessment is the clinical interview and examination, looking for pathognomonic signs and comorbidity that could explain sleep disorders. If there is a sleep problem, that's when you can schedule tests. Although the ICSD-3 specifies that the diagnosis of Insomnia is clinical, it also states that tools to objectify the Insomniac's complaint are necessary in complex or resistant cases. In our practice, this is generally always the case. In Section 2.1, we have already seen the importance of recording brainwaves to characterize sleep, but how sleep can be objectively and subjectively assessed has been the subject of research, resulting in numerous tools available for the clinician and the researcher. We'll take a look at the main ways of assessing sleep. The tools used in our different studies are marked with a † and won't have additional descriptions to avoid further repetitions. In the same spirit, the features extracted from these tools, which we will use to build our various datasets, are summarized in Tables in the Appendix B.3 for each type of data collected and marked with a ‡

First, **Polysomnography**† is the gold standard for an extensive sleep evaluation. In clinical practice, we add sleep-specific sensors to the EEG, like oxygen, respiration, and leg movement sensors, to evaluate respiratory and neurological sleep disorders, usually coupled with video and sound recording (see Figure B.8). However, it is a rather invasive examination, generally done in a sleep laboratory under the control of trained personnel. It is quite possible to do it at home, which is generally desirable when evaluating sleep quality, but at the cost of artifacts. When recording at home, intervening if an electrode becomes partially detached is impossible. There can also be electromagnetic artifacts from electronic devices, transportation, or sweat if the temperature is poorly regulated. The particularity of a PSG compared to a simple EEG is that the placement of the electrodes requires expertise so that it can record a physiological signal correctly for several hours. So, the different sleep states can be recorded and identified directly by electrodes collecting the electrical activity produced by the brain. However, as the electric signal must be conducted through the meninges (the three membranes that cover and protect the brain and spinal cord), the skull, and the skin, a strict protocol must be followed to obtain a good signal. In the case of sleep recordings, the electrodes must remain in place for an average of ten hours, cleaning and rubbing the skin, applying a conductive paste, and maintaining the electrodes with a strong glue to hope to have a good quality signal throughout the night. The ideal is to have an impedance of 100 Kohms and, in any case, lower than 5 Kohms for each electrode as the reflection of a good signal-to-noise ratio.

The international rules state that the EEG signal must be recorded on the scalp with electrodes located in the Frontal (F), Central (C), and Occipital (O) regions, usually F3, F4, C3, C4, O1, and O2. By convention, even numbers are on the right, and odd numbers are on the left (See Figures B.5, B.7 and B.6 in Appendix B.4). The choice of this setup is justified by the need to cover the whole scalp with a minimum of electrodes, avoid artifacts (Electrocardiogram, movement), and also because some EEG patterns are more specific to a given region.

We can see in Figures 2.14 and 2.15 the technical difficulties of a polysomnographic recording and an example of an artifact when an electrode becomes partially detached. In general, the electrodes must be completely detached to be unable to interpret the different stages of sleep. Still, rescheduling patients in case of a bad recording is part of the usual practice. When a sleep expert has manually scored a sleep recording, he can print the synthesis as a sleep report with the primary data on the patient's sleep, such as the total time of sleep, the number of awakenings, the percentages of each stage, etc. Although EEG is usually sufficient for labeling the different stages of sleep, the findings on REM sleep and its peculiarities, such as rapid eye movements and muscle atonia have led to the addition of muscle and eye sensors to reach better accuracy. As we used this test, a detailed description of the features is shown in Appendix B.3 in Table B.4‡.

- **Actigraphy**† is used to evaluate sleep over a longer period. It's a body-worn activity monitoring device used to document physical movement associated with applications in physiological monitoring. The device is intended to monitor body movements during daily living and sleep. Numerous studies have shown a good correlation with PSG. These devices allow a rough but relatively accurate evaluation of TST, Time in Bed (TIB), and SOL with a sensitivity of 0.9 and an accuracy of 0.8 [111]. For WASO, actigraphy tends to under-evaluate the time spent awake [122, 111]. Depending on the type of actigraph used, we could extract different information about the activity level or the inter-day stability. As we used this test, a detailed description of the features is shown in Appendix B.3 in Table B.5‡.

  For more accurate results and better assessment, a **sleep diary**† is recommended, theoretically providing information on sleep times and any awakenings during the night. In addition, the sleep diary provides information on the patient's life during the observation period. It's a tool for day-to-day self-assessment of sleep perception over several days, usually one or two weeks. The patient annotates on a sleep log the primary information about their sleep, especially Bedtime, time in bed (TIB), TST, SOL, and WASO. Researchers have tested the ability of different thresholds of these quantitative parameters to predict Insomnia. SOL and WASO are the most used parameters to diagnose Insomnia.

  However, even before testing, numerous questionnaires have been developed thanks to the experience acquired by sleep researchers in the field of sleep disorders. We won't describe them all here, but only those most common that will be used in this thesis. The different questionnaires

Figure 2.14: Polysomnography

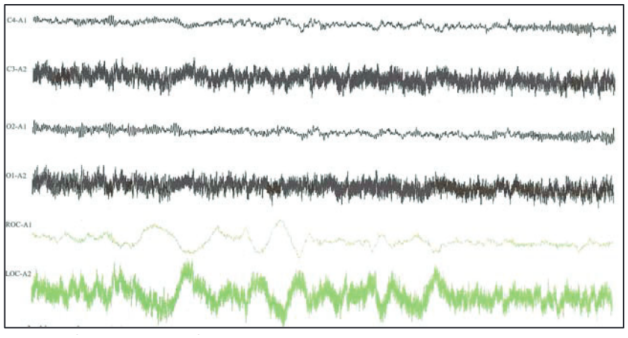


Figure 2.15: Noise secondary to partial detachment of the M2 electrode

could be general to assess sleeper typologies, global sleep normality, or specific symptoms like sleepiness. They could be more specific to understand disorders like sleep apnea syndrome or Restless Legs Syndrome. Additional questionnaires concerning frequent comorbidities associated with sleep disorders are used in clinical practice, such as personality, depression, or anxiety questionnaires. The benefits of all these validated questionnaires are to target the patient better, assess clinical and therapeutic progress, and compare studies with each other.

Transversal sleep questionnaires

- Describing extensively those questionnaires is useful to understand the meaning of the main features used in this work.
- **Horne and Ostberg questionnaire** [89], also known as the Morningness-Eveningness Questionnaire†. It's a self-assessment questionnaire designed to measure "morningness" or "eveningness", which is the preference for morning or evening activities.
- The Pittsburgh Sleep Quality Index Questionnaire is a subjective sleep self-report questionnaire over a one-month interval. The measure consists of seven items to calculate an overall score as an index of sleep quality and a screening of the major sleep disorder [27].
- The SLEEP-50 is designed to detect common sleep disorders in the general population. It could detect five sleep disorders: apnea, narcolepsy, restless legs/periodic limb movement disorder, circadian rhythm disorder, and sleepwalking [194].
- The **ESS**† [95] asks the respondent to rate their usual chances of dozing off or falling asleep. It's widely used to validate improvement in sleepiness. The higher the ESS score, the higher the average sleep propensity in daily life. A score of 11 or higher is usually significant for sleepiness. A study compared the ESS score between Insomniacs and the general population on more than 700 subjects and found a mean score of 9,4 ($\pm$4.65) for CID vs 8,05 ($\pm$3.86) [181].

Insomnia questionnaires

- **The Insomnia Severity Index (ISI)**† [65], is internationally used in Insomnia research for diagnostic, follow-up, and treatment outcome [10]. It is a 7-point self-report instrument commonly used in Insomnia research and clinics. A score of 10 or higher is very specific and sensitive in a community sample. In the clinical sample of Insomniacs, the results vary between studies, but the score is generally between 17.5 and 19 (SD$\pm$4) [41, 100]
- **Dysfunctional Beliefs and Attitudes about Sleep Scale**† [147] was designed to evaluate the belief about sleep and is also widely used in Insomnia assessment. It was designed to measure sleep-related cognitions. It consists of 30 questions intended to measure five dimensions:

  1. Misconceptions about the causes of Insomnia
  2. Misattributions or amplification of its consequences
  3. Unrealistic expectations
  4. Control, and predictability of sleep
  5. Dysfunctional beliefs about sleep-promoting practices.

  There is no clear cut-off score [147]. Other scales developed in Insomnia assessment were designed to evaluate a hyperarousal specific to bedtime. One is the Pre-Sleep Arousal Scale (PSAS) [205] created to measure somatic and cognitive arousal in the period that is right before sleep. This scale is highly correlated with stress measure [204]. A study shows that cognitive and physiological arousal is linked to sleep perception [196]. But in another study, pre-sleep worry (PSAS cognitive arousal) was associated with interpretational bias but not sleep misperception [72]. PSAS is correlated with DBAS and STAI in the 0.4 range [204, 72]. Another study shows that PSAS, ISI, and DBAS increased in the same proportion between normal sleepers and Insomniac patients [155].

Psychological questionnaires

- **Minnesota Multiphasic Personality Inventory 2 (MMPI2)**† [26]is a test widely used for general psychological assessment, as well as a treatment outcome predictor [1] and as psychometric evaluation in Insomnia [99]. It is a psychological test used to measure personality traits and psychopathology. Subjects are asked 567 true-or-false questions designed to detect various psychological problems. It also includes validity scales to determine whether a subject is willing to exaggerate or mask his psychological disorders. 120 scales and subscales are built from the initial questionnaire. The scales are built using T-score calculation. The T-score is derived from the raw scores obtained on the MMPI scales. It is calculated by converting the raw score into a standard score with a mean of 50 and a standard deviation of 10. The T-score allows for easy comparison of an individual's scores to a large normative sample. A T score above 70 is significant for a given scale, and the limit is between 65 and 70. As we used this test, a detailed description of the features is shown in Appendix B.3 in Table B.2‡.

- **Beck Depression Inventory-Second Edition (BDI2)**†is also widely used [14]. It contains 21 items that measure the severity of depressive symptomatology on a three-point scale (0 = absence of symptoms, 3 = most severe). Initially standardized for monitoring depression, the BDI2 is an effective psychometric instrument with standard cutoff scores to categorize depressive disorders. The total scores could be between 0 and 63, with proposals for increased severity:14–19 for mild, 20-28 for moderate" and 29–63 for severe. The cutoff is not homogeneous, but a score of 17 seems appropriate for significative mild depression with good specificity and sensitivity (81% and 79%) [207], especially in the Insomniac population [33].

- **State-Trait Anxiety Inventory Form (STAI-F)**† is also widely used in assessing anxiety as a state (present) or a trait (usual feeling)). The two forms include 20 self-descriptive statements on a four-point scale (1 = not at all, 4 = very much so) [193]. The score that differentiates significative anxiety in the Insomniac population seems to be around 47 for STAI-T and 55 for STAI-S when a score of 40.22 and 51 is not associated with significant stress or anxiety [119]

2.4.2 Paradoxical Insomnia: numerous propositions, low specificity

The last official definition of ParI could be found in the last ICSD-3 ([183]). ParI corresponds to underestimating real sleep time measured by PSG, sometimes with sub-estimates of sleep far from objective reality. The ICSD3 now recognizes that this characteristic is shared by most patients suffering from Insomnia, which tend to underestimate sleep duration and overestimate sleep latency and awakenings, unlike good sleepers. This subjective-objective mismatch is considered secondary to physiological hyperarousal and could be considered one of the main characteristics of Insomnia and no longer a subtype of Insomnia. There is no quantitative definition. This modest mention in the ICSD-3 should be seen in the context of the numerous publications trying to define it until recently. Indeed, ParI is perhaps the best example of the inability of the scientific community to harmonize definitions and diagnostic criteria to describe the same phenomenon until it disappears, drowned out by the diversity of definitions used. Apart from the fact that this disorder has changed its name greatly in the last few decades, it has been the subject of hundreds of publications purporting to study the populations concerned but with rarely identical definitions. In other words, to take a trivial example, it is like trying to categorize and understand obesity by systematically changing the definition and calculation of the body mass index. We will try to understand this paradox in many ways by first reviewing the evolution of the different classifications before reporting all the meaningful studies on the subject until 2023.

History of the definition

We will describe the history of this disorder by referring chronologically to the various classifications whose chronological appearance corresponds to the numbering below.

1. The first mention of this subtype appeared officially in 1979 in the Diagnostic Classification of Sleep and Arousal Disorders (first edition) [190] under the heading "Disorders of Initiating and Maintaining Sleep (DIMS or Insomnias) Complaint without Objective Findings". This name has the merit of being a definition in itself. The subjective nature of Insomnia is already at the heart of this definition with an idea of the acceptable thresholds that can be used to define a

"good objective sleep", namely a sleep latency (SL) of fewer than 20 minutes and a TST of more than 6.5 hours. One aspect of the diagnosis at the time concerns the honesty of the complaint, which is associated with the notion of "lack of psychopathology". It is even made a warning about the secondary benefits of this complaint in some cases It is necessary to separate the subjective DIMS patients from malingerers who claim they sleep poorly to obtain drugs and for other reasons. Therefore, from its first classification, this subtype already raises controversial questions. The adjective "paradoxical" appears in the text along with "perplexing" condition but not in the diagnosis. This condition is evaluated at 25% of all Insomnia complaints. This term emphasized the declarative aspect of this subtype of Insomnia compared to an objective examination such as PSG.

2. The term was then changed to nSSM in the first classification of sleep disorders in 1990. However, the definition is globally the same. Thus, the criteria include a complaint of Insomnia with sleep of normal duration and quality on PSG examination (PSG demonstrates a normal sleep pattern, with sleep latencies of less than 15 to 20 minutes, and sleep durations greater than 6,5 hours), without objective sleepiness, and without another disorder that could explain it.

3. Then, finally, this subtype took the name of ParI in the ICSD-2 in 2005, and the objective PSG parameters disappeared from the Diagnostic criteria. The latest officially recognized diagnostic categorization is presented in table 2.5

ICSD-2 Paradoxical Insomnia
A Patient's symptoms meet the criteria for Insomnia
B The Insomnia is present for at least one month
C One or more of the following criteria apply:
i: The patient reports a chronic pattern of little or no sleep most nights; with rare nights during which a relatively normal amount of sleep is obtained.;
ii: Sleep log data during one or more weeks of monitoring shows an average sleep time below the published age-adjusted normative value, often with no Sleep at all indicated for several nights per week and no nap;
iii The patient shows a consistently marked mismatch between objective findings from PSG or actigraphy and subjective sleep estimates derived from self-report;
D At least one of the following is observed:
i: The patient reports near-constant awareness of environmental stimuli throughout most nights;
ii:The patient reports a pattern of conscious thoughts or rumination throughout most nights.
E The daytime impairment reported is consistent with that reported by other insomnia subtypes, but it is much less severe than expected given the extreme level of sleep deprivation reported.
F No other disorders explain these symptoms better.

Table 2.5: ICSD-2 diagnostic criteria: All the capitalized criteria are mandatory to diagnose ParI

Nowadays, the community remains divided on this topic, with one side believing that the underestimation of sleep duration is a characteristic of all Insomniac patients, which led to its removal from the latest classification of sleep disorders. The other part, of which we are part, believes it is a clinical entity in its own right, for which the objective criteria for detecting and explaining it are not yet known. But diagnosing and explaining are two distinct issues that are sometimes mixed in some papers.

Before reviewing the main studies that have attempted to determine what ParIis, it is worth describing the tools used in most studies to better understand their scope and the features used.

From the extensive work of [35], we added new definitions published since then, extracted from 44 articles giving a quantification of ParI. This relative freedom of interpretation of the diagnostic criteria regarding the intensity of poor sleep perception has opened the way to a kind of competition between research teams to find the best diagnostic criterion supposed to define ParI, this is all the more true since several hypotheses have not yet been determined such as

1. Misperception of sleep as wakefulness;

2. Anxiety and selective attention to sleep-related threats;

3. Presence of brief awakenings;

4. Local arousal and local sleep [35].

We will describe the different definitions found in this study with another published since then with the terminology used in this review. Then, all reproducible definitions from 1979 to 2019 were labeled with a different alphabet letter for each different one [35]. For articles published later, we use the same procedure as that article by adding letters to designate them (we just added the Z). It should be noted that this alphabetical classification is not chronological since several articles published at different dates may refer to a particular definition. We can see in Table 2.6 the formula used and the studied sample's prevalence by publication.

2.4.3 Attempts to find biomarkers of ParI

Attempts to find specific biomarkers of ParI have mainly focused on analyzing brain waves during sleep to find proof of brain hyperarousability.

The most promising attempts at explaining ParI come from EEG spectral analysis trying to find some brain signature related to sleep misperception. Concerning the EEG spectral analyses, many studies have been interested in this aspect with the hypothesis that Insomnia could correspond to a cortical hyperarousal, which would be translated by a decrease in the power of the spectral waves of low frequencies (Delta, Theta) to the benefit of high-frequency waves of Beta type. This has been shown between chronic Insomniacs (not specifically ParI) and control subjects [139, 163]. However, studies have focused on the difference between Insomniacs with poor sleep perception (ParI or Subjective Insomnia) and other primary Insomniacs (Objective Insomnia or Psychophysiological Insomnia). One of them, [108], which corresponds to the J formula (see Table 2.6, evaluated the spectral analysis between subjective and objective Insomnia and found a link between the degree of objective-subjective sleep discrepancy and N-REM EEG relative delta activity, and an association with elevated high-frequency relative activity. Greater Delta relative power was associated with higher sleep-quality ratings, and greater relative N-REM alpha power was predictive of less TST. This study also evaluated these differences according to scaled cutoffs for the TST and the percentage of bad perception. Paradoxically, they found a loss of this difference for Delta rhythms when the percent underestimation of sleep is below 10%, which seems counter-intuitive if it is a marker of hyperarousability; in fact, one would expect an increase. It is even worse for beta rhythms since there is no significant relationship with TST underestimation. In the study [94], with the base for Formula Q, they compared a control group to ParI and PsyI on two nights. They found, on the totality of the analyzed sleep period, absolute delta activity at C3 and P3 in ParI compared with PsyI. They didn't find significant results in the Beta activity or gamma activity. But, they found higher sigma activity at P3 in N-REM sleep. It's interesting because it's the same frequency as spindles, so spindles could be of interest as a marker of ParI (this point will be discussed in Section 4.2). Also, absolute alpha activity was higher in ParI than PsyI at P3. However, the differences between groups in absolute did not translate in the sleep macro-structure (similar in duration and time spent in all sleep stages). The last significant study on the subject [117], which corresponds to formula S, selected a subgroup of ParI subjects with an underestimation of more than 40% compared to the norm in a cohort of 2092 participants. It should be noted that the authors use the term underestimator and not ParI, although they have a higher score on the PSQI (Pittsburgh Sleep Quality Index) scale and the DFA questionnaire (Difficulties Falling Asleep in < 30 minutes) and on the FNA questionnaire (Frequency of Nocturnal Awakenings). They found, with EEG acquired with a 256-channel system, that individuals who underestimate their total sleep time display a more 'wake-like', activated EEG (higher relative power in the beta band in central regions in N-REM sleep) and lower relative power in the delta band over the right frontal electrode in both N-REM sleep (N2-N3 combined and N2) and REM sleep, while opposite changes are observed in REM sleep in subjects who overestimate their sleep time. The authors suggested that Insomnia patients may correctly perceive subtle shifts toward wake-like brain activity. At the same time, they didn't find a higher number of scorable arousal than normal sleepers. Therefore, in the end, the results are contradictory according to the studies and the definitions, and the samples remain relatively small. It is, therefore, still difficult to explain or predict the intensity of ParI according to

23 different definitions of ParI published with sample prevalence (Prev)				
Name	**Formula Calcul**	**Prev**	**N**	**Ref**
A-1N	sSOL/oSOL > 1.5	60%	29	1979 [22]
B-3N	oSE > 90% AND oN2SL < 30 min	NA	16	1985 [195]
C-2N	sSOL < 30 min AND oSE > 87%	50%	16	1989 [109]
D-3N +7acti	oSE > 85% AND oSOL < 40	22%	36	1992 [85]
E-2N*3	oSE >90% AND oTST-sTST >=60 min	33%	21	1992 [180]
F	oSE > 80% AND (sSOL−oSOL)/oSOL > 0.2 AND (oTST-sTST)/oTST>=0.2	25%	28	1995 [138]
G-2N	sWASO >40 min AND sSOL >45 minAND oSOL <30min AND oSOL <30minAND oSE >90% AND oSE/sSE >2 AND oSOL >20 min	NA	18	1997 [21]
H	sSOL/oN2SL >1.5	50%	18	1997 [55]
J	oTST >390 min AND oSE >85% AND sTST <390 min (at home)	25%	57	2002 [177]
K	oTST >390 min AND oSOL <30 min AND oTST-sTST >120 min AND sSOL/oSOL >120%	NA	20	2012 [160]
L	oTST-sTST >120 min	NA	159	2010 [131]
L2	oTST-sTST >=60 min	NA	142	2011 [62]
M	(oTST-sTST)/oTST >= 0.9 AND oTST > 120 min	17%	159	2010 [131]
N	oSE > 85%	45%	205	2011 [62]
0	oTST >360 min	55%	444	2010 [61]
P	oTST >360 min AND oSE >85% AND oTST-sTST >60 min OR sSE-oSE >=15%	36%	112	2010 [103]
Q-3N	oTST >380 min OR oSE >= 80% AND sSOL−oSOL >= 60 min OR oTST-sTST >= 60 min OR oSE-sSE >= 15% 3n	NA	87	2013 [94]
R	oTST >390 min AND oSE >= 85% AND sSE-oSE >= 15% AND oTST-sTST >= 60 min 4n	NA	58	2013 [11]
S	(sTST/oTST) * 100 < 58.8%	NA	NA	1979 [195]
T	oTST >= 390 min AND oSE >= 85%	26%	250	2015 [144]
U	sSOL−oSOL (no cut-off)	NA	32	2015 [102]
V	300 min <oTST <600 min + oSOL <30 min + 50 min < REML <100 min + 55% <N1 + N2 <60% + 15% <N3 <25% + 15% <REM <25% + WASO <5% TST OR <30 wake ep	52%	255	2018 [210]
Z	oSOL ≤ 30 min AND WASO ≤ 30 min AND TST ≥ 360 min	18%	335	2020 [3]

Table 2.6: Main formulas proposed across publications. Each letter corresponds to a specific calculation explained in the "Formula Calcul" column. The prevalence, when available in the paper, is provided. N corresponds to the number of subjects in each study, and N corresponds to the number of sleep laboratory night recordings

the spectral analysis; good quality studies on many subjects with a homogeneous definition would be necessary.

However, all studies investigating sigma rhythms (11-14 Hz) have increased ParI. This frequency band corresponds to the frequency of spindles. Studies have specifically addressed this issue. In 2016, [152] compared ParI, PsyI, and Normal Sleeper. They used an algorithm for spindle detection [143]. They found that the duration of sleep spindles was shorter for ParI (mean shorter than good sleepers but not from PSY-I). Other characteristics were not significantly different. But surprisingly, in these studies, the subjects labeled PsyI had a longer TST than good sleepers and ParI, which is the opposite

of most studies on ParI. Moreover, the validity of the spindles detection algorithm is not discussed, nor is the formula validity. In 2020, a group [3] used spindles analysis and ML tools to evaluate chronic Insomnia and ParI. This study is the reference for the Z formula. They studied 288 CID patients (59 ParI). The different PSG features (hypnogram, EEG spectrum, and sleep microstructure) were extracted (slow waves, spindles). To detect slow waves and spindles, the algorithm is not mentioned. They used supervised (see definition in B.1.2) algorithms to differentiate the patients. They found that sleep spindles have reduced amplitude and increased frequency and duration in all Insomnia patients but no specificity concerning ParI. A more recent study studied spindles and personality traits in ParI compared to Healthy Subjects. The formula used is not clear. All subjects corresponding to three different definitions were included (at least one of the following equations should be met: E, L2, or M). The analysis of sleep spindles was conducted only onN-REM 2 sleep stages; artifact-free-epochs were selected at first and the last sleep cycle of the night on C3 and F4. The detection was made by a semi-automatic algorithm (not described) and revised by a Sleep Expert. They didn't find a significant difference in duration and frequency, but the density was significantly decreased in ParI. On the other hand, the duration of sleep spindles showed positive correlations with the extroversion dimension scores [188].

A study evaluated subjective and objective sleep features and psychometric measures in patients with primary Insomnia. The main objective was to find, in a population of primary Insomnia, one or more psychometric measures that could be correlated with sleep perception [54]. They analyzed nSSM as a dimensional value and confirm that a variable degree of misperception is observed in most patients without any clinical or psychometric measure that could differentiate these Insomnia subtypes. But they found that lower scores on the Self-Administered Anxiety scale were associated with nSSM. This finding diverged from precedent studies [197, 62] where the subjects with nSSM presented more anxiety on the subjective scales.

A recent study [124] tried to link poor sleep perception with chronic anxiety. In this study, all 305 patients were diagnosed with anxiety-related disorders. Their formula corresponds to the L2 formula described in Table 2.6. Patients were divided into normal sleep perception, positive sleep perception abnormality, and negative sleep perception (underestimation of TST $>$ 60 min). The PSG indicators significantly related to negative sleep perception (50% of the samples) compared to normal perception were an increase in objective TST, in the total number of awakenings, in spontaneous arousal times, and a decrease in Sleep Latency and WASO. In that study, 67.3% of patients with ParI, according to the L2 formula group, had used sedative-hypnotic drugs. Unfortunately, they didn't assess the level of anxiety during the study.

2.5 Launching this Thesis

In conclusion, after reviewing the different methods and explanations for ParI, it isn't easy to understand this concept accurately, so great are the divergences between studies. The main confounding factor is the choice of starting definition for the studies. As we have just seen, most studies we have cited use different formulas, making comparisons difficult. This is the purpose of [35] comparative study, which tested these different formulas on a data set of 200 Insomniacs subjects and controls to show the agreement between most of the formulas. In Figure 2.16, the agreement calculation on 16 formulas showed clearly the vast majority of no concordance between formulas. This shows that the first step before trying to explain a phenomenon is to try to define it.

They also concluded that the current state of the art indicates that TST should be preferred to define Paradoxical Insomnia rather than SOL, but above all, that evidence-based knowledge on Paradoxical Insomnia should be obtained with larger case-control observational studies assessing multiple subjective and objective sleep parameters (not only TST and SOL). Furthermore, they proposed adding sleep logs and prolonged actigraphy before the recording nights. All these recommendations aim to have as much available data as possible to achieve this difficult objective, define Paradoxical Insomnia.

It's exactly what we had in mind with our database collection to bring new insight into this subject.

But why is it so important to define ParI patient? The first reason concerns the patient's and the doctor's behavior according to the category type. When a patient is labeled ParI, they would likely be

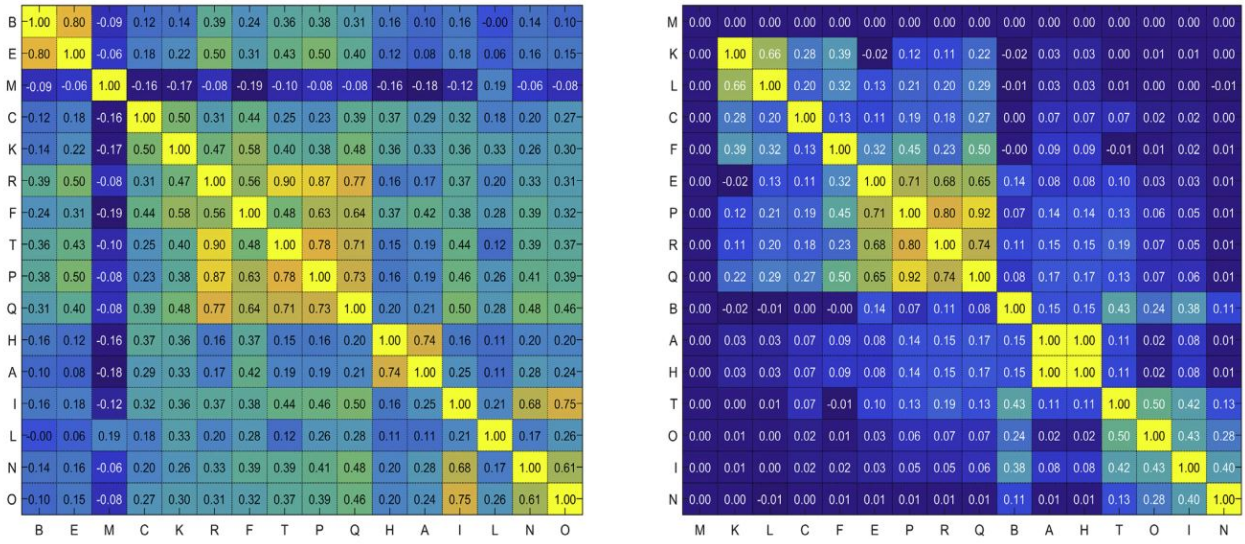


Figure 2.16: Poor agreement between formulas on the same dataset, on clinical(N=200, left panel) or controls (N=200, right panel) [35]

categorized as a complicated patient, or worse, a malingerer with all the negative a priori that could imply. This categorization can directly impact the therapeutic relationship or management, especially if the therapeutic response is negative or the patient already has ineffective treatments. As a second argument, we saw that different studies using causal inference protocol, assuming that the definition they chose is accurate, tried to explain the ParI with costly studies with EEG or MRI, without a discussion about the reality of the concept. In that sense, it must be wiser, as in the last studies we mentioned [117, 124], to use a kind of spectrum of nSSM instead of a categorial diagnosis, in the ideal case correlated with the intensity of the complaint to find the best cut off. But to do that, we need to tackle the high complexity and interrelation of all the available characteristics for these subjects without reproducing the existing ones to bring a new light on this topic. We will see in the next chapter why MLcould be a good choice to tackle this problem and the main algorithm that could be used.

Part II: Machine Learning for Applied Medicine

Chapter Highlights (Part II: Machine Learning)

1. Why Machine Learning? ML algorithms are designed to learn complex, meaningful relationships between variables to apprehend what defines the best target to be predicted under the defined model parameters. In our case, it is our purpose to know which variables could predict subjects with ParI.

2. Frequentist Statistics versus Machine Learning Frequentist Statistic FS is a paradigm designed to capture elements in a sample to reveal an explication about the target population (inference). ML is about prediction and refers to using a trained model to estimate or forecast an output value given some input data by learning from the data. We use ML methodologies to find new variables explaining ParI.

3. Considerations of Sample Size: In Defense of 'Small' Data We will focus on linear and classifier models widely used on medical datasets.

4. Data Mining and Exploratory Data Analysis (EDA) We will describe the techniques and algorithms used in this first step, also called Exploration Data Analysis, like non-supervised pattern detection or clustering methods.

5. Building Predictive (ML) Models We will cover the building of predictive models. Predictive models can be used as simple tools to guide a human diagnostic (indeed, this is our primary consideration); or as *prescriptive models*, i.e., the model's output will be taken as a recommendation.

6. In this thesis: ML as a tool In this thesis, we use ML to discover new relationships, in conjunction with expert medical knowledge, to reveal new insights about ParI.

Key Terms and concepts

Acronym/term	Definition	Ref.
DT	Decision Trees	p. 172 (B.1.2)
EDA	Exploratory Data Analysis	p. 45 (2.9)
KNN	K-Nearest Neighbors	p. 50 (2.10.1)
KMEANS	K-Means Clustering	p. 46 (2.9.3)
LASSO	Least Absolute Shrinkage and Selection Operator	p. 173 (B.1.2)
LR	Logistic Regression	p. 50 (2.10.1)
PCA	Principal Component Analysis	p. 47 (2.9.4)
OLS	Ordinary Least Squares	p. 39 (2.7)
RF	Random Forest	p. 49 (2.10.1)
SVM	Support Vector Machine	p. 50 (2.10.1)
tSNE	t-Distributed Stochastic Neighbor Embedding	p. 47 (2.9.5)

2.6 Why Machine Learning?

In this thesis, we use ML tools to answer our research questions. This choice was made to complement frequentist (classic) statistics. We aim to use complementary tools to gain the fullest possible understanding of the data. After explaining the difference and the complementarity of these two approaches (see Table 2.7 for the synthesis), in Part II of this current chapter, we describe some of these tools for medical experts who might be unfamiliar with them. We also refer this reader to our brief published review on the subject written for physicians for a global view of Artificial Intelligence and ML. [157].

2.7 Frequentist Statistics versus Machine Learning

The classical statistic is typically focused on obtaining an estimate of some parameter of a **population** (e.g., of *all* people with chronic Insomnia) based on data from a **sample** (e.g., diagnosed Insomniacs attended at a particular clinic). Part of classical statistics is frequentist statistics; the parameters are fixed, and the data are random, meaning the parameters are not considered random variables. This differs from Bayesian statistics, which allow for parameter probability distributions. Frequentist statistic is the most used in medical research. Indeed, we are referring to a paradigm about the definition of probability as the long-run frequency of an event occurring when an experiment is repeated an infinite number of times under identical conditions [38]. This branch of statistics focuses on data collection, analysis, interpretation, presentation, and organization. This interpretation is aligned with the practical, empirical view of probability used in many parts of science and medicine, for example, in a clinical trial [116].

A typical example in the field of sleep medicine could be the average hours of sleep per night,

Let's say that the parameter of interest is y (hours of sleep); statistical tests can be employed to estimate how close the estimated value is to the true value (which we can never have because it is not possible to accurately and unambiguously survey every single possible patient). Ordinary Least Squares (OLS) could be used to estimate the unknown parameters in a linear regression model.

From the point of view of **OLS** regression, we assume the true model is:

$$y_i = f(x_{i,1}, x_{i,2}; \theta) + \epsilon = \theta_1 x_{i,1} + \theta_2 x_{i,2} + \epsilon$$

where $x_{i,1}$ describes the 1-st **feature** of the i-th patient (e.g., frequency of spindles they experience per night); and y_i some aspect about this patient that we are interested in estimating (e.g., number of hours sleeping per night). The ϵ term refers to the **irreducible Bayes error**, which indicates that even the best possible diagnosis (even by a medical expert) based on the observations of these two features may not be correct.

So in statistics, one wishes to infer (i.e., perform **inference** of) **parameters**, producing estimates of $\hat{\theta}_1$ and $\hat{\theta}_2$ which defines our [estimated] model. These θ tell not only the function of that feature but also its relative **importance** (for estimating y). Once again, statistical tests are typically implied to obtain confidence regarding how close estimated $\hat{\theta}$ is to true [hypothetical] θ.

However, all this is based on major *assumptions* of the true concept (i.e., of what θ represents here, such as linearity, feature independence), including that there is a linear relationship between features $x_{i,j}$ and the target variable y_i. Such assumptions, more often than, do not hold in the real world. Some key concepts are:

- The population follows a predetermined distribution (Parametric models).
- The formulation of hypothesis testing to reject the null hypothesis (statement of no effect or difference) through the available data.
- After the use of a statistical test (t-test for comparing the means of two groups, ANOVA for several groups, Chi-Square test for categorical variables, etc..), the probability (p-value) of observing major difference from what we would expect under the null hypothesis is calculated, and in general in medicine, if the p-value is small (typically, less than 0.05), then the evidence against the null hypothesis is strong.

- To estimate the range within which the true population parameter lies with a certain confidence level, Confidence Intervals are used [38].

- The main goal of all these processes, except comparing differences between groups or clinical trials, is to capture some significant explicative elements in a sample of a given population to reveal some characteristic properties of the larger population it is supposed to represent. This goal is named **inference**, typically involving regression models, point estimation, and hypothesis testing [28].

In the medical field, this paradigm is especially fitted to evaluate the relevance of each candidate variable associated with each patient in its possibility to affect the outcome of interest. Most of the time, the variable was chosen based on existing knowledge. Then, this approach, in the context of rigorous experimental control like two-sample hypothesis testing, is the routine academic statistics for small to medium datasets [59].

Of course, this model could be very effective in establishing bio/physiological effects that provide insight into what leads or not to some disorders at the population level.

For example, in [6], to test if hypocretin deficiency is associated with abnormally low serum leptin levels as a consequence of increased Body Mass Index in type 1 Narcolepsy (see definition in B.1.1), they used Chi-2 tests and ANOVA to compare respectively categorical and continuous variables between groups. They used another ANOVA to identify determinant factors when a significant difference was found. In the end, they performed a linear regression. The significance level was < 0.05 for variance and linear regression analysis and < 0.01 for post hoc tests. This rigorous statistical protocol is a good example of the traditional frequentist statistic. So, most of the time and for the big discovery, medical experts, with access to these traditional statistical tools, provide the most significant insight into specific problems and can still outperform AI on most of the diagnoses and treatment procedures in most medical areas.

So, as there are already efficient statistics tools, what is the point of using ML tools without a clear experimental protocol? One of the reasons is that the exponential increase in computing power (Moore's Law) is at the root of the need to use ML. Indeed, regarding patient analytics, the expert task (in discovering and providing key insights) becomes impossibly lengthy due to the increasing amount and complexity of data provided by increasingly sophisticated monitoring.

At the same time, statisticians and computer scientists have developed many different methods in the last decades, and aligned with the great increase in computing power and electronic data collection, some tools designed initially for computer science issues started to be more suitable and usable for medical studies involving more data [64]. So ML is to be used as a set of investigative tools to assist, both in practice and research, the expert and certainly not a replacement of this expert, much more an assisting tool.

Let us consider an example: prediction versus inference. The interest in ML from the research point of view and its ability to discover new knowledge, even on an already studied dataset by using frequentist statistics, was addressed by [28]. Yet, in a publication in 2019, [29], they discussed this issue specifically in neuroscience after a major experiment in 2018 [30]. The main debate in these papers is not about the different statistical approaches or tools but about the sight of the goal to be achieved. The typical inferential approach wants to tackle, for example, which specific gene could impact or partly explain Narcolepsy [106] or an epileptic syndrome [20]. In the **prediction** case, we want to know which gene locations are collectively useful to discriminate subjects with or without diseases. Indeed, this is the key point: faced with a data set, do we want to move towards inferences or predictions? Machine Learning (ML), on the other hand, also produces a model, $\hat{\theta}$, from a sample (**training data**, in ML terminology) but is more concerned with the **accuracy** of **predictions**, $\hat{y} = f(x_1, x_2; \hat{\theta})$ from that model; or a **loss metric** (less is better; i.e., inverse to accuracy) $L(y, \hat{y}; \hat{\theta})$; and in many cases also the **confidence** in those predictions. Many more complex models overcome some of the simplistic assumptions of least squares (several of which we use and review below) but at the cost of increased model complexity ($\hat{\theta}$ maybe hundreds, millions, or billions of dimensions, intricately connected in diverse ways).

The challenge of ML is the fact that the **test instances** x (e.g., new patients) have *not been seen by the model before*; therefore, the model must generalize (in ML terminology: avoid **overfitting** the training data). Another important challenge, particularly in the context of domains such as medicine,

beyond simply better predictive performance (low $L(y, \hat{y}; \hat{\theta})$, is **interpreting** the model (which can be very complex) and **explaining** how it made such a prediction; in a way which is accessible for a human domain-expert (medical doctor, in our case).

These concepts are covered in standard ML textbooks, e.g., [84].

Comparison of Machine Learning Prediction and Frequentist Inference		
	Machine Learning (Prediction)	Frequentist Statistics (Inference)
Objective	Focus on predicting output y for new input x	Focus on drawing conclusions about population parameters
Process	Training a model on labeled data to learn the relationship between x and y	Formulating a hypothesis, collecting data, and using statistical tests for inference
Evaluation Metrics	Accuracy, precision, recall, mean squared error	P-values, confidence intervals, statistical significance
Nature of Data	Handles complex, high-dimensional data; captures complex nonlinear relationships	Deals more with structured data and linear relationships; focuses on causality
Goal	Making accurate predictions for new data	Making conclusions about population parameters based on sample data
Data Utilization	Data-driven, uses complex algorithms for high-dimensional data	Focuses on testing hypotheses and estimating parameters reliably
Nature of Analysis	May not provide insights into causality or relationships between variables	Aims to understand relationships and cause-and-effect
Uncertainty Representation	Prediction intervals or probabilities	Confidence intervals and p-values
Complementarity	ML can enhance predictive accuracy using complex data, while frequentist methods can provide robust statistical inference to validate ML predictions	Frequentist methods can benefit from ML's data-driven approaches for exploratory analysis and prediction, while ML can incorporate frequentist techniques for hypothesis testing and validation

Table 2.7: Comparison of Machine Learning Prediction and Frequentist Inference [29].

In ML, **prediction** refers to using a trained model to estimate or forecast an output value given some input data. This process involves leveraging patterns and relationships learned from the training data to make informed predictions on new, unseen data points.

In the context ofsupervised learning, where we have labeled training data consisting of input-output pairs (x, y) (let us recall that x may be multi-dimensional), the typical prediction process involves fitting a model to the training data and then using that trained model to predict the output (y) for new input data (x).

Mathematically, we can represent the prediction process as follows:

Training Phase

So, given a training dataset with n samples:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

We aim to find a model approximating the underlying mapping between input (x) and output (y).

Let's denote this model as[1] $h(x;\theta)$, where θ represents the model parameters.

The model is typically defined as a function that takes input features (x) and produces an output prediction ($\hat{y}$).

The goal is to find the optimal values of the model parameters (θ) that minimize the discrepancy between the predicted outputs ($\hat{y}$) and the true outputs (y) in the training data. This is often done by minimizing a loss function, denoted as $L(y, \hat{y}; \theta)$.

The training phase typically involves an optimization algorithm to update the model parameters and minimize the loss function iteratively.

The training phase aims to find the optimal parameters that minimize the average loss over the training data:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} L(y_i, h(x_i; \theta))$$

Prediction Phase

After the model is trained, we can use it to predict the output ($\hat{y}$) for new, unseen input data (x).

Given a new input sample (x_{n+1}), we use the trained model $h(x;\hat{\theta})$ to estimate the corresponding output ($\hat{y}_{n+1}$):

$$\hat{y}_{n+1} = h(X_{n+1}; \hat{\theta})$$

The predicted output ($\hat{y}_{n+1}$) is the model's estimation based on the learned patterns and relationships from the training data.

The prediction process can differ depending on the ML algorithm used. Here are a few common approaches:

Regression: In regression, the goal is to predict a continuous numeric outcome. The model aims to learn a function $f(x;\theta)$ to predict a real-valued output y for a given input x. The model is trained by minimizing the discrepancy between the predicted and true values of y. This can be accomplished by optimizing a loss function, such as Mean Squared Error (MSE). In medical applications, a regression can predict a patient's blood pressure based on various factors such as age, weight, diet, exercise habits, and genetics. This can be useful for determining the risk of conditions like hypertension and heart disease. In that sense, the medical community is already using the spirit of ML massively, for example, when risk factors or predictors are calculated on a big cohort (Like the Framingham risk score (1998) [212]). We can see in Figure 2.19 that this type of work (red circles 19) is the first level in ML using computer power, on four. Each level corresponds to an increase in computer assistance and a decrease in human action [13].

Classification: Unlike regression, classification predicts categorical outcomes. The function $h(x;\theta)$ maps input features to discrete classes. The loss function, often a cross-entropy loss for multiclass classification problems, measures the difference between the true class labels and the model's predictions. A common application of classification in medicine is disease diagnosis. For instance, based on symptoms and medical test results, an ML model can be trained to classify whether a patient has a certain disease. Another example would be categorizing tumors as benign or malignant based on characteristics such as size, shape, and growth rate.

Time Series Forecasting: Time series forecasting uses historical data to predict future values. Here, the input x typically includes data points from previous time steps, and the output y is the predicted value for the next time step. Many algorithms used for regression can be used for time series forecasting, with modifications to handle temporal dependencies. It could be applied to predict the spread of an infectious disease over time. By training a model on past infection rates, the model can forecast future trends and help public health officials prepare and respond more effectively. Similarly, in individual patient care, time series analysis can be used to forecast the progression of chronic diseases, helping healthcare providers to make more informed treatment decisions.

Anomaly Detection: Anomaly detection is a bit different, as it focuses on identifying abnormal or unusual data points in the dataset. In this scenario, the function $h(x;\theta)$ learns to capture the

[1] Typically f for regression; and h for a classification model – but this is just a question of notation)

'normal' patterns in the data, and any deviation from these patterns is considered an anomaly. The output y is often binary, indicating whether an instance is normal or an anomaly. This algorithm could be very useful in medical imaging to identify unusual patterns that may indicate disease. For example, a model could be trained to recognize normal brain scans and then used to detect stroke in new scans.

Despite their differences, all these approaches follow the same fundamental process: they train a model on existing data and then use that trained model to make predictions on new, unseen data. By minimizing a loss function, the model learns to find patterns in the input data that can help it make accurate predictions. ** In our work, we will mainly use the classification algorithm.

2.8 Considerations of Sample Size: In Defense of 'Small' Data

It is important to note that the predictive accuracy relies heavily on the quantity and quality of the training data (including the feature engineering process) as much as the choice of model.

Authors of [30] systematically investigated more than 100,000 simulated datasets, with n between 10 and 100000, and the number of features between 1 and 40, changing characteristics to compare models for inference and prediction. For inference, they used ordinary linear regression; for prediction, they also took a linear model LASSO (cf. B.1.2) on identical datasets. So, the interest of this experiment is comparing the ability to recover the "meaningful" variables simultaneously. They evaluated the significance of the subset of correctly detected variables with OLS and the positivity of the LASSO coefficients. These metrics allowed the comparison of the number of correctly identified variables, analogously for OLS and LASSO. In Figure 2.17, we can see the correlation between the inference and the prediction on recovering relevant variables across the different dataset scenarios. We can see the disagreement in many cases, especially on small datasets (<100 samples). Interestingly, for many datasets, there is a poor correlation, with the correlation increasing in proportion to the number of samples (yellow), but even for high numbers (10 to 100000), the results are not the same. Thus, they could demonstrate that **diverging conclusions can emerge from the same dataset even with both linear models, which implies that the meaningful variables detected are not the same**.

2.8.1 Low patient count (n) implies neither small nor meaningless data

Our data collected (which we describe in intricate detail in the next chapter) is relatively small (in the big data era) but only in terms of the number of samples n (between 300 and 1000), whereas the number of features is large, ranging easily into the thousands (depending on how many features are extracted from the signals data per patient). Furthermore, we meticulously curated our datasets to ensure they are clean and representative of a typical Insomniac *population*. Except for the recent trend of deep learning, our collection might be considered a relatively 'standard' size, indeed greater in dimension than many benchmark datasets like the Montreal dataset, for example, [153].

In Figure 2.17, we can see the influence of the sample size on the results. The triangles represent the different data sets, their size, the number of significant variables, and their color corresponds to the number of samples; the upward point means that the LASSO found more significant variables than OLS, and the rightward point is the opposite. We've added a few indications to this graph. Firstly, we've plotted the color spectrum (purple gradient with a red line) for the 'standard sample size' from 350 to 1000. The aim is to visualize better the likely corresponding recovery rate scores for LASSO and OLS (between 0.4 and 0.6). The red circle represents the most likely maximum given this gradient, i.e. 0.8 for OLS and 0.95 for LASSO. LASSO performed better with smaller sample sizes and more meaningful variables. At the same time, OLS tended to be more successful at recovering important variables with larger sample sizes and smaller numbers of relevant variables. They also concluded that even small predictive performances typically coincided with finding underlying significant statistical relationships in most cases when even statistically strong associations with very low p values often shed only modest light on their value for the goal of prediction. However, this figure also shows us the possibility of having low scores with high sample sizes, showing that understanding predictions and choosing the right algorithms is just as important as sample size and that for each dataset, pre-

litigation work to select the most suitable algorithms is essential and constitutes the art of the data scientist.

2.8.2 Choosing a suitable ML framework

According to [28], given that our dataset could be at a maximum of one thousand samples with numerous explicative variables, we have a high probability of getting significant results, especially with LASSO techniques. Indeed, LASSO used L1 regularization that promotes sparsity by driving some coefficients to zero, selecting a subset of the most important features, and effectively eliminating the other from the model [199].

Other studies like [79] had tested three commonly used classifiers (see definition in B.1.2): They compared performance by changing the feature's sample size and signal-to-noise levels. They also evaluated the effect of non-Gaussian (skewed) feature distributions, the correlation between biomarkers, the imbalance in class distribution, and the choice of metric for quantifying classifier performance. In one of the experiments, they tested a binary prediction in a Gaussian distribution biomedical dataset with 10 to 50 meaningful biomarkers and 990 to 950 noisy features (like at random) on samples from 100 to 400. They showed that the most important was the percentage of meaningful biomarkers more than the sample size, even with only 150 samples. These results align with [28] on different algorithms. In our case, several samples with hundreds of features are available for binary classification, and all the features are possibly linked to the class prediction. The main classifiers corresponding to this task, LASSO, RF, SVM, and KNN are described in Table 2.8. These methods can be effective for medical datasets, where accuracy and interpretability are critical. However, the dataset's specific characteristics, like the features' nature, the prevalence of noise, and the need for interpretability, should guide the final choice of the classifier.

Classifier	**Main characteristics**	**Limitations**
LASSO	Useful in high-dimensional datasets for feature selection. Able to perform both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model it produces. With many features, it can be useful in identifying the most relevant features, potentially aiding in binary classification tasks [199].	Important to consider the interpretability and relevance of selected features [199].
RF	Handles many features well; generally effective in binary classification. The performance of RF was observed to improve with an increasing number of features, indicating its suitability for datasets with a large feature set [172].	Potential biases in variable selection; interpretability can be a challenge [172].
SVM	Effective in binary classification, especially with a clear margin of separation. In medical datasets, where separability might not always be clear, SVM's soft margin approach can be beneficial [135].	In cases of unclear separability, it requires careful tuning of the soft margin [135].
KNN	Simple and flexible, suitable for a variety of datasets.	Performance can degrade with noisy/irrelevant features; requires careful choice of 'k' and feature scaling [142].

Table 2.8: Suitability of Various Classifiers for Binary Classification in Medical Datasets

Common methods include **RF, SVM, and KNN** (description in the next Section).

The predictive approach is fully compatible with our goal, as we just showed. This choice is further strengthened by the fact that the ParI diagnosis or even the treatment outcome does not meet stable criteria, as we have seen in 2.4, and so the question that arises first is who belongs to the ParI whatever the criteria involved more than what is the factor that explains ParI.

But in this case, the logical question is how to predict something that isn't clearly defined. This is where ML tools themselves would have a limit because they can only predict what has been learned,

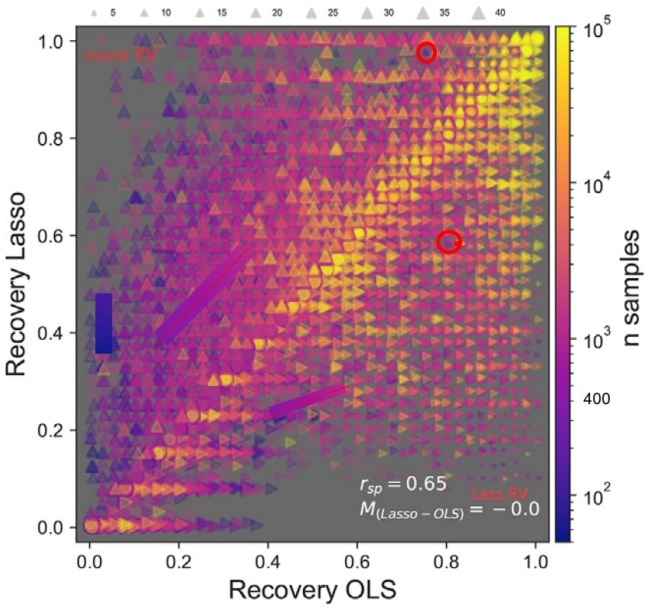


Figure 2.17: Difference in meaningful variables recovery between Statistical Inference tested with OLS and LASSO algorithm [28]

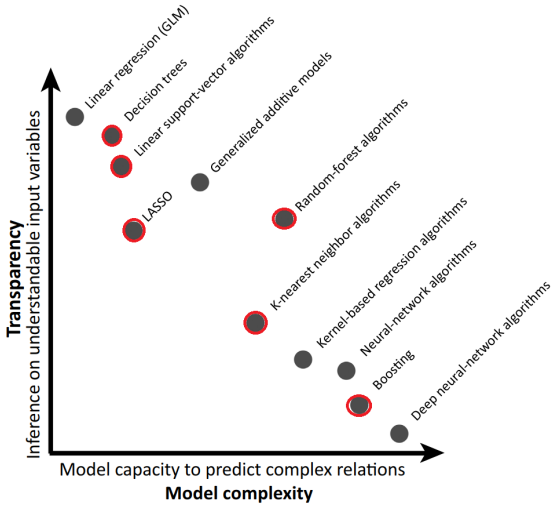


Figure 2.18: Transparency of the decision made by the algorithm versus complexity. The red circles correspond to the most used models in our work. [29]

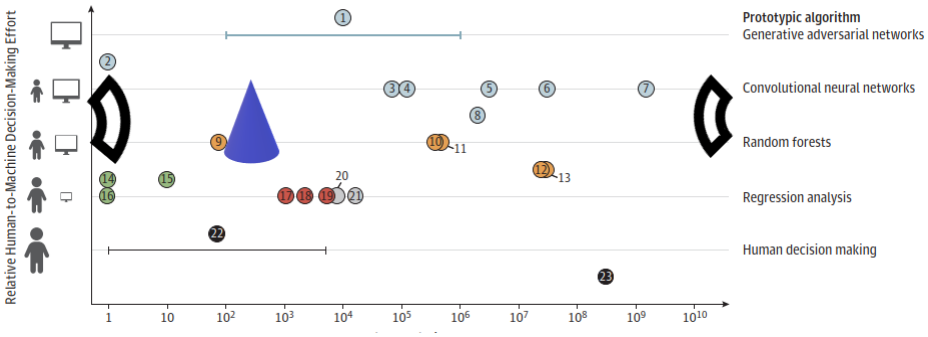


Figure 2.19: The schematic scale of ML spectrum and expert involvement adapted from [13]. The blue cone symbolizes the spectrum of our research, with the line of classical ML algorithms like RF at the base of the cone, and the tip of the cone corresponds to Deep Learning. On the x-axis, we can see the data size scales, and on the y-axis, the amount of expert effort required.

and therefore, on a unique dataset without ground truth reference, it wouldn't make much sense to learn something.

However, the hypothesis inside the hypothesis is that thanks to many detailed publications on ParI described in datasets comparable to ours, but generally with far fewer features, we could use these learning capabilities to see what makes a ParI patient according to each study on our dataset. And, as we have additional data not previously taken into account, we hope for a sort of homogeneity in the different predictions to define what ParI is. But not only in terms of objective PSG criteria, much more as a concept that pushed so many teams in the world for 40 years to put the "stick" ParI on the forehead of the patient.

Figure 2.20 presents the five essential steps in any data science project. Steps 3 and 4 will serve as a blueprint for briefly describing the main ML algorithms in this chapter.

2.9 Data Mining and Exploratory Data Analysis (EDA)

2.9.1 What is EDA ?

Once the data is collected and cleaned, we can start the data exploration or mining. This step is also named EDA. We won't go into all the statistical techniques used to visualize data, but the most useful

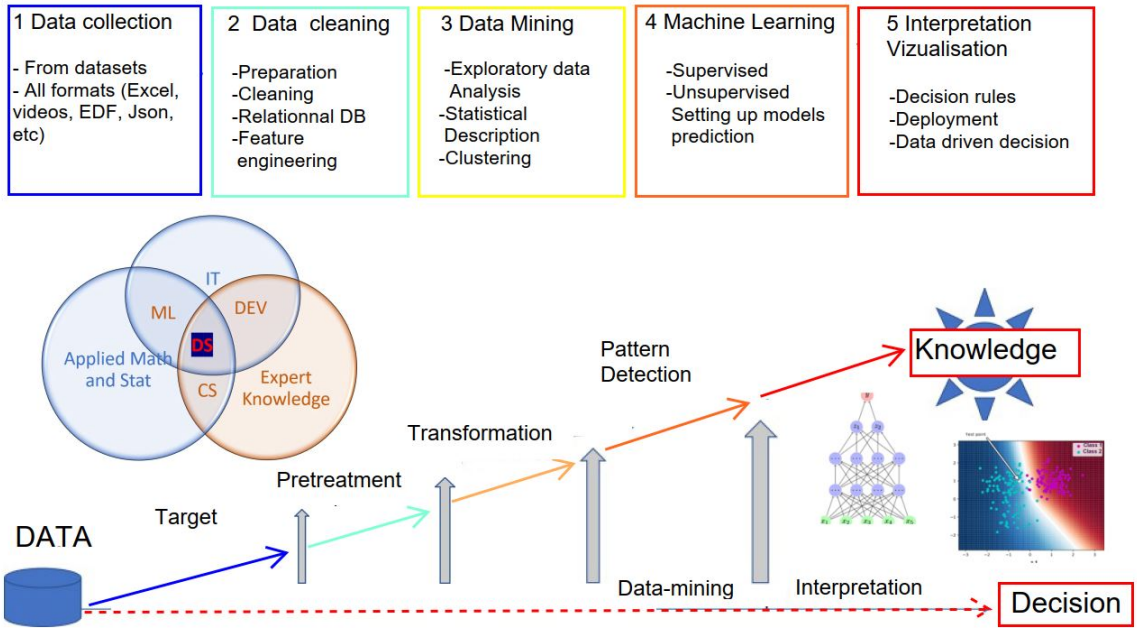


Figure 2.20: The schematic process in five steps of computer science to transform collected data into new knowledge and ultimately into decision support. Data science (DS) is at the crossroads of expertise, computing, Information Technology (IT), Software Development (DEV), ML, Applied Mathematics, and Statistic and Classical Statistics techniques

will be the display of means, standard deviations, correlations, and potential comparisons between groups using t-tests, for example. The aim is to understand how the data relate to each other and their impact on a possible output. This step is primarily devoted to understanding the data structure, finding outliers, and identifying patterns or relationships between variables. The methods we're going to use, which will form part of the description of our data set in the next chapter, are a mixture of descriptive statistics and unsupervised learning to visualize the data in the best possible way, both literally and figuratively (as a synonym for conceptualizing).

2.9.2 Simple statistical analysis

To understand the distribution, we will use simple descriptive statistics techniques such as mean, median, mode, standard deviation, skewness, and kurtosis, to have information about the average value of a dataset. This first analysis would be completed with histograms, bar charts, box plots, and scatter plots to visually examine data and identify trends, patterns, and outliers. This preliminary analysis could be very effective in identifying patterns or associations between variables; heatmaps can visually represent the correlation between variables. Then, to make inferences about a population based on sample data, t-tests, chi-square tests, and different types of analysis of variance or covariance, such as ANOVA, could be applied depending on the nature of the data and the research question [84]

2.9.3 Cluster analysis

In the context of EDA, the first step towards discovering hidden patterns is using unsupervised algorithms to discover group data points with similar characteristics or properties whose differences we'll then have to investigate without preconceived ideas.

Other algorithms could be used to investigate without preconceived ideas, like KMEANS, hierarchical clustering, or Density-Based Spatial Clustering of Applications with Noise [51]. The clustering techniques aim to identify groups or clusters in the data based on similarity measures. Clustering techniques aim to maximize intra-cluster similarity and minimize inter-cluster similarity.

We will develop only KMEANS here. This technique could be very useful for data segmentation into distinct groups or clusters based on similarity. We used this method to group similar observations and potentially show outliers. The k-means clustering algorithm is an iterative process that moves the cluster centers or centroids to the average position of their constituent points and reassigns the instances to their category [51].

2.9.4 Principal components analysis

Principal Component Analysis (PCA) [97] This old and classic dimensionality reduction technique transforms a large set of variables into a smaller one that keeps most of the information. This is achieved by reducing the dimensionality of the feature space by projecting the data onto a lower-dimensional subspace in vectorial space. This new set of variables, called principal components, is uncorrelated. This projection method uses the Mean and extreme variance of the data points, seeking their linear combinations. This clusterization can help visualize high-dimensional data, understand the underlying structure or remove redundancies.

We can see an example of PCA decomposition in 2D in Figure 2.21.

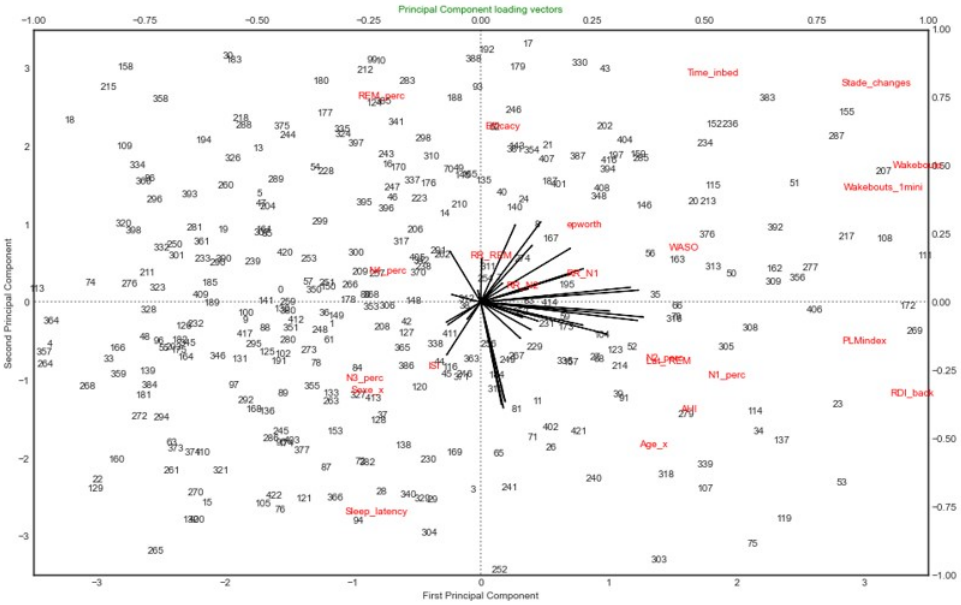


Figure 2.21: PCA1 (First Principal Component) and PCA2 (Second Principal Component) are the first two principal components. They are orthogonal (at right angles to each other) and represent the directions of maximum variance in the data. This 2D representation helps visualize the most significant patterns in the data, reducing the complexity of high-dimensional datasets to more understandable two-dimensional representations.

2.9.5 Tools to aid visualization

t-Distributed Stochastic Neighbor Embedding(tSNE) [202] is another technique for data visualization when we seek non-linear dimensionality reduction. It's an algorithm also used for exploring high-dimensional data and is particularly useful for visualization in two or three dimensions. In high-dimensional data, every data point can be considered a point in high-dimensional space. The "distance" between any two points can be considered a measure of how similar these points are. These distances can be calculated in various ways, like the Euclidean distance method. Unlike PCA, tSNE preserves small pairwise distances or local similarities, which could be useful when dealing with non-linear manifold structures in the data. Converting high-dimensional Euclidean distances between points into lower-dimensional space is done by conditional probabilities representing similarities. As a result, tSNE is performant at preserving the local structure of the data and then can help in visualizing clusters. But this technique is very sensitive to perplexity, a measure of the effective number of neighbors.

Chapter 3 will use all these algorithms.

2.10 Building Predictive (ML) Models

Our work will use classifier algorithms in the supervised context. As described above, the model learns a relation between input data (X) and corresponding output classes (Y) on labeled training

examples. The algorithm learns how to generalize from the labeled data to predict the same output classes on new datasets. When only two classes exist, this is a binary classification problem; if more, it's a multilabel classification.

2.10.1 Which model to use?

An extensive review of 120 classifiers from different families tested on 121 datasets (from 10 to 130000 samples) [60]. The whole characteristics and hyper-parameters tuning (see B.1.2 for the description) are described in [60] for all the classifiers. The interesting finding is that five classifiers from the RF family and two from the SVM are included among the top 10 best classifiers tested. Among the three left, there is a DT (C5.0Tree-t), a neural network classifier (mlp-t), and a direct kernel perceptron (dkp-C). The mean score of the maximum accuracy of these 20 classifiers and their names are presented in Figure 2.22.

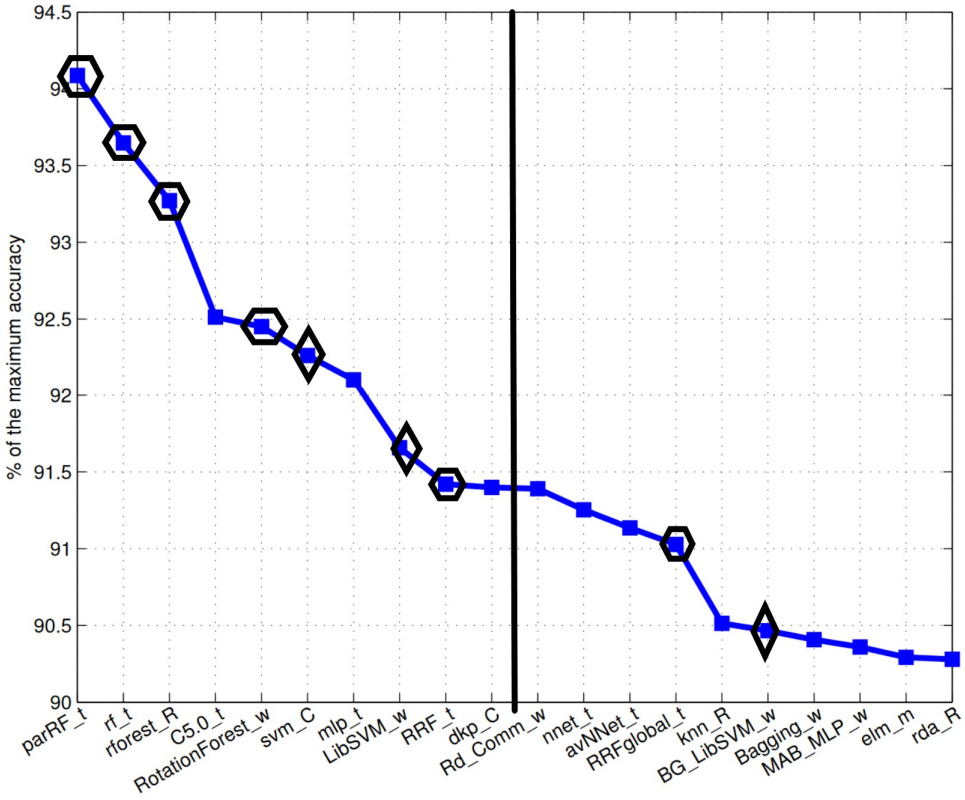


Figure 2.22: 20 best classifiers on 120 tested by [60]. The different families tested are DT (n = 15), Rule-based classifiers (n=12), Boosting (n=20), Stacking (n=2), Bagging (n=24), RF (n=8), Others Ensembles models (n=11), Generalized linear models (n=5), Nearest Neighbors (n=5), others methods (n=18). The vertical delimitation corresponds to the top 10 on the left side. The hexagons correspond to the RF family classifiers. The losanges correspond to the SVM family classifiers.

Those results were in some way replicated in [110]. A more recent paper [187] that tested SVM, KNN, Naive Bayes, and DT have also found SVM with the best results and DT as the second. Publications using ML on medical databases [93] also showed SVM and RF as the best algorithms, but also Extreme Gradient Boosting (XGB). LR was less efficient in predicting the target, but the scores were still competitive. Another study that compared RF, LR, KNN, Naive Bayes, and DT on the Breast cancer Wisconsin's dataset [9] found that LR had excellent accuracy after 10-fold cross-validation, and the best was RF. The main metrics used to establish accuracy were F-Measure and Matthews Correlation Coefficient [110]. We decided to test and compare all the algorithms described above for our predictions. We'll briefly describe them. A diagram with the level of transparency of the main algorithms cited is shown in Figure 2.18.

DT is used for classification and regression tasks. DT main principle is creating a flowchart-like model that makes predictions based on a series of binary decisions. Understanding their principles

and construction is a prerequisite to understanding RF. The general aim of DT is to explain a value from a series of discrete or continuous. If Y (the target) values are continuous, we will use a Regression DT; if the value is qualitative, it is a classification DT. The main principle of a DT is splitting the data according to selecting the most informative feature from the input dataset. The goal is to create subsets that are as pure as possible, meaning that the instances within each subset share similar characteristics. We have to choose the splitting criterion at each DT node to do that. The most commonly used criteria include Gini impurity and entropy for classification tasks and mean squared error or mean absolute error for regression tasks (See the definition for these terms in B.1.2). These criteria measure the impurity or the error in the subsets resulting from the split. After this process, a tree is built (see Figure for an example). Each internal node represents a decision based on a specific feature, and each leaf node represents a predicted output or a class label. The path from the root node to the leaf node represents the decision-making process for reaching a prediction. To avoid overfitting a pruning step may be performed to prevent overfitting. Once the DT is constructed, it can predict new, unseen instances. The predicted output or class label associated with that leaf node is then assigned to the instance. DT has many advantages, especially in the medical field, because the prediction process can be understood and interpretable. We can see an example of DT algorithm in Figure 2.23.

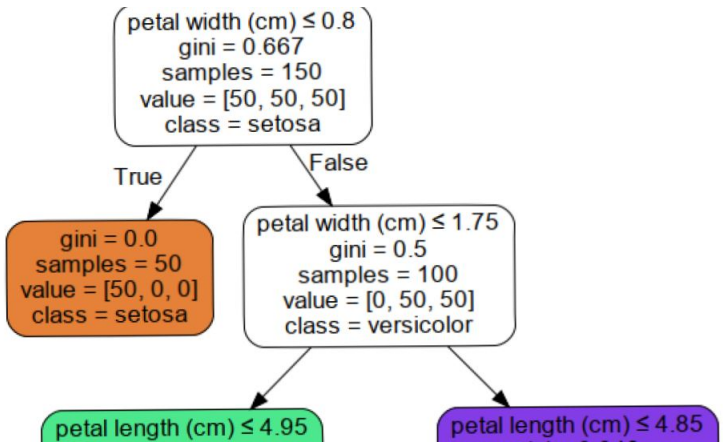


Figure 2.23: In this example, we took the Iris dataset containing flower characteristics to predict the class or species of an Iris flower based on its features. The DT algorithm is trained on the dataset with known class labels to learn the patterns and relationships between the features and the corresponding class labels. It's a multilabel classification of the three species (Setosa, Virginica, and Versicolor). As an example, we can take the first node. The node is split based on **the petal width (cm)** feature, and the threshold value is 0.8. The node contains 150 samples (instances) that satisfy this condition. The class distribution for this node is [50, 50, 50], indicating that the instances belong to the three classes the values mean. This first node classifies all the Setosa directly from that feature (orange color; class = setosa)

As they can handle numerical and categorical features, DT could be used in a wide range of problems, and data preprocessing is minimal, making them easy to use. However, they can be prone to overfitting if not properly regularized or if the tree grows too deep. It's then important to set the limit in the hyperparameters. Another algorithm, such as RF, can address these limitations and enhance the performance of DTs. The next paragraph will focus on this wide-use algorithm.

RF is part of the Ensemble Learning algorithms or Tree-Based ML Models. DT, Bagging, and Boosting are part of these models (see definition in B.1.2). So there are DT at the origin of the RF. This method combines multiple DT to make predictions. Instead of relying on a single DT, RF aggregates the predictions of multiple trees to achieve better overall performance. The main objective is to reduce overfitting. With a few more details, RF uses a bagging technique to create diverse subsets of the original dataset. Subsets of the studied sample will be replaced so that each subset is used to train an individual DT. In addition, RF randomly selects each DT node to create diversity in the resulting ensemble. For each subset of the data, a DT is constructed using a specific algorithm (typically, the Classification and Regression Trees (CART) algorithm [24]). The DT is trained to recursively split the data based on the selected features to maximize the homogeneity of the target variable within each resulting branch or leaf. In Scikit-learn [162], RF could be highly tunable, with many hyperparameters. This could be very time-consuming. The main parameters are the number of

DT to be used in the random forest, the criterion (quality of split), the maximum depth of each DT, the minimum number of samples required to split, the minimum number of samples required to be at a leaf node and the randomness of the algorithm. In the end, RF uses its collective predictions to make a final prediction by voting. In classification tasks, the class that receives the majority vote from the individual tree is selected as the final prediction. In regression tasks, the average or the median of the predicted values from all the trees is taken. One of the advantages of RF in medicine is that it can help visualize the algorithm's decision-making process.

SVM is an extended version of maximum margin classifiers, in which they have to find a decision boundary or hyperplane that maximally separates different classes in the data while maintaining a margin of separation. The margin refers to the distance between the decision boundary (hyperplane) and the nearest data points of each class. SVM aims to find the decision boundary that maximizes this margin, thus creating a clear separation between the classes. The support vectors are the closest data points to the hyperplane, influencing its position. They are linear classifiers designed for classes separated by a hyperplane. However, SVM can also handle non-linearly separable data by using the kernel trick. The kernel trick involves mapping the original feature space into a higher-dimensional space where the data becomes linearly separable. Commonly used kernels include the linear kernel, polynomial kernel, and radial basis function kernel. These kernels are part of the hyperparameters included in the algorithms. In Scikit-Learn, we can tune these hyperparameters to improve the prediction. The process of hyperparameter tuning could be done automatically. In the case of SVM, the other tunable parameters are the regularization parameters to find the best balance between maximizing the margin and minimizing the training error. [45]

K-Nearest Neighbors does not make any assumptions about the underlying distribution of the data (non-parametric ML model) and uses the training data directly for classification. The model memorizes the training observation for classifying the unseen test data. KNN compares the test observations with the nearest training observations based on a chosen distance metric (Euclidean distance) to determine the class or value. KNN does not learn a representation of the underlying data distribution. The theory behind it is far from new [47] . As KNN requires comparing each test data point with the training data to find the nearest neighbors, it can be computationally expensive and inefficient for large datasets. The time complexity of KNN grows linearly with the size of the training data, making it less suitable for big data scenarios. This curse of dimensionality can lead KNN to struggle a lot to find meaningful nearest neighbors due to the increased sparsity of the data in high-dimensional spaces. This can lead to a deterioration in performance as the distance-based similarity measures become less reliable. But in our case, this classifier could be useful or at least worth to be tested. The main tunable hyperparameters are the number of nearest neighbors to consider and the weights that could be proportional to the inverse of the distance.

Naive Bayes algorithm or Bayesian classifiers use Bayes' theorem (feature independence). They calculate the probability of a class given a set of features by considering the conditional probabilities of each feature given the class. This independence assumption simplifies the computation and allows the model to handle many attributes efficiently. The problem in our case is the amount of data necessary with this classifier. The interest with many features could be the joint probabilities of multiple features. Indeed, Bayesian classifiers can capture the collective influence of these attributes, which can contribute to very good classification. Thus, the Naive Bayes classifier is effective on tasks such as email classification based on words with high dimensions of vocabulary. But, if we have to tackle feature dependencies or interactions, especially with small datasets, this classifier could be inappropriate [175, 51].

Logistic regression is a linear model that estimates the relationship between the predictor variables and the probability of the binary outcome. It uses the logistic function (also known as the sigmoid function) to map the linear combination of the predictors to a value between 0 and 1, representing the probability. The model is trained using the maximum likelihood estimation method, which finds the parameters that maximize the likelihood of the observed data given the model [9].

2.10.2 Deep learning

In the context of our thesis, we avoided deep learning classifiers for the lack of transparency (except for Time series analyses). See Figure 2.18. Indeed, as we wanted to explain the predictions as much as possible, not only did we want the best trade-off between performance and transparency, but we added a model designed to explain which variables (or features) were involved in the decision to the classifier we will use as many as possible understandable algorithms that can provide important insights into what drives predictions. As we know the possible and well-known failings, we distrust ML and approach its use with abundant caution. Hence, our emphasis is on trustworthiness and explainability. Like in [28], we used the LASSO algorithm, and we added two more to increase the reliability in the prediction explanation, Morris sensitivity and Global Shap values. It helps in understanding which features are more important in model decision-making. We will develop this part extensively in Chapter 5.

2.10.3 Explainability methods

LASSO is a regularization technique used in linear regression models. It introduces a penalty term to the linear regression cost function, encouraging sparsity in the coefficient estimates and performing automatic feature selection. The L1 penalty term shrinks some coefficients to exactly zero, effectively excluding less important predictors from the model. LASSO is useful for improving model interpretability and handling high-dimensional datasets [199].

SHapley Additive exPlanations (SHAP) (Shapely value analysis) is a method for explaining the output of any ML model. It provides a unified framework for assessing the contribution of each feature to the prediction. SHAP values are based on Shapley values from cooperative game theory and provide a way to attribute the prediction to different features by quantifying their impact on the output. SHAP values offer local and global interpretability, allowing for individual instance explanations and overall feature importance analysis [130].

Morris sensitivity analysis is a global method to assess input variables' impact on a model's output. It is a variance-based method that measures the effect of one input variable at a time while keeping the others fixed at different levels. Morris sensitivity provides a qualitative ranking of input variables based on their influence on the output and helps identify influential factors for further analysis or model refinement [150].

2.11 In this thesis: ML as a tool

Then, the question here is not using ML as a magic tool with a black box but more using the math and the statistics behind some of these tools to understand hidden complex patterns involved in predicting the ParI concept and the outcome of CID. Explainable methods are recommended for this goal, such as generalized linear models [13]. We did this by using more ML algorithms than Deep Learning Algorithms. As we can see in Figure 2.19, ML tools are in a medium position in terms of human effort needed or, let's say, expertise. To situate it, we symbolized our work area with a blue cone on the ML spectrum. The human symbols correspond to the number of parameters predetermined by humans about the research question. The trade-off between the human specification of a predictive algorithm's properties vs learning those properties from data is what is known as the ML spectrum. In Figure 2.19, we can also see that our area of work is not really "big data" but sufficient to apply ML algorithms, as mentioned earlier.

Causal inference in the medical field is key to understanding the effectiveness of treatments, interventions, and policies. This concept refers to concluding a causal connection based on the conditions of the occurrence of an effect. Based on the observed data, the goal is to infer a causal relationship between a treatment (the cause) and an outcome (the effect). In a randomized study, the treatment assignment is random, which helps to balance confounding variables (variables that influence both the treatment and the outcome) across treatment groups. This helps to establish a causal relationship between the treatment and the outcome, given the assumption that there are no hidden biases due

to confounding. We hypothesized that we could learn and predict this causal relationship with a data-driven approach using ML classifiers and explainers. However, it's important to note that while ML has potential advantages, there are also challenges. These include the risk of overfitting, difficulty interpreting complex models, choosing a fitted algorithm to the task with good hyperparameter tuning, and, for some, the need for large amounts of data. These questions will be part of the discussion after each experiment in Chapters 5 and 6.

Chapter 3

Data Collection and Visualization

Chapter contents

We discuss our data collection, database curation, dataset construction, data mining, and visualization.

1. Data collection phase We collected data and built the different datasets used for this research with a three-step process: recovery, selection, and aggregation.
2. Difficulties Encountered During Database Collection and Recommendation from Experience Our strategy to overcome difficulties faced during database and dataset aggregation is described. We discuss here how we dealt with missing values and why we decided not to use missing value imputation methods.
3. Exploratory Data Analysis on the Different Databases We describe the data extracted from each database to do the first EDA before aggregation. All the features used will be described. Correlation analysis and dimensional reduction clustering methods will describe the potential informative knowledge.
4. Exploring the impact of respiratory and neurological events on sleep fragmentation We evaluated the impact of different thresholds of respiratory events and limb movements on sleep fragmentation in our sample.
5. Evaluation of the Relevance of Datasets to the Assumptions Made We evaluated the relevance of our sample dataset regarding the representativity of the CID population and the link with Paradoxical Insomnia.

Key Terms and concepts

Acronym/term	Definition	Ref.
AHI	Apnea-Hypopnea Index	p. 169 (B.1.1)
CID	Chronic Insomnia disorder	p. 193 (B.7)
EDA	Exploratory data Analysis	p. 45 (2.9)
EDF	European Data Format	p. 169 (B.1.1)
KMEANS	K-Means Clustering	p. 46 (2.9.3)
MMPI	Minnesota Multiphasic Personality Inventory	p. 32 (2.4.1)
OSA	Obstructive Sleep Apnea	p. 24 (2.4.1)
PCA	Principal Component Analysis	p. 47 (2.9.4)
PLM	Periodic Limb Movement	p. 24 (5)
tSNE	t-Distributed Stochastic Neighbor Embedding	p. 47 (2.9.5)

3.1 Data Collection and Datasets Aggregation Process

3.1.1 Data collection phase

Data were collected in the Insomnia investigation and treatment centre attached to the sleep disorders federation at the Pitié-Salpêtrière Hospital in Paris. This unit was an annex in the neurophysiologic department of the Sleep Disorders Federation and was mainly dedicated to treating CID. With the technical support of a neurophysiologic department, our multi-disciplinary approach, both psychiatric and neurological, has enabled us to systematically and comprehensively assess patients with CID in a multi-modal way.

Overall, the patients we saw over the years were middle-aged (between 40 and 50 y.o.), with a professional activity for the vast majority (nearly 80%), and about 66% of women. They were generally without severe organic or psychiatric disorders, apart from a few exceptions. Around 40% had sleep medication at the first interview, and a few percentage could present treated or not the primary complaint of Obstructive Sleep Apnea (OSA) (defined by Apnea-hypopnea index (AHI) $\geq$ 15 per hour) and Restless Legs Syndrome (RLS) with or without Periodic Limb Movements (PLM) (defined by PLM index $\geq$ 15 per hour).

How recruitment and assessment were organized had much to do with this. Patients had to fill in questionnaires at the time of application. These questionnaires then served as a first selection step, enabling patients with psychiatric, organic, or other sleep-related disorders besides Insomnia to be referred directly to the relevant departments, especially subjects with typical symptoms of OSA, hypersomnia or neurological sleep disorders like REM Sleep Behavior Disorders (See definitions in B.1.1).

So, the patients evaluated in our department were autonomous, often followed for CID from several months to several years before, with no other predominant sleep disorders apart from a few comorbidities with treated OSA presenting with Insomnia, comorbid, treated RLS. The subjects recruited were theoretically between 18 and 65 for the Insomnia assessment protocol, with older subjects referred to a senior sleep center. However, some older subjects managed to pass this filter.

The data in this thesis were collected with the patient's informed consent and written authorization in the frame of retrospective non-interventional research. Data was gathered from 2011 to 2017 during routine hospital care. We adhered to the regulations and recommendations of the Commission Nationale de l'Informatique et des Libertés (CNIL) regarding data mining. The data were anonymized, protected, and restricted to essential information for the study, and only the author (O. Pallanca) had access to the patient's personal information. Further details of the efforts are shown in Figure 3.1.

The features and the sample selected to be part of the final datasets will participate in understanding and determining the profile of CID and ParI. So, the aim was to have a dataset compatible with assessing CID as the primary complaint and not secondary to psychiatric or medical disorders. The data collected come from different databases corresponding to the tests, questionnaires, and clinical assessments made routinely in the department. The theoretical part of these tests and questionnaires were already described in the **Assessment of Insomnia** subsection (See 2.4.1). So here, we will talk about the data retrieval process and the number of features selected at each stage of the database retrieval and assembly process. The process will be explained for each database and the difficulties encountered in the next section.

3.1.2 From databases to datasets

We needed to put data in a format upon which we could deploy machine learning, i.e., to datasets with definitive instances, features, and (where possible) class labels.

Figure 3.1 presents the data retrieval process and aggregation in three steps:

- The first step corresponds to the recovery of all available data. Data sources are represented as data storage (marked sources one to five in orange), each containing specific information for up to 1,735 patients.

- The second step corresponds to the first data selection with source A as a reference. Indeed as we wanted the psychological profile of chronic Insomniacs patients, source A included selected MMPI-2 scores of CID patients (1182), we then tried to retrieve data from sources B, C, D and

E corresponding to the patient's ID in source A. This process led to four Databases (I, II, III, and IV) with the maximum sample size for each source without applying the definitive selection criteria (CID assumption was made only with clinical diagnosis).

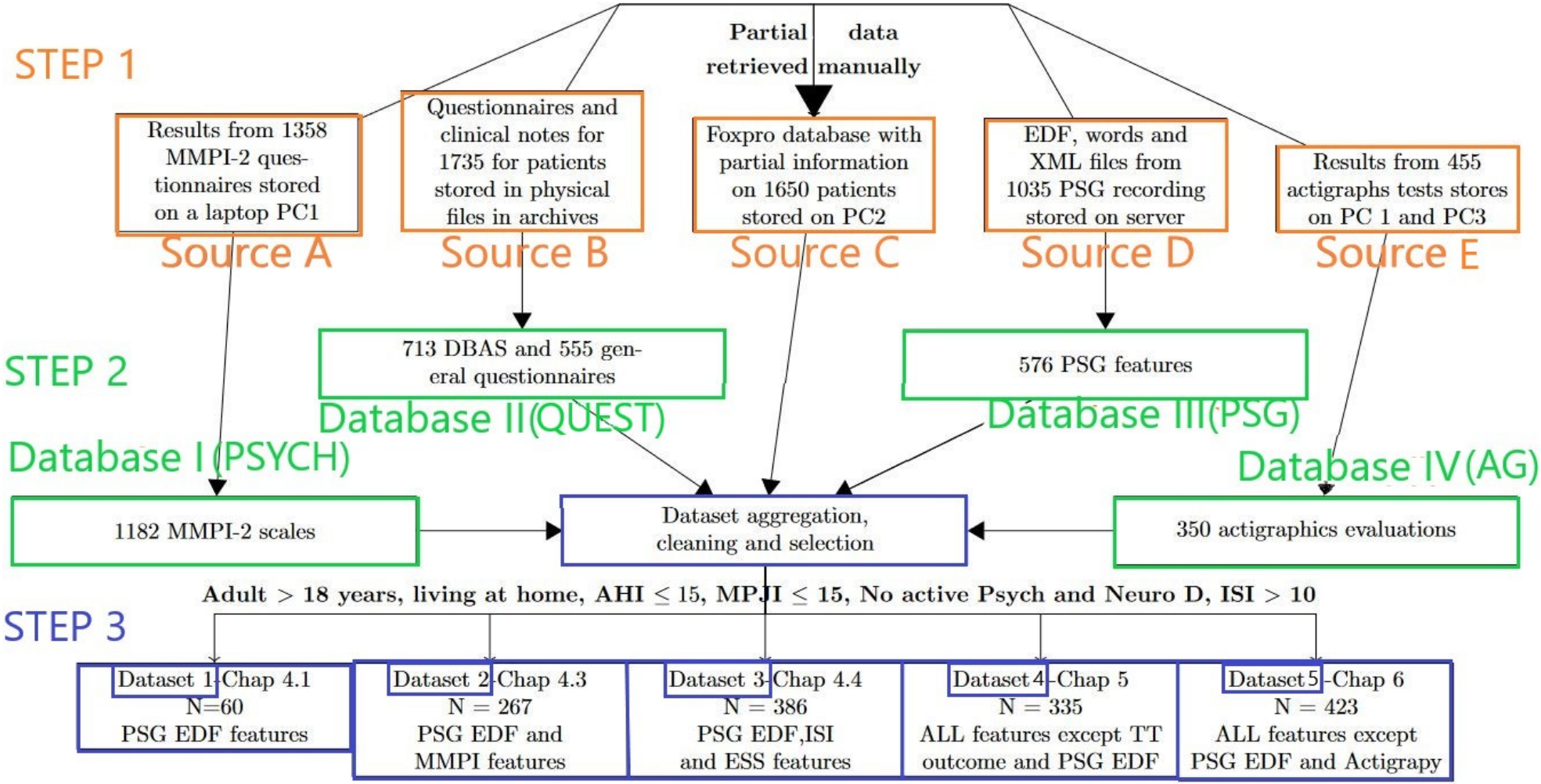
- The third step corresponds to the aggregation between the four Databases plus additional missing values provided by source C. Five datasets were built according to the different hypotheses. The first three used European Data Format (EDF) files from the EEG. Dataset two used MMPI-2 scores for prediction, and Dataset three ISI and ESS for clustering. Dataset four aggregated all the features from databases I to IV except EDF files and Treatment outcome features. Dataset five aggregated all the features from databases I to III except EDF files and most actigraphic features.

The 91 features selected in the DI-PSYCH are listed in the Appendix (see Table B.2) The 45 features selected in the DII-DBAS are listed in the Appendix (see Table B.3) The 38 features selected in the DIII-PSG are listed in the Appendix (see Table B.4) The 27 features selected in the DIV-AG are listed in the Appendix (see Table B.5) Seven additional features from the sleep log (LOG) are included and listed in the Appendix (see Table B.6):

The EEG features extracted from EDF files will be described in Chapter 4.

Then, if we remove the keys and duplicates for age and gender, a maximum of 202 original features were selected from the five data sources.

Figure 3.1: Selection process for datasets used in this thesis work



As we can see, our expectation in terms of sample size was lower than expected. The process and the main difficulties encountered in building our datasets are listed in the next section.

3.2 Difficulties Encountered During Database Collection and Recommendation from Experience

The process started with the MMPI-2 source available in one block on a secured computer. We were optimistic with 1182 CID MMPI-2 scores selected on 1345 total. We thought it would be easy to create a final dataset of at least 1000 patients with all the other features available in the other sources. But very quickly after the beginning of this work, difficulties inherent in our lack of knowledge of data storage, human errors, organizational problems, hospital regulations on data retention time, oversights, losses, inappropriate external interventions, patients who did not return usable data, etc., have somewhat dampened our hopes. The data collection process for this project experienced significant challenges, leading to considerable data loss. The main challenges are described in the next subsection.

3.2.1 Main challenges in the data collection process

1. **Data Fragmentation**: The datasets were fragmented and stored in different locations and formats. For instance, MMPI-2 was stored on PC1, while the DBAS questionnaire was stored in Excel on a technician's computer.

2. **Incomplete Records**: There were substantial missing data. Of the potential 1182 patients, only 713 had fully completed the DBAS questionnaire. Missing data was due to various reasons, including hand corrections instead of computer inputs, deletions, or patients simply not filling them out. However, the number of non-random missing data was relatively low. This was because patients had to fill in questionnaires on the spot, which the technicians collected. As a result, there was little useful information in the missing data.

3. **Data Retrieval Issues**: The retrieval process was difficult. Questionnaires filled in on paper had to be retrieved from archives, and many were either lost or discarded, reducing the usable questionnaire data to 519 records.

4. **Data Deletion**: Unforeseen circumstances like unannounced computer maintenance led to the deletion of some files. This resulted in only 350 out of 519 actigraph results being recoverable.

5. **Partial Salvaging**: Despite these obstacles, the team salvaged additional data from other patient databases, specifically average sleep times over a week. However, this was not sufficient to fully compensate for the lost data.

6. **Corrupted data**: Although we will address this issue in Chapter 4 regarding retrieving polysomnographic data in EDF format, we encountered several types of difficulty. First of all, EDF formats did not take expert scoring into account. We had to re-export files in XML format that could not be directly superimposed with the EDF formats, with a risk of error. At this point, we devised the idea of using automatic scoring algorithms after realizing that the one we had available in the EDF files was unreliable. Another problem was the portion of PSG data with artifacts. Although they appeared legible at the time of their interpretation, when we applied data extraction, we realized how difficult it was to obtain reliable data, which explains the smaller size of the data sets for the EEG study. Again, we prioritized quality to explain the study subjects better rather than adding corrupted data.

7. **Formatting errors:** As the results of each PSG were exported in Word format, some of the information was only available there. We, therefore, had to create an extremely laborious VBA code to retrieve all the data from all the fields in the Word documents. All these files had to be cleaned by hand to remove export errors, not to mention all the commas and periods hidden, added by hand during the test, making the CSV files extremely difficult to put back

in order. During maintenance work carried out by our IT staff, some of the archived scoring reports disappeared. We had to score the tracing and export it again for around fifty patients, corresponding to at least 50 hours of extra work.

3.2.2 Main solutions and proposals to solve these challenges

1. **Data cleaning**: Once all the data had been recovered and the formatting errors repaired, all the data sets had to be transformed. Variables were renamed. Data were anonymised. Keys were created for merging the databases. We transformed strings into float when necessary.

2. **Missing values**: When the missing values were random, we replaced them with the mean of the retrieved features (less than 5%). We had some missing values not at random, like some Sleep Latency when the patients didn't sleep; we had to represent the fact that the latency was long without any value; in that case, we chose to fill with a high generic number like 999.

3. **Crossing sources** Whenever possible, we searched the physical files for missing information. We even called some patients back to obtain information, particularly on their treatment outcomes. We created a special questionnaire mailed to them to evaluate it objectively when the data was missing.

But from our experiences, we wanted to formulate some recommendations that would probably be useful to other medical departments.

1. **Centralization of the database**: We regret not implementing a centralized data storage system to avoid fragmentation. All the data should be stored in one location accessible to all relevant parties. A cloud-based system could facilitate this and offer robust data protection and recovery options.

2. **Standardization of Data Entry**: This is also fundamental to establishing a standard procedure for data entry to ensure consistency and completeness. Using systematic computer-based forms instead of paper to store the data in the database directly. This reduces manual input errors and makes data retrieval easier. This needs the full compliance and training of technicians and medical doctors. In the ideal, this process could use Data Validation Checks during the data entry process to ensure the data is complete and in the correct format. This could include prompts for missing entries or warnings for data that does not fit expected formats or values.

3. **Regular Backups** are also fundamental when storing data on a specific computer, even in the hospital, to avoid loss from unexpected incidents like system failures or accidental deletion.

4. **Clear Communication with IT Staff**: is also an issue. We have to inform the IT staff of the importance of the data and the specific formats and software used. Clear communication and instructions can prevent unintended loss or corruption of data during maintenance or updates, as we experienced.

We will now make some EDA on the databases recovered and cleaned.

3.3 Exploratory Data Analysis on the Different Databases

We will present this EDA on all the databases to benefit from the maximal number of samples. We will describe them in the sorted order from Database I to IV, corresponding to the descending sample size from 1182 for Database I to 335 for Database IV.

We wanted to describe each database individually to have an insight into the data with unsupervised learning before aggregating them in datasets four and five used in 5 and 6. We will follow the same process of EDA for each database with descriptive statistics, PCA, t-SNE, and K-means clustering. Only meaningful results will be shown.

3.3.1 Database I (DI-PSYCH)

(DI-PSYCH) is the database gathering all the MMPI-2 questionnaire responses. These responses (567 in total - "yes" or "no" choices) are used to build some different psychological scales. The score used in this thesis is the T-score. In the MMPI-2, T-scores are a standardized scores used to interpret the results and better understand how an individual's score relates to the scores of others who took the same test on each test scale. The T-score is calculated to have a mean (average) of 50 and a standard deviation of 10 in the reference population. For most MMPI-2 scales, a T-score of 65 or above is usually considered clinically significant, meaning it may indicate the presence of a psychological problem. This cut-off can vary, though, depending on the specific scale and the context in which the test is being used. But in general, High T-scores (>70) indicate that the test taker similarly answered the questions to individuals in the clinical group (those with a diagnosed mental health disorder). Low T-scores indicate that the test taker answered the questions similarly to the normative group (people without diagnosed mental health disorders), But an extremely low T-score could have the same value as a very high T-score in some scales, especially in the validations scales.

The MMPI-2 questionnaire and its interest in Insomnia research were described in section 2.4.1. Compared with the studies described above that used only the main 10 general scales (see Table3.1), we decided to retain all the validated scales and subscales available for each analysis of the 567 responses given by each patient. In the end, 91 features corresponding to the 91 scales were retained, enabling a more detailed analysis of patients' psychological profiles. The features and their definitions are listed in the Appendix (Table B.2). The main results of this first EDA are listed below.

Distribution

The Distribution of each T-score for 80 scales is shown in Figure 3.23. We can see an interesting deviant peak for the Es scale in the third plot (DI-PSYCH 3), showing that a significant proportion of the sample is below 40 for this scale. That means all the patients concerned generally have a poor self-image, feel worthless, ruminate, feel powerless or maladjusted with old problems, tend to be inhibited, and have physical ailments, chronic fatigue, fears, or phobias. They often feel unable to cope with the pressures of their environment and are often a little rigid in solving problems. They often express a desire for change in their care but do nothing. This scale is very important in predicting good treatment adherence. Also, we could see a peak in the scale TRT (Negative Treatment Indicators).

General Statistics

The average age is 44.9 ±13.8 yo, and the female proportion is 65.4%.

The mean, standard deviation (SD), median, and quartiles of the T-score concerning the general scales are shown in Table 3.1.

scales	Hs5K	D	Hy	Pd4K	Mfm	Pa	Pt1K	Sc1K	Ma2K	Si
mean	**66.4**	**67.0**	**65.9**	58.5	51.4	59.6	63.2	60.5	50.4	56.1
std	12.7	12.6	13.7	12.5	10.4	13.8	12.3	12.6	10.8	10.4
25%	57.0	58.0	56.0	50.0	45.0	51.0	54.2	52.0	42.0	49.0
50%	66.0	66.0	66.0	56.0	50.0	58.0	62.0	59.0	48.0	56.0
75%	75.0	76.0	74.0	66.0	57.0	68.0	71.0	68.0	57.0	63.0

Table 3.1: T-scores for the 10 General scales on 1182 patients. This scale is theoretically sufficient to detect a pathological profile if the validity scales are in the normal range, as is the case here

Correlations analysis

The previous results showed that some specific scales increased the T-score means, like Hy, D, or H5K. Others showed particular distributions on the dataset and were linked to the treatment outcome, like Es and TRT. The plot DI-PSYCH 1 in 3.2 showed that the feature FB (one on the main validation scales related to fatigue and attentional issues) is almost sufficient to represent three clusters according to its T-Score on our dataset. We present in Figure 3.3 a specific focus on the Pearson correlation between these features.

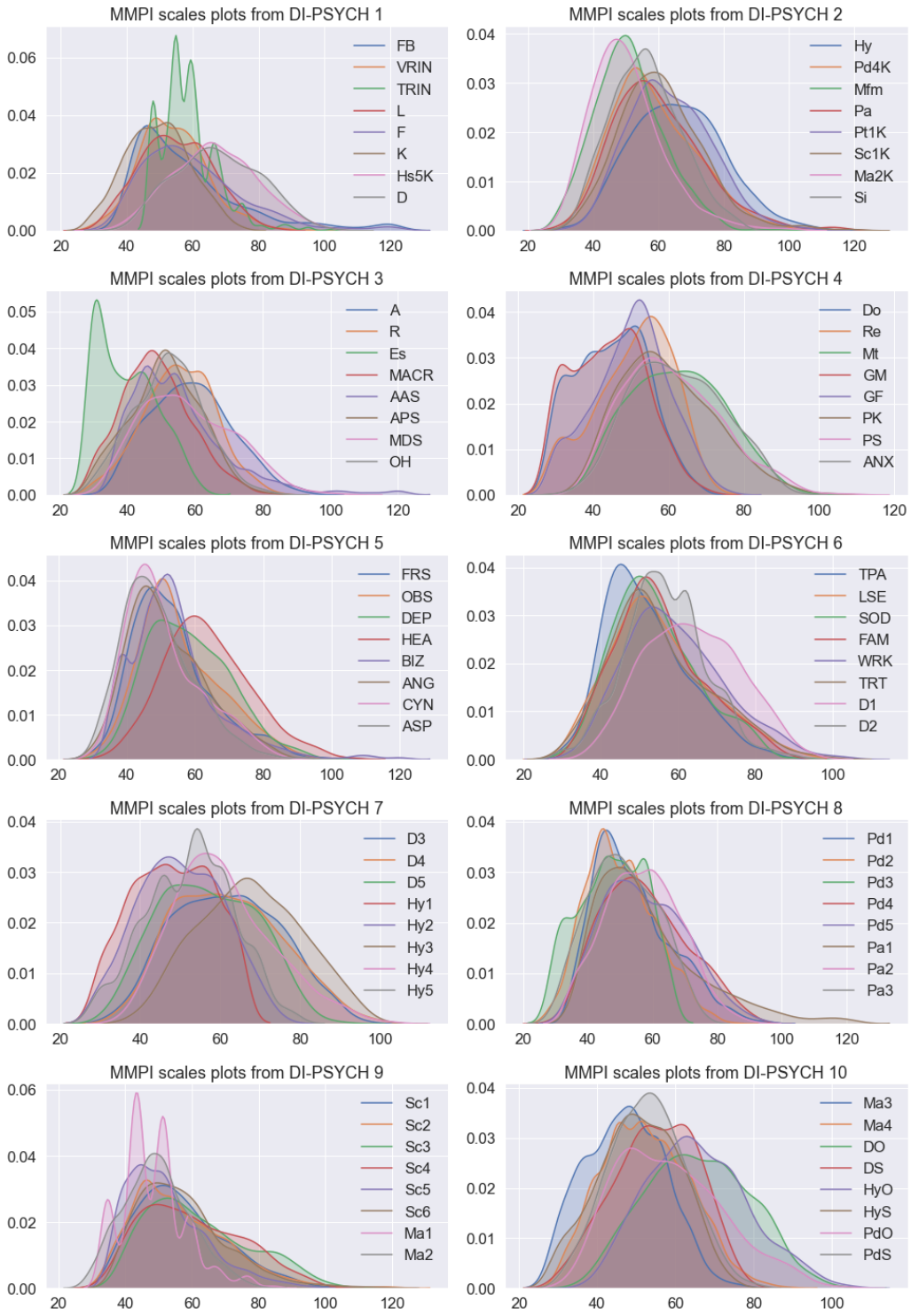


Figure 3.2: The Distribution for each scale is plotted with the T-score in the x-axis and the ratio to total sample size in the y-axis. 80 scales are represented. The main validity and general scales are represented in plots 1 and 2. Plots 3 to 10 represent the MMPI-2 subscales. We can see that the validity scale (VRIN, TRIN, L, and F) allowing the interpretation of the MMPI-2 is largely under a T-score of 70. We can see in Plot 3 that the Es scale T-score is separated in two around 40, with a peak below 40. This scale is related to mental ruminations and low self-esteem. We can see a strange distribution for the Ma1 scale in DI-PSYCH 9 with four peaks, all below the pathological threshold. Also of interest is a peak for the scale TRT that is also related to negative treatment

The results show a high correlation between FB and the feature TRT, so it's possible that the cluster with the highest FB T-score could be the one with the most treatment resistance. TRT and Es are anti-correlated in a significant way, which means that these two scales are linked to a profile. The correlation between Fb and Es is weaker, so Es must cover a wider field than the treatment outcome. On the other hand, the highest T-score mean in our sample, Hy, D, and Hs, are poorly correlated to FB, so we have effectively different clusters of psychological profiles. In detail, we could even observe a cluster in the cluster with Hy and Hs5K poorly correlated to TRT when D does. This observation

could lead to the interpretation that being depressed is a risk factor for poor treatment outcomes.

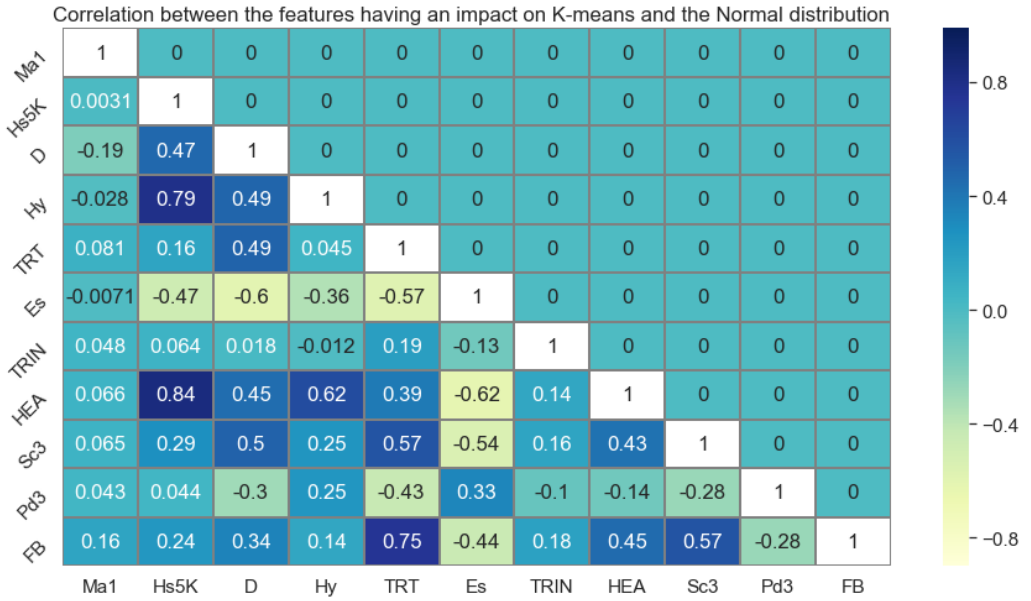


Figure 3.3: Pearson correlation analysis between scales with increased T-score means(Hy, D, and H5K), the ones linked to a negative treatment outcome (Es, TRT), and FB heavily involved in the three clusters found in our dataset

Dimensionality reduction and clustering

The main results for 3D PCA are presented in Figure 3.4. We couldn't visualize any clusters in the high-dimensional data for the t-SNE results after perplexity tuning from three to 100. For K-means clustering, we tested 3 to 6 clusters corresponding to the main numbers of insomnia subgroups published until now. Figure 3.26 presents the visually discriminative results.

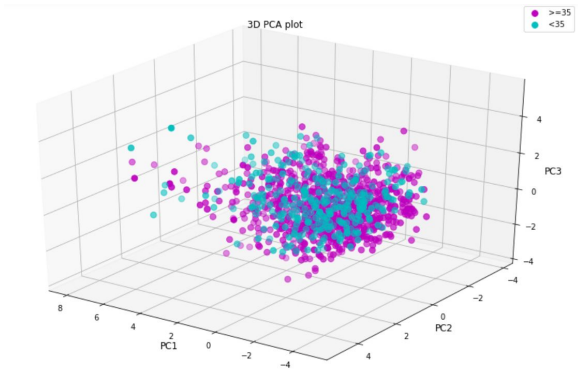


Figure 3.4: PCA on the 10 main scales on the MMPI-2 dataset with age and gender. We found after several experiments that only age could slightly affect the PCA representation, below or above 35.

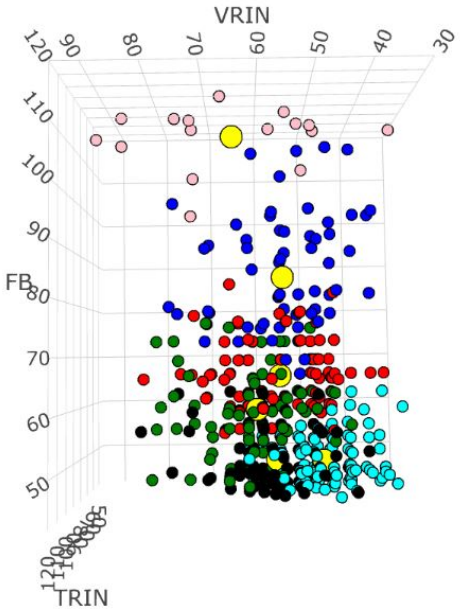


Figure 3.5: K-means clustering in 3D for K=6 for MMPI-2 Tscore with the validity scale VRIN, TRIN and Fb. We could see three distinct cluster regions almost entirely linked to the Fb scale. When the Fb T-score is elevated scale, it is potentially linked to fatigue, attention disorders, a tendency to show oneself in an unfavorable light, or a sign of severe psychopathology.

Discussion

- So, in our population of chronic Insomniacs, from 1182 subjects, visualization by K-means clustering enabled us to identify three clusters differentiated solely by their T scores on the 87 MMPI2 scales and subscales. We can see that Fb is heavily involved in discriminating the different clusters. This is the only study to have used so many patients and all the scales on such a large population of chronic Insomniacs. It confirms and further extends a similar study [56] involving only 100 patients and 13 scales. It also found three clusters using T scores on the three validity and 10 clinical MMPI scales with the Fortran clustering procedure available. Even if the method is not described in the paper [56], we could find it in the Fortran book for clustering [213], and centroid technique and Nearest-Neighbor were available in Fortran Programming. What is also of interest is that the author, Jack Edinger, was a pioneer in Insomnia subtyping, chair of the Insomnia sections of the ICSD-2 and ICSD-3 and headed AASM's Research Diagnostic Criteria for Insomnia Workgroup, and currently is leading the academy's Insomnia Treatment Guidelines Task Force. So, it seems that 35 years later, our results confirm these clustering premises. However, by the time of this study, one of the three groups was too small and removed from further descriptive or inferential statistical findings [56]; so they kept only two clusters without any similar attempt with MMPI. In 2017, he participated in a paper whose title is evocative: "Characterization of Patients Who Present With Insomnia: Is There Room for a Symptom Cluster-Based Approach?". In that paper [48], they found three clusters among 170 patients with Latent profile analysis from sleep logs, questionnaires, and PSG features again.

- We could see that two subscales (Es and TRT) used as a treatment adherence and outcome indicator showed elevation peaks in our sample. TRT reveals a significant positive relation between scores and readmission to the hospital for patients who may be at high risk for unsuccessful substance abuse treatment [73].

So from these first observations, we could see that some scales are specifically interesting in finding different clusters in CID, like Hs5K, D, Hy, Ma1, TRT, and Es. We will finish this first EDA with a correlation matrix for the scale of interest.

Conclusions

The findings suggest strong relationships among certain variables in the MMPI-2 (Minnesota Multiphasic Personality Inventory-2) psychological profiles. A notable positive correlation exists between the FB (Infrequency Back) scale and the TRT (Treatment Resistance) feature. This implies that the cluster with the highest FB T-score could potentially represent individuals with the most resistance to treatment.

A significant negative correlation exists between the TRT and Es (Ego Strength) scales. This anti-correlation suggests that these two scales collectively characterize a distinct psychological profile. However, the relationship between FB and Es is unclear, suggesting that Es may have broader implications beyond negative treatment outcomes.

Despite having the highest average T-scores in the sample, the Hy (Hysteria), D (Depression), and Hs (Hypochondriasis) scales show weak correlations with FB. This suggests the existence of distinct clusters of psychological profiles.

Further analysis reveals nuanced relationships within these clusters. For instance, while Hy and Hs5K show weak correlations with TRT, D exhibits a notable correlation. This differential relationship can be interpreted as indicating that depression may pose a significant risk factor for poor treatment outcomes independent of hysteria and hypochondriasis.

These results highlight the multifaceted nature of psychological profiles in the Chronic Insomniac population. The identified clusters and correlations might be analyzed with the sleep features to see if they could inform therapeutic strategies or help predict treatment resistance.

3.3.2 Database II (DII-QUEST)

Database II (DII-QUEST) is the aggregation of the Dysfunctional Beliefs and Attitudes about Sleep (DBAS) questionnaire with the other transversal questionnaires used to assess the severity of Insomnia

(ISI), sleepiness (ESS), anxiety (STAI-T and STAI-S), depression (BDI-II), and the circadian profile (HO) (see Section 2.4.1 for background). The DBAS is a 30-item self-report questionnaire designed to identify and assess various sleep/Insomnia-related cognitions (e.g., beliefs, attitudes, expectations, appraisals, attributions). The importance of the DBAS was emphasized in Dysfunctional sleep-related cognitions and attitudes as a model of Insomnia [81]. This study shows that negative emotions in cognitions could lead to arousal, activating selective monitoring of physiological and environmental factors related to sleep performance. Thus, Insomniacs could classify them as inefficient and could conclude there is a sleep deficit. This aspect is of great interest for ParI analysis. In this Database II, all the scores for each of the 30 individual questions of the DBAS are available. The interest is that each question is an assertion about sleep (for example, " I need to sleep 8 hours each night to feel refreshed") with an estimated agreement to each assertion between 1 (low) and 10 (high) (see the full questionnaire in Appendix in Figure B.4). This questionnaire allows a good understanding of the cognitive representation of Insomniacs and could help in understanding specific profiles.

Transversal questionnaires on sleep, 519 patients

Although 713 DBAS questionnaire IDs matched with the DATABASE I (DI-PSYCH), we could find only 519 corresponding transversal questionnaire IDs.

Distribution We can see the Distribution for each questionnaire in Figure 3.6

General statistics This DII-QUEST allowed the evaluation of Insomnia severity with the ISI questionnaire in our sample and compared it to the literature. The description of each score is presented in Table 3.3.2.

	count	mean	std	min	25%	50%	75%	max
Female	519.00	64.93	47.76	0.00	0.00	100.00	100.00	100.00
Age	519.00	45.46	13.92	18.00	35.00	45.00	56.00	84.00
ISI	519.00	19.49	4.06	11.00	17.00	19.00	22.00	28.00
ESS	519.00	8.13	5.14	0.00	4.00	8.00	12.00	23.00
DBAS	519.00	153.04	34.87	13.00	132.00	154.00	177.00	232.00
Ho	519.00	50.58	11.22	7.00	44.00	51.00	57.00	86.00
stai_etat	519.00	40.77	12.23	20.00	31.00	39.00	48.00	80.00
stai_trait	519.00	48.06	9.69	18.00	41.00	48.00	55.00	72.00
BDI_2	519.00	16.90	10.47	0.00	9.00	15.00	23.00	59.00

Table 3.2: Mean score, standard deviation, and quartiles for the different transversal questionnaires and the total score of DBAS

Correlation analysis The correlations between the transversal questionnaires and the total score of the DBAS questionnaire are presented in Tables 3.7 and 3.8.

Dimensionnality reduction We did a PCA on the merge of DI-PSYCH and DII-QUEST presented in Figure 3.9 with poor results.

We also ran a K-means clustering and a t-sne, but the results presented in 3.10 and 3.19 are not contributives.

Regressor Decision trees To see if we could predict the DBAS total score with the psychological profiles, we took the merge of DABS and MMPI-2 with 713 patients, and we ran a Decision Tree regressor presented in Figure 3.12.

Figure 3.6: This figure shows eight plots ((DII(DBAS) 1 to 8) distribution on our sample. Plots DII(DBAS) 1 to 6 show the subscore Distribution for each of the 30 questions of the DBAS questionnaire on the x-axis and the ratio to total sample size on the y-axis. Plots DII(DBAS) 7 to 8 show the score Distribution of the other sleep questionnaires. Epworth (ESS) shows a non-normal distribution that we need to investigate, as many questions of the DBAS. We could observe the discrepancy between the anxiety scale designed to evaluate anxiety on the entire life (stai-trait) and the present time evaluation (stai-etat). We could see the non-normal Distribution of the depression scale (BDI-II) showing a kind of subgroups of more depressed subjects and a similar one with the STAI-E

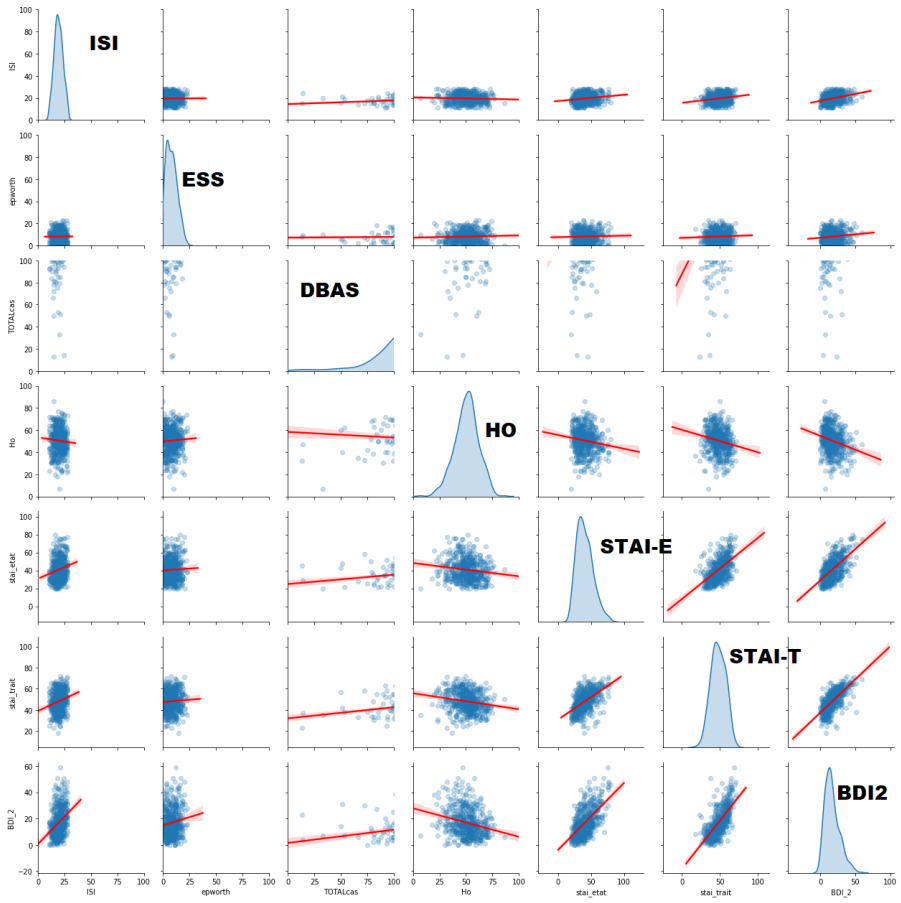


Figure 3.7: Graphical pairwise comparison and Distribution between the questionnaires. We could observe the high correlation between the questionnaires designed to assess the anxiety trait and state (STAI-T and STAI-E) and the questionnaire used to assess depression (BDI2).

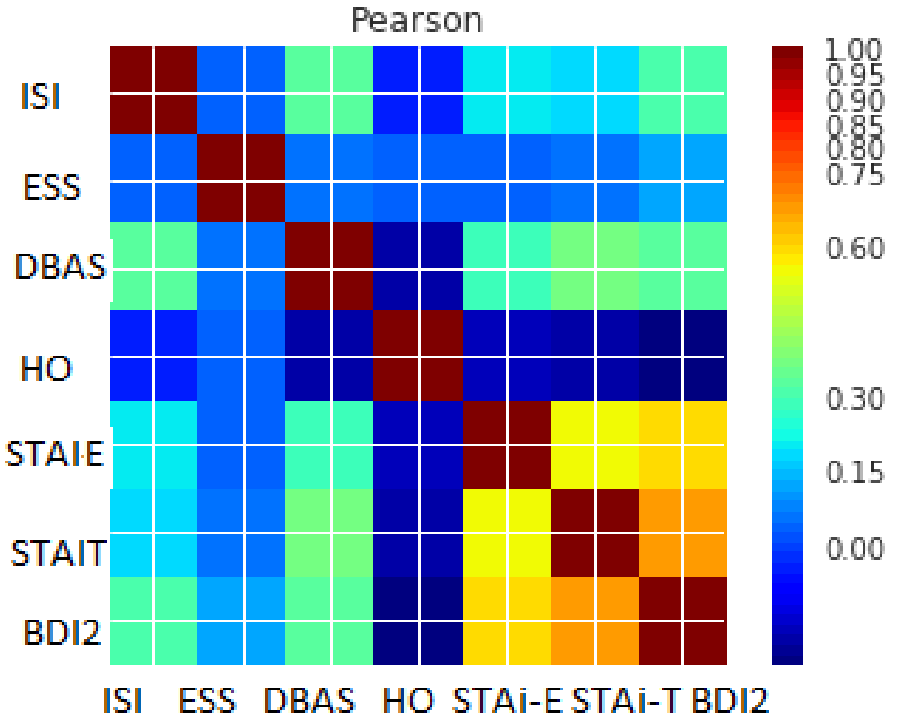


Figure 3.8: Pearson correlations between the questionnaires shown in Figure 2.7. We could observe the high correlation score between the questionnaires designed to assess the anxiety trait (STAI-T and STAI-E) and the questionnaire used to assess depression (BDI2).

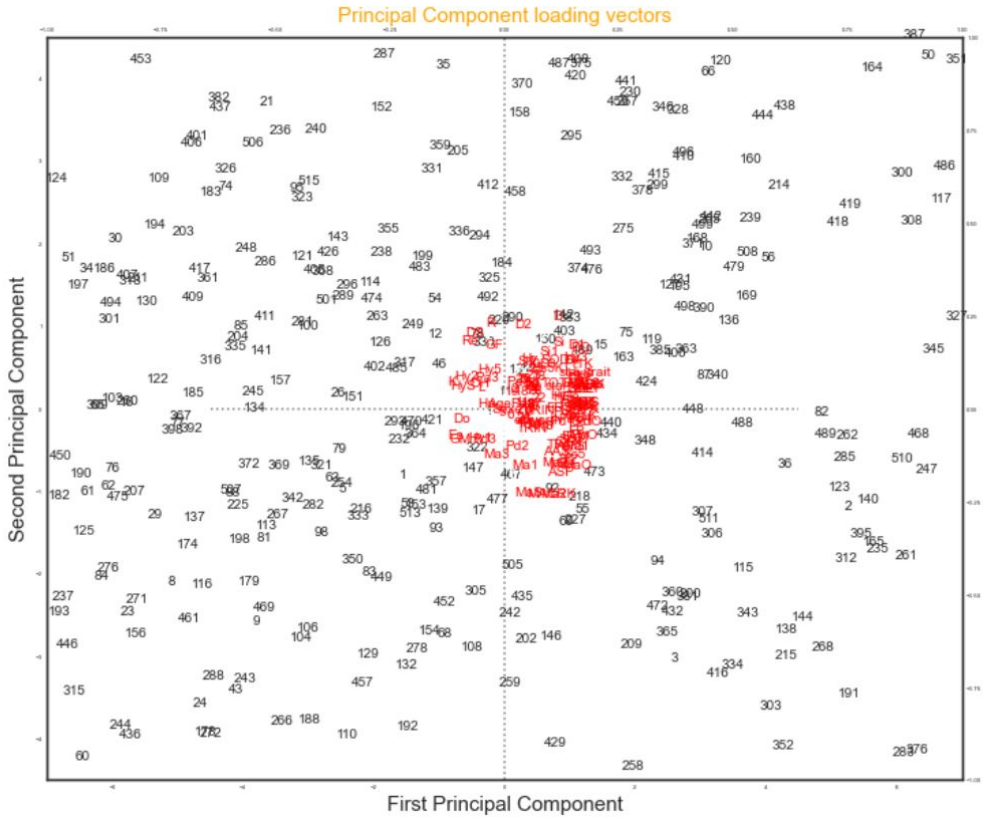


Figure 3.9: PCA on RangeIndex: 519 entries, 0 to 518 Columns: 126 entries with MMPI-2 scales and QUEST features. We can see the first two components on this plot; the data points appear scattered and do not form any discernible clusters or patterns. So, a very small percentage of the total variance is explained by the first two principal components, leading to the conclusion that PCA did not reveal meaningful structure in the data.

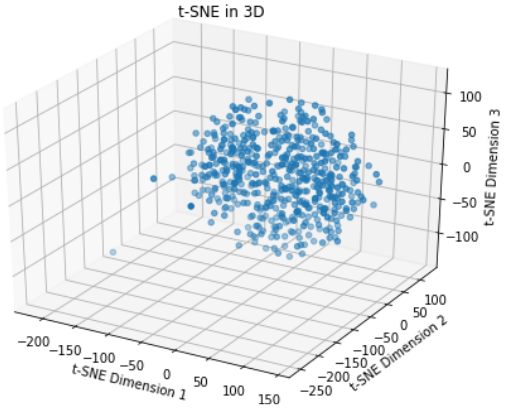


Figure 3.10: t-sne on the DII-QUEST

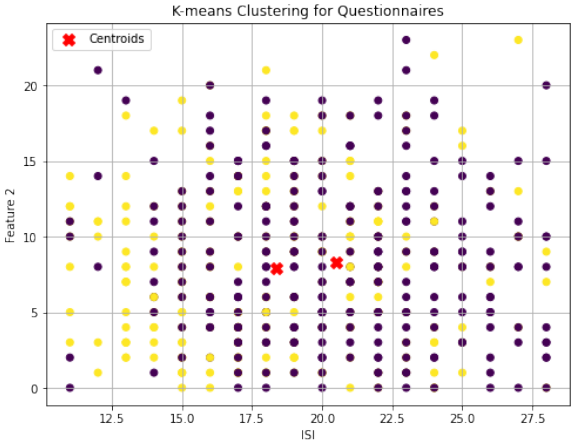


Figure 3.11: K-means on DII QUEST x-axis is ISI score and y-axis the ESS score

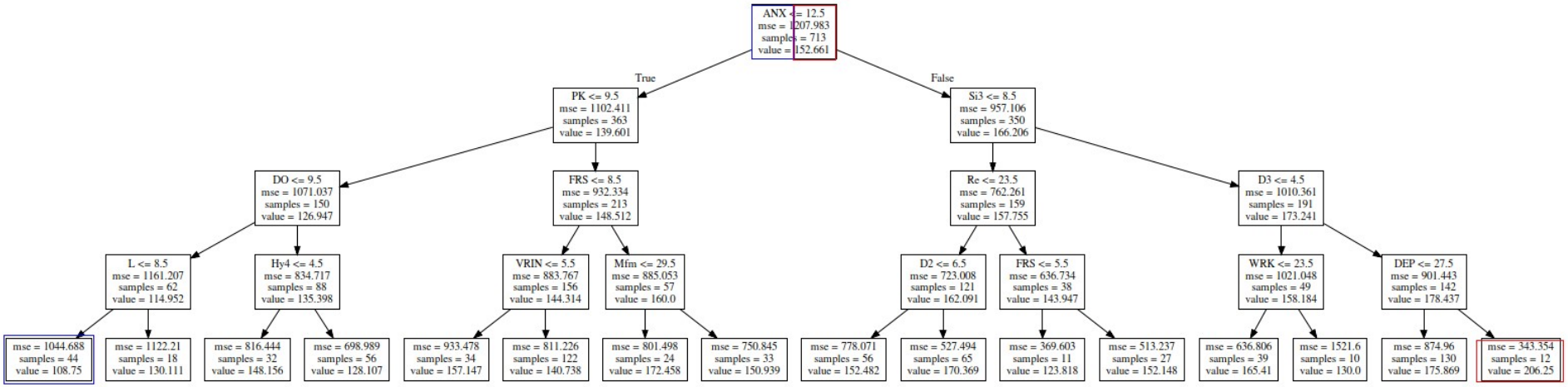


Figure 3.12

This decision tree shows that the main criterion for separating the sample was the MMPI anxiety scale (ANX). By following the leaves from ANX, we can see that the lowest score is on the left of the tree (Blue box), with 44 subjects having a mean DBAS value of 108, and the highest on the right (Red box) of the tree at the bottom, with a score of 206 for 12 subjects. This result already confirms the correlation of the DBAS with the anxiety scales, meaning that the lowest score on the anxiety scale of the MMPI could predict a low score on the DBAS scale.

We ran a PCA visualization to understand better the feature's influence with only the contributive features shown in the Decision Tree regressor. We can see the spatial representation of these features according to the total DBAS score (we reinforced the visualization with a proportional coloration gradient). This figure shows an interesting link between the anxiety scale (ANX) and problem at work scale (WRK) and the dysfunctional beliefs and attitudes toward sleep. On the contrary, the scales Do (Dominance) and Re (Responsability) are poorly linked with the high DBAS score.

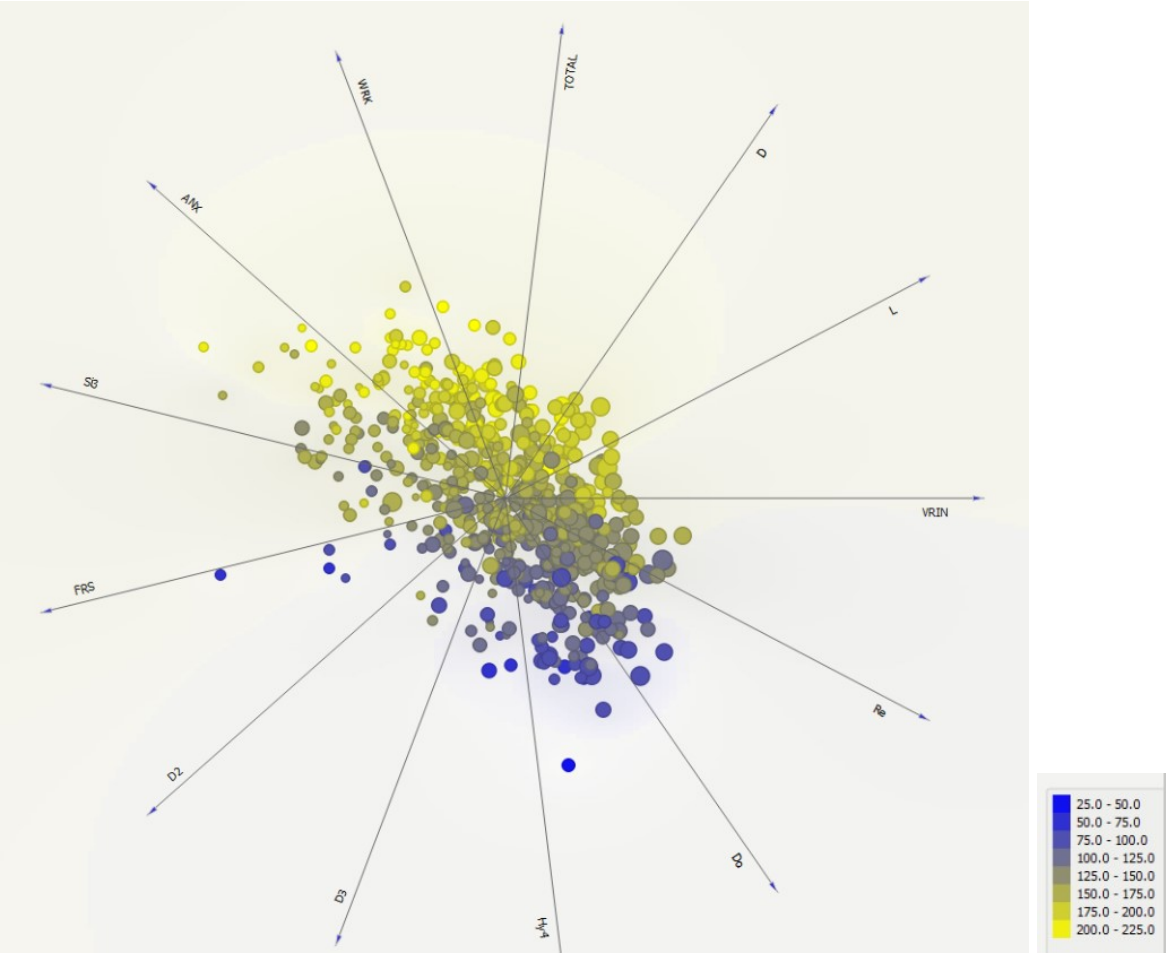


Figure 3.13: PCA with visualization of the main scales involved in the Decision Tree process to predict DBAS Total Score. Blue = Low DBAS score, and Yellow = High DBAS score. To better understand the Decision Tree shown in Figure 3.12, this graph confirms the influence of the MMPI anxiety scale in predicting the DBAS total score. DBAS total scores are distributed along the axis symbolized by the ANX (highest score) and Do (lowest score) scales, meaning that the higher the ANX score, the greater its influence in predicting the DBAS total score; conversely, the higher the Do score, the lower the DBAS score.

Discussion

In the dataset with 713 samples mixing the MMPI2 and DBAS scales, we found interesting and concordant results mixing a DT for a regression problem to predict the total score obtained on DBAS and a PCA using the features extracted from the DT. From these analyses, we could show with the DT that the Anxiety scale of MMPI2 was the most important node, which is not very new as we know that DBAS is correlated with anxiety scales [204], but most interestingly, we showed that the lower score, which means people with normal beliefs and attitude toward sleep, so in theory not the target for I-CBT, corresponded to the lower scores on the Lie scale in MMPI2 and high score in social responsibility and dominance. This profile suggests that this cluster of people does not feel vulnerable and is not afraid of being judged. Conversely, the patients with higher scores of DBAS are highly depressed, and they could feel alienated, inadequate, unattractive, unlikable, and vulnerable to judgment from others.

Conclusion

Thanks to the DT regressor and a PCA visualization technique, we found discriminant features associated with psychological profiles to predict the degree of preoccupation with sleep and possible insomnia (Anxiety versus Social responsibility). This opened new perspectives on more personalized cognitive treatment.

3.3.3 Database III (DIII-PSG)

DIII-PSG (D-PSG) is the database gathering features available in the sleep report generated by the experts after sleep scoring. Table B.4 shows the total feature list. The features extracted were from the reports after expert scoring. The features are the classic parameters explored in sleep studies but with significant additional features compared to most studies on Paradoxical Insomnia like

1. The distinction between spontaneous, respiratory, and related to periodic limb movements. The goal is to discriminate the sleep fragmentation origin.

2. The number of awakenings longer than one minute during the sleep episode. The goal is to have a specific assessment of the awakenings that could be theoretically remembered.

3. The number of awakenings between 15 and 60 sec.

4. The average HR and the RR interval in the different stages with the SD to evaluate the sympathetic activity during the sleep episodes.

5. The index of respiratory events, meaning the number of partial or total respiratory limitations per hour of sleep, specifically on the back position. Indeed, insomnia linked to respiratory disorders can sometimes only occur in the supine position and go unnoticed during an initial check-up for sleep apnea.

6. The number of stage changes during the sleep episode.

Distribution

We will show the Distribution of the central features displayed in a sleep report in Figure 3.14

General Statistics

The mean, standard deviation (SD), median, and quartiles of the T-score concerning the general scales are shown in Table 3.3.

	count	mean	std	25%	50%	75%
Efficacy %	576.0	74.75	15.44	67.00	78.30	86.20
TIB h	576.0	7.54	1.16	7.05	7.42	8.28
TPS h	576.0	6.66	1.39	6.09	6.58	7.43
TST h	576.0	5.65	1.33	5.02	6.00	6.44
WASO min	576.0	58.76	50.04	22.00	43.50	79.03
AHI /h	576.0	5.71	7.79	0.90	2.90	8.02
MicA /h	576.0	20.50	11.47	12.60	18.80	25.52
TotAr /h	576.0	23.84	12.40	15.15	22.10	30.00
PLM-index /h	576.0	9.57	14.61	0.90	3.70	12.05
Wake1min number	576.0	10.74	9.09	6.00	9.00	13.00
StageCh number	576.0	85.58	44.73	60.00	79.00	100.00
SOL min	576.0	31.97	34.69	10.28	21.05	40.02

Table 3.3: PSG features mean, standard deviation (SD), median and quartiles

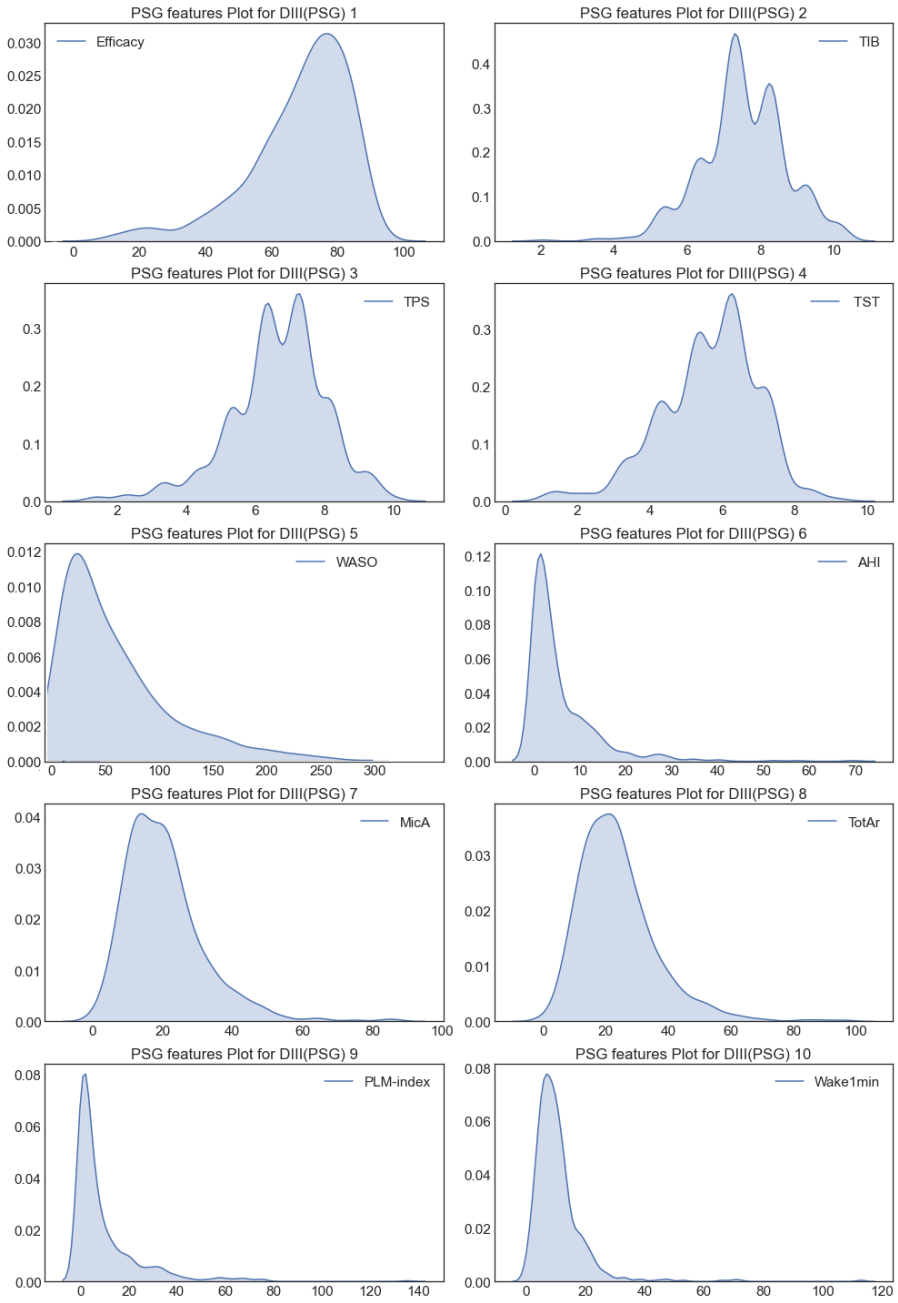


Figure 3.14: The Distribution for the main features used in PSG reports are displayed. The y-axis corresponds to the ratio of the total sample size, and the x-axis changes upon the features. For plot 1 (Sleep Efficiency), it corresponds to percentage; for plot 1 (Sleep Efficiency), it corresponds to percentage; for plots 2,3 and 4 (Time in Bed, Time Period of sleep, Total Sleep Time and) it corresponds to hours, For plot 5 (Wake after Sleep Onset), it corresponds to minutes, For plot 6,7,8 and 9 (Apnea Hypopnea Index, Micro-Arousal, Total Arousal, and Periodic Limb Movements) it corresponds to the number of events per hour of sleep and for plot 10 (Wake episodes above 1 minute) it corresponds to the total number of wake episode.

Exploring the impact of respiratory and neurological events on sleep fragmentation

In our inclusion criteria, unlike most studies on the Insomniac population, we discarded applying the usual selection criteria for respiratory events or abnormal movements, usually $< 15/h$, to be considered serious enough. Even if the results presented in 3.3 are quite normal concerning this aspect, we had some patients with higher indexes. We will concentrate our first EDA analysis on its influences to consider its influence. The justification for this choice is developed in the Appendix in B.3.4.

We would like to know how Respiratory events like obstructive sleep apnea (OSA) and Periodic Legs Movements (PLM) affect objective sleep quality and fragmentation in our specific samples of Insomniac patients.

The correlation matrix in figure 3.15 shows the interactions between the number of events detected in our sample (number of respiratory and movement events per hour) and the main features showing

sleep fragmentation. We could notice that only the Micro-arousal index is moderately correlated, but there is no correlation with WASO, TST, and SOL, the usual metrics used in Insomnia assessment, respectively the wake during the sleep episode, the Total sleep time and the sleep onset Latency.

In most papers studying Chronic Insomnia Disorder, the patients with more than 15 events per hour of sleep (respiratory (AHI) and movements (PLM)) are discarded from the studies. This threshold corresponds to mild disorder. They removed the subjects because these events could impact the sleep quality by themselves and then somehow change the observation of the sample, especially the sleep fragmentation, the sleep onset latency for PLM, and also WASO as the patients will wake up during the night. We just saw, surprisingly, that SOL and WASO were not correlated with the number of events.

We proceeded with an impact evaluation according to different thresholds to see the exact impact on sleep fragmentation and find a significant threshold that could have too much impact on our analysis. We removed the patients in our dataset according to their events index from above 40 to less than 5. The results are presented in the table 3.4 with the number of subjects remaining in the dataset after each drop of the corresponding sub-sample.

	index/h					
	> 40 Max 57	< 40	< 30	< 15	< 10	< 5
AHI	0.58	0.55	0.52	0.48	0.36	0.36
N	576	574	568	534	480	396
	index/h					
	> 40 Max 111	< 40	< 30	< 15	<10	< 5
MPJ	0.49	0.48	0.4	0.25	0.28	0.25
N	576	566	524	478	409	361

Table 3.4: Evolution of AHI and PLM index Pearson correlation with micro-Arousal index according to different index threshold

For the AHI index, we could see that the correlation decreased very slowly until the threshold of 10/h, and after that, there was a stabilization. There is a decrease between 30 and 15/h for the PLM index, but no difference after.

To see the influence of MPJ and AHI, we ran a PCA and t-SNE, and K means with the threshold of 10 for AHI and 15 for MPJ (see Figures 3.17 and 3.16)). We can see that an increase in PLM defines a subgroup of patients, more than AHI > 15. This effect disappears when we remove the subject with PLM >15. This effect is not observed for AHI. So it seems that PLM has more effect on our sample than AHI; we will see in our predictive models if this effect could impact sleep perception.

	mean	**std**	**min**	**25%**	**50%**	**75%**	**max**
Snore_index	37.27	98.08	0.0	0.00	0.2	15.85	705.7
MicAr_index	20.50	11.47	5.0	12.60	18.80	25.52	64.4
Mic_Ar_wakebouts_index	23.36	11.22	5.4	15.15	22.1	29.70	71.6
Mic_Ar_Respi_index	4.27	5.93	0.0	0.40	1.8	5.40	47.8
Mic_Ar_PLMS_index	4.71	6.95	0.0	0.40	2.1	6.15	58.9
RDI_back	9.13	15.01	0.0	0.40	3.1	11.05	101.5
Wakebouts_1mini	10.74	9.09	0.0	6.00	9.0	13.00	28.0
Stade_changes	85.58	44.73	11.0	60.00	79.0	100.00	208.0

Table 3.5: PSG features related to the sleep fragmentation, mean, SD, median, and quartiles

Dimensionality reduction and clustering

We did a PCA on the PSG features in datasets four and five. Indeed, as the number of samples is not so different, unlike Database I and II, we wanted to see the exact relationship between PSG features on the datasets built for explaining Paradoxical Insomnia and the Treatment outcome.

The PCA with the correlation with the two first components is presented in Figure 3.20

We also try to find clusters with K-Means. We can see the results in Figure 3.21

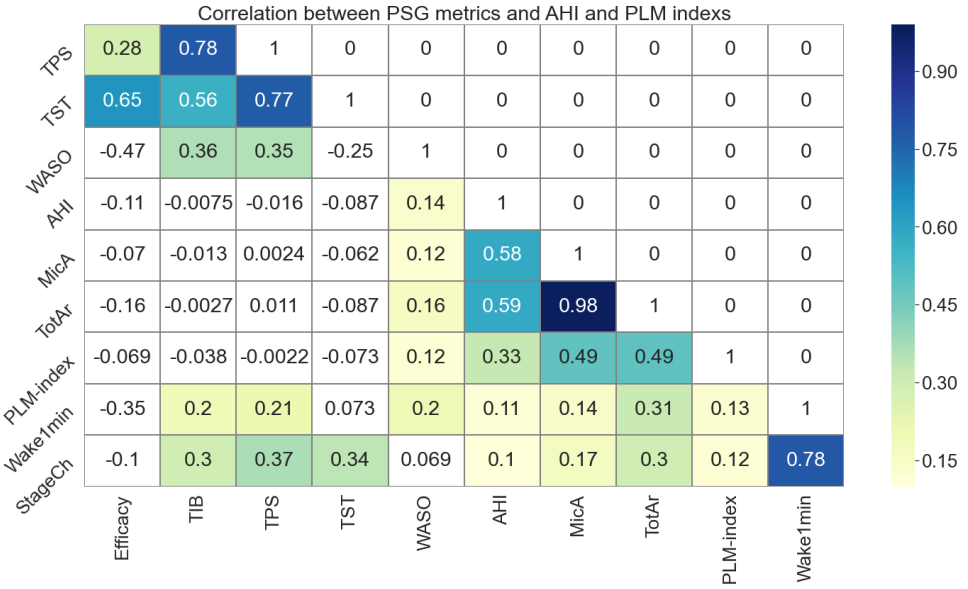


Figure 3.15: Pearson correlation between the index of respiratory events per hour of sleep (AHI) and the index of movements per hour of sleep (MPJ) and the features reflectin sleep fragmentation

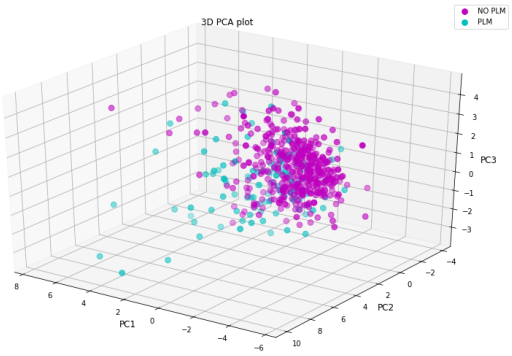


Figure 3.16: PCA with the transformation into class 0 or 1 of Periodic limb movement (PLM) > 15 (class 1 (blue points))

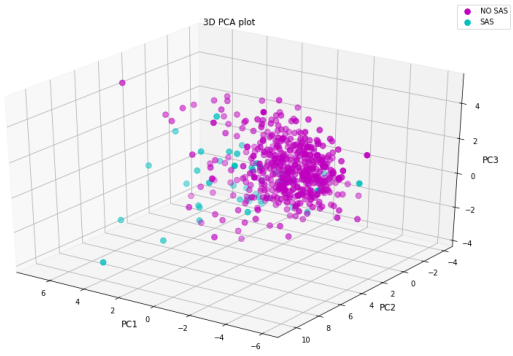


Figure 3.17: PCA with the transformation into class 0 or 1 of Apnea-Hypopnea Index (AHI) > 15 (class 1 (blue points))

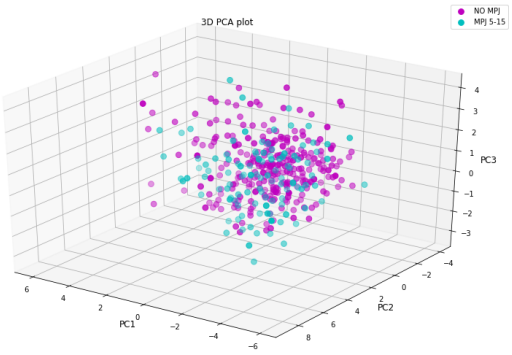


Figure 3.18: PCA with the transformation into class 0 or 1 of Periodic limb movement (PLM) > 5 but < 15 (class 1 (blue points)). The patients with an index > > 15 are removed

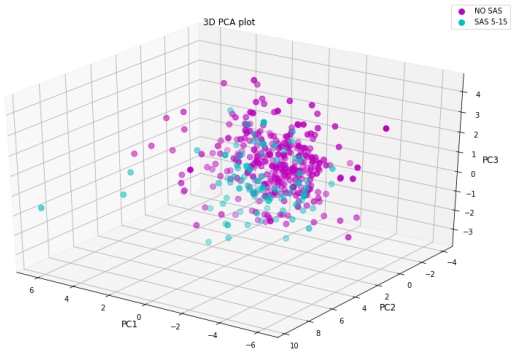


Figure 3.19: PCA with the transformation into class 0 or 1 of Apnea-Hypopnea Index (AHI) > 5 but < 15 (class 1 (blue points)). The patients with an index > 15 are removed

Discussion

Concerning the DIII-PSG, we want to emphasize that based on our correlation analysis described in 3.4, the impact of the threshold of AHI is almost the same with an index of 15 or 30 per hour on the sleep fragmentation, and that this is the threshold of 10 that showed a notable change in the correlation, but this impact doesn't change even if the AHI index is below 5. Thus, concerning the

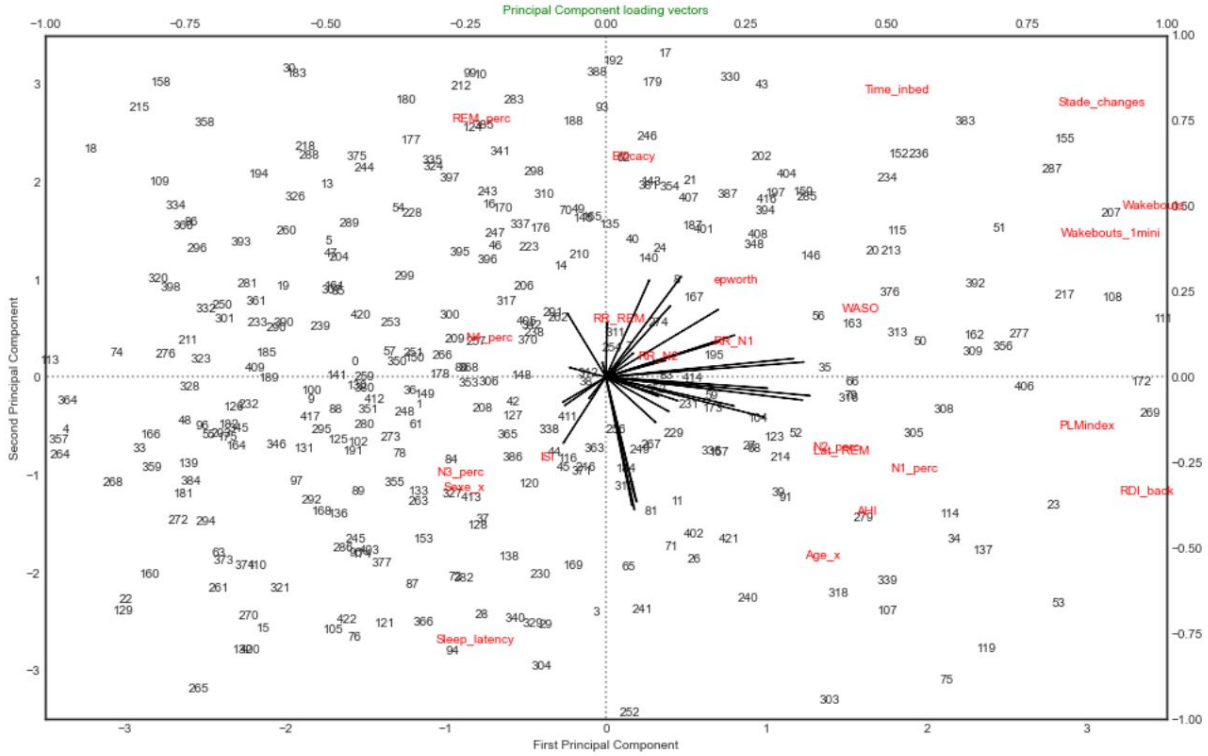


Figure 3.20: The PCA for DIII-PDG is presented in these figures. The x-axis corresponds to the first component and the y-axis to the second component. We can see that the five features on the upper right are positively correlated with each other and contribute significantly to the first and second principal components. It is interesting to note that Time in Bed (TIB) is correlated to the sleep fragmentation illustrated by Wakebouts above one minute. On the lower right, it's interesting that RDI back and PLM are positively correlated, but they show an inverse relation with the second principal component compared to the first group of features. These features might provide contrasting information to the first group. These features negatively correlate with the REM sleep percentage, possibly impacting sleep structure. Finally, we can see that sleep latency is negatively correlated with TIB. This negative correlation appeared very significant in our dataset.

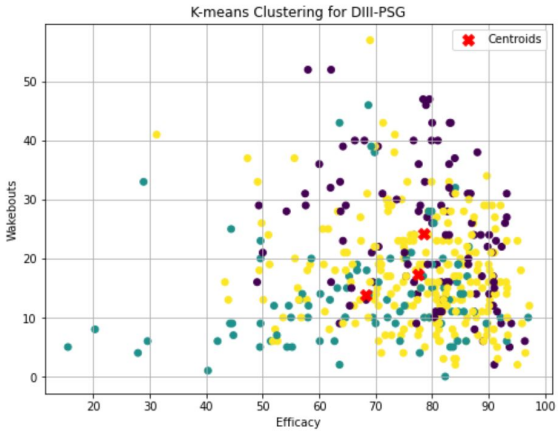


Figure 3.21: Kmeans clustering for the PSG features with K=3, x-axis = sleep efficacy in %, y-axis) wakebouts number

proportion of Sleep Apneas in the final datasets used for ParI prediction (dataset four) and treatment outcome prediction (dataset five), the percentage of AHI index > 10 are respectively 16.7% and 4.4%. Concerning the impact of MPJ, there is a clearer cut-off corresponding to the usual recommendation of 15 (See 3.4). The percentage of MPJ index > 15 is respectively 15.3% and 7%. We will see in the corresponding chapter the impact of having kept these subjects in the sample, mainly of the ParI prediction where the percentage is higher. It would be interesting to see if these factors would be considered.

Concerning the PCA and clustering techniques, we could observe subgroups of patients with dif-

ferent wakebout numbers and Time in Bed as discriminant features inversely correlated with sleep latency.

3.3.4 Database IV (DIV-AG)

D-ACTI is the database gathering all the feature's values recorded by an actigraph worn continuously during an entire week by 350 patients. For this EDA, we will describe the 335 patients selected in dataset four (see 3.1), and we added socio-economical features values and sleep logs features values to better understand the activity-rest cycle according to labor activity. The detailed feature's definition can be found in the Appendix in Tables B.5, B.6. We reduced the different information for the socio-economic status in four quantitative values with the label MDV (Mode De Vie). We assume to make arbitrary quantitative gradation corresponding to:

1. MDV = 0 for unemployed and single
2. MDV = 1 for unemployed and in a relationship
3. MDV = 2 for employed and single
4. MDV = 3 for employed and in a relationship

Distribution

This choice is arbitrary but aims to investigate the correlation between supposed heavy constraints for the MDV = 3 to relatively low in the case of MDV = 0.

Figure 3.22 shows the distribution of the various features.

General Statistics

	mean	std	min	25%	50%	75%	max
MDV	2.21	0.96	0.00	2.00	2.00	3.00	3.00
H_coucher	1421.20	74.77	1211.00	1378.00	1409.00	1450.50	1972.00
Esti_tps_eNaNormt	100.69	411.63	0.00	21.00	38.00	68.00	3420.00
NB_reveil/nuit	2.19	1.93	-1.00	1.00	2.00	3.00	18.00
Esti_duree_Tot_reveil_pdt_som	188.01	575.14	0.00	30.00	60.00	124.00	3420.00
Tps_Tot_Som_esti_pdt_nuit	324.63	93.45	13.00	262.50	330.00	390.00	585.00
Tps_passe_hors_lit	31.86	92.60	0.00	0.00	3.00	11.00	400.00
H_sortie_lit	487.53	85.18	244.00	443.00	476.00	520.00	865.00
Assumed_sleep	483.59	82.82	1.00	446.50	482.00	521.00	1314.00
Actual_sleep_time	416.24	66.20	21.00	386.00	419.00	454.00	661.00
Mean_sleep_last_day	366.98	75.58	30.00	326.50	378.00	416.50	555.00
log_lastnight_dur	277.48	108.54	0.00	210.00	300.00	360.00	545.00
Sleep_latency	21.33	50.70	0.00	6.00	13.00	23.00	830.00
Wake_bouts	26.30	9.62	0.43	19.26	25.88	31.36	87.57
Mean_sleep_bout_time	23.50	69.89	4.00	13.00	17.00	23.00	1280.00
Mean_wake_bout_time	2.49	9.27	0.00	1.00	2.00	2.00	170.00
Immibile_mins	405.81	60.70	221.40	373.94	407.25	442.38	635.43
Mouving_mins	77.82	31.90	11.50	56.59	71.86	96.48	201.00
Nb_of_immobile_phases	42.58	13.47	10.14	33.94	42.00	49.29	110.86
Mean_length_immobility	10.80	5.54	3.29	7.86	9.71	11.94	61.00
One_minute_immobiloty	8.39	5.71	0.00	4.94	7.43	10.57	66.14
Tot_activity_score	12963.52	57728.20	860.57	5002.00	7465.43	10636.19	860151.57
Mean_activity_score	26.84	130.24	1.86	9.54	14.52	22.27	1865.71
Mean_score_in_active_periods	182.65	860.43	31.88	77.74	95.50	118.32	10442.29
Fragm_index	33.00	11.64	2.88	25.20	31.43	40.14	82.80
Avg_wake_mvmt	252.26	316.99	0.00	169.88	218.29	262.62	4184.43
Interdaily_stability	0.51	0.12	0.07	0.43	0.53	0.59	0.84
Intradaily_variability	0.87	0.20	0.36	0.73	0.84	0.99	1.75
Lowest_5h_count	1111.89	918.20	94.00	561.00	855.00	1384.50	7321.00
Max_10h_count	15450.19	6068.29	864.00	11606.50	14543.00	18427.50	56424.00

Table 3.6: General statistics on the Database IV (AG) with additional features from sleep log (yellow) and work/marital status (grey)

Correlations analysis

The previous results showed that some specific scales increased the T-score means, like Hy, D, or H5K. Others showed particular distributions on the dataset and were linked to the treatment outcome, like Es and TRT. We saw that the feature FB is almost sufficient to represent three clusters according

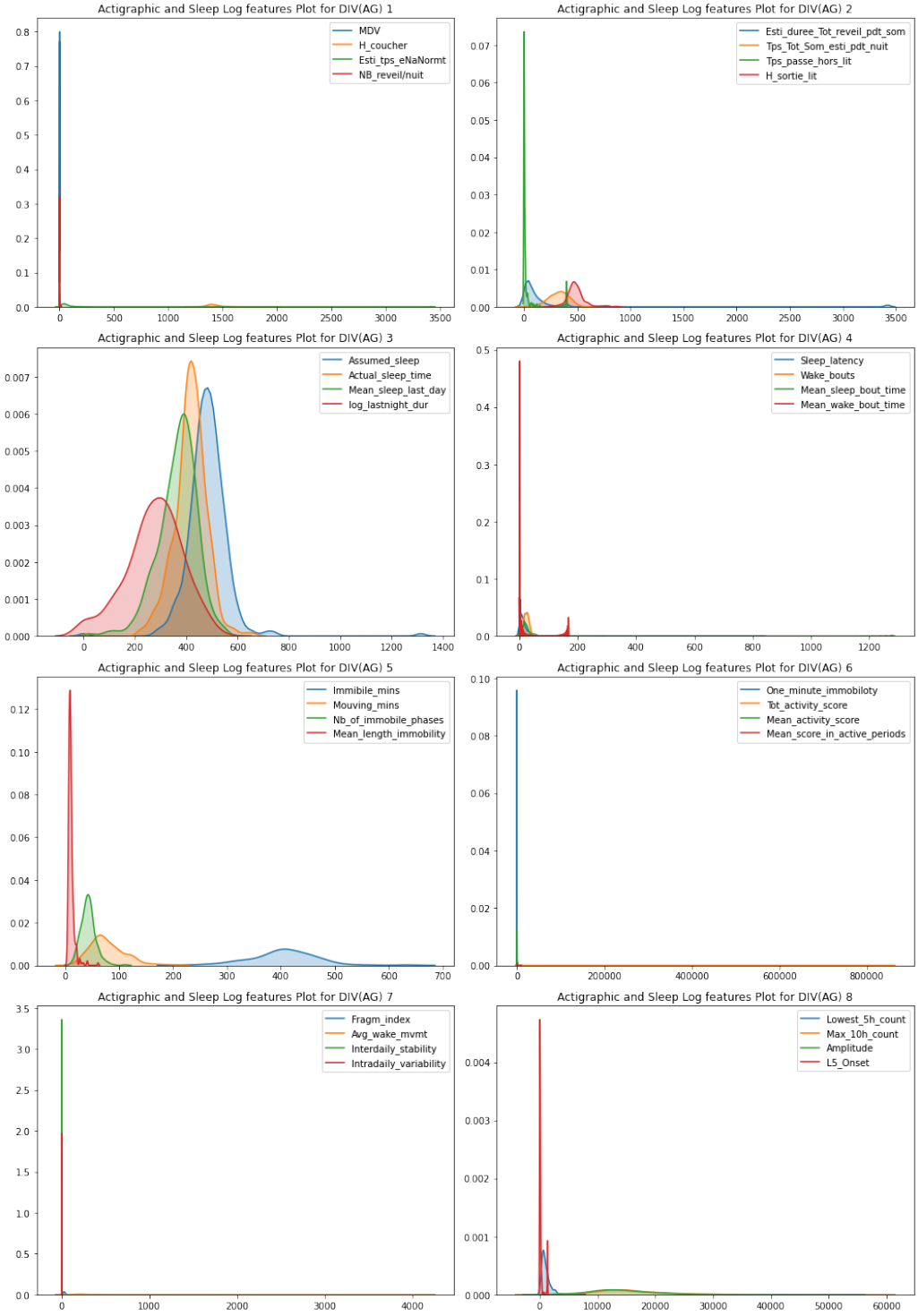


Figure 3.22: Non normalized Distribution of the raw values obtained with the actigraph. We can see that the visual distribution is difficult to assess due to the wide disparity of the value scales. Indeed, there is a mixture of cumulative values, such as the number of movements per 24 h, and transformed values, such as total sleep time, which makes it difficult to visualize the features together. In the next figure, we have normalized all the values to make them more comparable.

to its T-Score on our dataset. We present in Figure 3.3 a specific focus on the Pearson correlation between these features.

The results show a high correlation between FB and the feature TRT, so it's possible that the cluster with the highest FB T-score could be the one with the most treatment resistance. TRT and Es are anti-correlated in a significant way, which means that these two scales are linked to a profile. The correlation between Fb and Es is weaker, so iEs must cover a wider field than the treatment outcome. On the other hand, the highest T-score mean in our sample, Hy, D, and Hs, are poorly correlated to FB, so we have effectively different clusters of psychological profiles. In detail, we could even observe

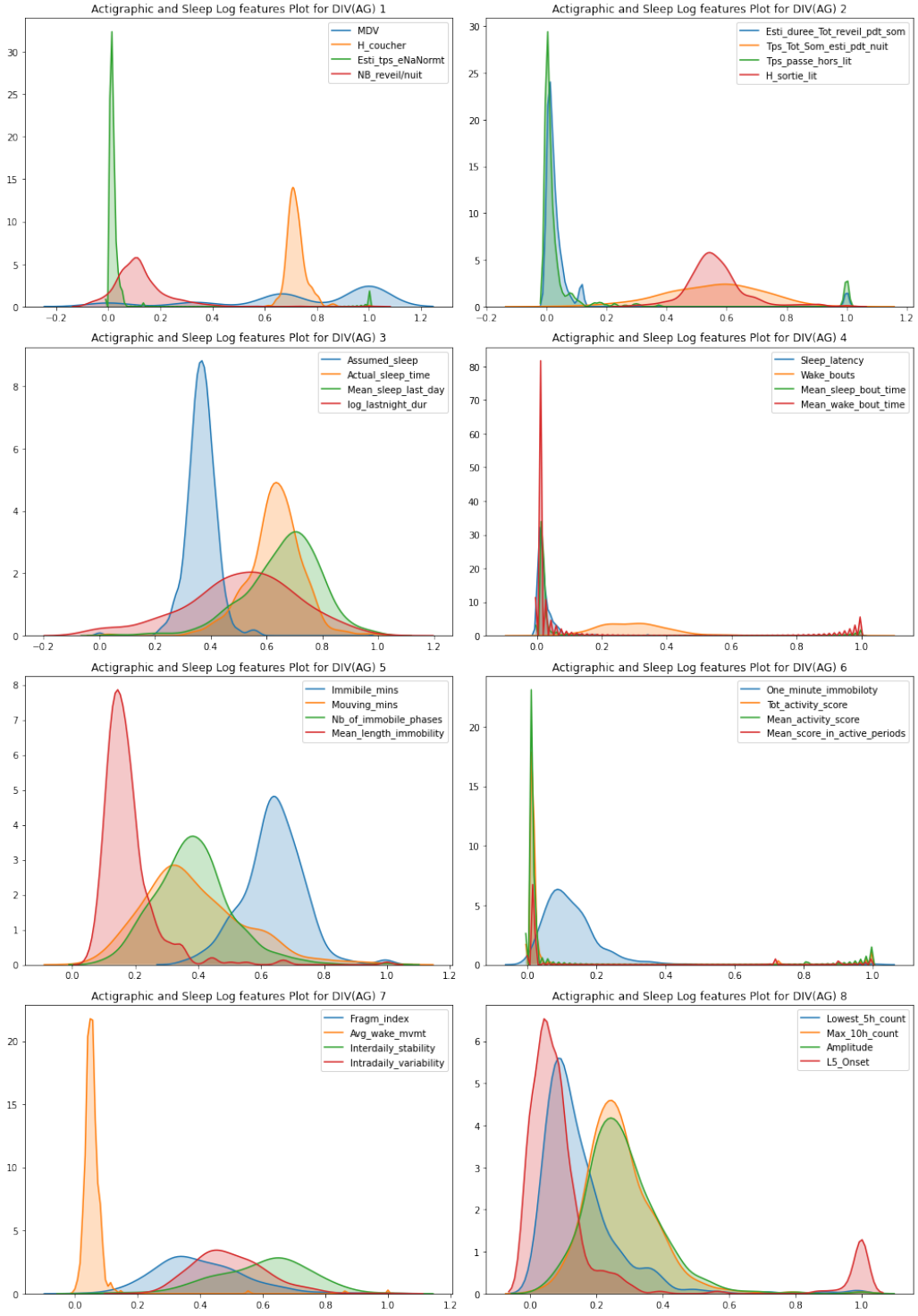


Figure 3.23: Distribution of the normalized values obtained with the actigraph. We can see here the visual distribution of the different values. In plot eight, we see two groups related to the L5onset (the period of the five more stable hours of sleep)

a cluster in the cluster with Hy and Hs5K poorly correlated to TRT when D does. This observation could lead to the interpretation that being depressed is a risk factor for poor treatment outcomes.

Dimensionality reduction and clustering

The main results for 3D PCA are presented in Figure 3.25. We couldn't find significant t-SNE results after perplexity tuning from three to 100. For K-means clustering, we tested 6 clusters. Figure 3.26 presents the visually discriminative results.

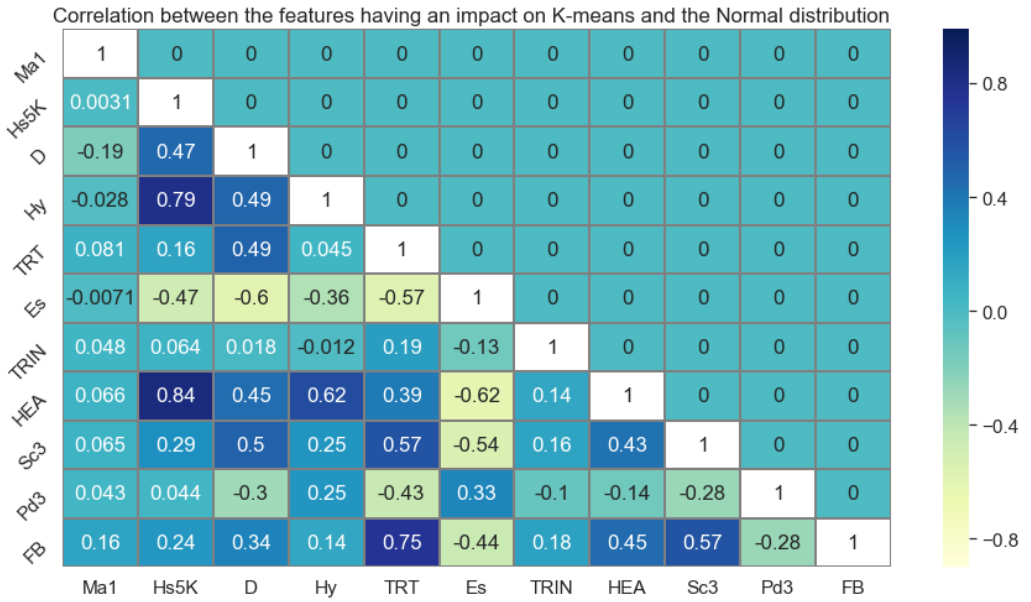


Figure 3.24: Pearson correlation analysis between scales with increased T-score means(Hy, D, and H5K), the ones linked to a negative treatment outcome (Es, TRT), and FB heavily involved in the three clusters found in our dataset

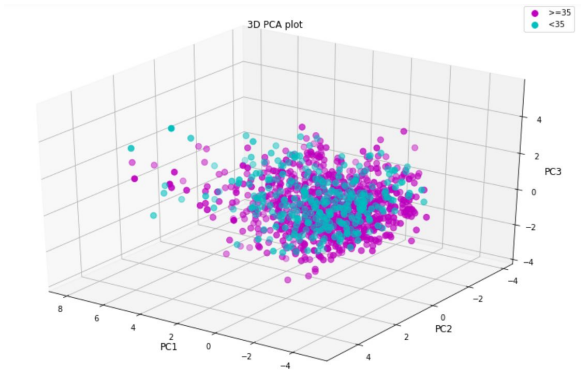


Figure 3.25: PCA on the 10 main scales on the MMPI-2 dataset with age and gender. We found after several experiments that only age could slightly affect the PCA representation, below or above 35. But the PCA here doesn't help much.

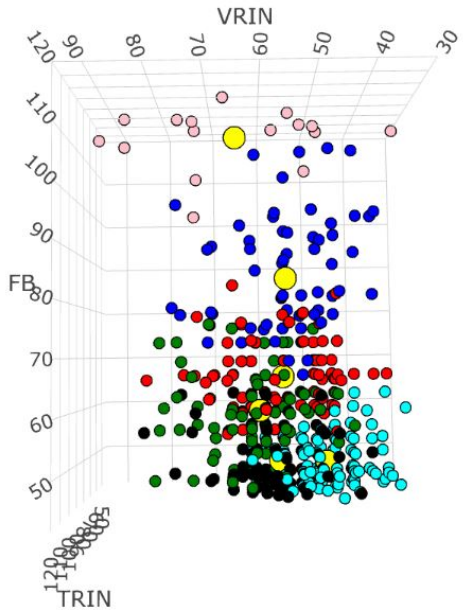


Figure 3.26: K-means clustering in 3D for K=6 for MMPI-2 Tscore with the validity scale VRIN, TRIN and Fb. We could see three distinct cluster regions almost entirely linked to the Fb scale. According to [26], when the Fb T-score is above 70, it is potentially linked to fatigue, attention disorders, a tendency to show oneself in an unfavorable light, or a sign of severe psychopathology.

3.3.5 Dataset 4 and 5

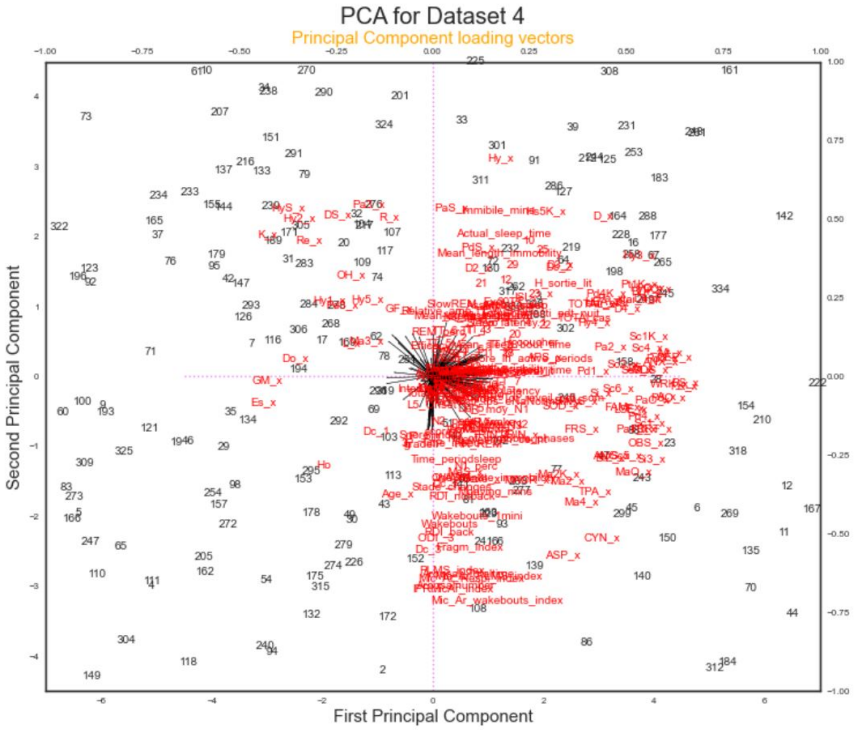
In the forthcoming Chapters 5 (Explaining Negative Sleep State Misperception) and 6 (Explaining Therapeutics Issues), a comprehensive exposition will be provided for datasets 4 and 5, which represent the culmination of the aggregation of databases I through IV, and databases I through III, respectively. Before this detailed discussion, exploratory unsupervised modeling was conducted to determine the preliminary cluster formations within these datasets. The initial findings are depicted in Figures 3.27 and 3.28.

Upon examination through Principal Component Analysis (PCA), datasets 4 and 5 exhibit consid-

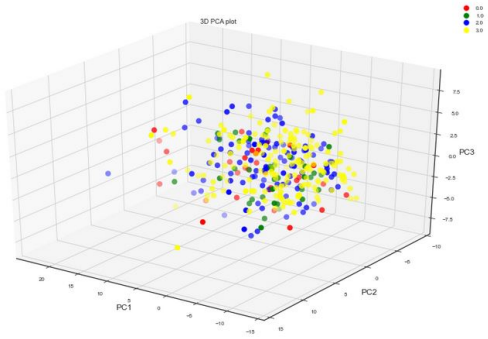
erable similarities, with minor differences. Specifically, dataset 4, which amalgamates data from 335 patients across databases I to IV, appears to be particularly influenced by the MMPI2 scales, especially the Dominance scale. Further investigation will ascertain the impact of this scale on Paradoxical Insomnia (ParI). However, when applied based on socioeconomic criteria, additional analytical techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE) and K-Means Clustering failed to yield further insights.

Conversely, dataset 5—which combines data from 423 patients across databases I to III—shares similarities with dataset 4 as indicated by PCA but does not demonstrate a notable influence from the Dominance scale. Interestingly, the t-SNE visualization with a perplexity setting of 30 suggested the presence of two minor clusters.

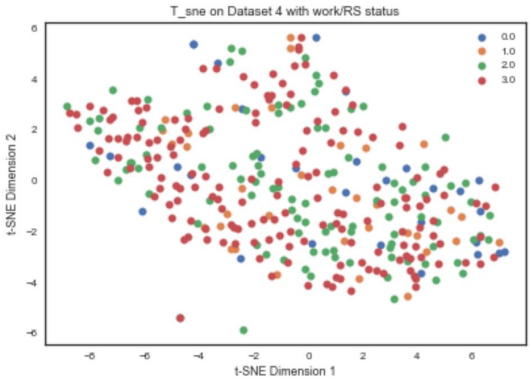
Collectively, these analytical approaches suggest the existence of at least two to three distinct patient subgroups within the datasets, providing a foundational understanding that will be elaborated upon in the detailed analyses in the aforementioned chapters.



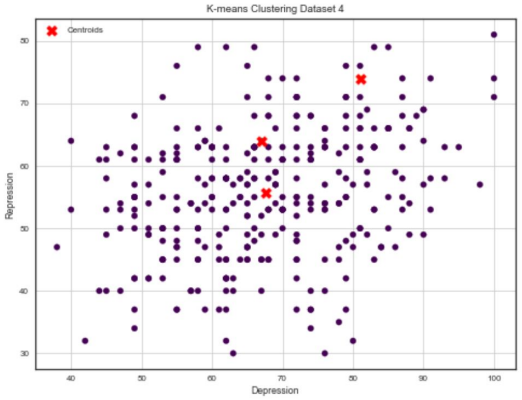
(a) PCA with all the features in the dataset 4



(b) PCA with all the features in dataset 4 according to the socio-economic status.

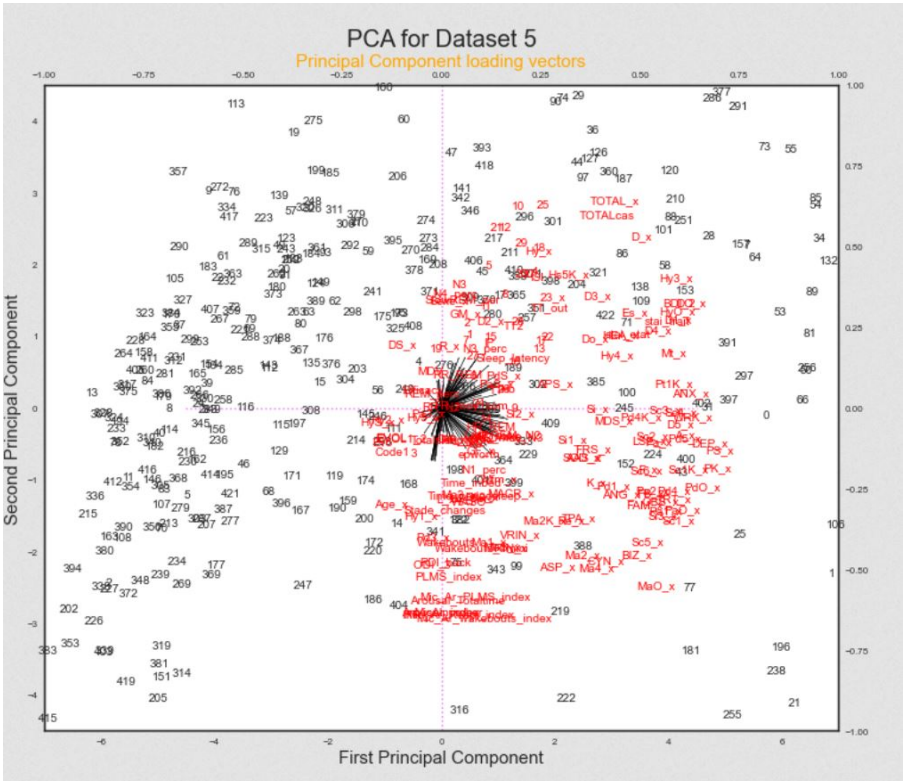


(c) T-sne with all the features in dataset 4

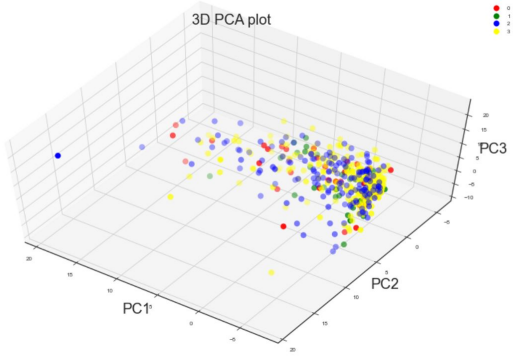


(d) K-means clustering according to MMPI scale D and R, K=3

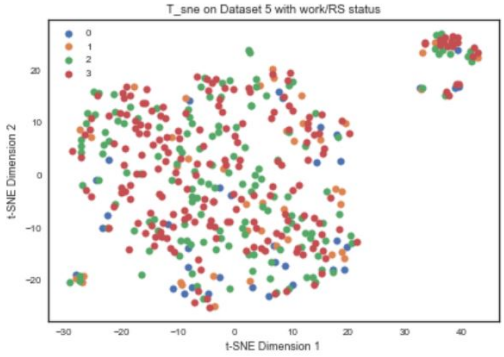
Figure 3.27: Dimensionality reduction and clustering on Dataset 4



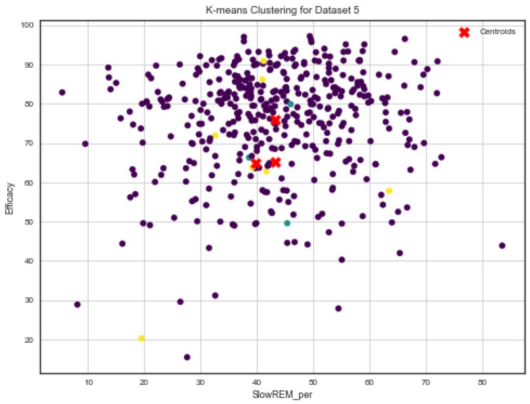
(a) PCA with all the features in the dataset 5



(b) PCA with all the features in dataset 5 according to the socio-economic status.



(c) T-sne with all the features in dataset 5



(d) K-means clustering according to percentage of SWS and REM sleep D and R, K=3

Figure 3.28: Dimensionality reduction and clustering on data 5

3.4 Evaluation of the Relevance of Datasets to the Assumptions Made

3.4.1 Sample representativeness

Age and gender

Significant studies in general population samples diagnosed with indistinct Insomnia based on standardized questionnaires showed across different countries: -In China, a study [32] on 14263 subjects found a mean age of 43.7 y.o of and 55% of female proportion. -In Canada, a study [148] on 2000 subjects found a mean age of 48.6 ±12 yo and 60.5% of female proportion. -In France, a study [15] on 7844 subjects found a mean age of 46.2 yo±13 yo and 56% of female proportion.

Significant studies specific to primary Insomnia samples (especially without psychiatric comorbidity) showed: -On 567 subjects with Primary Insomnia, a study [66] found a mean age of 58 y.o ±14of and 59% of female proportion. -On 429 subjects with severe Insomnia in the general population, a study [87] found a mean age of 45 yo (only age class available) and 65% of female proportion. -On 283 subjects (124 clinical patients (chronic primary Insomnia) and 159 research(primary Insomnia or comorbid Insomnia (mostly anxiety and depression)), a study [146] found a mean age of 46.6 yo (± 10.0, 20 to 71 yo) and 59.5% female proportion.

Significant studies of so-called comorbid Insomnia showed: - A meta-analysis of 23 studies involving 1379 patients [71] found a mean age of 53.0 yo (± 10.0 y), and the female proportion was 66.5%. So, our sample is close in terms of age and gender of the target population studies in the literature, especially those corresponding to severe Insomnia or mixed primary-secondary Insomnia with possible sub-clinical anxiety and depression issues.

Psychological profile

MMPI scores

A study [217] on 199 patients compared the MMPI-2 scales for four types of profiles, psychiatric, PsyI, ParI, and nocturnal myoclonus. The graphical results of the main scores are shown in Figure 3.29. The results from 3.1 were reported on this subtype differentiation. Our sample is highly superimposable on the two intermediate curves corresponding to the two primary types of Insomnia, PsyI (Pp on the figure) and ParI (NOFon the figure). With a more detailed analysis, our population doesn't share any highly significant score with the psychiatric Insomnia (Ps) profile with a T score above 70 in five scales. So, our sample aligns with age, gender, and psychological profile with the thesis objective, describing ParI, initially classified as primary Insomnia, as we discussed previously. A recent review found that the scale usually increased in chronic or primary Insomnia studies are Hs5K, D, Hy, and Pt1K, and the mean score for these scales. Finally, a few studies tried to make clustering, like [56], which found two subtypes with few patients. Nevertheless, as they found more than 20 subjects per group, the results are interesting according to [50]. Indeed, in their paper studying different clustering analyses, like K-Means or Fuzzy clustering, they found that clustering outcomes were mostly unaffected by differences in covariance structure. Sufficient statistical power was achieved with small samples (N=20 per subgroup). In Table 3.7, we compared the scores of this study with our global results. Again, we found very similar results except for the Hs scale. However, although this sample is the same in terms of mean age (45.7 yo (±15.5)), the gender repartition is not in line with most of the studies as the female proportion is only 45%. This could explain why Hs is lower, but still, this result aligns with the usual finding on Primary Insomniacs patients described in [112].

Insomnia type	Number	Hs	D	Hy	Pd	Pa	Pt	Sc	Ma	Si
PI	88	59.7	64.75	65.38	60.0	59.0	63.0	60.0	53.0	55.0
CI	1182	66.4	67.0	65.9	58.5	59.6	63.2	60.50	50.4	56.1

Table 3.7: CI = Chronic Insomnia PI = Primary Insomnia

	SOL	WASO	TST	SE	N1_perc	N2_perc	N3_perc	N4_perc	REM_perc
mean	33.47	64.52	351.80	75.63	4.37	52.23	18.19	6.14	19.00
std	33.52	56.95	79.33	13.71	4.10	12.27	8.57	7.84	7.24
min	0.00	0.00	74.00	15.50	0.00	8.80	0.00	0.00	0.00
25%	12.10	24.50	301.50	68.85	1.70	43.65	12.55	0.00	13.95
50%	23.80	48.00	358.00	79.10	3.00	52.20	17.10	2.00	18.50
75%	41.45	87.00	404.50	85.05	5.90	60.35	22.60	10.45	23.90
max	214.10	429.50	642.00	97.20	29.70	90.90	46.00	45.60	42.60

Table 3.8: Main PSG features description in the dataset five

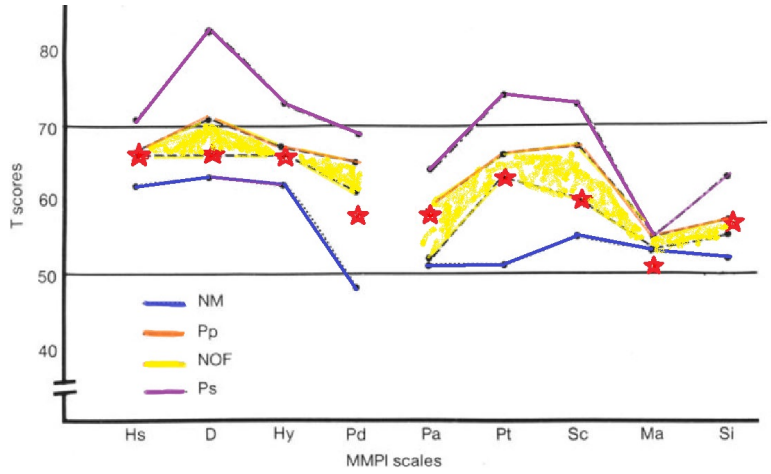


Figure 3.29: The diagram shows target diagnostic categories, denoted in yellow, which are nearly perfectly superimposed with our sample mean of MMPI elevations (red stars). The diagnostic categories include NM (Nocturnal Myoclonus), Pp (Psychophysiological Disorder), NOF (No Objective Findings), and the MMPI scales are Ps (Psychiatric Disorder), Hs (Hypochondriasis), D (Depression), Hy (Hysteria), Pd (Psychopathic Deviate), Mf (Masculinity-Femininity), Pa (Paranoia), Pt (Psychasthenia), Sc (Schizophrenia), Ma (Mania), and Si (Social Introversion) from [217].

Specific and general questionnaires usually used in CID evaluation

Subjective index of Insomnia severity on the sample On the DII-DBAS, our sample is in line with the major studies on Insomnia Indeed, with a mean of 19.4 ± 4 to the ISI scale, the mean score is comparable to [41] who found a mean score of 17.51 (± 4.41) on 250 clinical Sample patients (Mean age 49.6 (± 13.65), women 55%), and [100] who found a mean score of 18.4 (± 3.7) and 19.4 (± 4.1) on two clinical groups (N = 49 and 51) with a mean age of age 41.4 ± 10.5 and 41.3 ± 12.5, female 71% and 79%. Also in line with [149].

Subjective depression score on the sample For the depression inventory, our results are also in line with the literature. Indeed, as mentioned earlier, a score of 17 seems to be the cut-off for depression detection; with 16.8, our sample is just below. **Subjective anxiety score on the sample** For the STAI-trait and state, our sample is closer to the population of Insomniacs without too much stress described in [119], which is in line with all the others scores obtained in particular in the MMPI-2 scale, adding a stone to the representativeness of our sample to study primary Insomniacs subtypes.

Polysomnographic evaluation

We compared the main PSG values in dataset five (N=423) to several studies on CID and Primary Insomnia. The features used for the comparison are described in Table 3.8.

We compared the mean and SD to three studies on CID. We can see the visual comparison in Figures 3.30, 3.31 and 3.32. We can see that the mean overlapped with most of the studies on CID in terms of TST and sleep stage percentage. We can see that the SD of WASO is quite high in our sample compared to [10] and [94], but our sample size is almost 10 times each of these studies; this could explain this variability.

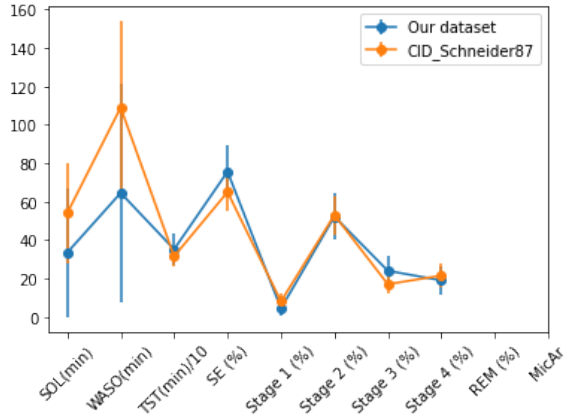


Figure 3.30: Visual comparison of the eight main PSG features between our dataset and a population of indistinct CID (N=16) [185]

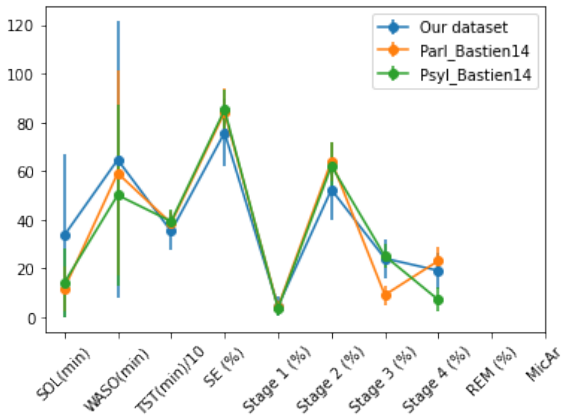


Figure 3.31: Visual comparison of the eight main PSG features between our dataset and a population of ParI and PsyI (N=30 for PsyI, N=28 for ParI)[11]

Figure 3.32: Visual comparison of the eight main PSG features between our dataset and a population of ParI and PsyI (N=26 for PsyI, N=20 for ParI) [94]

3.5 Conclusions

Our visualizations showed that our dataset highly represented the populations of chronic Insomniacs described in the literature, especially primary Insomnia. Indeed, we have shown that the psychological profiles of our patients correspond to so-called primary Insomniacs and that our sample is, therefore, suitable for studying the determinants of Paradoxical Insomnia. We showed that the linear (PCA) and non-linear (t-sne) visualization methods could allow us to find subgroups in datset4 and 5. For the MMPI-2 scales, we also found three clusters on 1182 patients.

On a representative dataset of a population of 713 chronic Insomniacs, we were able to find subgroups of patients in terms of rumination and preoccupation identified by the DBAS according to scores on the MMPI-2 psychological assessment scale, thanks to a prediction tool (Decision Tree). This is the first study to predict preoccupation and rumination around sleep using the MMPI-2. The next step will be to create a specific scale in the MMPI-2 for Insomnia. This finding seems confirmed by a clustering method finding three subgroups of patients similar to the clustering produced on the 1182 MMPI-2 results. We also show that the subscale of the MMPI-2 is useful to predict the different subgroups; until now, only the 10 basic scales have been used in the various studies cited previously.

Chapter 4

Investigating Machine Learning Tools for EEG Analysis in Sleep Medicine

The purpose of this chapter is to assess the feasibility and reliability of ML tools for analyzing electroencephalogram (EEG), particularly for standardized sleep scoring. A particular aspect we look at is spindle detection. Finally, we look into clustering the features we extracted with tools in our analysis. The global objective is to standardize the sleep analysis for ParI characterization and phenotyping.

Chapter Highlights : Five experiments on EEG, from automatic sleep stage scoring, to spindles detection until EEG features engineering for CID clustersization and subtyping

1. Automated Sleep Scoring In this exploratory study, we are testing the primary hypothesis that an EEG signal detection algorithm, which employs Empiric Mode Decomposition (EMD) and bandwidth filtering, can accurately score sleep stages based on varying qualities of sleep EEG (either clean or noisy), achieving results comparable to those of expert scoring. The data used are 2*1000 epochs from six EEG channels, one from corrupted and one from good PSG. The Data preprocessing include low and high band filter (0.3-30 HZ).

2. Convolutionnal Neural Network (CNN) for automated sleep scoring In this study, we are testing the primary hypothesis that an EEG signal detection algorithm, which employs Convolutional Neural Network (CNN) and bandwidth filtering, can accurately score sleep stages, achieving results comparable to those of expert scoring. The goal is to evaluate the reliability and transferability of such an algorithm. The algorithm is tested from different sources (our dataset, MASS, and Sleepphysionet) with 60 PSG recordings each. The Data preprocessing include low and high band filter (0.3-30 HZ).

3. Benchmark for Spindles Detection: In this study, we used micro biomarkers useful in phenotyping but fastidious to detect manually, specifically **sleep spindles**, which have already been used to characterize ParI in previous studies; we test and compare algorithms on our dataset to assess their reliability. The dataset here is one single night from a young patient, with high-quality signals and expert scoring of the ground truth. The Data preprocessing include low and high band filter (0.3-30 HZ).

4. Spindles and Personality Prediction: The main hypothesis is that sleep spindles, as a biomarker, **can predict predispositions towards certain psychopathological traits** important to understand CID clusters or subtypes. The characteristics of sleep spindles (including their density, average duration, average frequency, and the average number of oscillations) can be used to predict the occurrence of a patient's predisposition towards certain psychopathological traits as assessed by the MMPI-2 questionnaire. we used spindles detection algorithms on 267 subjects from our dataset with density and duration detection of spindles.The Data preprocessing include low and high band filter (0.3-30 HZ).

5. Subtyping Insomniacs with Significant Difference in Subjective Sleepiness using Graph Spectral Theory and clustering techniques on raw EEG and hypnogram scored by expert The main hypothesis is that we could find clusters (linked to specific subtypes) of CID patients with a significant difference in terms of subjective sleepiness (ESS questionnaire score) and insomnia severity (ISI questionnaire score) using Graph Spectral Theory and clustering techniques on raw EEG and hypnograms scored by sleep experts from 386 PSG recording. The Data preprocessing include low and high band filter (0.3-30 HZ).

Key Terms and concepts

Acronym/term	Definition	Ref.
AdamOptimizer	Adaptive Moment Estimation	p. 171 (B.1.2)
CNN	Convolutional Neural Networks	p. 171 (B.1.2)
EDF	European Data Format	p. 169 (B.1.1)
EMD	Empirical Method Decomposition	p. 172 (B.1.2)
ESS	Epworth Sleepiness Scale	p. 31 (2.4.1)
ICA	Independent Component Analysis	p. 173 (B.1.2)
IMF	Intrinsic Mode Functions	p. 172 (B.1.2)
ISI	Insomnia Severity Index	p. 31 (2.4.1)
MASS	MASS Sleep Dataset	p. 169 (B.1.1)
MMPI	Minnesota Multiphasic Personality Inventory	p. 32 (2.4.1)
PSD	Power Spectral Density	p. 175 (B.1.2)

4.1 Automated Sleep Scoring

The main objective of this section is to get an idea of the feasibility of automatic sleep stage detection on our dataset. At the start of our thesis, this need was in anticipation of a possible extension of our ParI and Treatment outcome predictions to other datasets. In this case, being able to replicate and ideally explain precisely how these sleep stages were scored would increase the reliability of our results and avoid criticism of the reliability of expert scoring. Indeed, our interest in the automatic scoring of sleep stages lies in the desire to standardize analysis so that they can be extended to larger databases and not depend on inter-rater variations, which can complicate comparisons between datasets. A recent meta-analysis [121] on the subject confirms analyzing 11 studies that Cohen's kappa for manual, overall sleep scoring was 0.76 (0.71-0.81). In another four studies by sleep stage, the agreement was 0.70, 0.24, 0.57, 0.57, and 0.69 for the W, N1, N2, N3, and R stages, respectively. These results reflect that sleep scoring can sometimes prove difficult due to the multiplicity of parameters to be considered and the variability of EEG aspect as a function of age, associated pathologies, medication, or artifacts. So, if we could have an algorithm powerful enough to exceed 0.8 (more than manual scoring agreement predictions), and one that we could also explain, this would be ideal for subsequent replication of our studies.

4.1.1 Characterization of sleep states with EEG pattern detection according to the quality: a proof of concept study

Introduction and hypothesis genesis

We did a first proof of concept (PoC) study as a preliminary investigation with small-scale data exploration. We chose this first experiment because our sleep acquisition software (BrainRT) was designed to enable interfacing between clinical recording and research and allow us to implement specific features to analyze the signal. After contacting the BrainRT RD team, we could access the algorithm based on empirical method decomposition (EMD). After the state-of-the-art review, we could see that this technique could be efficient in noisy recording. So we decided to make a brief PoC on a noisy and good-quality recording to see if we could reliably use this algorithm to standardize the analysis and earn some time.

Background

The automatic sleep scoring algorithm used in this experiment is a mix of automatic denoising algorithms using Empirical Method Decomposition (EMD), notch filtering for electrical currents, Electrocardiogram filtering, spindles, delta, and alpha detection with EMD, and handcraft denoising setting criteria to limit and classify as artifact part of the signal exceeding some limits in-band frequency spectrum analysis, power, duration, amplitude. The setting allows the implementation of AASM [90] criteria for sleep scoring, like the percentage of Slow Wave Sleep or alpha in one single epoch, to be classified in the proper sleep stages.

The issue was whether it made sense to test this particular algorithm about our goal of achieving relatively reproducible and reliable results, even on artifact-based tracings. We have to analyse the state-of-the-art literature on the subject to first answer this question.

The question here is to evaluate the interest of EMD in the case of noisy signals. Usually, conventional time-frequency representation algorithms, such as the short-time Fourier transform and continuous wavelet transform, have been commonly employed to help in EEG quantification. However, these methods limit the time-frequency representations by relying on predefined sets of basis functions. This fixed arrangement may not fully capture the characteristics of the data and especially in bad quality data where "non" desired figures named artifacts could appear, especially in ambulatory analysis. We already know from experience with expert scorers that the sleep EEG, especially in ambulatory recording, could be very noisy. Indeed, noise could come from multiple origins, internal or external. The internal origin could be eye blink or movement, ECG pulse, chewing, swallowing, clenching, sniffing, talking, scalp contraction, etc.. The external could be electrode displacement and pop-up, cable movement, poor ground, electrical or magnetic or sound waves, body movements, etc.[92] In figure 4.1, we can see a moderately corrupted EEG as an example in[92].

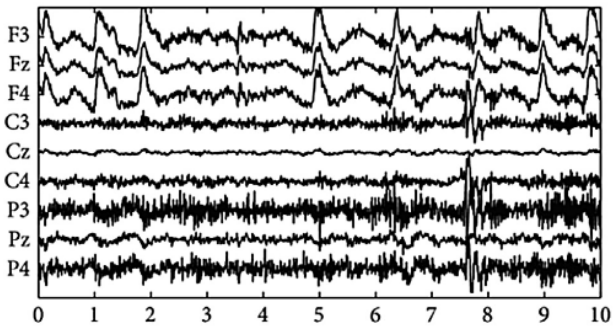


Figure 4.1: Visual appearance of a bad quality EEG trace on 10-sec scalp EEG segment (all channels corrupted with muscle activity) [92]

The hypothesis that we could use automatic scoring on the corrupted recording was encouraged by recent findings [44] suggesting that accurate power spectral density estimates strongly resembling the output of visual scoring can be achieved by very simple detection algorithms detection and claiming that the method was robust against inter-/intra-subjects and raters variability. At the time of our

preliminary study in 2017, a reference of [92] extensively reviewed 35 methods for artifact detection from scalp EEG. From this review, the recommendation for ambulatory EEG monitoring, when the number of channels is fewer than standard EEG, was wavelet-based Transform algorithms. It offers good time-frequency localization, making it suitable for analyzing non-stationary signals like EEG. Wavelet denoising techniques can help remove unwanted noise components from the signal while preserving EEG features. Independent Component Analysis was also described as a good statistical method to separate a multichannel signal into statistically independent components. It can effectively isolate EEG sources from artifacts and external noise sources. By assuming the sources are statistically independent, ICA can separate EEG components from other signals recorded simultaneously, such as muscle artifacts. But around the same period, a paper [83] found that using EMD could perform even better than these two techniques on EEG.

So without further research, we decided to test the performance of this algorithm on our data. The methodology and the algorithm pseudo-code will be described in the section methodology.

Methodology

Rationale for our methodology choice The work of Coppieters et al. inspires our study's methodology [44]. They devised a four-step protocol, examining the algorithm's performance on one to six different nights' data from 35 subjects. Our aim differs from theirs, as we intend to predict sleep stages while considering artifacts using heterogeneous data. The Coppieters protocol begins with six PSG recordings from five healthy young subjects under controlled laboratory conditions. Furthermore, spindle detection was also part of their algorithm. A human rater visually inspected each recording to identify and reject artifacts and arousals. They progressed to the subsequent step only if they achieved positive results in the current step. Their first step already delivered satisfactory results compared to expert scoring (Since there is neither any ground truth data available nor any universal nor standard quantitative metric(s) used in the literature that can capture both amounts of artifact removal and distortion), with sensitivity ($87 \pm 5\%$), Kappa coefficient (0.70 ± 0.15), and average overlap of detected events (0.70 ± 0.10). However, the False Discovery Ratio was less impressive ($39\% \pm 17\%$). We want to achieve similar results in our first phase before going further. Our methodology tested our algorithm on two datasets characterized by extreme differences in quality and artifact levels (Good versus bad). This first step aims to understand the accuracy range our algorithm can achieve compared to the ground truth (expert scoring) in these two recording quality scenarios. This information is crucial to understanding why certain parts of the recording could not be detected, allowing us to adjust the algorithm's tuning for subsequent steps.

Choice of the recording The PSG recording was chosen to represent our sample, i.e. women (mean age 48 ± 4 yo) with no other sleep disorders than CID. The algorithm used all the available signals and electrodes recorded by the two EOG (EOG 1 and EOG2), the EMG, the six EEG electrodes (C3, C4, F3, F4, O1, and 02), and the two Mastoids electrodes M1 and M2. The EEG montage is bipolar, with each EEG electrode associated with the opposed Mastoid electrode. The montage could be viewed in the illustrations in B.7 and B.6.

So one recording was of very good quality, with only minor artifacts when the other was highly corrupted but still interpretable by a sleep scorer expert. The visual aspect of these two recordings for the six EEG can be seen in Figures 4.2 and 4.3.

Description of the algorithmic protocol The algorithm based on Empirical Mode Decomposition (EMD), a data-driven method used for the time-frequency analysis of signals decomposes a signal into intrinsic mode functions (IMFs) or modes. Each IMF represents a specific oscillatory pattern in the signal, providing a localized representation of the signal in the time-frequency domain.

In the EMD algorithm, we performed an iterative decomposition process on an N-point EEG epoch, represented as X. The algorithm identified local extrema, constructed envelopes, and iteratively subtracted the local mean curve from the signal until convergence. The outcome is a set of IMFs representing the time-frequency components of the original EEG signal.

The general formula for EMD decomposition can be written as:

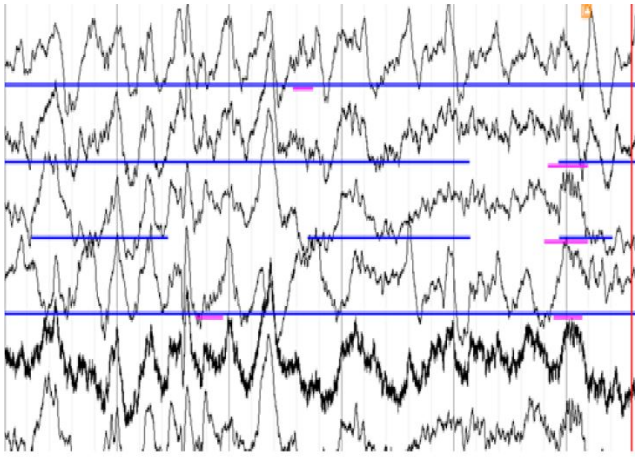


Figure 4.2: Visual appearance of a good-quality EEG trace on F3-M2, C3-M2, and O1-M2 (top) and F4-M1, C4-M1 and O2-M1 (bottom) during 15 seconds of stage N3 with bandwidths 0.3-30 Hz filter. Channel C4-M1 presents light but continuous artifacts secondary to partial electrode detachment, which is common during recording. The Blue lines correspond to Delta rhythm prediction and the Magenta to the sigma

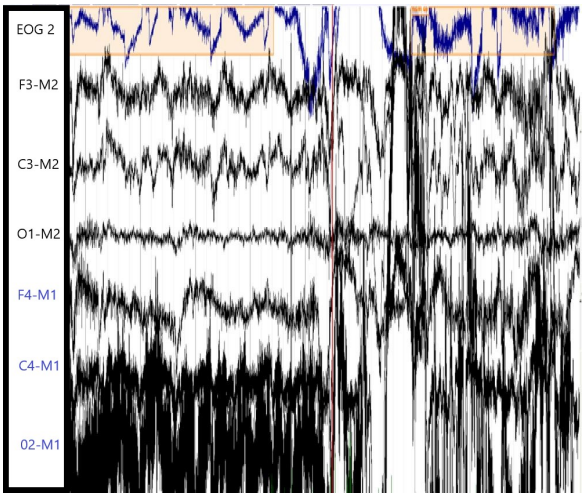


Figure 4.3: Visual appearance of a bad quality EEG trace on all channels but increasing in C4-M1 and 02-M1 (bottom) during 15 seconds of stage W with the application of bandwidths 0.3-30 Hz filter. Note the EOG at the top with artifacts; the orange line is the eye movement detection. These artifacts are secondary to partial to almost complete electrode detachment

$$x(t) = \sum_{i=1}^{N} c_i(t) + r(t) \tag{4.1}$$

Here, $x(t)$ represents the original signal, $c_i(t)$ denotes the i-th IMF, N is the total number of IMFs, and $r(t)$ is the residual signal.

The algorithm has been enhanced for the application of BrainRT. We changed the applied limits on the frequency threshold to be more selective on the artifacts detection. The entire pseudo code with all the parameters tuned is described in Appendix in B.2.1

Description of the performance evaluation Evaluating the performance of a classifier becomes quite challenging when the classes are imbalanced. Indeed, we will evaluate 16 hours of EEG with a majority of wake time and some discrepancy between each sleep stage. To assess the algorithm's performance in detecting sleep stages, we recovered all the 30-second epochs scored by the algorithm with the 5 classes corresponding to the different sleep stages (W, N1, N2, N3, and REM). The two EEG recordings correspond to around 1900 epochs, each assigned one of the 5 classes. We used confusion matrices with raw and percentage (after normalization) of accuracy between the algorithm prediction and the Ground Truth (expert scoring) to compare results between the automatic scoring algorithm and expert scoring.

The schematic generic confusion matrix is depicted in Figure 4.4

		Predicted Label	
		+	−
True Label	+	True Positive (TP)	False Negative (FN) Type II error
	−	False Positive (FP) Type I error	True Negative (TN)

Figure 4.4: Generic confusion matrix for a binary classification

Results

The results are presented in the four confusion matrix in Figures 4.5, 4.6, 4.7 and 4.8.

The following results were noted:

1. For the Good quality recording, the main results are:

    - Even for the good quality recording, the prediction was poor, with a global accuracy of 0.64, essentially due to the high accuracy in predicting W stages. Indeed, the balanced global accuracy is 0.4.
    - For the sleep stages, the accuracy is only 18% for stage N1, 27% for stage N2, 21% for stage N3 and 40% for REM sleep.
    - There was also a tendency for the algorithm to over-predict stage N3 instead of N2.

2. For the poor-quality recording, the main results are:

    - The results are surprisingly identical in terms of global accuracy than for the good quality. However, the balanced accuracy is only 0.32.
    - But looking at the detailed results, the W prediction is still good, but of practically 0% for stages N1, N3, and REM.
    - Stage N2, on the other hand, is very well predicted despite the artifacts with 77% accuracy.
    - More annoying is the prediction of false positives, notably for REM sleep, or N2, which means that the algorithm "invents" sleep in a high percentage during wake time.

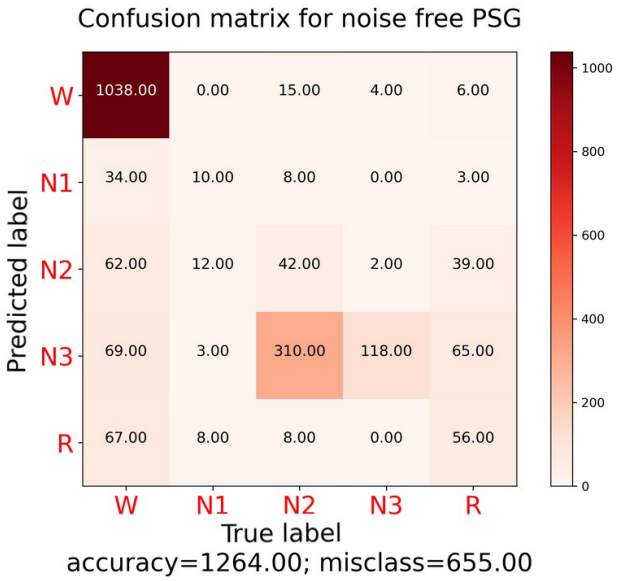


Figure 4.5: Non normalized Confusion Matrix between automatic scoring and visual expert scoring (Good quality recording)

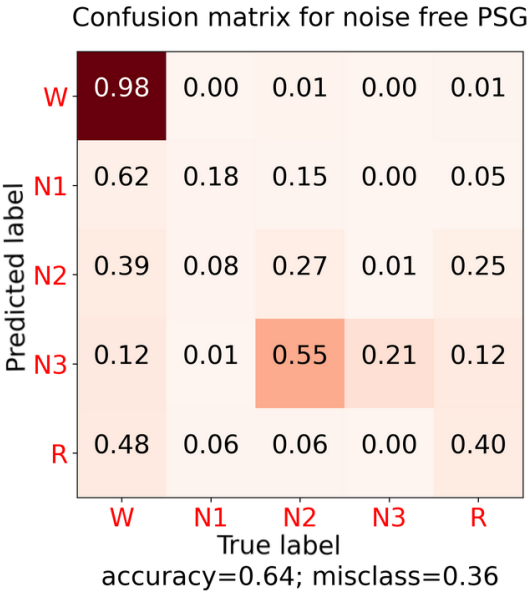


Figure 4.6: Normalized Confusion Matrix between automatic scoring and expert scoring (Good quality recording)

Discussion

The feasibility of automatic sleep stage detection based on our dataset was the primary aim of this study. Our initial anticipation was the potential expansion of our ParI and Treatment outcome predictions to other datasets, emphasising replicability and accuracy of sleep stage scoring to increase the reliability of our results. Our interest was predominantly in the automatic staging of sleep stages to standardise analyses for larger databases and to mitigate inter-rater variations that could potentially complicate comparisons between datasets. The results from our proof-of-concept (PoC) study revealed several important findings.

In our PoC study, we explored small-scale data and incorporated Empirical Mode Decomposition (EMD) for sleep stage scoring. EMD was chosen due to its known effectiveness in noisy recording. To evaluate the utility of EMD in our task, we examined the state-of-the-art literature and adopted an algorithm based on it. The selected algorithm had a mixed design, including automatic denoising algorithms using EMD, notch filtering for electrical currents, and other features.

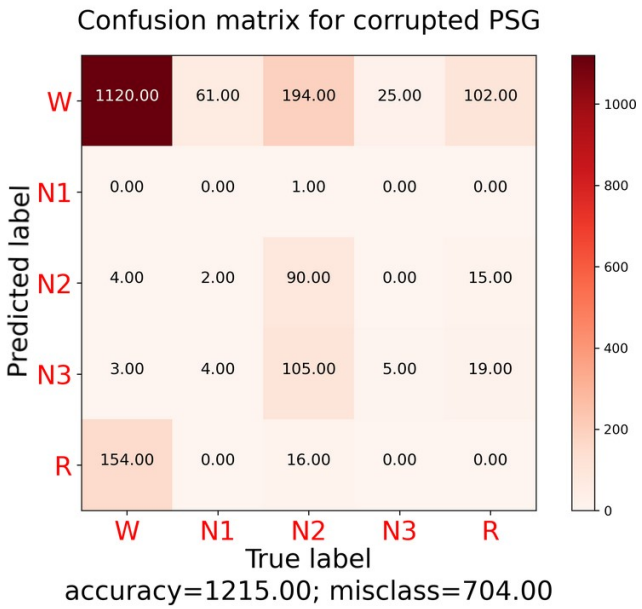


Figure 4.7: Non normalized Confusion Matrix between automatic scoring and visual expert scoring (Corrupted recording)

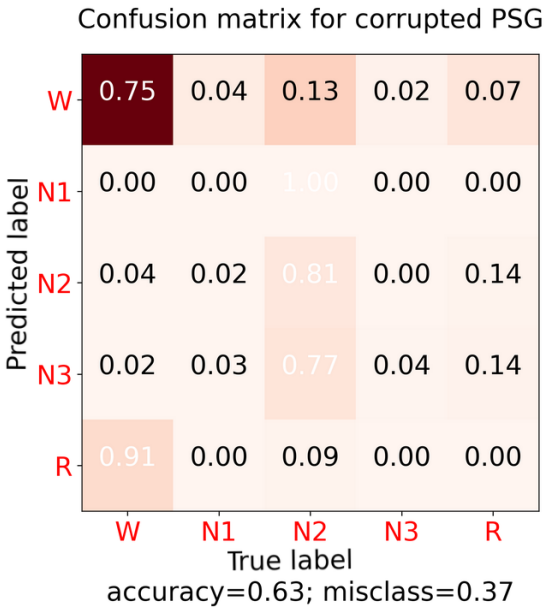


Figure 4.8: Normalized Confusion Matrix between automatic scoring and expert scoring (Corrupted recording)

Our findings were that even with good-quality recordings, the prediction was suboptimal, with an overall accuracy of 0.64 with the same result on noisy recordings. This subpar performance was primarily due to high accuracy in predicting W stages and a tendency to overpredict stage N3 instead of N2. This result brings into question the generalizability of these findings to our dataset.

The algorithm notably excelled at identifying the N2 stage in the noisy environment when it failed on good quality recording. Upon examining the spindle detection component of the algorithm, we found it exhibited strong accuracy in predicting sleep spindles. This consistency aligns with the original design of the algorithm, where considerable emphasis was put on EMD and the Viterbi algorithm to enhance spindle detection. This might also clarify the observed false positives when predicting N3 for N2, as spindles are a common element in both stages. In the case of insufficient detection of slow wave sleep, the algorithm may wrongly consider slow wave sleep and spindles as N3 instead of N2 if delta detection is not performant enough. Nevertheless, spindle detection remains accurate in both cases.

As stated in the hypothesis, if the first phase were unsuccessful, we would stop the experiment.

Limitations

The first limitation is our reliance on a single algorithm for our data analysis, but it was part of the hypothesis as we wanted to test the ones we already had. But of course, given the diverse algorithms available for sleep stage detection, our results may not represent the results achievable with other methods. Furthermore, our algorithm was based on EMD, a technique primarily used for noisy recordings. Although we did test the algorithm on both good-quality and noisy recordings, further research should examine the performance of different algorithms on various types of data.

Then, our sample size was limited, with our methodology based on a single PSG recording that was highly homogenous (i.e., women with a mean age of 48 years and no other sleep disorders than CID). The algorithm's performance on a larger and more diverse population remains untested.

Also, our study did not account for inter-rater variability in the scoring of sleep stages. Although we aimed to develop an algorithm that could mitigate this variability, we did not validate our algorithm against multiple raters to test this claim truly.

Conclusion

In conclusion, our study aimed to assess the feasibility of automatic sleep stage detection using an algorithm based on EMD. While the algorithm was successful in standardising the analysis and saving time, the accuracy was less than desired, especially considering previous findings in the literature. Given the limitations of our study, it is apparent that further research is required, including the use

of diverse algorithms, a larger and more diverse population, and validation against multiple raters to truly assess the potential of automatic sleep stage detection. But this EMD method could be efficient in sleep spindles detection alone. We then abandon the idea of using this algorithm, and in the next experiment, we will implement an ICA with CNN. algorithm to increase our chance of getting accurate results.

4.1.2 Convolutionnal Neural Network (CNN) for automated sleep scoring

4.1.3 Introduction, Background and hypothesis genesis

The primary aim of this second experiment is to leverage automated scoring for standardising outcomes and enhancing the reliability and reproducibility of results. Initially, we hesitated to employ a deep learning algorithm due to our interest in the explicability of predictions. However, the inferior performance of the EMD-based automatic scoring algorithm led us to consider an algorithm with a strong precedent in publications and practical application.

Background

Our selection fell upon an algorithm proposed by [36], recognized as one of the most advanced at the time of our research and fitting our aforementioned criteria. Numerous other algorithms using Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN), both or Attentional Networks, were published during this period. To sverify if our results are of interest regarding the recent publications, we will do the states of the arts of the algorithms tested on the same Sleep database, MASS (see the description in Appendix B.1.1). We synthesised the different results in Table 4.1

Ref Years	Input	Sequence encoder	Accuracy
Supratak et al (2017), Phan et al (2019)	Raw	RNN	0.800
U-Time Perslev et al (2019	Raw	CNN	0.800
TinySleepNet Supratak and Guo (2020)	Raw	RNN	0.782
GraphSleepNet Jia et al (2020)	Raw	Attention	0.834
SeqSleepNet Phan et al (2019)	Time freq	RNN	0.815
Algorithm used in our experiment	Raw	CNN	0.820

Table 4.1: Balanced accuracy score for sleep stages prediction on the MASS dataset used to train the algorithm used in our experiment since its publication in 2018. Only Attention sequence encoder performed a little bit better.

Hypothesis

Our objective was to study the reproducibility of a performant sleep-scoring neural network. The model we are studying was introduced in 2018 by [36]. It is a deep neural network that performs temporal sleep stage classification from multimodal (typically EEG, EMG and EOG) and multivariate time series. The model in question efficiently amalgamates information from various sensors using a linear spatial filtering operation. This helps construct a hierarchical representation of Polysomnography (PSG) data via temporal convolutions. Furthermore, the model also incorporates data from different modalities, which are processed via distinct pipelines. This model's structure can be observed in Figure 4.9

The algorithm we chose was specifically designed for home sleep helmets, with practical applications in military and space usage. Its development was facilitated by a sleep laboratory and INRIA. The algorithm is built on a deep convolutional neural network and performs temporal sleep stage classification using multivariate and multimodal time series. It was designed to work effectively with a limited number of electrodes and, as per the authors, attains 80% accuracy for multilabel balanced sleep stage classification, outperforming manual scoring.

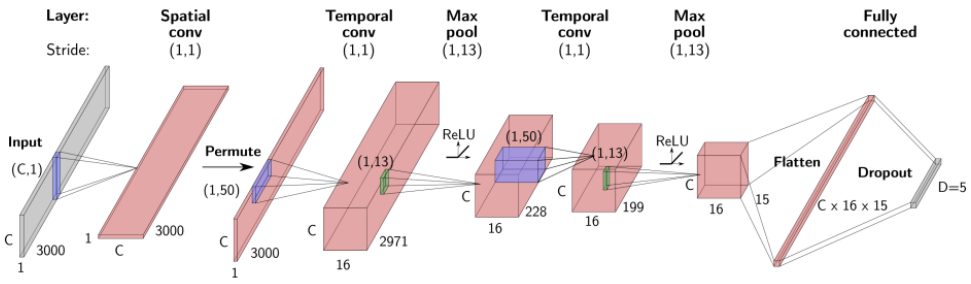


Figure 4.9: Graphical depiction of the sleep staging model's framework [36]

If successful on our data, this algorithm would not only provide more consistent sleep stage classification but would also potentially allow replication of our results on other datasets. This is particularly significant if sleep stages and their patterns play a role in explaining ParI or treatment responses in chronic insomnia. Unlike the first experiment, where we only used our data, we added two other datasets from different sources to evaluate the algorithm's performance. In addition to testing the algorithm, the idea here is to see if our data quality is comparable to datasets used in the literature. The two other datasets are the Montreal Archive of Sleep Studies (MASS) dataset [153] and the SleepPhysionet dataset [74].

4.1.4 Methodology

Data preparation and preprocessings

We aimed to evaluate our model's performance across various datasets to evaluate the reproducibility, which implied that these datasets needed to be comparable, possessing similar EEG/EOG/EMG channels concerning electrode placement. Regrettably, our datasets didn't contain identical channels, especially in terms of EEG channels, which hindered our ability to compare our model's performance across all datasets uniformly. Instead, we had to evaluate them on a pair-by-pair basis.

Figure 4.10 illustrates the diversity of the EEG channels across each dataset, highlighting their heterogeneity. We drew comparisons between MASS and SleepPhysionet and our Clinical dataset (Clinical in this section), but a comparison between SleepPhysionet and Clinical was unfeasible. For each pair of datasets, we selected various EEG, EOG, and EMG channels, which will be outlined below.

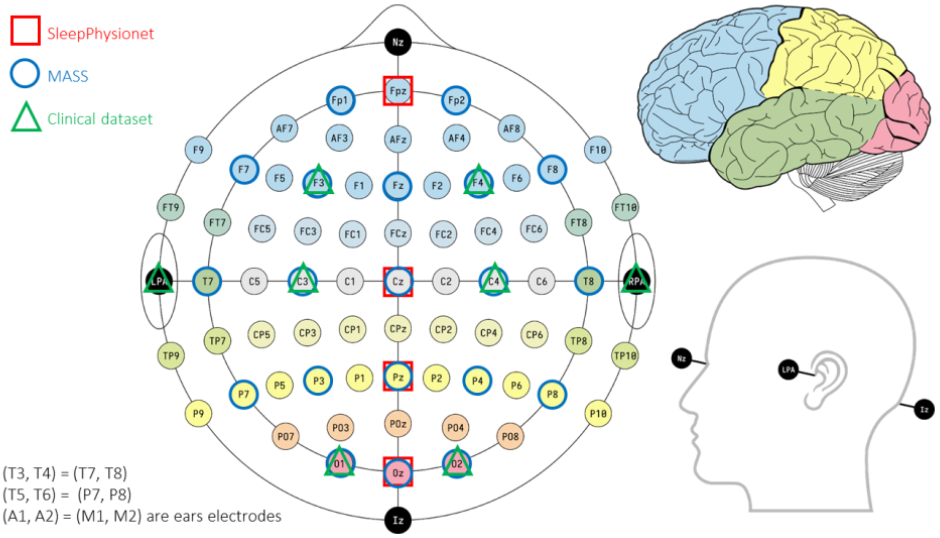


Figure 4.10: EEG electrodes positioning used in each dataset

SleepPhysionet dataset The SleepPhysionet dataset encompasses 153 full-night polysomnographic sleep recordings obtained from 78 individuals (typically two recordings per subject). The dataset includes two EEG channels (Fpz-Cz and Pz-Oz), one EOG channel (horizontal), and one EMG channel (submental chin). Accompanying these recordings are annotations of sleep patterns. The hypnograms were manually scored by a variety of skilled technicians in accordance with the Rechtschaffen and Kales manual [98]. The sleep stages are denoted as W, R, N1, N2, N3, N4, M (Movement time), and ? (not scored). In our analysis, we retained the annotations W, R, 1, and 2, while we combined stages 3 and 4 for consistency with other recordings scored according to the AASM rules [90]. During our investigation, we concentrated on subjects 0-60, excluding subject 39, and only the first recording from each subject was retained.

MASS dataset The Montreal Archive of Sleep Studies (MASS) dataset is a public resource. In our research, we emphasized the recordings from the third session, denoted SS3, comprising 62-night records from distinct subjects (28 males and 34 females). These records include 20 EEG channels (C3, C4, Cz, F3, F4, F7, F8, O1, O2, P3, P4, Pz, T3, T4, T5, T6, Fp1, Fp2, Fz and Oz), 2 EOG channels (left and right), and 3 EMG bipolar channels (chin). The EEG channels are referenced to either CLE (computed linked ear) or LER (linked ear reference with 10kΩ resistance). Sleep was classified by proficient PSG technicians according to the AASM guidelines. In our study, we focused on subjects 1-62, excluding subjects 43 and 49 due to preprocessing difficulties with their recordings. For comparing our model's performance between MASS and SleepPhysionet, we selected 2 EEG channels (Fpz-Cz and Pz-Oz, with Cz and Oz as respective references), 1 EOG channel (EOG horizontal, defined as the average between EOG left and EOG right), and 1 EMG channel (EMG Chin1). To contrast our model's performance between MASS and Clinical, we chose 6 EEG channels (C3, C4, F3, F4, O1 and O2), using the average of these channels as a reference, 2 EOG channels (EOG left and EOG right), and 1 EMG channel (EMG Chin1).

Our clinical dataset We used a subset of the EDF+ files available to compare to the two other datasets. After our first experiment, we focused on 60 subjects with artifact-free EEG channels. The patients were between 18 and 76 y.o. (mean 45.84 $\pm$ 13.05 with 66.5% females).

Data preprocessing To streamline data processing, we converted our datasets to the Brain Imaging Data Structure (BIDS) [76, 164]. Neuroimaging data is complex to arrange, as it typically originates from various experiments and generates multiple files for a single patient. Given the lack of consensus on how to organise and share such data, two researchers within the same lab may choose to organise their data differently. The BIDS standard offers a simple and adaptable way of organising neuroimaging data, using file formats compatible with existing software, unifying common practices in the field, and capturing metadata essential for most data processing operations. In particular, the BIDS standard substantially simplifies the analysis of neuroimaging data using Python, with the help of the *mne-python* [78] and *mne-bids* [4] libraries.

We used the *mne-python* package for preprocessing our datasets, following the same steps as [36]. As the most relevant information in sleep EEG data is below 30Hz; we applied a low-pass filter with a 30Hz cutoff frequency to reduce the impact of high-frequency noise. We downsampled to a sampling frequency of 100Hz (the SleepPhysionet dataset frequency, whereas the other two datasets were sampled at 256Hz) and converted signals from V to μV. We also eliminated 30 minutes of wake events before and after other sleep events. After applying these steps, we divided our signal into 30s windows, each corresponding to a specific sleep stage. Each window was individually standardised to have zero mean and unit variance. This standardization is crucial due to the variability in recording conditions over the nearly 8-hour recording period. Individual standardization addresses potential shifts, rescaling frequency power in every band without altering their relative amplitude.

Algorithmic methodology

Model description The full model description is detailed in [36], but we can see the detailed architecture in Figure 4.11

	Layer	#filters	#params	Size	Stride	Output dimension	Activation	Mode
	1. Input					(C, T)		
	2. Reshape					(C, T, 1)		
	3. Convolution 2D	C	C * C	(C, 1)	(1, 1)	(1, T, C)	Linear	
	4. Permute					(C, T, 1)		
Features	5. Convolution 2D	8	8 * 64 + 8	(1, 64)	(1, 1)	(C, T, 8)	Relu	same
Extractor	6. Maxpooling 2D			(1, 16)	(1, 16)	(C, T // 16, 8)		
	7. Convolution 2D	8	8 * 8 * 64 + 8	(1, 64)	(1, 1)	(C, T // 16, 8)	Relu	same
	8. Maxpooling 2D			(1, 16)	(1, 16)	(C, T // 256, 8)		
	9. Flatten					(C * (T // 256) * 8)		
	10. Dropout (50%)					(C * (T // 256) * 8)		
Classifier	11. Dense		5 * (C * T // 256 * 8)			5	Softmax	

Figure 4.11: This network has three key features: linear spatial filtering, to estimate virtual channels, convolutional layers, to capture spectral features and separate pipelines. It can handle various input channels and several modalities at the same time, iEEG/EOG channels and EMG channels through separate pipelines. It performs spatial filtering for each modality and applies convolutions, non-linear operations and max-pooling (MP) over the time axis. The outputs of the different pipelines are concatenated to feed a softmax classifier from [36]

Training Protocol This model was developed using the PyTorch library [161]. For each experiment, we studied 60 subjects, using stratified 10-fold cross-validation to ensure that approximately 60% of the events were part of the training set, 20% in the validation set, and 20% in the testing set. This protocol ensures that the model is trained, validated, and tested on datasets representative of the overall class distribution.

We initialized weights with a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.1$. The loss function (criterion) used was categorical cross-entropy, and optimization was performed using AdamOptimizer (see definition in B.1.2). Minimization was achieved with stochastic gradient descent with a learning rate of $lr = 0.0005$ and a batch size of 8. The model was trained over 10 epochs (we chose this number considering the small size of the dataset).

Performance evaluation

- We used a confusion matrix as described in 4.1.1, and we added the three most used metrics, Precision, Recall, and F1 score (see definitions in B.1.2,B.1.2 and B.1.2).
- Precision, recall, and F1-score are statistics used to study a binary classification. In the case of multi-class classification, we compute the precision, recall, and F1 score per class.
- We also used a balanced accuracy score for the average accuracy. Balanced accuracy is a metric often used for imbalanced class problems since it considers the varied nature of the classes. In multiple classes, balanced accuracy is defined as the recall average for each class.

4.1.5 Results

As mentioned in the introduction, we could not compare the three datasets simultaneously but only two by two. We will present the results of MASS and Physionet and then our Clinical dataset.

MASS and Physionet

The four confusion matrices are presented in Figure 4.12 and the metrics in Tables 4.13 and 4.14

MASS and Clinical Dataset

The four confusion matrices are presented in Figure 4.15 and the metrics in Figures 4.16 and 4.17

Figure 4.12: Confusion matrix table, comparing MASS and Physionet datasets average scores. The predicted label are on the Y-axis and the True lable on the X-axis. We can notice the low scores in the N1 column, showing that the model rarely predicts N1

Training set \ Testing set	MASS	SleepPhysionet
MASS	**0.802** *0.815*	**0.487** *0.384*
SleepPhysionet	**0.560** *0.488*	**0.634** *0.618*

Figure 4.13: Table of balance accuracy results, comparing the datasets MASS and SleepPhysionet. We can notice the loss in accuracy when the algorithm is trained on a dataset different from the testing set. This shows poor reproducibility.

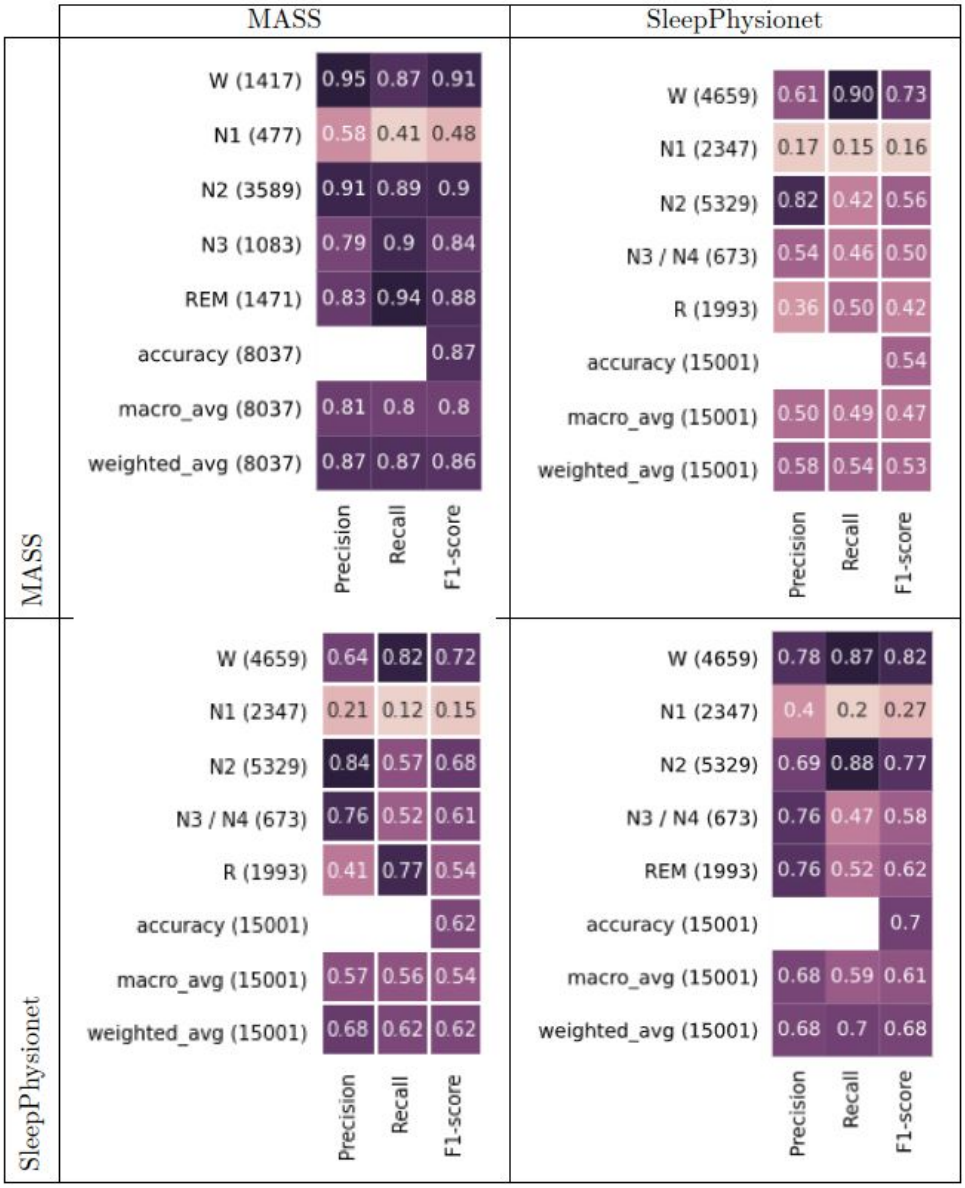


Figure 4.14: Table of classification reports, comparing the datasets MASS and SleepPhysionet. We can see the difficulty in predicting N1 stage.

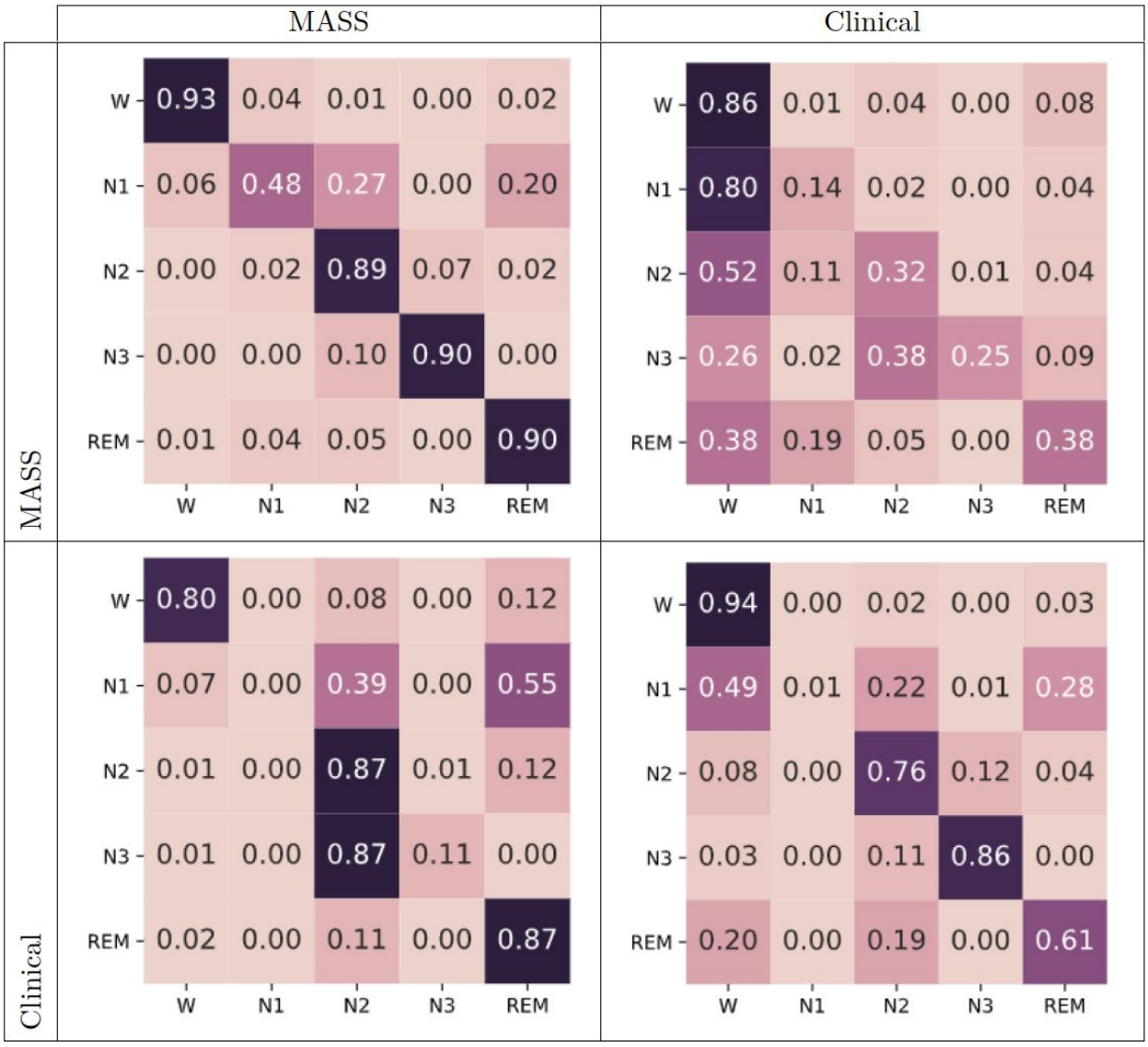


Figure 4.15: Confusion matrix table, comparing MASS and our Clinical Dataset

Training set \ Testing set	MASS	Clinical
MASS	**0.817** *0.822*	**0.390** *0.287*
Clinical	**0.530** *0.557*	**0.635** *0.697*

Figure 4.16: Table of results, comparing balance accuracy between datasets MASS and Clinical. On these balanced accuracy scores we notice an increase in the scores for MASS-MASS (0.817 compared to 0.802) and the low score for physionet when the algorithm is trained on clinical and applied on Clinical. The training on MASS and the test set on clinical is even lower with only a score of 0.39

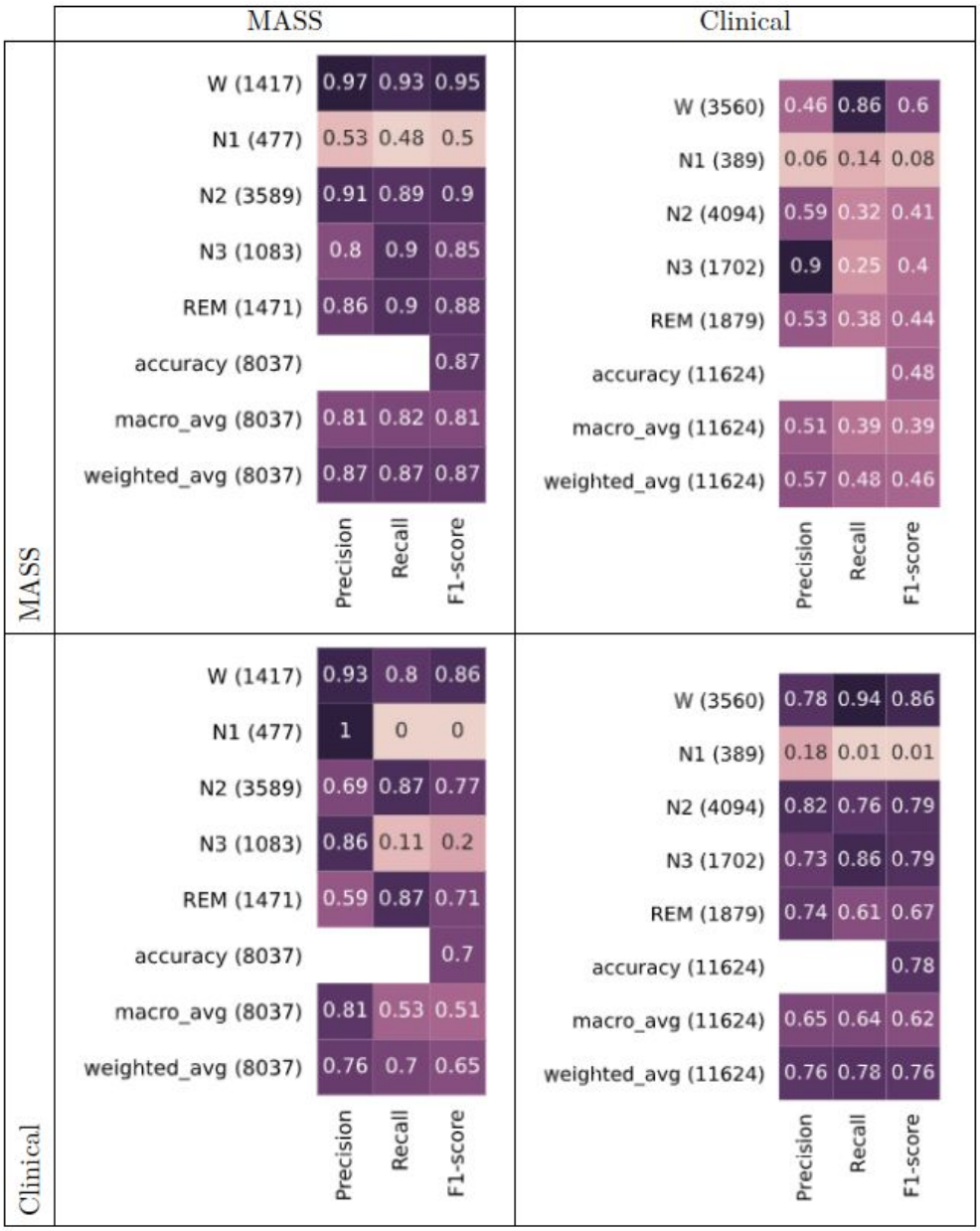


Figure 4.17: Table of classification reports, comparing the datasets MASS and Clinical. We notice a similar behaviour with regards to N1, this class is quite difficult to predict

4.1.6 Discussion

Our research has yielded crucial insights into the performance of the CNN model and the underlying characteristics of the utilized datasets. A major finding is that the model performs significantly better on the MASS dataset than the SleepPhysionet and Clinical datasets. This superior performance might be attributed to different reasons discussed below:

1. The preprocessing steps might not be appropriately tailored for the latter datasets. It involves a set of operations that prepare and transform the raw data into a format the model can more effectively process. These operations include data cleaning, normalization, transformation, and feature extraction, and if there are not optimised for a specific dataset, the model's ability to learn from that data could be compromised, leading to suboptimal performance.

2. As the algorithm was initially benchmarked using the MASS dataset, it may have led to a degree of overfitting to the characteristics specific to the MASS dataset, making the model less generalizable to other datasets with different characteristics. But it would be surprising because this algorithm was used massively to analyze the sleep of clients who bought the Dreem headset for ambulatory sleep analysis; we could imagine that they would have changed the algorithm if the results were not so good as claimed in the seminal paper[36].

3. Another explanation for such bad results may be non-optimal hyperparameters. Indeed, as they control the learning process and can significantly affect the model's performance, it is plausible that the chosen hyperparameters were not ideal for these datasets, thus undermining the model's performance.

4.1.7 Limitations

A notable limitation of our study is the difficulty in accurately classifying the N1 sleep stage. This challenge could manifest the class imbalance problem, a common issue in machine learning where the classes are unequal. In such scenarios, the learning algorithm may become biased towards the majority class, leading to poorer performance in the minority class. Future studies could explore strategies to mitigate this issue, such as implementing resampling techniques or adopting different performance metrics more resilient to class imbalance.

Interestingly, our model performed similarly on the Clinical and SleepPhysionet datasets. This similarity could be due to the increased number of channels used in our Clinical dataset offsetting the potentially higher noise level inherent in clinical data. This finding indicates the importance of leveraging multiple data sources and incorporating more comprehensive features in sleep stage classification tasks.

However, the model's performance varied substantially between datasets, indicating a degree of overfitting and poor generalizability. These findings highlight the need for adopting more sophisticated model architectures or regularization techniques that can effectively mitigate overfitting. Additionally, adaptive hyperparameter tuning methods could be explored to optimise model performance across different datasets.

4.1.8 Conclusion

To sum up, our study provides valuable insights into the complex task of sleep stage classification using machine learning techniques. While the CNN model demonstrates promising results, especially on the MASS dataset, there are several areas for improvement and further investigation. Pursuing this research could significantly advance the field of sleep stage classification and facilitate the development of more effective and generalizable models. But the conclusion for our specific study is that the CNN model didn't perform any better than the EDM model described previously, and then we concluded with this second experiment that we could not use a reliable and reproducible algorithm to harmonize the sleep stage scoring in our thesis.

4.2 Benchmark for Spindles Detection

4.2.1 Spindles and sleep

EEG sleep analysis can mix meaningful events for sleep classification and clinical assessment. The Macro-structured neural events refer to segments that are usually 30 s long and represent different sleep stages or epochs, or levels of sleep compared to the awake condition (Wake, N1, N2, N3 and REM sleep). On the other hand, micro-structured neural events refer to local and short segments, such as sleep spindles, K complex, alpha rhythm, etc. These micro-neural events are also important for finding the right sleep stages and better characterising the pathology.

Sleep spindles, which are the most typical sleep pattern, typically occur during sleep stage 2 and are believed to be generated from the Thalamus area of the brain. The definition of a spindle varies across studies, but we will consider two different consensual criteria. The first one is the bandwidth; a spindle could be defined as an increase in EEG power over consecutive NREM sleep stage two repeatedly found in the 11±16 Hz (sigma rhythm) but most of the time between 12 and 14 Hz, with a duration > 0.5 seconds up to 3 seconds [218, 17, 176].

To detect those spindles automatically, a lot of algorithms were built, essentially on bandwidth detection corresponding to the spindles bandwidth, so that could be different across studies. The most used methods rely on performing threshold on a filtered signal (fixed or not) with different techniques of signal analysis applied like Fourier transform or wavelet [206, 46, 173, 113, 63, 143, 151]. See [154] for review.

The second is the shape, with a typical waxing and waning. The third is the amplitude; the minimum peak-to-peak amplitude [176] of the spindle should be 10 micro-volts.

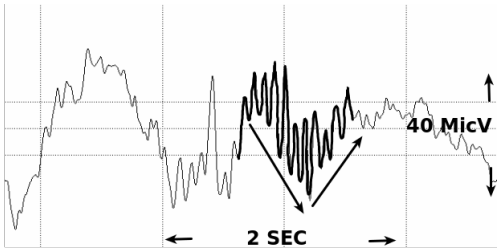


Figure 4.18: Typical spindle shape recorded on C3 derivation during typical N2 sleep stage. Illustration of the waxing and waning shape, the frequency is 13 Hz, the duration is about 1 sec, and the amplitude is about 40μV. The Spindle appears superimposed on a slower wave symbolized by the arrows.

Recently, new approaches using sparse optimization iterative shrinkage/thresholding algorithm (ISTA) [159] and deep learning was also proposed [36].

Besides the sleep stages characterization, the main interest in spindles is the possible characterization of different psychiatric or neurological disorders such as dementia, schizophrenia, depression, sleep disorders, or stroke recovery [208].

4.2.2 Hypothesis and design of experiments

Sleep dataset This preliminary study used a single night of a young, healthy subject, totally artifact-free. The hypothesis is that we could achieve at least 0.8 accuracy in spindle prediction with at least one of the published algorithms. [113], for example, declared very good results close to 0.8 accuracy.

Methods

Pre-processing and expert spindle detection Sleep experts first visually detected the evaluation of sleep stages and spindles. Sleep stages (N1, N2, N3, and REM sleep), awake time, and movement artifacts were scored offline for 30-sec intervals according to the AASM criteria [17]. EEG data from the sleep cycle were chosen, and analyses targeted the bipolar channels C4-M1 and C3-M2, where spindles are most pronounced. [208] observed that 14% of the spindles fell in the 0.3-0.5 s duration

and that 85 % of the spindles duration was between 0.5 and 2 seconds. Following this observation, we performed spindle detection according to AASM rules (Sigma rhythms (11–16 Hz) are visible on NREM, EEG for at least 0.3 sec (maximum 2 seconds) using a band-pass filter (0.5–30 Hz). A spindle event was included in the analysis only if it was validated 2 times by the expert using this rule. 616 spindles were detected in total on 276 epochs of N2 sleep (mean 5.46 ±1.15 spindles/ min), which is in line with the range of mean density (0-10/min) found in the "gold standard data set" spindle density found in 110 subjects [208] and with the mean spindle density (4±2 /min) found in average across 3 nights in 24 young adults[165]. A typical spindle's pattern is shown in figure 4.18

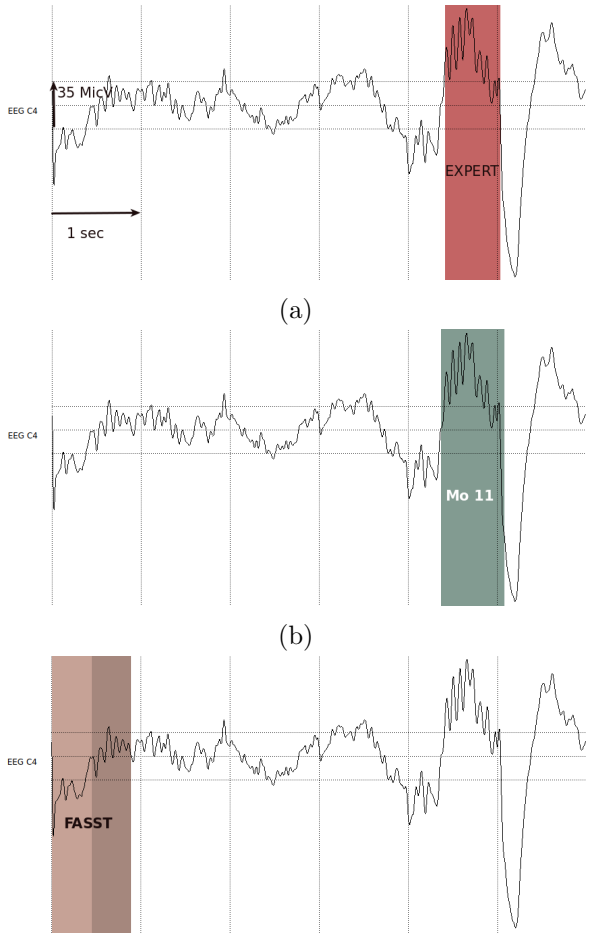


Figure 4.19: Typical NREM2 stage (6 sec) on C4-M1 with one typical spindle; 2a Spindle manually detected by a human expert, 2b Same Spindle detected by Mo11[143] that have the best F-Measure, 2c False positive spindle detected two times by FASST (fMRI artifact rejection and sleep scoring toolbox). At the same time, the right spindle is not detected. [118]

Automatic spindles detection Using nine different algorithms, we used an open-source tool for spindle detection (Gio Piantoni / Jordan O'Byrne). Seven are published Mo11 [143], Fe07 [63], Nir11 [151], Ray15 [173], FASST [118], mar13 [134], La18, [113] and two unpublished algorithms, the UCSD algorithm (the University of California based on wavelet analysis) and CONCORDIA algorithm (Concordia University) based on Root Mean Square (RMS) of the signal. We benchmarked those algorithms, choosing the same bandwidth for the spindle definition (11-16 Hz). Experts apply the same criterion to detect spindles. We will describe the main principle used by these algorithms in the next paragraph, mainly Wavelet convolution and detection threshold.

Wavelet convolution A wavelet is a wave-like oscillation with an amplitude that begins at zero, increases, and then decreases back to zero. also known as a wave packet in physics. Morlet's wavelet is designed to have the optimal properties for detecting spindle-like activity as it has the shape of a sleep spindle. (see figure 4.20). This method convolves the Morlet's wavelet with the EEG signal.

If a real spindle is present in the EEG signal, it will be multiplied by the spindle-like wavelet, thus resulting in a very high amplitude signal. Applying a threshold to the resulting amplitude allows us to detect the spindles. Methods such as RAY2015[173] and UCSD - University of California, San Diego (unpublished) use convolution.

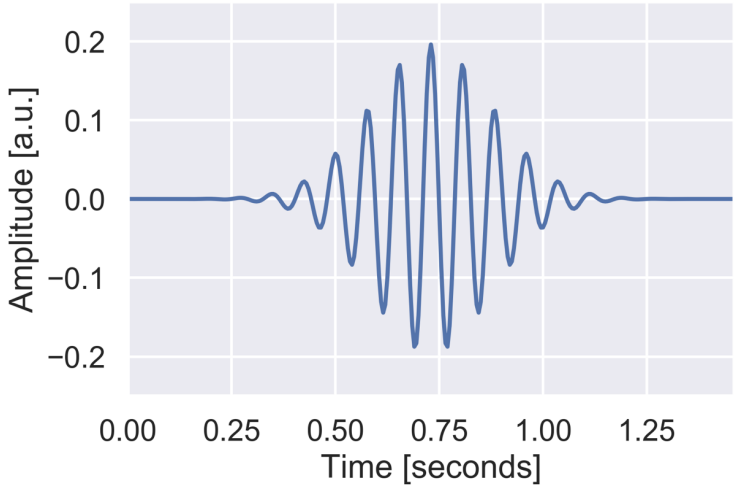


Figure 4.20: Morlet's wavelet

Detection thresholds A second more common and more precise way of detecting sleep spindles is through thresholds. This aims to progressively identify by eliminating which parts of the EEG signal compose sleep spindles. Many methods use thresholds; for instance, Moelle 2011[143] use thresholds based on the root mean square (RMS) value. They follow a different detailed protocol but with the same structure.

1. Detect signals within the sigma frequency range (11-16 Hz).
2. Compute the Root mean square (RMS) of the detected signals using an adjustable window size and step.
3. Compute the RMS threshold: $RMS_{thresh} = RMS_{mean} + 1.5 * RMS_{std}$
4. Spindles are detected whenever $RMS > RMS_{thresh}$
5. Only the spindles lasting between 0.5s and 3s are retained.

Depending on the algorithm, the core structure may include some more complex or detailed steps. For instance, step 1 can be achieved by applying a simple band pass filter. Or, by making use of a Short Term Fourier Transform (STFT) to detect whenever the signal has a relative power in the sigma frequency range ≥ 0.2 to ensure that the increase in sigma power is specific to the sigma frequency range and not just due to a global increase in power (e.g. caused by artifacts).

4.2.3 Results

Spindles detection: true positives and false negatives vs domain expert

We observed an overall sub-estimation for all algorithms except the FASST algorithm but with a very high cost in (i.e., many) false negatives. In total, **three algorithms significantly exceed the total number of spindles detected by the expert (250)**, the FASST (3 times), Mo11 (2 times), and Ray15 (less than two times).

Although the PSG recording was high quality, some periods are light artifacts by natural movements during NREM2 sleep. We were surprised to see how this very brief period of artifact could increase the false negatives for almost all the algorithms tested. We can see in 4.21 an example of an artifact, very different from a spindle; however, detected as a spindle by seven of the nine algorithms (except La18 and Mo11).

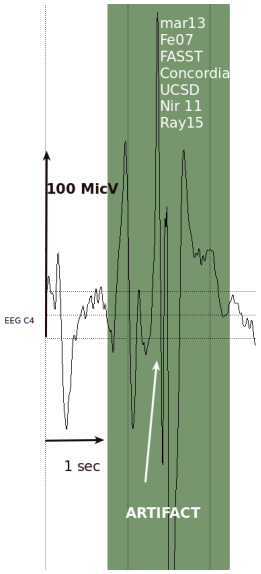


Figure 4.21: Artifact during NREM2 sleep, channelC4-M1. False detection as a spindle by mar13, FeO7, FASST, Concordia, UCSD, Nir11 and Ray15.

Precision/recall and F-measure

We see in Figures 4.22 and 4.23 the prediction results. There is a global homogeneity between Precision and Recall except for FASST, which shows the biggest discrepancy due to an overestimation of spindles with, at the same time, a lot of false negatives and double detection.

4.2.4 Discussion

Our study delineated performance disparities when juxtaposed with the outcomes from [113, 208]. Contrary to these studies, our inclusion of [143] revealed it as the superior algorithm with an F-measure of 54.8, showcasing a commendable balance between Precision and Recall. We further extended our analysis to compare La18[113] against Ray15, UCSD, and FASST, each surpassing La18. This was complemented by juxtapositions with methodologies from [63] and [134], affirming the consistency of our preliminary assessments with antecedent benchmarks, notwithstanding the smaller scale of our dataset.

While indicative of performance, the F-measure may mask underlying inconsistencies in detection accuracy. A case in point involves FASST and La18, where an equivalent F-measure of approximately 0.49 belies divergent Recall and Precision statistics. As [154] contends, a more comprehensive metric, such as the MCC (Matthew's Correlation Coefficient), is imperative for a nuanced comparison of algorithms.

Our empirical evaluations highlighted the imperative to discern between precision and the propensity for algorithms to replicate detections.

Furthermore, the persistence of artifacts poses a significant impediment in ambulatory EEG recordings, rarely devoid of such distortions. This challenge, as evidenced by our dataset, is prevalent even in meticulous hospital settings and is exacerbated in portable EEG devices employing dry electrodes.

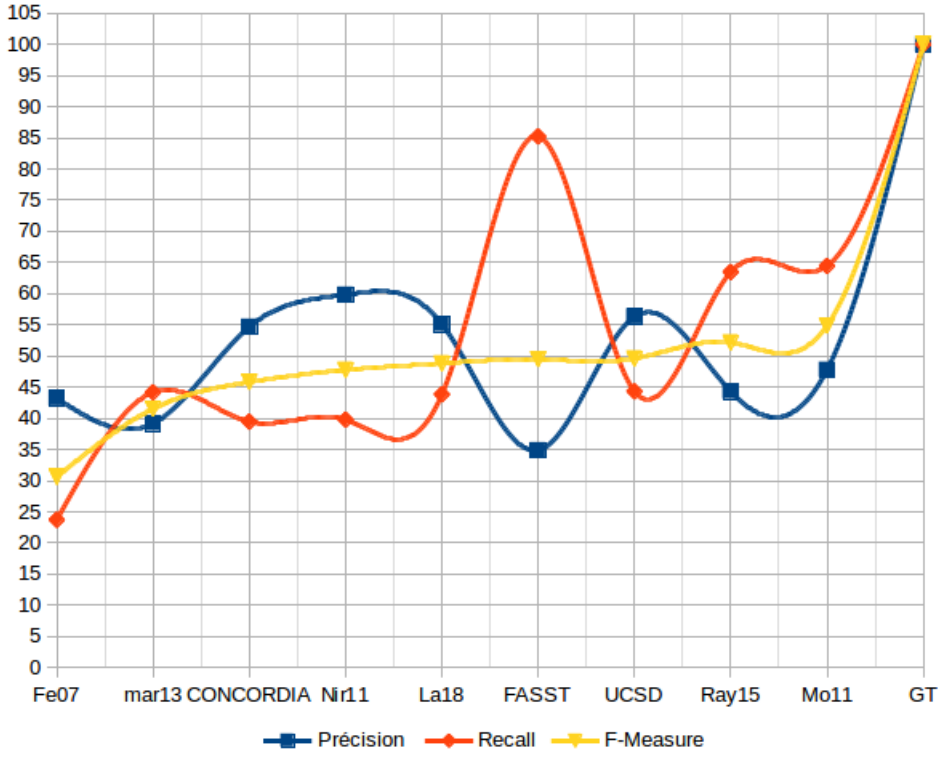


Figure 4.22: Comparison of the precision, the recall, and the F-measure for each algorithm against the expert detection (GT for "Ground Truth"). F-Measure is Fe07[30.6], mar13[41.5], Concordia[45.8], Nir11[47.8], La18[48.8], FASST[49.4], UCSD[49.6], Ray15[52.1], Mo11[54.8].

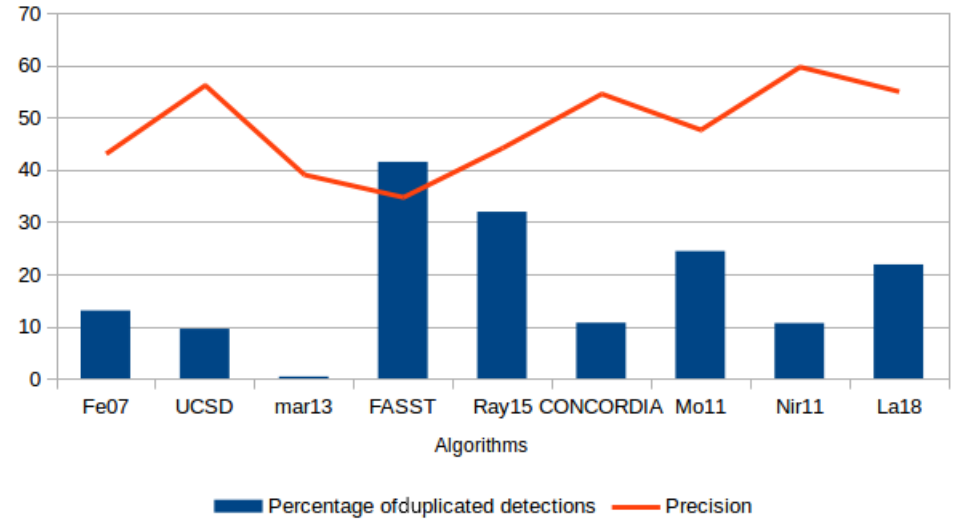


Figure 4.23: Comparison of the precision and the % of spindles detected at least two times on a single channel by the algorithm with overlap.

4.3 Spindles and Personality Prediction

4.3.1 Introduction and hypothesis genesis

Background

Following our evaluation of sleep spindle detection and analysis, we became interested in whether this could enhance the diagnostic process for sleep disorders and predict predispositions towards certain psychopathological traits. Previous research has endeavoured to characterize sleep spindles to understand their genetic architecture better [171]. Furthermore, sleep spindles have been characterized within specific populations, such as patients with major depressive disorder [167], intellectually disabled children [189], and schizophrenic patients along with their unaffected relatives [49]. However, the extent to which sleep spindles can be utilized to diagnose specific sleep disorders like paradoxical insomnia (ParI) or psychological conditions such as Hypochondria remains unexplored. Although our investigation focuses on ParI explanation, we wanted to test a first hypothesis linked to the psychological profile to see if spindles could discriminate very different psychological profiles. The selection of sleep spindles is driven by their integral role in sleep and the evidence showing strong associations between sleep disorders and psychopathological traits such as Depression, Anxiety, Post-traumatic Stress, Schizophrenia, Bipolar Disorder, Eating Disorders, Borderline Personality Disorder, among others [211, 133, 136]. To date, no previous study uses sleep EEG spindles to predict psychopathological traits labelled by the MMPI-2 questionnaire.

Hypothesis

From this observation, we hypothesise that Sleep spindles, as a biomarker, can predict predispositions towards certain psychopathological traits important to understand CID clusters or subtypes. The characteristics of sleep spindles (including their density, average duration, average frequency, and the average number of oscillations) can be used to predict the occurrence of a patient's predisposition towards certain psychopathological traits as assessed by the MMPI-2 questionnaire.

Protocol in brief

We begin by identifying sleep spindles in the patients' Electroencephalogram (EEG) data using three distinct algorithms for the highest level of accuracy. We then compile a spindle dataset, extracting parameters such as density, average duration, average frequency, and the average number of oscillations from sleep spindles occurring in stages 2 and 3 of non-rapid eye movement sleep. Next, we test the prediction of ParI using this dataset. ParI is a sleep disorder in which the patient's perception of insomnia severity significantly exceeds any observed sleep disturbance, often believing they are not sleeping well despite evidence to the contrary. Visualization of the dataset and application of classification techniques, however, suggest that spindle data cannot accurately predict ParI, and, thus, cannot aid in its diagnosis. Finally, we analyse if spindle data can predict a patient's predisposition towards a psychopathological trait. We accomplish this using the scales of the MMPI-2 questionnaire (explained further in the following sections) taken by all patients.

4.3.2 Methodology and experiment design

Dataset

EEG dataset From the EDF files available in our dataset, we kept 267 patients (mean age 47± 12, 56% are Female). We tried to be the most selective as possible and we kept only the EEG labelled as good or very good after expert visual inspection. This implied that all the PSR records with more than one noisy channel, irreducible EMG artifacts or interference were removed to reduce the risk of misclassification by the algorithm. As we saw in our benchmarking of spindles detection algorithm 4.2, we could have a lot of False positives even on a good quality recording, so we wanted to maximize the chance to have good results for this research question aimed to find qualitative results.

We used the six following EEG electrodes available in our dataset :

- $A_1 - C_4$ • $A_2 - C_3$ • $A_1 - F_4$ • $A_2 - F_3$ • $A_1 - O_2$ • $A_2 - O_1$

We apply a notch filter of 50Hz to the data to cancel any potential interference created by the electric plugs in the patient's room. The recordings last 6-14 hours, but we only focus on N2 and N3 sleep stages.

Feature Extraction On 267 recordings, we could extract 746.4 hours of sleep stage scored N2 (average of 169 minutes per patient) and 247.4 hours of sleep stage scored N3 (average of 55 minutes per patient).

We apply a notch filter of 50Hz to the data to cancel any potential interference from the electric plugs in the patient's room. The recordings last 6-14 hours, but we only focus on certain parts of the night.

MMPI-2 dataset The description is already done in 2.4.1 and the features used in this experiment in B.2 To test our hypothesis, we selected only five scales belonging to the Main clinical scales described in 3.1. These scales are Hy, D, Hs, Pa and Sc. Indeed these five features can already discriminate two major psychological profiles observed in daily practice, Neurotic and psychotic. We made these choices to keep the experiment simple and linked to daily practice. Indeed as we already described, the first three scales are the most significant in Insomnia care as described in 3.4.1, and the only ones with a mean > 65 in our sample (considered as significant threshold)[69]. These three scales are considered as "neurotic triad" [16].

- The Hs scale stands for Hypochondriasis and measures the preoccupation level for health and bodily functions. Individuals who score high on this scale are often seen as excessively worried about their health and may believe they have serious illnesses despite a lack of medical evidence.
- The D scale stands for Depression and measures pessimism and general dissatisfaction with their own life.
- The Hy scale stands for Hysteria and is used to evaluate histrionic behaviour, somatization, and defense mechanisms such as denial and repression.

So the neurotic triad (Hs, D, and Hy) often suggests difficulties in coping with stress, a tendency to internalise conflicts and potential vulnerability to stress-related physical or mental health issues. But in general, this triad is unrelated to psychotic disorder [52].

On the other hand, the psychotic profiles are essentially linked to two scales, Pa and Sc [52]. We will briefly describe these two scales.

- The Sc Scale stands for Schizophrenia scale. This scale assesses a person's tendency towards schizophrenic behaviours and thoughts. High scores on the Sc scale can indicate unusual thought processes, bizarre fantasies, difficulties in concentration, and social withdrawal. While this scale was initially intended to identify individuals with schizophrenia, it is now seen as a measure of a person's general "strangeness" or "unusualness" and can be high in other conditions such as bipolar disorder, severe anxiety, and depression.
- The Pa Scale stands for the Paranoia scale. It measures an individual's level of paranoia. High scores may suggest a person is overly suspicious, sensitive, feels persecuted, or is experiencing delusions of grandeur.

So these two scales could be elevated in case of psychotic disorder like Schizophrenia or bipolar disorder [52].

Sleep spindles detection The process of accurately identifying sleep spindles, defined by unique bursts of oscillatory brain activity, is quite easy for sleep experts. They primarily appear during the N2 sleep phase and fall within the 11-16 Hz frequency range. Sleep spindles have been associated with numerous functions, such as maintaining the disconnection from the external environment during sleep, aiding in sleep-dependent memory consolidation, and playing a role in cortical development. Nevertheless, the detection of sleep spindles can be arduous due to their ambiguous definition, the lack

of consensus among experts in scoring them, the absence of standard automated detection techniques, and inconsistencies in the methods used to assess the performance of automated detectors. [209, 189, 191].

Various methods have been developed to detect sleep spindles, predominantly around wavelet convolution and threshold detection. Morlet's wavelet, designed with a shape similar to a sleep spindle, is particularly conducive for detecting spindle-like activity in EEG signals. However, our previous study 4.2 found that this method fell short in accuracy due to the variations in spindle shapes, frequencies, and oscillations, making it challenging to find a one-size-fits-all wavelet for spindle detection. From the previous experiment described before 4.2, we could conclude that extracting spindles properly is not an easy task. For this experiment, we wanted to increase our accuracy in the detection by using multiple detection algorithms.

From our first experiment on spindles detection 4.2, even if the results with the wavelet transform were not fantastic, we could at least confirm the results published by [143] with their algorithm (named Molle2011 here), and it was the best on our dataset, so we decided to keep it. Furthermore, this method allows for adjusting all parameters, offering significant flexibility. From our experiment using the BrainRT algorithm with Empiric Mode Decomposition (EMD) in Section 4.1.1, we could observe that the N2 sleep stages was the only stage keeping an accurate classification even in case of corrupted EEG. The efficient spindle detection provided by the EMD algorithm could explain these results, so we also decided to keep it for the experiment. As we wanted explainable models, we didn't search Deep Learning based spindle algorithms. To add another spindles detection method, we explored numerous techniques such as Yet Another Spindle Algorithm (YASA) [201], and FAAST2 [186], which set thresholds based on the Root Mean Square (RMS) value of detected signals. YASA emerged as an attractive choice due to its precise detection capabilities and the output of valuable parameters regarding the spindles. It applies three different thresholds: relative σ power to detect signals within the sigma frequency range, correlation to detect spindles visible on the raw EEG signal, and RMS threshold to detect an increase of energy in the EEG signal.

So we decided to keep Molle2011, BREMD (that we renamed violet) and YASA for our experiment.

Tuning hyper parameters We used a domain expert to detect spindles on one EDF file: EDF-MANUELX4. Indeed, the detection is extremely time-consuming as the sleep recording is over 8 hours long while spindles last between 0.5 and 2 seconds which explains we only had one such annotated file. The domain expert's detection is used as the ground truth to tune the parameters of the three methods.

Merging spindles The initial data comprises 6 channels. Moelle 2011 and Violet detected spindles on each channel separately and did not include a way of merging these results. Indeed, by analysing this detection, we noticed many spindles overlapped in time; thus, computing the spindle density per 30 seconds by simply summing the spindles found on each channel would have been error-prone. Instead, we decided on a merge rule presented in figure 4.24 where the red line represents the duration of the merged spindle. The merge rule decided was the one used by YASA and was added to the code of Moelle 2011 and Violet so that all 3 methods could be in accordance. Note we use density per 30 seconds for historical reasons. Indeed before the use of computers, the scoring of sleep was done on paper. One paper corresponded to 30 seconds; thus, 30 seconds is considered an epoch for sleep scoring.

Working with phase segments For each patient, we first set out to obtain the nine features seen in figure 4.25. It is important to note that the overall density over N2 or N3 is not the average of the segment densities since the density is measured concerning the duration of each segment; hence it is a weighted average.

Checking for errors Errors can occur due to an error in the code or a misunderstanding of the data. Therefore, it is extremely important to check the results' coherence. Units, for instance, can be a source of mistakes as the duration and time in the original data are given in microseconds while the data we extract uses seconds.

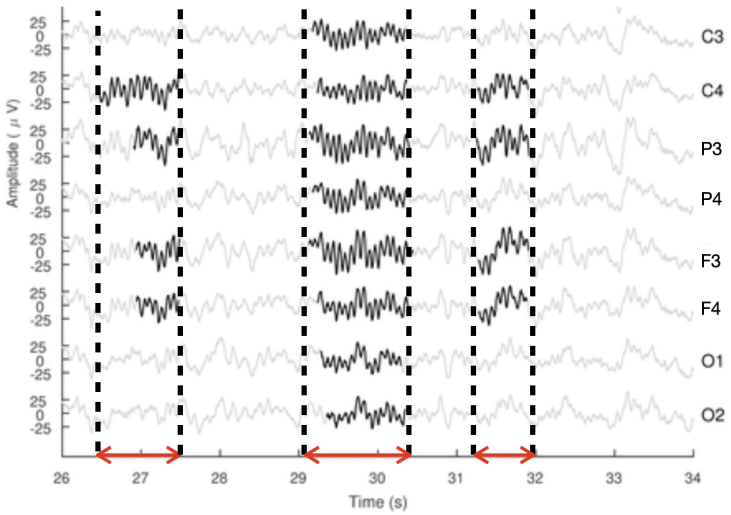


Figure 4.24: Spindle merging rule

File	mo dens	mo av dur	violet dens	violet av dur	seg dur	yasa dens	yasa av dur	yasa av freq	yasa av osci nb
04_1	1.581395	0.886432	0.976744	0.659722	1290.0	0.046512	0.656250	12.966834	7.750000
04_4	2.000000	0.879232	0.750000	0.629296	360.0	0.083333	0.752604	13.809815	8.833333
04_7	2.016667	0.811435	1.900000	0.632812	1800.0	0.016667	0.546875	13.530694	7.200000
04_8	1.781250	0.865063	1.750000	0.634364	960.0	0.062500	0.761719	12.895143	9.625000
04_9	2.200000	0.892116	1.880000	0.648011	750.0	0.120000	0.713728	13.196670	8.642857
4_12	2.266667	0.754969	1.783333	0.685571	1800.0	0.050000	0.605078	13.317776	7.500000
1004	2.030405	0.818519	1.496622	0.638580	8880.0	0.040541	0.701282	13.218796	8.255119

Figure 4.25: Feature extraction for the EDF recording used in phase N2. Files with underscore correspond to N2 segments. missing segments were dropped due to missing values. den = density, av dur = average duration, seg dur = segment duration, freq = frequency, osci = oscillation

Visualization is a key method to spot errors; this is how an error was found in the code extracting the features using the Violet algorithm. We plotted the densities in decreasing order; this gave the plot in figure 4.26. Indeed, as a spindle duration is usually between 0.5 and 2 seconds, on a 30-second epoch, it's quite rare to have more than 10 spindles. This fact led to notice that some densities were much too high. Indeed, the densities should be no bigger than 10 spindles per 30 seconds. Moreover, the densities should be more or less the same amongst all segments, while here we see huge disparities (standard deviation amongst N2 segments of patient 0228 - the first patient to the left of the x-axis in figure 4.26 - was off of 317,62).

This error was due to a misunderstanding of the annotation file, which caused some N2 segments to last the entirety of the recording rather than the found duration. The code was fixed, and we obtained the results in 4.27, which are much more coherent.

Classification protocol The best results were achieved using the data from all algorithms (Mo, yasa, and violet); we merged all the spindle detection from N2 and N3.

Once the spindles were extracted with our protocol, we used an RF classifier for binary classification to predict each of the five scales according to its characterization in normal (≤ 65) or significant $>$ 65. This allows for a binary problem (Class 1 for scores above 65, 0 otherwise). However, this cutoff does not, in all cases, create a class balance; thus, the F1 score was the most suitable measurement to interpret the results. We chose this instead of a regression problem to see if we could find a clear cluster according to the pathological aspect.

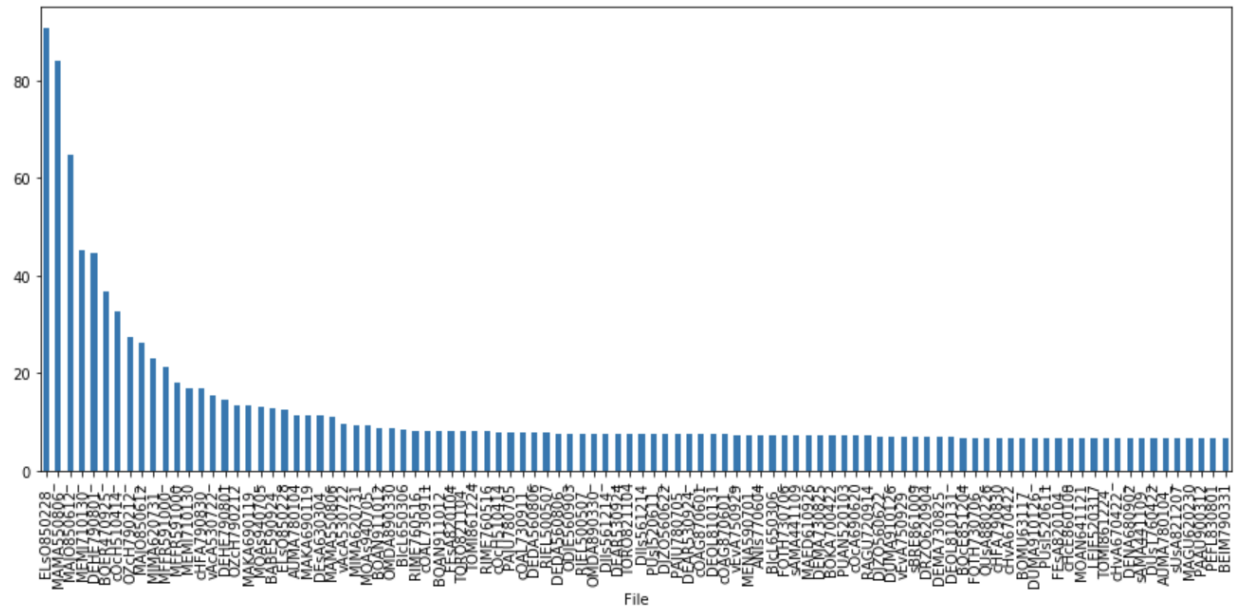


Figure 4.26: Densities found by Violet in decreasing order, in y-axis the density of spindles per 30 seconds, in x axis N2 segments with their file name

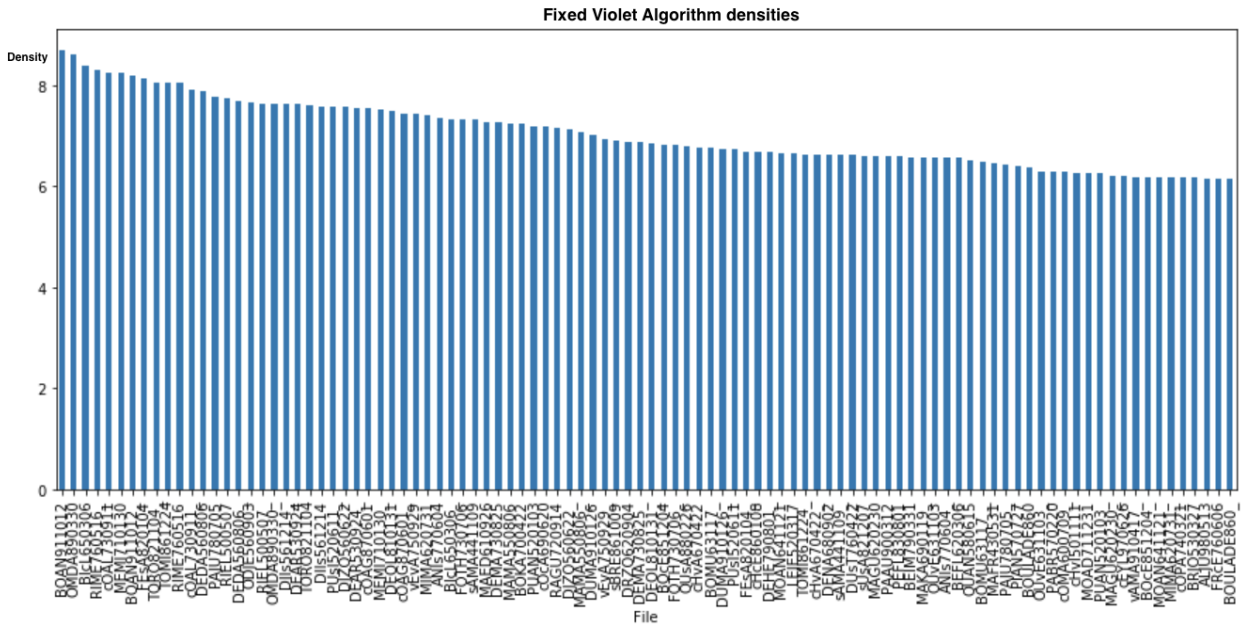


Figure 4.27: Densities found by fixed Violet, in y-axis the density of spindles per 30 seconds, in x axis N2 segments with their file name

4.3.3 Results

In Table 4.2, we can separate the interpretation into two groups:

1. The three neurotic scales: Hypochondria, Depression, and Somatization.

2. The two psychotic scales: Paranoia and Schizophrenia.

The F1 scores remaining around 0.5 for the neurotic scales, we may conclude that sleep spindles cannot help to predict any of the three typical neurotic scales. In the case of psychotic scales, the F1 score for Class 0 is above 0.8 for the two typical psychotic scales.

This is a very interesting result as it could help domain experts in their diagnosis as they could eliminate these possibilities to focus their diagnosis better. It is interesting to note that we can interpret this differently for psychotic and non-psychotic. This leads to the question of whether the

Mean F1 score	Hypochondria	Depression	Somatisation	Paranoia	Schizophrenia
Class 0 = False	0.478± 0.27	0.657± 0.32	0.514± 0.38	**0.818**± 0.29	**0.822**± 0.36
Class 1 = True	0.614± 0.28	0.528 ± 0.32	0.557± 0.35	0.226± 0.33	0.139± 0.4

Table 4.2: Mean F1 scores over ten folds for the prediction of 5 psychopathological traits with Violet algorithm. The dataset is balanced (51% for Class 1, 49% for Class 0).

link between the psychopathological traits and the characteristics of the sleep spindles is proportional to the severity of the trait. Indeed, psychotic traits are considered to be more severe diagnoses.

4.3.4 Discussion

This study is conducted on a relatively small data set comprising 267 patients. Moreover, these patients had consulted the domain expert for sleep-related issues. Hence, these patients do not represent a broad enough population. The study should be conducted on patients from a wider background to be properly generalized. Moreover, not all parameters were taken into account. For instance, our study did not consider a patient's medical treatments, which may affect sleep spindles.

The spindle detection has its limits as the three detection algorithms used gave drastically different measures for the density and duration of the spindles. However, they conserved the order, meaning that if patient A had a higher density than patient B, as evaluated by Moelle 2011, YASA would also show a higher density of spindles for A than B. Then, these values can be used within the study, but any given value may not be extracted and interpreted. For instance, stating that female patients have an average spindle density of 3.5 is false, but saying that female patients have a higher spindle density than men may be true.

Finally, there is a limitation that may be linked to our choice to consider only stages N2 and N3 for the spindle detection. Although spindles normally only appear in these two stages, this choice could theoretically impact the number of spindles in absolute terms when some spindles are part of another epoch labeled N1, W, or R but not in terms of density. As we considered essentially the spindle density per epoch and the spindle duration as the main features and not the number of spindles per night, we considered this potential source of error marginal.

4.4 Subtyping Insomniacs with Significant Difference in Subjective Sleepiness using Graph Spectral Theory and clustering techniques on raw EEG and hypnogram scored by expert

4.4.1 Introduction and hypothesis genesis

In this last study, given the results we have so far from using EEG features, we designed a protocol to predict the two most important subjective features in our populations, namely insomnia severity and perceived daytime sleepiness potentially associated with sleep deprivation. We have decided to include this section in this chapter because of the signal analysis work involved, which goes beyond simple data exploration. Indeed, we have a specific hypothesis: the prediction of insomnia severity and sleep perception and alertness with EEG signal. Depending on the results, we hope to be able to cross-reference these findings with the prediction of Paradoxical Insomnia. For this study, we took a subsample of CID patients with only the ISI and ESS scores to see what could predict these scores from the sleep EEG and hypnogram organization. For this experiment, we need to extract EEG microstructure, mainly usual bandwidth frequencies, and chronological scoring from the expert. The main objective is to find clusters of insomniacs, verify if they are related to the main subjective scales, and perhaps reveal some non-evident characteristics of different groups [48].

Theoretical background

As described extensively in 2.4.1, the ISI questionnaire is associated with insomnia severity perception and the ESS questionnaire with sleepiness perception. Then, these two questionnaires evaluate two cardinal values in sleep medicine, namely the presence of wakefulness in sleep (ISI) and the presence of sleep in wakefulness (ESS). So, they are theoretically designed to give a correct and symmetrical reflection of a physiological disturbance, both to measure clinical evolution and to assess a treatment's efficacy. For ESS, a recent review on physiological correlates of ESS [127] reveals on a community-based sample, using ML models, that standard measures of sleep are not predictive of ESS scores, nor are these scores well correlated with measures of sleepiness. They used RF regression analysis and LASSO to predict the ESS score. They used two types of variables: medical and sleep variables. The sleep variables (18) included time spent in stage REM, N1, N2, N3, PSG, WASO, stage REM-N1 shifts per hour, stage REM-N2 shifts (per hour), stage REM-N3 shifts (per hour), stage W-sleep shifts (per hour), stage N2–N1 shifts (per hour), stage N3-N2 and N1 shifts (per hour), REMLatency, RDI, Arousal Index, NREM Arousal Index, REM Arousal Index, and TIB. Raw EEG was not used. They chose to explain the variance of ESS according to 55 variables, including the 18 sleep variables described. 7.15%–10.0% of the variance of ESS scores could be explained. The most important predictor was the self-reported frequency of not getting enough sleep, age, and gender. A study could test the impact of the AHI range on the EES score and find a significant relationship between the AHI index and the snoring status with the ESS score. The other PSG features and EEG were not available in this study [77]. Same research in Pubmed for ""Physiological correlates of Insomnia Severity Index"" and ""EEG" AND "Insomnia Severity Index"" didn't retrieve any publication on the subject. So, our study would be the first to explore this relationship.

Hypothesis genesis

The main hypothesis here is that we could find clusters (linked to specific subtypes) of CID patients with a significant difference in terms of subjective sleepiness (ESS questionnaire score) and insomnia severity (ISI questionnaire score) using Graph Spectral Theory and clustering techniques using raw EEG and hypnograms scored by sleep experts. We also hypothesize that this could help better understand ParI subtypes. The advantage of using these two questionnaires is that they are the two most closely related to the perception of wakefulness and sleepiness, so we hope to find a link between sleep state misperception (SSM), EEG, and ISI/ESS scores to help refine its clinical and physiological definition. Recently, a study [101] supports our hypothesis of finding neurophysiological insomnia subtypes after sleep deprivation. Indeed, they could find three subtypes derived from the data-driven classification of PSG, EEG spectral power, and interhemispheric EEG asymmetry index. They also claimed that this subtyping process could be linked to SSM, with a subtype named Short Sleep

Deficiency. However, a quick analysis of the results presented shows that there are some limitations in this study, such as the small sample size (N=26 subjects) and the fact that this small sample subtype presents a huge SD with a significant proportion of subjects presenting positive sleep misperception. Anyway, the interest of this study is showing that we could find some clusters based on EEG features and hypnogram features and that with more subjects we could probably find some physiological subtypes.

So, we are making the assumption, despite the results described in [127] that didn't study specifically CID population, that we could find a link between these two central questionnaires, the raw EEG and the Sleep report features given by experts to find some clusters of insomniac patients. Indeed there is no similar attempt in the literature to find some correlations between quantitative values between subjective questionnaires theoretically designed to give a correct and symmetrical reflection of a physiological disturbance, both to measure clinical evolution and to assess the efficacy of a treatment.

4.4.2 Methodology and experimental design

Dataset used

In this research, EEG data were procured from the 576-participant PSG Database DIII-PSG, as detailed in Table 3.3. A subset of 386 records was meticulously chosen based on superior data quality. Furthermore, the inclusion of three participants with Insomnia Severity Index (ISI) scores below ten was intentional, to underscore the potential disparities between ISI scores and derived EEG characteristics.

Feature extraction protocol

Macro Features From the hypnogram Analysis generated by two experts, the different steps to use these macro features were done as follows:

We used the following standard features:

Feature	Description
W	30 s of EEG scored Wake
N1	30 s of EEG scored N1
N2	30 s of EEG scored N2
N3	30 s of EEG scored N3
REM	30 s of EEG scored REM
Time in bed (min)	Time from go to bed until get up
Sleep time (min)	Total time sleep (from first episode of sleep to last Wake episode)
Stage time(min)	Time spent in W, N1, N2, N3, REM sleep stages (min)
Stage percentage	Percentage spent in W, N1, N2, N3, REM sleep stages (%)
Awake frequency	Frequency of awakening during Total sleep time
Awake Percentage e	Percentage of awakening during Total sleep time
Frequency of stage	Stage shifting frequency

Table 4.3: 15 Standard hypnogram features used to generate standard hypnogram features, graph spectral features and Levenshtein feature

We used a similar experimental design concerning graph spectral theory described in [37] that already attempts to describe complex sleep dynamics throughout transition networks and scalar measures in insomnia. This study used EEG and sleep stages to quantify and differentiate control and insomnia on the sleep onset periods period only. From this study, we kept the same protocol concerning the transition Networks and Graph Spectral Theory described as follows, but we didn't apply classification, instead, we performed a K-Means clustering.

Transition Networks and Graph Spectral Theory design

Hypnograms were converted into sleep transition networks analyzed by spectral graph theory to represent sleep stage interactions regarding matrix properties and spectra. This approach avoids visual graphical representations in favor of isomorphic network comparisons based on eigenvalues and eigenvectors. Subsequently, similarity distances derived from graph spectral metrics were calculated to quantify the similarity between each subject's sleep transition networks.

Then, to compare this graph similarity, if G and H are transition networks of two subjects with the same structure as in Figure 4.28, they will have different edge weights according to their transitions amongst sleep stages. We could call it a vector signature. Initially, each network's degree matrix D_G, adjacency matrix A_G, and incidence matrix C_G are derived for two networks G and H.

Degree matrix D is a diagonal matrix with $D(i,i) = 0$ if vertex i has no self-directing shifts or $D(i,i) = 1$.

Adjacency matrix A depicts the connection between different vertices and is a 5×5 matrix (true for fixed model size) with $A(i,j) = 0$ if vertex i has no shift to vertex j, or else $A(i,j) = 1$.

The incidence matrix shows the connection relationship between vertices and edges. In our project, the incidence matrix C is a 5×25 matrix, where each row represents one vertex and each column represents one edge. For vertex i and edge jk,

$$C(i,jk) = \begin{cases} \frac{2}{w(j,k)}, & \text{if } i = j = k \text{ and } w(j,k) \neq 0, \\ \frac{-1}{w(j,k)}, & \text{if } i = j, j \neq k \text{ and } w(j,k) \neq 0, \\ \frac{1}{w(j,k)}, & \text{if } i = k, j \neq k \text{ and } w(j,k) \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Three similarity distance measures based on these graph-related matrices were calculated for the experiment based on [37].

The similarity distance $d_1(G,H)$ between G and H; correspond to subtraction between the diagonal matrix of noninverted weights (degree matrix $D_{G|H}$) and the full-rank matrix of inverted directed weights (adjacency matrix $A_{G|H}$) to produce a Laplacian matrix $L_{G|H}$.

Upon the Laplacian matrix, an Eigenvalue Decomposition (EVD) finds the corresponding eigenvalue for each network (See B.1.2). So at the end, we have the eigenvalues

$$\lambda_i \text{ from G network and } \nu_i \text{ from H network.}$$

$$L_{G|H} = D_{G|H} - A_{G|H} \overset{\text{EVD}}{=} Q_{G|H} \Lambda Q_{G|H}^T$$

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}, \quad d_1(G,H) = \begin{cases} \sqrt{\sum_{i=1}^N (\lambda_i - \nu_i)^2} / \sum_{i=1}^N \lambda_i^2, & \text{if } \sum_{i=0}^{N-1} \lambda_i^2 \leq \sum_{i=1}^N \nu_i^2, \\ \sqrt{\sum_{i=1}^N (\lambda_i - \nu_i)^2} / \sum_{i=1}^N \nu_i^2, & \text{if } \sum_{i=1}^N \lambda_i^2 \geq \sum_{i=0}^{N-1} \nu_i^2. \end{cases}$$

The similarity distance $d_2(G,H)$ use the adjacency matrices$A_{G|H}$ to perform an EVD. The total transformation is presented below.

$$A_{G|H} \text{ EVD} = Q_{G|H} \Lambda Q_{G|H}^T, \quad \Delta = A_G - Q_G Q_H A_H Q_H^T Q_H$$

$$\Delta = \begin{bmatrix} \delta_1 & 0 & 0 \\ 0 & \delta_2 & 0 \\ 0 & 0 & \delta_3 \end{bmatrix}, \quad d_2(G,H) = \frac{1}{\sqrt{N}} \sqrt{\sum_{i=1}^N \delta_i^2}.$$

The similarity distance $d_3(G,H)$ follows the same steps as $d_2(G,H)$, however, an incidence matrix $C_{G|H}$ triggers the calculation rather than the adjacency matrix.

$$C_{G|H} \text{ SVD} = U_{G|H} \Sigma V_{G|H}^T, \quad \hat{\Delta} = C_G - Y_G^T U_H C_H V_H^T V_G$$

$$\hat{\Delta} = \begin{bmatrix} \hat{\delta}_1 & 0 & 0 \\ 0 & \hat{\delta}_2 & 0 \\ 0 & 0 & \hat{\delta}_3 \end{bmatrix}, \quad d_3(G,H) = \frac{1}{\sqrt{N}} \sqrt{\sum_{i=1}^N \hat{\delta}_i^2}.$$

From the sleep stage sequence, we generated a network of sleep stage transitions (See Figure 4.28). We can see a network of 5 vertices, each representing one sleep stage. The weight of the directed edge, e.g. from vertex N 1 to vertex N 2, represents the shift frequency from sleeping stage N1 to N2. We applied graph spectral theory to describe sleep patterns in graph-related matrices and spectra. The comparison of two hypnograms is turned into the comparison between two networks with homogeneous structures (See Figure 4.28).

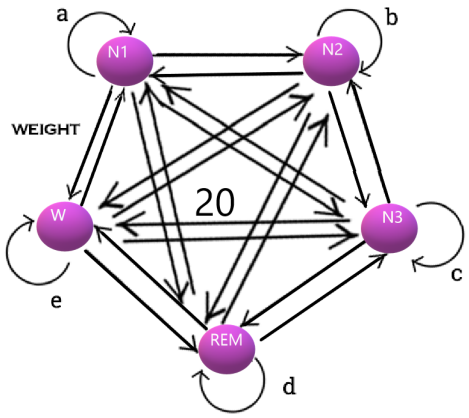


Figure 4.28: Sleep stage transition network of 5 vertices. The arrows represent the potential edge from vertex i to vertex j. The weight of the edge measures correspond to the shifting frequency from vertex i to vertex j.

String similarity feature

Based on the generated sleep stage sequence, the Levenshtein distance [67] was used to measure the similarity of two hypnogram patterns. It calculates the number of operations (insertion, replacement, deletion) to transform string1 into string2. We used the same reference patient mentioned before and calculated the Levenshtein distance between the reference sequence and all the other 386 patient sequences*. The process used is as follow :

To calculate the Levenshtein distance between two strings, let's say string A and string B, the algorithm follows these steps:

1. Initialize a matrix (often called the Levenshtein matrix) with dimensions $(m+1) \times (n+1)$, where m and n are the lengths of strings A and B, respectively.

2. Initialize the first row and the first column of the matrix with values 0 to m and 0 to n, respectively. These values represent the number of insertions or deletions required to convert an empty string to A or B.

3. Iterate through the matrix, starting from the second row and the second column.

4. At each position (i, j) in the matrix, calculate the cost of transforming $A[1 : i]$ to $B[1 : j]$ as follows:

   - If the characters $A[i]$ and $B[j]$ are the same, the cost is equal to the value at position $(i-1, j-1)$ in the matrix.
   - Otherwise, the cost is the minimum of the following three values:
     - The value at position $(i-1, j)+1$, representing the cost of deleting the character $A[i]$.
     - The value at position $(i, j-1)+1$, representing the cost of inserting the character $B[j]$.
     - The value at position $(i-1, j-1)+1$, representing the cost of substituting the character $A[i]$ with $B[j]$.

5. Once the iteration is complete, the value at the bottom-right corner of the matrix represents the Levenshtein distance between strings A and B.

Microfeatures selection

Power Spectral Density features EEG signals obtained was first filtered (0.3-30Hz), with Notch denoising and artifacts were removed by the expert. Only EEG signal of channel 'F3-M2','C3-M2','C4-M1','O1-M2' were used in our analysis

The power spectral density (PSD) was calculated using the Welch Method for an entire frequency range of 1-30Hz with the MNE-Python package. Then for the particular EEG rhythms: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–11 Hz), sigma (11-16hz), beta (16–30 Hz), their average power of frequency band were calculated respectively.

Clustering

We conducted **clustering** based on standard hypnogram feature embedding, graph spectral distances, and Levenshtein distances, respectively. Our embedding was first transformed to a 3-dimension vector if their dimension was larger than 3, and then a K-means clustering method with k=2,3,4,5 was used after the optimal number of clusters determination by the "elbow" method.

Statistical analysis

Once the clusters were detected, we carried out a one-way ANOVA and the Kruskal test for the ISI and ESS scores of different clusters, respectively. We accept that there are at least two clusters with a significant difference in their mean score with a confidence level of 5%. A further two-pair T-test was conducted to determine if there was a significant difference for each pair of clusters.

Embedding Paradigm

For each patient, we have constructed the following features with the following dimensions, for a total of 2200 possible features:

1. Standard hypnogram features: 15
2. Graph Spectral Distances: 3
3. Levenshtein distances: 2
4. Entire EEG signal PSD (4 channels)
5. Entire EEG signal Average Band power (4 channels)
6. Sleep stage (5 stages) EEG signal PSD (4 channels)
7. Sleep stage (5 stages) EEG signal Average Band power (4 channels)
8. First 20% EEG signal Sleep stage (5 stages) EEG signal PSD (4 channels)
9. First 20% EEG signal Sleep stage (5 stages) EEG signal Average Band power (4 channels)
10. Last 20% EEG signal Sleep stage (5 stages) EEG signal PSD (4 channels)
11. Last 20% EEG signal Sleep stage (5 stages) EEG signal Average Band power (4 channels)

4.4.3 Results

ISI and ESS description on the dataset

A patient with an ISI score lower than 11 is regarded as having slightly no insomnia, and an Epworth score lower than 11 indicates that the patient has very little tendency to feel tired after a night's sleep. Among the 386 patients with the two scores, we separated them into 4 groups:

- Group 1: ISI < 11, Epworth ≤ 11
- Group 2: ISI < 11, Epworth > 11

- Group 3: ISI ≥ 11, Epworth ≤ 11
- Group 4: ISI ≥ 11, Epworth > 11

From Figure 4.29, it could be observed that the majority of the patients fall into the 3rd and 4th groups, with an ISI score mean of 19.47 (std 4.50) and an Epworth score mean of 8.14 (std 5.10).

The description of the sample concerning ISI and ESS score plot is presented in 4.30.

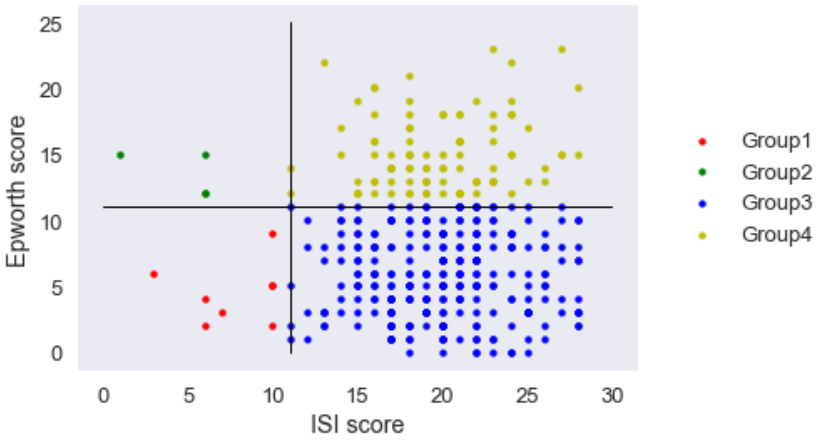


Figure 4.29: Scatter plot for ISI and ESS score

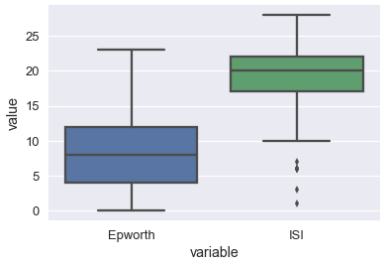


Figure 4.30: Distribution of ISI and Epworth (ESS) Scores

Impact of Respiratory Disturbance Index (RDI) on ESS severity score

From the first section, we saw that RDI (Respiratory Disturbance Index) could have an impact on ESS score via the sleep microfragmentation[77]. To see if our protocol could have only detected clusters of patients presenting more RDI, we ran a simple Pearson Correlation analysis on our sample. The results are presented in Table 4.4. We could see that there is no correlation in our sample. We will discuss this observation later.

	RDI	**ESS severity**
RDI	1.000000	0.084168
ESS severity	0.084168	1.000000

Table 4.4: Pearson correlation between the ESS score severity and the Respiratory Disturbance Index (RDI). The RDI take into account all the respiratory events detected (Sleep apneas and respiratory limitations (without oxygen desaturation >3%)).

Clustering with Hypnogram features

As mentioned earlier, we generated three types of hypnogram features: standard hypnogram features (15-dimensional vector), spectral graph features (3-dimensional vector), and Levenshtein features (2-dimensional vector). Considering all the hypnogram features, we have a data embedding of 20 dimensions.

We obtained different clusters using K-means directly on this embedding, with cluster numbers ranging from 2 to 5. After conducting a one-way ANOVA test at a confidence level of 5%, we found that none of the cluster results showed a significant group difference in the patient group or ISI score. However, for cluster numbers 2, 4, and 5, there was a significant difference between groups in terms of Epworth score.

The results of the ANOVA for ISI and ESS are presented in Table 4.5 and 4.6.

Clustering with EEG features

Using the EEG PSD features, we conducted the same analysis method with the 2180 dimension embedding data. Unfortunately, The ANOVA test showed that none of the patient scores of obtained clusters manifests a significant group difference. So the result is not presented to save space.

Cluster number	F-statistic	p-value	H-statistic	p-value
2	0.78	0.37	0.72	0.39
3	2.22	0.10	1.96	0.37
4	2.46	0.06	5.65	0.12
5	2.20	0.06	5.44	0.24

Table 4.5: ANOVA test and Kruskal test of different cluster number on Patient ISI score data

Cluster number	F-statistic	p-value	H-statistic	p-value
2	14.64	**0.00015***	13.15	**0.00028***
3	2.47	0.08	4.09	0.12
4	4.94	**0.0022***	13.47	**0.0037***
5	3.31	**0.01***	12.19	**0.01***

Table 4.6: ANOVA test and Kruskal test related to the number of clusters on ESS scores

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
7.6 ± 4.9 (2)	**10.03 ± 5.3 (1)**	None	None	None
8.57 ± 5.4	7.67 ± 4.7	10.07 ± 5.1	None	None
7.88 ± 5.1 (2)	**10.15 ± 5.3 (1,3)**	**7.25 ± 4.5 (2)**	8.57 ± 4.30	None
7.22 ± 4.7 (4)	**7.30 ± 4.6 (4)**	8.57 ± 4.3	**9.85 ± 5.3 (1,2,5)**	**8.06 ± 5.1 (4)**

Table 4.7: Comparison of the Mean of ESS scores with t-test when we applied k=2

4.4.4 Discussion

The study aimed to predict insomnia severity (ISI) and perceived daytime sleepiness (ESS) in CID patients using EEG features. The clustering analysis was conducted based on hypnogram, and EEG features to identify potential clusters of insomniac patients and explore their relationship with subjective scales. The main points of discussion are :

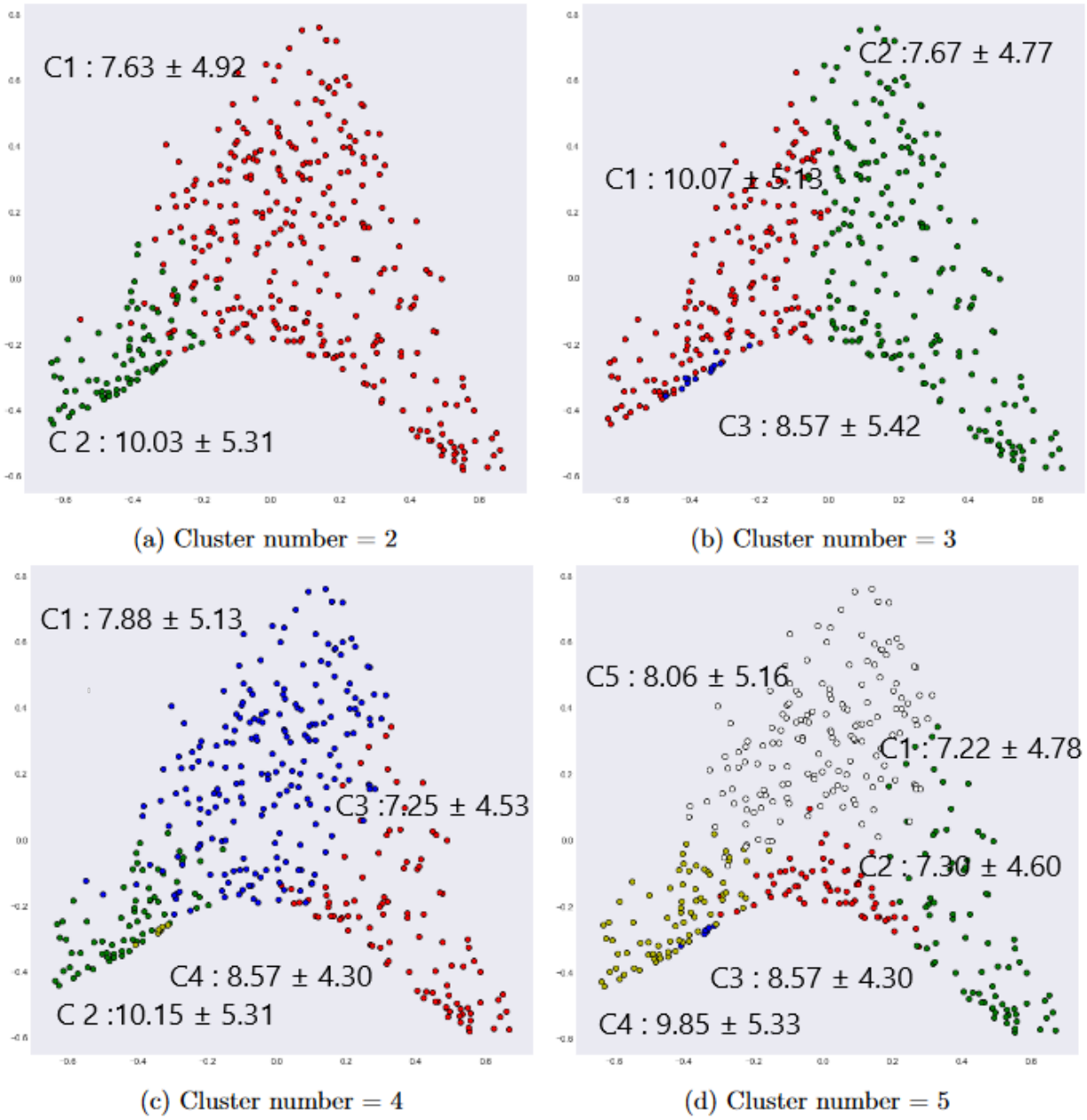
- The study utilised standard hypnogram features, spectral graph features, and Levenshtein features for clustering analysis. While no significant differences were observed in ISI scores among the clusters, significant differences were found in Epworth scores for certain cluster numbers. This is the first study showing such a result in a sample of CID. This finding highlights the potential of clustering techniques, combined with the analysis of sleep-related features, to uncover distinct subgroups or phenotypes within the CID population based on subjective sleepiness. This novel insight can contribute to a better understanding of the subjective experiences and symptomatology of individuals with CID.

- However, it is worth noting that no significant differences were observed in Insomnia Severity Index (ISI) scores among the clusters. This finding is surprising because ISI is one of the most validated measures that assess the severity of insomnia symptoms. The absence of significant differences in ISI scores suggests that the clustering analysis may have been less effective in capturing variations in insomnia severity within the CID population or that this subjective scale is not correlated to the objective measure of sleep fragmentation.

- These findings underscore the complexity and heterogeneity of CID as a sleep disorder. Insomnia manifests in various ways, with diverse underlying causes and subjective experiences. The significant differences in ESS scores indicate that subjective sleepiness may be a more prominent and discernible feature among CID patients, while insomnia severity may be influenced by many factors not easily captured by the selected features and clustering approach.

4.4.5 Limitations

The study acknowledged limitations, including the small sample size and potential heterogeneity within the patient population. Future studies with larger sample sizes and diverse populations could



(a) Cluster number = 2 (b) Cluster number = 3

(c) Cluster number = 4 (d) Cluster number = 5

Figure 4.31: Projected cluster visualization (for different values of k) using Spectral Embedding for Hypnogram features with each ESS mean and SD score

provide more robust insights into the relationship between EEG features, hypnogram organization, and subjective sleep measures. Additionally, exploring alternative EEG analysis techniques, such as Deep Learning algorithms, may perhaps offer deeper insights into the physiological characteristics of different insomnia subtypes.

4.4.6 Conclusion

In conclusion, this study aimed to predict insomnia severity and daytime sleepiness using EEG features and hypnogram organization in a population of CID patients. The clustering analysis based on hypnogram features revealed significant differences in perceived daytime sleepiness among certain clusters. However, the clustering analysis using EEG features did not yield significant group differences in patient scores. The correlation analysis did not show a significant relationship between RDI and ESS severity scores. These findings suggest that the selected EEG features and clustering approach may have limited predictive power for insomnia severity and daytime sleepiness in this population.

Further research with larger sample sizes and alternative EEG analysis techniques is necessary to gain a better understanding of the physiological characteristics underlying different insomnia subtypes and their associations with subjective sleep measures.

Chapter 5

Explaining Negative Sleep State Misperception: A Machine-Learning Approach

Chapter Highlights

We used a dataset with 335 samples and all the features available to assess CID (see dataset four in 3.1.1) to apply the formulas published in the literature for defining ParI (cf. Table 2.6), we want to test the following hypotheses:

1. First hypothesis: The implementation in our dataset of the main formulas published to define ParI and their prevalence analysis will confirm the poor overlap between formulas: We will confirm that there is poor or no overlap between the subjects diagnosed as ParI, depending on the formulas used.

2. Second hypothesis: Finding the more accurate predictive model on each formula prediction will allow their explanation We determined the most accurate predictive models (among most used binary classifiers) selected after benchmarking (plus LASSO, for its feature-selection characteristics), and we did feature importance analysis and feature selection to explain the prediction. Our methodology is named "Ensemble Type Method for Prediction Explanation" (ETMPE) for final feature selection.

3. ParI, is there a possible harmonization across formula definitions? By adding the selected variables found by ETMPE, we proposed new meta bio-markers to reinvent the diagnostic and characterization of ParI.

4. Proposal for a new definition of Paradoxical Insomnia including seven nights sleep analysis We consolidate the results above to propose a new definition of ParI.

Key Terms and concepts

Acronym	Definition	Ref.
AG	Actigraphy	p. 28 (2.4.1)
CV	Cross-validation	p. 172 (B.1.2)
CID	Chronic Insomnia Disorder	
EEG	Electroencephalogram	p. 169 (B.1.1)
FI	Feature Importances	p. 173 (B.1.2)
LASSO	Least Absolute Shrinkage and Selection Operator	p. 173 (B.1.2)
MMPI	Minnesota Multiphasic Personality Inventory	p. 32 (2.4.1)
ParI	Paradoxical Insomnia	p. 32 (2.4.2)
PsyI	Psychophysiological Insomnia	p. 170 (B.1.1)
MoSA	Morris Sensitivity Analysis	p. 170 (B.1.1)
SOL	Sleep Onset Latency	p. 170 (B.1.1)
SHAP	SHapley Additive exPlanations	p. 175 (B.1.2)
XAI	explainable AI	p. 176 (B.1.2)

5.1 Introduction and hypothesis genesis

This chapter will test our first main hypothesis described in 1.3, i.e.: **Can we provide an improved definition of ParI** using a data-driven approach with machine learning tools?

We discussed in previous chapters why this question is of interest: firstly for the sleep medicine community, but also in terms of a new approach to making causal inferences in medicine using ML tools as a new way to bring new insight on unresolved issues like ParI (see 2.6).

So, in this chapter, we will use predictive and explainability methods (see B.1.2 for definition), making the general hypothesis that we can find a new way to explain ParI thanks to the explanation of complex feature interaction in the predictive models. In particular, we aim to study the use of feature explanation algorithms associated with classifier models to evaluate this hypothesis.

This hypothesis came after seeing the interest in using ML as a new tool to solve or improve medical issues is directly linked to the transparency of the algorithms and the techniques involved in explainability. These methods belong to a relatively new field, the explainable AI (XAI, see definition and development in B.1.2), which has been growing since the late nineties [105]. So, these tools could make the perfect intermediary between complex algorithm understanding with good results in predicting an outcome and the need for understanding the prediction by the final user, especially in medicine. We also think that these tools could also participate in discovering new causal inferences in sleep medicine research.

Our approach aims to provide a holistic understanding of the model behavior rather than focusing on individual predictions. This is exactly what the Global Explanation model is doing, trying to explain the overall logic, decisions, or rules the model uses on the entire data to predict the target. On the contrary, we had no interest in our protocol to use local explanations for individual prediction as we wanted to explain the concept of ParI and not why a specific subject is classified ParI. Indeed, Global Explanations provide an overview of how a model makes decisions or predictions across all data instances. This contrasts with local explanations, which focus on specific individual predictions. Global explanations aim to describe the overall behavior and logic of the model, offering insights into feature importance, decision rules, and the model's structure. These explanations are crucial in contexts where understanding the model's decision-making process is as important as the predictions, especially in high-stakes areas like healthcare. Then, in the case of a model-centered global explanation, the FI analysis could analyze the weight of each feature and its influence on the output. So, depending on the model and the explainer used, the weight of a given feature on the result can change; it is the uncertainty of FI. [179].

From a mathematical perspective, depending on the model used, Global Explanation involves summarizing the overall decision-making process of the model. This explanation describes the model's behavior on the entire data distribution. Unlike local explanations focusing on individual predictions or instances, Global Explanations provide insights into the model's general rules and patterns. Depending on the model type, these can be mathematically represented in various forms. For example, in the case of the Linear model or Decision Tree, the Global explanation could be explained as follows:

- **Linear Models:** For a linear model like linear regression, global explanation can be directly interpreted from the model coefficients. For a model f with features $x_1, x_2, \ldots, x_n$ and coefficients $\beta_1, \beta_2, \ldots, \beta_n$, the importance of each feature can be understood from the magnitude and sign of the coefficients. The model can be represented as:

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{5.1}$$

- **Decision Trees:** In decision trees, global explanations can be derived from the structure of the tree itself. The splits at the top (near the root) have more global importance as they affect more data points. The structure can be represented as a series of if-else conditions leading to a decision.

This Chapter aims to evaluate the possibility of explaining the ParI diagnosis using a Global Explanation. As we have already explained in the section 2.4, there are many definitions of this syndrome, diagnosis, or subtype of insomnia to choose from, although all clinicians dealing with CID are confronted at least every week to this paradox. So, the problem is using supervised ML on a fluctuating definition from one formula to another and doing systematic feature extraction with many

possible explicative features to see if other features not initially involved in the definition process could bring a new angle of understanding on each definition.

Despite the development of many so-called XAI methodologies, most of this work does not directly consider the utility of methods to a practitioner who may not be well versed in probability theory or game theory and may remain (quite rightfully) distrustful of the most novel developments. This concept, under the name "causability", was addressed by [88], who proposed using an explanation interface after the explainable part to facilitate experts' use and trust. Indeed, in the ideal, the experts could and probably should participate in the understanding, the learning, and why not correcting the algorithm [88].

Our experiment will involve five steps symbolized in Figure 5.1 corresponding to the four hypotheses formulated to interpret the knowledge extracted from the predictive model (Steps two and three are part of the second hypothesis).

1. Implementation and comparison with [35] in our dataset four of the different formulas used to define ParI described in 2.6 (Hypothesis one).

2. Each formula will give a target on our dataset for each sample (0 or 1), six classifiers will be trained on each target, and the one with the best performance will be kept to apply the explainers (Part I of hypothesis two).

3. Two Global Explainer models will use the most performant classifier, LASSO does an additional explanation, and the top ten features among the three are compared (Part II of hypothesis two).

4. Our explainable interface consists of generating a Reliability Score to improve the uncertainty of FI. The metric involved in this evaluation is the Mathews correlation coefficient (MCC) (Hypothesis not developed here).

5. Once the important features are selected, we will cross the results with group comparison by t-test on all the features to discriminate the ones involved in subjects labeled ParI, whatever the formula used (Hypothesis three).

6. Could we have new reliable insight on ParI after this process and perhaps explain or propose a new definition of ParI?

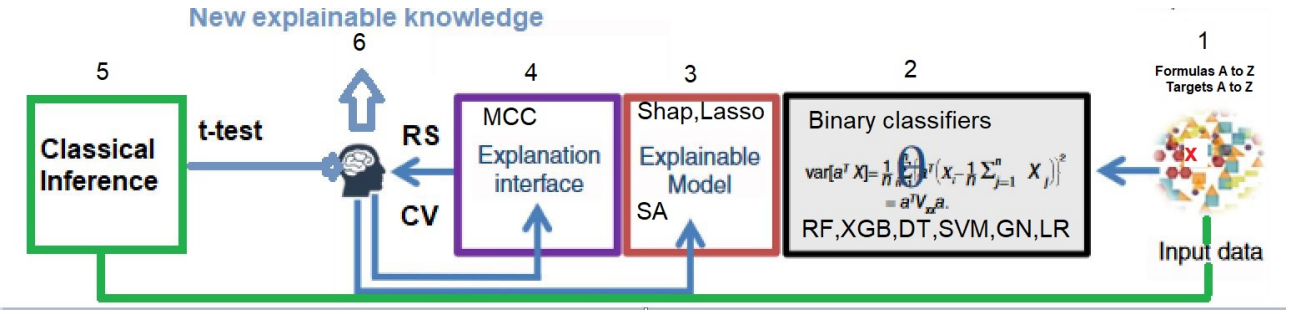


Figure 5.1: Six steps of our experiment protocol: Moving from our datasets to new conclusions about ParI.

5.2 First hypothesis: The implementation in our dataset of the main formulas published to define ParI and their prevalence analysis will confirm the poor overlap between formulas

5.2.1 Many proposed formulas; too much diversity, insufficient agreement

We have already described the vast diversity of definitions used to define ParI from 1979 to 2020 (See 2.6 in the subsection 2.4.2), with most of them already implemented in a single dataset of chronic insomniac patients (N=200) and a control group (N=200) by [35] to evaluate the overlap between most of the formulas (See correlation matrix in Figure 2.16). The agreement ranged from −0.19 to 0.9. A

brief analysis of the results presented shows that certain formulas are highly correlated in the CID sample, such as B and E (>0.8), A and H (>0.7), I and O (>0.7), or the R, T, P, Q, and F formulas > 0.8 between each other. Others are not at all, like M with almost all the other formulas, or B, which is negatively correlated with all the others except E, and the vast majority of the other formulas are weakly to moderately correlated with each other (from 0.2 to 0.4). In the same study (right correlation matrix in Figure 2.16), we can see the correlation matrix applied to a control population. Surprisingly, there is a big change in the correlations compared to the CID sample, especially in a decreasing way, like B and E, which are no longer correlated at all (<0.2), or T and R (from 0.9 to 0.14), or less correlated like I and O (<0.5). Conversely, formulas uncorrelated in insomniac subjects become much more correlated in control subjects, such as L and K (from 0.33 to 0.66) or E and Q (from 0.4 to 0.65). If these results are confirmed, they highlighted that the ParI definition doesn't uniquely define a measure, which may seem surprising because these are a priori formulas with clear cutoffs. This observation raises the question of the indistinct use of ParI and SSM, which could cover two different concepts. The first could refer to pathology, and the second to the perception of sleep. Indeed, more and more studies are now trying to explain SSM negatively and positively, assuming that the negative SSM is the equivalent of ParI. On the other hand, numerous studies claimed that SSM is present in most CIDs, and that's why the subtype ParI should be removed from the classification, but the observation from these findings could lead to separating these two concepts. The second observation from the results presented is the presence of clusters(c), defined by a strong correlation inter-formula > 0.6 or a weak < 0.4 with all the other formulas for a singleton. Applying these criteria to the CID population, six clusters could describe c1(B, E), c2(M), c3(L), c4(R, T, F, Q, P), c5(A, H) and c6(N, I, O). So, at least six different populations of insomniacs could be classified with the same diagnosis across different studies according to these results.

Logically, given the results described above, the prevalence of ParI in a CID sample varies hugely among the studies, between 8% and 66% according to the review of [35]. The prevalence found in their 16 formula implementation ranged from 12 to 64%; They didn't evaluate the correlation between the prevalence and the agreement between formulas. In the studies used to generate the formula calculation (27 papers judged as relevant for the analysis of 282 publications), the prevalence varies from 16 to 60% of the CID sample studies. Some studies pre-selected specifically ParI subjects; in that case, the prevalence is not available for comparison (NA in the table 2.6).

5.2.2 Methodology and tools to test the hypothesis

The initial results described before suggest that our hypothesis is likely to be confirmed. However, the variability of correlations observed in the two data sets prompts us to repeat the same prevalence and correlation study on our dataset. We have to check if we will obtain results close to the one described by [35] on our CID population to reinforce the legitimacy of our dataset for studying this issue versus leading studies on the subject, and thus the power of any results found subsequently. However, we demonstrated in Chapter 3 that our dataset represented a population of chronic insomniacs, so it would be surprising to find divergent results. Furthermore, the prevalence and correlation of formulas not implemented by [35] could be implemented in our dataset (including the new formula published), which would complement the work done by [35] in a sort of update.

We will describe the dataset used and the 20 formulas implemented in the following.

Dataset: 335 CID-subjects described by 198 features

The dataset used for this experiment is dataset four (cf. 3.1.1).

Once data was collected and aggregated, we reduced the number of features by removing duplicates and colinear variables. We made feature engineering removing high correlated value (≥ 0.95). The final dataset used to implement the different formulas contains 198 features and 335 subjects with CID according to the selection process described in 3.1.

The main characteristics of our dataset are a mean age of 46 ± 12 yo, 66% of women, a TST of 355±75 min, and a mean ISI score of 19.7±4.

Implementing the Paradoxical Insomnia formulas/definitions from the literature

From the 23 formulas described in Table 2.6, we have the features needed to reproduce the formula on our dataset for 20. We conveniently used the same name as in [35]. So we kept 18 formulas described in [35], namely A, B, C, D, E, F, J, K, L, L2, M, N, O, P, Q, R, T, and V formulas. We used the original publication to implement the calculation ([22, 195, 109, 85, 180, 138, 58, 57, 160, 131, 152, 62, 103, 94, 10, 144, 102]. We added a formula used in a recent publication [3] named Z. We used the definition S described recently in [117]. After implementation, each subject in our sample was categorized as ParI negative or positive according to each feature's threshold described in each formula publication.

Table B.1 in the appendix shows the distribution by age range, sex ratio, total sleep time, and insomnia severity scale score for the dataset and each formula computation on our dataset. The main characteristics are a mean age of 47±10 yo, 66% of women, a TST of 358±80 min, and a mean ISI score of 19.7±4.

5.2.3 Results (analysis of prevalence and overlap of ParI diagnoses)

We used a representation tool (UpSet plots [123]) to show the prevalence for each formula on our dataset and the overlapping subject by subject (See Figure 5.3a). This plot shows the cardinality of every category combination seen in our data.

We also calculated Pearson's correlations between each formula; the correlation matrix is shown in Figure 5.5.

Correlation matrix We could find the same range of correlation and clusters described earlier by comparing the results obtained on our dataset to [35]. Indeed, Formula M has no significative correlation with other formulas, Formula L neither, except with L2 not implemented in [35]. We also found high correlations between T, P, Q, R, F; between I and O; and B and E. We didn't implement the H formula, so the comparison was impossible for this formula. The J formula is assimilated to the T formulas on the dataset (correlation = 1).

The fact that the differences between the different formulas are quite similar to [35] allows us to draw two conclusions: 1) Our dataset is representative of the target population of CID (considering the validity of the paper [35]). 2) The fact that the results are similar to those studied on another sample makes it even more important to understand what predicts each formula, as this can be reproduced in different clinical samples from different sleep centers.

Overlapping An overlapping visualization technique evaluates the correspondence in our population between the different formulas. We used a graphical representation in Figure 5.3a. This first plot aimed to find a group of patients who could belong to several different definitions of unification. Unfortunately, this plot confirms the discrepancy between formulas, showing that very few subjects on our dataset shared common ParI diagnoses across formulas. Not a single subject had a ParI status common to the 20 formulas. A total of 139 combinations are found, with most combinations shared only by one or two subjects. When we look at the combinations shared by at least ten subjects, there are only five, with the biggest involving 29 subjects. But these combinations concerned a few common formulas; for the most part, only one formula is part of this combination. Ultimately, this plot shows that the biggest sample sharing common definitions are the subjects never categorized as ParI (N=62). But at the same time, taking all the combinations, 82% of our sample was categorized as ParI with at least one formula.

These findings confirm the poor agreement between formulas and between subjects, even with new formulas implemented (L2, S, and Z).

Prevalence The prevalence found on our dataset for each formula aligns with the previous results and shows a very different percentage of subjects categorized as ParI. Indeed, the prevalence presented in 4.19b shows a minimum of 3% of subjects with the Formula M to 51% for the formulas A and L2.

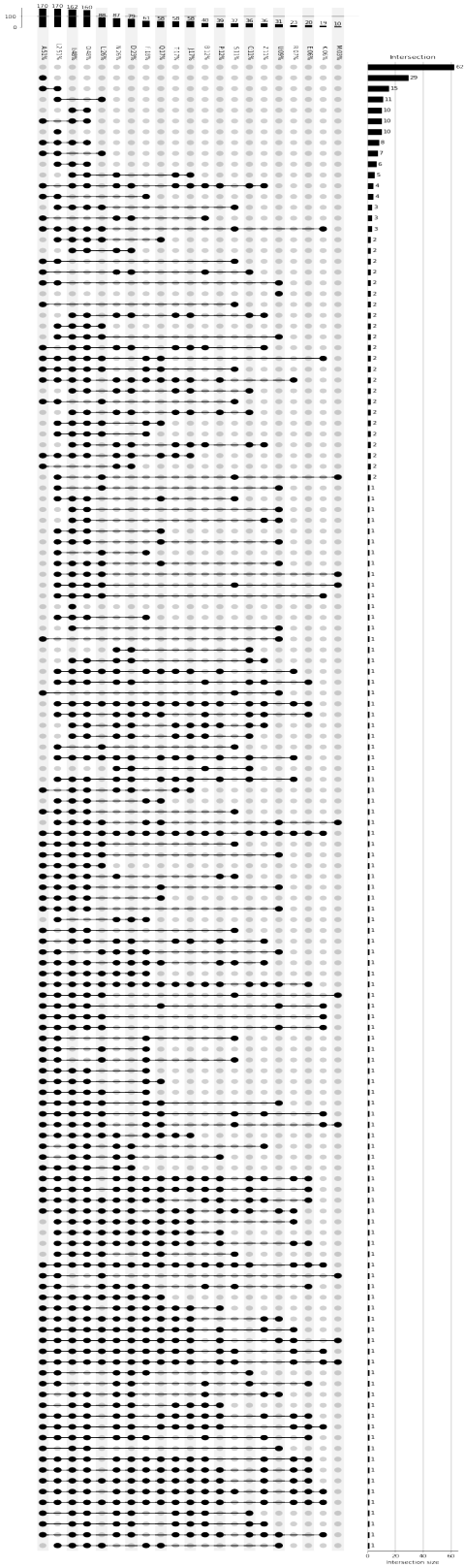
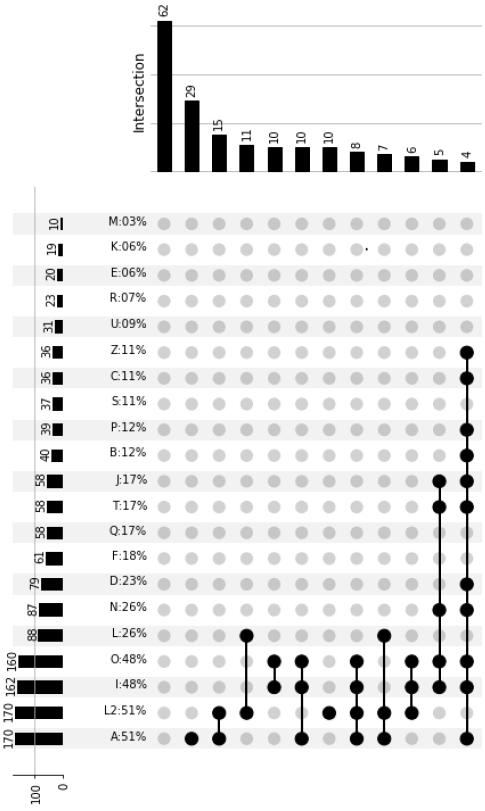


Figure 5.2: Graphical representation of each subject according to their negative or positive categorization as ParI and the common categorization between formulas. A total of 138 combinations for a positive categorization are described for 273 subjects and 62 subjects are categorized as negative ParI as a result of the negative consensus between the 20 formulas



(a)

Figure 5.3: **Figure(a)** is an enlargement of the graphical overlap representation showing the prevalence according to each formula and the 20 first combinations.

5.2.4 Discussion

- We could confirm that our dataset highly represents the typical CID samples used to describe ParI and that our next hypothesis is feasible, at least regarding the target population. Even if

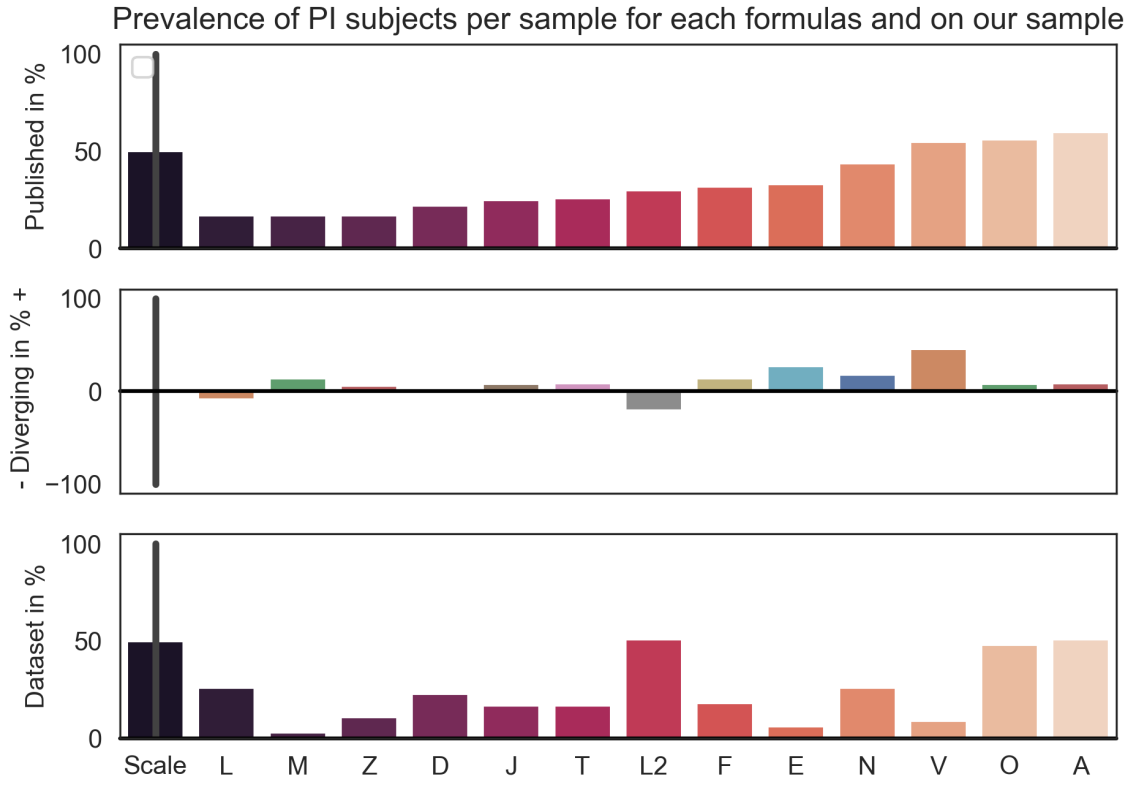


Figure 5.4: Comparison between the original paper prevalence for 13 of the 20 formulas implemented and the prevalence found in our dataset. For the seven left, the information was not available

we don't have a perfect match in terms of prevalence with some of the original papers, we have a perfect agreement with some of the reference papers as [58]. The results are comparable with the literature, especially the study from [35] with similar prevalence in our dataset. We found that the J and T formulas applied to our dataset covered the same subjects [58]; this overlapping was not tested by the previous study by [35]. Figure 5.4 shows the differential comparison with the prevalence found in the original paper. Our dataset shows an under-representation of subjects classified as ParI for formulas E and V. In the case of formula V, this is explained by the sample, which is relatively young and from a military population with few women, which does not correspond to a usual clinical population (17% women and majority of men between 17 and 35 years). This is the biggest difference. For Formula E, the population was recruited by advertisement and did not correspond to a CID consultation. In addition, the mean age in this study is about 35 years, which does not correspond to a typical population of chronic insomniacs either.

- We confirmed the great variability between formulas in their ParI definition and the poor correlation on average between formulas. The best indicator of this very low specificity is the huge percentage of subjects detected as Paradoxical Insomniacs in our sample, which rises to 82% for all formulas combined, even though prevalence varies from 3% to 50% depending on the formula.
- There are no clear clusters corresponding to different subgroups of formulas. Indeed, on 273 patients classified as ParI in our sample, there are 139 mini clusters or, let's say, different ways to classify ParI according to the different formulas. We can see in Figure 5.3a that the bigger cluster corresponds to a unique formula (A).
- There is not a single patient sharing all the formula definitions.
- The most surprising is the lack of agreement between the less specific formulas, A and L2. The correlation between these two formulas is very weak (0.2). Indeed, the number of subjects

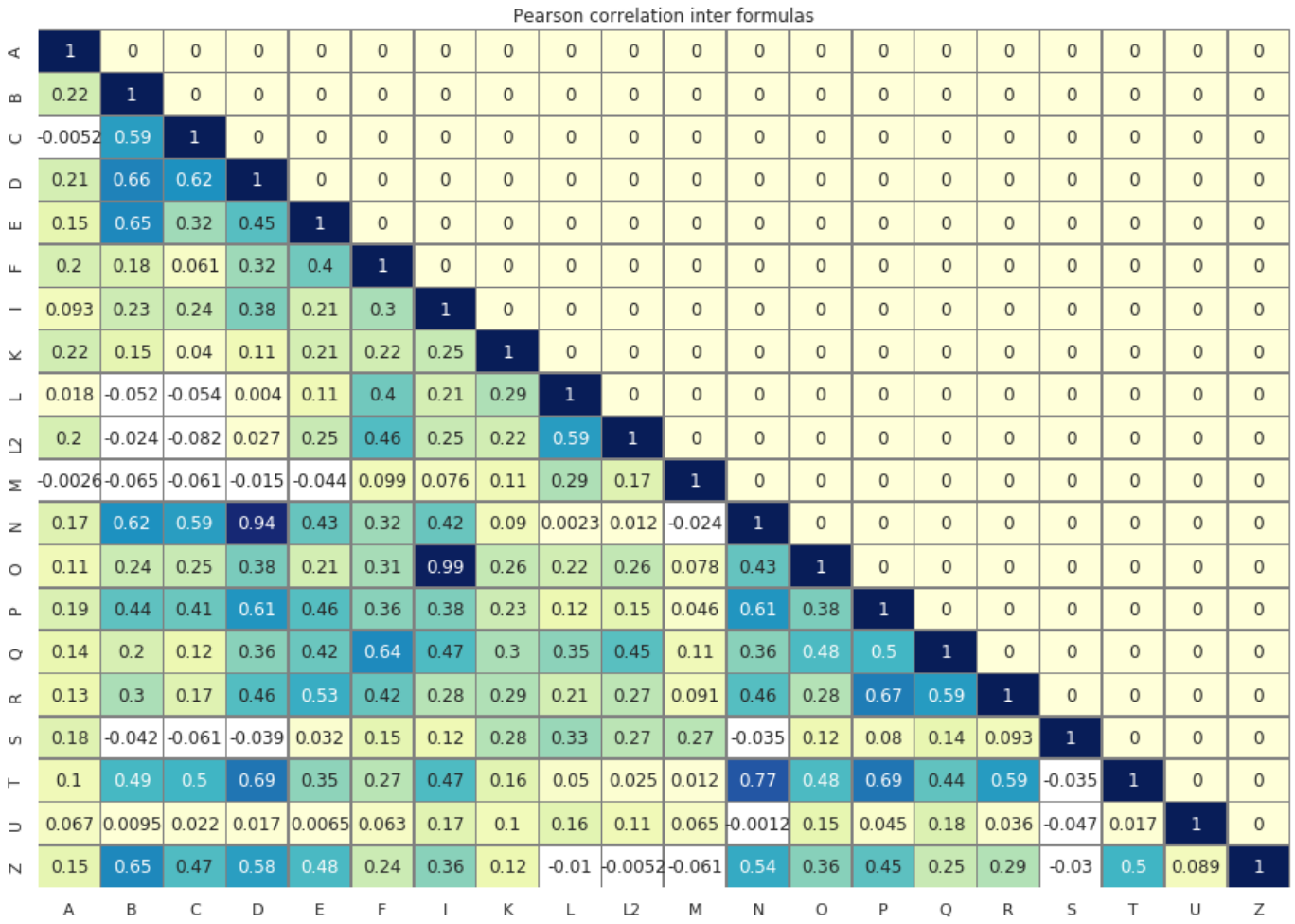


Figure 5.5: Correlations inter-formulas. As the formulas J and T are perfectly overlapping, we removed the J formula from the correlation matrix.

sharing these two formulas is only 17 out of 165, which shows the great diagnosis disparity. This observation is very important, as it means there is no correlation between the prevalence and recovery of formulas. In other words, in that case, this discrepancy in prevalence is not due to a change in specificity or sensitivity but to different criteria for diagnosing ParI.

5.2.5 Limitations

The main limitations of this first experience are:

- We could not implement all the published formulas, as some of the parameters required to calculate them were unavailable in our dataset. However, the features used were broadly the same, but the duration criterion could not be met for one, formula G, which set the criterion of a subjective WASO observed four times a week for one year, a piece of information we did not have. We didn't have the cut-off provided for formula U, but the features were the same as for formula A.

- The number of subjects, although larger than most studies on the subject, often limited to a few dozen patients, fell short of our initial target of 1,000 subjects to increase the power of our study.

5.2.6 Conclusions

We were able to confirm our first initial hypothesis showing that the implementation of the most reliable formulas described in [35] and the most recent papers [117, 3] have a poor global overlap or no overlap at all. We also showed that some formulas differentiated by [35] as O and I are almost perfectly correlated on our dataset, and J and T are fully correlated. Finally, we have shown that our

dataset is very representative of the target population, i.e. a clinical population of CID, as evidenced by the similarities with [35], and the same ParI prevalence in our dataset as in [58] who is the most recognized reference in the field.

As our sample is representative, as we could reproduce the very low specificity of almost all the used ParI definitions on a single representative dataset, we could test our second hypothesis claiming that using additional features available in our dataset, coupled with the use of ML tools; we could identify better the type of profiles to which these different sub-groups of subjects categorized under the single label ParI belong. This is all the more true as we have shown that identical prevalences detected by different formulas did not always overlap between subjects, which means that the different formulas used are not only different thresholds of a single continuum but could correspond to different profiles. The difficulty here will be to differentiate the categorical aspect from the dimensional.

5.3 Second hypothesis: Finding the more accurate predictive model on each formula prediction will allow their explanation

We demonstrated in the previous section that our sample was representative of the target population of CID. Compared to most studies, it's a large sample with additional characteristics. So, by predicting each formula from all the available features except the one involved in the initial calculation and explaining each prediction, we must shed new light on the different subgroups corresponding to each formula, especially when there is no correlation between them.

5.3.1 Hypothesis background and definition

Hypothesis definition

The general hypothesis we want to test here is that a particular ensemble learning protocol might reveal reliable relationships between the dataset features and the target prediction after applying the best possible predictor according to a previous benchmark.

The two following conditions will serve as the tests for this hypothesis:

- If the ensemble learning protocol retrieves at least one shared feature, it suggests the model correctly identifies key relationships.
- If the model never selects a randomly generated feature (within the quantitative range of a common feature in the dataset), this would suggest that the model appropriately and reliably distinguishes between relevant and irrelevant features. By randomly generated features, we refer to a variable or attribute in a dataset created by random processes rather than derived from real-world observations or measurements. These features can serve several purposes, but in our case, a randomly generated feature can be included in feature selection or feature importance analysis to act as a control. If a machine learning model assigns significant importance to this random feature, it might indicate overfitting or issues with the feature selection process.

If both conditions are met, we could support the hypothesis and use the results to gain insight into the ParI explanation. Our objective is to find an acceptable balance between reliability and explanation.

Hypothesis background

The following rationale could support the choice of this hypothesis:

- The algorithms used in this task are supervised learning algorithms, specifically those that deal with binary classification problems, as we want to predict and explain the prediction of each of the formulas defining being ParI positive or negative.

  To increase the chance of success in our experiment, we need to implement tried-and-tested performant classifiers, which can also be implemented in explainers, on the understanding that their decision mode is accessible and explainable, at least to another algorithm. This excludes black box models such as Neural Network classifiers from the outset, especially for the lack of feature attribution. Indeed, unlike models that use easily interpretable decision rules (like decision trees), neural networks do not inherently provide clear attribution of how much each

input feature contributes to the final output. So we chose the non-neural network classifiers from an extensive review of different algorithms on hundreds of different databases and datasets [60] showed that six Random Forests classifiers (RF) and three Support Vector Machine (SVM) are included among the 20 most performant classifiers after hyperparameters tuning, which were the two best families. Those results were in some way replicated in [110]. A more recent paper [187] that tested SVM, K-Nearest Neighbours (K-NN), Naive Bayes, and Decision Tree (DT) have also found SVM with the best results and DT as the second one. Publications using ML on medical databases [93] showed SVM and RF as the best algorithms, but also Gradient Boosting (XGB). Logistic regression (LR) was less efficient in predicting the target, but the scores were still competitive. Another study that compared RF, LR, K-NN, Naive Bayes, and DT on the Breast cancer Wisconsin's dataset[9] found that LR had excellent accuracy after 10-fold cross-validation, and the best was RF. Based on these publications, we could conclude that six algorithms (RF, SVM, DT, XGB, LR, and NB) will likely be the predictive models to test our hypothesis.

- As we want to explain relationships between features from the predictive model, we also have to choose carefully the good metric to assess the classifier performance. Our goal is to focus on a reliable prediction of positive instances (ParI positive), so having an idea of the true positive prediction accuracy is critical. The F1 score (see definition in B.1.2) is well suited to measuring a binary classifier's performance, particularly when dealing with imbalanced datasets [169, 203]. The F1 score is the harmonic mean of precision and recall. Specifically, the F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. So, using the F1 score on positive predictions would measure how well the model predicts positive instances (ParI positive) in terms of both precisions (how many of the instances predicted as positive are positive) and recall (how many of the actual positive instances the model can capture). But more recently, a paper highlighted the advantage of Matthews Correlation Coefficient (MCC) (see definition in B.1.2 methods [40]. To rank the classifiers, we needed the most reliable metric to detect the variables with the greatest impact on the positive prediction of ParI, but also to detect True Negative. MCC is a reliable statistical rate that produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset. Given the selectivity of our criteria for rejecting the hypothesis, we need the most reliable metric to choose the best classifier, so we decided to keep MCC as the main metric for the prediction accuracy, and we added F1 on positive value as an indicator and the classic AUROC (area under the receiver operating characteristic curve).B.1.2

- Once the classifier and the metric are set, the explainability of the prediction needs additional tools named explainer models. These algorithms belong to a complex area of research with many different proposed models. To explain here means determining which inputs have the most influence on the output of the model used. Those algorithm developments took place in response to the need to understand better the decisions made by the increasing implementation of ML algorithms, especially in medicine. So, we want to use these models to help us explain the ParI prediction on our dataset. The choice of good models is not an easy task. We aim to be as selective as possible to avoid selecting meaningless features that could invalidate our hypothesis while keeping sufficient information to explain the prediction. As we want to explain the prediction's main output, we need a Global Explainer model that allows us to understand the average impact of the variables on the prediction. Our literature review focuses on finding complementary models to triangulate the feature selection. At the same time, the explainer should allow us to implement the predictive model or should be adapted to our dataset. We can divide the main explainers models into Interpratables models, surrogate models, and Explanations Generation [25]. The Interpretable models are interpretable alone (white box); the typical example is a Decision Tree or a Linear Model such as LASSO (see the definition in B.1.2). The surrogate models, known to be more flexible and accurate, use the Supervised Machine Learning algorithm (black box) for prediction, but a white box surrogate model of the black box model is provided for explicability. The typical examples are LIME or SHAP [130].

The third category, Explanation generation, is an explainer function that takes a Supervised Machine Learning algorithm (black box or interpretable) and explains it. They are independent of certain model predictions and try to reveal the properties of the black box model. The Leave-One-Out approach and the Global Sensitivity Analysis are two examples [25].

Our conclusion after this review is that we must take one explainer from each category.

We will describe our choices in the methodology section

5.3.2 Methodology and tools to test the hypothesis

Classification tools

- We kept the following six algorithms:
  - Random Forest from scikit-learn
  - Support Vector Machine from scikit-learn
  - Decision trees from scikit-learn
  - Logistic Regression from scikit-learn
  - Naive Bayes from scikit-learn
  - Gradient Boosting (XGB) from the XGBoost library

We added Gradient Boosting (XGB) from the XGBoost library, which recently proved its good performance in medical datasets on accuracy and feature selection [39] at the cost of transparency. We also kept the LR and Naive Bayes classifier as described in Section 2.5.

  - Before training each classifier, we used the grid-search protocol available in Scikit-Learn [162] for automatic hyperparameter tuning.
  - Then we trained each classifier on the dataset for each of the formulas described (A to Z) after removing all the features involved in the definitions except the sleep stages percentage used only in the V definition (see Table 2.6). Indeed, the risk of seeing only these features as the main predictor is high and of poor interest to our research question. A good example to avoid such truism in the explanation of ParI could be found in the B formula paper [195], where the authors have chosen a formula including objective SE $\geq$ 90 % and OSOL$\leq$ 30 min, and in conclusion, they found statistically that what differentiates the most the different groups were these two variables! We aim to see behind this first line of basic definitions which other features could emerge. The last point explains why we are very selective in the feature selection process because we know that removing the main features involved in the construction of the target could lead to low prediction accuracy, and the risk of selecting meaningless features for the explanation is high. The Ensemble type method for prediction explanation described in Chapter 5 and conditions one and two described previously are designed to avoid such a pitfall.
  - For each classifier, we used 10-fold cross-validation (CV) for hyperparameter tuning and prediction. The use of CV is a means of predicting the effectiveness of a model on a hypothetical test set of new patients (results averaged from 10 models, each trained on 90% and tested on 10% of the data, ensuring that the model is always tested on instances it has not been built on). This ensures a generalization of performance. Although we were unable to use a separate Test Set due to the limited number of samples, we limited the risk of overfitting by applying the following protocol:
    1. Class unbalance adjustment of the hyperparameters for each classifier to avoid bias towards the majority class.
    2. Cross-validation of hyperparameter tuning to find the optimal hyperparameters while using the entire dataset. It ensures that the tuning process is not biased towards a specific subset of the data.

3. Repetition of 10*10 cross-validations with additional random columns for each turn and stratified sampling. This step helps to get a more stable estimate of the model's performance. It reduces the variability that might arise from a single round of cross-validation, especially when the sample size is small. The random column addition is designed to test the robustness of the model. If the model starts to assign importance to these random features, it might indicate overfitting. We chose this approach to ensure that the model is not just memorizing the data. The stratified sampling ensures that each cross-validation fold maintains the same proportion of each class as in the full dataset.

Metrics used to assess the predictive model performance

We quantify the performance via MCC, area under the receiver operating characteristic curve (AUROC), and F1 score on the positive prediction. To optimize the classification metrics score (AUROC, F1, and MCC) obtained on each new data set for each classifier, we also tuned the class-weight option of Scikit-learn [162] with the corresponding ratio to avoid bias on the results, especially F1 but also MCC who has shown that he is also sensitive to unbalanced dataset [216].

Explainers and Feature importances Selection

* To ensure the minimum robustness and reliability in the predictive model's explanation, we then chose to use the three categories of explainers (surrogate, interpretable, and interpretation generation), and we chose from them LASSO, Shap, and Sensitivity Analysis.
* This approach is a voting protocol between three different models, a sort of algorithmic triangulation designed to eliminate a feature from the FI selection. As there is similitude with the ensemble models, we named it the Ensemble type method for prediction explanation (ETMPE).
* To understand the minimum to consider a feature as meaningful, we added a random feature in the dataset.
* The protocol adopted was to keep the best ten mean scores associated with each feature after ten runs of explanation after dataset sampling.
* Once the top 10 means for each explainer are kept, we compared the features and kept only the shared features. This approach is an existing measure of faithfulness already published and described as "Correlation to others", a common method of testing explainers. The principle is to compare their output to other popular existing methods [2].
* The originality of our methods is that we applied a systematic 10-fold repetition for each explainer with the randomly filled column. So, we ran 30 Feature importance determinations per formula, then 600 in total, to evaluate the rate of random feature selection and the score we could consider significant.

5.3.3 Results

Best predictions for each Formula

The performance results in predicting ParI are shown in Table 5.1. The best performance on the F1 score on the positive prediction on MCC and AUC is shown for each classifier. We can see that few formulas reach satisfactory prediction scores. Indeed, only formulas B, D, I, and O seem to be able to be predicted with AUC scores $\geq 85\%$ and an F1 score $\geq 70\%$. Formula N is at the limit, with an AUC score of 84% and an F1 score of 68%. Interestingly, this corresponds to the most correlated formulas.

Formula	Model	AUC	F1	MCCb
A	XGB	.57±.01	.58±.02	.15±.03
B	LR	.91±.02	.67±.02	**.53±.03**
C	SVM	.84±.01	.52±.02	**.31±.03**
D	LR	.87±.00	.70±.01	**.51±.02**
E	SVM	.88±.01	.48±.03	.28±.04
F	SVM	.64±.02	.47±.02	.18±.03
K	RF	.85±.00	.40±.01	.27±.02
L	XGB	.68±.01	.50±.01	.18±.03
L_2	LR	.69±.01	.61±.01	.26±.01
M	RF	.80±.01	.29±.01	.15±.02
N	LR	.84±.00	.68±.00	**.47±.01**
O	XGB	.88±.00	.78±.01	**.56±.01**
P	LR	.84±.01	.57±.01	**.38±.02**
Q	RF	.80±.00	.49±.00	.29±.01
R	RF	.82±.01	.38±.01	.22±.01
S	RF	.76±.01	.45±.01	.25±.01
T	LR	.80±.01	.60±.01	**.40±.02**
V	RF	.68±.00	.38±.00	.18±.01
Z	SVM	.84±.01	.58±.02	**.40±.03**

Table 5.1: Best score obtained on Matthews Correlation coefficient after 10*10 Fold cross Validation for each of the 19 formulas among the six classifiers tested (we removed the I and the J formulas for their high correlations with the O and T formulas). The satisfactory prediction scores for MMC (above 0.3) are in bold.

Results after the protocol of Ensemble Type method for prediction explanation (ETMPE)

After applying the protocol, the results are as follows:

* The number of definitions predicted by the set of features present in our dataset (minus the ten features used to calculate them) by the best classifier and which could be explained by our protocol is 17 out of 19 (or 19 out of 21 if we considered the I and J formulas).
* The total number of features selected is small; only 11 features account for the 17 formula predictions according to our protocol.

The raw results for each formula after the ETMPE selection are presented in Figure 5.6 and the synthesis in Table 5.2. The complete Feature importance selection is available as supplementary content in the Appendix (Figure B.1 for LASSO, Figure B.2 for SHAP, and Figure B.3 for Sensitivity Analysis)

We can see that for each formula, there are, on average, two or three features in common. We can see that 16 formulas out of 19 have in common the first features **Time in Bed** and **Time Period of Sleep**. The meaning full results are listed below :

* Concerning the two definitions without reliable explicative variables, M and S, we could assume that for these two formulas, the removed features contain almost all the information useful for their prediction. This observation is confirmed by the low scores obtained in the prediction assessed by MCC or F1 score, especially for the M formula with the lowest scores compared to the others (0.15 and 0.29, respectively). These two definitions use exclusively objective TST and subjective TST. In the case of Formula M, each of these two features is used twice for the calculation (see 2.6). It was also in our dataset the lowest prevalence with only three % of the sample, so it is possible that the relatively unbalanced class could explain these results even if we adjusted the hyperparameters according to the prevalence. For the formula S, we can apply the same observations even if the MCC score is not the worst with 0.25, nor

	PSG	ActiG	MMPI	Formulas
TIB	**11**			**A, B, C, D, E, F, N, Q, T, Z**
TPS	**16**			**A, B, C, D, E, F, K, L, N, O, P, Q, R, T, Z**
Arousal	**2**			**O,C**
MicEv	**1**			**O**
Stadechanges	**1**			**V**
Meansleeplastday		**3**		**C,O,Z**
Totalactivityscore		**1**		**L2**
assumedsleep		**1**		**C**
Hsortielit		**1**		**C**
Low5		**1**		**L2**
Hs			1	**R**

Table 5.2: Total features selected to explain the ParI among the 18 formulas. The features are grouped according to the Pre-dataset origin (PSG, AG, and MMPI). TIB=Time in Bed, TPS = Total Period of Sleep, Arousal = Total Number of Arousal, MicEV =Micro-arousal index, Stadechanges = Total number of sleep stade changes, Meansleeplastday = TST measured by AG the same night that PSG, Totalactiivity score = AG activity measured during the day, assumed sleep = total elapsed time between "Fell Asleep" and "Woke Up" times, Hsortielit= Get up time, Low5 = lowest activity during the night measured by AG, Hs = hypochondriac scale.

is the sample percentage with a prevalence of 7 %, but as they chose a very selective cut-off, as they did for the formula M, it's possible that in both case all the useful information are contained mainly in the ratio oTST/sTST chosen by the authors. But, more surprisingly, even with this similarity, these two formulas are barely correlated on our dataset with a Pearson correlation of 0.27. So for these two formulas, the low perception percentage of the objective TST seems to be the main explicative criterion.

* Concerning the 17 remaining formulas with explicative features, 13 have two features in common, TPS and TIB as main predictors, 11 have TPS as the most important feature, five have TIB as the most important feature, and 16 have at least one of these two features as the most important.
* The other features remain related to objective sleep measures, and very few are (only one) related to the psychological profile or the questionnaires, even the most specific to sleep, like ISI, ESS, or HO. There is no great influence from age, gender, or belief and attitude towards sleep. These results are quite surprising as they show that the main predictor of ParI is the relation between the time spent in bed (TIB) and the total sleep period (TPS). So, studying the relationship between these two variables is essential to define better how they are closely linked to the prediction of ParI.
* TIB refers to the total duration a person spends in bed, regardless of whether they are actively attempting to sleep or awake. It includes both the time spent asleep and the time spent awake when a person is in bed.
* TPS, on the other hand, is the time spent in bed specifically for sleeping. It includes the time spent asleep and any time spent awake during sleep, contrary to TST, which considers only the time of sleep. So here, TPS is the time from the first episode of sleep to the last episode of waking.

Although this finding is of great interest to our research questions, we can see a huge selection of FI with our ETMPE protocol, with a probable loss in the complex interaction between features. To understand the potential information loss, we calculated the total normalized weight of FI from the three explainers for each feature to see how much information has been lost.

As 82 features were selected, we will detail only the meaningful features in Table 5.7a. The cutoff for meaningful features is calculated using the random feature mean score, and SD presented in Table 5.8a. All the features below, or with a mean score in the range of the random feature score SD or with an SD overlapping the random feature SD, are removed. In the end, only 58 features left. We calculated the summed score ratio for each compared to the random feature summed score.

Figure 5.6: Mean and sd normalized score for each feature selected with our ETMPE protocol for the 18 formulas for which we could have at least one feature

We will synthesize the meaningful ones by feature's origin category and importance from the results described previously to see if we can define a "Generic" ParI patient. Indeed, from the 58 contributive features described in 5.7a, the global feature's summed score by pre-dataset described in 3.1.1 is used. So the four categories correspond to MMPI-2 T-scores (Database I), questionnaire scores (Database II), PSG features (Database III), and Actigraphic features (Database IV), and the data coming from the sleep logs added in the end (see B.6. We included the dataset origin as a variable, as ML could sometimes learn a dataset structure. The global weights are presented in Table 5.3.

Discussion

The Ensemble Type method for prediction explanation (ETMPE) protocol results revealed that 17 out of 19 (or 19 out of 21 if considering the I and J formulas) formulas could be

Features	Sum Sco	Sum SD	IF	r	Origin
Timeperiodsleep	4.233248	2.576903	14K	1	PSG
Timeinbed	2.440031	2.058738	7K	2	PSG
LatREM	2.282124	2.059576	7K	3	PSG
Avgwakemvmt	1.364706	1.527187	4K	4	ACTI
Meansleeplastday	1.011161	0.905045		5	ACTI
Totactivityscore	0.842664	1.024457	3K	6	ACTI
Arousalnumber	0.725582	0.758414		7	PSG
tempsaulitagenda	0.684860	0.608199	2K	8	LOG
MicArindex	0.404562	0.417502		9	PSG
Hcoucher	0.345849	0.347014		10	LOG
Hsortielit	0.328778	0.437622		11	LOG
M10Onset	0.287455	0.401189	1K	12	ACTI
N3tot	0.284374	0.088259		13	PSG
Max10hcount	0.282445	0.427656		14	ACTI
Lowest5hcount	0.266474	0.416496		15	ACTI
ArousalTotaltime	0.260922	0.201383		16	PSG
Sleeplatency	0.198865	0.301632		17	PSG
Relativeamp	0.184833	0.147801		18	ACTI
Actualsleeptime	0.171682	0.208601		19	ACTI
Assumedsleep	0.151588	0.201966	500	20	ACTI
actiefficacy	0.141913	0.206234		21	ACTI
19	0.121106	0.150179		22	QUEST
ISI	0.108288	0.153929		23	QUEST
25	0.095746	0.084429		24	QUEST
Intradailyvariability	0.093158	0.101075		25	ACTI
Ho	0.092388	0.106839		26	QUEST
Stadechanges	0.092311	0.115564		27	PSG
Immibilemins	0.087855	0.161363		28	ACTI
2	0.081100	0.100964		29	QUEST
N2perc	0.069948	0.046		30	PSG
17	0.057522	0.033516		31	QUEST
MDSx	0.054178	0.057735		32	MMPI
PdSx	0.044586	0.057777	100	33	MMPI
REMperc	0.041415	0.018739		34	PSG
Hy4x	0.039494	0.020301		35	MMPI
Dox	0.037313	0.057309		36	MMPI
epworth	0.036100	0.038516		37	QUEST
MaSx	0.035656	0.026347		38	MMPI
Wakebouts1mini	0.034948	0.030934		39	PSG
WRKx	0.034758	0.057143		40	MMPI
1	0.031873	0.059793		41	QUEST
Hs5Kx	0.031640	0.037146		42	MMPI
30	0.029731	0.041106		43	QUEST
staietat	0.028432	0.050624		44	QUEST
FRmoysom	0.022619	0.015391		45	PSG
N1perc	0.020044	0.026745		46	PSG
Wakebouts	0.018779	0.020619		47	ACTI
Lx	0.018393	0.024967		48	MMPI
Esx	0.015294	0.020136		49	MMPI
HRREM	0.014092	0.033490		50	PSG
Ma4x	0.012972	0.013896		51	MMPI
Agex	0.008449	0.002617	10	52	PHYSIO
Rx	0.007222	0.001938		53	MMPI
RDIback	0.004660	0.001189		54	PSG
HRwake	0.003413	0.001434		55	PSG
IPR	0.001568	0.000164		56	PSG
Fragmindex	0.001548	0.000275		57	ACTI
Dx	0.001484	0.000883		58	MMPI

(a) Summed normalize scores of the 58 significant features sorted by Impact factor scaled on the Random features score mean + SD. This table ranks features in descending order according to the sum of normalized scores after the explainer sorting process (Lasso, Shap, and SA). IF corresponds to the impact factor, the proportional factor of importance about the random feature (e.g., 1 K = a score 1000 times higher than that of the random feature). Boxes colored red correspond to features detected by our ETMPE protocol. Finally, the database origin of each feature is noted on the right.

Table 5.3: Total summed score weight of the main features selected by the three explainers by dataset origin

ORIGINE	Summed Score	Summed SD
PSG	11.134726	8.752056
ACTI	4.906261	5.749964
LOG	1.359487	1.392835
QUEST	0.682286	0.819895
PSYCH	0.334070	0.376179

predicted and explained using the selected features. This indicates that the predictive models, in combination with the explainers, were able to provide insights into the relationships between features and the formula predictions.

The complementary analysis of all the meaningful features selected by each explainer shows that some important features were removed by our EMTPE protocol but proportionally to the summed scores of FI across the three different explainers. Indeed 70% of the top 10 features were kept by our EMPTE protocol.

After, we could observe a reduction, with 30% of the biggest summed scores of FI from 11 to 20, and then less than 5% of the following with an increased dispersion. So **we could confirm that our ETMPE protocol could select the major part of the meaningful features and highlight the ones involved in specific formula prediction**, like Stade changes that are the only meaningfully detected for the V formula.

But this ETMPE protocol removed important FI like LatREM when we looked at the explainer-by-explainer comparison for LatREM; it was removed, for example, for the formula C as the main FI to explain the formula for SA and Shap but not selected by LASSO, so it was excluded from the final selection. We could see different examples of such divergence in the explainer's selection in Figure 6.2.

We wanted to understand **why this important feature, REM latency, was not selected** with a summed score almost equivalent to TIB. After analysis, this parameter's problem was that very few patients had no REM sleep during the night. As a result, during database construction, the value 999 was set to reflect the absence of available latency while at the same time reflecting reality, i.e., a much longer duration to reach REM sleep. This is the only feature with this characteristic. This example is interesting because it shows how the different explainers reacted to the choice of this variable. In the selection process, both Shap and SA chose this feature, and the cross-selection between both put this feature as the most important. In the meantime, LASSO didn't keep it, so our EMTPE protocol removed these features. An explanation could be that the LASSO penalty term encourages sparsity in the model by promoting some coefficients to exactly zero, effectively excluding corresponding features from the model. In our case, where a particular feature has a value significantly higher than the other features, it may still be chosen as important by LASSO if it strongly impacts the prediction outcome. However, the fact that the feature represents less than 1% of the dataset could potentially diminish its importance in the LASSO feature selection process.

Indeed, in LASSO, the feature selection is primarily driven by the coefficients associated with each feature. If the feature with high values is not strongly correlated with the target variable or if its influence on the predictions is not substantial relative to other features, LASSO may assign its coefficient a value close to zero, excluding it from the model. LASSO aims to balance prediction accuracy and model simplicity by selecting a subset of features that explain the data well. In contrast, sensitivity analysis and Shap with SVM focus on the impact of features on the predictions rather than the coefficients. These methods consider the overall contribution or sensitivity of each feature, irrespective of the coefficient values; in that specific case, the huge value of this feature seems to have influenced these two explainers. So, thanks to our ETMPE protocol, we could understand why a feature was not chosen, and our initial hypothesis on the triangulation of the feature selection by different algorithmic approaches is quite effective in increasing the reliability of FI selection

Features	Summed score	Summed SD
Si3	0.001080	0.000601
Pd4K	0.000998	0.000696
GF	0.000948	0.000698
Ma3	0.000741	0.000131
Sc6	0.000714	0.000178
Snoreindex	0.000691	0.000364
D4	0.000687	0.000769
PLMSindex	0.000642	0.000827
Mouvingmins	0.000575	0.000025
Pt1K	0.000527	0.000074
D3	0.000476	0.000365
LSE	0.000422	0.000250
Hy	0.000400	0.000645
randNumCol	0.000344	0.000674
activeperiods	0.000307	0.000200
APS	0.000278	0.000448
F	0.000265	0.000313
L5Onset	0.000209	0.000135
Sc2	0.000204	0.000430
DBAStotal	0.000027	0.000057
RRN2	0.000000	0.000000
VRIN	0.000000	0.000000
Sexe	0.000000	0.000000
TRIN	0.000000	0.000000
staitrait	0.000000	0.000000

(a) Summed score of each feature selected by at least one of the three explainers in the random column range

in that case.

The findings from this study contribute to understanding the different subgroups corresponding to each formula and highlight the importance of accurate predictive models and explainers in gaining insights from complex datasets.

The last aspect to be discussed is the surprising importance of random features, selected in the top 10 features by the three explainers 10% of the FI selection process. It was in 60% by LASSO and 20% by the two others each. The three explainers never chose the Random feature simultaneously, and only LASSO could select several times the Random Features for a given formula. This finding raises a real question about the reliability of the FI explanation with only one explainer. We can see the random feature summed score in Table 5.8a.

Limitations

There are two major limitations in our experiment:

1. The absence of a mathematical demonstration of our empirical approach is a major limitation. Even if our choice is empirically logical and based on the theoretical background, our approach is more practical, based on the assumption that three algorithms built on different mathematical foundations will be able to overcome each other's shortcomings. Although it seems to work in our case and provide new information on the subject of interest, we need to mathematically formalize this approach before a possible generalization to other datasets (at least to optimize the number of explainers and the complementary between each other).
2. The 2nd limitation is the lack of systematization. We haven't tested every possible combination of explainers, choosing to associate three, even without exploring the mathematical underpinnings.

5.3.4 Conclusion

This second experiment, based on an empirical feature selection protocol called the Ensemble type method for prediction explanation, enabled us to select the important features explaining the prediction of each formula defining Paradoxical Insomnia on our data set. We could show that most of the features selected on our core dataset corresponded to two variables from the Polysomnography report, Time In Bed and Time Period of Sleep, which were the most important features for 16 of the 19 formulas. However, since we cannot mathematically assert that our empirical protocol is reliable, our next experiment is to compare these results with classical inferential statistics by T-test on most of the variables present in the dataset.

5.4 ParI, is there a possible harmonization across formula definitions?

To go further in the explanation, we built two classes on the dataset: the subjects never categorized ParI, whatever the formula, and those categorized as ParI at least once.

We will see in the first subsection a systematic analysis of the value of the features between the two categories to get some insight into the no Paradoxical Insomniac patients. We will also see the main feature correlations with Paradoxical Insomnia.

The second subsection will cross the ML and classical statistics results to describe the Paradoxical Insomniac profile.

As there is no doubt left about the most Important Features, TIB and TPS, to understand the relation between these two features better, we created a new feature, TIME, which is the simple difference between them. The idea was to see the impact of time awake before the sleep period, which must be correlated to sleep latency.

TIME = TIB - TPS

5.4.1 Systematic feature impact analysis between subject classified ParI positive or negative on our dataset

1. T-tests with post-hoc Bonferroni [42] correction on all the features in the dataset. We included the new feature TIME described above. We listed all the significant and non-significant results of interest in Table 5.4.
   .
2. Pearson correlation analysis to see the most correlated features with ParI.

T-test analysis

The t-test for the main features involved between the two groups is presented in Table 5.4

From the Table 5.4 we could make some conclusions :

1. The main features detected by EMTPE on each formula are highly discriminative when applied to the global distinction between subjects classified ParI or Not, whatever the formulas. Then, there is a clear link between these features and the global concept of ParI.
2. Objective TST on one night is the most discriminative feature between all the subjects classified ParI or Not, whatever the formulas. This feature shows that the ParI subject slept much more on one night than the other CID. This feature is used by some formulas as the only feature to explain the ParI concept with the assumption that restorative sleep is linked to the number of minutes spent in sleep, and if this sleep is not more fragmented, the hypothesis is that ParI subjects don't perceive it because of a sort of "active state" concomitant to sleep. Our results confirmed that microsleep fragmentation detected by the MicArousal index is no different in our sample, but when we look at the features linked to cumulative sleep fragmentation (not per hour), we can see that the number of episodes of wake, less or above one minute long are much more frequent in the ParI population, as the number of stages changes.
3. There is no difference between the two main scales, ISI and ESS.
4. There is a clear difference in chronotype with the ParI subject toward the morning type more than the other CID. This tendency is confirmed by the Actigraph sensor (AG), showing that the ParI subjects go to bed 35 minutes earlier than the others (see Bedtime-H-acti in the **Circadian profile** cell in Table 5.4).
5. There is no big psychological difference. Still, some are significant, like less tendency toward addiction acknowledgment and more inhibition of aggressivity and anger,
6. So TST on night PSG has the main impact on differentiating the ParI and NoParI subjects. Still, we look at the mean of TST on seven nights, symbolized by the Actual

Features	ParI+	ParI-	ttest-ind	p-value	sign
Percentage of the sample	81.5	18.5			
gender and age					
Women(%)	69	59	-1.50	0.13	NS
Age(y.o.)	45.8±12	47.7±13	-1.00	0.31	NS
PSG one night					
Timeperiodsleep(TPS in min)	432±73	352±77	7.41	7.10e-11	***
Totalsleeptime(TST in min)	370±68	290±70	8.04	3.45e-12	***
Timeinbed(TIB in min)	474±71	437±65	3.95	0.0001	***
LatenceN1(min)	25.4± 21.9	63.8±47.3	6.26	3.03e-08	***
LatenceN2(min)	29.4± 23.5	69.2±51.5	5.92	1.18e-07	***
LatenceREM(min)	106.7± 58.5	108.5±78.9	-0.16	0.86	NS
TIME(min)	42.3±37.5	84.8±56.8	5.61	3.29e-07	***
WASO(min)	61.5±52	60.8±45	-0.11	0.9	NS
PSGWakebouts	18.5±9.8	13.2±8.7	4.22	5.37e-05	***
PSGWakbouts1min	10.3±5.8	7.8±4.2	3.95	0.0001	***
MicArindex(per hour)	20.5±10.7	19.7±9.6	0.59	0.55	NS
Arousalnumber	126±67	94±52	4.08	8.37e-05	***
Stadechanges	84±32	63±31	4.68	9.43e-06	***
N1(%)	4.2±3	4.3±4	0.39	0.69	NS
N2(%)	52±12	48±12	2.13	0.03	*
N3(%)	23.6±10.7	28.3±11.7	-2.86	0.005	**
REM(%)	19.4±6.7	18.7±8.4	0.60	0.54	NS
ACTI seven nights					
Actualsleepime(min)	414±63	424±77	-0.96	0.33	NS
Avg_wake_mvmt	265±348	196±58	3.08	0.002	**
Meansleeplastday(min)	374±77	335±59	4.35	3.01e-05	***
Fragmindex	33±12	33±10	0.53	0.59	NS
ActiWakebouts	26.5±8.7	25.2±12.6	0.78	0.43	NS
Interdailystability	0.51±0.12	0.47±0.11	2.65	0.009	**
Lowest5hcount	1102±916	1151±934	0.37	0.71	NS
Circadian profile					
Ho	52.5±10.5	48±10	3.09	0.0025	**
Bedtime-H-acti	23:34±75	0:09±66	-3.61	0.0004	***
Psychological profile					
Addiction Acknowledgement	51±10	54±10	2.19	0.03	*
Masc-Fem	51±10	54±10	-1.99	0.048	*
Inhib of Aggression	53±11	50±11	2.03	0.04	*
Anger	50±11	53±11	-2.07	0.04	*
Rscale	56±10	53±10	2.21	0.029	*
Sleep Questionnaire profile					
ISI	20±4.2	18.4±6.35	1.89	0.06	NS
Epworth	7.3±5.2	7.6±4.6	-0.40	0.68	NS

Table 5.4: Main features Mean and comparison between ParI positive or negative with t-test and post-hoc Bonferroni correction.

Sleep time feature, and we can see no difference between the two groups. Indeed, The ParI subjects slept more or less the same amount of time, and the other CIDs slept much more than the single PSG night. But on the last night of AG, "meansleeplastday", the difference is very significant. So, it means that last night's recording with the PSG device changed the TST for CID without ParI much compared to CID with ParI. This observation leads to reconsidering the choice of objective TST by a single night PSG to discriminate ParI patients in a sample. Our observation is strengthened by the fact that we used Ambulatory PSG, which is theoretically less prone to sleep disturbance

than Laboratory PSG, especially in Insomniac samples, which are usually very sensitive to new environments.

7. Our analysis confirms the main hypothesis that ML algorithms could bring new knowledge compared to Classical inference for some features. One of the main discoveries from this difference is the selection of the Micro-arousal index as one of the top 10 features detected by our EMPTE protocol to predict ParI, but with no significant difference in the t-test comparison. For example, we could observe less obvious differences for the ISI scale still in the quiet, meaningful range of FI detected but with no significant difference with the t-test.
8. Our analysis also shows the overlap between ML and Inferential statistics, especially for the most important features like TPS, TIB, or arousal numbers, but with different magnitudes showing that the different approaches are not detecting the same relation between features. We also analyzed why Latency to REM sleep was discarded by our protocol, and going back to the dataset; we saw that two samples had a very high value left in the dataset to traduce the fact that they didn't have any REM sleep, so no calculation of the latency was possible. Interestingly, this anomaly was detected by the ETMPE that discarded this feature even if the cumulative score was high. Only LASSO couldn't discard it.

Correlation impact analysis

As mentioned, we implemented TIME in our dataset to see the relationship with being ParI positive. We also used all the features, including the ones necessary for formula calculation—only the correlations above 0.3 or below -0.3 were kept. The results are presented in Correlation Matrix 5.9. We can see that this featured biomarker, equivalent to global sleep latency but with the additional time between going to bed and turning off the light, is more anti-correlated than WASO and slightly more than SL to N1 or N2. So it means that the most correlated variable, not ParI, is this time between going to bed and the first episode of sleep, whatever the sleep stages, N1 or N2. We can see that TST and SE on one night have better scores than TPS for the positive correlation.

We also calculated the correlations of TIME with the prevalence of the 18 explainable formulas. We can see in Figure 5.10 the high correlation between TIME in minutes and the percentage of prevalence on our dataset. Even if the correlation is not causality, we could hypothesize that the different definitions of ParI (at least the 18 studies here) could be a continuum of the same phenomenon, leading to different groups depending on the chosen threshold.

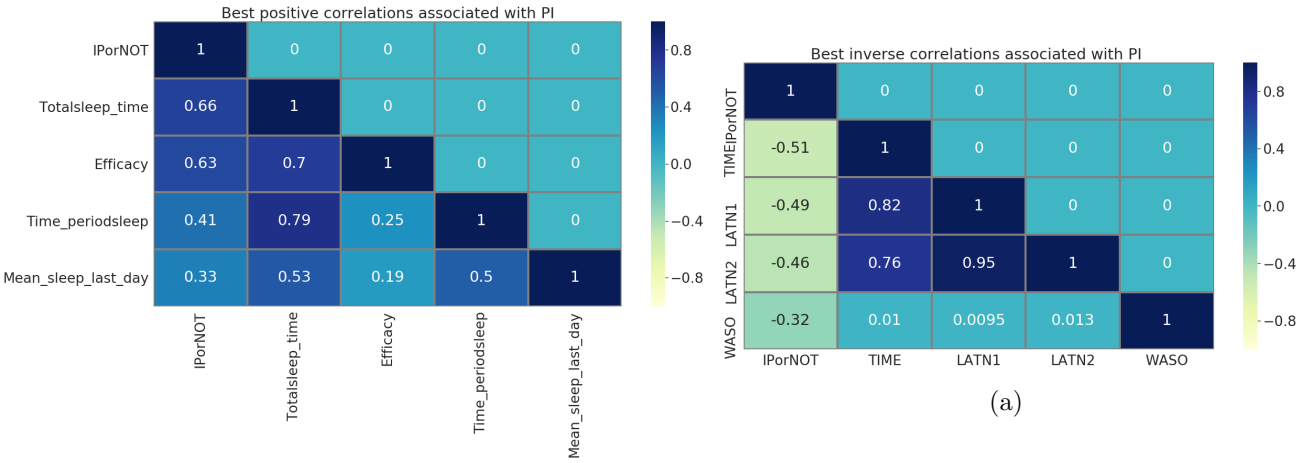


Figure 5.9: Best positive and negative correlation to ParI positive on our sample.

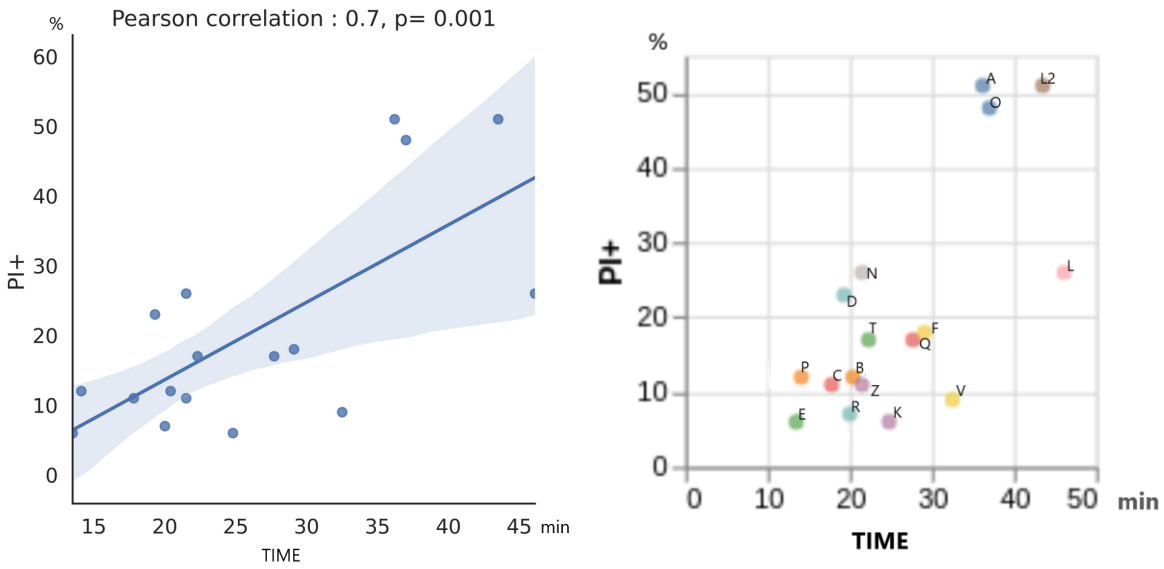


Figure 5.10: Correlation of the feature TIME with the percentage of prevalence of each of the 18 formulas.

5.4.2 Mixing ML and inferential statistics to describe the generic ParI concept

We will analyze according to the origin of the features. In this analysis, we will call "Generic Paradoxical patient" the mean profile corresponding to the formula overlap. This analysis aims to see the big picture of a Paradoxical Patient looks like using the different meaningful features.

PSG features

From these results, we can see that PSG features have the strongest contributions. It is quite logical since the formulas all use at least a PSG feature in their calculations. The main results are listed below:

1. From TIME, SL(N1, N2), and HO, we can deduce that the Generic ParI goes to bed earlier, switches off the light and falls asleep far more quickly than the other CID.
2. From the results presented in 5.10 and 5.9 we can observe a correlation between the duration of TIME in minutes and the probability for a subject to be diagnosed as ParI.
3. TST and SE are the most positively correlated to the positive ParI diagnosis, so the Generic ParI sleeps in total much more than the other chronic insomniac patients and more efficiently.
4. But the Generic ParI patient sleeps with significantly more wake bouts, especially above one-minute duration.
5. The very light sleep, N1, is the same percentage, but as patients sleep more, the duration of N1 sleep becomes higher for the Generic ParI.
6. The Generic ParI patient has more stage N2, arousal number, and stage changes.

From all these, we can retain that the Generic ParI sleep more and in total, fall asleep more quickly, BUT with far more arousals and more light sleep but without an increase in the microsleep fragmentation detected by the Micrarousal index.

All these PSG findings suggest that the Generic ParI has not a sleep problem but a wake problem and that probably the sum of all these wake-bouts gives him the sensation of being awake all night. This fact is to be put in perspective because one can't memorize his sleep but can memorize wake-bouts, especially when they are over one minute long. Furthermore, spending more time in light NREM sleep can also give a feeling of drowsiness without real sleep, especially in N1 sleep.

Actigraphic features

1. From actualsleeptime corresponding to the average TST calculated by the actigraphy, we can see no difference between the two groups. This is a huge surprise because we could observe the significant difference only on the last night corresponding to the PSG night, i.e., the night serving to calculate the ParI insomnia.
   This finding is extremely important because it calls into question calculating the perception of sleep over a single night. So, the Generic parI sleeps more or less simultaneously as the No parI patient sleeps much less during the PSG recording night in our sample.
2. From the feature Interdailystability, we could infer that the Generic ParI have a more rest/activity cycle than the No ParI patients.

All these actigraphic findings suggest that the Generic ParI has a more stable activity/rest cycle and sleep duration during a whole week. But the main finding here is the opposite sleep duration during seven days and the last night corresponding to the PSG recording for the subjects categorized non-ParI. These findings question in a frontal way the relevance of the definitions used to date, based on the last night of polysomnographic recording, for distinguishing between ParI and non-ParI subjects.

Questionnaires features

1. From the circadian features, we can see that the Generic ParI is more of an early bird and goes to bed earlier than the other Chronic Insomnia Disorders. .
2. From the Specific sleep questionnaires, there is no difference in ISI and ESS.

Psychological features

There are no significant differences between the two groups but only some tendencies. So, compared to the no-ParI, the Generic ParI acknowledges less addiction tendency, less gender affirmation, more aggressivity inhibition, less expressed anger and more emotional repressed, and more conformism.

5.4.3 Discussion

In this section, we intend to harmonize the different ParI formula Definitions. By analyzing the dataset using various formulas, we could identify common features that are highly discriminative for distinguishing ParI patients from non-ParI patients. This suggests that there may be a shared underlying concept of ParI, regardless of the specific formula used. We could show that the TIME features are highly correlated with the prevalence of each formula on our dataset. Even if a correlation is not causality, we could infer that this TIME feature, as a reflection of the sleep latency plus the time in bed, is a sign that the faster insomniac subjects fall asleep, the longer they stay in bed, the more they will have multiple long wake bouts, and the end, the less they would perceive sleep.

Of great interest in these results is the significance of total sleep time (TST) on one night as the most discriminative feature between ParI and noParI and the inversion of this value on a whole week.

We could also find some interesting findings concerning the psychological profile, even if only tendencies; the Generic Paradoxical patient seems less introverted and globally repressed his anger and emotional state.

Finally, concerning a possible explanation of Paradoxical Insomnia across the different studies, our findings go toward a specific sleep profile shared by all the definitions, i.e., a longer Total Period of sleep than the other insomniacs, a shorter sleep latency, but also more wakes episodes during the night and especially above one minute. This finding is important since we intentionally introduced this feature with a hidden hypothesis that it could be linked to memorizing wake episodes.

Indeed, it was shown in an old study now, [214], that the closer a stimulus (in this case, words) was presented to sleep and, in particular, in the three minutes preceding it, the more it was forgotten in the case of implicit memory, i.e., without instruction. This case figure could be the closest to an awakening in an insomniac subject. The results of their experiment are presented in Figure 5.11. This would mean that physiologically, the longer an awakening, the more likely it is to leave a trace if it is followed by a sleep of more than 10 minutes. Therefore, brief awakenings of less than one minute are unlikely to leave any trace. This aspect alone could explain why a subject awakening several times during the night, with long episodes but in between numerous episodes of sleep (that we can't memorize), could have a more important memory of this night, in terms of awakenings and thus have the impression of not having slept [214].

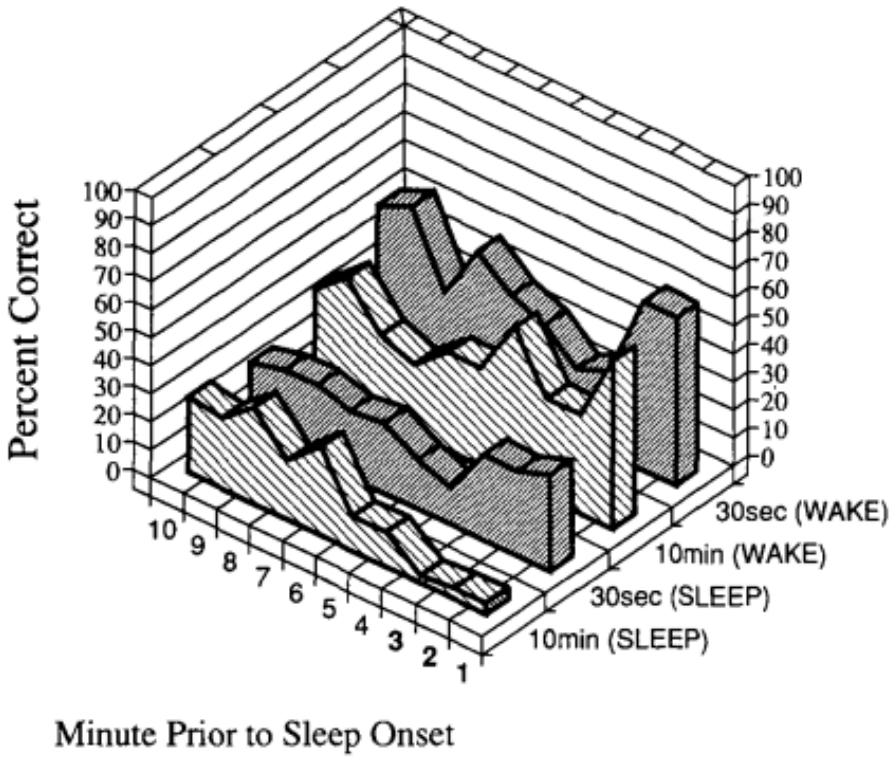


Figure 5.11: Free recall task performance: This figure shows that the closest a stimulus is to the sleep onset, the less the recall, or that you need time, in general, more than three minutes to recall something from [214]

5.4.4 Limitations

As always in this thesis, the sample size is relatively modest and could avoid generalization, but we could show that our dataset was representative of the target population. Even if we provide a selective empirical model for feature importance, we can't prove that our protocol has selected the best features. That's why we made the statistical comparisons with t-tests and correlation analysis. But even in that case, t-test results and correlations are given without providing effect sizes, confidence intervals, or statistical power information. These additional measures would have strengthened our results. But, the fact that two different approaches are going in the same direction is already a good result, even if replication and further statistical analysis are needed.

5.4.5 Conclusions

Our two main conclusions (under the abovementioned limitations) are:

* that Paradoxical Insomniacs shared a longer sleep period and more wake episodes than the other insomniacs. Time in Bed and Time period of Sleep are the main predictors of Paradoxical Insomnia, and the difference between these two features is highly correlated with the percentage of subjects categorized as Paradoxical Insomniacs in our sample. We then support one of the three hypotheses proposed by [82] to explain the sleep misperception, i.e., the presence of brief awakenings. A deficit in time estimation ability was not supported. At this step of our work, we don't have any argument to support the two others: misperception of sleep as wake and worry,
* that objective Total Sleep Time, the main feature involved in the definitions of Paradoxical Insomnia and the biggest difference in statistical terms between Paradoxical and Not Paradoxical Insomniacs, is no longer statistically different if we take seven days of analysis instead of a single night. This finding opens the path to a new definition based on an entire week instead of only one night. This is the last section of the experiment.

5.5 Proposal for a new definition of Paradoxical Insomnia including seven nights sleep analysis

Because of the previous results and observations on the difference between the objective TST on seven days versus one night, it appeared necessary to change at least the method to define ParI, if not the current features. Indeed, as we have shown, the TST between one night and one week is different for the Non-ParI insomniacs when it stayed equivalent for PArI, with an inversion in terms of comparison, from much less TST to even more TST than ParI. This observation, on the same dataset, on such a fundamental feature involved in the ParI diagnosis is very surprising. The validity of these results is supported by comparing the TST for PSG and AG on the same night that completely coincides with the ParI subgroup (370±68 for PSG versus 374±77 for ActiG). The conclusion is that the predominant discrimination by objective TST measured by PSG over one or two nights in the sleep laboratory does not seem to represent the reality of sleep at home over one week without PSG.

As we have concomitantly one week of actigraphic measures and one polysomnographic evaluation, we applied a corrective factor based on the ratio between psgTST and actiPSG to increase the reliability of the objective TST calculated by actigraphic. Thus, the estimated seven-day TST is labeled woTST (for Week objective TST). The formula definition is described in the formula 5.3

$$woTST = \frac{oTST}{actlastnightTST} \times actimeanTST \tag{5.2}$$

We then took the same formula L2 described in 2.6 but applied for seven days. We took it because these formulas involved only TST, and we expected a major change in the subject detected by our formula.

$$IP = woTST - wsTST > 60min \tag{5.3}$$

To evaluate the impact of different cut-offs on the correlation with the other formulas, we created 12 derivatives features listed in Table 5.5 to cover an 8-hour night in increments of 30 minutes.

	Formula
IP	woTST-wsTST >60 min
IP2	woTST-wsTST >90 min
IP3	woTST-wsTST >120 min
IP4	woTST-wsTST >150 min
IP5	woTST-wsTST >180min
IP6	woTST-wsTST >210 min
IP7	woTST-wsTST >240 min
IP8	woTST-wsTST >270 min
IP9	woTST-wsTST >300 min
IP10	woTST-wsTST >330 min
IP11	woTST-wsTST >360 min
IP12	woTST-wsTST >390 min
IP13	woTST-wsTST >420 min

Table 5.5: ParI formulas scaled on a 30-minute basis. woTST = objective Total Sleep Time on seven days, wsTST = subjective Total Sleep Time on seven days

In Figure 5.12, we can see the correlations with the other formulas, except for the formula S; almost no correlation exists between our formula and the 19 others. The prevalence in our dataset is 43% for IP.

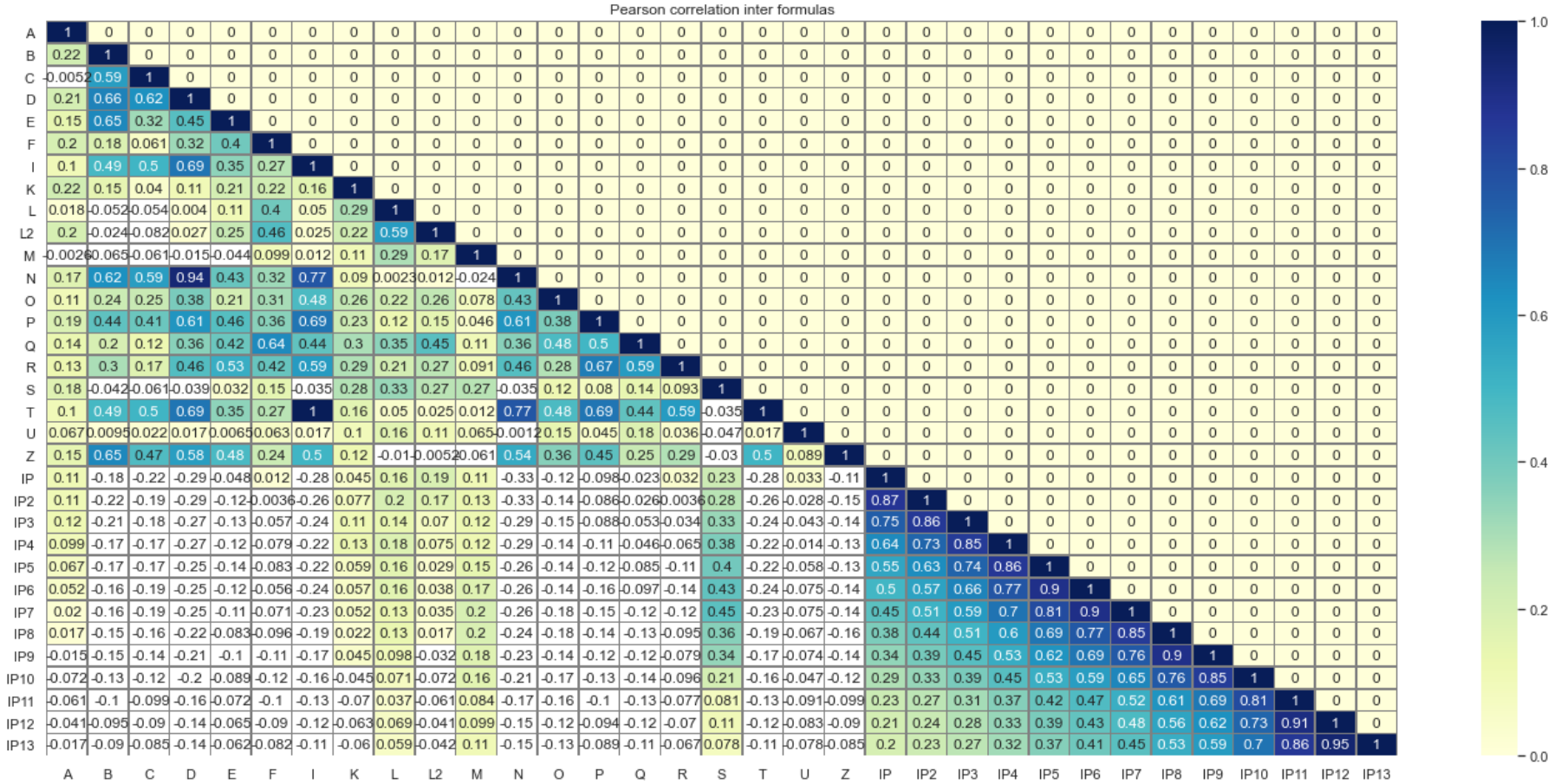


Figure 5.12: Correlations between the new formula and its derivatives to the current formulas to define Paradoxical Insomnia.

Chapter conclusion on ParI understanding

This work on the explicability of classifiers using the Paradoxical Insomnia model has also led to a better understanding of this clinical entity. Indeed, on the 19 formulas we reproduced on our database and predicted from 200 variables, including objective data such as polysomnography and psychological data, the main explanatory variables found were the high time spent in bed. Indeed, The features Time In Bed TIB and Total Sleep Time (TST) were the most explicative. Surprisingly, there was no significant difference in the microarousal index between subjects classified PArI or not - (20.5 vs. 19.7), which means that the poor perception of sleep is not linked to the direct impact of microarousal. On the other hand, if we take the cumulative arousals over the night, they are significantly more important in the Paradoxical Insomnia group. On the other hand, the percentage of stage N1 is not different between the two groups either. Thus, Paradoxical Insomnia appears to correlate with a higher absolute number of awakenings. In terms of psychological profiles, four scales of the MMPI 2 seemed to be significantly different between the two groups, but only in a modest way, namely the masculinity-femininity scale, the aggression inhibition scale, the anger scale, and the Repression scale, but the scores remained within the norm on average, so it isn't easy to draw any conclusions. Concerning the scales classically used in sleep medicine, the ISI and Epworth scores were not significantly different. On the other hand, the Horne and Hosberg Circadian typology scale was significantly larger in the ParI group, which theoretically translates into a tendency to go to bed earlier. On the other hand, all the formulas are the major determinant of the prediction. Thus, our work allows us to harmonize the various published works on Paradoxical Insomnia. Indeed, objective sleep data does not explain the different formulas intended to represent the perception of sleep. They all spend too much time in bed, which seems to be the main determinant of Paradoxical Insomnia. Insofar as the arousability indexes are identical, the recovery sleep times are similar, but only the number of cumulative awakenings is statistically different; one could wonder if Paradoxical Insomnia is not simply the reflection of a threshold of awakenings or micro-awakenings from which the insomniac has a perception of having woken up too much and thus gives him this sensation of not sleeping, in the end not so paradoxical since it corresponds well to an experience of more important cumulative awakenings without cutting down the recovery.

The other finding was the difference in terms of TST between one night and 7 seven days, especially in the non-ParI patients. We took an existing formula, the L2 formula, using only the difference between subjective and objective sleep perception, but on one single night. We calculated the formula on 7 seven nights and compared it to the other formula. After this process, the correlation with L2 was only 0.19. So, we conclude that our formula detects something different than just sleep perception in one night. To see the impact of our new formula, we implemented it in dataset 6 for the last experiment, predicting the treatment outcomes in CID.

Chapter 6

Explaining Therapeutic Issues: A Machine-Learning Approach

Chapter Highlights

We use similar methods (machine learning classifiers) as the previous chapter, but here, we insist upon a certain level of accuracy before explainability analysis.

1. **The first hypothesis** tested in this experiment is the predictability of treatment outcome in a binary problem (label 1 for a positive outcome, 0 for a negative outcome). We will use the six classifiers described in the previous chapter (RF, SVM, DT, XGB, LR and NB) 5, but to increase our probability, we add a neural network classifier (MLP) (see B.1.2), an Adaboost classifier(see B.1.2) and a KNN classifier (see 2.10.1). As our primary outcome is the response to treatment prediction, we will use classical metrics such as AUROC, Classification Accuracy, Precision, Recall, and F1 score (see B.1.2 for definitions). Based on the usual literature on prediction accuracy, we will consider our hypothesis validated if the main metrics are > 0.8, whatever the model used.
2. **The second hypothesis** is dependent on the primary hypothesis validation, as we want to explain the prediction; in this case, it only makes sense if the accuracy of the prediction is validated. So, suppose the classification accuracy and AUROC are above 0.8. In that case, we are postulating that we could explain the treatment outcome prediction, at least partially, by using the same Feature Importance detection described in the previous chapter.
3. **The dataset used** is dataset five with a sample size of 423 CID with the pharmacologic treatment taken at the time of the test and clinical evaluation at least six months later for the outcome categorization.

Key terms and concepts

Acronym/term	Definition	Ref.
CBT-I	Cognitive and Behavioral Therapy for Insomnia	p. 169 (B.1.1)
MLP	Multilayer Perceptron	p. 174 (B.1.2)
CA	Classification Accuracy	p. 171 (B.1.2)
XGB	Extreme Gradient Boosting	p. 176 (B.1.2)
AdaBoost	Adaptive Boosting	p. 171 (B.1.2)
ROC	Receiver Operating Characteristic	p. 171 (B.1.2)

6.1 Hypothesis one: Predicting treatment outcome in CID using ML classifiers

6.1.1 Hypothesis and background

The interest in predicting treatment outcomes in CID could have several impacts. Among others, perhaps the most important could be a shift toward personalized treatment. Predicting who will respond to which treatments could allow more comprehension of the patient, which would likely improve patient outcomes in any case. For instance, cognitive-behavioral therapy for Insomnia (CBT-I) may work well for some individuals, while others might benefit more from pharmacotherapy or alternative treatments. Predictive models could help determine the most effective treatment approach for each individual. Another big interest in predicting treatment outcomes for Insomnia, as for that last mentioned point on CBT-I, is the need to make the most of the limited time and caregivers available at a time when psychiatry and healthcare, in general, are suffering from a lack of resources. Predicting treatment response could help optimize resources by reducing the time and cost associated with trial-and-error treatment approaches. Instead of trying multiple treatments to see what works, predictions could guide the initial treatment choice, giving the patient more confidence and compliance. If patients can see the results of treatment predictions showing a high likelihood of success, they may be more motivated to adhere to treatment plans. This could significantly improve treatment outcomes and be therapeutic if we could reassure the patient. Indeed, the stress associated with relapse or loss of control of the disease could be an aggravating or perpetuating factor. Undermining the impact of the stress associated with sleep dissatisfaction is, per se, a potential risk factor for psychiatric and general health conditions disorders. Besides the clinical aspect, predicting treatment response could provide insight into the underlying mechanisms of Insomnia and its treatment, potentially leading to the development of new, more effective treatments, or at least to use the most efficient with the best chance of success.

So, we hypothesize that ML algorithms could achieve this difficult task of predicting the treatment outcome in our dataset. As for Paradoxical Insomnia, the great difficulty here is defining a good or a bad treatment outcome. We will discuss that point later, but as mentioned before, sleep satisfaction is the main factor here. Although no equivalent hypothesis was tested in a publication at the beginning of the thesis, studies have since been published in 2021 and 2022 using MRI design (see Table 6.1). The design of these studies has nothing to do with our own, neither in terms of the number of subjects nor the data used; anyway, these studies using ML on insomniacs reinforce our hypothesis insofar as they manage to find subgroups of subjects more likely to respond to treatment presenting functional connectivity disturbances [129] and different spatial covariance pattern of blood oxygen level-dependent (BOLD) in PsyI compared to healthy subjects [120]. In both cases, these findings support our hypothesis first to predict the treatment outcomes in our samples and try to explain the prediction a second time.

Year	N of CI	Data	ML Model	Ref
2022	51	T1 MRI, rsfMRI, DWI	HoTS	[129]
2021	19	fMRI (BOLD)	SVM	[120]

Table 6.1: Studies on Predicting Treatment Outcome in Chronic Insomnia with ML. HoTS: Hollow Tree Super, rsfMRI:resting-stage functional magnetic resonance imaging, DWI: diffusion-weighted imaging

6.1.2 Methodology and tools to test the hypothesis

Dataset

The dataset used is five, described in 3.1.1, with 423 CID subjects and almost 200 features. All the features described in Section3.3 are included to predict the treatment outcome, ex-

cept the actigraphic features are unavailable for all patients. But even if the whole features were lost for 88 subjects, we could retrieve at least the mean sleep duration for the last night and the mean for the week in our database necessary to calculate the new ParI definition described in Chapter5. As already mentioned, the complete list of features used can be found in Section 3.3 and in Appendix B.3 in the Tables B.4, B.3, and B.2. We are using 180 features in total, considering the objective sleep measurements, the psychological aspects, the socio-economic aspect, the gender, the sex, the comorbidities, the different treatments, and the chronotype. The few missing data, only numerical, were imputed via the median of the training data. We used feature analysis to process data and select the most important predictors. We preserve an optimal ratio between features and labels in the dataset to avoid overfitting. We mostly used correlation-based feature selection filtering, especially in the features issued from polysomnographic recording to remove repeated features. We used dimensionality reduction algorithms like PCA (linear) and T-SNE (Nonlinear) for visualization, intending to have the smallest number of independent features in the training dataset to create our model to have a more stable and robust model to minimize the risks of overfitting the data. At the end of the process, 166 features were kept.

Class determination: how did we define treatment outcome ?

The primary outcome was an improvement or no improvement after evaluation and treatment in a prospective way six months to 24 months after the initial evaluation. The main goal here was to predict the profile of patients that the preconized treatment could efficiently treat according to the sleep European Guidelines [174].

The definition of what should be considered a successful outcome in treating Insomnia is not an easy task. The first paper on the subject appeared in 2003 [145]—the recommendations of this seminal paper listed candidates for assessing outcomes in insomnia treatment studies. The conclusion for the practice was that treatment outcomes should be evaluated with daily sleep diaries and selected self-report questionnaires targeting sleep/insomnia symptoms, psychological and fatigue symptoms, and more global measures of treatment satisfaction/acceptability. Several assessment instruments were proposed in this first attempt to standardize the measurement of insomnia treatment outcomes. From this paper, the recommendations were using different tools and symptoms to assess in the interview like:

.

* A sleep diary to assess the sleep/wake parameters,
* assessing the insomnia symptoms by the Insomnia Severity Index questionnaire,
* assessing daytime functioning by interview
* assessing the psychological symptoms by Beck Depression Inventory 2, State-trait anxiety Inventory, or interview.

The evaluation was made at least six months after the treatment. Because there is still no established clinical significance to evaluate the outcome [137], we focused on the following categories:

For this study, we chose the following criteria :

* The evaluation was made at least six months after the treatment
* Principal criteria: At least one clinical improvement on the three main criteria used to define Insomnia according to ICSD 3 assessed by clinical interview and ISI improvement (at least 3 points) with Sleep-Diary evaluation and at least one of the minor criteria
* Minor criteria:2) satisfaction/acceptability of the treatment or 3)improvement in psychological and fatigue symptoms assessed by interview.

Machine Learning Approach

As in the previous experiment described in 5, we used Supervised learning algorithms dealing with binary classification.

From the previous experiment, we kept the same six classifiers :

1. RF: Random Forest classifier (see 2.10.1),
2. SVM: Support Vector Machine (see 2.10.1),
3. DT: Decision Tree classifier (see B.1.2),
4. XGB: XGBoost, which stands for eXtreme Gradient Boosting (see B.1.2),
5. LR: Logistic Regression (see 2.10.1),
6. NB: Naive Bayes classifier.

To increase the chance of success in predicting the accuracy of the treatment outcome, we added three other classifiers (the definitions for each model are available in the Appendix in B.1.2):

1. Adaboost: Adaptive Boosting (see B.1.2)
2. A Neural network classifier based on Multilayer Perceptron (see B.1.2)
3. KNN: K-Nearest Neighbors classifier (see 2.10.1)

For each classifier, we tuned the hyperparameters to increase the chance of success.

The main parameters used for each classifier are listed below and were chosen after hyper-parameters tuning using a grid search method (see B.1.2):

* The DT is based on the Gini quality measure,
* The NB classifier assumes a Gaussian distribution
* The SVM was defined with the radial bias function kernel
* The RF classifier used 100 trees.

We developed a prediction model using 70% of the dataset to train and 30% to test the model internally. This process was repeated ten times. We used the mean AUROC and CA to determine which model performed best, which was then tested with the testing dataset.

So, in this first experimental part, we compared these nine algorithms using scikit-learn library [162]

6.1.3 Results

The data set is relatively well balanced with 188 output = 1 (44.4%) and 235 output = 0 (55.6%). This first result shows a relatively low success rate in positive treatment outcomes. Table 6.2 shows results after 10-fold cross-validation;

Table 6.2: Performance metrics of various models

Model	AUC	CA	F1	Precision	Recall
RandomForest	0.857	0.827	0.822	0.798	0.861
XGBoost	0.851	0.814	0.810	0.777	0.862
SVM	0.848	0.801	0.802	0.801	0.818
NeuralNetwork	0.849	0.787	0.763	0.755	0.771
Tree	0.794	0.787	0.783	0.791	0.783
LogisticRegression	0.836	0.768	0.746	0.727	0.766
AdaBoost	0.723	0.723	0.698	0.678	0.718
NaiveBayes	0.694	0.624	0.622	0.562	0.697
kNN	0.568	0.574	0.508	0.522	0.495

We could see that four models scored above 0.8 for the three metrics targeted (ROC-AUC, CA, and F1). An ANOVA test was conducted among the four most performant classifiers on F1 metric and showed statistical differences (F-Statistic: 147.67, P-Value: 7.16e-12). We then used Tukey's Honestly Significant Difference (HSD) test as a post-hoc analysis to compare the different means. The results are presented in Table 6.3.

We generated the ROC curves for RF and XGBoost in Figure 6.1 to have a graphical representation of the prediction.

Group 1	Group 2	Mean Diff	P-Adj	Lower	Upper	Reject
Neural Network	Random Forest	0.1542	0.0091	0.0355	0.2728	True
Neural Network	SVM	-0.5651	0.001	-0.6838	-0.4465	True
Neural Network	XGBoost	0.217	0.001	0.0984	0.3357	True
Random Forest	SVM	-0.7193	0.001	-0.8379	-0.6007	True
Random Forest	XGBoost	0.0628	0.4528	-0.0558	0.1815	False
SVM	XGBoost	0.7821	0.001	0.6635	0.9008	True

Table 6.3: Pairwise Comparisons of means for RF, XGBoost, NeuralNetwork and SVM. We can see that only RF and XGB, the two best classifiers, are not statistically different

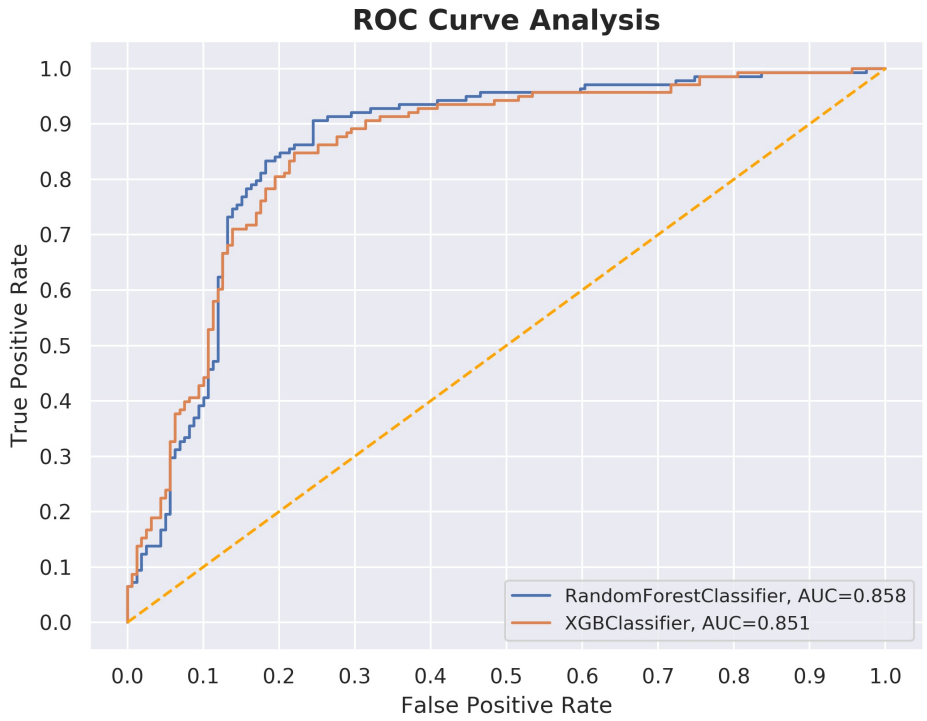


Figure 6.1: ROC Curves for the two models of classification algorithms tested on the positive outcome predictions from 10 Folds Cross-Validation

6.1.4 Discussion

The results from this study reinforce the potential role of ML algorithms in predicting treatment outcomes for chronic Insomnia (CID). The results validated our hypothesis that ML algorithms could predict treatment outcomes using the available dataset. The RandomForest (RF) and XGBoost (XGB) models, which are ensemble learning algorithms, emerged as the top performers in this task (See Table .

The selection of an appropriate algorithm was a critical aspect of this study. Machine learning algorithms function differently based on their inherent mathematical and statistical principles. Some may be more sensitive to outliers or skewed distributions, while others may be better equipped to handle imbalanced data. In our case, ensemble models proved to be the most robust and consistent performers. Ensemble methods, like RF and XGB, operate by constructing many decision trees during training and outputting the class that is the mode or mean prediction of the individual trees. This 'wisdom of crowd' effect allows them to make more accurate predictions and avoid overfitting.

Another factor that may have contributed to the effectiveness of the RF and XGB models is their ability to handle high-dimensional data, a key characteristic of our dataset. A total of 166 features were considered in our models, presenting a challenge of high dimensionality.

RF and XGB algorithms are well-suited for this, as they can effectively handle many input variables without variable deletion, providing a clear view of the critical variables.

However, the size of the dataset can also influence the performance of the ML algorithms. With 423 CID subjects, our study had a moderate-sized dataset. It is known that ML algorithms, particularly complex ones, perform better when trained with larger datasets because they can learn more complex patterns without overfitting. While our dataset wasn't small, it was not extremely large, which might explain the lower performance of models like SVM or Neural Networks. These models often require larger datasets to learn and generalize adequately, even if their results are not so far from the ensemble learning algorithms.

Looking at the performance metrics, our top classifiers (RF and XGB) showed acceptable levels of accuracy, precision, and recall. This indicates not only the capacity to make accurate predictions (as indicated by high accuracy) but also the ability to minimize false positives (as indicated by high precision) and false negatives (as indicated by high recall). The balanced performance across these three metrics suggests that our models can effectively identify patients who would respond to the treatment without overestimating or underestimating the treatment outcomes.

6.1.5 Limitations

We could find several limitations in our study that could impact the reproducibility and reliability of our findings. The first one is our sample size, which is relatively small, which limits the complexity of models we can reliably use, potentially impacting the performance of our models. Then, we utilized a single set of criteria to define successful treatment outcomes; we could imagine other criteria that could change the prediction. However, the clinical reality of insomnia disorders is complex and subjective, with different patients possibly having different definitions of what they consider a successful treatment outcome. While our criteria were based on expert recommendations and patient-reported outcomes, they may not capture all the nuances of a successful treatment. The features used in the models were based on the data available in our dataset. Other unmeasured variables could be influential in predicting treatment outcomes, such as genetic markers or specifics about individual treatment plans or, unfortunately, the actigraphic features. Lastly, while we have attempted to validate our models using cross-validation, our results still need to be tested in an external dataset to assess the reproducibility of our findings to other patient populations.

6.1.6 Conclusion

The use of ML algorithms, specifically RF and XGB, effectively predicted treatment outcomes with an overall accuracy of > 0.8 in our specific dataset of CID using a moderately sized dataset with high dimensional features. The results of this study signify an important step towards personalized treatment strategies for Insomnia, allowing for more efficient use of resources and potentially leading to better patient outcomes. However, further research is needed, ideally with larger and more diverse datasets, to validate these findings and fine-tune the prediction models. Future studies may also explore more advanced or specialized machine learning and statistical techniques to tackle the complexities inherent in treatment prediction tasks. Also, the definition of treatment outcome must be more investigated. Despite all these limitations, as we validated our first hypothesis, we could test the second one, explaining the prediction.

6.2 Hypothesis two: Explaining the treatment outcome

6.2.1 Hypothesis background and definition

This experiment continues the use of ML models to predict treatment outcomes. We could demonstrate, with all the limitations, that we could predict the treatment outcome with a

relatively good accuracy assessed by F1 score, AUROC, and Classification Accuracy > 0.8. We also confirm the relatively low rate of therapeutic success in our clinical population, with only 44% of positive outcomes at least six months after the initial treatment. This prevalence is in line with the existing studies showing that even in the case of efficient recommended treatment(medication and Cognitive Behavioral Treatment for Insomnia (CBT-I)) [174], the rate of relapse is high with more than 50 % of chance in the four years [137]. In our case, some patients were followed for several years, so our success rate is in line with the literature, although not very satisfactory. Although some authors were interested in describing why some patients dropped out from standard therapy, giving information about the insomniac characteristics, there is still a lack of studies evaluating personalized treatment and assessing in detail the profile of responder or no responder to the different components of pharmacological or behavioral treatment. One of the reasons could be the difficulty of processing all the data generated by an exhaustive assessment according to the international recommendation using clinical evaluation, psychiatric evaluation, psychological evaluation, questionnaires, demographics data, physiological recording of one or several nights (± video), pharmacological evaluation, personality evaluation, and wrist actigraphy for one or two weeks. Altogether, it could be a huge amount of parameters for only one patient.

So, our experiment attempts to bring new knowledge on the variables that could explain treatment resistance.

6.2.2 Methodology

As we had some encouraging results in the prediction, we have a good chance to find some interesting knowledge using the same protocol Ensemble Type Model for Prediction explanation (ETMPE) used in the Paradoxical Insomnia chapters (see 5.3.2). So, we will apply our EMPTPE protocol with the same RF predictive model used in the precedent section. We decided to add two feature importance Explanations using the Feature importance algorithm available in Scikit-Learn to compare with our protocol.

6.2.3 Features extraction from Random Forest

The results of our process with the RF are presented in Figure 6.2 for LASSO, Figure 6.3 for Shap, and Figure 6.4 for Sensitivity Analysis. We could see clearly that the three explainers share one feature, our New definition of Paradoxical Insomnia (IP on the three figures). With a small effect, the Hysteria scale of the MMPI-2 (Hy) is also slightly involved in the prediction.

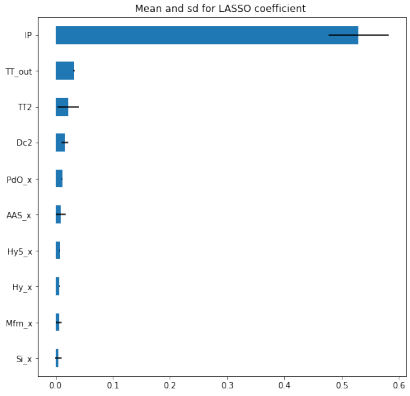


Figure 6.2: Top 10 features selected by LASSO explainer

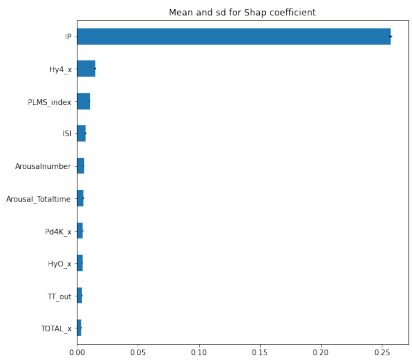


Figure 6.3: Top 10 features selected by Shap explainer

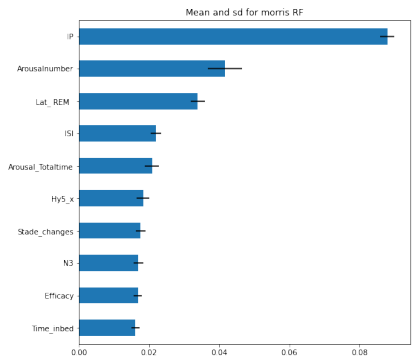


Figure 6.4: Top 10 features selected by SA explainer

To have another view, we ran FI on RF and XGB; the results are presented in Tables 6.4a and 6.4b. We could see the same results, with the new definition of Paradoxical Insomnia as the main explicative features and the Hysterical Scale (named Som for somatoform).

To understand better these features and their interaction, we used visualization techniques.

Table 6.4: Features Importances for Random Forest and XGboosting

(a) Random Forest

Features	Mean	Sd
IP	0.327	0.023
Som_C	0.020	0.003
TTout	0.020	0.003
Arousalnumber	0.016	0.002
LatRem	0.015	0.002
Efficacy	0.015	0.002
ArousalTotaltime	0.014	0.001
StadeChanges	0.014	0.001
PLMSindex	0.013	0.003
Sleeplatency	0.013	0.002
ISI	0.012	0.002
SA	0.012	0.003
N1perc	0.011	0.002
REMperc	0.011	0.002
TTent	0.011	0.002
HyO	0.011	0.010
Hs5K	0.011	0.002
RRN2	0.010	0.001
FRmoysom	0.010	0.001
Hy	0.010	0.001
RDIback	0.010	0.001
TTS	0.010	0.001
HRwake	0.010	0.001
Wakebouts	0.010	0.001

(b) XGBoosting

Features	Mean	Sd
IP	0.08	0.005
Som_C	0.026	0.004
SA	0.021	0.012
Ds	0.018	0.001
SC1	0.018	0.008
Wakebouts	0.018	0.006
TTent	0.018	0.002
25	0.018	0.003
PdO	0.017	0.003
Dc2	0.017	0.003
Sexe	0.017	0.001
MACR	0.017	0.003
HyO	0.016	0.002
PLMAr	0.016	0.007
FB	0.015	0.002
Hy2	0.015	0.001
15	0.015	0.002
MicAr	0.015	0.006
StadesChanges	0.015	0.005
PLMSindex	0.014	0.004
LatREM	0.013	0.001
RDIback	0.013	0.002
TOTAL	0.013	0.001
1	0.013	0.001

6.2.4 Data visualization

Linear Projection

The first is a PCA to confirm the clear separation between a positive and negative treatment outcome. We can see the results in Figure 6.5

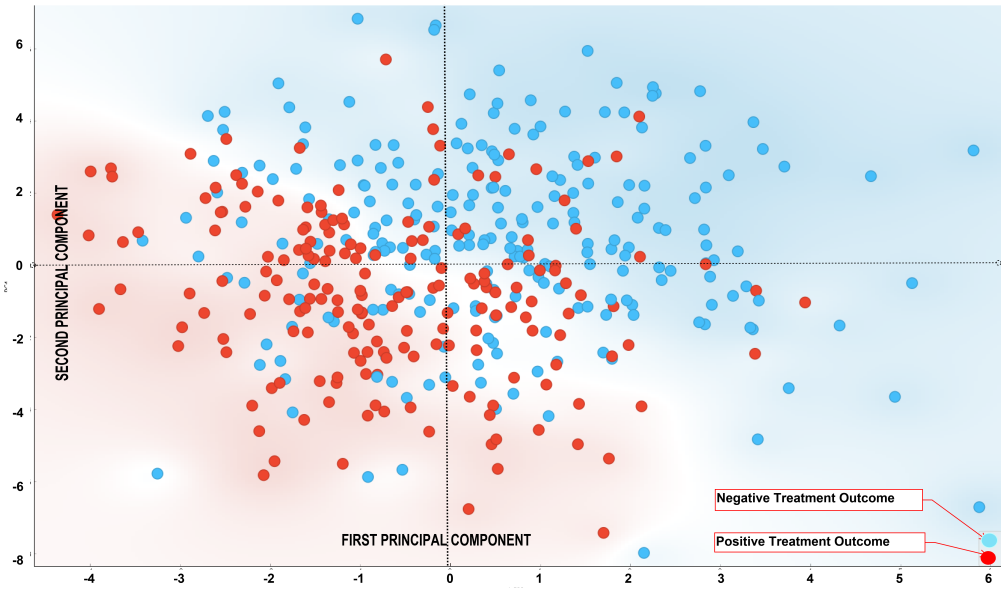


Figure 6.5: PCA with a clear separation between the patients with a positive outcome and negative outcome

FreeViz and DT We used two other visualization techniques to understand the relationship between the different features better.

1. The FreeViz algorithm has a different data representation. This method finds a good two-dimensional linear projection of the given data, where the quality is defined by separating the data from different classes and the proximity of the instances from the same class [86].
2. The DT, as described previously in B.1.2 and 2.10, is a very useful technique to see the clear relationship between the features. We must remember that the representation of a single tree is one view of the problem and not the true relationship between the different features represented.

We can see the relationship of the features with FreeViz in Figure 6.6 and one example of DT in Figure 6.7 with still Paradoxical Insomnia as the most important feature detected.

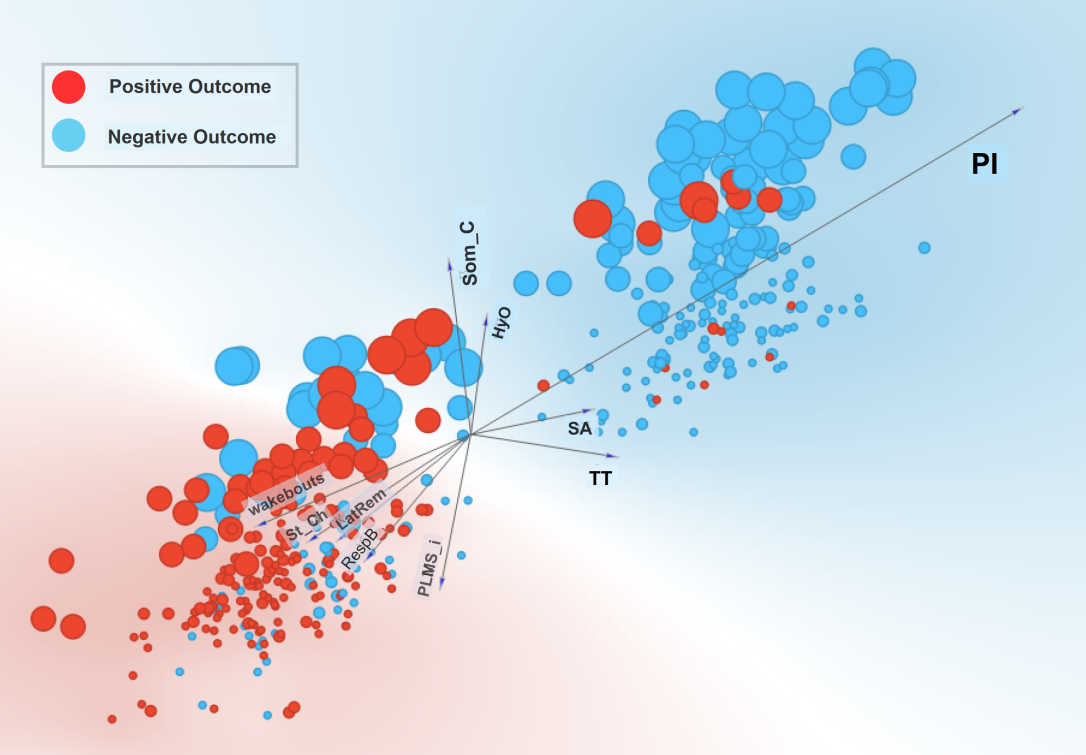


Figure 6.6: Linear Projection through FreeViz Algorithm showing the great interaction between the Paradoxical Insomnia feature positive and the negative treatment outcome.

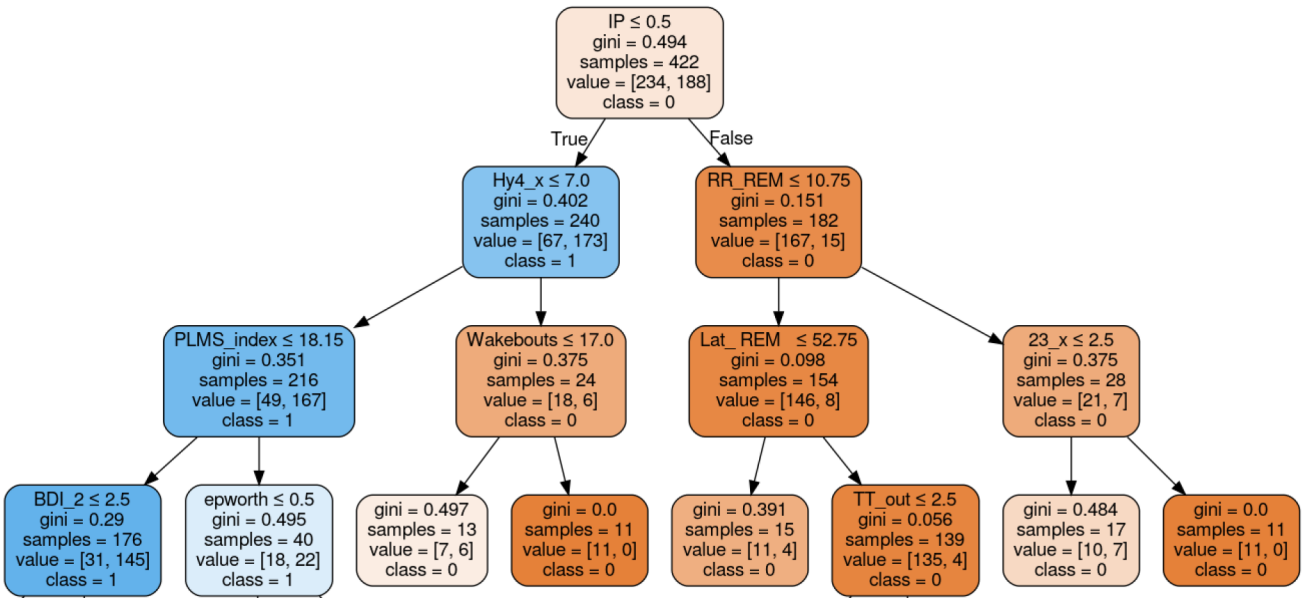


Figure 6.7: One example of Decision Tree used in Random Forest

6.2.5 Discussion and Limitations

The feature selection subsection reveals that the most influential factor identified by our ETMPE protocol, the FI for the best-performing algorithms (RF and XGB) and all the visualization techniques, is the new definition of Paradoxical Insomnia.

All the results presented above showed the major influence of being classified as Paradoxical or not on the treatment outcome. The other parameters pointing toward the negative outcome seem much more anecdotal, such as the hysteria scales of MMPI-2 or the type of treatment used. On the other hand, having awakenings during the night, having numerous sleep stage changes, seeing a shortened REM sleep latency, and respiratory events on the back or periodic leg movements are good prognoses for management. These elements are also reflected in the DT classifier. All these findings require further investigation to understand their relationships.

Indeed, consistently finding the same feature's importance across different methods can offer compelling evidence of its importance in predicting the outcome variable. However, the consistency of a feature's importance across different models does not automatically indicate its causal impact on the outcome variable. We are keeping this fact in mind until a better understanding. At least we can assume a strong correlation, but the underlying reasons may be more complex and could involve other confounding variables not included in the models. Therefore, while it provides strong grounds for further investigation, it does not definitively prove causality. Nevertheless, the diversity of techniques used in our protocol, with different approaches, confirms that this feature is very important in predicting the treatment outcome, even if it is not the causal factor.

We must also keep in mind that the ParI feature is positive for half of the sample, and there is perhaps a risk of overfitting. In any case, a deeper investigation may be necessary to discover the actual cause-effect relationships.

6.3 Conclusion

Despite these limitations, the study has provided valuable insights into the potential use of ML models for predicting and explaining treatment outcomes. The study confirms the complexity of insomnia treatment outcomes and highlights the importance of personalized treatment strategies.

Our findings, in line with the literature, affirm the relatively low therapeutic success rate and high relapse rate. Despite this, we established the influence of a new definition of Paradoxical Insomnia as a major factor in predicting treatment outcomes.

The models and techniques used in this study, such as the ETMPE protocol and the Feature Importance Algorithm, have shown their potential to handle the high dimensionality of patient data. Moreover, visualization techniques like PCA, FreeViz, and DT contributed to a better understanding of data patterns and interactions.

The study underscores the importance of further research into personalizing treatment and understanding the predictors of therapeutic success. Future studies should focus on improving the predictive power of machine learning models, investigating the underlying mechanisms for the identified predictors, and identifying new predictors for better personalizing treatments for Insomnia.

Chapter 7

Conclusions and Future Perspectives

The main contributions will be listed according to the three main research questions (RQ) developed in the Chapter 1, in the Sections 7.1, 7.2 and 7.3.

At the beginning of each contribution, we chose a color code according to their impact in terms of:

1. Generating a new hypothesis
2. Confirming an existing hypothesis
3. Invalidating an existing hypothesis
4. Confirming a recognized and admitted hypothesis
5. Invalidating a recognized and admitted hypothesis.

Each contribution will be introduced with a colored C1.7 and a number corresponding to one of the three hypotheses.

7.1 Hypothesis 1: An Improved Definition of Paradoxical Insomnia using a Data-Driven Approach with Machine Learning Tools

C1.0 is the reinforced confirmation of the heterogeneity of the Paradoxical Insomnia definitions in clinical research studies. We could find similar results from the first study on the topic published in 2019 [35] and add some new results. Our results confirm the increased heterogeneity over time by adding two recent publications since 2019 using two new definitions ([117, 3]. We could show that these two last studies are adding more confusion in the definition with even a correlation equal to 0 between them. So, in total, after implementing 20 different formulas on our dataset, 82% of our sample was diagnosed with Paradoxical Insomnia (ParI) with at least one definition, and we could find 139 combinations (among 335 subjects) between the different formulas to diagnose a subject. But the biggest sample combination was the subjects never classified as ParI, whatever the formulas, meaning that there is a subgroup of Chronic Insomnia Disorder (CID) patients who are not ParI. In conclusion, the different definitions used are not specific to a single subgroup of Paradoxical Insomnia nor different homogenous subgroups, but they could define a clear subgroup of CID that is not Paradoxical Insomniac. So, this first analysis could be used to state that Sleep State Misperception is not a shared symptom by all CIDs but that the definitions used have very low specificity.

We could demonstrate that our dataset was representative of CID in terms of questionnaires, MMPI psychological profiles, and PSG features with similar results to the reference paper on Paradoxical Insomnia [58]

We used an ML algorithm to provide a robust and explainable definition with a minimum number of features for reasons of interpretability and reliability. In particular:

C1.1 We showed on our dataset that it could avoid selecting random features in the top 10 explicative features that would have been selected by an explainer only. In total, 10% of the whole top Feature's Importance (FI) was random. In conclusion, using an explainer only among the most popular ones could not ensure reliability.

C1.2 We found two features (TIB and TPS) and an engineered feature TIME corresponding to the difference between them that could explain 80% of the definitions predictions. We could also show that this new feature, TIME, is correlated at 0.7 with the prevalence of each definition in our sample. These findings lead to the notion of the continuum or

dimensionality of Paradoxical Insomnia instead of categorization. Further discussion between experts is required in the next steps.

C1.3 We showed that the main criteria used until now to define Paradoxical Insomnia, Total Sleep Time (TST) calculated with the PSG, was different between the PSG recording on one night and the six days before for the no Paradoxical Insomnia patients. These findings could lead to abandoning the usual definition to systematically adopt an entire week of analysis.

C1.4 We proposed a new definition of Paradoxical Insomnia based on seven days with a protocol to correct the TST of the actigraphic measures with the PSG recording personalized for each patient. We could show that our definition found very different subgroups of subjects compared to all the other formulas, meaning that criteria applied on a single night do not represent a full week of sleep, no matter which TST is used. This new definition demonstrates significant differences when applied to subgroups of subjects compared to existing definitions with only a light correlation with the S formula. This finding highlights that criteria based on a single night's sleep data may not represent the reality of the Insomniacs studied neither of the Paradoxical Insomnia

C1.5 We brought out, using Features Importance-based explanations and inferential statistics, useful information that overlaps in the case of a major impact of a predictor on the outcome and could be complementary for the more modest contributions needed to gain a finer understanding of the complex interactions between the features used for predictions, especially when there are numerous.

C1.6 In three points:

(a) Thanks to our ETMPE protocol used on states-of-the-art predictive models, the use of unsupervised learning algorithms such as PCA and classical inferential statistics, we were able to show that the explanatory variable TIME was associated with a highly significant increase in periods of arousal, particularly of more than one minute in the group of subjects categorized as Paradoxical Insomniac whatever the formula.

(b) At the same time, we confirmed that the index of micro-arousals per hour of sleep was not significantly different between ParI and non-ParI subjects.

(c) So we claimed that the hypothesis already evoked in some publications that ¨Paradoxical Insomnia may reflect an accumulation of memorized arousals in parallel with a normal sleep time in terms of quality and duration is the most likely explanation for this disorder and not a different sleep leading to abnormal vigilant consciousness.

C1.7 is the demonstration that Sleep State misperception is not a general trait found in all CID (unlike as stated in the last ICSD-3; p 35 of [183]). Further explorations are needed to delimit the normal from the pathologic. However, we argue that the term sleep state misperception is incorrect and should instead be called **Wake State Overperception**.

7.2 Hypothesis 2: Better understanding of treatment outcome, especially the resistance factor and relapses in Chronic Insomnia Disorder and the factors determining its negative evolution

C2.2 was to show, thanks to the systematic benchmarking, grid search, and hyparameters tuning of nine popular predictive models, including one neural network model, we could predict the treatment outcome for 423 patients with an Accuracy > 0.8 with three of them: Random Forest, Extreme Gradient Boosting and Support Vector Machine with similar scores around 0.82. These findings are a step toward treatment personalization (but this needs replication and further discussion, especially regarding the choice of a good or a bad treatment outcome).

Once the Accuracy score of the prediction was validated, we used the same ETMPE protocol described in the 5, and we added a Feature Importances (FI) extraction from

the RF and the XGBoost to compare the results.

C2.1 was to show that, thanks to the FI selection process, the main feature involved in the prediction was the new definition of Paradoxical Insomnia implemented in the dataset. Whatever the method used to explain the prediction, this new definition of ParI is a great predictor of treatment response and, as we could expect, negatively correlated with a positive outcome. Of course, these findings need replication and more analysis to be understood, but it's going in the way to rehabilitate this subtype of Chronic Insomnia.

Our third interest in this work is time series analysis, specifically in the form of EEG brain recording signals. Indeed, as we use brain recording daily, we wanted to know if the ML approach could bring some new insight into the two first hypotheses from a brain analysis perspective. But as we know, extracting useful features from a brain recording could be hard, especially when the recording is ambulatory like ours.

7.3 Hypothesis 3: Identifying a Reliable ML Algorithm to Extract Meaningful Features from Raw EEG Data

C3.1 is that using an algorithm mixing Power Spectral Density extraction with Empirical Mode Decomposition implementation on spindles, SWS, and alpha rhythm is unable to provide enough reliable results on corrupted or not corrupted recording with an overall prediction accuracy of 0.63, below our expected minimum accuracy of 0.8 to use such an algorithm in our protocol. Our other small contribution shows that accuracy matters and the temporal distribution of the detection. Indeed, as we tested the automatic detection algorithm on 16h long datasets, we could see that sleep episodes, whatever the stages (N1, N2, N3, or REM), could be falsely detected seven hours before the real first sleep, both, on good quality or corrupted EEG. So, beyond the accuracy in sleep stage prediction, what poses the most problems here is the failure to consider the sleep episode itself and its structure. This is especially true if we want to predict characteristics of insomniac patients like TIB, TST, TPS, or nap episodes. Indeed, having distorted information on the data type could jeopardize or even mislead insomnia subtype characterization or treatment plans for a given subject.

C3.2 was that after reproducing CNN, automated sleep scoring could achieve a global accuracy score of 0.8 on the original dataset used for its design (MASS dataset), we couldn't reproduce this score on our or another open-source medical dataset (physionet). Like the previous experiment, we obtained an overall accuracy score of 0.63 for these two data sets. So, we found that the DL base learning algorithm was not better than EMD on our dataset or a classical dataset used for sleep studies. These results enhance the danger of assuming that a result obtained after training on a single database can be taken at face value in the real world, especially with complex data such as EEG. Indeed, in the special case of this algorithm, a device intended to be a medical device used these results to claim its reliability in sleep detection of insomnia and even proposed treatment based on these results.

C3.3 was to find an unsatisfactory prediction of spindles by nine algorithms, most of which have been published in sleep research studies and do not achieve the same performance as advertised in most cases, except for the less under-performing algorithm [143]. F-measure scores ranged from 30 to 55 in our sample, making it difficult to trust their eventual contribution if used alone without expert verification. These preliminary results have some important limitations. First, the limited amount of data that we could not increase for this first preliminary study. Indeed, scoring spindles manually is time-consuming, and this time couldn't be taken. Second, there is a limit in the global benchmarking because we didn't test all the algorithms available, especially the new techniques like deep learning techniques.

C3.4 We could accurately predict the low probability for a given subject in our dataset of being detected as psychotic by the MMPI-2 scales Pa and Sc. The study is conducted on a relatively small data set comprising 267 patients. The patients are all CID, so we

must generalize to other datasets to reproduce these results. More parameters should be included as treatment, which may affect sleep spindles detection.

C3.5 is that using K-means clustering methods (K = 2,3,4,5) on 20 embedded vectorial dimensions generated from sleep hypnograms features, after conducting a one-way ANOVA test at a confidence level of 5%, we found that cluster numbers 2, 4, and 5 showed a very significant difference between groups in terms of ESS score (F-statistic and p-Value respectively 14.6 (0.0001), 4.94 (0.002) and 3.31 (0.01).

C3.6 is that using K-means clustering methods (K = 2,3,4,5) on 20 embedded vectorial dimensions generated from sleep hypnograms features, after conducting a one-way ANOVA test at a confidence level of 5%, we could not found a significant difference between groups in terms of ISI score (Best F-statistic and p-Value for cluster number 4 with 2.46 (0.06) and 5 with 2.2 (0.06).

C3.7 is that using K-means clustering methods (K = 2,3,4,5) on 2180 embedded vectorial dimensions generated from EEG Power Spectral Dimensional features, after conducting a one-way ANOVA test at a confidence level of 5%, we could not found a significant difference between groups in terms of ISI and ESS scores

7.4 Other Contributions Emerging from the Initial Hypotheses

The following contributions are somehow linked to Paradoxical Insomnia into the insomnia subtypes classifications removed from the last ICSD-3([183]. As paradoxical Insomnia was a subtype of Primary Chronic Insomnia, discoveries or confirmation of previous discoveries leading to the identification of homogeneous subcategories of chronic insomniacs are consistent with the need for a more refined semiology and a more personalized diagnostic and therapeutic approach to chronic insomnia. Therefore, these discoveries support the idea of including subtypes of chronic insomnia in the international classification of sleep disorders, including Paradoxical Insomnia, however defined.

7.4.1 Finding concerning clusters of CID

C4.1 was to find three clusters on 1182 insomniac patients with K-means clustering on the 87 scales and subscales of the MMPI2. These findings confirm and further extend results found in an old, similar study [56] involving only 100 patients and 13 scales also found three clusters using T-scores on the three validity and ten clinical MMPI scales with the Fortran clustering procedure available. What is also of interest is that the author, Jack Edinger, was a pioneer in insomnia subtyping, chair of the insomnia sections of the ICSD-2 and ICSD-3 and headed AASM's Research Diagnostic Criteria for Insomnia Workgroup, and currently is leading the academy's Insomnia Treatment Guidelines Task Force. So, it seems that 35 years later, our results confirm these clustering premises. However, by the time of this study, one of the three groups was too small and removed from further descriptive or inferential statistical findings [56]; so they kept only two clusters without any similar attempt since then. In 2017, this author participated in a study whose title is evocative: "Characterization of Patients Who Present With Insomnia: Is There Room for a Symptom Cluster-Based Approach?". In that paper [48], they found three clusters among 170 patients with Latent profile analysis from sleep logs, questionnaires, and PSG features again. So, our findings massively support that at least three clusters of CID, mainly based on Psychological profiles, could be found.

C4.2 was to find a link between MMPI-2 subscales (Es and TRT) used as a treatment adherence and outcome indicator showed elevation peaks in our sample and are correlated with one of the three clusters we just mentioned. Even if this discovery needs more research to be confirmed, it could pave the way for a specific approach to these patients, who seem most resistant to the various treatments from a purely psychological angle.

C4.3 was to find very distinct psychological profiles provided by the MMPI2 scales (different from the previous ones) explaining high and low scores on the Dysfunctional Beliefs and Attitude toward Sleep (DBAS) scale thanks to a DT for a regression problem. The Anxiety (ANX) scale of the MMPI-2 and the Lie (L) scale could help discriminate subjects with low scores on the DBAS questionnaire from the ones with high scores. This finding could help personalize the cognitive Behavioral Treatment for insomnia, especially in the cognitive part of the treatment, by providing more personalized insight into the psychological profile of the CID patients.

7.5 Final Words

Machine Learning tools have helped identify key polysomnographic features never used before to categorize Paradoxical Insomnia.

From these findings, we proposed a new definition of Paradoxical Insomnia. Overall, we conclude that. Paradoxical Insomnia is not a "Sleep State Misperception" but a "Wake State Overperception".

We also obtained a better understanding of the factors contributing to treatment outcomes, particularly resistance and relapse in Chronic Insomnia Disorder.

We could not find a way to use an automated sleep scoring reliably on our data set (regardless of data quality). Neither could we predict sleep spindles accurately, even after an extensive benchmarking of the different predictive models published in the literature. But this itself is an interesting outcome. We thus could not involve spindle detection in support of our two main hypotheses. However, we could find clusters of significantly more sleepy insomniac patients thanks to a graph-spectral method based on EEG and hypnogram features.

This thesis was at the interface between medical and computer science research with a constant tradeoff between algorithm transparency, explainability, and expert knowledge. We hope this work could lead to increased collaboration between Medical Doctors and Computer Scientists, which will become essential in the face of the tide of medical data, patient demands, and the personalization of medicine.

The main limitation of our study is the sample size. Although we have defended that it is sufficient to advance our hypotheses, it also limits how far we have advanced them and the certainty we can have around them. We recommend more reliable data-collection practices to anyone undertaking a similar research experience and before commencing, along with an adequate background in data science – even for medical professionals.

Appendix A

Synthèse du Manuscrit en Français

A.1 Introduction

La thèse présentée vise à approfondir la compréhension de l'insomnie paradoxale (IP), un des sous types frequent de l'insomnie chronique (IC) qui affecte 10 à 20 % de la population générale. Cette condition se caractérise par une perception erronée de l'état de sommeil. Bien que fréquent, ce trouble reste mal compris et constitue encore un défi thérapeutique. Notre recherche s'articule autour de l'utilisation d'outils d'apprentissage automatique (AA) pour caractériser l'IC et en particulier l'IP. De plus nous évaluons l'hypothèse de l'apport de l'AA pour prédire la réponse thérapeutique en incluant ce sous-type, dans le but de mieux comprendre les rechutes fréquentes qui affectent 50% des patients avec IC traités.

A.2 Contexte et Problématique

L'IP se manifeste par une discordance entre la perception subjective du sommeil et les mesures objectives obtenues par polysomnographie (PSG). Suivant les définitions utilisées, les patients souffrant d'IP ont la perception de ne pas dormir une plus moins grande partie de leur nuit de sommeil quand un enregistrement de leur sommeil montre une quantité de sommeil en général normale. Ainsi, malgré des traitements adaptés, cette perception peut persister et conduire à un sentiment d'échec et d'impuissance de la part du thérapeute et du patient, entrainer une majoration de l'anxiété et à une surenchère de traitements, parfois iatrogènes, avec des risques psycho-sociaux accrus. A ce jour, il existe encore une compréhension parcellaire de ce trouble et une controverse concernant son existence propre. En effet il n'est pas encore tranché si l'IP est un sous-type de l'IC ou un simple symptôme (mauvaise perception du sommeil) commun à tous les patients insomniaques. Cette controverse pose la question de l'existence même du diagnostic d'IP, qui, si il était confimrmé, nécessiterait donc une définition claire et un prise en charge spécifique. Mais même dans l'hypothèse où la mauvaise perception du sommeil serait uniquement un symptôme ubiquitaire de l'IC, il persiste une question additionnelle non résolue concernant la définition d'un seuil de mauvaise perception du sommeil considéré comme anormal. Pour essayer de répondre à ces questions, nous avons décidé d'utiliser des outils d'AA pour utiliser sans à priori toutes les données disponibles concernant un groupe d'insomniaques et prédire le degré de perception du sommeil grâce aux différentes définitions de l'IP publiées jusqu'ici. Nous avons voulu également étudier l'impact de ces problématiques sur la réponse thérapeutique. Enfin, nous avons voulu savoir si les outils d'AA pouvaient nous permettre d'exploiter de manière plus fiables et reproductibles les séries temporelles enregistrées lors des PSG, théoriquement objectives mais soumises à une variabilité de l'interprétation humaine. Ces problématiques correspondent aux 3 hypothèses de recherche décrites ci-après.

La première hypothèse testée dans cette thèse est qu'il est possible d'améliorer la définition d'un seuil de perception anormal du sommeil utilisable en clinique et de l'utiliser pour définir l'IP comme un sous-type clinique à l'aide d'une approche fondée sur les données et l'apprentissage automatique. Cette première hypothèse inclue de tester l'hétérogénéité des définitions déjà publiées sur un dataset représentatif d'insomniaques chroniques, et la proposition d'une unification de la définition basée sur une analyse de sommeil sur sept nuits consécutives au lieu d'une nuit habituellement.

La deuxième hypothèse de recherche est que nous pouvons obtenir une meilleure compréhension des facteurs responsables de l'efficacité ou de la résistance à un traitement classique de l'IC à l'aide d'une approche fondée sur les données et l'apprentissage automatique. Cette deuxième hypothèse inclut la possibilité d'une prédiction fiable du succès ou de l'échec

thérapeutique sur des nouveaux patients.

La troisième hypothèse générale est que l'on peut utiliser un algorithme d'AA fiable pour extraire des caractéristiques significatives à partir de séries temporelles brutes, en particulier issues de l'électroencéphalographie (EEG) et ainsi automatiser les interprétations et les prédictions pour pouvoir uniformiser la recherche sur le sommeil sans dépendre de la variabilité inter-experts.

A.3 Matériels et Méthodes

La première hypothèse est testé sur une base de données multimodale de 335 patients souffrant d'IC constituée dans un centre spécialisé dans le diagnostic et la prise en charge de l'IC (Service de Neurophysiologie Clinique, Hopital de la Pitié-salpétrière, Pris). Cette base inclut des données cliniques, psychométriques, actimétriques et polysomnographiques incluant des enregistrements EEG sur huit canaux. Chaque patient inclus a été suivi pendant au moins six mois, permettant une évaluation précise du diagnostic et de la réponse au traitement standard. En utilisant des outils d'AA, l'étude a cherché à identifier des sous-groupes de patients et à tester des hypothèses existantes concernant les profils d'IP à travers l'analyse des données présentes pour chaque patient. Pour prédire chaque définition nous avons sélectionné le modèle le plus performant parmi les plus utilisés (notamment, Random Forest, Extreme Gradient Boosting ou Support Vector Machine), puis nous avons utilisé des modèles globaux d'explicabilité de la prédiction pour comprendre les variables impliquées. La deuxième hypothèse concernant la réponse au traitement et l'implication des différents sous-types de l'IC a été conduite sur la meme base élargie à 423 patients. Le protocole est comparable, mais la prédiction porte sur l'appartenance ou non au groupe des patients ayant répondu positivement à un traitement standardisé. Nous avons également appliqué trois modèles d'explicabilité pour essayer de comprendre de manière fiable ce qui conduisait à un échec thérapeutique. La troisième hypothèse concerne la fiabilité des outils d'AA appliqués à l'EEG pour prédire les stades de sommeil et détecter des fuseaux de sommeil mieux que ne le ferait un expert.

A.4 Résultats

Nous retrouvons une grande hétérogénéité dans les définitions existantes de l'IP sur notre dataset et donc des seuils de mauavaises perceptions du sommeil. L'application des différentes définitions quantitatives utilisées pour diagnostiquer l'IP sur notre dataset ont montré que la majorité des patients étudiés étaient classifiés comme souffrant d'IP selon au moins une définition, mais il n'y avait pas de consensus général. La recherche a également indiqué qu'un groupe homogène de patients atteints d'IC n'était jamais classé IP, quelles que soient les 20 définitions utilisées. Cette dernière observation suggère que la perception erronée du sommeil n'est pas pathognomonique de l'IC et donc que l'IP est bien un sous-type de l'IC. En utilisant l'apprentissage automatique, notre travail a permis de proposer une nouvelle définition de l'IP, basée sur une analyse temporelle plus longue et moins sujette aux aléas d'une seule nuit d'enregistrement, qui semble mieux refléter la réalité des patients et permet de distinguer plus clairement le sous type IP des autres formes d'IC.

Nos résultats sur la prédiction de la réponse thérapeutique ont atteint une précision supérieure à 0,8 grace à un modèle ajusté de Random Forest. L'analyse des variables explicatives impliquées dans cette prédiction mettent en évidence l'importance de l'IP définie comme prédicteur majeur de la réponse au traitement. Cette découverte ouvre la voie à une approche plus personnalisée dans le traitement de l'insomnie chronique, bien que des études supplémentaires soient nécessaires pour une compréhension plus approfondie et pour valider ces résultats.

Concernant les outils d'AA pour harmoniser les analyses de séries temporelles nos résultats n'ont pas montré une performance suffisante pour la prédiction fiable des états de sommeil, avec une précision inférieure à l'objectif fixé correspondant à l'agreement inter-scoreur (précision de 0,8).

A.5 Implications Cliniques et Perspectives

Les résultats de cette recherche ont des implications cliniques importantes. Ils suggèrent que l'approche actuelle de traitement de l'IP pourrait nécessiter une révision, en mettant davantage l'accent sur la perception subjective du sommeil sur plusieurs nuits. Cette approche pourrait aider à identifier plus précisément les patients souffrant réellement d'IP et à leur fournir des traitements plus ciblés et efficaces.

En outre, cette recherche ouvre la voie à de futures études utilisant l'AA pour mieux comprendre et traiter d'autres troubles du sommeil. La capacité de l'AA à analyser de grandes quantités de données et à identifier des modèles complexes peut révolutionner la manière dont nous abordons les troubles du sommeil, conduisant à des diagnostics plus précis et à des traitements plus personnalisés.

A.6 Conclusion

Cette thèse représente une avancée significative dans la compréhension et le traitement de l'IP. En utilisant des outils d'AA pour analyser des ensembles de données complexes, cette recherche contribue à une meilleure caractérisation de l'IP et à une prédiction plus précise de la réponse au traitement de l'IC.

Appendix B

B.1 Definitions

B.1.1 Sleep and Medicine Definitions

Apnea-Hypopnea Index is a measure used in sleep medicine to evaluate the severity of sleep apnea. The AHI quantifies the average number of apneas and hypopneas per hour of sleep. To calculate the AHI, the total number of apneas and hypopneas observed during a sleep study (usually obtained through polysomnography) is divided by the total number of hours of sleep. The result is the number of apnea and hypopnea events per hour. The AHI is used to classify the severity of sleep apnea: AHI < 5: Normal, AHI 5-15: Mild, AHI 15-30: Moderate, AHI > 30: Severe

BDNF The protein brain−derived neurotrophic factor (BDNF) is a member of the neurotrophin family of growth factors involved in the plasticity of neurons in several brain regions. There is evidence that BDNF expression is decreased by experiencing psychological stress and that a lack of neurotrophic support causes major depression. [184]

Cognitive and behavioral therapy for Insomnia is a non-pharmacological technique that has shown its efficacity in CID. [149]. The main axes of this therapy are Sleep education: Learning about the factors that influence sleep, sleep patterns, and the impact of lifestyle choices on sleep. Sleep restriction: Establishing a consistent sleep schedule and limiting time spent in bed to match actual sleep time helps improve sleep efficiency. Stimulus control: Adjusting the sleep environment and bedtime routine to associate the bed with sleep and relaxation, reducing factors that may interfere with sleep. Sleep hygiene: Adopting healthy sleep habits, such as maintaining a regular sleep schedule, avoiding stimulants close to bedtime, and creating a comfortable sleep environment. Cognitive therapy: Identifying and challenging negative thoughts and beliefs about sleep that contribute to insomnia and replacing them with more positive and realistic ones. Relaxation techniques: Practicing relaxation exercises, such as progressive muscle relaxation or deep breathing, to reduce physical and mental tension before bedtime. CBT-I is typically delivered in a structured format over several sessions with a trained therapist.

European data Format is a file format commonly used for storing and exchanging medical time series data, particularly physiological signals such as EEG. The EDF format allows for the standardized representation of data collected during medical examinations or research. It provides a structured way to organize and store multiple channels of time series data, along with relevant metadata and annotations.

Electroencephalography is a method of cerebral exploration that measures the brain's electrical activity using electrodes placed on the scalp, often represented as an electroencephalogram trace. Comparable to the electrocardiogram, which studies the functioning of the heart, the EEG is a painless, non-invasive examination that provides information on the neurophysiological activity of the brain over time, and of the cerebral cortex in particular, either for diagnostic purposes in neurology or for research in cognitive neuroscience. The electrical signal at the origin of the EEG is the summation of synchronous post-synaptic potentials from many neurons.

Montreal Archive of Sleep Studies (MASS) is an open-access and anonymized polysomnographic dataset that contains sleep studies on various patient groups, including healthy subjects and patients with various sleep disorders. The dataset includes raw EEG, demographic, and hypnograms, which are sleep stage scoring data. MASS is highly useful for developing and evaluating automatic sleep staging algorithms or studying various sleep disorders and conditions. There are different subsets of the sample.

Narcolepsy is a neurological disorder characterized by severe, irresistible daytime sleepiness and sudden loss of muscle tone (cataplexy) and can be associated with sleep−onset or sleep−offset paralysis and hallucinations. These sudden sleep attacks may occur during any activity at any time of the day. This disorder is secondary to the early loss of neurons in the hypothalamus that produce Orexin, a wakefulness−associated neurotransmitter. The cause of neural loss could be autoimmune since most patients have the HLA DQB1*0602 allele that predisposes individuals to the disorder

Psychophysiologic insomnia, also known as psychophysiological insomnia or learned insomnia, is a type of sleep disorder characterized by difficulty falling asleep or staying asleep that is primarily caused by psychological or emotional factors. A chronic condition often develops due to a person's negative thoughts, worries, and anxieties surrounding sleep. Individuals with psychophysiological insomnia typically experience hyperarousal and awareness of their sleep-related thoughts and bodily sensations. This heightened state of vigilance can make it difficult for them to relax and fall asleep, leading to chronic insomnia. The condition often develops through a process known as conditioned arousal. This occurs when an individual begins associating their bed or sleep environment with frustration, anxiety, and wakefulness instead of relaxation and sleep. This negative association creates a cycle of sleeplessness and further reinforces the individual's difficulties with sleep. Psychophysiologic insomnia is often linked to psychological and emotional factors, such as stress, anxiety, depression, and traumatic experiences. It can also be influenced by maladaptive sleep habits, poor sleep hygiene, irregular sleep schedules, excessive time spent in bed while awake, and an overall preoccupation with sleep. Treatment for psychophysiological insomnia typically involves a combination of cognitive-behavioral therapy (CBT) and sleep hygiene practices. [149]

REM sleep behavior disorder Patients with REM sleep behavior disorder (RBD) enact violent dreams without normal muscle atonia during REM sleep. This disorder is highly frequent in patients with synucleinopathies (60%−100% of patients) and rare in other neurodegenerative disorders. The disorder is detected by interview plus video and sleep monitoring [5].

Sleep efficiency is another important parameter that refers to the percentage of total time in bed spent in sleep. It is calculated as the sum of Stage N1, Stage N2, Stage N3, and REM sleep, divided by the total time in bed and multiplied by 100. Sleep efficiency gives an overall sense of how well the patient slept but does not distinguish frequent, brief episodes of wakefulness. A low sleep efficiency percentage could result from long sleep latency and long sleep offset to lights on time with otherwise normal quantity and quality of sleep in between. **Sleep onset latency (SOL)** is the duration of time between when the lights are turned off (lights out) as the patient attempts to sleep until the time patient falls asleep, as evidenced by EEG and behavioral parameters changes consistent with sleeping (three epoch of Stage N1 sleep or one epoch of other sleep stages) **Sleep state misperception** in The International Classification of Sleep Disorders, Revised [198], sleep state misperception (also known as pseudo insomnia or subjective insomnia) is a disorder in which a complaint of insomnia arises when polysomnography demonstrates a "normal sleep pattern" with sleep onset latencies of less than 15 to 20 minutes, sleep durations over 6.5 hours, and an average number and duration of awakenings. **total sleep time** is the total sleep time scored during the total recording time. This includes the time from onset to offset and is distributed throughout the sleep time as minutes of Stage N1 sleep, Stage N2 sleep, Stage N3, and REM sleep. Recurrent awakenings, define as high sleep fragmentation levels and stage shifts, may result in complaints of non-restorative sleep even when a normal total sleep time is present.

Wake After Sleep Onset refers to periods of wakefulness occurring after sleep onset latency. This parameter measures wakefulness, excluding the wakefulness occurring before sleep onset.

B.1.2 Machine Learning Definitions

AdaBoost for Adaptive Boosting is an ensemble learning algorithm that constructs a classifier by fitting multiple weak classifiers on various data distributions and then combines them into a weighted sum to form a final single strong classifier. [68] **Adam** for "Adaptive Moment Estimation" is an optimization algorithm used in deep learning applications, which can be used to replace the classical stochastic gradient descent procedure to update network weights iteratively based on the training data. Adam is known for its computational efficiency and has little memory requirements. It is particularly suitable for problems with large data or many parameters. Adam maintains a per-parameter learning rate that improves performance when dealing with sparse gradients on noisy problems. It uses estimations of the first and second moments of the gradient to adapt the learning rate for each weight of the neural network. Adam includes bias correction estimates to handle the issues of sparse gradients and noisy data [104].

Area Under the ROC Curve (AUC−ROC) The AUC−ROC is a performance metric used for binary classification. It measures the model's ability to discriminate between positive and negative instances across different probability thresholds. A higher AUC−ROC value indicates better classification performance [70]. **Bagging** Stand for Bootstrap Aggregating. It is an ensemble learning technique that aims to improve the stability and accuracy of ML models by combining predictions from multiple models trained on different subsets of the original training data. Each model is trained independently on a randomly sampled subset of the training data with replacement. Bagging reduces the variance and helps mitigate overfitting by averaging the predictions of individual models [23]. **Black Box** Refers to a model or system whose internal workings or decision-making process is not transparent or easily interpretable. Although the model can provide accurate predictions or outputs, it may not clearly understand how it arrives at those results. Black box models are essentially referred to as deep learning models [18].

Boosting is an ensemble learning technique that combines multiple weak or base learners to create a strong predictive model. Unlike bagging, boosting trains models sequentially, where each subsequent model is trained to correct the mistakes made by the previous models. The final prediction is a weighted combination of the predictions from all the models. Boosting focuses on reducing bias and variance, improving overall performance [68].

CART stands for Classification and Regression Trees, a machine learning algorithm used for classification and regression tasks. A decision tree−based algorithm recursively partitions the input space into smaller regions, creating a tree−like model for making predictions.

Classification Accuracyis a metric accuracy measures the proportion of correct predictions to the total number of predictions. It is a simple and widely used metric, but it may not be suitable for imbalanced datasets [70].

Classifier A machine learning model used in supervised machine learning. It is designed to assign input data points to predefined categories or classes based on their features or attributes. Examples of classifiers include logistic regression, decision trees, support vector machines, and neural networks [70].

Convolutional Neural Networks are a type of artificial neural network specifically designed for analyzing visual data. Inspired by the visual cortex in animals, CNN utilizes convolutional layers to learn local patterns and spatial hierarchies in input data automatically. CNN detects features such as edges and textures by sliding filters over the data. Pooling layers reduce dimensionality while retaining important information. Non-linear activation functions introduce non-linearity, fully connected layers learn high-level representations, and backpropagation enables training. CNN excels in image-related tasks by extracting meaningful representations and can be adapted for other data types like time series analysis. [75]

Cross−validation A technique used to evaluate the performance of a machine learning model. It involves dividing the dataset into multiple subsets or folds. The model is trained on some data (training set) and tested on the remaining data (validation set). This process

is repeated multiple times, with each fold being the validation set once. The results are then averaged to estimate the model's performance [51].

Decision Tree The decision tree algorithm is used for both classification and regression tasks. We used the Gini index as the splitting criterion to build the DT. The Gini index measures the impurity of a set of samples by computing the probability of misclassifying a sample in that set if it were randomly assigned to a class. The Gini index ranges from 0 to 1, with 0 indicating a completely pure set and 1 indicating a completely impure set.

The algorithm for building a decision tree using the Gini index can be described as follows: If all samples in the current node belong to the same class, stop and return that class label. For each feature, calculate the Gini index of the split resulting from splitting on that feature. Select the feature that results in the lowest Gini index and split the node based on that feature. Recursively apply steps 1−3 to the resulting child nodes until a stopping criterion is met.

The formula for the Gini index is:

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2 \tag{B.1}$$

where D is the set of samples being considered, c is the number of classes, and p_i is the proportion of samples in D that belong to class i [84].

Data Mining is the process of discovering patterns, relationships, and insights from large datasets. It involves extracting valuable information from raw data using various techniques, such as statistical analysis, machine learning, and pattern recognition. Data mining aims to uncover hidden patterns or knowledge that can be useful later in building a prediction model or understanding complex systems [170].

Data Science refers to the interdisciplinary field that involves extracting insights and knowledge from data using various techniques, including statistical analysis, machine learning, data visualization, and data mining. It combines elements from mathematics, statistics, computer science, and domain expertise to uncover patterns, make predictions, and gain actionable insights from complex and large datasets [170].

Deep Learning is a subfield of ML that focuses on training deep neural networks, which are artificial neural networks (ANN) with multiple layers. Deep learning models are designed to automatically learn hierarchical representations of data by stacking layers of artificial neurons [**DLbishop2006pattern**].

Eigenvalue Decompositionis a method used to decompose a square matrix into its constituent parts. Specifically, given a matrix A, its Eigenvalue Decomposition represents A as the product of three matrices: V, Λ, and V^{-1}. − V is a matrix whose columns are the eigenvectors of A. − Λ is a diagonal matrix containing the eigenvalues of A. − V^{-1} is the inverse of V. The Eigenvalue Decomposition of a matrix A can be expressed as:

$$A = V\Lambda V^{-1}$$

Eigenvalue Decomposition is particularly useful because it provides insights into the properties and behavior of the original matrix. It can reveal important characteristics such as the eigenvalues (representing scaling factors) and eigenvectors (representing directions of transformation) associated with the matrix.

Empirical Mode Decomposition is a signal processing technique that decomposes a given signal into a set of Intrinsic Mode Functions (IMFs). Each IMF represents a specific oscillatory mode contained within the signal. EMD does not rely on predefined basis functions and adapts to the local characteristics of the signal. It has been widely used for analyzing nonlinear and non-stationary data.

Intrinsic Mode Functions are the building blocks of EMD. In EMD, a given signal is decomposed into a set of IMFs representing the different oscillatory modes or components present in the signal. Thus, IMF captures the intrinsic oscillatory behavior of a signal at a specific scale. The first IMF generally represents the highest-frequency oscillation, while subsequent IMFs capture progressively lower-frequency components.

F1 Score The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance by considering both precision and recall [70].

Feature Importances correspond to the assessment process of the relevance or contribution of individual features or variables in a dataset towards predicting a target variable or outcome. It is commonly used in machine learning and statistical modeling to understand which features significantly impact the model's predictions..

Freeviz Freeviz, like Radviz, is a visualization method. The data instances are plotted inside a circle, the position of each determined by the value of its features and the positions of the corresponding anchors. Informally, each anchor pulls the instance towards itself with a strength proportional to the value of the corresponding feature, so the position of an example depends upon the relative values of features (e.g., if all features have equal values, the instance is placed in the center). Despite a slightly different mathematical formulation, RadViz and FreeViz are similar, with the essential difference that in FreeViz, the "anchors" can be anywhere in the projection plane and are not placed evenly around the circle. To use FreeViz visualization in classification, the projection is used to find the coordinates of a new, unclassified instance and let the instances from the original training set "vote" for its class, with the weight inversely proportional to their distance to the new instance. The classifier can either predict a class or normalize the distribution of votes to obtain a class probability estimate [53]

Hyperparameter Tuning Involves selecting the optimal values for the hyperparameters of a machine learning model. Hyperparameters are parameters not learned from the data but are set by the practitioner before training the model. Hyperparameter tuning is important as it can significantly impact the model's performance. **Independent Component Analysis**is a statistical signal processing technique that separates a set of mixed signals into their underlying independent components. It assumes that the observed signals are linear combinations of statistically independent source signals and aims to estimate the mixing matrix and the source signals by maximizing statistical independence. [43]

K-Nearest Neighbors or KNN is a non-parametric and lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. KNN does not make any assumptions on the underlying data distribution, but it does assume that data is in a feature space and the distance metric can be calculated. [80]

LASSO (Least Absolute Shrinkage and Selection Operator) LASSO is a feature selection and regularisation technique in linear regression models. It works by imposing a penalty on the sum of the absolute values of the regression coefficients, which encourages sparse solutions and helps to avoid overfitting. LASSO is particularly useful when dealing with high-dimensional datasets with potentially irrelevant or redundant features. It can help identify the most relevant features for the prediction and improve the model's interpretability and generalization. LASSO is a powerful technique for feature selection in binary classification problems, which can provide insights into the most important features for prediction. [84].

Loss Function A loss function, also known as a cost function or error function, is a function that maps a set of parameter values for a model to a scalar value that represents the cost, error, or "loss" of the model's predictions with those parameters, compared to the true values of the target variable. The goal of a machine learning algorithm is typically to find the model parameters that minimize the loss function. There are many types of loss functions [84], and the choice of the loss function can depend on the specific machine learning task. Here are a few examples:

* **Mean Squared Error**is commonly used in regression tasks. It is the average squared difference between the predicted and actual values.

$$L(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

* **Cross−Entropy Loss**: This is often used in classification tasks, particularly with probabilistic outputs. It measures the dissimilarity between the predicted probability

distribution and the actual distribution.

$$L(y, \hat{y}) = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

* **Hinge Loss**: This is used for "maximum−margin" classification tasks, most notably with Support Vector Machines.

$$L(y, \hat{y}) = \max(0, 1 - y_i \cdot \hat{y}_i)$$

Matthews correlation coefficient measures the quality of binary classifications, which considers true and false positives and negatives. It ranges from−1 to +1, with +1 indicating a perfect prediction, 0 indicating a random prediction, and−1 indicating a completely incorrect prediction. The formula for MCC is:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{B.2}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

MNE for MNE-Python is an open-source software package designed to process, analyze, and visualize functional neuroimaging data (specifically MEG, EEG, sEEG, ECoG, and fNIRS data). [78]

Multilayer Perceptron (MLP) is a type of artificial neural network model often used for classification tasks, including in medical datasets where the goal might be to predict the presence or absence of a disease given a set of symptoms or other similar tasks. An MLP consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. Each node in one layer connects with a certain weight to every node in the following layer. These networks are called 'fully connected'.MLPs use a supervised learning technique called backpropagation for training. [12]

a Model is a mathematical or computational representation of a real−world process, system, or phenomenon. It captures the relationships and patterns within the data to make predictions, classifications, or other forms of analysis. Models can be created using various algorithms and techniques and are trained or optimized based on available data [84].

Naive Bayes is a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. It is a simple and efficient classification task method, particularly for large datasets. The "naive" assumption of this classifier is that the presence of a particular feature in a class is unrelated to the presence of any other feature, even if these features are dependent on each other. This simplifies computation, and that's why it is considered 'naive'. [132]

Optimization An optimization algorithm in machine learning is a procedure or method used to improve a model or function at a given task. Optimization algorithms navigate the landscape of the chosen model's loss function to find parameters that minimize the loss. Several optimization algorithms are used in machine learning, and the choice depends on the specific task, model, and sometimes even the data. A common example is Gradient Descent optimization [178].

In all these examples, $f(\theta)$ is the objective function, $\nabla f(\theta)$ is the gradient of the function at θ, H is the Hessian matrix (matrix of second derivatives), and α is the learning rate, which determines the step size during the iterative process.

These algorithms aim to find the model parameters that minimize the loss function, making the model's predictions as accurate as possible.

Out−of−Sample Prediction Refers to evaluating a machine learning model's performance on data it has not seen during the training phase. It involves predicting new, unseen data points to assess how well the model generalizes and performs in real-world scenarios.

Power Spectral density is denoted as $S(f)$ and is calculated as the squared magnitude of the Fourier Transform of the signal $x(t)$:

$$S(f) = |X(f)|^2$$

where $X(f)$ is the Fourier Transform of $x(t)$.

Parameters The variables within a model that are learned or estimated during the training process. These values define the specific configuration or behavior of the model. In machine learning, parameters are adjusted iteratively to minimize the difference between predicted and actual outputs. They are often learned from the data and can affect the model's performance and generalization ability [84].

Precision Precision is the ratio of true positive predictions to the total number of positive predictions. It measures the model's ability to identify positive instances correctly. High precision indicates a low rate of false positives [84].

Prediction Estimating or forecasting an unknown or future outcome based on available data and learned patterns. Prediction typically involves using a trained model to make inferences or generate outputs for new or unseen data points. It can be made for various types of problems, such as regression (predicting a continuous value) or classification (predicting a categorical label) [84].

Pruning Refers to reducing a decision tree's size by eliminating unnecessary branches or nodes. The goal of pruning is to improve the generalization capability of the tree by reducing overfitting, where the model becomes too specific to the training data and performs poorly on new, unseen data [84].

Recall Recall is the ratio of true positive predictions to the total number of positive instances. It measures the model's ability to identify all positive instances. High recall indicates a low rate of false negatives [84].

Root Mean Squared Error RMSE is the square root of MSE, providing a measure of the average prediction error in the original units of the target variable. It is commonly used for regression problems and measures the average prediction error [84].

Shap Shapley Values is a technique used to explain the predictions of a model by assigning contributions to each input feature in the prediction. It is a model-agnostic technique that provides a unified framework for feature importance measurement, even when the features are correlated. Shapley values are based on cooperative game theory and assign the contribution of each feature to the prediction by comparing its inclusion in different subsets of features. This technique can help to understand how different features contribute to the model's predictions and to identify potential biases or confounding factors.

Sensitivity Analysis is a technique used to determine how sensitive a model's predictions are to changes in the input variables. It involves varying the values of one or more input variables while keeping the other variables constant and observing the corresponding changes in the model's output. Sensitivity analysis can help identify the most important input variables that drive the model's predictions and evaluate the robustness of the model.

Specificity Specificity is the ratio of true negative predictions to the total number of actual negative instances. It measures the model's ability to identify negative instances correctly. High specificity indicates a low rate of false positives [84].

Supervised Learning A paradigm in which an algorithm learns a mapping between input data and corresponding output labels by being provided with labeled training examples. The algorithm learns to generalize from labeled data and can predict unseen data based on the learned patterns [84].

Unsupervised Learning An approach where the algorithm learns patterns and structures in the input data without being provided with explicit labels or supervision. The objective is to explore the data's inherent structure and identify patterns, clusters, or relationships within it [84].

Explainable AI is invested in making the decisions taken by algorithms understandable to humans. In brief, algorithms could be categorized as white-box if explainable or black-box if hardly understood by domain experts [128]. So understanding the reasoning behind decisions or predictions is the holy grail of this discipline, where research focuses on news algorithms designed to "interpret" or "explain" mainly black-box algorithms and, in any case, to achieve more transparency in the decision-making process leading to a given prediction. These three terms correspond to different definitions.

1. **Interpretability** refers to the degree to which a human can understand the cause of a decision made by a machine learning model. An interpretable model can explain the relationship between the input features and the predictions [126].
2. **Explainability** focuses more on providing insights into the factors that led to a particular decision. In this context, an explanation might be a set of features in the interpretable domain that have contributed to a decision for a given example [88].
3. **Transparency** is more related to explaining the whole process of transformation behind the algorithm from the input data into training features, how the learning process works, and how predictions are generated from the testing data. This term is the most related to ethical aspects. Theoretically, The points that need to be addressed under this concept are the algorithms, data, goals, outcomes, compliance, influence, and usage [114].

Extreme Gradient Boosting (XGBoost) is a scalable and improved version of the gradient boosting algorithm designed for speed and performance. It is a machine learning algorithm that belongs to the ensemble learning method, and it constructs new classifiers that aim to predict more accurately than existing ones and is often used for supervised learning problems. [39]

The XGBoost algorithm can be described using the following formula:

$$\hat{y_i} = \phi(\mathbf{x}i) = \sum k = 1^K f_k(\mathbf{x}_i) \tag{B.3}$$

In this equation, each $f_k(\mathbf{x}_i)$ represents the prediction made by the k-th decision tree in the ensemble for the input vector $\mathbf{x}i$. The sum of these individual predictions $\sum k = 1^K f_k(\mathbf{x}_i)$ produces the final prediction $\hat{y_i}$ where $\hat{y_i}$ is the predicted value for the *i-th* sample, $\mathbf{x}_i$ is the input vector for the *i-th* sample, K is the number of decision trees in the ensemble, and $f_k(\mathbf{x}_i)$ is the prediction of the *k-th* decision tree.

Each decision tree f_k is constructed to minimize a loss function L with an additional regularization term that penalizes the complexity of the tree:

$$\mathcal{L} = \sum_{i=1}^{n} L(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \tag{B.4}$$

where y_i is the true value for the i−th sample, $\hat{y_i}$ is the predicted value, and $\Omega(f_k)$ is the regularization term that penalizes the complexity of the tree f_k.

The regularization term can be expressed as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{B.5}$$

where T is the number of leaves in the tree, w_j is the weight of the j-th leaf, γ is a hyperparameter that controls the complexity of the tree, and λ is the regularization parameter.

The prediction of each decision tree f_k is computed as a sum of the predicted values of its leaves:

$$f_k(\mathbf{x}) = \sum_{j=1}^{T} w_{q(\mathbf{x})_j} \tag{B.6}$$

where $w_{q(\mathbf{x})_j}$ is the weight of the j−th leaf for the input vector $\mathbf{x}$, and $q(\mathbf{x})_j$ is the index of the leaf node that $\mathbf{x}$ falls into.

B.2 Results

B.2.1 Pseudocode algorithm EMD

```
Input: EEG signal, EOG signal EXG signal
Output: Spindle detection and Sleep scoring

Procedure RemoveEXG(EEG signal, EXG signal):
 Combine EEG signal and EXG signal
 Perform Empirical Mode Decomposition on the combined signal to obtain IMFs
 Remove IMFs associated with the EXG signal
 Extract modified EEG signal from the remaining IMFs
 Return modified EEG signal

Procedure EMD(EEG signal):
 Perform Empirical Mode Decomposition on the EEG signal to obtain IMFs
 Return IMFs

Procedure CalculateLikelihood(IMFs):
 Extract robust additional features from the decomposed IMFs
 Calculate the likelihood of each signal segment belonging to a spindle
 using the additional features
 Return likelihood values

Procedure ViterbiDecoder(likelihood values):
 Train the Viterbi decoder using the labeled spindle and non-spindle segments
. Decode the spindle likeness based on the likelihood values using the
 trained Viterbi decoder
 Return spindle classification result

Procedure SpindleAnalysis(EEG signal):
 IMFs = EMD(Modified EEG signal)
 Likelihood values = CalculateLikelihood(IMFs)
 Spindle classification = ViterbiDecoder(Likelihood values)
 Return spindle classification
Procedure Slowwavesleep Analysis(EEG signal):
 IMFs = EMD(Modified EEG signal)
 Likelihood values = CalculateLikelihood(IMFs)
 SWS classification = ViterbiDecoder(Likelihood values)
 Return Slowwavesleep classification
Procedure AlphawaveAnalysis(EEG signal):
 IMFs = EMD(Modified EEG signal)
 Likelihood values = CalculateLikelihood(IMFs)
 Alphawave classification = ViterbiDecoder(Likelihood values)
 Return Alphawave classification

Procedure ApplyLimitsforsleepclassification(EEG signal, EOG signal, spindles):
 If AlphawaveAnalysis(EEG signal) = True
 If alpha frequency (7.5 and 12.5 Hz), 10 uV < amplitude < 50 uV
 AND/OR  beta frequency (15 and 30 Hz),
 0 uV < amplitude < 250 uV  > 15 sec
 then epochs = Wake.
 Else if theta > 15 sec, with theta frequency between 3.5 and 7.5 Hz
  Apply limits on amplitude: 25 uV < amplitude < 100 uV,
  duration > 1 sec
  Return 30 sec epochs = N1
```

```
Else if theta > 15 sec, with theta frequency between 3.5 and 7.5 Hz
 Apply limits on amplitude: 25 uV < amplitude < 100 uV, duration > 1 sec
 if SpindleAnalysis(EEG signal) return = true
 Apply limits on sigma power: rhythm frequency between 11.5 and 14.5 Hz
 Apply limits on sigma amplitude: 10 uV < amplitude < 70 uV,
 duration between 0.5 and 2 sec
 Apply limits on delta power: delta < 20% in 0.35-2 Hz band
 Apply limits on delta amplitude: 75 uV < amplitude < 300 uV,
 duration > 0.6 sec
 Classify as an artifact if amplitude > 300 uV
 Return 30 sec epochs = N2
Else if delta > 6 sec with delta frequency between 0.3 abd 2 Hz
 Apply limits on delta amplitude: 75 uV < amplitude < 300 uV,
 duration > 0.6 sec
 Classify as an artifact if amplitude > 300 uV
 if SpindleAnalysis(EEG signal) return = true
 Apply limits on sigma power: rhythm frequency between 11.5 and 14.5 Hz
 Apply limits on sigma amplitude: 10 uV < amplitude < 70 uV,
 duration between 0.5 and 2-sec Return 30 sec epochs = N3
Else if theta > 15 sec, with theta frequency between 3.5 and 7.5 Hz
 Apply limits on theta power: theta > 60 uV
 Apply limits on amplitude: 25 uV < amplitude < 100 uV, duration > 1 sec
 if SpindleAnalysis(EEG signal) return = False
 if DeltaAnalysis(EEG signal) return = False
 Apply limits on eye movement: EOG eye movement duration
 between 0.09 sec < x < 500 ms
 Ensure REM episodes are separated by at least 5 sec
 Apply limits on beta power: beta frequency between 15 and 30 Hz,
 0 uV < amplitude < 25 uV
End If
Return modified EEG signal
```

B.2.2 Top 10 features for LASSO, Shap and SA

B.2.3 Sample description for each paradoxical Insomnia formula

	Dataset	.....	meanA1	meanA0	meanB1	meanB0	meanC1	meanC0
Age(yo)	46 ± 12		46 ± 12	46 ± 13	40 ± 10	47 ± 13	43 ± 12	46± 13
Wo(%)	66		69	64	77	65	75	66
TST(min)	355± 75		362± 73	347 ± 76	416 ± 68	346 ± 72	418 ± 61	347 ± 73
ISI(score)	19.7±4.75		20 ± 4	19.4 ± 5	19.4 ± 4	19.7± 5	19.2 ± 4	19.7± 5
	meanD1	meanD0	meanE1	meanE0	meanF1	meanF0	meanJ1	meanJ0
Age(yo)	43 ± 11	47 ± 13	43 ± 12	46± 13	45 ± 12	46± 13	41 ± 11	47 ± 13
Wo(%)	72	65	80	66	77	64	74	65
TST(min)	409 ± 59	338 ± 71	434± 65	349 ± 73	403 ± 56	344 ± 75	440 ± 40	337± 68
ISI(score)	19.6 ± 4	19.7 ± 5	20.4 ± 4	19.6 ± 5	20.4 ± 4	19.5± 5	19.2± 4	19.8 ± 5
	meanK1	meanK0	meanL1	meanL0	meanL21	meanL20	meanM1	meanM0
Age(yo)	44 ± 11	46 ± 13	47± 13	45 ± 13	47 ± 12	45± 13	49 ± 10	45 ± 13
Wo(%)	74	66	69	66	69	66	70	66
TST(min)	460± 67	348 ± 71	386 ± 70	343 ± 74	374 ± 67	343 ± 74	373 ± 79	354± 75
ISI(score)	20.8± 3	19.3 ± 5	20.7 ± 4	19.3± 5	20.6 ± 4	19.3±5	20.6 ± 3	19.7 ± 5
	meanN1	meanN0	meanO1	meanO0	meanP1	meanP0	meanQ1	meanQ0
Age(yo)	42 ± 11	47 ± 13	44 ± 12	48 ± 13	44 ± 10	46± 13	43 ± 11	47 ± 13
Wo(%)	71	65	73	61	74	66	76	65
TST(min)	410± 57	335 ± 71	413 ± 44	301 ± 55	443 ± 47	343 ± 70	430 ± 39	339 ± 71
ISI(score)	19.4 ± 4	19.8 ± 5	19.9 ± 4	19.5± 5	20.3 ± 3	19.6±5	20.4 ± 4	19.5 ± 5
	meanR1	meanR0	meanT1	meanT0	meanV1	meanV0	meanZ1	meanZ0
Age(yo)	42 ± 11	46 ± 13	41 ± 11	47 ± 13	44 ± 10	46± 13	41 ± 11	47 ± 13
Wo(%)	78	64	74	65	77	66	78	65
TST(min)	448± 48	348 ± 72	440 ± 40	337 ± 68	399 ± 65	350 ± 75	420 ± 41	347 ± 74
ISI(score)	21.5 ± 4	19.5 ± 5	19.2 ± 4	19.8± 5	18.2 ± 5	19.9±5	18.7 ± 4	19.8 ± 5

Table B.1: Comparisons for each formula implemented in our dataset of the mean and the standard deviation for the Age, the Total Sleep Time recorded with the polysomnography, the score on the Index of Severity of Insomnia (ISI) and the percentage of women in each subgroup. For each formula the subgroup corresponding to the Paradoxical Insomnia group is labeled $mean_formula_1$

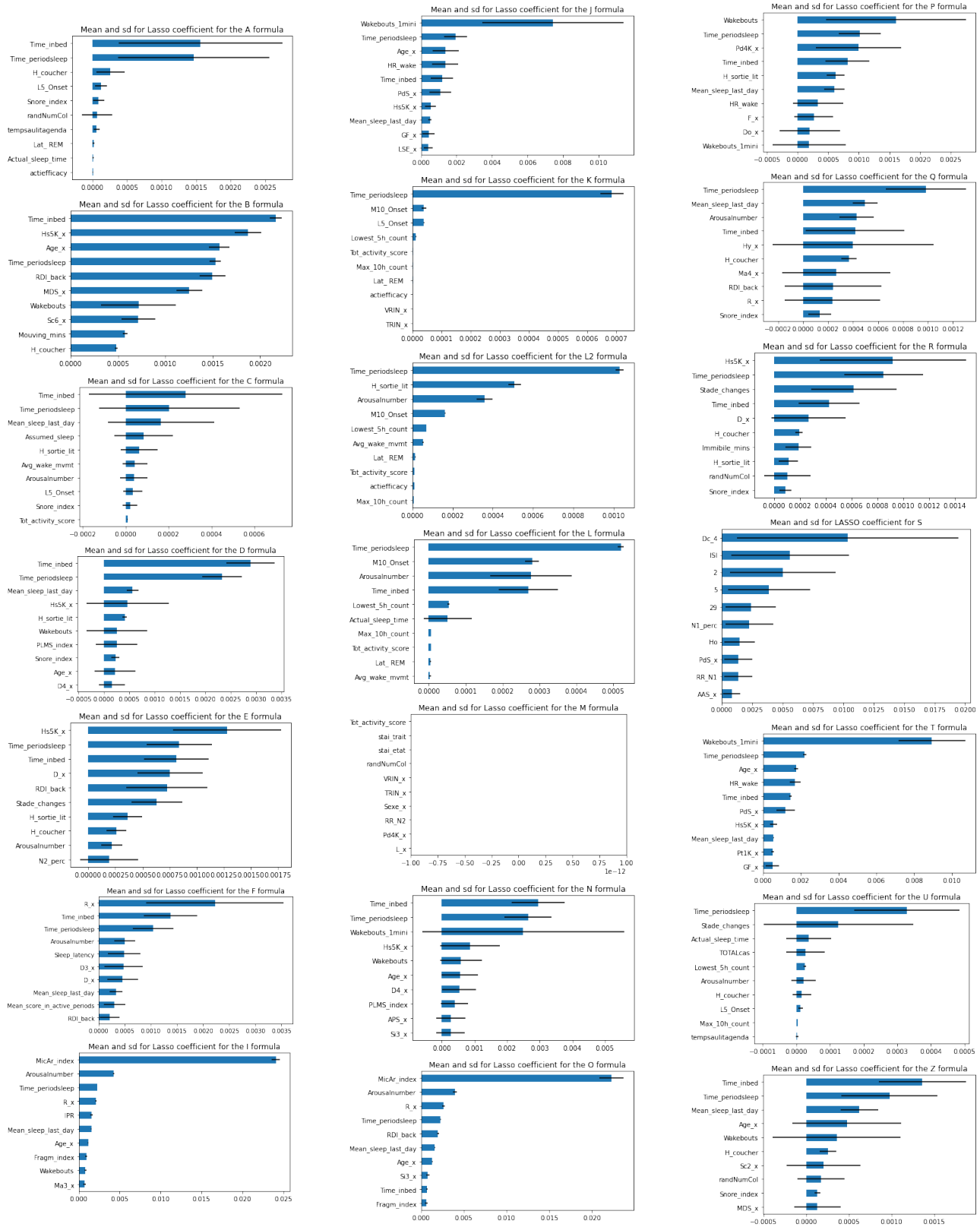


Figure B.1: Top 10 features selected by LASSO regularization

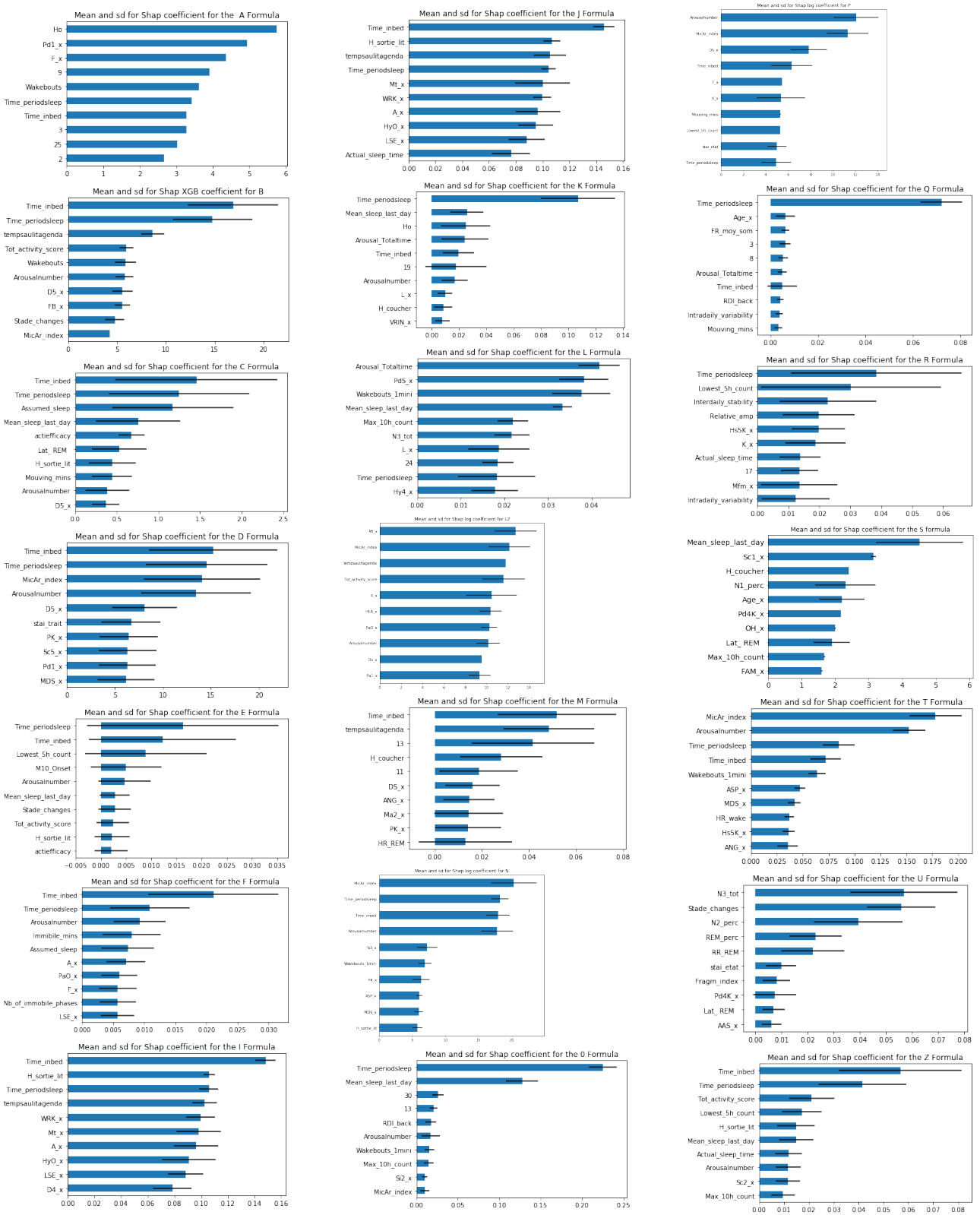


Figure B.2: Top 10 features selected by Shap

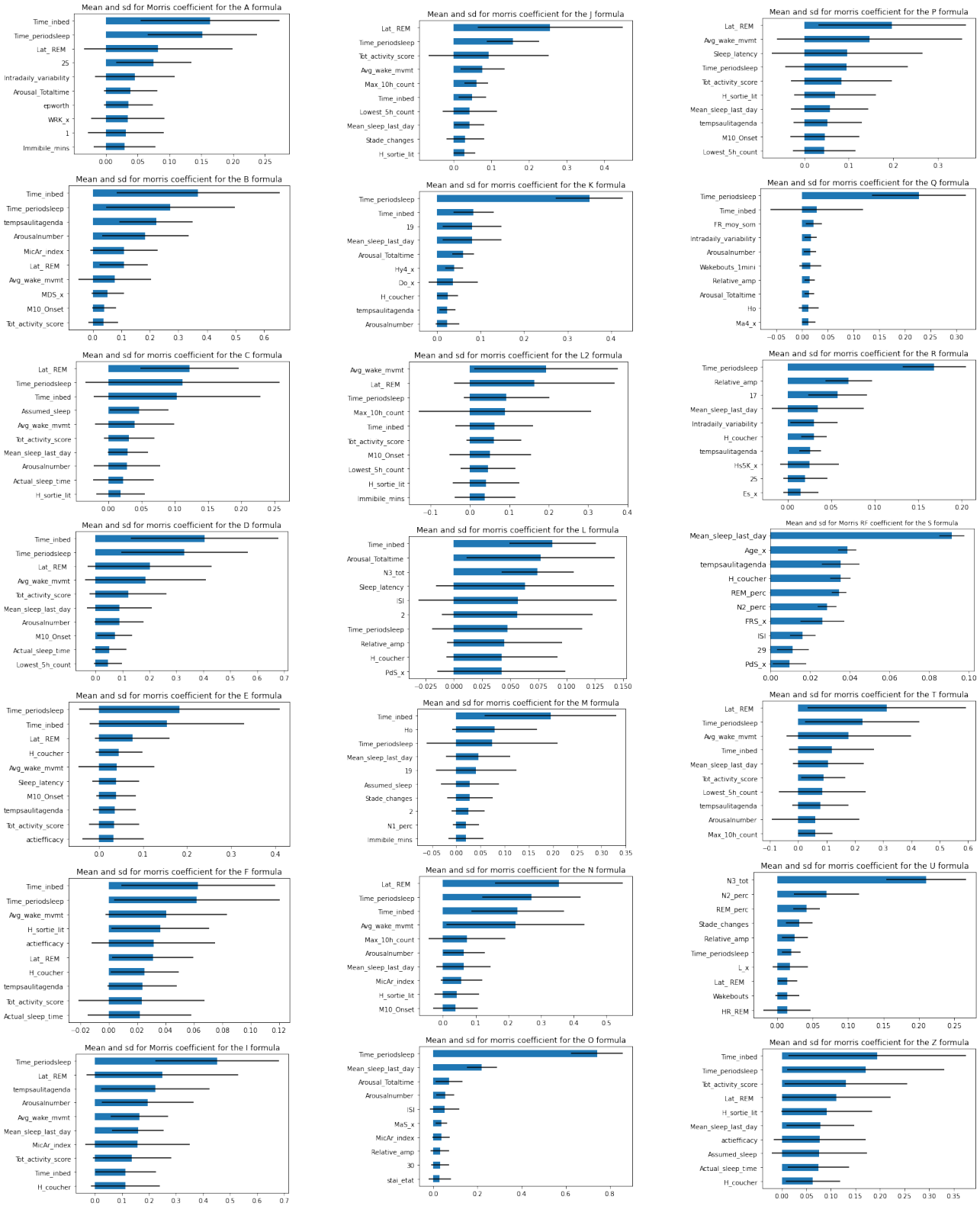


Figure B.3: Top 10 features selected by sensitivity analysis

B.3 Data

B.3.1 DI-PSYCH features definitions

DI-PSYCH			
Feature	**Definition**	**Feature**	**Definition**
A	Anxiety	AAS	Addiction Admission
Age	yo	ANX	Anxiety
ANG	Anger	APS	Addiction Potential
ASP	Antisocial Practices	BIZ	Bizarre Mentation
CYN	Cynicism	D	Depression
D1	Subjective Depression	D2	Psychomotor Retardation
D3	Physical Malfunctioning	D4	Mental Dullness
D5	Brooding	DEP	Depression
DO	Depression Objective	Do	Dominance
DS	Depression subjective	Es	Ego Strength
F	Infrequency	FAM	Family Problems
FB	Back	FRS	Fears
Fb	Back F	Gender	two categories
GF	Gender Role – Feminine	GM	Gender Role – Masculine
HEA	Health Concerns	Hs5K	Hypochondriasis (Hs)
Hy	Hysteria (Hy)	Hy1	Denial of Social Anxiety
Hy2	Need for Affection	Hy3	Lassitude-Malaise
Hy4	Somatic Complaints	Hy5	Inhibition of Aggression
HyO	Hysteria Objective	HyS	Hysteria Subjective
ID	Coded key	K	Correction
L	Lie	LSE	Low Self-Esteem
MACR	MAC-R- (Alcool)	Ma1	Amorality
Ma2	Psychomotor Acceleration	Ma2K	Hypomania (Ma)
Ma3	Imperturbability	Ma4	Ego Inflation
MaO	Hypomania Objective	MaS	Hypomania subjective
MDS	Marital Distress	Mfm	Masculinity-Femininity (Mf)
Mt	College Maladjustment	OBS	Obsessiveness
OH	Overcontrolled Hostility	P	Psychopathic Deviate (Pd)
Pa	Paranoia (Pa)	Pa1	Pa1- Persecutory Ideas
Pa2	Pa2- Poignancy	Pa3	Pa3- Naïveté
PaO	Paranoia Objective	PaS	Paranoia subjective
Pd1	Familial Discord	Pd2	Authority Problems
Pd3	Social Imperturbability	Pd4	Social Alienation
Pd4K	Psychopathic Deviate (Pd)	Pd5	Self-Alienation
PdO	Psychopathic	PdS	Psychopathic subjective
PK	Post-Traumatic Stress Disorder	PS	Post-Traumatic Stress Disorder
Pt1K	Psychasthenia (Pt)	R	Repression
Re	Social Responsibility	Sc1	Social Alienation
Sc1K	Schizophrenia (Sc)	Sc2	Emotional Alienation
Sc3	Lack of Ego Mastery-Cognitive	Sc4	Lack of Ego Mastery-Conative
Sc5	Lack of Ego Mastery-Defective Inhibition	Sc6	Bizarre Sensory Experiences
Si	Social Introversion (Si)	Si1	Shyness / Self-Consciousness
Si2	Social Avoidance	Si3	Alienation- Self and Others
SOD	Social Discomfort	TPA	Type A
TRIN	True Response Inconsistency	TRT	Negative Treatment Indicators
VRIN	Variable Response Inconsistency	WRK	Work Interference

Table B.2: DATABASE II (PSYCH) extracted from MMPI-2 database: 1182 samples; 91 features

B.3.2 DII-QUEST features definitions

DII-QUEST			
Feature	**Definition**	**Feature**	**Definition**
ID	Coded key	Gender	2 categories
Age	yo	SES	Work/marital status (4 categories)
ISI	0-28 score	ESS	0-24 score
H0	16-86 score	STAI_T	20-80 score
STAI_S	20-80 score	BDI_2	0-63 score
DBAS1	0-10 score	DBAS2	0-10 score
DBAS3	0-10 score	DBAS4	0-10 score
DBAS5	0-10 score	DBAS6	0-10 score
DBAS7	0-10 score	DBAS8	0-10 score
DBAS9	0-10 score	DBAS10	0-10 score
DBAS11	0-10 score	DBAS12	0-10 score
DBAS13	0-10 score	DBAS14	0-10 score
DBAS15	0-10 score	DBAS16	0-10 score
DBAS17	0-10 score	DBAS18	0-10 score
DBAS19	0-10 score	DBAS20	0-10 score
DBAS21	0-10 score	DBAS22	0-10 score
DBAS23	0-10 score	DBAS24	0-10 score
DBAS25	0-10 score	DBAS26	0-10 score
DBAS27	0-10 score	DBAS28	0-10 score
DBAS29	0-10 score	DBAS30	0-10 score
DBASTotal	0-300 score	TT_outcome	2 categories
TT_ent	8 categories, cumulative	Dc_ent	9 categories, cumulative
TT_out	8 categories, cumulative		

Table B.3: DATABASE II (QUEST) with 713 samples with DBAS scale scores and 519 with the others features. 45 features. DBAS questions from 1 to 30, Treatment outcome positive or negative at least six months after the assessment, Treatment taken during the assessment, Comorbid Diagnostic (treated or not) at the time of the assessment, and Treatment given after the assessment

B.3.3 DBAS questionnaire

Several statements reflecting people's beliefs and attitudes about sleep are listed below.
Please indicate to what extent you personally agree or disagree with each statement.
There is no right or wrong answer. For each statement, circle the number that corresponds to your own *personal belief*. Please respond to all items even though some may not apply directly to your own situation.

		Strongly Disagree ← → Strongly Agree
1.	I need 8 hours of sleep to feel refreshed and function well during the day.	1 2 3 4 5 6 7 8 9 10
2.	When I don't get the proper amount of sleep on a given night, I need to catch up on the next day by napping or on the next night by sleeping longer.	1 2 3 4 5 6 7 8 9 10
3.	Because I am getting older, I need less sleep.	1 2 3 4 5 6 7 8 9 10
4.	I am worried that if I go for 1 or 2 nights without sleep, I may have a "nervous breakdown."	1 2 3 4 5 6 7 8 9 10
5.	I am concerned that chronic insomnia may have serious consequences on my physical health.	1 2 3 4 5 6 7 8 9 10
6.	By spending more time in bed, I usually get more sleep and feel better the next day.	1 2 3 4 5 6 7 8 9 10
7.	When I have trouble falling asleep or getting back to sleep after nighttime awakening, I should stay in bed and try harder.	1 2 3 4 5 6 7 8 9 10
8.	I am worried that I may lose control over my abilities to sleep.	1 2 3 4 5 6 7 8 9 10
9.	Because I am getting older, I should go to bed earlier in the evening.	1 2 3 4 5 6 7 8 9 10
10.	After a poor night's sleep, I know that it will interfere with my daily activities on the next day.	1 2 3 4 5 6 7 8 9 10
11.	In order to be alert and function well during the day, I believe I would be better off taking a sleeping pill rather than having a poor night's sleep.	1 2 3 4 5 6 7 8 9 10
12.	When I feel irritable, depressed, or anxious during the day, it is mostly because I did not sleep well the night before.	1 2 3 4 5 6 7 8 9 10
13.	Because my bed partner falls asleep as soon as his/her head hits the pillow and stays asleep through the night, I should be able to do so too.	1 2 3 4 5 6 7 8 9 10
14.	I feel that insomnia is basically the result of aging and there isn't much that can be done about this problem.	1 2 3 4 5 6 7 8 9 10

		Strongly Disagree ← → Strongly Agree
15.	I am sometimes afraid of dying in my sleep.	1 2 3 4 5 6 7 8 9 10
16.	When I have a good night's sleep, I know that I will have to pay for it on the following night.	1 2 3 4 5 6 7 8 9 10
17.	When I sleep poorly on one night, I know it will disturb my sleep schedule for the whole week.	1 2 3 4 5 6 7 8 9 10
18.	Without an adequate night's sleep, I can hardly function the next day.	1 2 3 4 5 6 7 8 9 10
19.	I can't ever predict whether I'll have a good or poor night's sleep.	1 2 3 4 5 6 7 8 9 10
20.	I have little ability to manage the negative consequences of disturbed sleep.	1 2 3 4 5 6 7 8 9 10
21.	When I feel tired, have no energy, or just seem not to function well during the day, it is generally because I did not sleep well the night before.	1 2 3 4 5 6 7 8 9 10
22.	I get overwhelmed by my thoughts at night and often feel I have no control over this racing mind.	1 2 3 4 5 6 7 8 9 10
23.	I feel I can still lead a satisfactory life despite sleep difficulties.	1 2 3 4 5 6 7 8 9 10
24.	I believe insomnia is essentially the result of a chemical imbalance.	1 2 3 4 5 6 7 8 9 10
25.	I feel insomnia is ruining my ability to enjoy life and prevents me from doing what I want.	1 2 3 4 5 6 7 8 9 10
26.	A "nightcap" before bedtime is a good solution to sleep problem.	1 2 3 4 5 6 7 8 9 10
27.	Medication is probably the only solution to sleeplessness.	1 2 3 4 5 6 7 8 9 10
28.	My sleep is getting worse all the time and I don't believe anyone can help.	1 2 3 4 5 6 7 8 9 10
29.	It usually shows in my physical appearance when I haven't slept well.	1 2 3 4 5 6 7 8 9 10
30.	I avoid or cancel obligations (social, family) after a poor night's sleep.	1 2 3 4 5 6 7 8 9 10

Figure B.4: DBAS questionnaire with 30 questions scored from 1 to 10 [147]

B.3.4 DIII-PSG features definitions

DIII-PSG	
Features	**Definition**
ID	Coded Key
Arousalnumber	Total arousal number (Microarousal + Wakebouts)
AHI	Apnea−Hypopnea Index
HRmoyN1	Average Heath Rate in N1
HRmoyN2	Average Heath Rate in N2
HRmoyN3	Average Heath Rate in N3
HRmoyREM	Average Heath Rate in REM
HRmoywake	Average Heath Rate during wake
LatREM	Latency to REM
MicArPLMindex	Mircorousal index secondary to Periodic limb movement
NREM	Non−rapid eye movement sleep (N1,N2 and N3)
N1SL	stage−1 latency
N1perc	Percentage of N1
N2SL	sleep−onset latency + stage−2 latency
N2perc	Percentage of N2
N3tot	Percentage of N3
PLMindex	Periodic limb movement index
RDI	Respiratory Disturbance Index, including apneas and hypopneas
RDIback	Respiratory Disturbance Index in the supine position
RDInotback	Respiratory Disturbance Index in non supine position
REM	Rapid eye movement sleep
REMperc	Percentage of REM
RESP	Respiration (the rate of breathing)
RF	Respiratory rate during sleep
RREM	Heart Rate variation in REM
RRN1	Heart Rate variation in N1
RRN2	Heart Rate variation in N2
SE	Sleep Efficiency
SOL	Sleep Onset Latency
SpO2	Blood oxygen saturation
Stade Changes	Total number of stades changes during the sleep episode
SWS	Slow−Wave Sleep
TIB	Time In Bed
TPS	Time Period of Sleep
TST	Total Sleep Time
WASO	Wake After Sleep Onset
Wakebouts	Number of awakening ($>$ 15 sec and $<$ 60 s)
Wakebouts1mini	Number of awakening $\geq$ 1 minute

Table B.4: DATABASE III (PSG) PSG features-578 samples, 38 features

Justification for keeping the respiratory events and periodic limb movements

1. As we saw in section 2.4.2, an a-priori categorical approach to insomnia seems to create more confusion than real diagnostic criteria. Moreover, although there is evidence that a high number of sleep apneas or periodic movements can impact sleep fragmentation, our clinical experience has taught us that there is not necessarily a correlation between sleep perception and microsleep fragmentation. So as we have the origin of each microarousals, we wanted to see the influence of such events on sleep perception and the treatment outcome.
2. We are considering the complaints of insomnia first, of course, provided that the patient has no complaint neither of his legs nor his nocturnal breathing, and even more if he is already treated for it.
3. This dimensional approach is highly compatible with machine learning algorithms, which learn from data and can detect whether poorly or insufficiently treated respiratory events or periodic movements have a role to play in sleep perception or response to treatment. It might even be said that knowing this in the context of resistant insomniacs would be desirable. The aim of our work is precisely to take into account as much data as possible.
4. Furthermore, cluster studies like [19] lack objective data, meaning that it is possible that the subjects included also suffer from respiratory disorders or PMJ, so we consider that retaining this data coupled with ML tools could enable us to understand the links between OSA, PMJ, and insomnia from a new angle.

B.3.5 DIV-AG features definitions

DIV-AG	
Features	**Definition**
ID	Coded key
Time in bed	Time between “Lights Out” and “Got Up”
Assumed sleep	total elapsed time between “Fell Asleep” and “Woke Up” times.
Actual sleep time	total time spent in sleep
Actual sleep (%)	expressed as a percentage of assumed sleep time.
Actual wake time	total time spent in wake
Actual wake (%)	Actual wake time as a percentage of assumed sleep time.
Sleep efficiency (%)	Actual sleep time as a percentage of time in bed.
Sleep latency	time between “Lights Out” and “Fell Asleep.”
Sleep bouts	several adjacent sections categorized as sleep
Wake bouts	several adjacent sections categorized as wake.
Mean sleep bout	average length of each of sleep bouts.
Mean wake bout	average length of each of the wake bouts.
Immobile mins	total time categorized as Immobile
Immobile time (%)	immobile time expressed as a percentage of assumed sleep time.
Mobile mins	total time categorized as mobile
Mobile time (%)	mobile time expressed as a percentage of assumed sleep time.
Immobile bouts	several adjacent sections are categorized as immobile
Mean immobile bout	average length of each of immobile bouts.
Immobile bouts <=1min	number of immobile bouts less 1 minute
Immobile bouts <=1min (%)	expressed as a percentage of the total number of immobile bouts.
Total activity score	total of all activity counts during the assumed sleep period.
Mean activity /epoch	total activity score divided by epochs in the assumed sleep period.
Fragmentation Index	sum of “Mobile time (%)” and “Immobile bouts 1 min (%).

Table B.5: DATABASE IV (AG) extracted from actigraphic features: 350 samples, 27 features

B.3.6 SLEEP LOG features definitions

Sleep log features	
Features	**Definition**
Bedtime	“Lights Out”
Get up time	“Lights on” and definitive get up from bed
TSTs	subjective total sleep time
Time spent outside the bed	declarative time spent outside bed.
total wake time	total time spent in wake estimated
SOLs	subjective sleep onset latency
Wake bouts estimated	subjective count of period categorised as wake.

Table B.6: Features extracted from sleep-log, seven features

B.4 Illustrations

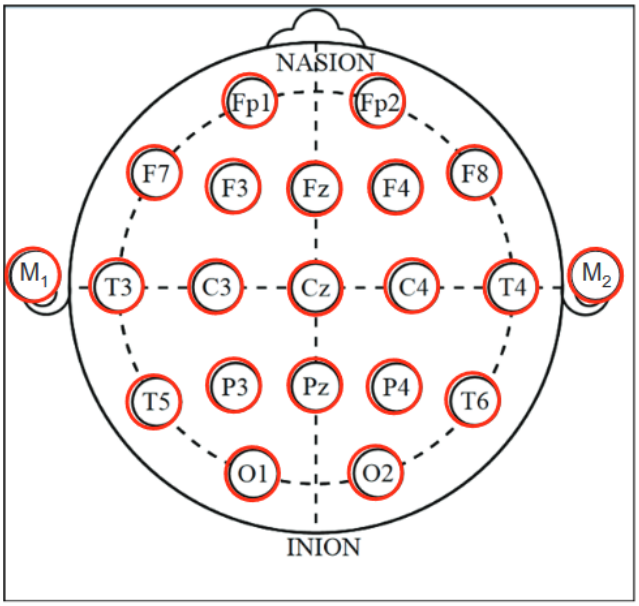


Figure B.5: 10−20 system

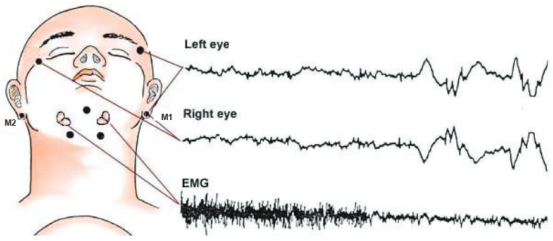


Figure B.6: Location of EOG and EMG

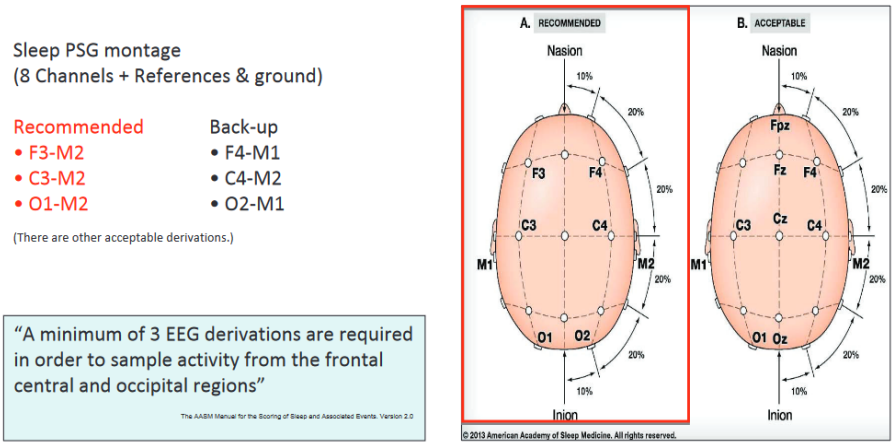


Figure B.7: Recommended EEG Sleep PSG montage

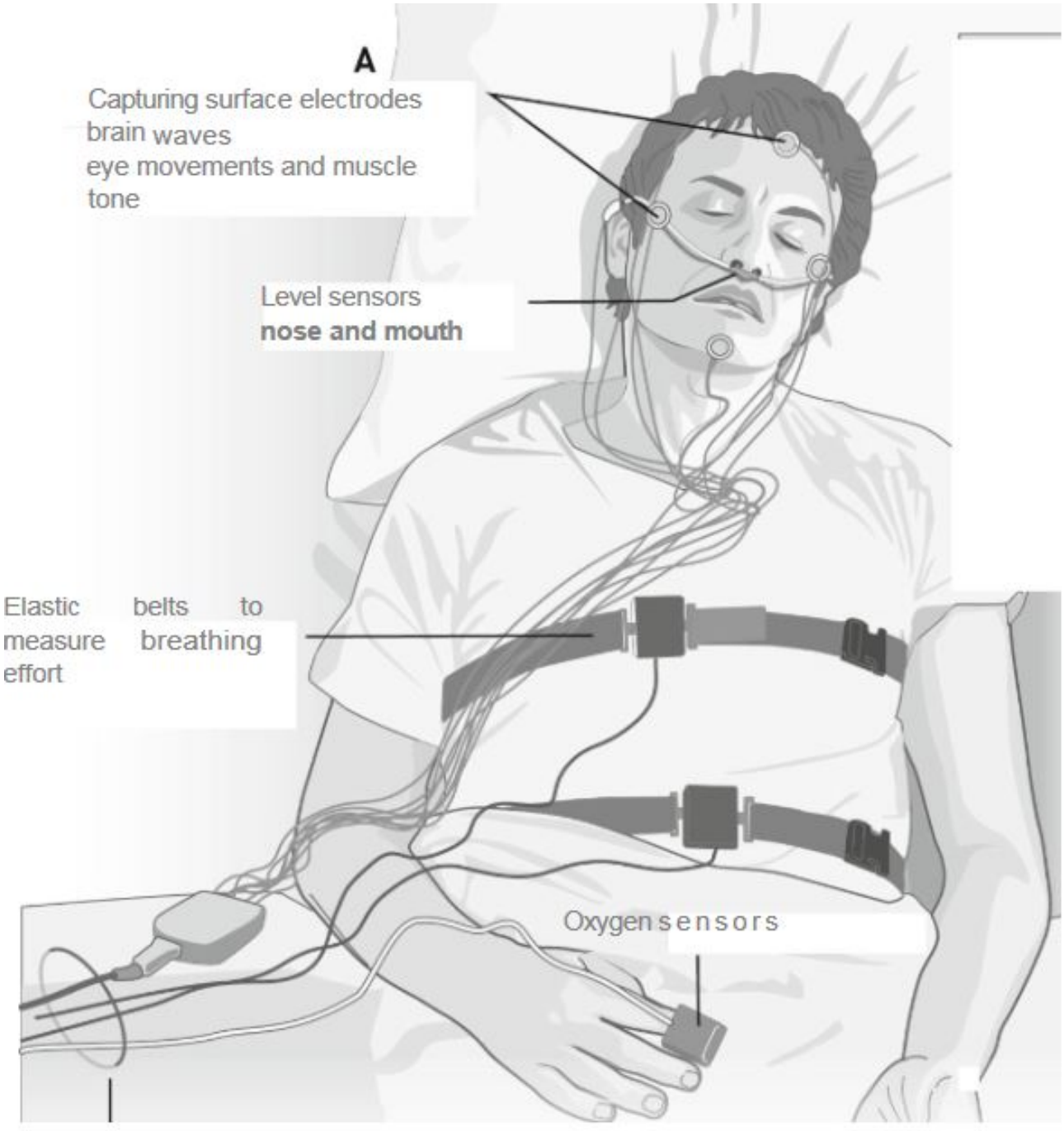


Figure B.8: PSG illustration adapted from [**pallanca**]

B.5 Supplement Material to find out more about Paradoxical Insomnia Concept

History of Paradoxical Insomnia across the differents international classifications

1. The **first classification** approach, the **DCSAD** [190] was mainly based on symptoms with a detailed description of the disorders and their semiology. In this first classification, insomnia is classified under the banner *Disorders of Initiating and Maintaining Sleep* or DIMS. Nineteen sub-types of the current concept of chronic insomniac were described dispatched in nine categories (see Table B.7 in the Appendix). Each category was extensively described by detailed clinical descriptions, useful to the clinician but lacking systematization. It should also be understood that it was at this time that the ICD-9 was published under the aegis of the World Health Organization (WHO) to list and categorise all known diseases. But, as the United States did not recognize it in the '70s, it was revised in the '80s, especially to be adapted to US constraints.
2. Thus, the **second classification**, **ICSD**, published in 1990 [198] began to build bridges with the revised ICD-9 (ICD-9-CM), and some categories described in the DCSAD were removed from the section "Insomnia". So although the ICSD was published as a continuation of the DCSAD, the authors used a multiaxial system for stating and coding diagnoses in clinical reports and database purposes. The main difference is that sleep disorders were classified according to presumed pathophysiological mechanisms and not centred on the main symptoms. Insomnia disorder is then classified as extrinsic and intrinsic dyssomnia (see Table B.7 in the Appendix). The sleep disorders associated with medical or psychiatric conditions were transferred to a new third section and replaced the categories 2a-c and 3d. Categories 8a and 8b describing atypical PSG features were described in the Axis-B, which comprises the ICD-9-CM's classification of procedures. So the remaining Insomnia category is then divided into Dyssomnias-Sleep Disorder Intrinsic or extrinsic (see Table B.7 in the Appendix). Intrinsic must be understood as primarily sleep disorders that either originate or develop within the body or arise from causes within the body. The list of intrinsic sleep disorders includes multiple items, such as PsyI, nSSM, and idiopathic insomnia, primarily producing insomnia. **So this is the first time that insomnia is recognised as a disorder in itself**, implicitly naming the direct consequence of a disturbance of the sleep-wake systems or primary. Two of the three categories of insomnia meeting this criterion were already present under the same name in the DCSAD, except category 9b, corresponding to *Subjective DIMS Complaint without Objective Findings* - which becomes nSSM. This effort to classify disorders jointly continued with the **third classification**, the **ICSD-2** in 2005 [192], which extended the relationship to the new, 10th edition of the ICD and the DSM-4-TR. This is the first classification with a chapter called Insomnia, and in this movement of harmonisation of diagnoses, part of the disorders previously described is referred either to the ICD-10 when it comes to somatic pathologies or the DSM-IV-TR when it comes to mental pathologies. However, the classification remains relatively close to that of the ICSD and ICSD-R (see Table B.7, with ten categories of CID that are broadly similar to those of the ICSD-R. We find the notion of primary insomnia with the same names as PsyI and idiopathic insomnia but a name change for nSSM, which becomes ParI. The notion of secondary insomnia is mainly reflected in five categories whose causal character is mentioned by "Due to", either a mental disorder, drug/substance, or medical condition, Non-organic not otherwise specified, or organic not otherwise specified. Although this classification has the merit of pooling all the categories with insomnia as a common and principal complaint, when reading the diagnostic criteria carefully, one realises many fuzzy areas. For example, concerning insomnia due to mental disorders, the fourth diagnostic criterion states that insomnia must be more important than that typically associated with the mental disorders, which leaves a wide latitude of interpretation both in terms of the diagnosis and in terms of the treatment to be introduced depending on the psychiatric disorder. Moreover, it can

be observed in this classification that there are still no mandatory validated objective criteria for the disorders described in the Insomnia chapter.

3. This lack of objectivity in the diagnostic criteria of the different categories of insomnia maintained until then has led to a drastic simplification of Insomnia description in the **fourth classification**, the **ICSD-3**, in 2014. Indeed, of the ten categories, only one remains, regrouping in an indistinct way all the categories described until now. This last evolution is shown in Tables 2.1 and 2.2. This classification is harmonized with the ICD-11 and DSM-5, which abandoned any attempt to categorize insomnia by presumed pathology and left aside the primary/secondary and organic/nonorganic dichotomies. The reason exposed was that there is too much overlap between primary and secondary insomnia symptoms. In addition, many people with insomnia have multiple medical and psychiatric comorbidities making causal attribution difficult. As a result, discrimination between these subtypes has proven difficult given their current definitions and available methods, and they have been removed. Then, only the subtype, CID, is now available. The diagnostic criteria from ICSD-3 are included in Appendix. Six criteria must be met from A to F, with criteria A and B related to the clinical symptoms, C to the sleep condition, D the frequency (at least three times a week), E the duration (at least three months), and F a mention on other sleep disorder that must be not the better explanation for insomnia.

A. The patient reports, or the patient's parent or caregiver observes, one or more of the following:[1]
   1. Difficulty initiating sleep.
   2. Difficulty maintaining sleep.
   3. Waking up earlier than desired.
   4. Resistance to going to bed on appropriate schedule.
   5. Difficulty sleeping without parent or caregiver intervention.

B. The patient reports, or the patient's parent or caregiver observes, one or more of the following related to the nighttime sleep difficulty:
   1. Fatigue/malaise.
   2. Attention, concentration, or memory impairment.
   3. Impaired social, family, occupational, or academic performance.
   4. Mood disturbance/irritability.
   5. Daytime sleepiness.
   6. Behavioral problems (e.g., hyperactivity, impulsivity, aggression).
   7. Reduced motivation/energy/initiative.
   8. Proneness for errors/accidents.
   9. Concerns about or dissatisfaction with sleep.

C. The reported sleep/wake complaints cannot be explained purely by inadequate opportunity (i.e., enough time is allotted for sleep) or inadequate circumstances (i.e., the environment is safe, dark, quiet, and comfortable) for sleep.

D. The sleep disturbance and associated daytime symptoms occur at least three times per week.

E. The sleep disturbance and associated daytime symptoms have been present for at least three months.[2]

F. The sleep/wake difficulty is not better explained by another sleep disorder.

Figure B.9: Diagnostic criteria for CID in ICSD3 [183]

DCSAD(1979)	ICSD−R[1990-1997]	ICSD−2(2005)	ICSD−3TR[2014-2023]
A−DIMS	1−Dyssomnias A−ISD B−ESD 3−SDA (with M,N, or 0)	Insomnia Primary and Secondary	Insomnia
1−Psychophysiological 1a−Persistent 2 + Psychiatric D 2a−Symptom & Personality D 2b−Affective D 2c−Other Functional Psychoses 3+ Use with Drug and OH 3a−Tolerance or Withdrawal from CNS Depressants 3b−Sustained Use of CNS Stimulants 3c−Sustained Use or Withdrawal from Other Drugs 3d−Chronic Alcoholism 4+ Sleep−Induced RI 5+ Sleep−Induced Myoclonus and RLS 6+ other Medical, Toxic, Environmental Conditions 7+ Childhood−Onset DIMS 8+ Other DIMS Conditions 8a−Repeated REM Sleep Interruptions 8b−Atypical PSG Features 8c−NOS 9+ No DIMS Abnormality 9a−Short Sleeper 9b−Subjective DIMS Complaint without Objective Findings 9c−NOS	1A1−Psychophysiologic I. 1B1−Inadequate sleep hygiene 3M3−Anxiety and Panic D 3M2−Mood D 3M1−Psychoses 1B10−Hypnotic−Dependent D 1B11−Stimulant−Dependent D Use Axis B 3M5−Alcoholism 1A13 Extrinsec Environmental Conditions Idiopathic Insomnia Use Axis B Use Axis B 4.1−Short Sleeper 1A2−Sleep State Misperception	Psychophysiological I. Insomnia due to Mental D Inadequate sleep hygiene I. due to drug or substance I. due to medical condition I. not due to a substance or known physiological condition, unspecified Physiological (organic) insomnia unspecified Idiopathic Insomnia Behavioral I. of Childhood Short Sleeper Paradoxical I.	Chronic I.D Short−Term I.D Other I.D Excessive TIB Short Sleeper

Table B.7: DIMS: Disorders of Initiating and Maintaining Sleep,+: Associated with, D: Disorder, U: Usage, OH: Alcohol, CNS: Nervous System, I.: Insomnia, Central, RI: Respiratory Impairment, RLS: Restless Legs Syndrome, PSG: Polysomnographic, NOT: Not Otherwise Specified, ISD: Intrinsic Sleep Disorders, ESD: Extrinsic Sleep Disorders, SDA: Sleep Disorders Associated, M: Associated with Mental Disorders, N: Associated with Neurologic Disorder, O: associated with Other Medical Disorders,

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
N3 stages	Lower 63.6 vs 92.0 min (Psycho−I)	17	[22]	±20	–	?
SOL	Lower 20 vs 45 min (Psycho−I)	17	[22]	±20	–	?
Personnality/cognitive features						
MMPI scales > 65	score globaly higher than Psycho−I Ma(69.6),Pt(67),Sc(68.6)	17	[22]	±20	–	?
Treatment features						
Progressive relaxation Training	No efficient	17	[22]	± 20	–	?

Table B.8: Formula A explanation (compared to psychopgysiological Insomnia)− N= number of ParI subjects, Au = Authors,T = Presence of treatment (+ =yes,−=No),W% = percentage of women

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
TST	Higher 420 vs 379.0 min	8	[195]	32(10)	–	75
N2 stages	Higher 61.9 vs 53.6 %	8	[195]	32(10)	–	75
Personality/cognitive features						
AVT omissions	score globally higher 3h after awakenings	8	[195]	32(10)	–	75

Table B.9: **Formula B explanation** (compared to objective DIMS)− N= number of ParI subjects, Au = Authors,T = Presence of treatment (+ =yes,−=No),W% = percentage of women,AVT=Auditory Vigilance Task

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
NA		8	[109]	32(12)	–	50

Table B.10: **Formula C explanation** (compared to objective DIMS)− N= number of ParI subjects, Au = Authors,T = Presence of treatment (+ =yes,−=No),W% = percentage of women,

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
TST	Higher 373 vs 339(1) vs 313(2) min	8	[85]	45(24−69)	–	66
Actigraph						
TST	Lower 337 vs 364(1) vs 341(2) min	8	[195]	45(24−69)	–	66

Table B.11: **Formula D explanation** (compared to objective (1)=Psychophysiological Insomnia,(2)=Psychiatric Insomnia)− N= number of ParI subjects, Au = Authors,T = Presence of treatment (+ =yes,−=No),W% = percentage of women,(1)=Psychophysiological I,(2)=Psychiatric Insomnia

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
TST	Higher 456 vs 418(1) vs 452 (2) min	7	[180]	35(6)	–	45
Wake N	Lower 3.5 vs 7.2 (1) vs 3.5 (2)	7	[180]	35(6)	–	45
N2 stages	Lower 55 vs 59(1) vs 52 (2) %	7	[180]	35(6)	–	45
N3 stages	Higher 12 vs 7(1) vs 15 (2) %	7	[180]	35(6)	–	45
cognitive features						
MMPI scales > 65	score higher than (1) and (2) Pd(70) and Hy(65)	7	[180]	35(6)	–	45

Table B.12: **Formula E explanation** (compared to objective (1)=Psychophysiological Insomnia, (2)=Psychiatric Insomnia)– N= number of ParI subjects, Au = Authors, T = Presence of treatment (+ = yes, – = no), W% = percentage of women, (1) = other Insomnia, (2) = Control

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
TST	Higher 366 vs 288(1) min	9	[138]	35(6)	–	65
SOL	Lower 9.6 vs 25.3(1) min	7	[138]	35(6)	–	65

Table B.13: Formula F explanation
(compared to objective (1)=Psychophysiological Insomnia,– N= number of ParI subjects, Au = Authors,T = Presence of treatment (+ =yes,-=No),W% = percentage of women)

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
TST	Higher 451 vs 433(1) min	9	[21]	31.7(8)	–	40
Personnality/cognitive features						
MMPI scales > 65	score higher than (1)	9	[21]	31.7(8)	–	40

Table B.14: **Formula G explanation** (compared to (1)= Controls)– N= number of ParI subjects, Au = Authors,T = Presence of treatment (+ =yes,–=No),W% = percentage of women

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
TST	equal 414 vs 392(1)vs 303(2) min	9	[55]	20(18–25)	–	30
SLN2	lower 15.6 vs 61.6(1)vs 40(2) min	9	[55]	20(18–25)	–	30
awake N	equal 2.2 vs 1.7 (1)vs 2(2)	9	[55]	20(18–25)	–	30
Stage 1/2	lower/equal 6/60 vs 12/60 (1) vs 10/65(2) %	9	[55]	20(18–25)	–	30
Stage N3	Higher 62 vs 31 (1)vs 40 (2)	9	[55]	20(18–25)	–	30
cognitive features						
EPI	NS	9	[55]	20(18–25)	–	30

Table B.15: **Formula H explanation**(compared to objective (1)=Psychophysiologic Insomnia,(2)=Control,– N= number of ParI subjects, Au = Authors,T = Presence of treatment (+ =yes,–=No),W% = percentage of women, EPI = Eysenck Personality Inventory

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
TPS (min)	463 vs 442(1)vs 439(2) vs 453(3)	22	[58]	57(10)	–	55
TST (min)	392 vs 330(1)vs 342(2) vs 399(3)	22	[58]	57(10)	–	55
Questionnaires features						
STAI trait	NS	22	[58]	57(10)	–	55
BDI	NS	22	[58]	57(10)	–	55
DBAS	NS	22	[58]	57(10)	–	55

Table B.16: **Formula I explanation** compared to objective (1)=Psychophysiologic Insomnia,(2)= Control with SSM, (3) = Control without SSM,– N= number of ParI subjects, Au = Authors,T = Presence of treatment (+ =yes,–=No),W% = percentage of women, BDI=Beck Depression Incentory, STAI=State–Trait Anxiety Inventory, DBAS=Dysfunctional Beliefs and Attitudes About Sleep

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
TPS (min)	463 vs 442(1)vs 439(2) vs 453(3)	22	[58]	57(10)	–	55
TST (min)	392 vs 330(1)vs 342(2) vs 399(3)	22	[58]	57(10)	–	55
	Higher 427 vs 303 (1)vs 395 (4)	12	[108]	56(12)	–	66
REM (min)	Higher 112 vs 63(1)vs 84(4)	12	[108]	56(12)	–	66
SOL(min)	equal 12 vs 19(1)vs 16(4)	12	[108]	56(12)	–	66
Questionnaires features						
STAI trait	NS	22	[58]	57(10)	–	55
BDI	NS	22	[58]	57(10)	–	55
DBAS	NS	22	[58]	57(10)	–	55
EEGq features in NREM						
Mean and DS of RSP						
Delta	Low/(4) 67 vs 70(1) vs 74(4)	12	[108]	56(12)	–	66
Theta	equal 16 vs 15(1)vs 14(4)	12	[108]	56(12)	–	66
Alpha	equal 9 vs 8(1)vs 7(4)	12	[108]	56(12)	–	66
Sigma	Higher 5.6 vs 4.5(1)vs 3.4 (4)	12	[108]	56(12)	–	66
Beta	Higher 2.4 vs 1.9(1)vs 1.7(4)	12	[108]	56(12)	–	66
Gamma	equal 0.4 vs 0.4(1)vs 0.4(4)	12	[108]	56(12)	–	66

Table B.17: **Formula J explanation** compared to objective (1)=Objective Insomnia,(2)= Control with SSM,(3)= Control Without SSM, (4) = Control Global– N= number of ParI subjects, Au = Authors,T = Presence of treatment (+ =yes,–=No),W% = percentage of women, BDI=Beck Depression Inventory, STAI=State–Trait Anxiety Inventory, DBAS=Dysfunctional Beliefs and Attitudes About Sleep,RSP:Relative Spectral power,Relative power was computed as the power within a frequency band (in micV2 /Hz) divided by the power across all frequencies (0.5–60 Hz) (also in micV2 /Hz) and is therefore dimensionless. In this table the values are multiplied by 100, indicating the percentage of power in the frequency band of interest

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
TPS (min)	415 vs 424(1)	17	[152]	41(7)	–	80
TST (min)	356 vs 376 (1)	10	[117]	41(7)	–	80
N3 (min)	Lower 60 vs 87(1)	10	[117]	41(7)	–	80
Awakenings	equal 19.75 vs 19.70 (1)	10	[117]	41(7)	–	80

Table B.18: **Formula S explanation** compared to (1) Healthy Subject− N= number of ParI subjects, Au = Authors,T = Presence of treatment (+ =yes,−=No),W% = percentage of women

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
TST (min)	410 vs 395 (1) vs 415 (2)	26	[94]	41(10)	–	55
N3 dur (min)	41 vs 29 (1) vs 31 (2)	26	[94]	41(10)	–	55

Table B.19: **Formula Q explanation** compared to Formula Q explanation compared to objective (1)= Psychophysiological Insomnia (2)= Good Sleepers, N= number of ParI subjects, Au = Authors,T = Presence of treatment (+ =yes,−=No),W% = percentage of women

Features Origin	Findings	N	Au	Age(σ)	T	W%
PSG features						
TPS (min)	415 vs 424(1)	10	[117]	41(7)	–	80
TST (min)	356 vs 376 (1)	10	[117]	41(7)	–	80
N3 (min)	Lower 60 vs 87(1)	10	[117]	41(7)	–	80
Awakenings	equal 19.75 vs 19.70 (1)	10	[117]	41(7)	–	80

Table B.20: **Formula S explanation** compared to (1) Healthy Subject− N= number of ParI subjects, Au = Authors,T = Presence of treatment (+ =yes,−=No),W% = percentage of women

Appendix C

List of publications

C.1 Refeered Journal Paper

* Agrigoroaie Roxana, Pallanca Olivier, Tapus Adriana, Impact of Insomnia and User Profile on Cognitive Performance, IEEE Transactions on Affective Computing Under Review
* Pallanca Olivier, Read Jesse, General principles and definitions in artificial intelligence,Archives des maladies du coeur et des vaisseaux Pratique Vol 2021 - N° 294 P. 3-10 - janvier 2021 Doi : 10.1016/j.amcp.2020.11.002

C.2 Refereed Workshop and Symposia Paper

* Pallanca Olivier, Khalife Sammy and Read Jesse, "Detection of sleep spindles in NREM 2 sleep stages: Preliminary study and benchmarking of algorithms," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),Madrid, 2018, pp. 2652-2655, doi: 10.1109/BIBM.2018.8621305.
* Agrigoroaie Roxana; Pallanca Olivier; Tapus Adriana, Impact of insomnia and morningness-eveningness type on cognitive performance, Journee Fedev 2018

C.3 Refereed Poster Papers

* Pallanca Olivier, Boniol Paul, Read Jesse, Characterization of sleep states with EEG pattern detection and impact of signal quality,DS3-2018

Bibliography

[1] Arthur L Aaronson, Oran B Dent, and Christopher D Kline. "Cross-validation of MMPI and MMPI-2 predictor scales". In: *Journal of clinical psychology* 52.3 (1996), pp. 311–315.

[2] David Alvarez-Melis and Tommi S Jaakkola. "On the robustness of interpretability methods". In: *arXiv preprint arXiv:1806.08049* (2018).

[3] Thomas Andrillon et al. "Revisiting the value of polysomnographic data in insomnia: more than meets the eye". In: *Sleep medicine* 66 (2020), pp. 184–200.

[4] Stefan Appelhoff et al. "MNE-BIDS: Organizing electrophysiological data into the BIDS format and facilitating their analysis". In: *Journal of Open Source Software* 4.44 (2019), p. 1896. DOI: 10.21105/joss.01896. URL: https://doi.org/10.21105/joss.01896.

[5] Isabelle Arnulf. "REM sleep behavior disorder: motor manifestations and pathophysiology". In: *Movement Disorders* 27.6 (2012), pp. 677–689.

[6] Isabelle Arnulf et al. "CSF versus serum leptin in narcolepsy: is there an effect of hypocretin deficiency?" In: *Sleep* 29.8 (2006), pp. 1017–1024.

[7] Eugene Aserinsky and Nathaniel Kleitman. "Regularly occurring periods of eye motility, and concomitant phenomena, during sleep". In: *Science* 118.3062 (1953), pp. 273–274.

[8] American Psychiatric Association. "12 Sleep-wake disorders -Chronic insomnia". In: *In Diagnostic and statistical manual of mental disorders* 5-text rev (2022).

[9] Vedant Bahel, Sofia Pillai, and Manit Malhotra. "A comparative study on various binary classification algorithms and their improved variant for optimal performance". In: *2020 IEEE Region 10 Symposium (TENSYMP)*. IEEE. 2020, pp. 495–498.

[10] C. H. Bastien, A. Vallieres, and C. M. Morin. "Validation of the Insomnia Severity Index as an outcome measure for insomnia research". In: *Sleep Med* 2.4 (2001), pp. 297–307. ISSN: 1878-5506 (Electronic) 1389-9457 (Linking).

[11] Célyne H Bastien et al. "Information processing varies between insomnia types: measures of N1 and P2 during the night". In: *Behavioral sleep medicine* 11.1 (2013), pp. 56–72.

[12] William G Baxt. "Use of an artificial neural network for the diagnosis of myocardial infarction". In: *Annals of internal medicine*. Vol. 115. 11. American College of Physicians. 1991, pp. 843–848.

[13] Andrew L Beam and Isaac S Kohane. "Big data and machine learning in health care". In: *Jama* 319.13 (2018), pp. 1317–1318.

[14] Aaron T Beck, Robert A Steer, and Gregory K Brown. "Manual for the Beck Depression Inventory-II". In: (1996).

[15] F Beck, JB Richard, and D Léger. "Insomnia and total sleep time in France: prevalence and associated socio-demographic factors in a general population survey". In: *Revue neurologique* 169.12 (2013), pp. 956–964.

[16] Joan B Beckwith, Samuel Battle Hammond, and Ian Matthew Campbell. "Homogeneous scales for the neurotic triad of the MMPI". In: *Journal of personality assessment* 47.6 (1983), pp. 604–613.

[17] Richard B Berry et al. "The AASM manual for the scoring of sleep and associated events". In: *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine* (2012).

[18] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[19] Tessa F Blanken et al. “Insomnia disorder subtypes derived from life history and traits of affect and personality”. In: *The Lancet Psychiatry* 6.2 (2019), pp. 151–163.

[20] Morgane Boillot et al. “Glutamatergic neuron-targeted loss of LGI1 epilepsy gene results in seizures”. In: *Brain* 137.11 (2014), pp. 2984–2996.

[21] Michael H Bonnet and DL Arand. “Physiological activation in patients with sleep state misperception”. In: *Psychosomatic medicine* 59.5 (1997), pp. 533–540.

[22] TD Borkovec. “Pseudo (experiental)-insomnia and idiopathic (objective) insomnia: theoretical and therapeutic issues”. In: *Advances in behaviour research and therapy* 2.1 (1979), pp. 27–55.

[23] Leo Breiman. “Bagging predictors”. In: *Machine learning* 24.2 (1996), pp. 123–140.

[24] Leo Breiman et al. *Classification and regression trees*. CRC press, 1984.

[25] Nadia Burkart and Marco F Huber. “A survey on the explainability of supervised machine learning”. In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 245–317.

[26] James N Butcher et al. *MMPI-2: Manual for administration and scoring*. 1989.

[27] Daniel J Buysse et al. “The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research”. In: *Psychiatry research* 28.2 (1989), pp. 193–213.

[28] Danilo Bzdok, Denis Engemann, and Bertrand Thirion. “Inference and prediction diverge in biomedicine”. In: *Patterns* 1.8 (2020), p. 100119.

[29] Danilo Bzdok and John PA Ioannidis. “Exploration, inference, and prediction in neuroscience and biomedicine”. In: *Trends in neurosciences* 42.4 (2019), pp. 251–262.

[30] Danilo Bzdok et al. “Prediction and inference diverge in biomedicine: Simulations and real-world data”. In: *BioRxiv* (2018), p. 327437.

[31] Ian G Campbell. “EEG recording and analysis for sleep research”. In: *Current protocols in neuroscience* 49.1 (2009), pp. 10–2.

[32] Xiao-Lan Cao et al. “The prevalence of insomnia in the general population in China: a meta-analysis”. In: *PloS one* 12.2 (2017), e0170772.

[33] Colleen E Carney et al. “Assessing depression symptoms in those with insomnia: an examination of the beck depression inventory second edition (BDI-II)”. In: *Journal of psychiatric research* 43.5 (2009), pp. 576–582.

[34] Mary A. Carskadon and William C. Dement. “The Multiple Sleep Latency Test: What does it measure?” In: *Sleep Medicine Reviews* 9.5 (2005), pp. 395–406.

[35] Anna Castelnovo et al. “The paradox of paradoxical insomnia: a theoretical review towards a unifying evidence-based definition”. In: *Sleep medicine reviews* 44 (2019), pp. 70–82.

[36] Stanislas Chambon et al. “A deep learning architecture to detect events in EEG signals during sleep”. In: *arXiv preprint arXiv:1807.05981* (2018).

[37] Ramiro Chaparro-Vargas et al. “Insomnia characterization: from hypnogram to graph spectral theory”. In: *IEEE Transactions on Biomedical Engineering* 63.10 (2016), pp. 2211–2219.

[38] Brigitte Chaput, Jean-Claude Girard, and Michel Henry. “Frequentist approach: Modelling and simulation in statistics and probability teaching”. In: *Teaching statistics in school mathematics-Challenges for teaching and teacher education: A joint ICMI/IASE study: The 18th ICMI study* (2011), pp. 85–95.

[39] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794.

[40] Davide Chicco and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC genomics* 21.1 (2020), pp. 1–13.

[41] Vanda Clemente et al. "The European Portuguese version of the insomnia severity index". In: *Journal of Sleep Research* 30.1 (2021), e13198.

[42] Ton J. Cleophas and Aeilko H. Zwinderman. "Bonferroni t-Test". In: *Statistical Analysis of Clinical Data on a Pocket Calculator*. Dordrecht: Springer, 2011. DOI: 10.1007/978-94-007-1211-9_15. URL: https://doi.org/10.1007/978-94-007-1211-9_15.

[43] Pierre Comon. "Independent component analysis, a new concept?" In: *Signal processing* 36.3 (1994), pp. 287–314.

[44] Dorothée Coppieters't Wallant et al. "Automatic artifacts and arousals detection in whole-night sleep EEG recordings". In: *Journal of neuroscience methods* 258 (2016), pp. 124–133.

[45] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[46] João Costa et al. "An Automatic Sleep Spindle detector based on WT, STFT and WMSD". In: *World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering* 6.8 (2012), pp. 397–400.

[47] Thomas Cover and Peter Hart. "Nearest neighbor pattern classification". In: *IEEE transactions on information theory* 13.1 (1967), pp. 21–27.

[48] Megan R Crawford et al. "Characterization of patients who present with insomnia: is there room for a symptom cluster-based approach?" In: *Journal of Clinical Sleep Medicine* 13.7 (2017), pp. 911–921.

[49] Armando D'Agostino et al. "Sleep endophenotypes of schizophrenia: slow waves and sleep spindles in unaffected first-degree relatives". In: *npj Schizophrenia* 4.1 (2018), pp. 1–8.

[50] Edwin S Dalmaijer, Camilla L Nord, and Duncan E Astle. "Statistical power for cluster analysis". In: *BMC bioinformatics* 23.1 (2022), pp. 1–28.

[51] Pratap Dangeti. *Statistics for machine learning*. Packt Publishing Ltd, 2017.

[52] Tam K Dao, Frances Prevatt, and Heather Leveta Horne. "Differentiating psychotic patients from nonpsychotic patients with the MMPI–2 and Rorschach". In: *Journal of Personality Assessment* 90.1 (2008), pp. 93–101.

[53] Janez Demšar, Gregor Leban, and Blaž Zupan. "FreeViz—An intelligent multivariate visualization approach to explorative analysis of biomedical data". In: *Journal of biomedical informatics* 40.6 (2007), pp. 661–671.

[54] Serena Dittoni et al. "Psychological functioning measures in patients with primary insomnia and sleep state misperception". In: *Acta neurologica Scandinavica* 128.1 (2013), pp. 54–60.

[55] Cynthia M Dorsey and Richard R Bootzin. "Subjective and psychophysiologic insomnia: an examination of sleep tendency and personality". In: *Biological Psychiatry* 41.2 (1997), pp. 209–216.

[56] Jack D Edinger, Anna L Stout, and Timothy J Hoelscher. "Cluster analysis of insomniacs' MMPI profiles: relation of subtypes to sleep history and treatment outcome." In: *Psychosomatic medicine* 50.1 (1988), pp. 77–87.

[57] Jack D Edinger et al. "Derivation of research diagnostic criteria for insomnia: report of an American Academy of Sleep Medicine Work Group". In: *Sleep* 27.8 (2004), pp. 1567–1596.

[58] Jack D Edinger et al. “Insomnia and the eye of the beholder: are there clinical markers of objective sleep disturbances among adults with and without insomnia complaints?” In: *Journal of consulting and clinical psychology* 68.4 (2000), p. 586.

[59] Bradley Efron and Trevor Hastie. *Computer age statistical inference*. 2013.

[60] Manuel Fernández-Delgado et al. “Do we need hundreds of classifiers to solve real world classification problems?” In: *The journal of machine learning research* 15.1 (2014), pp. 3133–3181.

[61] Julio Fernandez-Mendoza et al. “Insomnia with objective short sleep duration is associated with deficits in neuropsychological performance: a general population study”. In: *Sleep* 33.4 (2010), pp. 459–465.

[62] Julio Fernandez-Mendoza et al. “Sleep misperception and chronic insomnia in the general population: the role of objective sleep duration and psychological profiles”. In: *Psychosomatic medicine* 73.1 (2011), p. 88.

[63] Fabio Ferrarelli et al. “Reduced sleep spindle activity in schizophrenia patients”. In: *American Journal of Psychiatry* 164.3 (2007), pp. 483–492.

[64] Samuel G Finlayson, Andrew L Beam, and Maarten van Smeden. “Machine Learning and Statistics in Clinical Research Articles—Moving Past the False Dichotomy”. In: *JAMA pediatrics* 177.5 (2023), pp. 448–450.

[65] Kathleen A Foley et al. “Subtypes of sleep disturbance: associations among symptoms, comorbidities, treatment, and medical costs”. In: *Behavioral Sleep Medicine* 8.2 (2010), pp. 90–104.

[66] Daniel E Ford and Douglas B Kamerow. “Epidemiologic study of sleep disturbances and psychiatric disorders: an opportunity for prevention?” In: *Jama* 262.11 (1989), pp. 1479–1484.

[67] Mary E Frame. *Quantifying Eye Movement Trajectory Similarity for Use in Human Performance Experiments in Intelligence, Surveillance, and Reconnaissance (ISR) Research*. Tech. rep. Wright State Research Institute Beavercreek United States, 2018.

[68] Yoav Freund and Robert E Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”. In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.

[69] Alan F Friedman et al. *Psychological assessment with the MMPI-2/MMPI-2-RF*. Routledge, 2014.

[70] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer, 2001.

[71] Jeanne M Geiger-Brown et al. “Cognitive behavioral therapy in persons with comorbid insomnia: a meta-analysis”. In: *Sleep medicine reviews* 23 (2015), pp. 54–67.

[72] Finja Gerlach et al. “Insomnia-related interpretational bias is associated with pre-sleep worry”. In: *Journal of sleep research* 29.1 (2020), e12938.

[73] Jerome D Gilmore et al. “Adherence to substance abuse treatment: Clinical utility of two MMPI-2 scales”. In: *Journal of Personality Assessment* 77.3 (2001), pp. 524–540.

[74] Ary L Goldberger et al. “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals”. In: *Circulation* 101.23 (2000), e215–e220. DOI: 10.1161/01.CIR.101.23.e215. URL: https://www.ahajournals.org/doi/abs/10.1161/01.CIR.101.23.e215.

[75] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[76] Krzysztof J Gorgolewski et al. “The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments”. In: *Scientific Data* 3 (2016), p. 160044. DOI: 10.1038/sdata.2016.44. URL: https://www.hal.inserm.fr/inserm-01345616.

[77] Daniel J Gottlieb et al. “Relation of sleepiness to respiratory disturbance index: the Sleep Heart Health Study”. In: *American journal of respiratory and critical care medicine* 159.2 (1999), pp. 502–507.

[78] Alexandre Gramfort et al. “MEG and EEG data analysis with MNE-Python”. In: *Frontiers in Neuroscience* 7 (2013), p. 267. DOI: 10.3389/fnins.2013.00267. URL: https://www.frontiersin.org/article/10.3389/fnins.2013.00267.

[79] Yu Guo et al. “Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms”. In: *BMC bioinformatics* 11.1 (2010), pp. 1–19.

[80] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Elsevier, 2011.

[81] Allison G Harvey. “A cognitive model of insomnia”. In: *Behaviour research and therapy* 40.8 (2002), pp. 869–893.

[82] Allison G Harvey and Nicole KY Tang. “(Mis) perception of sleep in insomnia: a puzzle and a resolution.” In: *Psychological bulletin* 138.1 (2012), p. 77.

[83] Ahnaf Rashik Hassan and Mohammed Imamul Hassan Bhuiyan. “Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting”. In: *Computer methods and programs in biomedicine* 140 (2017), pp. 201–210.

[84] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[85] Peter J Hauri and Joyce Wisbey. “Wrist actigraphy in insomnia”. In: *Sleep* 15.4 (1992), pp. 293–301.

[86] Patrick Hoffman et al. “DNA visual and analytic data mining”. In: *Proceedings. Visualization’97 (Cat. No. 97CB36155)*. IEEE. 1997, pp. 437–441.

[87] F Hohagen et al. “Prevalence and treatment of insomnia in general practice: a longitudinal study”. In: *European archives of psychiatry and clinical neuroscience* 242 (1993), pp. 329–336.

[88] Andreas Holzinger et al. “Causability and explainability of artificial intelligence in medicine”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.4 (2019), e1312.

[89] James A Horne and Olov Östberg. “A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms”. In: *International Journal of Chronobiology* 4.2 (1976), pp. 97–110.

[90] Conrad Iber et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine, 2007.

[91] John PA Ioannidis. “Why most published research findings are false”. In: *PLoS medicine* 2.8 (2005), e124.

[92] Md Kafiul Islam, Amir Rastegarnia, and Zhi Yang. “Methods for artifact detection and removal from scalp EEG: A review”. In: *Neurophysiologie Clinique/Clinical Neurophysiology* 46.4-5 (2016), pp. 287–305.

[93] Su-Kyeong Jang et al. “Reliability and clinical utility of machine learning to predict stroke prognosis: comparison with logistic regression”. In: *Journal of stroke* 22.3 (2020), p. 403.

[94] Geneviève St-Jean et al. “REM and NREM power spectral analysis on two consecutive nights in psychophysiological and paradoxical insomnia sufferers”. In: *International Journal of Psychophysiology* 89.2 (2013), pp. 181–194.

[95] Murray W Johns. “A new method for measuring daytime sleepiness: the Epworth sleepiness scale”. In: *sleep* 14.6 (1991), pp. 540–545.

[96] Sandra LJ Johnson. “AI, machine learning, and ethics in health care”. In: *Journal of Legal Medicine* 39.4 (2019), pp. 427–441.

[97] Ian T Jolliffe and Jorge Cadima. “Principal component analysis: a review and recent developments”. In: *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150202.

[98] Anthony Kales and Allan Rechtschaffen. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. UCLA Brain Information Service. Brain Research Institute, 1968.

[99] Anthony Kales et al. “Biopsychobehavioral correlates of insomnia: II. Pattern specificity and consistency with the Minnesota Multiphasic Personality Inventory.” In: *Psychosomatic Medicine* (1983).

[100] Håvard Kallestad et al. “Mode of delivery of Cognitive Behavioral Therapy for Insomnia: a randomized controlled non-inferiority trial of digital and face-to-face therapy”. In: *Sleep* 44.12 (2021), zsab185.

[101] Chien-Hui Kao et al. “Insomnia subtypes characterised by objective sleep duration and NREM spectral power and the effect of acute sleep restriction: an exploratory analysis”. In: *Scientific reports* 11.1 (2021), p. 24331.

[102] Daniel B Kay et al. “Subjective–objective sleep discrepancy among older adults: associations with insomnia diagnosis and insomnia treatment”. In: *Journal of sleep research* 24.1 (2015), pp. 32–39.

[103] Habibolah Khazaie et al. “Insomnia treatment by olanzapine. Is sleep state misperception a psychotic disorder?” In: *Neurosciences Journal* 15.2 (2010), pp. 110–112.

[104] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[105] Ansgar Koene et al. “A governance framework for algorithmic accountability and transparency”. In: (2019).

[106] Birgitte R Kornum et al. “Common variants in P2RY11 are associated with narcolepsy”. In: *Nature genetics* 43.1 (2011), pp. 66–71.

[107] H. et al Kryger Meir. “Section 1: Normal sleep and its variance”. In: 2021, pp. 18–22.

[108] Andrew D Krystal et al. “NREM sleep EEG frequency spectral correlates of sleep complaints in primary insomnia subtypes”. In: *Sleep* 25.6 (2002), pp. 626–636.

[109] Linda A Kuisk, Amy D Bertelson, and James K Walsh. “Presleep cognitive hyperarousal and affect as factors in objective and subjective insomnia”. In: *Perceptual and Motor Skills* 69.3-2 (1989), pp. 1219–1225.

[110] Roshan Kumari and Saurabh Kr Srivastava. “Machine learning: A review on binary classification”. In: *International Journal of Computer Applications* 160.7 (2017).

[111] Clete A Kushida et al. “Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients”. In: *Sleep medicine* 2.5 (2001), pp. 389–396.

[112] Merijn van de Laar et al. “The role of personality traits in insomnia”. In: *Sleep medicine reviews* 14.1 (2010), pp. 61–68.

[113] Karine Lacourse et al. “A sleep spindle detection algorithm that emulates human expert spindle scoring”. In: *Journal of neuroscience methods* (2018).

[114] Stefan Larsson and Fredrik Heintz. "Transparency in artificial intelligence". In: *Internet Policy Review* 9.2 (2020).

[115] Daniela Latorre et al. "Narcolepsy: a model interaction between immune system, nervous system, and sleep-wake regulation". In: *Seminars in immunopathology*. Vol. 44. 5. Springer. 2022, pp. 611–623.

[116] David LeBlond. "FDA Bayesian statistics guidance for medical device clinical trials-application to process validation". In: *Journal of Validation technology* 16.4 (2010), p. 24.

[117] Sandro Lecci et al. "Electroencephalographic changes associated with subjective under-and overestimation of sleep duration". In: *Sleep* 43.11 (2020), zsaa094.

[118] Yves Leclercq et al. "fMRI artefact rejection and sleep scoring toolbox". In: *Computational intelligence and neuroscience* 2011 (2011).

[119] Chang Woo Lee et al. "Depression and anxiety associated with insomnia and recent stressful life events". In: *Chronobiology in Medicine* 1.3 (2019), pp. 121–125.

[120] Mi Hyun Lee et al. "Multitask fMRI and machine learning approach improve prediction of differential brain activity pattern in patients with insomnia disorder". In: *Scientific Reports* 11.1 (2021), p. 9402.

[121] Yun Ji Lee et al. "Interrater reliability of sleep stage scoring: a meta-analysis". In: *Journal of Clinical Sleep Medicine* 18.1 (2022), pp. 193–202.

[122] H Matthew Lehrer et al. "Comparing polysomnography, actigraphy, and sleep diary in the home environment: The Study of Women's Health Across the Nation (SWAN) Sleep Study". In: *Sleep Advances* 3.1 (2022), zpac001.

[123] Alexander Lex et al. "UpSet: visualization of intersecting sets". In: *IEEE transactions on visualization and computer graphics* 20.12 (2014), pp. 1983–1992.

[124] Yingjie Liang et al. "Sleep misperception and associated factors in patients with anxiety-related disorders and complaint of insomnia: a retrospective study". In: *Frontiers in Neurology* 13 (2022).

[125] Mengting Liao et al. "A novel predictive model incorporating immune-related gene signatures for overall survival in melanoma patients". In: *Scientific reports* 10.1 (2020), pp. 1–12.

[126] Paulo JG Lisboa. "Interpretability in machine learning–principles and practice". In: *Fuzzy Logic and Applications: 10th International Workshop, WILF 2013, Genoa, Italy, November 19-22, 2013. Proceedings 10*. Springer. 2013, pp. 15–21.

[127] Renske Lok and Jamie M Zeitzer. "Physiological correlates of the Epworth Sleepiness Scale reveal different dimensions of daytime sleepiness". In: *Sleep Advances* 2.1 (2021), zpab008.

[128] O. Loyola-González. "Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View". In: *IEEE Access* 7 (2019), pp. 154096–154113.

[129] Qian Lu et al. "Connectomic disturbances underlying insomnia disorder and predictors of treatment response". In: *Frontiers in Human Neuroscience* (2022).

[130] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*. 2017, pp. 4765–4774.

[131] Mauro Manconi et al. "Measuring the error in sleep estimation in normal subjects and in patients with insomnia". In: *Journal of sleep research* 19.3 (2010), pp. 478–486.

[132] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.

[133] Dara S Manoach et al. “Reduced sleep spindles in schizophrenia: a treatable endophenotype that links risk genes to impaired cognition?” In: *Biological psychiatry* 80.8 (2016), pp. 599–608.

[134] Nicolas Martin et al. “Topography of age-related changes in sleep spindles”. In: *Neurobiology of aging* 34.2 (2013), pp. 468–476.

[135] MathWorks. *Support Vector Machines for Binary Classification*. https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html.

[136] Catherine A McCall et al. “Sleep and psychiatric disease”. In: *Atlas of Clinical Sleep Medicine E-Book: Expert Consult-Online* (2022), p. 396.

[137] Christina S McCrae et al. “Cognitive behavioral treatments for insomnia and pain in adults with comorbid chronic insomnia and fibromyalgia: clinical outcomes from the SPIN randomized controlled trial”. In: *Sleep* 42.3 (2019), zsy234.

[138] Wallace B Mendelson. “Long-term follow-up of chronic insomnia”. In: *Sleep* 18.8 (1995), pp. 698–701.

[139] Helli Merica, Robert Blois, and J-M Gaillard. “Spectral characteristics of sleep EEG in chronic insomnia”. In: *European Journal of Neuroscience* 10.5 (1998), pp. 1826–1834.

[140] Christopher Miller et al. “Chapter 4. Methodology for the Assessment of Sleep”. In: Dec. 2015. ISBN: 9780124171886. DOI: 10.1016/B978-0-12-417188-6.00004-9.

[141] Christopher B Miller et al. “Clusters of insomnia disorder: an exploratory cluster analysis of objective sleep parameters reveals differences in neurocognitive functioning, quantitative EEG, and heart rate variability”. In: *Sleep* 39.11 (2016), pp. 1993–2004.

[142] Tom Michael Mitchell et al. *Machine learning*. Vol. 1. McGraw-hill New York, 2007.

[143] Matthias Mölle et al. “Fast and slow spindles during the sleep slow oscillation: disparate coalescence and engagement in memory processing”. In: *Sleep* 34.10 (2011), pp. 1411–1421.

[144] Hye-Jin Moon, Mei Ling Song, and Yong Won Cho. “Clinical characteristics of primary insomniacs with sleep-state misperception”. In: *Journal of Clinical Neurology* 11.4 (2015), pp. 358–363.

[145] Charles M Morin. “Measuring outcomes in randomized clinical trials of insomnia treatments”. In: *Sleep medicine reviews* 7.3 (2003), pp. 263–279.

[146] Charles M Morin, Annie Vallières, and Hans Ivers. “Dysfunctional beliefs and attitudes about sleep (DBAS): validation of a brief version (DBAS-16)”. In: *Sleep* 30.11 (2007), pp. 1547–1554.

[147] Charles M Morin et al. “Dysfunctional beliefs and attitudes about sleep among older adults with and without insomnia complaints.” In: *Psychology and aging* 8.3 (1993), p. 463.

[148] Charles M Morin et al. “Prevalence of insomnia and its treatment in Canada”. In: *The Canadian Journal of Psychiatry* 56.9 (2011), pp. 540–548.

[149] Charles M Morin et al. “The Insomnia Severity Index: psychometric indicators to detect insomnia cases and evaluate treatment response”. In: *Sleep* 34.5 (2011), pp. 601–608.

[150] Max D. Morris. “Factorial Sampling Plans for Preliminary Computational Experiments”. In: *Technometrics* 33.2 (1991), pp. 161–174.

[151] Yuval Nir et al. “Regional slow waves and spindles in human sleep”. In: *Neuron* 70.1 (2011), pp. 153–169.

[152] Marie-Pier Normand, Patrick St-Hilaire, and Célyne H Bastien. “Sleep spindles characteristics in insomnia sufferers and their relationship with sleep misperception”. In: *Neural plasticity* 2016 (2016).

[153] Christian O’Reilly et al. “Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research”. In: *Journal of Sleep Research* 23.6 (2014), pp. 628–635. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.12169.

[154] Christian O’Reilly and Tore Nielsen. “Automatic sleep spindle detection: benchmarking with fine temporal resolution using open science tools”. In: *Frontiers in human neuroscience* 9 (2015), p. 353.

[155] Isa Okajima et al. “Development and Validity of the Japanese Version of the Pre-Sleep Arousal Scale”. In: *The Tohoku Journal of Experimental Medicine* 252.2 (2020), pp. 169–176.

[156] World Health Organization. “E07 Sleep-wake disorders -Chronic insomnia.” In: *In International statistical classification of diseases and related health problems* 11 (2019).

[157] O Pallanca and J Read. “Principes généraux et définitions en intelligence artificielle”. In: *Archives des Maladies du Coeur et des Vaisseaux-Pratique* 2021.294 (2021), pp. 3–10.

[158] Olivier Pallanca. “Archive of the Center for the Investigation and Treatment of Insomnia”. In: (2023).

[159] Ankit Parekh et al. “Multichannel sleep spindle detection using sparse low-rank optimization”. In: *Journal of neuroscience methods* 288 (2017), pp. 1–16.

[160] Liborio Parrino et al. “Cyclic alternating pattern (CAP): the marker of sleep instability”. In: *Sleep medicine reviews* 16.1 (2012), pp. 27–45.

[161] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: (2019), pp. 8024–8035.

[162] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[163] Michael L Perlis et al. “Beta/Gamma EEG activity in patients with primary and secondary insomnia and good sleeper controls”. In: *Sleep* 24.1 (2001), pp. 110–117.

[164] Cyril Pernet et al. “BIDS-EEG: an extension to the Brain Imaging Data Structure (BIDS) Specification for electroencephalography”. In: (2018).

[165] Kevin R Peters et al. “Age differences in the variability and distribution of sleep spindle and rapid eye movement densities”. In: *PloS one* 9.3 (2014), e91047.

[166] Christelle Peyron et al. “A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains”. In: *Nature medicine* 6.9 (2000), pp. 991–997.

[167] DT Plante et al. “Topographic and sex-related differences in sleep spindles in major depressive disorder: a high-density EEG investigation”. In: *Journal of affective disorders* 146.1 (2013), pp. 120–125.

[168] Shi-Hui Poon, Shin-Yi Quek, and Tih-Shih Lee. “Insomnia Disorders: Nosology and Classification Past, Present, and Future”. In: *The Journal of Neuropsychiatry and Clinical Neurosciences* 33.3 (2021), pp. 194–200.

[169] David Martin Powers. “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”. In: *arXiv preprint arXiv:2010.16061* (2011).

[170] Foster Provost and Tom Fawcett. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O’Reilly Media, 2013.

[171] SM Purcell et al. “Characterizing sleep spindles in 11,630 individuals from the National Sleep Research Resource”. In: *Nature communications* 8.1 (2017), pp. 1–16.

[172] "Random forest versus logistic regression: a large-scale benchmark experiment". In: *BMC Bioinformatics* (). URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2264-5.

[173] Laura Ray et al. "Expert and crowd-sourced validation of an individualized sleep spindle detection method employing complex demodulation and individualized normalization". In: *Frontiers in human neuroscience* 9 (2015), p. 507.

[174] Dieter Riemann et al. "European guideline for the diagnosis and treatment of insomnia". In: *Journal of sleep research* 26.6 (2017), pp. 675–700.

[175] Irina Rish. "Empirical comparison of supervised learning algorithms". In: *Proceedings of the International Conference on Machine Learning*. Vol. 18. Morgan Kaufmann. 2001, pp. 577–584.

[176] Andrea Rodenbeck et al. "A review of sleep EEG patterns. Part I: a compilation of amended rules for their visual recognition according to Rechtschaffen and Kales". In: *Somnologie* 10.4 (2006), pp. 159–175.

[177] Timothy Roehrs et al. "Disturbed sleep predicts hypnotic self-administration". In: *Sleep Medicine* 3.1 (2002), pp. 61–66.

[178] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[179] Mirka Saarela and Susanne Jauhiainen. "Comparison of feature importance measures as explanations for classification models". In: *SN Applied Sciences* 3.2 (2021), pp. 1–12.

[180] Rafael J Salin-Pascual et al. "Long-term study of the sleep of insomnia patients with sleep state misperception and other insomnia patients". In: *Am J Psychiatry* 149.7 (1992), pp. 904–908.

[181] Stacy D Sanford et al. "The influence of age, gender, ethnicity, and insomnia on Epworth sleepiness scores: a normative US population". In: *Sleep Medicine* 7.4 (2006), pp. 319–326.

[182] Clifford B Saper, Thomas E Scammell, and Jun Lu. "Hypothalamic regulation of sleep and circadian rhythms". In: *Nature* 437.7063 (2005), pp. 1257–1263.

[183] Michael J Sateia. "International classification of sleep disorders". In: *Chest* 146.5 (2014), pp. 1387–1394.

[184] Karen Schmitt, Edith Holsboer-Trachsler, and Anne Eckert. "BDNF in sleep, insomnia, and sleep deprivation". In: *Annals of medicine* 48.1-2 (2016), pp. 42–51.

[185] Dietrich Schneider-Helmert. "Twenty-four-hour sleep-wake function and personality patterns in chronic insomniacs and healthy controls". In: *Sleep* 10.5 (1986), pp. 452–462.

[186] Julie Seibt et al. *Role of spindle oscillations across lifespan in health and disease*. 2016.

[187] Pratap Chandra Sen, Mahimarnab Hajra, and Mitadru Ghosh. "Supervised classification algorithms in machine learning: A survey and review". In: *Emerging technology in modelling and graphics*. Springer, 2020, pp. 99–111.

[188] Gülçin Benbir Şenel et al. "Changes in sleep structure and sleep spindles are associated with the neuropsychiatric profile in paradoxical insomnia". In: *International Journal of Psychophysiology* 168 (2021), pp. 27–32.

[189] Masamitsu Shibagaki, Sigehiro Kiyono, and Kazuyoshi Watanabe. "Spindle evolution in normal and mentally retarded children: a review". In: *Sleep* 5.1 (1982), pp. 47–57.

[190] Association of Sleep Disorders Centers. *Diagnostic classification of sleep and arousal disorders*. Raven, 1979.

[191] American Academy of Sleep Medicine. “AASM manual for the Scoring of Sleep and Associated Events: Rules, Terminology and technical Specifications”. In: *aasm* 2.1 (2014).

[192] American Academy of Sleep Medicine et al. “International classification of sleep disorders”. In: *Diagnostic and coding manual* (2005), pp. 51–55.

[193] Charles D Spielberger, Richard L Gorsuch, and Robert Lushene. *Manual for the State-Trait Anxiety Inventory*. Consulting Psychologists Press, 1983.

[194] Victor I Spoormaker et al. “Initial validation of the SLEEP-50 questionnaire”. In: *Behavioral sleep medicine* 3.4 (2005), pp. 227–246.

[195] Jeffrey L Sugerman, John A Stern, and James K Walsh. “Daytime alertness in subjective and objective insomnia: some preliminary findings”. In: *Biological Psychiatry* 20.7 (1985), pp. 741–750.

[196] Nicole KY Tang and Allison G Harvey. “Effects of cognitive arousal and physiological arousal on sleep perception”. In: *Sleep* 27.1 (2004), pp. 69–78.

[197] Nicole KY Tang and Allison G Harvey. “Time estimation ability and distorted perception of sleep in insomnia”. In: *Behavioral Sleep Medicine* 3.3 (2005), pp. 134–150.

[198] Michael J Thorpy. “Classification of sleep disorders.” In: *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society* 7.1 (1990), pp. 67–81.

[199] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58.1 (1996), pp. 267–288.

[200] *UK Biobank Resource*. http://www.ukbiobank.ac.uk/. Accessed: yyyy-mm-dd. Year.

[201] R Vallat. “YASA (yet another spindle algorithm): A fast and open-source sleep spindles and slow-waves detection toolbox”. In: *Sleep Medicine* 64 (2019), S396.

[202] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[203] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, 1979.

[204] Hoda Doos Ali Vand et al. “Validity and reliability of the dysfunctional beliefs and attitudes about sleep scale-10 in iranian clinical population”. In: *Iranian Journal of Psychiatry and Behavioral Sciences* 12.2 (2018).

[205] Juliana Vochem et al. “Pre-sleep arousal scale (PSAS) and the time monitoring Behavior-10 scale (TMB-10) in good sleepers and patients with insomnia”. In: *Sleep medicine* 56 (2019), pp. 98–103.

[206] Erin J Wamsley et al. “Reduced sleep spindles and spindle coherence in schizophrenia: mechanisms of impaired memory consolidation?” In: *Biological psychiatry* 71.2 (2012), pp. 154–161.

[207] Yuan-Pang Wang and Clarice Gorenstein. “Psychometric properties of the Beck Depression Inventory-II: a comprehensive review”. In: *Brazilian Journal of Psychiatry* 35 (2013), pp. 416–431.

[208] Simon C Warby et al. “Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods”. In: *Nature methods* 11.4 (2014), p. 385.

[209] Oren M Weiner and Thien Thanh Dang-Vu. “Spindle oscillations in sleep disorders: a systematic review”. In: *Neural plasticity* 2016 (2016).

[210] M Wenigmann et al. “Sleep state misperception in psychiatric patients”. In: *Somnologie* 23 (2019), pp. 43–48.

[211] Ines Wilhelm et al. “Widespread reduction in sleep spindle activity in socially anxious children and adolescents”. In: *Journal of psychiatric research* 88 (2017), pp. 47–55.

[212] Peter WF Wilson et al. “Prediction of coronary heart disease using risk factor categories”. In: *Circulation* 97.18 (1998), pp. 1837–1847.

[213] D Wishart. “FORTRAN 2 PROGRAMS FOR 8 METHODS OF CLUSTER ANALYSIS (CLUSTAN 1).” In: (1969).

[214] James K Wyatt et al. “Sleep onset is associated with retrograde and anterograde amnesia”. In: *Sleep* 17.6 (1994), pp. 502–511.

[215] Jie Zhang and Zong-ming Zhang. “Ethics and governance of trustworthy medical artificial intelligence”. In: *BMC Medical Informatics and Decision Making* 23.1 (2023), pp. 1–15.

[216] Qiuming Zhu. “On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset”. In: *Pattern Recognition Letters* 136 (2020), pp. 71–80.

[217] F Zorick et al. “Polysomnographic and MMPI characteristics of patients with insomnia”. In: *Sleep, Benzodiazepines and Performance: Experimental Methodologies and Research Prospects*. Springer. 1984, pp. 2–10.

[218] Jarosław Żygierewicz et al. “High resolution study of sleep spindles”. In: *Clinical Neurophysiology* 110.12 (1999), pp. 2136–2147.

Titre : Une meilleure compréhension de l'insomnie paradoxale (IP) : une approche basée sur la connaissance utilisant des outils d'apprentissage automatique (AA)

Mots clés : médecine du sommeil, insomnie paradoxale, modélisation prédictive, explicabilité, électroencéphalogramme (EEG)

Résumé : Cette thèse à pour but la caractérisation de l'insomnie et la prédiction de la réponse thérapeutique afin de mieux comprendre les rechutes fréquentes de ce troublequi affecte 10-20% de la population générale. En particulier, l'IP, qui est une perception erronée de l'état de sommeil. L'IP est fréquente mais est encore mal comprise et difficile à gérer. Cette thèse vise à apporter un éclairage nouveau sur ce sujet grâce à des outils d'AA. Nous avons créé une base de données décrivant 423 patients diagnostiqués avec une insomnie chronique (IC), sur lesquels nous avons effectué des examens cliniques, psychométriques, actimétriques et polysomnographiques (tels que l'EEG) au début d'une étude prospective et un suivi d'au moins six mois avec une évaluation de la réponse au traitement standard; ce qui donne autour de 200 caractéristiques par patient. Nous avons utilisé des outils d'AA pour identifier des groupes particuliers Nous avons testé les hypothèses existantes selon lesquelles les profils d'IP pouvaient être identifiés par l'analyse EEG, afin de déterminer leur fiabilité. L'application des outils d'AA sur les tracés EEG n'étaient pas suffisamment fiables pour prédire la perception erronée de l'état de sommeil. Nous avons trouvé des sous-groupes de patients avec présentant une somnolence subjective. Nous avons reproduit et confirmé sur notre ensemble de données la faible concordance entre 20 formules publiées dans la littérature pour définir l'IP (qui consiste essentiellement à préciser le seuil de discordance considéré comme pathologique entre la perception subjective et objective du sommeil). Grâce à des outils d'AA, nous avons montré que seules deux caractéristiques intervenaient dans la prédiction de l'IP par la plupart des formules. Cette constatation contribue à l'harmonisation de la définition de l'IP; Nous avons également montré que les patients souffrant d'IP présentaient une augmentation significative des longues périodes d'éveil au cours du sommeil, ce qui explique dans une certaine mesure le paradoxe de l'IP. Nous avons proposé une nouvelle définition de l'IP, en étendant la période d'analyse du sommeil d'une à sept nuits afin d'améliorer la fiabilité de la perception du sommeil. Au-delà de l'IP, nous sommes les premiers à utiliser l'AA pour prédire avec précision l'amélioration de l'IC après traitement à six mois (évaluée selon l'échelle ISI [Index Severity of Insomnia]), et nous avons montré que notre nouvelle définition de l'IP en était le principal facteur prédictif.

Title : An improved understanding of ParI: A knowledge-based approach using Machine Learning (ML) tools

Keywords : sleep medicine, paradoxical insomnia, predictive modeling, explainability, electroencephalogram (EEG)

Abstract : This thesis sets out to improve the characterization of insomnia, in order to provide a better understanding of treatment outcome and avoid frequent relapses in this disorder, which affects 10-20% of the general population. In particular, ParI, a decreased sleep-state misperception (the patient is unable to accurately estimate their objective sleep length and quality). PI is common in clinical practice, yet is still not well understood and difficult to manage. This thesis aims to shed new light on this subject. We curated a new database describing 423 patients diagnosed with chronic insomnia, on whom we performed clinical, psychometric, actimetric, and polysomnographic analysis (such as EEG) examinations at the beginning of the prospective study and a follow-up of at least six months with response to standard treatment; resulting in 200 features per patient. We used ML tools to identify distinct groups among chronic insomniacs, particularly identifying characteristics of those with ParI and the influence on treatment outcome. We tested existing hypotheses that ParI profiles could be identified via EEG analysis, in order to determine their reliability. We determined that EEG profiles were insufficiently reliable to be used as predictors of PArI. However, we did find subgroups of insomniac patients with subjective sleepiness (as per the Epworth sleepiness-scale questionnaire). We confirmed on our dataset the poor agreement among 20 published formulas for defining ParI.. With ML tools, we showed that only two features were involved in the prediction of PI by most of the formulas. This finding leads us to harmonize the definition. We also found that the PI patients had a significant increase in long wake bouts during the sleep episode, which explains to some extent, the paradox of ParI. We propose a new definition of PI, extending the sleep analysis period from one to seven nights to improve the reliability of sleep perception. Beyond PI, we are the first to use ML to predict the improvement of insomnia accurately (all types) after treatment at six months (evaluated with the Insomnia Severity Index), and we showed our new definition of ParI was, in fact, the main predictive factor.

Institut Polytechnique de Paris
91120 Palaiseau, France