学位論文（博士）

Explainable AI-Driven Identification of Multimodal Biomarkers for Early Prediction of Cognitive Decline

（説明可能 AI によるマルチモーダル・バイオマーカーの同定と認知機能低下の早期予測）

氏名　A H M Fahad

所属　山口大学大学院医学系研究科
医学専攻 (システムバイオ
インフォマティクス講座)

令和 8 年 5 月

Graduate School of Medicine

Yamaguchi University



Explainable AI-Driven Identification of Multimodal Biomarkers for Early Prediction of Cognitive Decline

（説明可能 AI によるマルチモーダル・バイオマーカーの同定と認知機能低下の早期予測）

A Dissertation

Submitted by

A H M Fahad

in Partial Fulfillment of the Requirement for the Degree of

Doctor of Philosophy (Medicine)

May 2026

Declaration

I hereby declare that the performed in silico experiment reported on in this research project was undersigned. I also declare that the appropriate acknowledgement to source of information of other workers results in reference.

May 2026

Yamaguchi University

Author

A H M Fahad

Preface

Dementia and age-related cognitive decline pose growing clinical and societal challenges, partly because meaningful biological changes can precede overt symptoms and because individuals follow highly variable trajectories. Improving early risk stratification and practical monitoring requires approaches that are accurate, scalable, and interpretable enough to support real-world decision-making in community and primary-care contexts.

This thesis is positioned at the intersection of biomarker discovery and applied machine learning, with a focus on whether a two-stage, clinically aligned hierarchical prediction framework can better capture the progression from normal cognition to early impairment and more advanced decline than single-step approaches. In an explicitly exploratory biomarker-discovery manner, the work further examines whether explainable inference can identify reliable, stage-specific predictors and whether scalable, non-invasive parameters can achieve high predictive performance suitable for broader deployment.

The chapters are organized to introduce the clinical and methodological background, describe the study design and analytical pipeline, present predictive and explainability results, and discuss translational implications, limitations, and directions for future validation. Gratitude is extended to supervisors, collaborators, and all participants whose contributions made this research possible.

Acknowledgments

This thesis was made possible by the intellectual foundations laid by countless philosophers, researchers, mathematicians, and scientists whose cumulative work has shaped modern civilization and the foundations of science, mathematics, statistics, machine learning, and related disciplines. Their ideas and persistent efforts continue to guide rigorous inquiry into complex questions concerning the universe and humanity and have created the conceptual and technical pathways that sustain contemporary human civilization.

I extend my sincere gratitude to Professor Yoshiyuki Asai, Principal Investigator at the Department of Systems Bioinformatics, Graduate School of Medicine, Yamaguchi University, for his supervision, guidance, and continuous support throughout the research process. I am also grateful to Yamaguchi University for waiving my tuition fees, which provided essential financial stability and allowed me to devote myself fully to doctoral research. I would like to express my deep appreciation to the people of Japan for their kindness, generosity, and continuous support throughout my life in Japan.

I would like to express my heartfelt gratitude to the Rotary Club of Ube Nishi and Rotary Hiroshima for providing the Rotary Yoneyama Scholarship, and to the Japan Educational Exchange and Services (JEES) for their invaluable scholarship support. Without this generous assistance, it would not have been possible for me to continue my life and academic pursuits in Japan. Their support has been essential not only for my studies but also for sustaining my family during this period.

Gratitude is also extended to all faculty members and students who contributed through discussions, feedback, and by fostering an intellectually supportive research environment. I am thankful to the secretary of Professor Asai for continuous administrative assistance, and to the faculty members of the Department of Public Health for their encouragement and support during this work.

Finally, my deepest thanks are reserved for my parents for their unwavering sacrifices and steadfast belief in my education. I offer my heartfelt appreciation to my wife Ayesha, and my children Farishta and Ayash, for their patience, understanding, and constant support during the demanding periods of study and writing, without which the completion of this thesis would not have been possible.

Summary of the study

Cognitive decline and dementia are major challenges in aging societies, but predicting who will develop meaningful impairment remains difficult. One reason is that cognitive decline is not driven by a single factor- it reflects the combined effects of metabolism, vascular health, inflammation, nutrition, and everyday behavior. As a result, relying on one test or one biomarker often fails to identify risk early enough for prevention.

In this study, a single community-based cohort of older adults in Japan was analyzed to develop an interpretable, data-driven method for predicting 18-month cognitive status using a broad set of baseline measurements. The study combined routine blood clinical chemistry, targeted metabolomics (including chiral amino acids), and wearable-derived physical activity features, and then applied an explainable two-stage machine-learning strategy to mirror clinical decision-making. In the first stage, the models screened for "any cognitive abnormality" (Possible MCI or Impaired) versus Normal. In the second stage, we focused only on those with abnormal cognition and precisely separated Possible MCI from Impaired.

The models identified a coherent set of predictors that are easy to interpret biologically. In Stage 1, both models highlighted renal and systemic metabolic markers (e.g., creatinine, uric acid, blood urea nitrogen), amino acid and redox-related metabolites (including D-serine, D-amino acid proportions, L-asparagine, alanine, L-glutamic acid, cysteine, methionine sulfoxide), and wearable-derived activity variability (e.g., fluctuation coefficients and steps per minute), with the Simpson index of gut microbiome diversity also contributing in the logistic model. In Stage 2, the models converged on a distinct signature involving glucose and albumin, uric acid and uridine, choline and carnitine, multiple amino acids (such as phenylalanine, proline, ornithine, tryptophan, threonine, and short-chain amino acids), oxidative/energy markers (niacinamide, methionine, methionine sulfoxide, ergothioneine), hematologic indices, and high-MET activity fluctuation metrics. Collectively, these results support a stage-dependent, multisystem view of cognitive aging in which broad renal–metabolic, amino acid, and behavioral vulnerabilities characterize early abnormality, whereas more pronounced alterations in energy metabolism, nucleotide and choline pathways, oxidative stress, and activity irregularity accompany progression from Possible MCI to Impaired status. By combining routine clinical chemistry, targeted metabolomics, gut microbiome diversity, and wearable-derived behavioral measures within an explainable AI framework, this two-stage approach illustrates a scalable, biologically grounded strategy for stage-aware risk stratification and monitoring of cognitive decline in community settings.

Overall, this work shows that combining inexpensive clinical blood tests, targeted metabolites, and wearable activity data—interpreted through explainable machine learning—can help detect early cognitive risk and more precisely classify severity. The long-term goal is to support scalable community screening and earlier, more personalized prevention strategies for

cognitive decline. This dissertation therefore presents an interpretable, stage-aware machine-learning framework for multimodal biomarker discovery in cognitive aging, emphasizing both predictive performance and biological explanation.

Contents

Abstract 1
1 Introduction 2
2 Literature review 4
2.1 Cognitive disorders and Dementia 4
2.2 Mini Mental State Examination (MMSE): 6
2.3 Biomarkers in healthcare 7
2.4 Multimodal biomarkers in Dementia research 8
2.5 Digital biomarkers: physical activity and behavioral phenotyping 9
2.6 Artificial intelligence and machine learning in dementia research 9
2.7 Unsupervised Learning 11
2.8 Supervised Learning 12
2.9 Explainable AI (XAI) in dementia research and multimodal models 12
2.10 Hyperparameter 13
2.11 Random Forest 14
2.12 Logistic Regression, Ridge, and Lasso 15
2.13 Bayesian optimization (BO) 18
2.14 Imbalanced-learn 19
2.15 SHAP Analysis for biomarker discovery 20
3 Problem Statement 22
4 Research Gaps 23
5 Hypotheses 25
6 Conceptual framework 26
7 Methods 28
7.1 Original Cohort Study Design 28
7.2 Participants 28
7.3 Ethical Committee Review 30
7.4 Secondary Analysis Framework 30
7.5 Data 30

7.6 Cognitive Outcome Classification Based on MMSE 31
7.7 Biological Rationale for Selected Predictors 33
7.8 Data Preprocessing 34
7.8.1 Variable Scaling 34
7.8.2 Encoding of Categorical Variables 34
7.8.3 Missing Data Handling 34
7.9 Machine Learning Model Architecture 34
7.10 Validation Strategy 35
7.11 Hyperparameter Optimization 35
7.12 Feature Attribution via SHAP Analysis 38
7.13 Post Hoc Group Comparison 38
7.14 Software Packages 38
8 Results 39
8.1 Stage 1 Performance: Normal vs Abnormal 39
8.2 Stage 2 Performance: Possible MCI vs Impaired 42
8.3 Feature Importance Based on SHAP Analysis 45
8.4 Post Hoc Statistical Validation 54
9 Discussion 59
9.1 Principal Findings and Rationale of the Two-Stage Framework 59
9.2 Biological Signatures in Stage 1 59
9.3 Physical Activity Variability in both Stages 61
9.4 Biological Signatures in Stage 2 62
9.5 Integrative Interpretation and Clinical Implications: 64
10 Limitation of the study 66
11 Conclusion 67
12 References 68
13 List of Training parameters: 80

List of Tables

1. **Table 1:** Inclusion and exclusion criteria for the study 29
2. **Table 2:**Mini-Mental State Examination (MMSE) thresholds used to classify participants' cognitive status at the end of follow-up 32
3. **Table 3:** Optimized hyperparameter settings used for the Random Forest classifiers and Penalized Logistic Regression at each stage of the hierarchical pipeline 37
4. **Table 4:** Stage 1 Classification performance metrics and Clinical Metrics for Normal vs Abnormal cognitive status for both ML pipeline 41
5. **Table 5:** Stage 2 Classification performance metrics and Clinical Metrics for Possible MCI vs Impaired cognitive status for both ML pipeline 44
6. **Table 6:** Descriptive Statistics of Top 20 parameters according to MMSE category from Stage 1 and Stage 2 SHAP analysis of Random Forest classifier 49
7. **Table 7:** Details of Games-Howell Test on Cognitive Categories with corrected p-val 56

List of Figures

1. Figure 1: (a) Confusion matrix for Stage 1 leave-one-out cross-validation (LOOCV) showing predicted vs true labels for Normal vs Abnormal classification for Random Forest.
   (b) Receiver operating characteristic (ROC) curve for the same Stage 1 LOOCV model, reporting overall discrimination with AUC = 0.715 for Random Forest
   (c) Confusion matrix for Stage 1 leave-one-out cross-validation (LOOCV) showing predicted vs true labels for Normal vs Abnormal classification for Penalized Logistic Regression
   (d) Receiver operating characteristic (ROC) curve for the same Stage 1 LOOCV model, reporting overall discrimination with AUC = 0.705 for Penalized Logistic Regression. 40
2. Figure 2: (a) Confusion matrix for Stage 1 leave-one-out cross-validation (LOOCV) showing predicted vs true labels for Normal vs Abnormal classification for Random forest.
   (b) Receiver operating characteristic (ROC) curve for the same Stage 2 LOOCV model, reporting overall discrimation with AUC = 0.95 for Random forest
   (c) Confusion matrix for Stage 2 leave-one-out cross-validation (LOOCV) showing predicted vs true labels for Normal vs Abnormal classification for Penalized Logistic Regression
   (d) Receiver operating characteristic (ROC) curve for the same Stage 2 LOOCV model, reporting overall discrimination with AUC = 0.705 for Penalized Logistic Regression 43
3. Figure 3: (a) SHAP beeswarm plot plot visualizes the top 20 features from Stage 1 that contribute most to the classification between "Normal" and "Abnormal" cognitive status using the optimized random forest model

(b) SHAP beeswarm plot plot visualizes the top 20 features from Stage 1 that contribute most to the classification between “Normal” and “Abnormal” cognitive status using the optimized penalized logistic regression model.. 46

4. Figure 4: (a) Stage 2 beeswarm plot displays the top 20 predictive features for distinguishing between “Possible MCI” and “Impaired” cognitive states.
(b) SHAP beeswarm plot plot visualizes the top 20 features from Stage 2 that contribute most to the classification between “Normal” and “Abnormal” cognitive status using the optimized penalized logistic regression model 48

5. Figure 5: Games- Howell post-hoc comparisons of significant biomarkers and activity parameters ($p < 0.05$, p is corrected by Welch’s degrees of freedom correction and the studentized range distribution) across cognitive groups.
(a) Significant biochemical parameters; (b) significant amino acid markers; (c) significant metabolic parameters; and (d) significant physical activity features (including fluctuation coefficients) are shown. Only parameters with statistically significant group differences after multiple testing correction are displayed, highlighting multi-domain signatures distinguishing Impaired, Normal, and Possible MCI groups.. 55

List of Abbreviations

AAN	American Academy of Neurology
AD	Alzheimer's disease
AI	Artificial intelligence
ALB	Albumin
AMP	Adenosine monophosphate
AMPK	AMP-activated protein kinase.
Asn	Asparagine
ATP	Adenosine triphosphate
AUC	Area under the ROC curve
BASO	Basophils
BMI	Body mass index
BO	Bayesian optimization
bp	Base pairs
BUN	Blood urea nitrogen
CK	Creatine kinase
CRE	Creatinine
CSF	Cerebrospinal fluid
CTP	Cytidine triphosphate
Ala	Alanine
Pro	Proline
Ser	Serine
DNA	Deoxyribonucleic acid
ECM	Extracellular matrix
EI	Expected improvement
EO	Eosinophils
EX	Wearable-derived activity feature label used in the thesis
FDA	Food and Drug Administration
BEST	Biomarkers, EndpointS, and other Tools framework
FTD	Frontotemporal dementia
GABA	Gamma-aminobutyric acid
GLU	Glucose
Gly	Glycine
HbA1c	Hemoglobin A1c
HCT	Hematocrit
HGB	Hemoglobin
HJA-750C	Wearable accelerometer model used for physical activity recording
Arg	Arginine
Asp	Aspartic acid
Cit	Citrulline
Gln	Glutamine
His	Histidine
Ile	Isoleucine
Leu	Leucine
Lys	Lysine
Met	Methionine
Orn	Ornithine
Phe	Phenylalanine

Pro	Proline
Thr	Threonine
Trp	Tryptophan
Tyr	Tyrosine
Val	Valine
LC-MS/MS (LC-MSMS)	Liquid chromatography–tandem mass spectrometry
LBD	Lewy body dementia
LOOCV	Leave-one-out cross-validation
LYMPH	Lymphocytes
MCH	Mean corpuscular hemoglobin
MCHC	Mean corpuscular hemoglobin concentration
MCI	Mild cognitive impairment
MCV	Mean corpuscular volume
MET / METs	Metabolic equivalent(s) of task
ML	Machine learning
MMSE	Mini-Mental State Examination
MONO	Monocytes
MPV	Mean platelet volume
NAD	Nicotinamide adenine dinucleotide
NEUT	N-methyl-D-aspartate
NIH	National Institutes of Health
NMDA	N-methyl-D-aspartate
NRBC	Nucleated red blood cells
P-LCR	Platelet large cell ratio
PCT	Plateletcrit
PDW	Platelet distribution width
PET	Positron emission tomography
PLT	Platelet count
RBC	Red blood cell count
RDW-CV	Red cell distribution width (coefficient of variation)
RDW-SD	Red cell distribution width (standard deviation)
ROC	Receiver operating characteristic
ROC-AUC	Area under the ROC curve
SHAP	SHapley Additive exPlanations
TP	Total protein
UA	Uric acid
UCB	Upper confidence bound (acquisition function)
UMIN-CTR	University Hospital Medical Information Network Clinical Trials Registry
UTP	Uridine triphosphate
VaD	Vascular dementia
WBC	White blood cell count
WHO	World Health Organization
XAI	Explainable artificial intelligence

Abstract

This study developed a two-stage, explainable machine learning framework to predict 18-month MMSE-based cognitive status from baseline multimodal data in community-dwelling older adults in Japan. A hierarchical design was used in which Stage 1 distinguished cognitively Normal participants from those with any abnormality (Possible Mild Cognitive Impairment (MCI) or Impaired), and Stage 2 further separated Possible MCI from Impaired within the abnormal subgroup. Both an Imbalanced-Learn Random Forest and a penalized logistic regression baseline were trained under Leave-One-Out Cross-Validation, yielding fair discrimination in Stage 1 (Random Forest AUC = 0.72, accuracy = 0.71; logistic regression AUC = 0.71, accuracy = 0.76) and apparently strong separability in Stage 2 (Random Forest AUC = 0.95, accuracy = 0.96; logistic regression AUC = 0.82, accuracy = 0.92) in a small sample size with high class imbalance. SHapley Additive exPlanations (SHAP) with TreeExplainer for Random Forest and LinearExplainer for logistic regression were used to identify interpretable biomarkers at each stage through feature attribution. In Stage 1, both models highlighted renal and systemic metabolic markers (e.g., creatinine, uric acid, blood urea nitrogen), amino acid and redox-related metabolites (including D-serine, D-amino acid proportions, L-asparagine, alanine, L-glutamic acid, cysteine, methionine sulfoxide), and wearable-derived activity variability (e.g., fluctuation coefficients and steps per minute), with the Simpson index of gut microbiome diversity also contributed in the logistic model. In Stage 2, the models converged on a distinct signature involving glucose and albumin, uric acid and uridine, choline and carnitine, multiple amino acids (such as phenylalanine, proline, ornithine, tryptophan, threonine, and short-chain amino acids), oxidative/energy markers (niacinamide, methionine, methionine sulfoxide, ergothioneine), hematologic indices, and high-MET activity fluctuation metrics. Collectively, these results support a stage-dependent, multisystem view of cognitive aging in which broad renal–metabolic, amino acid, and behavioral vulnerabilities characterize early abnormality, whereas more pronounced alterations in energy metabolism, nucleotide and choline pathways, oxidative stress, and activity irregularity accompany progression from Possible MCI to Impaired status. By integrating routine clinical chemistry, targeted metabolomics, gut microbiome diversity, and wearable-derived behavioral measures within an explainable AI framework, this two-stage approach provides a biologically interpretable and potentially scalable strategy for stage-aware risk stratification and monitoring of cognitive decline in community settings.

1 Introduction

Dementia and Mild cognitive impairment (MCI) in aging population have put significant challenges all over the world from poor to developed nations. These conditions are mostly neurodegenerative disorders associated with multiple physiological systems and lifestyles affect different faculties of brain function. Among dementia Alzheimer's disease (AD) is the most common and caused mainly by brain cell death. Mild cognitive impairment (MCI) is widely regarded as an intermediate clinical state between normal aging and dementia, yet its progression is highly heterogeneous, with only a subset of individuals transitioning to Alzheimer's disease (AD) within a few years of diagnosis. However, statistics show that only 20–40 % of individuals with MCI will transition to AD within three years of diagnosis [1] . Researchers are continuously investigating promising biomarkers for prediction, prognosis and prevention of the onset of these conditions. While literature suggests biomarker changes align with dementia development, no single biomarker adequately forecasts the conversion of MCI (Mild cognitive impairments) patients or healthy individuals to dementia with sufficient accuracy and ample lead time before the initial appearance dementia. Although many candidate biomarkers have been investigated for prediction, prognosis, and prevention, no single biomarker has shown sufficient accuracy and lead time to reliably forecast future cognitive deterioration in community-dwelling older adults. One reason is that dementia is not a single disease entity but a biologically and clinically heterogeneous syndrome involving multiple interacting pathways, including vascular dysfunction, metabolic dysregulation, neuroinflammation, and synaptic failure. A critical aspect of current diagnostic approaches is their reliance on costly tools (such as brain imaging) and invasive clinical procedures (like amyloid-PET (Positron Emission Tomography) scans and CSF (Cerebro Spinal Fluid) analysis), often necessitating highly specialized personnel [2,3]. Diagnostic standards, such as amyloid positron emission tomography and cerebrospinal fluid analysis, provide valuable pathophysiological information but remain costly, invasive, and difficult to scale for broad community screening. Artificial Intelligence (AI) and Machine Learning have changed the world, developing real world solutions and applications in all domains of our life. However, the adoption of machine learning approaches in healthcare has been slower than in other fields despite the increasing pressure on healthcare systems despite it can accelerate personalized and precision care [4,5]. Recent advances in artificial intelligence (AI) and machine learning have created new opportunities for precision medicine by enabling the integration of large-scale multimodal data, including biobanks, electronic health records, medical imaging, omics technologies, and wearable sensor data. Biomedical data from large biobanks, electronic health records, medical imaging, wearable and ambient biosensors, and the lower cost of genome and microbiome sequencing have created the window to develop multi domain based artificial intelligence solutions which will be capable of fulfilling the intricate complexity of human health and disease[6]. Biomarkers are defined as measurable properties that represent any biological processes, pathogenic processes, or responses to different event[7]. In today's world of information, with possible large multimodal cohorts and high-throughput technologies, the potential for discovering novel biomarkers is no longer limited by data and conventional study designs[8]. At the same time, the central challenge has shifted from data scarcity to the extraction of clinically meaningful and interpretable patterns from complex, high-dimensional

datasets. Improving the clinical management of dementia requires reliable biomarkers that can aid in identifying high-risk populations, early diagnosis, accurate subtyping, and predicting prognosis, drug response, or adverse events. Biomarkers from different biological processes have been investigated in people with dementia, including neuroimaging[9], electrophysiological[10], genetic[11], gene expression[12], protein[13], metabolic[14], gut microbial[15], sleep[16], gait [17], and digital biomarkers[18]. Biomarkers from cerebrospinal fluid (CSF), as well as minimally invasive collectable biological fluids including blood [19] have shown promise in improving dementia diagnosis. Artificial intelligence (AI) and machine learning approaches have been developed to detect novel biomarkers and interactions in complex data sets. Machine learning approaches have increasingly been applied to predict cognitive decline from multidomain data, but many existing studies rely on single-step classification strategies that directly map baseline measurements to final diagnostic categories. Such designs may overlook stage-specific biological and behavioral mechanisms that emerge during progression from normal cognition to MCI and more severe impairment. MMSE (Mini Mental State Examination) is commonly used to stage impairment severity and to track progression in conditions such as Dementia (Alzheimer's disease etc.) because declines in total score reflect clinically meaningful loss of function across multiple domains relevant to neurodegeneration. The MMSE score was proposed by Marshal Folstein is one of the most widely used instrument for cognitive screening in clinical surveys and epidemiologic surveys[20]. The Mini-Mental State Examination (MMSE) is a 30-point, standardized screening tool that assesses global cognition across orientation, attention, memory, language, and simple visuospatial construction, making it well suited for longitudinal monitoring of age-related cognitive change. In dementia research, for this study, longitudinal MMSE served as the primary outcome because it is brief, interpretable at the bedside, repeatedly obtainable at multiple visits, and harmonized across participants and timepoints, enabling consistent labeling for predictive modeling. Studies have been conducted to investigate longitudinal rate of decline in cognitive function using machine learning based on data from multiple domains[21]. The black box nature of Machine learning models can be explained by SHAP (SHapley Additive exPlanations) to gain insights on various features' predictive effect on models. It is a popular feature-based model agnostic interpretability method for regression and classification problem[22]. SHAP analysis is rooted in Shapley values, a concept from collaborative game theory. Shapley values provide a fair distribution of a payout among players in a collaborative game where players work together for a common goal, even if the players may have contributed unequally. Recently machine learning techniques have been widely used for integrative analysis of complex diseases for biomarker discovery[23].

2 Literature review

2.1 Cognitive disorders and Dementia

Dementia represents a major and growing global health burden, characterized by progressive decline in memory and other cognitive domains that ultimately interfere with independent functioning. Among dementia subtypes, Alzheimer's disease (AD) is the most common neurodegenerative cause and is biologically heterogeneous, involving amyloid and tau pathology, neuroinflammation, vascular contributions, synaptic dysfunction, and systemic metabolic changes[24,25]. Mild cognitive impairment (MCI) is widely recognized as an intermediate clinical state between cognitively normal aging and dementia, but conversion rates to AD vary substantially and a considerable proportion of individuals remain stable or improve, complicating prognosis in clinical practice[26,27]. Dementia represents one of the most consequential challenges in aging societies because it progressively compromises cognition and independent living, producing large personal, healthcare, and socioeconomic burdens. Although Alzheimer's disease (AD) is commonly described as the most frequent neurodegenerative cause of dementia, AD itself is biologically heterogeneous and often coexists with other processes such as vascular injury, neuroinflammation, metabolic dysfunction, and synaptic failure, which together shape symptom onset and rate of decline[25,28]. This heterogeneity complicates clinical prognostication: individuals with apparently similar cognitive presentations may follow markedly different trajectories due to differences in underlying pathophysiology and systemic vulnerability[27,28]. Many individuals classified as MCI remain stable for years or even revert to normal-range performance, while others progress to dementia. This variability highlights a limitation of symptom-based categorization: cognitive performance at any single time reflects not only neuropathology but also cognitive reserve, education, mood, sleep, systemic illness, and measurement noise.

Dementia is an umbrella term for a syndrome of progressive cognitive and functional decline that can be caused by multiple distinct brain diseases, and the type matters because each disease tends to affect different brain systems and produces different symptom profiles. The most commonly discussed subtypes in clinical and research settings include Alzheimer's disease, vascular dementia, Lewy body dementia, and frontotemporal dementia, and mixed forms are also common. Alzheimer's disease (AD) is characterized by abnormal protein deposits in the brain, including amyloid plaques and tau tangles, and it commonly presents with progressive impairment in memory and other cognitive domains over time. Vascular dementia (VaD) / vascular cognitive impairment is associated with disease of brain blood vessels and reduced blood supply to brain tissue, sometimes with a more sudden onset after vascular events, and mixed AD–vascular pathology can occur in the same person[28]. Lewy body dementia (LBD) involves abnormal alpha-synuclein deposits ("Lewy bodies"), and it can feature cognitive decline along with problems like reduced attention/alertness and other characteristic neuropsychiatric/behavioral features. Frontotemporal dementia (FTD) is linked to abnormal accumulation of proteins such as tau and TDP-43 in frontal/temporal lobes and is often marked

by early changes in behavior, personality, executive control, and/or language rather than an amnestic presentation. Some resources describe Lewy body disease as a spectrum that includes Parkinson's disease dementia, with the clinical label partly depending on whether movement symptoms or cognitive/behavioral symptoms appear first. This heterogeneity is a key reason symptom-only labels can be unstable early on: the same "mild" cognitive presentation may reflect different underlying pathologies with different progression patterns and intervention targets. Mild cognitive impairment (MCI) is widely regarded as an intermediate clinical state between cognitively normal aging and dementia and is clinically important because it is associated with elevated risk of future decline. However, progression from MCI is highly variable: some individuals remain stable for years, some revert to normal-range performance, and others progress to dementia, which makes prognosis difficult when based on symptoms alone. This variability reflects the fact that cognitive performance at a single time point is shaped not only by neuropathology but also by cognitive reserve, education, mood, sleep, systemic illness, comorbidities, and measurement variability. Accordingly, symptom-based categorization without complementary biological or functional markers may be insufficient for identifying which individuals are on a high-risk trajectory. Mild cognitive impairment is clinically important because it is associated with elevated risk of future dementia, but progression is variable, which makes prognosis difficult without additional biological and functional markers. A robust meta-analysis of inception cohorts reported adjusted annual conversion rates from Mayo-defined MCI to dementia of about 5–10% (with higher rates in specialist settings and lower rates in community studies), showing both meaningful risk and substantial heterogeneity. This variability creates a practical clinical need: identifying which "suspected MCI" patients are on a higher-risk trajectory enables prioritization of monitoring, work-up, and preventive action while avoiding over-medicalization of those likely to remain stable.

Contemporary clinical guidance frames mild cognitive impairment (MCI) not only as a diagnostic label, but as a management opportunity in which clinicians should conduct a structured evaluation and address potentially modifiable contributors to cognitive symptoms, including a careful medication and comorbidity review and broader clinical risk profiling. This "clinical management plus prevention" orientation is particularly important because MCI is heterogeneous in cause and trajectory, and early-stage care can focus on reducing avoidable cognitive stressors while simultaneously implementing evidence-informed strategies that may slow decline. In parallel, global public-health guidance emphasizes risk reduction as a medium-term prevention focus that can be deployed even at the MCI stage, explicitly treating early identification as a window to act on modifiable pathways rather than waiting for irreversible functional deterioration. The WHO dementia risk-reduction guidelines describe several intervention domains relevant to adults with normal cognition and/or MCI, including physical activity, dietary patterns (including Mediterranean-like approaches), management of vascular and metabolic risks (notably hypertension, diabetes, and dyslipidaemia using established clinical guidelines), tobacco cessation, reduction of hazardous alcohol use, and cognitive training, with recommendation strength varying by domain and evidence base. Consistent with this approach, the American Academy of Neurology (AAN) practice guideline update states that for patients diagnosed with MCI, clinicians should recommend regular exercise (twice per

week) as part of management. This broader preventive perspective strengthens the rationale for developing multimodal biomarkers that can improve early risk stratification, distinguish heterogeneous pathways of decline, and support more personalized monitoring in community and clinical settings.

2.2 Mini Mental State Examination (MMSE):

The Mini-Mental State Examination (MMSE), originally introduced by Marshall F. Folstein, Susan E. Folstein, and Paul R. McHugh in 1975, is a brief, standardized instrument used to quantify global cognitive status and is widely applied in both clinical practice and population-based research to support the assessment of cognitive impairment[20]. It is typically administered as part of a structured cognitive evaluation when clinicians assess individuals with cognitive complaints or suspected mild cognitive impairment (MCI), alongside history, functional assessment, and broader medical review[29]. In longitudinal cohort studies, repeated MMSE measurements are especially useful because they provide a consistent numeric outcome that can be tracked over time and operationalized as either a continuous endpoint (change in total score) or a categorical endpoint (e.g., severity bands or transitions between cognitive states).

The MMSE is scored on a 30-point scale, with higher scores indicating better performance, and it samples multiple cognitive domains within a short administration time. Its items broadly cover orientation, immediate registration, attention/calculation, delayed recall, language functions, and a simple visual construction task, thereby producing a compact summary of performance across core cognitive systems rather than an in-depth neuropsychological profile. Because of its brevity and standardization, MMSE has historically been attractive for large studies and routine practice where time and resources constrain more comprehensive testing.

Despite these advantages, MMSE has important methodological limitations that become particularly salient when MMSE is used as a labeling variable for predictive modeling. First, a single total score compresses heterogeneous cognitive profiles into one number, meaning that different patterns of domain impairment can yield the same MMSE total, which reduces etiologic specificity and can obscure subtype-relevant signals. Second, the instrument is vulnerable to boundary instability when cutoffs are used (e.g., “normal” vs “possible MCI”), because small score fluctuations—driven by fatigue, mood, sleep, intercurrent illness, testing conditions, or rater effects—can shift an individual across a categorical threshold even if underlying neurobiology is unchanged[30]. Third, ceiling effects in high-functioning individuals and reduced granularity at lower scores can limit sensitivity to subtle early decline and restrict resolution in advanced impairment, respectively, which can distort both longitudinal change estimation and class separation near the extremes[31].

2.3 Biomarkers in healthcare

A biomarker is defined as a measured characteristic indicating normal biological processes, pathogenic processes, or responses to an exposure or intervention, and this definition intentionally spans molecular, histologic, radiographic, and physiologic measurements[7]. In healthcare delivery, biomarkers are valuable because they can add objective information to clinical assessment, but they should be distinguished from clinical outcome assessments (COAs), which directly measure how a person feels, functions, or survives[32].

The FDA-NIH BEST-aligned framework classifies biomarkers by purpose, including diagnostic biomarkers (detect/confirm disease or subtype), monitoring biomarkers (serial assessment of disease status or exposure), pharmacodynamic/response biomarkers (change with exposure to an intervention), predictive biomarkers (identify who is more likely to respond or be harmed by an intervention), prognostic biomarkers (likelihood of progression or clinical events in those with disease), safety biomarkers (likelihood/presence/extent of toxicity), and susceptibility/risk biomarkers (likelihood of developing disease in currently unaffected individuals). A single measurement can sometimes serve more than one role, but evidence must be generated for each specific "context of use," meaning the same biomarker can be appropriate in one setting and misleading in another. In health systems, these categories map naturally to key workflows-screening/risk assessment, diagnosis/subtyping, longitudinal monitoring, treatment selection, and adverse-event surveillance-each requiring different performance characteristics and clinical thresholds[33].

Moving a biomarker from discovery to routine healthcare requires more than demonstrating association; it requires analytical validation (reliable measurement), qualification based on evidentiary assessment, and demonstration of appropriate utilization in the intended setting[34]. When a biomarker is used as a study endpoint, it must be precisely defined, including how and when it is measured and how repeated measures are handled, because endpoint rigor determines reproducibility and interpretability[33]. A recurring pitfall is assuming that correlation with clinical outcomes makes a biomarker a valid surrogate endpoint; surrogate use requires strong evidence that changes in the biomarker explain changes in true clinical outcomes, typically across multiple therapies and contexts[35].

Healthcare is increasingly adopting complex and composite biomarkers, motivated by the recognition that single measures may be insufficient to represent multidimensional disease biology and that combined signals can improve prediction and decision-making[36]. Digital biomarkers—derived from sensors and personal devices-are an additional frontier, enabling more frequent, lower-burden measurements of functions such as activity, cognition-related behaviors, or physiological rhythms, although standards linking these measures to meaningful outcomes are still evolving[37]. In dementia and related disorders, the expansion of multimodal cohorts and high-throughput technologies has accelerated biomarker discovery, and AI/ML methods are increasingly used to interrogate high-dimensional data and interactions, with translation constrained by feasibility, invasiveness/cost of some modalities, dataset bias, and the need for robust real-world validation[38].

2.4 Multimodal biomarkers in Dementia research

Multimodal biomarkers in Alzheimer's disease and dementia research refer to integrated measurement strategies that combine heterogeneous data streams- such as neuroimaging, fluid biomarkers, genetics and other omics, electrophysiology, cognitive and clinical measures, and increasingly digital phenotyping- to capture different levels of disease biology and functional impact within a unified framework[39]. This approach is driven by the recognition that dementia syndromes are biologically and clinically heterogeneous, and that reliable susceptibility, diagnostic, and prognostic biomarkers are unlikely to emerge from a single modality alone[40]. Large multimodal cohorts such as ADNI, PREVENT-AD, and the Deep and Frequent Phenotyping studies explicitly collect structural and molecular imaging (MRI, PET), CSF and blood analytes (including Aβ42/40, t-tau, p-tau, neurofilament light and other proteins or metabolites), genomic and transcriptomic profiles, and repeated cognitive assessments, sometimes augmented by wearable- or sensor-derived digital measures of gait, sleep, and daily behavior[41]. Similarly, the PREVENT-AD program provides an open-science longitudinal dataset that includes multimodal imaging, genetics, fluid biochemistry, neurosensory measures, and cognition in at-risk older adults[42]. The Deep and Frequent Phenotyping (DFP) study protocol also emphasizes repeated, multi-domain phenotyping- including PET, CSF, structural/functional MRI, electrophysiology, and bio sampling- while additionally evaluating smartphone- and wearable-based approaches for high-frequency functional monitoring[43]. Empirical analyses in these datasets show that combining modalities can improve classification and prediction performance- for example, models integrating MRI, PET, CSF proteins, and APOE genotype outperform single-modality approaches for distinguishing AD from MCI, and combinations of relatively accessible measures (plasma biomarkers, genetic risk scores, cognitive tests) can reach performance comparable to imaging-based models for predicting conversion from MCI to AD. At the same time, mobile and wearable technologies provide continuous or high-frequency readouts of motor function, speech and language, eye movements, autonomic physiology, sleep, and neuropsychiatric behavior, offering a rich layer of digital biomarkers that can be passively collected and fused with traditional molecular and imaging markers to support earlier detection and longitudinal monitoring at scale[44]. Conceptually, this trend towards multimodal panels aligns with regulatory biomarker frameworks that emphasize composite and digital biomarkers and stress that each candidate marker (or panel) must be validated for a specific context of use- diagnostic, prognostic, predictive, monitoring, safety, or susceptibility- with appropriate analytical validation, evidentiary qualification, and demonstration of clinical utility[33]. Methodologically, the integration of high-dimensional multimodal data has accelerated the adoption of AI and machine-learning techniques, including supervised and unsupervised models, but these gains are counterbalanced by challenges related to sample size, overfitting, missingness across modalities, cost and invasiveness of some measures, and limited diversity and real-world representativeness of existing cohorts[33].

2.5 Digital biomarkers: physical activity and behavioral phenotyping

Digital biomarkers for physical activity and behavioral phenotyping leverage mobile/wearable sensors to capture objective, high-frequency, real-world measures of movement, mobility, routines, and social behavior that may shift years before overt clinical presentation of Alzheimer's disease (AD) [18]. For physical activity, inertial sensors (accelerometer/gyroscope), geo-positioning, and actigraphy-derived measures can quantify gait speed, stride length, symmetry, variability, and overall activity level either via short, prompted tasks or passively during everyday life, with longitudinal trajectories potentially revealing subtle change points that could contribute to composite early-warning scores when combined with other domains[45]. For behavioral phenotyping, passive streams such as GPS/location variance and habitual mobility patterns, device usage logs, and inferred social interaction signals (e.g., communication frequency or changes in social circle size) can operationalize neuropsychiatric and functional manifestations—apathy, depression-related withdrawal, reduced time outside the home, and disruption of daily routines—that are common in AD and can precede diagnosis[46]. Framed within contemporary biomarker terminology, these sensor-derived measures can function as monitoring biomarkers when tracked serially, and they increasingly appear as complex/composite and "digital" biomarkers that require a clearly defined context of use and rigorous validation to ensure they are reliable and clinically meaningful[33]. Across both physical activity and behavioral phenotyping, the main promise is dense longitudinal measurement with low patient burden, while key implementation challenges include managing large continuous datasets and establishing ground truth through well-designed longitudinal studies that link digital phenotypes to established clinical and biological assessments[47].

2.6 Artificial intelligence and machine learning in dementia research

Machine learning (ML) methods are increasingly used in dementia research to model high-dimensional, multimodal data and to detect complex nonlinear interactions that are difficult to capture with classical statistical modeling[39,48–50]. ML can support risk stratification, early detection, subtype differentiation, and prognosis prediction, particularly when combining heterogeneous data such as clinical labs, omics profiles, and sensor-derived signals. However, ML deployment in healthcare remains constrained by issues of interpretability, generalizability, bias, and clinical utility, requiring careful validation strategies and transparent reporting[51].

Multimodal biomedical AI has expanded rapidly alongside growth in biobanks, electronic health records, and sensor data, creating opportunities to move from single biomarker hypotheses toward data-driven biomarker discovery[52,53]. For cognitive outcomes, ML is well-suited to settings where risk is distributed across multiple systems (metabolism, vascular health, inflammation, behavior), and where early signals may be subtle and dispersed. Yet dementia modeling often faces practical challenges, including small sample sizes, missing data, and highly imbalanced outcomes (few severe impairment cases), which can distort model training and produce misleading performance if not addressed with appropriate methods[54].

A recurring theme is that dementia is heterogeneous (mixed pathology, variable trajectories), so AI is leveraged to detect multivariate patterns rather than relying on a single biomarker or a single test. AI/ML is used to develop susceptibility/risk models (who is likely to develop dementia), diagnostic models (who currently have dementia/subtype), and prognostic models (how quickly decline will progress), which aligns with standard biomarker categories used in dementia research. These papers also emphasize that AI/ML's practical value increases when models can use accessible measures (e.g., blood analytes, wearables, smartphone-derived behavior) rather than expensive or invasive measures alone (e.g., PET, CSF), even if those remain important reference standards[55].

Modern dementia research increasingly uses large cohorts with repeated measures over time and multiple modalities (imaging, omics, clinical data, digital measures), which creates the scale/complexity where AI/ML is most useful. Examples of modalities explicitly highlighted include neuroimaging (MRI/PET), electrophysiology (e.g., EEG), genetics and gene expression, proteomics, metabolomics, gut microbial measures, sleep and gait, and "digital biomarkers" from mobile/wearable devices. Digital biomarkers (from phones/wearables) are framed as a way to obtain objective, high-frequency longitudinal data with relatively low participant burden, enabling passive "digital phenotyping" of mobility, activity, sleep, and behavior that may shift before overt clinical diagnosis. The papers consistently organize ML into supervised learning (learn mapping from predictors to known labels such as diagnosis/MMSE category) versus unsupervised learning (discover structure, clusters, subtypes, or latent representations without labels). Unsupervised methods are used for tasks like clustering patient subgroups (e.g., biomarker-guided clustering of heterogeneous AD-related profiles) and dimensionality reduction to help explore structure in high-dimensional datasets. Because longitudinal labels are expensive and follow-up may be incomplete, the biomarker-discovery review notes the role of semi-supervised/unsupervised learning and deep learning to handle limited labeled outcomes and complex interactions. From a broader clinical-AI framing, AI models can also be used for prediction from electronic health records and multimodal clinical data, but the need for prospective, real-world validation is repeatedly emphasized before clinical adoption. A major translational requirement in dementia is moving from "black-box" prediction toward explainable and auditable evidence, so that candidate biomarkers are interpretable, testable, and more likely to generalize. Across the dementia biomarker discovery review, translation is positioned as requiring replication across cohorts, real-world validation (not only internal cross-validation), and attention to practical feasibility (cost, invasiveness, scalability) of the final biomarker panel. Biomarker terminology also matters for translation: digital and composite biomarkers still require clear "context of use" (diagnostic vs monitoring vs prognostic), with analytical validation and evidentiary qualification to avoid misapplication. High-dimensional biomarker discovery is statistically fragile: ML can overfit, especially when features far exceed sample size (a common situation in omics), producing models that look strong internally but fail to generalize. The dementia biomarker discovery review highlights dataset limitations as a major barrier—especially lack of participant diversity and cohort bias—which can cause performance to drop sharply when models are tested outside the original dataset. Relatedly, the broader medical-AI review underscores limitations such as bias, privacy/security, and lack of transparency, and stresses

that high retrospective performance does not imply clinical efficacy. Reporting and evaluation standards are treated as essential for trustworthy dementia AI (e.g., emphasis on reproducibility, appropriate validation, and movement toward structured reporting guidance for prediction models). Models should be judged by real-world benefit (clinical outcomes, workflow impact, equitable access), not just summary metrics like AUC from retrospective datasets[56].

Common supervised algorithms in dementia biomarker work include regression models, support vector machines, random forests, gradient boosting, and deep learning, with model choice depending on the modality (e.g., image-based deep learning vs. tree-based methods for tabular biomarkers)[57].

2.7 Unsupervised Learning

Unsupervised learning refers to a family of machine learning methods that infer patterns and structure from unlabeled data, meaning that no explicit outcome (e.g., "Alzheimer's disease vs control") is provided during training. In dementia research, it is used to address clinical and biological heterogeneity by discovering latent structure—such as data-driven subtypes (endotypes), spectra, or clusters of individuals with similar multimodal biomarker profiles—rather than forcing all participants into predefined diagnostic categories. This paradigm is particularly relevant because modern dementia cohorts often contain high-dimensional and multimodal measurements (e.g., imaging, omics, clinical features, and digital measures), where relationships among variables may be nonlinear and not well captured by traditional single-modality or hypothesis-driven analyses[58,59].

In practice, unsupervised pipelines commonly combine dimensionality reduction (to compress thousands of features into a lower-dimensional latent space) with clustering algorithms (to group participants into coherent profiles), after which clusters are interpreted post hoc using clinical characteristics and biomarker patterns. These approaches can support biomarker-guided stratification and the identification of patient subgroups that may reflect distinct underlying mechanisms, which is important for precision medicine and for improving the design of downstream predictive models and clinical trials. Unsupervised methods are also useful in settings where diagnostic labels are missing, noisy, or inconsistently defined across sites (e.g., population surveys and real-world clinical datasets), enabling discovery and screening based on data structure rather than label quality. However, unsupervised results can be sensitive to preprocessing choices, hyperparameters, and cohort artifacts (including confounding and site effects), and clusters may not replicate across independent datasets if they capture dataset-specific biases rather than generalizable biology. Consequently, robust validation across cohorts, transparent reporting, and careful interpretability analyses are essential before treating unsupervised clusters as clinically meaningful dementia subtypes[60,61].

2.8 Supervised Learning

Supervised learning is a machine learning paradigm in which models are trained on labeled examples to learn a mapping from input features (predictors) to a known outcome (label), such as diagnosis (AD/MCI/control), dementia subtype, conversion from MCI to dementia, or future cognitive decline. In dementia research, supervised learning is widely used because it can integrate large numbers of variables across modalities—neuroimaging, fluid biomarkers (CSF/plasma), genetics/omics, neuropsychological measures, and digital biomarkers—and quantify how well these combined signals predict clinically meaningful endpoints. This makes supervised learning especially valuable for developing susceptibility/risk models, diagnostic classifiers, and prognostic models that aim to predict progression or trajectories rather than relying on single markers or single time-point assessments[62].

A typical supervised learning workflow includes feature preprocessing and selection, model training on a labeled dataset, and rigorous evaluation using hold-out testing or cross-validation to estimate generalization to unseen data; this is critical because dementia datasets are often heterogeneous and can be prone to overfitting, particularly in high-dimensional settings. Common algorithm families include regularized regression, support vector machines, and tree-based ensemble methods (e.g., random forest and gradient boosting) for tabular biomarker data, as well as deep learning architectures for imaging and other high-dimensional inputs where representation learning can be advantageous. Model performance is usually quantified using metrics such as sensitivity/specificity and ROC-AUC, but multiple metrics are recommended because accuracy alone can be misleading under class imbalance or asymmetric clinical costs of false positives vs false negatives. Supervised pipelines are increasingly paired with explainability methods (e.g., SHAP) to rank predictors and provide per-individual feature attributions, supporting biomarker discovery and improving the transparency needed for clinical trust and translation.

2.9 Explainable AI (XAI) in dementia research and multimodal models

Explainable artificial intelligence (XAI) refers to methods that make the outputs of AI/ML models understandable to humans by providing evidence for why a particular prediction was produced (e.g., which variables, modalities, or regions of an image contributed most). XAI is especially important in dementia research because models are increasingly trained on complex, high-dimensional, multimodal datasets (imaging, omics, fluid biomarkers, clinical variables, and digital biomarkers), where strong predictive performance alone does not guarantee biological validity, clinical trust, or generalizability. A recurring barrier to real-world adoption of AI in medicine is the “black box” nature of many models, so transparency and auditability are needed to support clinician confidence, reduce harmful failures, and bridge the gap between retrospective performance and true clinical utility[63].

In multimodal dementia models, explainability plays a distinct role because predictions often emerge from feature interactions across modalities rather than a single marker, and XAI helps disentangle which data sources drive the decision (and whether that reliance is clinically

plausible). For example, explanation methods can reveal whether a model is dominated by one convenient modality (e.g., demographics or site-related imaging artifacts) instead of disease-relevant signals, which is critical given known risks of dataset bias and poor generalization when models are trained on limited or non-representative cohorts. In imaging-based dementia models, visual explanation techniques (e.g., saliency/attention maps, Grad-CAM, and relevance propagation families) are used to highlight which brain regions most influenced classification or staging, supporting both scientific interpretability and quality control of model behavior. In non-imaging tabular or multi-omics settings, feature-attribution methods such as SHAP and LIME are commonly used to quantify global biomarker importance and provide patient-level explanations, turning a predictive model into a structured biomarker ranking and hypothesis-generation engine[64,65].

From a biomarker discovery standpoint, XAI is valuable because it can reduce "blind" reliance on performance metrics by showing which predictors are being used, enabling mechanistic scrutiny before costly downstream validation and translation. This is particularly relevant in high-dimensional biomedical data where ML can overfit and identify spurious signals that appear strong in internal validation but fail to replicate, making interpretability a practical safeguard as well as a scientific tool. In your multimodal, two-stage cognitive decline modeling work, SHAP-based interpretation is explicitly used to identify and communicate which routine clinical chemistry markers, metabolite/amino-acid measures, and wearable-derived behavioral features contributed to each stage of prediction, illustrating how XAI supports both scientific narrative and clinical plausibility in multimodal models. More broadly, recent dementia-focused XAI literature emphasizes that explanations can be used for "clinical alignment" (checking whether model reasoning matches known clinical patterns) and for quality assurance to detect biases or inappropriate shortcuts, which are key requirements for trustworthy deployment of dementia AI systems[66].

2.10 Hyperparameter

Hyperparameters are configuration choices set before model training that control either the model's structure (e.g., number of layers/units in a neural network, maximum depth/number of trees in tree ensembles) or the learning process itself (e.g., learning rate, batch size, regularization strength), and they differ from model parameters (weights/coefficients) that are learned directly from the training data. In biomarker and multimodal clinical research, hyperparameters are especially influential because they strongly determine model capacity and inductive bias, thereby shaping the underfitting–overfitting trade-off and the stability of discovered "signals" in high-dimensional settings. For example, in tree-based methods, deeper trees and weaker regularization can increase apparent training performance but often reduce generalization, while in neural networks, architecture and optimization hyperparameters can determine whether training converges reliably and whether the learned representation captures disease-relevant patterns or dataset artifacts. Hyperparameters also matter beyond supervised prediction: in unsupervised methods commonly used in biomedical pipelines (e.g., t-SNE),

outputs can be sensitive to user-set hyperparameters, meaning that inferred clusters or separations may change substantially with different settings and should be treated as analysis choices rather than ground truth[67].

Hyperparameter tuning (hyperparameter optimization) is the iterative process of selecting hyperparameter values that maximize a predefined objective (e.g., cross-validated AUC, sensitivity at a fixed specificity, calibration error) while maintaining generalization to unseen data. Common tuning strategies include manual/heuristic tuning, grid search (systematic enumeration over a predefined set), random search (sampling combinations from distributions), and Bayesian optimization (probabilistic search that uses past evaluations to choose promising new settings), with the choice driven by the size of the search space, computational budget, and the expected interactions among hyperparameters. A key methodological requirement is to separate data used for tuning from data used for final evaluation to prevent optimistic bias; in practice, tuning is performed using internal cross-validation or a validation split, and the final model is then assessed once on a held-out test set, with "nested" cross-validation often preferred when sample size is limited and the risk of overfitting is high. This is particularly important in omics and multimodal dementia-related datasets where the feature dimension can be very large relative to the number of participants, making it easy for flexible models and extensive hyperparameter searches to fit noise, inflate performance estimates, and generate non-reproducible biomarker rankings without rigorous validation design[68,69].

2.11 Random Forest

Random Forest (RF) is a supervised ensemble learning algorithm originally proposed by Leo Breiman in 2001, who provided a rigorous mathematical foundation demonstrating that the generalization error of a forest converges to a limit as the number of trees increases, and is bounded by the strength of individual classifiers and the correlation between them. It builds a collection of decision trees and combines their predictions into a single output, typically using majority voting for classification and averaging for regression. The core idea is to reduce the instability and overfitting tendency of a single decision tree by training many trees on different bootstrap samples of the training data (sampling with replacement) and injecting additional randomness by restricting the set of candidate features considered at each split. This "bagging + random feature selection" mechanism decorrelates individual trees so that their errors are less likely to align, which often improves generalization performance on unseen data compared with a single tree[70].

A practical advantage of RF in biomedical and dementia-related biomarker modeling is that it handles nonlinear relationships and higher-order interactions without requiring explicit specification of interaction terms, which can be valuable when combining heterogeneous clinical chemistry, metabolomics, wearable-derived features, and other modalities. Tree-based ensembles (including RF and gradient boosting) are frequently strong baselines for tabular biomarker datasets and, in many real-world settings, can outperform more complex neural models when sample sizes are limited or when the signal-to-noise ratio is modest. In dementia biomarker discovery work, RF-style models are also attractive because they can operate on

mixed feature types, tolerate collinearity better than many linear models, and provide built-in routes to ranking variables for downstream biological interpretation (with the caveat that importance measures require careful validation)[71,72].

Key RF hyperparameters control model capacity, bias–variance trade-offs, and computational cost, and therefore strongly influence performance and reproducibility. The most important tuning parameters include the number of trees (often called n_estimators), the maximum tree depth (or related constraints such as minimum samples per leaf), and the number of candidate features evaluated at each split (often called max_features). A commonly used default heuristic is to consider p features at each split for classification (where p is the total number of input features), while regression often uses more features per split, though optimal settings are data-dependent. Because RF uses bootstrap sampling, many implementations also enable out-of-bag (OOB) evaluation, where each tree can be tested on samples not included in its bootstrap draw, providing an internal estimate of generalization error without requiring a separate validation split (though external validation remains preferred for publication-quality claims)[73].

Interpretability in RF is often addressed via feature-importance analyses, but these must be handled carefully in biomarker discovery where overfitting and cohort effects are common risks. A widely used approach is permutation importance, which measures the increase in prediction error when a feature's values are randomly permuted in out-of-bag samples, thereby estimating how much the model depends on that feature for predictive performance. However, RF importance rankings can be unstable across resamples and can be biased by correlated predictors or by variables with many possible split points, so robust reporting typically requires repeated cross-validation, sensitivity analyses, and ideally replication in an independent cohort before interpreting "top biomarkers" as biologically meaningful[74].

2.12 Logistic Regression, Ridge, and Lasso

Logistic regression is one of the most widely used statistical models for binary outcomes and is a standard method when the dependent variable has two categories, such as disease versus no disease, responder versus non-responder, or normal versus abnormal status. In scientific research, it is valued because it models probabilities directly, produces interpretable coefficients on the log-odds scale, and provides a statistical baseline against which more complex machine-learning models can be compared. In modern applied statistics, logistic regression is treated as a classical model for binary classification, while its penalized extensions became especially important in high-dimensional and correlated data settings. The mathematical foundation of logistic regression begins with the idea that, for a binary outcome $Y_i \in \{0, 1\}$, the conditional probability of the event can be modeled as a smooth function of the predictors rather than by ordinary linear regression. Specifically, logistic regression assumes that the probability of the event is linked to a linear predictor through the logistic function.

$$p(y_i = 1|x_i) = \pi_i = \frac{1}{1 + exp[-(\beta_0 + x_i^T \beta)]}$$

Equivalently, the model is linear on the log-odds scale, which is called the logit transformation.

$$log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + x_i^T \beta$$

This formulation is philosophically attractive because it maps any real-valued linear combination of predictors to a valid probability between 0 and 1, while preserving a simple linear structure in the transformed space. As a result, each coefficient describes the change in log-odds associated with a one-unit increase in the corresponding predictor, holding other variables constant. Under the Bernoulli sampling assumption, the model parameters are estimated by maximizing the likelihood, or equivalently by minimizing the negative log-likelihood. For *n* independent observations, the log-likelihood can be written as follows.

$$l(\beta) = \sum_{i=1}^{n} [y_i(\beta_0 + x_i^T \beta) - log(1 + exp(\beta_0 + x_i^T \beta))]$$

Unlike ordinary least squares, logistic regression does not generally have a closed-form solution, so estimation is performed iteratively, commonly through Newton-Raphson or iteratively reweighted least squares procedures. This likelihood-based foundation is one reason logistic regression became so influential in biostatistics, epidemiology, and clinical research, where binary outcomes are common and probabilistic interpretation is essential. Although classical logistic regression is powerful, it can become unstable when predictors are highly correlated, when the number of predictors is large relative to the sample size, or when the data contain weak signals and noise. Regularization addresses this problem by adding a penalty term to the objective function, thereby shrinking coefficient estimates toward zero and reducing overfitting. In this sense, penalized logistic regression can be viewed as a restricted or regularized maximum-likelihood framework. Ridge regression was proposed by Hoerl and Kennard in 1970 as a biased estimation strategy for non-orthogonal problems, and its central idea is to improve stability by penalizing the squared magnitude of the coefficients. When translated to logistic regression, ridge penalization augments the negative log-likelihood with an $L2$ penalty term.

$$\widehat{\beta_{\text{ridge}}} = \arg\min_{\beta} \left[-l(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \right]$$

The ridge penalty shrinks coefficients continuously toward zero but usually does not make them exactly zero, which means that all predictors remain in the model, albeit with reduced magnitude. Its mathematical philosophy is variance control: by accepting a small amount of

bias, ridge often reduces estimator instability and improves generalization when predictors are correlated or numerous. For this reason, ridge-type models are especially useful when many variables may contain partial information and the main goal is stable prediction rather than aggressive variable elimination. The least absolute shrinkage and selection operator, or lasso, was introduced by Robert Tibshirani in 1996. The lasso modifies the optimization problem by adding an L_1 penalty, that is, the sum of the absolute values of the coefficients.

$$\widehat{\beta_{\text{lasso}}} = \arg\min_{\beta} \left[-l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

The key difference from ridge is geometric as well as statistical: the L_1 constraint has corners, and these corners encourage some coefficients to become exactly zero. As Tibshirani emphasized, this gives the lasso an appealing combination of shrinkage and variable selection, producing models that are often both more interpretable and more stable than unrestricted estimation. In high-dimensional research settings, this sparsity property is especially valuable because it can reduce noise, simplify interpretation, and identify a smaller subset of potentially important predictors. A useful way to compare the two penalties is through their behavior in the presence of multicollinearity and high dimensionality. Ridge tends to distribute shrinkage across correlated predictors and is often preferred when many variables carry overlapping information. Lasso, in contrast, tends to retain only a subset of predictors and set others to zero, which can be advantageous for feature selection but may be less stable when several predictors are strongly correlated and carry similar signals. Thus, ridge is often viewed as a stabilization method, whereas lasso is often viewed as a sparsity-inducing selection method. Another important conceptual point is that penalized logistic regression reflects the broader statistical principle of the bias-variance tradeoff. Ordinary maximum-likelihood estimation aims to fit the observed data as closely as possible, but in noisy or small-sample settings this can yield estimates with large variance and poor out-of-sample performance. Regularization deliberately introduces bias through coefficient shrinkage in order to reduce variance and improve predictive performance on new data. These methods are widely used in scientific research because many real-world questions are naturally binary and because researchers require both prediction and interpretability. In medicine and public health, logistic regression is commonly used to estimate risk, identify associated factors, and provide interpretable effect sizes, while penalized forms are particularly useful when the predictor space is large, correlated, or noisy. Regularized logistic models are therefore popular in biomarker discovery, omics, epidemiology, and clinical prediction research, where overfitting, collinearity, and limited sample size are common analytical challenges. In the context of the present study, penalized logistic regression serves as a transparent and clinically interpretable comparison model alongside the tree-based imbalanced-learn Random Forest framework. Its inclusion is methodologically important because it allows evaluation of whether the predictive signal is detectable not only by a nonlinear ensemble model but also by a sparse linear classifier operating on the log-odds scale. Furthermore, when combined with SHAP-based explanation, penalized logistic regression

contributes to interpretable biomarker discovery by clarifying which variables most strongly influence classification at each stage of cognitive decline prediction.

2.13 Bayesian optimization (BO)

Bayesian optimization (BO) is a sequential, sample-efficient strategy for optimizing an objective function that is expensive to evaluate, treated as a "black box," meaning the optimizer does not require gradients or an explicit analytical form of the function. In machine learning, that objective is commonly a validation score (e.g., cross-validated AUC or loss) produced after training a model under a particular hyperparameter configuration, which makes each function evaluation computationally costly. Compared with grid or random search, BO is designed to use information from previous trials to choose the next hyperparameter setting more intelligently, concentrating evaluations in regions that are more likely to improve performance while still probing uncertain regions.

The key mechanism is that BO maintains a probabilistic surrogate model that approximates the relationship between hyperparameters and the objective value, and then uses an acquisition function to decide which hyperparameter configuration to try next. A common surrogate is a Gaussian process model, and common acquisition functions include Expected Improvement (EI) and Upper Confidence Bound (UCB), which formalize the exploration–exploitation trade-off (trying uncertain regions versus refining known good regions). The process is iterative: propose a candidate configuration using the acquisition function, train/evaluate the model to obtain the objective score, update the surrogate model with the new observation, and repeat until a stopping criterion is met (e.g., evaluation budget or convergence).

In hyperparameter tuning, BO is particularly useful when the search space is large (many hyperparameters, wide ranges, and interactions) and when each training run is time-consuming, because it aims to reduce the number of evaluations needed to find near-optimal settings. This efficiency is valuable in high-dimensional biomedical modeling, where the combination of complex models and limited sample sizes increases the risk of overfitting and makes rigorous validation (e.g., cross-validation and ideally nested validation when tuning) essential to avoid overly optimistic performance estimates. Practically, BO requires carefully defining (i) the search space (ranges/types of hyperparameters), (ii) the objective function (what metric is optimized and how it is computed), and (iii) the evaluation protocol (how folds/splits are performed and whether repeated runs are needed to stabilize noisy estimates). Because many clinical and multimodal pipelines introduce additional stochasticity (random initialization, sampling, missingness handling, and modality-fusion training dynamics), it is often necessary to reduce noise in the objective by fixing random seeds when appropriate and/or averaging across folds or repeated runs so that the optimizer does not "chase" random variation.

For multimodal AI models (e.g., combining clinical chemistry, metabolomics/omics, imaging, and digital biomarkers), BO becomes even more important because there are usually hyperparameters at multiple levels: modality-specific preprocessing, representation-learning architectures, fusion strategy, regularization strengths, and training dynamics. By iteratively

learning which regions of this joint hyperparameter space are promising, BO can make multimodal model development more computationally feasible and can improve generalization when paired with robust validation and transparent reporting, rather than relying on ad hoc manual tuning. However, BO does not remove the need for external replication: even well-tuned models can fail to generalize if training data are biased or not representative, so the optimized configuration should still be tested on independent cohorts and assessed for stability and clinical plausibility.

2.14 Imbalanced-learn

Imbalanced-learn (imported as imblearn) is a Python library from the scikit-learn-contrib ecosystem that extends the scikit-learn API with methods specifically designed for classification problems where the outcome classes are highly skewed (e.g., many "Normal" cases and few "Impaired" cases). Its core value is that it provides consistent, pipeline-friendly components for resampling and imbalance-aware modeling, so that class-imbalance handling can be treated as a first-class step in a machine-learning workflow rather than an ad hoc preprocessing trick. In practice, imbalanced-learn is used to reduce bias toward the majority class by either changing the training distribution (resampling) or modifying the learning algorithm (ensemble methods that integrate sampling), while keeping the familiar fit, predict, and cross-validation patterns used in scikit-learn.

Within imbalanced-learn, resampling methods are designed to transform the training set into a more learnable representation for minority classes, commonly via over-sampling (increasing minority examples), under-sampling (reducing majority examples), or combined strategies that apply both. These resampling approaches are exposed in a way that fits the scikit-learn mental model (e.g., "fit then transform/resample"), which makes it straightforward to embed them into reproducible pipelines and validation schemes. This is particularly important in biomedical prediction settings because the modeling goal is often not "overall accuracy," but clinically meaningful detection of rare outcomes, which requires evaluation strategies and metrics that do not get dominated by the majority class.

A balanced forest model in imbalanced-learn most commonly refers to the BalancedRandomForestClassifier, which adapts the standard Random Forest idea to imbalanced data by constructing each tree using a class-balanced bootstrap sample (typically by under-sampling the majority class for each tree). The motivation is that standard Random Forest bootstrapping can yield training subsets containing too few minority samples- especially under extreme imbalance- so individual trees may learn decision rules that largely ignore minority patterns. By forcing more balanced per-tree training data, the forest tends to allocate more modeling capacity to minority-class structure, often improving recall/sensitivity for rare classes while still retaining the robustness of tree ensembles. In imbalanced-learn, this estimator also exposes imbalance-specific controls such as sampling_strategy (how balancing should be performed) alongside standard forest hyperparameters (e.g., number of trees and tree complexity), enabling the user to tune both the resampling behavior and the model capacity. For reporting and validation, it is generally important to emphasize imbalance-aware metrics

(e.g., balanced accuracy and precision/recall-oriented measures) rather than relying on raw accuracy, because accuracy can remain high even for uninformative classifiers that mostly predict the majority class.

2.15 SHAP Analysis for biomarker discovery

SHAP (SHapley Additive exPlanations) was first proposed by Lundberg and Lee (2017) as a unified framework for interpreting complex machine learning models[75]. It relies on a rigorous mathematical foundation rooted in cooperative game theory, originally introduced by Lloyd Shapley in 1953, to fairly distribute a model's prediction among its input features. The method is considered one of the leading XAI (Explainable AI) approaches because it is the unique feature-attribution method that satisfies three desirable axioms: local accuracy (the sum of feature attributions equals the model output), missingness (features not present have zero impact), and consistency (if a model changes so that a feature relies more on it, its attribution does not decrease)[76].

In a simplified example, imagine three people (features A, B, and C) working together to generate a profit (prediction) of $100 compared to a baseline of $0. To determine exactly how much credit Feature A deserves, SHAP calculates the marginal contribution of A in every possible coalition:

How much does profit increase if A joins an empty team?

How much does it increase if A joins just B?

How much if A joins just C?

How much if A joins B and C together?

SHAP then averages these marginal contributions to assign a single "fair" value to Feature A. In a biomarker model, this means calculating how much a specific patient's high plasma tau level increased their predicted Alzheimer's risk, considering all possible combinations with other variables like age, genetics, and cognitive scores[77].

In biomarker discovery, SHAP is typically applied after training a predictive model (e.g., tree ensembles, gradient boosting, or neural networks) so that each candidate biomarker (laboratory value, metabolite, omics feature, wearable-derived metric, or microbiome feature) receives an attribution showing how much it pushed a specific subject's predicted risk higher or lower relative to a baseline expectation. By aggregating these individual-level attributions across the cohort (e.g., mean absolute SHAP values), SHAP produces global rankings that can be interpreted as model-based "importance" measures, while still retaining the ability to drill down to subject-specific explanations that are essential for clinical interpretability and hypothesis generation.In biomarker discovery workflows, SHAP's practical strength is that it supports both local and global interpretability: local explanations help clinicians or researchers understand why a particular participant is classified as high risk (useful for clinical plausibility checks and case-based reasoning), and global summaries help prioritize candidate biomarkers

for downstream statistical testing, pathway interpretation, and external validation. This dual view is especially useful in multimodal or multi-omics settings where traditional coefficients (e.g., from linear models) may be hard to interpret due to nonlinearity and interactions, while SHAP can reveal nonlinear effects and interaction patterns that are learned by complex models without requiring the model itself to be simplified. When used carefully, SHAP can therefore accelerate hypothesis generation by translating high-performing but opaque predictors into a ranked and visualizable set of biomarker candidates, narrowing a large feature space into a smaller set of biologically and clinically interpretable signals for further validation[78].

Shapley values are computationally expensive because the original formulation scales exponentially with the number of features, so practical SHAP implementations often rely on approximations (e.g., KernelSHAP) or model-specific accelerations (e.g., TreeSHAP), and approximation error can affect stability of rankings in high-dimensional biomarker spaces. Second, SHAP explanations are fundamentally associational: they explain how the trained model uses features to make predictions, but they do not establish causal effects, so high SHAP importance should be interpreted as "predictive relevance within this dataset/model" rather than proof of a biomarker being mechanistically causal. Third, feature dependence and correlation are challenging: in correlated biomarker panels, SHAP attributions can be distributed across correlated variables in ways that depend on the background distribution and the method used to handle "missing" features, which can lead to apparent importance being assigned to a proxy feature rather than the underlying causal signal. Finally, SHAP can be misinterpreted if it is treated as a universal measure of clinical importance; robust biomarker discovery still requires careful study design, leakage avoidance, appropriate cross-validation (often nested when tuning), and—critically—external validation to mitigate false discovery and overfitting in small or high-dimensional biomedical datasets[79,80].

3 Problem Statement

Dementia and age-related cognitive decline present a major clinical and public health challenge because cognitive impairment develops gradually, arises from heterogeneous biological and behavioral pathways, and is often difficult to detect early with sufficient precision for prognosis. In community and primary-care settings, this creates a practical detection and prognosis challenge because symptom-driven categorization- even when standardized- often identifies impairment relatively late and provides limited precision for forecasting whether an individual trajectory will remain stable, decline slowly, or progress rapidly.

A closely related translational challenge is that the most pathophysiologically specific Alzheimer's disease biomarkers (for example, CSF amyloid/tau or PET imaging) are resource-intensive, invasive, or both, which limits scalability for routine screening and longitudinal monitoring in large populations. As a result, many real-world settings still depend heavily on clinical criteria and brief cognitive instruments, while biomarker-driven workflows remain unevenly adopted across typical healthcare infrastructures. Minimally invasive biomarker strategies that can support earlier risk stratification using measurements that are more feasible in real-world care and community-based settings.

A further challenge lies in the analytical framework required to model cognitive decline. Realistic dementia prediction tasks may require combining high-dimensional multimodal inputs, dealing with missingness and irregular sampling, and learning from small-to-moderate cohorts which may increase the risk of overfitting and reduce generalizability if not handled rigorously. These concerns align with broader evidence from machine-learning biomarker discovery showing that powerful models can find patterns in complex biological datasets yet still fail to generalize without strong validation and replication practices, thereby increasing false discovery risk. Therefore, the central applied problem is how to develop scalable, non-invasive, multimodal biomarker panels- paired with clinically aligned machine learning and transparent, explainable inference- to improve prediction of near-future cognitive status (e.g., 18-month MMSE-derived categories) and reduce ambiguity near decision boundaries such as normal aging versus early impairment. Prediction with multimodal data from routine clinical chemistry, metabolomics, gut microbiome indices, and wearable-derived activity measures, while also dealing with a relatively small cohort, marked class imbalance, and clinically subtle boundaries between Normal, Possible MCI, and Impaired groups is quite challenging with the risk of overfitting & unstable estimates.

Accordingly, the central problem addressed in this thesis is how to develop a scalable, biologically interpretable, and clinically aligned framework that uses non-invasive multimodal baseline data to predict near-future cognitive status more effectively.

4 Research Gaps

Despite substantial progress in dementia research, several important gaps remain in the development of scalable, interpretable, and clinically useful biomarker-based prediction frameworks. A major gap is that the scalability–specificity trade-off remains unresolved in many dementia biomarker pipelines. Highly specific biomarkers may be less deployable at scale, while scalable signals (routine blood tests, wearable-derived behavior, and other minimally invasive measures) may be less disease-specific on their own, motivating the need for principled multimodal combinations that can regain specificity without sacrificing feasibility in population-level screening contexts.

A second gap is that multimodal integration remains methodologically and operationally difficult. Clinical reasoning naturally fuses heterogeneous information streams (history, routine labs, functional status, and- when available- specialized tests), yet many deployed AI systems still focus on narrow tasks with single modalities, and multimodal medical AI faces persistent barriers such as data heterogeneity, missingness, and dataset bias that complicate real-world generalization. This gap is particularly relevant in dementia because cognitive decline can be shaped by interacting systemic processes- metabolic, vascular, inflammatory, and behavioral- that are unlikely to be captured by any single modality in isolation. Highly specific approaches such as advanced imaging and cerebrospinal fluid-based biomarkers can provide valuable pathophysiological information, but their cost, invasiveness, and operational burden limit their suitability for routine screening and repeated monitoring in community settings.

A third gap concerns generalizability and dataset bias. Biomedical machine-learning systems often degrade when tested outside their development cohorts, and multimodal linkage can exacerbate selection bias because individuals who complete all modalities (e.g., labs + microbiome + wearables) may systematically differ from those who do not, threatening transportability across populations and care settings. Addressing this requires validation designs and reporting practices that explicitly evaluate robustness and describe cohort composition, rather than relying solely on within-cohort performance. Cognitive decline is shaped by interacting metabolic, vascular, inflammatory, behavioral, and possibly microbiome-related processes, making it unlikely that any single data type can adequately capture the full biological and functional spectrum of risk. Although clinical reasoning naturally integrates heterogeneous information sources, many analytical studies and deployed artificial intelligence systems still focus on narrower single-modality tasks, leaving the practical integration of multimodal signals insufficiently developed.

A fourth gap is the small-sample, high-dimensional false discovery problem that frequently occurs in biomarker discovery, especially in settings resembling $p \gg n$ regimes. In such contexts, machine learning can produce apparently strong predictive performance while still learning unstable or non-reproducible feature associations unless strict safeguards (appropriate cross-validation design, leakage prevention, external validation, and biological triangulation) are applied. This is not a theoretical concern but a recurrent reason why promising biomarker signatures fail to replicate.

A fifth gap is that interpretability is often insufficiently aligned with clinical and biological questions. Many machine-learning studies aim to maximize predictive accuracy, but the resulting models may function as black boxes and provide limited insight into which baseline features are driving predictions or whether different predictors matter at different stages of decline. Although explainable AI is widely discussed, explanations are often treated as post hoc add-ons rather than as tools for clinical reasoning, bias auditing, and stage-specific biological insight; moreover, explanation methods can be misunderstood as causal evidence if interpretation is not careful. SHAP is one of the best-known post hoc explanation approaches and is grounded in Shapley-value theory to attribute feature contributions to individual predictions, but even SHAP explanations should be interpreted as describing model behavior rather than causal mechanisms.

There is therefore an important gap in stage-aware, interpretable modeling approaches that can support biomarker discovery while remaining transparent enough to inform clinical reasoning. A remaining research gap, therefore, is the development of clinically aligned hierarchical modeling frameworks that first detect abnormality and then distinguish severity within the abnormal subgroup. Addressing label noise and boundary sensitivity in clinical cognitive categories, particularly when using thresholded scores such as MMSE. MMSE-based categories are practical and repeatable, but they compress multiple cognitive domains into a single score and can create "boundary noise" around cutoffs (e.g., Normal vs Possible MCI), which can distort supervised targets and penalize models for clinically ambiguous cases. This motivates modeling strategies that explicitly reduce boundary ambiguity, such as staged or hierarchical decision pipelines that mirror clinical screening and severity-grading workflows.

Taken together, these gaps indicate that dementia biomarker research still lacks sufficiently scalable, multimodal, interpretable, and stage-aware prediction frameworks for near-future cognitive status in community-dwelling older adults.

5 Hypotheses

This study investigates, in an exploratory biomarker-discovery framework leveraging Machine Learning techniques three (03) research questions –

- H1 (Stage-dependent performance difference): The two-stage clinical hierarchical (staged) prediction pipeline will show different performance across stages, such that Stage 2 (Possible MCI vs Impaired among screen-positive individuals) will achieve higher discriminability (e.g., ROC-AUC and macro-F1) than Stage 1 (Normal vs Abnormal), reflecting improved separability after restricting the decision space to clinically flagged participants.
- H3(Model robustness across algorithm classes): The two-stage clinical hierarchical (staged) prediction pipeline will show different performance across algorithms in both stages. Both the Imbalanced-Learn Random Forest model and the penalized logistic regression baseline will capture meaningful predictive signal from the multimodal baseline data, indicating that the association between baseline biomarkers and near-future cognitive status is not confined to a single model class.
- H3 (Stage-specific interpretable biomarker discovery): Explainable inference applied to both models will identify biologically interpretable predictors, with stage-specific differences in the features associated with early abnormality detection and severity discrimination.
- H4 (Utility of scalable multimodal predictors): Minimally invasive and scalable baseline measures, including routine clinical chemistry, targeted blood-based metabolites, gut microbiome diversity indices, and wearable-derived physical activity features, will provide sufficient information to support clinically meaningful risk stratification in community-dwelling older adults.

6 Conceptual framework

The conceptual framework of this thesis is based on the view that cognitive decline is a multisystem, multimodal, staged modeling approach in which cognitive decline is treated as an emergent phenotype arising from interacting biological and functional systems rather than a single-pathway process. Accordingly, the framework assumes that near-future cognitive status can be better understood and predicted by integrating complementary baseline signals from multiple domains, including routine clinical chemistry, targeted metabolomics, gut microbiome diversity, and wearable-derived physical activity measures. At the biological level, central neural vulnerability and pathology-related processes interact with systemic metabolic and vascular conditions (e.g., renal clearance, lipid metabolism, redox balance, inflammatory tone) that can be reflected in routine blood biochemistry and targeted metabolomics, while gut–brain axis signals (microbiome composition and derived metabolites) may index immune-metabolic modulation relevant to neurodegeneration. In parallel, functional phenotype expression-activity volume, regularity, and variability- offers downstream behavioral readouts that can be captured continuously through wearable devices, potentially improving sensitivity to subtle early changes that are difficult to detect with occasional clinic-based assessments.

At the measurement level, the rationale for combining these modalities is that they operationalize the same heterogeneous evidence integration that clinicians perform implicitly, but in a structured and scalable manner. Multimodal medical AI literature emphasizes that medicine is intrinsically multimodal and that real-world performance is often constrained by heterogeneity, missingness, and bias, which supports designing pipelines that treat multimodal fusion and robustness as primary objectives rather than secondary enhancements. Consequently, a panel such as routine labs plus targeted (including chiral) metabolomics, microbiome profiling, and wearable activity features can be framed as a practical approximation of clinical fusion logic using deployable measurements. By combining scalable and relatively low-burden measurements, the framework aims to approximate this multidomain clinical logic in a structured and analytically reproducible manner.

At the modeling level, the framework is organized as a two-stage hierarchical prediction pipeline that reflects the clinical progression of cognitive impairment. Stage 1 functions as a screening step that separates “likely normal cognition” from “possible impairment,” aiming to reduce false reassurance and to prevent subtle abnormal signals from being overwhelmed by the dominant normal class in imbalanced datasets. Stage 2 then performs severity discrimination among flagged individuals, distinguishing mild versus more advanced impairment within a narrower and more homogeneous decision space. This staged structure is conceptually important because it reduces the burden of forcing a single model to learn all category boundaries at once and is intended to better capture stage-dependent patterns of biomarker relevance. Finally, the framework links prediction to discovery through an explicit explainability-and-validation pathway. The predictive output provides an individual-level probability of near-future cognitive category membership, while the explanatory output uses SHAP-style feature attributions to identify which biomarkers push predictions upward or downward, ideally in a stage-specific manner so that screening biomarkers and severity biomarkers can be distinguished mechanistically. Because machine-learning biomarker

discovery is prone to overfitting and false discovery without replication, the framework also includes statistical and clinical triangulation steps- such as post hoc group comparisons, longitudinal consistency checks, and external validation where feasible- to increase credibility and translational readiness.

The analytical component of the framework includes both an Imbalanced-Learn Random Forest model and a penalized logistic regression model, allowing comparison between a nonlinear ensemble learner and a more statistically linear baseline under the same staged prediction structure. This dual-model design supports assessment of whether the predictive signal is robust across different algorithmic assumptions and helps distinguish model-specific findings from patterns that are more consistently detectable in the data.

Finally, the framework connects prediction with biomarker discovery through explainable inference and statistical validation. Predictive outputs provide estimated classification of near-future MMSE-based cognitive status, while SHAP-based feature attribution is used to identify which variables contribute most strongly to predictions at each stage. In this way, the framework is designed not only to classify cognitive status, but also to reveal stage-specific biomarker signatures that may support biological interpretation. To strengthen credibility, model-derived findings are further complemented by post hoc statistical comparisons across cognitive groups, thereby linking machine-learning explanations with group-level evidence from the observed data.

7 Methods

7.1 Original Cohort Study Design

The present analysis utilized data from a prospective, non-randomized controlled cohort study conducted in the Ajisu region of Yamaguchi city, Japan. The primary objective of the original cohort study was to evaluate whether multidomain exercise and yogurt-based nutritional interventions could attenuate decline in cognitive and motor function over 18 months compared with a usual-care control group.

Participants were allocated to three groups (Exercise+, Yogurt+, Control) by preference and feasibility, and all procedures, assessments, and follow-up schedules were predefined in the original protocol and approved by the institutional ethics committee.

The cohort included extensive baseline measurements spanning several biological and behavioral domains, including routine biochemistry, targeted metabolomics, chiral amino acids and glycine, gut microbiome profiling, and wearable-derived physical activity data. This comprehensive dataset provides foundation for subsequent investigations into biomarkers predictive of cognitive outcomes as a secondary endpoint.

7.2 Participants

Participants to the cohort study were recruited from Ajisu, a district in Yamaguchi city, Yamaguchi Prefecture, Japan. Eligible participants were community-dwelling older adults aged 75 to 83 years who were able to provide written informed consent after receiving a comprehensive explanation of the study procedures. All participants were required to be free from any serious medical conditions.

Exclusion criteria included: (1) suspected dementia, defined as a Mini-Mental State Examination (MMSE) score of 23 or lower; (2) certification as requiring nursing care at level 2 or higher; or (3) any condition deemed unsuitable for participation as determined by the investigators. Such conditions include inability to stand unassisted, neurological ot musculoskeletal disorders, serious medical illnesses (e.g., cancer or renal failure), or regular use of antibiotics at the time of recruitment. In addition, participants in the exercise group were excluded if they had contraindications to physical activity, including epilepsy, hernia, recent surgery, or joint prostheses. Participants undergoing standard treatment for chronic conditions, including diabetes, hypertension, and dyslipidemia, were eligible to remain in the study.

Initially, 104 volunteers expressed interest in participating in the 18-month intervention study. Four individuals withdrew from the study prior to enrollment due to procedural difficulties (n=2) or medical requirements (n=2). During the study, one participant voluntarily withdrew, and one participant died, resulting in 98 individuals who completed the study. The inclusion and exclusion criteria are shown in **Table 1.**

Table 1
Inclusion and exclusion criteria for the study

Inclusion and Exclusion Criteria of the Subjects
Inclusion criteria
Subjects Who:
1. Community dwelling older individuals aged 75 to 83 years at the time of informed consent
2. Able to provide voluntary written informed consent
3. Not suffering from any serious medical condition
4. Participants receiving standard treatment for chronic diseases such as diabetes, hypertension and dyslipidemia were eligible to remain in the study
Exclusion criteria
Subjects who:
1. Suspected Dementia (A score of 23 or lower in MMSE examination)
2. Certification of requiring nursing care at level 2
3. Presence of any condition deemed unsuitable for participation by the investigators such as inability to stand unassisted, neurological, musculoskeletal, or connective tissue disorder, serious medical illnesses such as cancer or renal failure, and regular use of antibiotics at the time of recruitment
4. Additionally, participants of exercise group were excluded if they had contradictions to physical activity including epilepsy, hernia, recent surgery or joint prostheses.

This table summarizes the eligibility criteria used to recruit community-dwelling older adults for the study. Inclusion criteria ensured participants were aged 75–83 years, capable of providing written informed consent, and free of serious acute medical conditions, while allowing stable chronic disease management (e.g., diabetes, hypertension, dyslipidemia). Exclusion criteria removed individuals with suspected dementia (MMSE ≤23), higher long-term care needs (care level ≥2), or medical/functional conditions likely to interfere with safe participation or valid assessments (e.g., inability to stand unassisted, serious neurological/musculoskeletal disorders, cancer, renal failure, or current antibiotic use).

7.3 Ethical Committee Review

The study protocol, informed consent forms, and related documentation were reviewed and approved by the Ethics Review Committee of Yamaguchi University. This study was conducted in accordance with the Declaration of Helsinki and the Japanese Ethical Guidelines for Medical and Health Research Involving Human Subjects, including the guidelines on human genome and genetic analysis research. This study was registered in the University Hospital Medical Information Network Clinical Trials Registry (UMIN-CTR) prior to participant recruitment.

7.4 Secondary Analysis Framework

This study constitutes an exploratory secondary analysis using data from the original cohort. We included all participants with complete baseline predictors and Mini-Mental State Examination (MMSE) scores assessed 18 months after baseline. The analysis aimed to identify baseline biomarkers associated with subsequent cognitive status, defined by MMSE outcomes at 18 months. Participants were categorized into three cognitive groups- Normal, Possible MCI, and Impaired- and these categories were used as supervised learning targets within a hierarchical machine-learning framework designed to reflect clinically meaningful progression in cognitive decline.

A two-stage binary classification approach was applied. In Stage 1, the model distinguished Normal from Possible MCI or Impaired. In Stage 2, the model further classified Possible MCI versus Impaired. This staged design was intended to reduce ambiguity at adjacent category boundaries and to allow each classification step to address a more specific clinical decision problem. Classification was performed using two model classes: an Imbalanced-Learn Random Forest model and a penalized logistic regression model used as a baseline comparator. Given the relatively small sample size, we adopted Leave-One-Out Cross-Validation (LOOCV) to estimate model performance. Bayesian hyperparameter optimization was used to tune model parameters, and SHapley Additive exPlanations (SHAP) values were computed to evaluate feature im-portance and interpretability.

Importantly, the set of explanatory predictors included not only the original study group assignments (Exercise+, Yogurt+, Control) but also a wide range of baseline multi-modal features collected at month 0. These predictors span demographic factors, routine clinical chemistry, targeted metabolomics including chiral amino acids, gut microbiome-related indices, and wearable-derived physical activity features. This design enabled the analysis to assess the relative contribution of both intervention exposure and baseline biological and behavioral characteristics to future cognitive classification.

7.5 Data

We included a range of baseline predictors in the model, spanning demographic, intervention, biochemical, behavioral, and microbiome-related data.

Demographic predictors include age and sex.

Intervention-related information consisted of group assignments in the original Ajisu cohort study: Exercise+ (n=40), Yogurt+ (n=20), and Control (n=40). The Exercise Group was given structured 90-minute group sessions conducted once weekly at a local health and welfare center. Participants in the Yogurt Group were instructed to consume one cup (100 g) of yogurt per day. These were implemented as categorical variables.

Biochemical and metabolic predictors were derived from 55 mL of fasting venous blood collected at baseline. Measurements included general blood biochemistry (liver and kidney function, lipid profile, glucose metabolism, anemia and targeted metabolites), and plasma concentrations of 24 chiral amino acids and Glycine: DL-Alanine (DL-Ala), L-Arginine (L-Arg), DL-Asparagine (DL-Asn), L-Aspartic acid (L-Asp), L-Citrulline (L-Cit), L-Glutamine (L-Gln), L-Glutamate (L-Glu), Glycine (Gly), L-Histidine (L-His), L-Isoleucine (L-Ile), L-Leucine (L-Leu), L-Lysine (L-Lys), L-Methionine (L-Met), L-Ornithine (L-Orn), L-Phenylalanine (L-Phe), DL-Proline (DL-Pro), DL-Serine (DL-Ser), L-Threonine (L-Thr), L-Tryptophan (L-Trp), L-Tyrosine (L-Tyr), and L-Valine (L-Val). These were quantified using chiral tandem liquid chromatography-tandem mass spectrometry (LC-MS/MS).

Cognitive function was evaluated using the standard Mini-Mental State Examination (MMSE) questionnaire at both baseline and 18 months.

Physical activity was recorded from 83 participants using wearable accelerometers (HJA-750C, OMRON Healthcare) worn for one month.

Gut microbiome profiles were obtained from fecal samples collected using FS-0007 collection kits (TechnoSuruga Laboratory Co., Ltd.). DNA libraries were sequenced across four lanes of an Illumina NovaSeq 6000 platform (Illumina, Inc.) with paired-end 150 bp reads.

7.6 Cognitive Outcome Classification Based on MMSE

Cognitive status at 18 months was categorized into three groups based on MMSE scores: Normal (MMSE > 26), Possible MCI (24 < MMSE ≤ 26) or a drop of in score more than 3 or equal from the initial time point to end time point, and Impaired (MMSE ≤ 23). This classification yielded 73 Normal, 20 Possible MCI, and 5 Impaired cases. These thresholds capture clinically meaningful cutoffs commonly used to characterize cognitive severity in older adults, and were used to define outcome labels for supervised machine learning (Table 2). However, the MMSE alone has limited sensitivity for subtle cognitive changes and there is no universally accepted standard for defining MCI based solely on the MMSE. Accordingly, our outcome categories should be viewed as pragmatic severity strata rather than definitive clinical diagnoses within a cohort that may already include individuals with mild baseline impairment, rather than incident conversion from entirely normal cognition (**Table 2**).

Table 2
MMSE Category

MMSE Interpretation	MMSE Interpretation	Number of Individuals
End Score > 26	End Score > 26	73
24< End Score ≤ 26 or initial score - end score ≥ 3	24< End Score ≤ 26 or initial score - end score ≥ 3	20
End Score ≤ 23	End Score ≤ 23	5

This table defines the Mini-Mental State Examination (MMSE) thresholds used to classify partici-pants' cognitive status at the end of follow-up. Participants with an MMSE score ['18 month'] > 26 were categorized as Normal (n = 73), those with 23 < MMSE score ['18 month'] ≤ 26 or (MMSE score ['0 month'] − MMSE score ['18 month']) ≥ 3 as Possible MCI (n = 20), and those with MMSE score ['18 month'] ≤ 23 as Impaired (n = 5). These categories were used as thise study's cognitive out-come labels for subsequent statistical comparisons and the two-stage classification pipeline.

7.7 Biological Rationale for Selected Predictors

A range of multimodal predictors, including biochemical markers, chiral amino acids, metabolites, physical activity metrics, gut microbiome alpha diversity, and β-amyloid, was selected to capture complementary biological, behavioral, and pathological pathways that may jointly influence cognitive status. These features were considered suitable for predicting MMSE-based cognitive categories due to their mechanistic relevance to cognitive decline.

Systemic metabolic dysfunction and low-grade inflammation, assessed via routine blood panels (e.g., lipids, glucose, renal function, uric acid, and leukocyte-derived ratios), are linked to amyloid burden and cognitive impairment. Their inclusion enables identification of vascular, metabolic, and immunologic contributor to cognitive changes. Variables such as lipid-related markers, glucose-related measures, renal function indices, uric acid, and leukocyte-derived measures can reflect broader metabolic, vascular, and immunologic states that may contribute to cognitive vulnerability. Their inclusion therefore enabled evaluation of whether accessible clinical laboratory measurements contain meaningful information relevant to later cognitive status.

Targeted plasma concentrations of chiral amino acids and related metabolites provide insights into neurotransmitter pathways (e.g., glutamate/GABA and tryptophan-kynurenine), host-microbiome co-metabolism, energy metabolism, and oxidative stress. These pathways are directly related to cognitive and mood regulation and provide sensitive molecular signatures for cognitive phenotyping[81–84]. In particular, amino acid-related pathways involving glutamatergic signaling, D-amino acid proportions, tryptophan-related metabolism, and oxidative stress have potential relevance to cognitive and neurobehavioral regulation. These measurements may therefore serve as sensitive molecular indicators of physiological changes associated with cognitive phenotyping and disease-related progression.

Physical activity metrics, such as daily steps, metabolic equivalents of tasks (METs), and exercise volume, are objectively associated with preserved cognitive function and represent modifiable behavioral signals that complement static clinical measures. Physical activity also interacts with vascular risk, metabolic regulation, and inflammatory processes [85–88]. Wearable-derived measures offer objective and ecologically relevant indicators of daily function that may capture subtle behavioral changes not easily detected during occasional clinic visits. Because physical activity is also a modifiable behavioral domain, its inclusion strengthens the translational relevance of the predictor set for risk stratification and monitoring.

Lastly, gut microbiome alpha diversity (e.g., Shannon index) and microbial community composition have been associated with cognitive performance and future cognitive trajectories, providing a gut–brain axis that complements both blood biomarkers and behavioral metrics [89]. Although microbiome-derived signals are complex and not disease-specific in isolation, they may complement blood-based and behavioral markers by capturing aspects of host-microbial interaction that are relevant to cognitive health and future decline. Taken together, the selected predictors were intended to provide a biologically grounded and multimodal

representation of the diverse processes that may contribute to MMSE-based cognitive outcomes.

7.8 Data Preprocessing

7.8.1 Variable Scaling

All predictor variables were standardized prior to model training to align their distributions and ensure comparability across features with different units and scales. Standardization adjusts each feature to have zero mean and unit variance, which prevents variables with large absolute magnitudes from disproportionately influencing model fitting. This procedure improves numerical stability, facilitates convergence during optimization, and enhances performance for scale-sensitive algorithms by ensuring balanced gradients.

7.8.2 Encoding of Categorical Variables

Categorical variables were encoded using one-hot encoding. This encoding is broadly compatible with tree-based machine learning models, which can inherently handle sparse and non-ordinal inputs without requiring additional normalization.

7.8.3 Missing Data Handling

Missing values were imputed using the median strategy, which is robust to outliers and skewed distributions, which are common in small, heterogeneous clinical datasets. Unlike the mean or model-based approaches, the median preserves central tendency without being distorted by extreme values, making it a reliable baseline for multimodal data imputation in low-sample settings.

7.9 Machine Learning Model Architecture

A two-stage hierarchical machine-learning framework was developed to predict 18-month cognitive outcomes, as defined by MMSE-based categories, using baseline multimodal data. The architecture was designed to reflect the clinical progression of cognitive impairment and to separate broad screening from finer severity discrimination. In the first stage, the model identified cognitively normal participants from those with any abnormality (Possible MCI or Impaired). In the second stage, it further differentiated between Possible MCI and Impaired cases. This stepwise approach reduces misclassification between adjacent cognitive categories and allows each stage to be optimized for its specific clinical trade-offs.

The MMSE-derived categories exhibit ordinal structure and clinically subtle boundaries, particularly between Normal and Possible MCI. A staged formulation helps isolate this boundary uncertainty by first separating broad abnormality before learning finer distinctions within a pre-filtered subgroup. Prior studies have shown that addressing class imbalance through resampling or class weighting improves model performance in small clinical datasets with skewed label distributions.

Two complementary classifier types were evaluated within this framework: an Imbalanced-Learn Random Forest model and a penalized logistic regression model used as a baseline comparator. The Random Forest model was selected because it can capture nonlinear relationships and higher-order interactions across heterogeneous multimodal predictors, while the penalized logistic regression model provided a more transparent linear alternative under the same classification structure. The inclusion of both models allowed assessment of whether the predictive signal was robust across different algorithmic assumptions rather than being dependent on a single modeling approach.

Because the dataset was characterized by limited sample size and substantial class imbalance, especially the low prevalence of Impaired cases, imbalance-aware modeling strategies were incorporated into the architecture. For the Random Forest classifier, integrated resampling within the imbalanced-learn framework was used to improve sensitivity to minority classes. For the logistic regression baseline, penalization and class weighting were used to stabilize estimation and improve classification of underrepresented outcome groups. These design choices were important because high-dimensional, low-sample clinical datasets are vulnerable to overfitting, unstable parameter estimates, and reduced generalizability. To mitigate these risks, the framework combined tree-based ensemble learning with feature subsampling and imbalance handling, while also incorporating penalized logistic regression as a complementary baseline model with implicit coefficient shrinkage and variable selection properties. Together, this machine-learning architecture provided a clinically aligned and methodologically balanced framework for multimodal prediction of cognitive status and for subsequent explainable biomarker analysis under Leave-One-Out Cross-Validation.

7.10 Validation Strategy

To evaluate model performance under the limited sample size, we applied Leave-One-Out Cross-Validation (LOOCV) for both stages and for both the Random Forest and logistic regression classifiers. In this framework, each iteration trains the model on all but one participant and evaluates it on the held-out individual, cycling through all samples so that every participant serves once as an independent test case. Using the same LOOCV scheme across models ensured directly comparable performance estimates and approximates the intended clinical use case, in which predictions must generalize to pre-viously unseen individuals with similar baseline assessments.

7.11 Hyperparameter Optimization

Bayesian optimization was used for hyperparameter tuning in both classification stages and for both model families, as it efficiently identifies high-performing configurations with fewer evaluations, which is particularly advantageous under LOOCV, where each evaluation is computationally intensive. For the Imbalanced-Learn Random Forest, a 1000-iteration search was conducted in Stage 1; LOOCV accuracies converged around 0.77–0.78 while macro-level precision, recall, and F1 fluctuated more, and the final model favored a relatively shallow forest with entropy splitting, max_depth = 19, and max_features = sqrt. In Stage 2, the Random Forest search converged within approximately 32 iterations, with optimization-stage LOOCV accuracy exceeding 0.96 and precision and recall approaching 0.92–0.98, leading to the

selection of a deeper tree ensemble (max_depth = 17, n_estimators = 256) that evaluates all features at each split. During these searches we relied on the internal resampling strategy of Balanced Random Forest Classifier and did not specify additional class weights, allowing the algorithm's built-in balancing to handle label skew while hyperparameters were tuned. For the logistic regression baselines, separate Bayesian searches were run in each stage over regularization strength (C), penalty type, and class-weighting schemes, yielding in Stage 1 an L1/Lasso- based penalized model with moderate regularization ($C \approx 1.48$) and no explicit class weighting, and in Stage 2 a weakly regularized L1/Lasso- based model ($C \approx 634.35$) with manual up-weighting of the minority Impaired class (class_weight = {Impaired: 3.0, Possible MCI: 1.0}). Hyperparameter search and performance estimation were both con-ducted within the same LOOCV framework without additional nested cross-validation or bootstrap resampling because further partitioning the data or extensive resampling (e.g., nested CV or bootstrap) would have resulted in extremely small and unstable training folds, particularly for the small classes. The resulting LOOCV metrics represent internally optimized, exploratory estimates of model performance rather than fully unbiased external validation. Optimization trajectories and the final tuned parameters for both Random For-est and logistic regression models are summarized in **Table 3**.

Table 3
Optimized hyperparameters for the two-stage hierarchical Imbalanced Learn Random Forest classification model.

Parameter	Stage 1	Stage 2
Splitting criterion	entropy	gini
max_depth	19	17
max_features	sqrt	None
min_impurity_decrease	0.05	N/A
min_samples_leaf	1	3
min_samples_split	11	10
n_estimators	361	256
replacement	True	N/A
sampling_strategy	all	N/A
Optimized hyperparameters for the two-stage base Logistic regression classification model		
C	1.4828	634.3524
penalty	l1	l1
Class_weight	'Impaired': 3.0, 'Possible MCI': 1.0	None

This table reports the final optimized hyperparameter settings for both the two-stage Imbalanced-Learn Random Forest classifiers and the corresponding baseline logistic regression models at each stage of the hierarchical pipeline. For Random Forest, Stage 1 (Normal vs. Abnormal screening) used an entropy splitting criterion, deeper trees (max_depth = 19), square-root feature subsampling (max_features = sqrt), and imbalance-handling settings (re-placement = True, sampling_strategy = all), whereas Stage 2 (Pos-sible MCI vs. Impaired) used a giniGini criterion, shallower but still relatively deep trees (max_depth = 17), evaluation of all fea-tures at each split (max_features = None), and larger terminal nodes (min_samples_leaf = 3) to limit overfitting within the small abnor-mal subset. The number of trees differed by stage (361 vs. 256), indicating distinct ensemble complexity requirements. For logistic regression, Stage 1 was tuned to an L1-regularized model with moderate regularization strength (C≈1.48) and no explicit class weighting, while Stage 2 favored a strongly regularized L1 model (C≈634.35) with manual up-weighting of the minority Impaired class (class_weight = {'Impaired': 3.0, 'Possible MCI': 1.0}) to improve sensitivity to more severe cognitive impairment. The N/A refers to Not applicable.

7.12 Feature Attribution via SHAP Analysis

To interpret model predictions and identify contributing features, we applied SHAP (Shapley Additive Explanations) in both classification stages. For the Imbalanced-Learn Random Forest models, we used the TreeExplainer module, and for the logistic regression baselines we used the LinearExplainer module, allowing SHAP values to be computed in a manner tailored to each model class while retaining a common additive framework. SHAP provides feature-wise attribution scores that quantify the contribution of each input variable to individual predictions, supporting both local interpretability at the participant level and global feature importance ranking across the cohort.

7.13 Post Hoc Group Comparison

We conducted Games–Howell tests to compare baseline features across MMSE-defined cognitive categories (Normal, Possible MCI, Impaired), as a post hoc statis-tical complement to the model-derived feature importance. This test is appropriate for un-equal variances and unbalanced sample sizes, providing more reliable Type-I error control than pooled-variance methods in heterogeneous clinical data.

7.14 Software Packages

All analyses were performed using the Python programming language and open-source Python packages. The software environment consisted of Python version 3.10 and the following major libraries (with versions), which were primarily used in the analyses: Pandas 2.2.2, NumPy 1.26.4, Matplotlib 3.5.1, Seaborn 0.13.2, Pingouin 0.5.5, Statsmodels 0.14.0, Scikit-bio 0.6.0, Scikit-learn 1.6.1, Scikit-optimize 0.10.2, SciPy 1.13.0, and SHAP 0.46.0.

8 Results

8.1 Stage 1 Performance: Normal vs Abnormal

Stage 1 demonstrated a moderate discriminative performance in separating Normal participants from those with an Abnormal cognitive status (Possible MCI and Impaired) in both the Imbalanced-Learn Random Forest and logistic regression models under Leave-One-Out Cross-Validation (LOOCV). The optimized Random Forest achieved an overall accuracy of 71.43% with a ROC–AUC of 0.72, while the tuned penalized regularized logistic regression baseline reached a higher accuracy of 76.53% with a comparable ROC–AUC of 0.71, indicating that the underlying signal is robust to differences in model class.

For the Random Forest, the confusion matrix showed high specificity (83.56%) but limited sensitivity (36.0%) for detecting Abnormal cognition, yielding a substantially stronger performance for the Normal class (precision = 79.22%, recall = 83.56%, F1 = 81.33%) than for the Abnormal class (precision = 42.86%, recall = 36.0%, F1 = 39.13%). Logistic regression improved Abnormal detection while maintaining similar specificity, with sensitivity increasing to 48.0% and Normal/Abnormal class metrics (Normal: precision = 84.93%, recall = 84.93%, F1 = 84.93%; Abnormal: precision = 52.17%, recall = 48.0%, F1 = 50.00%), reflecting a slightly more balanced screening profile. Overall, both models provided fair discrimination at the screening stage, with logistic regression offering modest gains in sensitivity and macro-averaged performance while Random Forest retained marginally higher AUC values (**Figure 1; Table 4).**

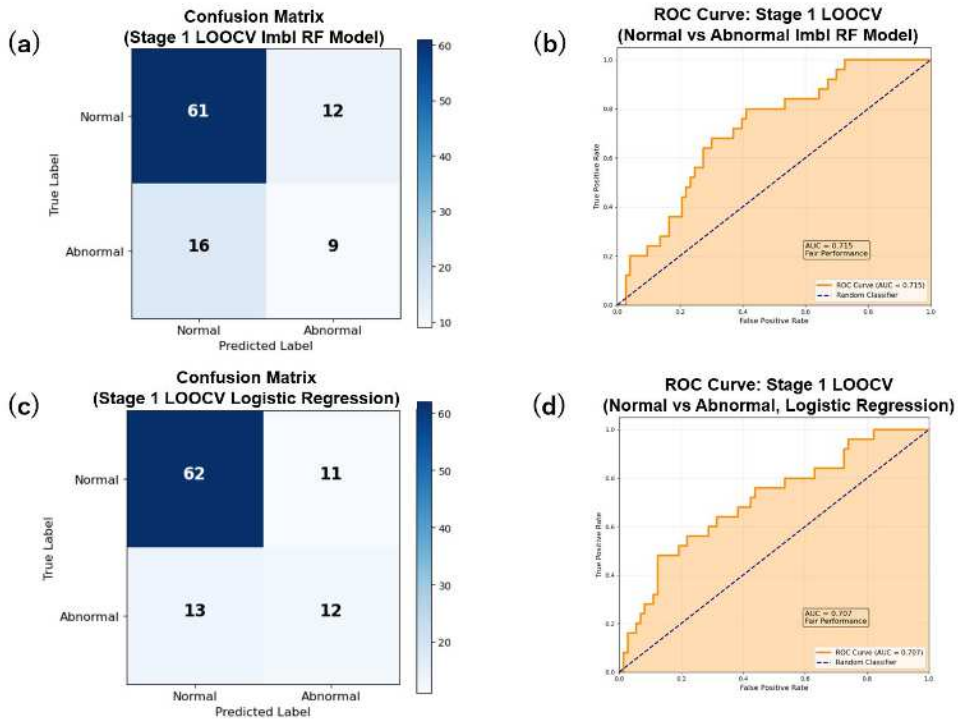


Figure 1.

Stage 1 classification performance for Normal vs. Abnormal cognitive status using two modeling approaches under Lleave-Oone-Oout Ccross-Vvalidation (LOOCV). Panels (a) and (b) show the confusion matrix and ROC curve, respectively, for the tuned Imbalanced-Learn Random Forest model (AUC = 0.715), while panels (c) and (d) show the corresponding confusion matrix and ROC curve for the tuned logistic regression baseline (AUC = 0.707). Together, these plots illustrate that both models achieve comparable overall discrimination, with logistic regression yielding slightly improved sensitivity and macro-averaged metrics, whereas the Random Forest provides margin-ally higher ROC–AUC.

Table 4
Stage 1 Classification performance metrics for Normal vs Abnormal cognitive status

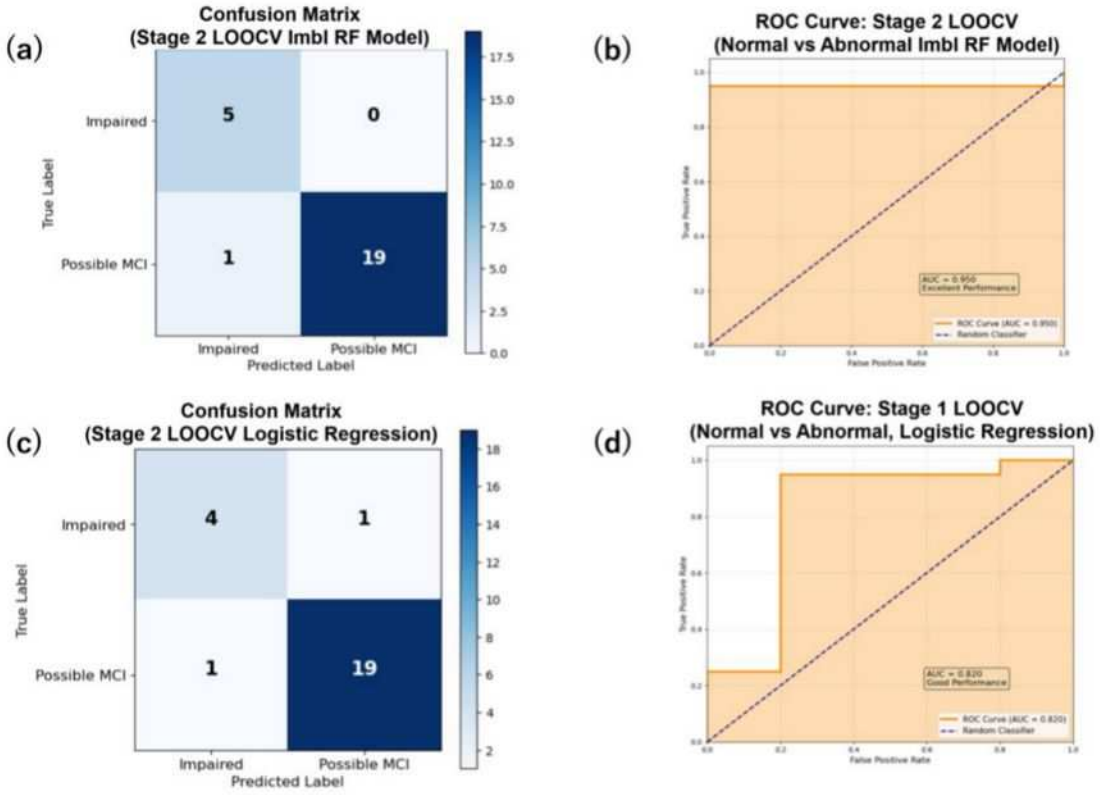
Stage 1 Performance Metrics		
	Imbalanced learn Random Forest	**Logistic Regression**
Accuracy	0.71	0.76
ROC-AUC	0.72	0.71
Macro Precision	0.61	0.67
Macro Recall	0.60	0.66
Macro F1-Score	0.60	0.67
Clinical Metrics		
Sensitivity (Abnormal detection)	0.36	0.48
Specificity (Normal detection)	0.84	0.85
Positive Predictive value	0.43	0.52
Negative Predictive value	0.79	0.83

This table summarizes the predictive performance of the Stage 1 classifiers in distinguishing cogni-tively Normal from Abnormal participants, comparing the tuned Imbalanced-Learn Random Forest with the tuned logistic regres-sion baseline under the same LOOCV procedure. Both mod-els achieved moderate discriminative ability, with the Random Forest showing slightly higher ROC–AUC (0.72 vs. 0.71), while logistic regression achieved higher overall accuracy (0.76 vs. 0.71) and improved macro-averaged precision, recall, and F1-score. Clinically oriented metrics indicate that logistic regression provided higher sensitivity for Abnormal detection (0.48 vs. 0.36) and slightly higher specificity (0.85 vs. 0.84).

8.2 Stage 2 Performance: Possible MCI vs Impaired

Stage 2 demonstrated excellent discriminative performance in distinguishing Possible MCI from Impaired cognitive states in both modeling approaches under LOOCV. The optimized Imbalanced-Learn Random Forest achieved an accuracy of 96.00% with a ROC–AUC of 0.95, while the tuned L1-regularized logistic regression baseline attained an accuracy of 92.0% and a ROC–AUC of 0.82, indicating that severity separation remained strong even in a simpler linear model despite the small Impaired subgroup.

The Random Forest model exhibited a near-perfect clinical performance, with 100% specificity for detecting Impaired cases and 95% sensitivity for identifying Possible MCI, yielding exceptional class-specific metrics for both Impaired (precision = 83.33%, recall = 100%, F1 = 90.91%) and Possible MCI (precision = 100%, recall = 95%, F1 = 97.44%). Logistic regression preserved high sensitivity for Possible MCI (95%) but showed lower specificity for Impaired (80%), with corresponding class metrics of Impaired (precision = 80.0%, recall = 80.0%, F1 = 80.0%) and Possible MCI (precision = 95.0%, recall = 95.0%, F1 = 95.0%), reflecting slightly less sharp separation yet still strong severity discrimination (**Figure 2; Table 5**).



Figure]2.

Stage 2 classification performance for Impaired vs. Possible MCI cognitive status under Leave-One-Out Cross-Validation (LOOCV), comparing the tuned Imbalanced-Learn Random Forest and logistic regression models. Panels (a) and (b) display the confusion matrix and ROC curve, respectively, for the Random Forest model, which demonstrates excellent discrimination with AUC = 0.950 and near-perfect separation between severity categories. Panels (c) and (d) show the corresponding confusion matrix and ROC curve for the tuned logistic regression baseline, which achieves good overall accuracy but somewhat lower ROC–AUC and specificity for Impaired detection compared with the tree-based model.

Table 5
Stage 2 classification performance metrics for Possible MCI vs Impaired cognitive severity discrimination

Stage 2 Performance Metrics		
	Imbalanced Learn Random Forest	**Logistic Regression**
Accuracy	0.975	0.92
ROC-AUC	0.950	0.82
Macro Precision	0.92	0.88
Macro Recall	0.98	0.88
Macro F1-Score	0.94	0.88
Clinical Metrics		
Sensitivity (Possible MCI detection)	0.95	0.95
Specificity (Impaired detection)	1.00	0.80
Positive Predictive value	1.00	0.95
Negative Predictive value	0.83	0.80

This table summarizes the predictive performance of the Stage 2 classifiers in distinguishing Possible MCI from Impaired cognitive status within the abnormal subgroup, comparing the tuned Imbalanced-Learn Random Forest with the tuned logistic regression baseline under the same LOOCV procedure. The Random Forest achieved very high overall accuracy (0.96) and excellent discriminative ability (ROC–AUC = 0.950), with higher macro-averaged precision, recall, and F1-score than logistic regression (0.92/0.98/0.94 vs. 0.88/0.88/0.88), indicating stronger separation between severity categories in this small sample. Clinically oriented metrics show that both models maintained high sensitivity for Possible MCI detection (0.95), while the Random Forest provided perfect specificity for Impaired detection (1.00 vs. 0.80) and higher positive and negative predictive values, suggesting superior rule-in performance for Impaired status compared with the simpler logistic model.

8.3 Feature Importance Based on SHAP Analysis

SHapley Additive exPlanations (SHAP) were applied to the optimized models in both stages to quantify the contribution of individual baseline features to classification out-comes. Feature importance was summarized using SHAP beeswarm plots, with the top 20 features ranked by mean absolute SHAP values for each stage. For the Imbalanced-Learn Random Forest we used TreeExplainer, and for logistic regression we used LinearEx-plainer, which allowed us to compare feature attribution patterns across non-linear and linear classifiers within a common additive framework (**Figures 3 and 4; Table 6**).

Stage 1: In the Random Forest model, the most influential features included creatinine (CRE), uric acid (UA), and exercise-related fluctuation metrics, followed by other physical activity measures and amino -acid–related variables. The top-ranked features were CRE, UA, EX fluctuation coefficient 1, and number of steps per minute, with additional contributors such as D-serine (D-Ser), D-amino -acid proportions (Ser(D/(D + L) × 100), Ala(D/(D + L) × 100)), L-asparagine (L-Asn), cysteine, alanine (Ala), L-glutamic acid (L-Glu), L-methionine (L-Met), adenosine monophosphate, hemoglobin (HGB), red cell distribution width–coefficient of variation (RDW-CV), EX, Activity(Ex), walking time, total steps, and steps per minute. The logistic regression model highlighted a partially overlapping but distinct set of top features, including threonine, number of steps per minute, D-amino -acid proportions, aspartic acid, L-phenylalanine, L-citrulline, glutamic acid, albumin (ALB), mean corpuscular hemoglobin (MCH), LDL-cholesterol, D-asparagine, glucose (GLU), height, glyoxylic acid, methionine sulfoxide, and the Simpson index of gut microbiome diversity, indicating that amino -acid, lipid, and microbiome-related measures also contribute to linear separation of Normal vs. Abnormal cognition. Together, these analyses show that both models converge on renal function, systemic metabolism, amino -acid balance, hematologic indices, and activity variability as key determinants of Stage -1 predictions (**Figure 3**).

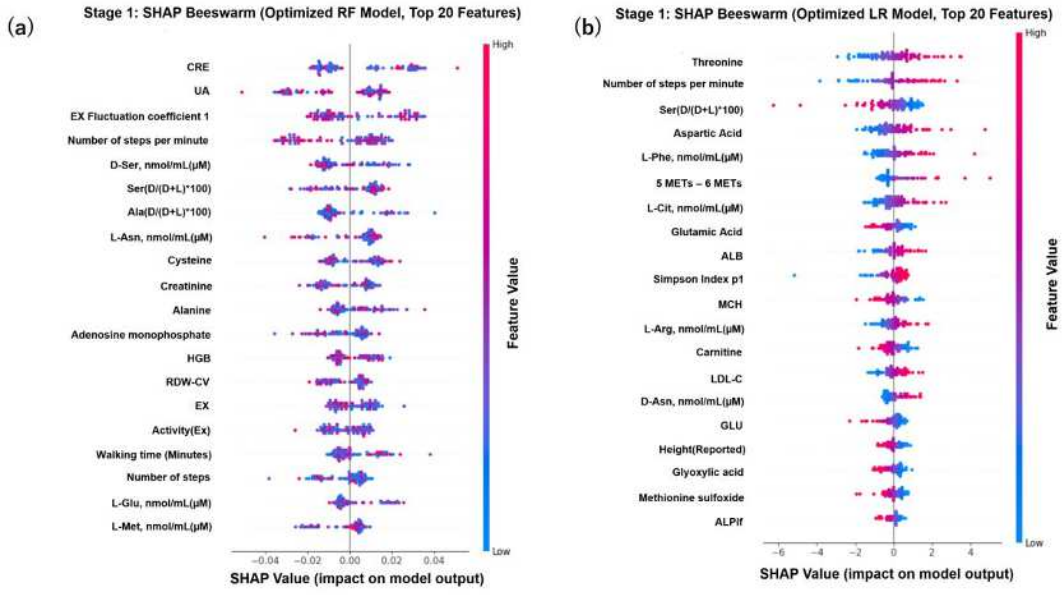


Figure 3.

SHAP beeswarm plots summarizing the top 20 features contributing to Stage 1 classification be-tween Normal and Abnormal cognitive status for both the optimized Imbalanced-Learn Random Forest and logistic regression models under LOOCV. Panel (a) shows the Random Forest model, while panel (b) shows the logistic regression model, with features ordered on the y-axis by mean absolute SHAP value (global importance) and each point representing an individual participant. Point colors indicate the underlying feature value (red = high, blue = low), and the x-axis SHAP values quantify each feature's impact on the predicted probability of Abnormal status, allowing the comparison of shared and model-specific multimodal signals across the two modeling ap-proaches. The (*) in the figure represents multiplication to convert the Ratio of D-Serine to Total Serine into percentage.

Stage 2: A different pattern of feature importance was observed when distinguishing Possible MCI from Impaired cognitive states. In the Random Forest model, the highest-ranked features were glucose (GLU), albumin (ALB), uric acid, and uridine, followed by choline and carnitine, emphasizing energy metabolism, nutritional status, and nucleotide/choline pathways. Several wearable-derived activity fluctuation metrics, including 3 METs or more fluctuation coefficient 5 and 4 METs or more fluctuation coefficient 5, were also ranked among the top features, indicating that high-intensity activity variability carries information about severity. Within the top 20 Random Forest features, multiple amino acids and related metabolites were identified, including methionine sulfoxide, methionine, phenylalanine, proline, ornithine, tryptophan, threonine, 4-aminobutyric acid, and 2-aminobutyric acid, as well as hematological parameters such as lymphocyte percentage (LYMPH%) and monocyte count (MONO#), and niacinamide. The logistic regression Stage -2 model selected complementary predictors, such as EX fluctuation coefficient 2 and 3, measured height, total protein (TP), L-tryptophan, L-asparagine, red blood cell count, creatinine, adenosine monophosphate, 8 METs or higher activity, MCH, age, choline, L-histidine, 4-hydroxyproline, 2 METs-level activity, D-serine, and ergothioneine. Collectively, these SHAP results indicate that both non-linear and linear models attribute Stage -2 discrimination to a coherent multimodal signature spanning glucose and protein metabolism, amino -acid and redox pathways, hematologic indices, and activity-pattern variability (**Figure 4**).

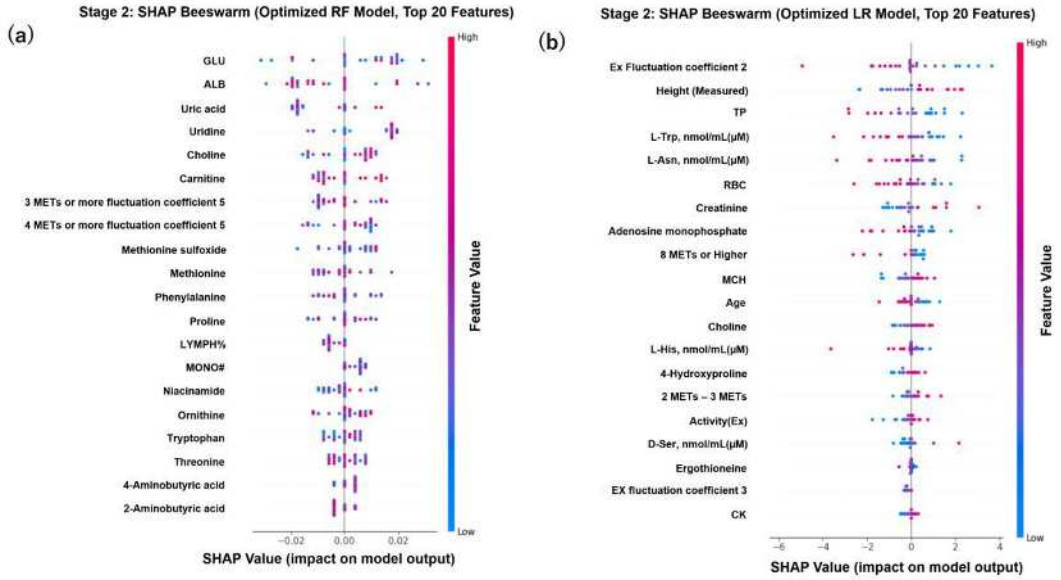


Figure 4:

Stage 2 SHAP beeswarm plots illustrating the top 20 predictive features for distinguishing Possible MCI from Impaired cognitive states for both the optimized Imbalanced-Learn Random Forest and logistic regression models under LOOCV. Panel (a) shows the Random Forest model, and panel (b) shows the logistic regression model, with features ordered by mean absolute SHAP value and each point representing an individual participant colored by the underlying feature value (red = high, blue = low). The horizontal SHAP value axis reflects the magnitude and direction of each feature's contribution to the predicted probability of Impaired versus Possible MCI status, enabling a comparison of shared and model-specific severity-related signals across the two modeling approaches. The # after MONO/ Monocyte represents absolute monocyte count during bio-chemical analysis of the blood.

Table 6

Descriptive Statistics of Top 20 parameters according to MMSE category from Stage 1 and Stage 2 SHAP analysis

Feature	MMSE Category	Count	Mean ± SD	Median	IQR_Lower	IQR_Upper	IQR_Range
UA	Impaired	5	5.48 ± 1.12	6.1	5.3	6.2	5.30-6.20
	Normal	73	5.47 ± 1.28	5.5	4.8	6.2	4.80-6.20
	Possible MCI	20	5.72 ± 1.71	5.7	5.3	6.9	5.28-6.93
CRE	Impaired	5	0.72 ± 0.10	0.8	0.7	0.8	0.72-0.76
	Normal	73	0.89 ± 0.30	0.8	0.7	1.0	0.71-0.97
	Possible MCI	20	0.94 ± 0.38	0.8	0.8	0.9	0.76-0.94
Number of steps per minute	Impaired	5	60.10 ± 26.18	71.2	41.5	72.0	41.49-71.95
	Normal	63	66.83 ± 15.59	66.7	58.3	78.0	58.25-77.99
	Possible MCI	15	50.24 ± 12.97	51.0	40.4	59.4	40.38-59.43
EX fluctuation coefficient 1	Impaired	5	0.44 ± 0.14	0.4	0.4	0.4	0.35-0.44
	Normal	63	0.36 ± 0.12	0.4	0.3	0.4	0.27-0.43
	Possible MCI	15	0.28 ± 0.11	0.3	0.2	0.4	0.19-0.36
D-Ser, nmol/mL(uM)	Impaired	5	2.68 ± 0.75	2.4	2.1	2.9	2.13-2.90
	Normal	73	2.52 ± 0.68	2.4	2.0	3.0	2.00-2.96
	Possible MCI	20	3.09 ± 1.31	2.7	2.4	3.2	2.40-3.21
Ser(D/(D+L)*100)	Impaired	5	2.60 ± 0.54	2.7	2.1	3.0	2.11-3.01
	Normal	73	2.16 ± 0.52	2.1	1.8	2.4	1.76-2.45
	Possible MCI	20	2.85 ± 1.31	2.4	2.2	2.8	2.19-2.78
Ala(D/(D+L)*100)	Impaired	5	1.11 ± 0.52	1.4	0.7	1.5	0.68-1.51
	Normal	73	0.80 ± 0.53	0.6	0.5	1.0	0.45-1.02
	Possible MCI	20	0.94 ± 0.95	0.6	0.5	0.9	0.52-0.86
L-Asn, nmol/mL(μM)	Impaired	5	60.09 ± 13.28	53.8	52.0	62.3	51.96-62.33
	Normal	73	53.02 ± 7.92	51.8	47.8	57.3	47.77-57.25
	Possible MCI	20	48.48 ± 6.36	48.5	45.8	50.7	45.78-50.74
Cysteine	Impaired	5	0.21 ± 0.08	0.2	0.2	0.3	0.16-0.27

	Normal	73	0.21 ± 0.08	0.2	0.2	0.3	0.17-0.26
	Possible MCI	20	0.20 ± 0.06	0.2	0.2	0.2	0.17-0.24
	Impaired	5	49.35 ± 7.15	47.0	45.7	54.1	45.66-54.08
Creatinine	Normal	73	61.60 ± 13.41	59.5	52.8	67.1	52.81-67.09
	Possible MCI	20	70.04 ± 18.62	64.6	59.3	75.4	59.25-75.37
	Impaired	5	52.26 ± 14.42	47.9	42.1	64.0	42.08-63.99
Alanine	Normal	73	50.64 ± 11.22	49.3	42.4	58.5	42.35-58.53
	Possible MCI	20	52.39 ± 7.74	54.3	49.0	57.8	49.01-57.76
	Impaired	5	0.38 ± 0.30	0.5	0.1	0.6	0.11-0.60
Adenosine monophosphate	Normal	73	0.39 ± 0.25	0.4	0.2	0.5	0.19-0.51
	Possible MCI	20	0.32 ± 0.23	0.3	0.1	0.4	0.14-0.41
	Impaired	5	14.98 ± 2.05	14.6	14.2	16.4	14.20-16.40
HGB	Normal	73	13.88 ± 1.32	13.9	13.0	14.8	13.00-14.80
	Possible MCI	20	13.29 ± 1.49	13.6	12.2	14.2	12.15-14.20
	Impaired	5	12.80 ± 0.35	12.9	12.8	13.0	12.80-13.00
RDW-CV	Normal	73	12.93 ± 0.75	12.9	12.5	13.2	12.50-13.20
	Possible MCI	20	12.87 ± 0.59	12.8	12.5	13.2	12.50-13.20
	Impaired	5	3.19 ± 2.21	3.9	1.8	4.4	1.78-4.35
EX	Normal	63	4.63 ± 2.66	4.2	2.6	6.2	2.61-6.23
	Possible MCI	15	4.55 ± 1.54	5.1	3.5	5.6	3.49-5.60
	Impaired	5	1.95 ± 1.26	1.8	1.6	2.8	1.57-2.75
Activity(Ex)	Normal	63	2.74 ± 1.44	2.4	1.8	3.4	1.83-3.39
	Possible MCI	15	3.49 ± 1.34	3.7	2.5	4.6	2.48-4.56
	Impaired	5	62.00 ± 43.10	76.0	34.7	76.7	34.67-76.67
Walking time (Minutes)	Normal	63	85.47 ± 46.97	76.7	46.3	118.2	46.33-118.17
	Possible MCI	15	89.11 ± 26.38	79.0	71.7	100.7	71.67-100.67
Number of steps	Impaired	5	4624.13 ± 4113.53	5408.7	1438.3	5516.3	1438.33-5516.33
	Normal	63	5770.38 ± 3584.58	4866.7	2727.5	7354.5	2727.50-7354.50

	Possible MCI	15	4397.16 ± 1497.80	4127.0	3457.2	5099.8	3457.17-5099.83
	Impaired	5	45.30 ± 17.66	50.1	43.0	57.6	43.01-57.64
L-Glu, nmol/mL(μM)	Normal	73	36.00 ± 15.08	33.9	24.1	45.4	24.13-45.37
	Possible MCI	20	37.82 ± 14.86	37.7	24.8	45.4	24.80-45.40
	Impaired	5	28.21 ± 7.67	30.2	21.0	32.9	21.04-32.91
L-Met, nmol/mL(μM)	Normal	73	26.01 ± 4.25	25.5	23.1	28.6	23.11-28.64
	Possible MCI	20	24.69 ± 3.52	24.2	22.6	27.7	22.61-27.67
	Impaired	5	4.52 ± 0.11	4.6	4.4	4.6	4.40-4.60
ALB	Normal	73	4.64 ± 0.30	4.7	4.5	4.8	4.50-4.80
	Possible MCI	20	4.42 ± 0.24	4.4	4.3	4.6	4.28-4.60
	Impaired	5	104.00 ± 12.59	102.0	100.0	102.0	100.00-102.00
GLU	Normal	73	102.41 ± 16.50	98.0	92.0	109.0	92.00-109.00
	Possible MCI	20	103.85 ± 13.86	100.0	96.0	108.3	96.00-108.25
	Impaired	5	1.06 ± 0.40	0.8	0.8	1.3	0.79-1.31
Uridine	Normal	73	1.27 ± 0.34	1.3	1.0	1.5	1.04-1.46
	Possible MCI	20	1.21 ± 0.26	1.1	1.0	1.3	1.04-1.35
	Impaired	5	1.74 ± 0.41	1.8	1.4	2.0	1.36-2.00
Uric acid	Normal	73	1.80 ± 0.41	1.8	1.5	2.1	1.52-2.11
	Possible MCI	20	1.95 ± 0.47	2.0	1.7	2.2	1.74-2.23
	Impaired	5	5.60 ± 1.83	5.0	4.8	5.9	4.80-5.95
Choline	Normal	73	7.41 ± 1.65	7.3	6.0	8.5	6.04-8.45
	Possible MCI	20	7.52 ± 1.57	7.3	6.4	9.0	6.35-8.97
	Impaired	5	50.10 ± 10.09	51.6	45.7	56.6	45.74-56.57
Carnitine	Normal	73	49.24 ± 12.14	50.7	40.9	57.9	40.89-57.88
	Possible MCI	20	55.75 ± 8.92	58.0	50.0	61.3	49.97-61.35
	Impaired	5	0.47 ± 0.30	0.4	0.3	0.4	0.32-0.43
3METs or more fluctuation coefficient 5	Normal	63	0.86 ± 4.15	0.3	0.3	0.4	0.26-0.42
	Possible MCI	15	0.26 ± 0.08	0.3	0.2	0.3	0.23-0.31
	Impaired	4	0.77 ± 0.72	0.4	0.4	0.8	0.40-0.77

4METs or more fluctuation coefficient 5	Normal	63	0.67 ± 1.26	0.5	0.4	0.7	0.38-0.66
	Possible MCI	15	0.39 ± 0.17	0.4	0.3	0.5	0.25-0.51
	Impaired	5	0.24 ± 0.07	0.3	0.2	0.3	0.19-0.28
Methionine sulfoxide	Normal	73	0.21 ± 0.09	0.2	0.2	0.3	0.15-0.26
	Possible MCI	20	0.24 ± 0.15	0.2	0.2	0.2	0.16-0.25
	Impaired	5	10.87 ± 2.69	11.9	8.0	12.9	8.00-12.94
Methionine	Normal	73	10.70 ± 2.31	10.3	8.9	12.0	8.91-11.97
	Possible MCI	20	11.11 ± 1.78	10.7	9.7	12.2	9.75-12.22
	Impaired	5	302.99 ± 40.85	286.4	275.8	308.8	275.79-308.81
Phenylalanine	Normal	73	327.05 ± 67.01	315.9	278.9	352.6	278.91-352.58
	Possible MCI	20	336.63 ± 51.60	330.5	310.1	357.1	310.08-357.10
	Impaired	5	4.07 ± 1.36	3.7	3.5	5.1	3.51-5.12
Proline	Normal	73	4.18 ± 1.24	4.0	3.4	4.5	3.42-4.51
	Possible MCI	20	4.17 ± 1.23	3.9	3.5	4.4	3.49-4.44
	Impaired	5	12.61 ± 2.92	13.8	11.6	14.7	11.63-14.70
Ornithine	Normal	73	14.22 ± 3.95	15.0	11.2	16.7	11.19-16.69
	Possible MCI	20	15.58 ± 7.36	15.1	11.4	17.3	11.35-17.27
	Impaired	5	0.36 ± 0.09	0.4	0.3	0.4	0.30-0.36
Niacinamide	Normal	73	0.47 ± 0.17	0.5	0.3	0.6	0.34-0.60
	Possible MCI	20	0.39 ± 0.20	0.4	0.3	0.4	0.28-0.40
	Impaired	5	5.00 ± 1.52	4.1	4.1	5.8	4.10-5.80
MONO#	Normal	73	4.73 ± 1.23	4.7	3.8	5.4	3.80-5.40
	Possible MCI	20	5.00 ± 1.72	4.8	3.9	5.8	3.90-5.80
	Impaired	5	33.92 ± 6.94	35.9	28.5	38.8	28.50-38.80
LYMPH%	Normal	73	37.24 ± 8.40	36.7	32.7	43.1	32.70-43.10
	Possible MCI	20	34.35 ± 7.33	35.2	29.3	39.7	29.32-39.70
	Impaired	5	12.12 ± 1.42	12.1	11.9	13.3	11.91-13.27
Threonine	Normal	73	12.96 ± 2.63	12.8	11.1	14.6	11.10-14.61
	Possible MCI	20	11.76 ± 2.80	11.5	10.3	13.2	10.33-13.21
Tryptophan	Impaired	5	82.77 ± 20.66	74.1	69.7	90.3	69.71-90.28

	Normal	73	81.01 ± 14.80	77.3	71.3	89.2	71.31-89.20
	Possible MCI	20	77.88 ± 16.86	76.6	65.7	87.5	65.68-87.47
	Impaired	5	3.48 ± 1.32	3.0	2.6	4.0	2.58-3.99
2-Aminobutyric acid	Normal	73	4.58 ± 1.22	4.3	3.6	5.3	3.64-5.27
	Possible MCI	20	4.76 ± 1.57	4.6	3.9	5.2	3.93-5.15
	Impaired	5	0.03 ± 0.01	0.0	0.0	0.0	0.02-0.04
4-Aminobutyric acid	Normal	73	0.03 ± 0.01	0.0	0.0	0.0	0.02-0.04
	Possible MCI	20	0.03 ± 0.01	0.0	0.0	0.0	0.02-0.04

Please note that the activity metrics were collected from 83 individuals because of the failure of measurement device for 1 month.

8.4 Post Hoc Statistical Validation

Games-Howell tests were performed to identify pairwise differences between MMSE-defined cognitive categories (Normal, Possible MCI, and Impaired) across multimodal variables, accounting for unequal variances and unbalanced sample sizes.

Several standard clinical biomarkers exhibited significant group differences. Creatine kinase (CK) showed highly significant differences in both Normal vs Impaired ($p < 0.001$) and Possible MCI vs Impaired ($p=0.01$) comparisons. Total protein (TP), albumin (ALB), creatinine (CRE), red blood cell count (RBC), and mean corpuscular volume (MCV) displayed significant differences between Normal vs Possible MCI groups ($p < 0.05$).

Amino acid-related measures also showed cognitive category-dependent differences. L-asparagine (L-Asn) and L-threonine (L-Thr) showed significant differences between Normal and Possible MCI ($p < 0.05$), while glycine exhibited significant contrasts in both Normal vs Impaired ($p < 0.01$) and Possible MCI vs Impaired ($p < 0.05$) comparisons.

Several metabolites displayed distinct pairwise differences across cognitive categories. Carnitine levels differed significantly between Normal vs Possible MCI ($p < 0.05$). Creatinine showed multiple significant contrasts: Normal vs Impaired ($p < 0.05$) and Possible MCI vs Impaired ($p < 0.01$). Kynurenine, a tryptophan metabolite linked to neuroinflammation, demonstrated a significant difference between Possible MCI and Impaired ($p < 0.05$).

Physical activity parameters showed significant group differences, primarily between Normal and Possible MCI categories. SMETS/METs (metabolic equivalent tasks) were significantly different between Normal vs Impaired ($p < 0.05$). In addition, total walking exercise, steps per minute, and exercise fluctuation coefficients (EX1, EX2, EX4, and EX5) demonstrated significant difference in Normal vs Possible MCI comparisons ($p < 0.05$). These differences are visualized in **Figure 7**. Detailed results of the Games-Howell analyses are summarized in **Table 9**.

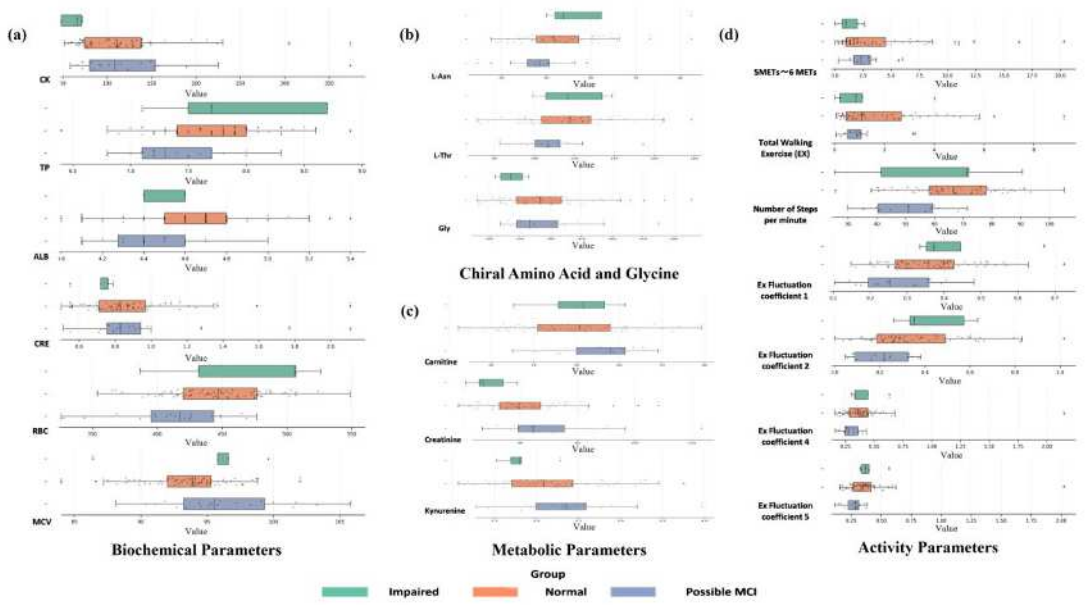


Figure 5. Games-Howell post -hoc comparisons of significant biomarkers and activity parameters ($p < 0.05$, p is corrected by Welch's degrees of freedom correction and the studentized range distribution) across cognitive groups. (a) Significant bio-chemical parameters; (b) significant amino acid markers; (c) significant metabolic parameters; and (d) significant physical activity features (including fluctuation coefficients) are shown. Only parameters with statistically significant group differences after multiple testing correction are displayed, highlighting multi-domain signatures distinguishing Impaired, Normal, and Possible MCI groups. Please note that the activity metrics were collected from 83 individuals.

Table 7
Details of Games-Howell Test on Cognitive Categories with corrected p-val

Data mode wise Games Howell Multi-comparison test between Cognitive Categories										
Biochemical Data										
Name of the Parameter	**Column A**	**Column B**	**Mean(A)**	**Mean(B)**	**Diff**	**SE**	**T**	**df**	**pval**	**hedges**
	Impaired	Normal	61	116.49	-55.49	8.48	-6.54	21.63	0.00004	-1.024
CK	Impaired	Possible MCI	61	130.3	-69.3	17.47	-3.97	22.06	0.0018	-0.99
TP	Normal	Possible MCI	7.73	7.43	0.31	0.10	2.97	34.85	0.0144	0.678
ALB	Normal	Possible MCI	4.63	4.43	0.21	0.06	3.32	36.36	0.0057	0.74
CRE	Impaired	Normal	0.72	0.89	-0.17	0.06	-3.10	10.63	0.0261	-0.59
RBC	Normal	Possible MCI	446.95	411.55	35.40	10.81	3.27	27.75	0.0077	0.88
MCV	Normal	Possible MCI	93.45	96.28	-2.82	1.10	-2.56	24.90	0.0432	-0.76
Amino Acid										

	Column A	Column B	Mean(A)	Mean(B)	Diff	SE	T	df	pval	hedges
L-Asn	Normal	Possible MCI	53.02	48.49	4.53	1.70	2.67	36.80	0.03	0.59
L-Thr	Normal	Possible MCI	118.86	107.54	11.31	4.39	2.57	41.95	0.036	0.53
	Impaired	Normal	236.09	288.38	-52.30	13.05	-4.01	10.20	0.006	-0.77
Gly	Impaired	Normal	236.09	284.92	-48.84	17.34	-2.82	18.87	0.03	-0.82

Metabolites

	Column A	Column B	Mean(A)	Mean(B)	Diff	SE	T	df	pval	hedges
Carnitine	Normal	Possible MCI	49.24	55.75	-6.51	2.45	-2.66	40.46	0.0295	-0.56
	Impaired	Normal	49.35	61.60	-12.25	3.56	-3.44	6.14	0.031	-0.92
Creatinine	Impaired	Possible MCI	49.35	70.04	-20.69	5.25	-3.94	18.09	0.0026	-1.16
Kynurenine	Impaired	Possible MCI	1.82	2.37	-0.54	0.18	-2.95	14.36	0.026	-0.94

Activity

	Column A	Column B	Mean(A)	Mean(B)	Diff	SE	T	df	pval	hedges
5 METs～6 METs	Impaired	Normal	1.26	3.50	-2.23	0.70	-3.18	17.35	0.014	-0.55

Total walking exercise (Ex)	Normal	Possible MCI	1.89	1.06	0.83	0.34	2.44	46.05	0.045	0.46
Number of steps per minute	Normal	Possible MCI	66.83	50.24	16.59	3.88	4.27	24.63	0.0007	1.08
EX fluctuation coefficient 1	Normal	Possible MCI	0.36	0.28	0.08	0.03	2.62	22.86	0.039	0.70
EX fluctuation coefficient 2	Normal	Possible MCI	0.35	0.21	0.14	0.04	3.41	40.34	0.004	0.68
EX fluctuation coefficient 4	Normal	Possible MCI	0.40	0.30	0.11	0.04	2.91	71.86	0.013	0.47
EX fluctuation coefficient 5	Normal	Possible MCI	0.37	0.27	0.10	0.04	2.84	66.56	0.016	0.47

9 Discussion

9.1 Principal Findings and Rationale of the Two-Stage Framework

This study developed and evaluated a two-stage, explainable machine learning framework to predict cognitive status at 18 months using baseline multimodal data from a community-dwelling elderly cohort. Across both the Imbalanced-Learn Random Forest and a penalized logistic regression baseline, the models consistently indicated that combinations of renal and systemic metabolic markers, amino -acid and redox-related metabolites, and wearable-derived physical activity features carry informative signals about cognitive abnormality and severity. Structuring prediction into two sequential stages, first separating cognitively Normal individuals from those with any abnormality (Possible MCI or Impaired), which already represents a generalized binary classifier, and then distinguishing Possible MCI from Impaired, aligns the modeling strategy with the clinical progression of cognitive decline and allows stage-specific biomarker patterns to emerge. It is also important to note that our models were trained in a cohort defined by baseline MMSE ≥ 24, a criterion that allows the inclusion of individuals with possible mild impairment at enrollment. As a result, the classifiers primarily capture associations between baseline multimodal profiles and MMSE-defined severity status at 18 months, on top of pre-existing variability in cognitive function, rather than predicting purely de novo onset of impair-ment from a uniformly normal baseline. Accordingly, our findings should be interpreted in terms of stage-dependent differences in 18-month MMSE severity, not as estimates of strict conversion rates from normal cognition to MCI or dementia.

In Stage 1, both models achieved moderate discrimination: the Random Forest pro-vided slightly higher ROC–AUC values, whereas logistic regression achieved somewhat higher accuracy and sensitivity, suggesting that the underlying signal is robust to different modeling assumptions. In Stage 2, both approaches revealed strong separability between Possible MCI and Impaired cognition, with the Random Forest achieving very high accuracy and AUC and logistic regression still performing well despite the small Impaired subgroup. SHAP analyses applied separately to the Random Forest and logistic regression models provided convergent evidence that the same broad biomarker groups,-renal and metabolic indices, amino acid and oxidative-stress markers, and activity-variability measures, drive predictions at each stage, supporting an integrative view of cognitive ag-ing as a multisystem process.

9.2 Biological Signatures in Stage 1

Stage 1 of our framework, which differentiated Normal participants from those of Possible MCI or Impaired, was primarily influenced by markers of renal function, nitrogen–purine metabolism, and systemic metabolic status in both modeling approaches. Uric acid, creatinine, and blood urea nitrogen emerged as the most influential predictors, indicating that kidney-related metabolic processes offer valuable insights into early cognitive abnormalities. The tuned logistic regression model likewise assigned substantial weight to creatinine. These

routinely measured clinical markers may reflect the combined effects of vascular burden, metabolic homeostasis, and systemic clearance mechanisms might be relevant to brain aging [90,91]. In practical terms, these routine clinical measures may capture systemic clearance, vascular burden, and metabolic homeostasis that collectively influence brain aging, making them attractive for scalable community screening panels.

In Stage 1, amino -acid-related signatures, with multiple chiral and proteinogenic amino -acid measures including D-serine, D-amino -acid proportions, and L-asparagine, alanine, and L-glutamic acid,- together with adenosine monophosphate and cysteine, pointing toward variation in neurotransmission-relevant pools, host–microbiome co-metabolism, and antioxidant capacity ranked among the top contributors. Logistic regression emphasized a partially overlapping subset, including threonine, glutamic acid, tryptophan, and methionine sulfoxide, reinforcing the importance of amino acid and redox pathways under a linear modeling assumption as well. These amino acid features can be interpreted as reflecting broader systemic metabolic and neurotransmission-relevant variation already captured by the multimodal panel, rather than being specific markers of established neurodegeneration at the screening stage. This interpretation is consistent with prior reports indicating that circulating amino -acid profiles, including alanine/asparagine panels and D-amino -acid proportions, differ across the cognitive spectrum and may relate to cognitive phenotypes in a domain dependent manner.

Cysteine emerged as a potential predictor of cognitive function in Stage 1, highlighting their crucial physiological role in neuronal redox homeostasis. As the primary rate-limiting substrate for glutathione synthesis, cysteine is essential for maintaining antioxidant defenses and protecting neural tissues from oxidative stress—an early driver of cognitive decline and neurodegeneration[92]. Recent experimental and clinical literature underscores that tightly regulated cysteine is fundamental to brain aging resilience, supporting their importance as predictive biomarkers for cognitive impairment [93]. In Stage 1 **L-Glu (Glutamic Acid**) as chiral amino acid as a predictor for early-stage screening. Glutamic acid is the brain's principal excitatory neurotransmitter and is essential for synaptic plasticity and memory formation, but sustained elevations in glutamatergic signaling can promote NMDA-receptor–mediated excitotoxicity, a widely discussed contributor to neurodegeneration in Alzheimer's disease and related dementias [94,95].

Chiral amino acids such as **Alanine** as metabolites, **D-Serine(D-Ser)**, the **proportion of D-Serine (D-Ser/(D-Ser + Ser))**, **D-Alanine (D-Ala)**, the **proportion of D-Alanine (D-Ala/(D-Ala + L-Ala))**, the **proportion of D-Asparagine (D-Asn/(D-Asn + L-Asn)**, and **L-Asparagine (L-Asn**) were identified as important predictors of early cognitive decline in our Stage 1 model. These amino acids are implicated in neurotransmission-relevant pathways and systemic energy metabolism, and multiple studies report that circulating amino-acid profiles shift across the continuum from normal cognition to MCI and Alzheimer's disease. Lin et al. showed that blood D-amino acids relate to cognition in a domain-specific manner across MCI and Alzheimer's disease, reporting that D-serine tracked worse global cognitive scores while D-alanine (and its D/total ratio) was associated with more severe behavioral components, whereas D-glutamate showed associations in the opposite direction—supporting the idea that

distinct D-amino acids capture different neurobiological processes relevant to cognitive decline [96]. Targeted amino-acid panels in clinical cohorts include alanine and asparagine among the quantified metabolites and demonstrate that amino-acid patterns can discriminate cognitive states and track disease-related metabolic remodeling [97,98]. Additionally, chiral metabolomics studies show that D-amino-acid proportions in blood can correlate with early cognitive decline, supporting the use of enantiomeric ratios as potential biomarkers [99] . These amino acids participate in central nervous system neurotransmission, synaptic plasticity, and energy metabolism; disturbances in their serum or plasma levels may reflect perturbed metabolic and neurochemical states characteristic of early neurodegeneration [100]. Threonine came as an important predictor in Stage 2 which is an important precursor to **Acetyl-CoA** which is an important neurotransmitter[101]. Threonine has been implicated as a cognitive predictor mainly through AD-related threonine-phosphorylated tau biomarkers and through metabolomics pathways that repeatedly flag glycine/serine/threonine metabolism as altered in cognitively impaired states[102].

Adenosine Monophosphate (AMP) emerges as an important predictor Adenosine monophosphate (AMP) is a core energy-state signal: when cellular ATP falls, AMP rises and activates AMP-activated protein kinase (AMPK), a master regulator that helps restore neuronal energy balance by shifting metabolism toward ATP production. Because synaptic transmission and memory formation are highly energy-demanding, disruption of AMP/AMPK signaling can translate into impaired synaptic plasticity and poorer cognitive performance, which is why AMP-related pathways are frequently discussed in Alzheimer's disease and cognitive decline [103].

HDL-C (High Density Lipid-Cholesterol) in Stage-1, establishing lipid metabolism as the primary biological pathway for cognitive severity discrimination. This finding extends recent evidence linking cardiovascular biomarkers to "brain free water" and cognitive decline [104].

In addition, the Simpson index of gut microbiome diversity appeared among the influential predictors in the Stage 1 logistic regression model, suggesting that overall microbial community structure may also contribute to baseline differences between cognitively Normal and Abnormal participants in this cohort.

9.3 Physical Activity Variability in both Stages

Wearable-derived physical activity metrics also contributed significantly to the predictive signal across both stages, particularly those capturing variability in activity patterns, highlighting their relevance to cognitive status. In Stage 1, the Random Forest model ranked exercise fluctuation coefficients, steps per minute, total steps, and walking time among the top features, while logistic regression also selected steps per minute and MET-based indicators within its most informative variables. In Stage 2, high-MET fluctuation coefficients (e.g., 3 METs or more and 4 METs or more) remained influential in the Random Forest SHAP results, and logistic regression highlighted additional activity-related variables such as EX fluctuation coefficients and MET-derived categories. The prominence of activity variability suggests that

irregular daily movement patterns may provide an early functional signature of instability associated with cognitive decline, complementing static biochemical measures. These wearable-derived features are objective, non-invasive, and scalable, and their consistent appearance among the leading predictors in both models underscores the value of integrating behavioral data with clinical and metabolomic biomarkers for community based cognitive risk assessment.

SHAP analysis also revealed that exercise fluctuation coefficients-including **EX fluctuation coefficients 1**, step-related metrics (**Number of Steps per Minute**), **EX** (Exercise) and **Activity (Ex)** in Stage-1 and **3 METs(Metabolic Equivalent of Task) or more fluctuation coefficient 5**, **4 METs or more fluctuation coefficient 5** from stage 2 are leading digital biomarkers with strong discriminative value in cognitive screening. Our approach validates recent evidence that temporal activity variability and patterns captured with wearable technology or accelerometry data can detect mild cognitive impairment and monitor functional transitions with high reliability. These digital metrics offer scalable, non-invasive tools for early cognitive risk prediction. Recent validation studies confirm that physical activity features from accelerometry data achieve 75.8% accuracy in detecting mild cognitive impairment [105,106]. Our findings extend recent multimodal lifestyle engagement studies showing that physical activity patterns account for differential rates of cognitive decline beyond neuropathological burden[107,108].

9.4 Biological Signatures in Stage 2

In Stage 2, which distinguished Possible MCI from Impaired cognitive conditions, the machine learning classifier and logistic regression baseline both indicated a distinct pattern of key features compared with Stage 1., In the Random Forest model, glucose and albumin, together with uric acid and uridine, emerged as important predictors, highlighting the association of glycemic control, protein and nutritional status, and nucleotide-related metabolism in more advanced impairment. Markers of mitochondrial and energy metabolism were also prominent. Carnitine, involved in fatty acid transport and mitochondrial energy production, and niacinamide, a precursor in NAD+ metabolism, contributed to severity classification. Oxidative stress—related metabolites, such as Methionine, and methionine sulfoxide, further supported the role of redox imbalance in advanced cognitive impairment. Together, these features suggest that the transition from Possible MCI to Impaired status involves coordinated changes in energy metabolism, neurotransmitter balance, and oxidative stress pathways Logistic regression converged on related pathways, emphasizing tryptophan, asparagine, histidine, 4-hydroxyproline, adenosine monophosphate, and ergothioneine among its top features. Across both models, amino acid profiles,- including phenylalanine, proline, ornithine, tryptophan, threonine, and short-chain amino acids,- recurrently appeared, suggesting that coordinated shifts in amino -acid metabolism, neurotransmitter precursors, and inflammatory or nitrogen-handling pathways accompany the transition from Possible MCI to more clearly Impaired status

In Stage 2, **uric acid** remains important, but **uridine** emerges as a key discriminator for separating Possible MCI from Impaired, indicating a shift toward nucleotide-related metabolism in severity classification[109,110]. Prior work in Alzheimer's disease-focused cohorts has suggested that blood uridine levels may relate to clinical progression in a stage-dependent manner, supporting its plausibility as a severity-linked metabolic marker rather than only a general health correlate. More broadly, uridine is biologically connected to synaptic membrane phospholipid pathways (via UTP/CTP-dependent processes) and has been discussed as potentially relevant to neuroprotection and synaptic maintenance in AD/MCI contexts, providing a mechanistic rationale for its prominence in the later-stage mode [111].

Phenylalanine and **proline** emerging as Stage-2 predictors sit at the intersection of neurotransmitter precursors, inflammation-related metabolism, and mitochondrial/oxidative pathways that tend to become more dysregulated with increasing cognitive impairment severity. Phenylalanine is upstream of tyrosine and catecholamine synthesis, and multiple metabolomics studies in AD brain and blood report dysregulated phenylalanine metabolism in association with AD pathology, supporting its relevance for severity discrimination (Possible MCI vs Impaired) [112]. Proline is tightly linked to collagen/ECM turnover (via hydroxyproline), mitochondrial redox balance, and arginine–proline metabolism; pathway-level metabolomics syntheses repeatedly identify arginine and proline metabolism among pathways implicated in progression from normal aging to MCI and AD [113].

Albumin (ALB) is a clinically routine marker that integrates nutritional status, systemic inflammation, and overall frailty, and multiple longitudinal studies show that lower serum albumin (even within the "normal" range) is associated with increased risk of cognitive decline and incident mild cognitive impairment/dementia [114–116]. In Alzheimer's disease–oriented cohorts, low serum albumin has also been linked to higher cerebral β-amyloid burden, supporting a mechanistic pathway in which albumin may influence amyloid transport/clearance ("peripheral sink" hypothesis) and thereby relate to cognitive deterioration[117]. These findings provide biological plausibility for why ALB emerged as an important Stage-2 predictor in our model, where distinguishing Possible MCI from Impaired likely depends on markers that capture systemic resilience, inflammatory burden, and AD-related pathology.

Choline prominence in Stage 2 directly implicates cholinergic neurotransmitter represents direct neurochemical targets for cognitive interventions[118,119].

The convergence of **niacinamide (NAD+ precursor)** and **Methionine, Methionine Sulfoxide (oxidative damage marker)** in Stage 2 reveals a critical energy metabolism-oxidative stress axis in cognitive severity. This represents novel validation of cellular energetics in cognitive classification [120,121]. **L-Met (L-Methionine)** also came as important predictor in Stage 1.

Carnitine, Ornithine, and **Tryptophan** emerging as important Stage-2 predictors index three processes that become increasingly relevant with worsening cognitive status: (i) mitochondrial energy handling (carnitine/acylcarnitines), (ii) nitrogen disposal and polyamine/urea-cycle flux (ornithine), and (iii) neuroinflammation-linked amino-acid catabolism (tryptophan → kynurenine pathway). Acylcarnitine is ester of L-carnitine that emerges from fatty acid

metabolism pathways in mitochondria and peroxisomes, playing a crucial role in energy transfer. Acylcarnitine abnormalities are also increasingly reported in neurodegenerative disorders, including Alzheimer's disease, where they are interpreted as markers of disrupted brain/systemic energy metabolism and mitochondrial dysfunction [122]. Ornithine has been reported as significantly altered across AD/MCI comparisons in plasma metabolomics work, implicating arginine/ornithine (urea-cycle) and polyamine biology, which has also been mechanistically connected to memory impairment through astrocytic urea-cycle activation and downstream ornithine metabolism [123]. Uric Acid, Blood urea Nitrogen and Uridine also came as important predictors which are related to urea-cycle. For tryptophan, systematic reviews/meta-analyses and longitudinal biomarker studies support that peripheral tryptophan is reduced and kynurenine-pathway activity is shifted in AD, and that kynurenine-pathway metabolites track with plasma AD biomarkers and cognition over time, making tryptophan-related metabolism a plausible severity discriminator [124].

In Stage 2, **4-aminobutyric acid (GABA)** and **2-aminobutyric acid (2-AB)** represents two complementary processes relevant to cognitive deterioration: neurotransmitter balance (GABAergic inhibition) and oxidative-stress/redox compensation (2-AB–linked glutathione pathways). GABA is the brain's primary inhibitory neurotransmitter, and converging evidence indicates that GABAergic interneuron dysfunction and an altered excitatory–inhibitory balance contributes to network hyperexcitability and cognitive impairment in Alzheimer's disease [125]. 2-AB is closely tied to cysteine/glutathione biology: it can rise under oxidative stress and has been described as a metabolite that reflects activation of glutathione-related responses [126].

Games-Howell statistical validation revealed kynurenine as a critical neuroinflammatory biomarker, with a significant biomarker in Stage 2 SHAP analysis confirmed by significant elevation in cognitively impaired individuals ($p=0.026$). This finding represents novel validation of the kynurenine pathway in cognitive aging, extending recent discoveries linking kynurenine metabolites to microglial senescence and neuroinflammation. Contemporary research confirms kynurenine as a peripheral inflammatory marker associated with central neuroinflammation, brain gray matter volume decrease, and cognitive dysfunction[127].

This SHAP-driven biomarker discovery represents a significant insight toward precision medicine for cognitive aging, providing both predictive capability and mechanistic interpretability essential for clinical translation and therapeutic development.

9.5 Integrative Interpretation and Clinical Implications:

Overall, findings from both stages support a stage-dependent interpretation of cognitive aging as a systems-level transition captured by multimodal biomarkers. Broad Normal vs. abnormal classifiers (Stage 1) are linked to broad metabolic and behavioral vulnerabilities, reflected in renal and systemic metabolic markers, amino -acid balance, and activity variability, whereas Possible vs. Impaired Classifier (Stage 2) is associated with more pronounced alterations in energy metabolism, nucleotide and choline pathways, oxidative stress, and behavioral

irregularity. The convergence between the Imbalanced-learn Random Forest and logistic regression models, together with SHAP-based explanations, strengthens confidence that these biomarker groups represent persistent signals in the da-ta rather than model-specific artefacts. Another important consideration is the balance between model complexity and sample size. The multimodal predictor space in this study is large compared with the number of participants, which increases the likelihood of overfitting and sensitivity of feature rankings to small perturbations in the data. Although the use of LOOCV, regularized logistic regression, and ensemble methods helps to stabilize estimates, the specific feature sets highlighted by SHAP should be regarded as one plausible model-based representation of the signal in this dataset rather than as a definitive, stable list of biomarkers.

While larger cohorts, longer follow-up, and broader cognitive batteries will be needed to validate and refine these signatures, by identifying feature contributions through SHAP analysis in hierarchal stages, offers transparency that is vital for clinical insights and hypothesis generation. Using accessible biomarkers, such as routine blood tests and wearable derived measures, increases the feasibility of applying similar frameworks in real world settings.

The excellent Stage -2 performance metrics should be interpreted with caution given to the small size of the Impaired subgroup (n = 5), which substantially increases the risk of overfitting despite the use of LOOCV and imbalanced-learning strategies. We acknowledge that overfitting cannot be definitively ruled out, and validation in larger, independent cohorts remains essential. However, when interpreted as exploratory trends rather than definitive quantitative findings, these results offer meaningful methodological insights. The convergence of biomarker signatures across two independent modeling approaches (Random Forest and Lasso logistic regression), the biological plausibility of identified pathways (renal and systemic metabolic indices, glucose metabolism, oxidative stress, amino acid dysregulation), and supporting evidence from Games-Howell post- hoc comparisons collectively suggest that observed patterns may reflect genuine biological signals worthy of further investigation. The primary contribution of this work lies in demonstrating a two-stage, explainable AI framework that integrates minimally invasive multimodal data for stage-aware cognitive assessment, serving as a methodological template for hypothesis- driven validation studies rather than providing immediately actionable clinical biomarkers.

10 Limitation of the study

Several limitations should be considered when interpreting these results. This study was conducted as a secondary, exploratory analysis of a cohort originally designed for intervention evaluation. The sample size was small, with class imbalance; especially, the Impaired group was small. Although leave-one-out cross-validation and imbalanced-learning techniques were employed, the results should be viewed as hypothesis-generating rather than confirmatory.

Future studies should verify these outcomes in larger, independent cohorts with more balanced cognitive categories. Longitudinal analyses of biomarker changes over time may further clarify their roles in cognitive trajectories. External validation and integration with interventional studies are also needed to assess generalizability and clinical impact. Regardless of these limitations, this work offers a structured, interpretable framework for multimodal biomarker discovery in cognitive aging.

This study was conducted on a novel dataset with a relatively small sample size for machine learning analysis. Here the methodology was based on Machine learning based biomarker discovery which is a deviation from the primary case control design of the aforementioned project. Despite employing a robust Leave-One-Out Cross-Validation (LOOCV) approach, the Stage 1 model did not achieve satisfactory level of accuracy. The classification problem and subsequent statistical tests were complicated by pronounced class imbalance, which was partially addressed using algorithms from imbalanced-learn library and the Games-Howell multiple group comparison test. In this study's secondary, post hoc approach does not adhere to the original intervention structure and involves outcome-based reclassification should be interpreted as hypothesis-generating biomarker discovery rather than confirmatory. Overall, this investigation focused on the discovery of candidate biomarkers associated with cognitive function prediction in older adults, highlighting promising directions for future research but requiring further validation in larger and more balanced cohorts.

11 Conclusion

This study offers a two-stage, explainable machine learning framework to predict cognitive status at 18 months using baseline multimodal data. Early cognitive abnormality is mainly linked to systemic metabolic and behavioral vulnerability, while advanced impairment is characterized by neurochemical, energetic, and oxidative metabolic changes. Key predictors include routine blood chemistry, targeted metabolomics, and variability in wearable-derived physical activity. Their contributions were interpreted using SHAP analysis.

By aligning predictive modeling with clinically meaningful stages, this framework delivers a systems-level perspective on cognitive aging that goes beyond single-domain biomarkers. While exploratory and limited by sample size, the use of accessible and minimally invasive measures supports scalable cognitive risk screening and early intervention planning. Validation in larger, independent cohorts will be necessary to confirm generalizability and clinical use.

12 References

1. Petersen, R.C.; Parisi, J.E.; Dickson, D.W.; Johnson, K.A.; Knopman, D.S.; Boeve, B.F.; Jicha, G.A.; Ivnik, R.J.; Smith, G.E.; Tangalos, E.G.; et al. Neuropathologic Features of Amnestic Mild Cognitive Impairment. *Arch. Neurol.* **2006**, *63*, 665, doi:10.1001/archneur.63.5.665.

2. D'Amore, F.M.; Moscatelli, M.; Malvaso, A.; D'Antonio, F.; Rodini, M.; Panigutti, M.; Mirino, P.; Carlesimo, G.A.; Guariglia, C.; Caligiore, D. Explainable Machine Learning on Clinical Features to Predict and Differentiate Alzheimer's Progression by Sex: Toward a Clinician-Tailored Web Interface. *J. Neurol. Sci.* **2025**, *468*, 123361, doi:10.1016/j.jns.2024.123361.

3. De Vugt, M.E.; Verhey, F.R.J. The Impact of Early Dementia Diagnosis and Intervention on Informal Caregivers. *Prog. Neurobiol.* **2013**, *110*, 54–62, doi:10.1016/j.pneurobio.2013.04.005.

4. Topol, E.J. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nat. Med.* **2019**, *25*, 44–56, doi:10.1038/s41591-018-0300-7.

5. Krones, F.; Marikkar, U.; Parsons, G.; Szmul, A.; Mahdi, A. Review of Multimodal Machine Learning Approaches in Healthcare. *Inf. Fusion* **2025**, *114*, 102690, doi:10.1016/j.inffus.2024.102690.

6. Acosta, J.N.; Falcone, G.J.; Rajpurkar, P.; Topol, E.J. Multimodal Biomedical AI. *Nat. Med.* **2022**, *28*, 1773–1784, doi:10.1038/s41591-022-01981-2.

7. Califf, R.M. Biomarker Definitions and Their Applications. *Exp. Biol. Med.* **2018**, *243*, 213–221, doi:10.1177/1535370217750088.

8. Winchester, L.M.; Harshfield, E.L.; Shi, L.; Badhwar, A.; Khleifat, A.A.; Clarke, N.; Dehsarvi, A.; Lengyel, I.; Lourida, I.; Madan, C.R.; et al. Artificial Intelligence for Biomarker Discovery in Alzheimer's Disease and Dementia. *Alzheimers Dement.* **2023**, *19*, 5860–5871, doi:10.1002/alz.13390.

9. Márquez, F.; Yassa, M.A. Neuroimaging Biomarkers for Alzheimer's Disease. *Mol. Neurodegener.* **2019**, *14*, 21, doi:10.1186/s13024-019-0325-5.

10. Meghdadi, A.H.; Stevanović Karić, M.; McConnell, M.; Rupp, G.; Richard, C.; Hamilton, J.; Salat, D.; Berka, C. Resting State EEG Biomarkers of Cognitive Decline Associated with Alzheimer's Disease and Mild Cognitive Impairment. *PLOS ONE* **2021**, *16*, e0244180, doi:10.1371/journal.pone.0244180.

11. Alzheimer Disease Genetics Consortium (ADGC),; The European Alzheimer's Disease Initiative (EADI),; Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium (CHARGE),; Genetic and Environmental Risk in AD/Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease Consortium (GERAD/PERADES),; Kunkle,

B.W.; Grenier-Boley, B.; Sims, R.; Bis, J.C.; Damotte, V.; Naj, A.C.; et al. Genetic Meta-Analysis of Diagnosed Alzheimer's Disease Identifies New Risk Loci and Implicates Aβ, Tau, Immunity and Lipid Processing. *Nat. Genet.* **2019**, *51*, 414–430, doi:10.1038/s41588-019-0358-2.

12. Soleimani Zakeri, N.S.; Pashazadeh, S.; MotieGhader, H. Gene Biomarker Discovery at Different Stages of Alzheimer Using Gene Co-Expression Network Approach. *Sci. Rep.* **2020**, *10*, 12210, doi:10.1038/s41598-020-69249-8.

13. Pedrero-Prieto, C.M.; García-Carpintero, S.; Frontiñán-Rubio, J.; Llanos-González, E.; Aguilera García, C.; Alcaín, F.J.; Lindberg, I.; Durán-Prado, M.; Peinado, J.R.; Rabanal-Ruiz, Y. A Comprehensive Systematic Review of CSF Proteins and Peptides That Define Alzheimer's Disease. *Clin. Proteomics* **2020**, *17*, 21, doi:10.1186/s12014-020-09276-9.

14. Lin, C.; Huang, C.; Huang, K.; Lin, K.; Yen, T.; Kuo, H. A Metabolomic Approach to Identifying Biomarkers in Blood of Alzheimer's Disease. *Ann. Clin. Transl. Neurol.* **2019**, *6*, 537–545, doi:10.1002/acn3.726.

15. Saji, N.; Murotani, K.; Hisada, T.; Kunihiro, T.; Tsuduki, T.; Sugimoto, T.; Kimura, A.; Niida, S.; Toba, K.; Sakurai, T. Relationship between Dementia and Gut Microbiome-Associated Metabolites: A Cross-Sectional Study in Japan. *Sci. Rep.* **2020**, *10*, 8088, doi:10.1038/s41598-020-65196-6.

16. Mander, B.A.; Winer, J.R.; Jagust, W.J.; Walker, M.P. Sleep: A Novel Mechanistic Pathway, Biomarker, and Treatment Target in the Pathology of Alzheimer's Disease? *Trends Neurosci.* **2016**, *39*, 552–566, doi:10.1016/j.tins.2016.05.002.

17. Montero-Odasso, M.M.; Sarquis-Adamson, Y.; Speechley, M.; Borrie, M.J.; Hachinski, V.C.; Wells, J.; Riccio, P.M.; Schapira, M.; Sejdic, E.; Camicioli, R.M.; et al. Association of Dual-Task Gait With Incident Dementia in Mild Cognitive Impairment: Results From the Gait and Brain Study. *JAMA Neurol.* **2017**, *74*, 857, doi:10.1001/jamaneurol.2017.0643.

18. Kourtis, L.C.; Regele, O.B.; Wright, J.M.; Jones, G.B. Digital Biomarkers for Alzheimer's Disease: The Mobile/Wearable Devices Opportunity. *Npj Digit. Med.* **2019**, *2*, 9, doi:10.1038/s41746-019-0084-2.

19. Teunissen, C.E.; Verberk, I.M.W.; Thijssen, E.H.; Vermunt, L.; Hansson, O.; Zetterberg, H.; Van Der Flier, W.M.; Mielke, M.M.; Del Campo, M. Blood-Based Biomarkers for Alzheimer's Disease: Towards Clinical Implementation. *Lancet Neurol.* **2022**, *21*, 66–77, doi:10.1016/S1474-4422(21)00361-6.

20. Folstein, M.F.; Folstein, S.E.; McHugh, P.R. "Mini-Mental State." *J. Psychiatr. Res.* **1975**, *12*, 189–198, doi:10.1016/0022-3956(75)90026-6.

21. Ahmadzadeh, M.; Cosco, T.D.; Best, J.R.; Christie, G.J.; DiPaola, S. Predictors of the Rate of Cognitive Decline in Older Adults Using Machine Learning. *PLOS ONE* **2023**, *18*, e0280029, doi:10.1371/journal.pone.0280029.

22. Ponce-Bobadilla, A.V.; Schmitt, V.; Maier, C.S.; Mensing, S.; Stodtmann, S. Practical Guide to SHAP Analysis: Explaining Supervised Machine Learning Model Predictions in Drug Development. *Clin. Transl. Sci.* **2024**, *17*, e70056, doi:10.1111/cts.70056.

23. Huang, H.-H.; Li, J.; Cho, W.C. Editorial: Integrative Analysis for Complex Disease Biomarker Discovery. *Front. Bioeng. Biotechnol.* **2023**, *11*, 1273084, doi:10.3389/fbioe.2023.1273084.

24. Cummings, J. The National Institute on Aging—Alzheimer's Association Framework on Alzheimer's Disease: Application to Clinical Trials. *Alzheimers Dement.* **2019**, *15*, 172–178, doi:10.1016/j.jalz.2018.05.006.

25. Gauthier, S.; Zhang, H.; Ng, K.P.; Pascoal, T.A.; Rosa-Neto, P. Impact of the Biological Definition of Alzheimer's Disease Using Amyloid, Tau and Neurodegeneration (ATN): What about the Role of Vascular Changes, Inflammation, Lewy Body Pathology? *Transl. Neurodegener.* **2018**, *7*, 12, doi:10.1186/s40035-018-0117-9.

26. Practice Guideline Update Summary: Mild Cognitive Impairment [RETIRED]: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology. *Neurology* **2018**, *90*, 126–135, doi:10.1212/WNL.0000000000004826.

27. Mitchell, A.J.; Shiri-Feshki, M. Rate of Progression of Mild Cognitive Impairment to Dementia – Meta-analysis of 41 Robust Inception Cohort Studies. *Acta Psychiatr. Scand.* **2009**, *119*, 252–265, doi:10.1111/j.1600-0447.2008.01326.x.

28. Chui, H.C.; Ramirez-Gomez, L. Clinical and Imaging Features of Mixed Alzheimer and Vascular Pathologies. *Alzheimers Res. Ther.* **2015**, *7*, 21, doi:10.1186/s13195-015-0104-7.

29. Arevalo-Rodriguez, I.; Smailagic, N.; Ciapponi, A.; Sanchez-Perez, E.; Giannakou, A.; Roqué I Figuls, M.; Pedraza, O.L.; Bonfill Cosp, X.; Cullum, S. Mini-Mental State Examination (MMSE) for the Detection of Alzheimer's Disease and Other Dementias in People with Mild Cognitive Impairment (MCI). In *Cochrane Database of Systematic Reviews*; The Cochrane Collaboration, Ed.; John Wiley & Sons, Ltd: Chichester, UK, 2013; p. CD010783.

30. Truong, Q.C.; Cervin, M.; Choo, C.C.; Numbers, K.; Bentvelzen, A.C.; Kochan, N.A.; Brodaty, H.; Sachdev, P.S.; Medvedev, O.N. Examining the Validity of the Mini-Mental State Examination (MMSE) and Its Domains Using Network Analysis. *Psychogeriatrics* **2024**, *24*, 259–271, doi:10.1111/psyg.13069.

31. Gluhm, S.; Goldstein, J.; Loc, K.; Colt, A.; Liew, C.V.; Corey-Bloom, J. Cognitive Performance on the Mini-Mental State Examination and the Montreal Cognitive Assessment Across the Healthy Adult Lifespan. *Cogn. Behav. Neurol.* **2013**, *26*, 1–5, doi:10.1097/WNN.0b013e31828b7d26.

32. Fleming, T.R.; Powers, J.H. Biomarkers and Surrogate Endpoints in Clinical Trials. *Stat. Med.* **2012**, *31*, 2973–2984, doi:10.1002/sim.5403.

33. Cagney, D.N.; Sul, J.; Huang, R.Y.; Ligon, K.L.; Wen, P.Y.; Alexander, B.M. The FDA NIH Biomarkers, EndpointS, and Other Tools (BEST) Resource in Neuro-Oncology. *Neuro-Oncol.* **2018**, *20*, 1162–1172, doi:10.1093/neuonc/nox242.

34. Chau, C.H.; Rixe, O.; McLeod, H.; Figg, W.D. Validation of Analytic Methods for Biomarkers Used in Drug Development. *Clin. Cancer Res.* **2008**, *14*, 5967–5976, doi:10.1158/1078-0432.CCR-07-4535.

35. Prentice, R.L. Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria. *Stat. Med.* **1989**, *8*, 431–440, doi:10.1002/sim.4780080407.

36. Vasudevan, S.; Saha, A.; Tarver, M.E.; Patel, B. Digital Biomarkers: Convergence of Digital Health Technologies and Biomarkers. *Npj Digit. Med.* **2022**, *5*, 36, doi:10.1038/s41746-022-00583-z.

37. Daniore, P.; Nittas, V.; Haag, C.; Bernard, J.; Gonzenbach, R.; Von Wyl, V. From Wearable Sensor Data to Digital Biomarker Development: Ten Lessons Learned and a Framework Proposal. *Npj Digit. Med.* **2024**, *7*, 161, doi:10.1038/s41746-024-01151-3.

38. Ng, S.; Masarone, S.; Watson, D.; Barnes, M.R. The Benefits and Pitfalls of Machine Learning for Biomarker Discovery. *Cell Tissue Res.* **2023**, *394*, 17–31, doi:10.1007/s00441-023-03816-z.

39. Jasodanand, V.H.; Kowshik, S.S.; Puducheri, S.; Romano, M.F.; Xu, L.; Au, R.; Kolachalama, V.B. AI-Driven Fusion of Multimodal Data for Alzheimer's Disease Biomarker Assessment. *Nat. Commun.* **2025**, *16*, 7407, doi:10.1038/s41467-025-62590-4.

40. Weiner, M.W.; Veitch, D.P.; Aisen, P.S.; Beckett, L.A.; Cairns, N.J.; Green, R.C.; Harvey, D.; Jack, C.R.; Jagust, W.; Liu, E.; et al. The Alzheimer's Disease Neuroimaging Initiative: A Review of Papers Published since Its Inception. *Alzheimers Dement.* **2012**, *8*, doi:10.1016/j.jalz.2011.09.172.

41. Veitch, D.P.; Weiner, M.W.; Miller, M.; Aisen, P.S.; Ashford, M.A.; Beckett, L.A.; Green, R.C.; Harvey, D.; Jack, C.R.; Jagust, W.; et al. The Alzheimer's Disease Neuroimaging Initiative in the Era of Alzheimer's Disease Treatment: A Review of ADNI Studies from 2021 to 2022. *Alzheimers Dement.* **2024**, *20*, 652–694, doi:10.1002/alz.13449.

42. Tremblay-Mercier, J.; Madjar, C.; Das, S.; Pichet Binette, A.; Dyke, S.O.M.; Étienne, P.; Lafaille-Magnan, M.-E.; Remz, J.; Bellec, P.; Louis Collins, D.; et al. Open Science Datasets from PREVENT-AD, a Longitudinal Cohort of Pre-Symptomatic Alzheimer's Disease. *NeuroImage Clin.* **2021**, *31*, 102733, doi:10.1016/j.nicl.2021.102733.

43. Koychev, I.; Lawson, J.; Chessell, T.; Mackay, C.; Gunn, R.; Sahakian, B.; Rowe, J.B.; Thomas, A.J.; Rochester, L.; Chan, D.; et al. Deep and Frequent Phenotyping Study Protocol:

An Observational Study in Prodromal Alzheimer's Disease. *BMJ Open* **2019**, *9*, e024498, doi:10.1136/bmjopen-2018-024498.

44. Tosun, D.; Veitch, D.; Aisen, P.; Jack, C.R.; Jagust, W.J.; Petersen, R.C.; Saykin, A.J.; Bollinger, J.; Ovod, V.; Mawuenyega, K.G.; et al. Detection of β-Amyloid Positivity in Alzheimer's Disease Neuroimaging Initiative Participants with Demographics, Cognition, MRI and Plasma Biomarkers. *Brain Commun.* **2021**, *3*, fcab008, doi:10.1093/braincomms/fcab008.

45. Xie, H.; Wang, Y.; Tao, S.; Huang, S.; Zhang, C.; Lv, Z. Wearable Sensor-Based Daily Life Walking Assessment of Gait for Distinguishing Individuals With Amnestic Mild Cognitive Impairment. *Front. Aging Neurosci.* **2019**, *11*, 285, doi:10.3389/fnagi.2019.00285.

46. Zeitzer, J.M.; David, R.; Friedman, L.; Mulin, E.; Garcia, R.; Wang, J.; Yesavage, J.A.; Robert, P.H.; Shannon, W. Phenotyping Apathy in Individuals With Alzheimer Disease Using Functional Principal Component Analysis. *Am. J. Geriatr. Psychiatry* **2013**, *21*, 391–397, doi:10.1016/j.jagp.2012.12.012.

47. Winer, J.R.; Lok, R.; Weed, L.; He, Z.; Poston, K.L.; Mormino, E.C.; Zeitzer, J.M. Impaired 24-h Activity Patterns Are Associated with an Increased Risk of Alzheimer's Disease, Parkinson's Disease, and Cognitive Decline. *Alzheimers Res. Ther.* **2024**, *16*, 35, doi:10.1186/s13195-024-01411-0.

48. James, C.; Ranson, J.M.; Everson, R.; Llewellyn, D.J. Performance of Machine Learning Algorithms for Predicting Progression to Dementia in Memory Clinic Patients. *JAMA Netw. Open* **2021**, *4*, e2136553, doi:10.1001/jamanetworkopen.2021.36553.

49. Jeong, W.; Chung, W. Development of Machine Learning Models for Prediction of Current and Future Dementia. *PLOS One* **2025**, *20*, e0330213, doi:10.1371/journal.pone.0330213.

50. Martin, S.A.; Townend, F.J.; Barkhof, F.; Cole, J.H. Interpretable Machine Learning for Dementia: A Systematic Review. *Alzheimers Dement.* **2023**, *19*, 2135–2149, doi:10.1002/alz.12948.

51. Vlontzou, M.E.; Athanasiou, M.; Dalakleidi, K.V.; Skampardoni, I.; Davatzikos, C.; Nikita, K. A Comprehensive Interpretable Machine Learning Framework for Mild Cognitive Impairment and Alzheimer's Disease Diagnosis. *Sci. Rep.* **2025**, *15*, 8410, doi:10.1038/s41598-025-92577-6.

52. Ports, K.; Dai, J.; Conniff, K.; Corrada, M.M.; Manson, S.M.; O'Connell, J.; Jiang, L. Machine Learning to Predict Dementia for American Indian and Alaska Native Peoples: A Retrospective Cohort Study. *Lancet Reg. Health - Am.* **2025**, *43*, 101013, doi:10.1016/j.lana.2025.101013.

53. Lee, M.-W.; Kim, H.W.; Choe, Y.S.; Yang, H.S.; Lee, J.; Lee, H.; Yong, J.H.; Kim, D.; Lee, M.; Kang, D.W.; et al. A Multimodal Machine Learning Model for Predicting Dementia Conversion in Alzheimer's Disease. *Sci. Rep.* **2024**, *14*, 12276, doi:10.1038/s41598-024-60134-2.

54. Hao, Y.; Cheng, C.; Li, J.; Li, H.; Di, X.; Zeng, X.; Jin, S.; Han, X.; Liu, C.; Wang, Q.; et al. Multimodal Integration in Health Care: Development With Applications in Disease Management. *J. Med. Internet Res.* **2025**, *27*, e76557–e76557, doi:10.2196/76557.

55. Lyall, D.M.; Kormilitzin, A.; Lancaster, C.; Sousa, J.; Petermann-Rocha, F.; Buckley, C.; Harshfield, E.L.; Iveson, M.H.; Madan, C.R.; McArdle, R.; et al. Artificial Intelligence for Dementia—Applied Models and Digital Health. *Alzheimers Dement.* **2023**, *19*, 5872–5884, doi:10.1002/alz.13391.

56. García-Gutiérrez, F.; Matias-Guiu, J.A.; Ayala, J.L. Deep Multimodal Learning for Domain-Level Cognitive Decline Prediction in Alzheimer's Disease. *Front. Artif. Intell.* **2025**, *8*, 1731062, doi:10.3389/frai.2025.1731062.

57. Katsimpras, G.; Aisopos, F.; Garrard, P.; Vidal, M.-E.; Paliouras, G. Improving Early Prognosis of Dementia Using Machine Learning Methods. *ACM Trans. Comput. Healthc.* **2022**, *3*, 1–16, doi:10.1145/3502433.

58. Campagner, A.; Marconi, L.; Bianchi, E.; Arosio, B.; Rossi, P.; Annoni, G.; Lucchi, T.A.; Montano, N.; Cabitza, F. Uncovering Hidden Subtypes in Dementia: An Unsupervised Machine Learning Approach to Dementia Diagnosis and Personalization of Care. *J. Biomed. Inform.* **2025**, *165*, 104799, doi:10.1016/j.jbi.2025.104799.

59. Bucholc, M.; James, C.; Al Khleifat, A.; Badhwar, A.; Clarke, N.; Dehsarvi, A.; Madan, C.R.; Marzi, S.J.; Shand, C.; Schilder, B.M.; et al. Artificial Intelligence for Dementia Research Methods Optimization. *ArXiv* **2023**, arXiv:2303.01949v1.

60. Prakash, J.; Wang, V.; Quinn, R.E.; Mitchell, C.S. Unsupervised Machine Learning to Identify Separable Clinical Alzheimer's Disease Sub-Populations. *Brain Sci.* **2021**, *11*, 977, doi:10.3390/brainsci11080977.

61. Prevot, E.; Shand, C.; Oxtoby, N.; Initiative, for A.D.N. How Reproducible Are Data-Driven Subtypes of Alzheimer's Disease Atrophy? 2024.

62. Ali, S.B.; Ghannam, N.E.; Mancy, H.; Elkilany, B.Gh. Multimodal Self-Supervised Learning for Early Alzheimer's: Cross-Modal MRI–PET, Longitudinal Signals, and Site Invariance. *Diagnostics* **2025**, *15*, 3135, doi:10.3390/diagnostics15243135.

63. Peng, C.; Li, L.; Peng, D. A Method-Oriented Review of Explainable Artificial Intelligence for Neurological Medical Imaging. *Expert Syst.* **2025**, *42*, e70111, doi:10.1111/exsy.70111.

64. Govindarajan, R.; Thirunadanasikamani, K.; Napa, K.K.; Sathya, S.; Murugan, J.S.; Priya, K.G.C. Development of an Explainable Machine Learning Model for Alzheimer's Disease Prediction Using Clinical and Behavioural Features. *MethodsX* **2025**, *15*, 103491, doi:10.1016/j.mex.2025.103491.

65. Vimbi, V.; Shaffi, N.; Mahmud, M. Interpreting Artificial Intelligence Models: A Systematic Review on the Application of LIME and SHAP in Alzheimer's Disease Detection. *Brain Inform.* **2024**, *11*, 10, doi:10.1186/s40708-024-00222-1.

66. Shankar, R.; Goh, Z.; Devi, F.; Xu, Q. A Systematic Review of Explainable Artificial Intelligence Methods for Speech-Based Cognitive Decline Detection. *Npj Digit. Med.* **2025**, *8*, 724, doi:10.1038/s41746-025-02105-z.

67. Dai, Y.; Zou, F.; Zou, B. High-Dimensional Biomarker Identification for Scalable and Interpretable Disease Prediction via Machine Learning Models. *BioRxiv Prepr. Serv. Biol.* **2024**, 2024.10.04.616748, doi:10.1101/2024.10.04.616748.

68. Calle, P.; Bates, A.; Reynolds, J.C.; Liu, Y.; Cui, H.; Ly, S.; Wang, C.; Zhang, Q.; De Armendi, A.J.; Shettar, S.S.; et al. Integration of Nested Cross-Validation, Automated Hyperparameter Optimization, High-Performance Computing to Reduce and Quantify the Variance of Test Performance Estimation of Deep Learning Models. *Comput. Methods Programs Biomed.* **2025**, *272*, 109063, doi:10.1016/j.cmpb.2025.109063.

69. Li, H.; Govindarajan, V.; Ang, T.F.; Shaikh, Z.A.; Ksibi, A.; Chen, Y.-L.; Ku, C.S.; Leong, M.C.; Shabaruddin, F.H.; Wan Ishak, W.Z.; et al. MSPO: A Machine Learning Hyperparameter Optimization Method for Enhanced Breast Cancer Image Classification. *Digit. Health* **2025**, *11*, 20552076251361603, doi:10.1177/20552076251361603.

70. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.

71. Lebedev, A.V.; Westman, E.; Van Westen, G.J.P.; Kramberger, M.G.; Lundervold, A.; Aarsland, D.; Soininen, H.; Kłoszewska, I.; Mecocci, P.; Tsolaki, M.; et al. Random Forest Ensembles for Detection and Prediction of Alzheimer's Disease with a Good between-Cohort Robustness. *NeuroImage Clin.* **2014**, *6*, 115–125, doi:10.1016/j.nicl.2014.08.023.

72. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data? 2022.

73. Probst, P.; Wright, M.; Boulesteix, A.-L. Hyperparameters and Tuning Strategies for Random Forest. **2018**, doi:10.48550/ARXIV.1804.03515.

74. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics* **2010**, *26*, 1340–1347, doi:10.1093/bioinformatics/btq134.

75. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions 2017.

76. van Batenburg, D. Mathematically Rigorous Proofs for Shapley Explanations 2025.

77. Bhattarai, P.; Thakuri, D.S.; Nie, Y.; Chand, G.B. Explainable AI-Based Deep-SHAP for Mapping the Multivariate Relationships between Regional Neuroimaging Biomarkers and Cognition. *Eur. J. Radiol.* **2024**, *174*, 111403, doi:10.1016/j.ejrad.2024.111403.

78. Bifarin, O.O.; Fernández, F.M. Automated Machine Learning and Explainable AI (AutoML-XAI) for Metabolomics: Improving Cancer Diagnostics. *J. Am. Soc. Mass Spectrom.* **2024**, *35*, 1089–1100, doi:10.1021/jasms.3c00403.

79. Svensson, D.; Hermansson, E.; Nikolaou, N.; Sechidis, K.; Lipkovich, I. Overview and Practical Recommendations on Using Shapley Values for Identifying Predictive Biomarkers via CATE Modeling 2025.

80. Lamsaf, A.; Samale, P.; Neves, J. Causal-SHAP: Feature Selection with Explainability and Causal Analysis. In *Pattern Recognition and Image Analysis*; Gonçalves, N., Oliveira, H.P., Sánchez, J.A., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, 2026; Vol. 15937, pp. 429–440 ISBN 978-3-031-99564-4.

81. Sears, S.M.; Hewett, S.J. Influence of Glutamate and GABA Transport on Brain Excitatory/Inhibitory Balance. *Exp. Biol. Med.* **2021**, *246*, 1069–1083, doi:10.1177/1535370221989263.

82. Arnone, D.; Saraykar, S.; Salem, H.; Teixeira, A.L.; Dantzer, R.; Selvaraj, S. Role of Kynurenine Pathway and Its Metabolites in Mood Disorders: A Systematic Review and Meta-Analysis of Clinical Studies. *Neurosci. Biobehav. Rev.* **2018**, *92*, 477–485, doi:10.1016/j.neubiorev.2018.05.031.

83. Chen, X.; Xu, D.; Yu, J.; Song, X.-J.; Li, X.; Cui, Y.-L. Tryptophan Metabolism Disorder-Triggered Diseases, Mechanisms, and Therapeutic Strategies: A Scientometric Review. *Nutrients* **2024**, *16*, 3380, doi:10.3390/nu16193380.

84. Teruya, T.; Chen, Y.-J.; Fukuji, Y.; Kondoh, H.; Yanagida, M. Whole-Blood Metabolomics of Dementia Patients Reveal Classes of Disease-Linked Metabolites. *Proc. Natl. Acad. Sci.* **2021**, *118*, e2022857118, doi:10.1073/pnas.2022857118.

85. Kimura, N.; Sasaki, Y.; Masuda, T.; Ataka, T.; Eguchi, A.; Kakuma, T.; Matsubara, E. Lifestyle Factors That Affect Cognitive Function–a Longitudinal Objective Analysis. *Front. Public Health* **2023**, *11*, 1215419, doi:10.3389/fpubh.2023.1215419.

86. Del Pozo Cruz, B.; Ahmadi, M.; Naismith, S.L.; Stamatakis, E. Association of Daily Step Count and Intensity With Incident Dementia in 78 430 Adults Living in the UK. *JAMA Neurol.* **2022**, *79*, 1059, doi:10.1001/jamaneurol.2022.2672.

87. Zlatar, Z.Z.; Godbole, S.; Takemoto, M.; Crist, K.; Sweet, C.M.C.; Kerr, J.; Rosenberg, D.E. Changes in Moderate Intensity Physical Activity Are Associated With Better Cognition in

the Multilevel Intervention for Physical Activity in Retirement Communities (MIPARC) Study. *Am. J. Geriatr. Psychiatry* **2019**, *27*, 1110–1121, doi:10.1016/j.jagp.2019.04.011.

88. Healy, G.N. Balancing Our Day for Heart Health. *Eur. Heart J.* **2024**, *45*, 472–474, doi:10.1093/eurheartj/ehad824.

89. Kossowska, M.; Olejniczak, S.; Karbowiak, M.; Mosiej, W.; Zielińska, D.; Brzezicka, A. The Interplay between Gut Microbiota and Cognitive Functioning in the Healthy Aging Population: A Systematic Review. *Nutrients* **2024**, *16*, 852, doi:10.3390/nu16060852.

90. Stocker, H.; Beyer, L.; Trares, K.; Perna, L.; Rujescu, D.; Holleczek, B.; Beyreuther, K.; Gerwert, K.; Schöttker, B.; Brenner, H. Association of Kidney Function With Development of Alzheimer Disease and Other Dementias and Dementia-Related Blood Biomarkers. *JAMA Netw. Open* **2023**, *6*, e2252387, doi:10.1001/jamanetworkopen.2022.52387.

91. Leeuw, F.A.; Tijms, B.M.; Doorduijn, A.S.; Hendriksen, H.M.A.; Rest, O.; Van Der Schueren, M.A.E.; Visser, M.; Den Heuvel, E.G.H.M.; Wijk, N.; Bierau, J.; et al. LDL Cholesterol and Uridine Levels in Blood Are Potential Nutritional Biomarkers for Clinical Progression in Alzheimer's Disease: The NUDAD Project. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **2020**, *12*, doi:10.1002/dad2.12120.

92. Hajjar, I.; Hayek, S.S.; Goldstein, F.C.; Martin, G.; Jones, D.P.; Quyyumi, A. Oxidative Stress Predicts Cognitive Decline with Aging in Healthy Adults: An Observational Study. *J. Neuroinflammation* **2018**, *15*, 17, doi:10.1186/s12974-017-1026-z.

93. Paul, B.D.; Sbodio, J.I.; Snyder, S.H. Cysteine Metabolism in Neuronal Redox Homeostasis. *Trends Pharmacol. Sci.* **2018**, *39*, 513–524, doi:10.1016/j.tips.2018.02.007.

94. Wang, R.; Reddy, P.H. Role of Glutamate and NMDA Receptors in Alzheimer's Disease. *J. Alzheimer's Dis.* **2017**, *57*, 1041–1048, doi:10.3233/JAD-160763.

95. Hynd, M. Glutamate-Mediated Excitotoxicity and Neurodegeneration in Alzheimer?S Disease. *Neurochem. Int.* **2004**, *45*, 583–595, doi:10.1016/j.neuint.2004.03.007.

96. Lin, C.-H.; Yang, H.-T.; Lane, H.-Y. D-Glutamate, D-Serine, and D-Alanine Differ in Their Roles in Cognitive Decline in Patients with Alzheimer's Disease or Mild Cognitive Impairment. *Pharmacol. Biochem. Behav.* **2019**, *185*, 172760, doi:10.1016/j.pbb.2019.172760.

97. Corso, G.; Cristofano, A.; Sapere, N.; la Marca, G.; Angiolillo, A.; Vitale, M.; Fratangelo, R.; Lombardi, T.; Porcile, C.; Intrieri, M.; et al. Serum Amino Acid Profiles in Normal Subjects and in Patients with or at Risk of Alzheimer Dementia. *Dement. Geriatr. Cogn. Disord. Extra* **2017**, *7*, 143–159, doi:10.1159/000466688.

98. Youn, C.; Caillaud, M.L.; Li, Y.; Gallagher, I.A.; Strasser, B.; Fuchs, D.; Tanaka, H.; Haley, A.P. Association between Large Neutral Amino Acids and Brain Integrity in Middle-Aged Adults at Metabolic Risk 2024.

99. Kimura, R.; Tsujimura, H.; Tsuchiya, M.; Soga, S.; Ota, N.; Tanaka, A.; Kim, H. Development of a Cognitive Function Marker Based on D-Amino Acid Proportions Using New Chiral Tandem LC-MS/MS Systems. *Sci. Rep.* **2020**, *10*, 804, doi:10.1038/s41598-020-57878-y.

100. Holeček, M. Aspartic Acid in Health and Disease. *Nutrients* **2023**, *15*, 4023, doi:10.3390/nu15184023.

101. Buerger, K.; Teipel, S.J.; Zinkowski, R.; Blennow, K.; Arai, H.; Engel, R.; Hofmann–Kiefer, K.; McCulloch, C.; Ptok, U.; Heun, R.; et al. CSF Tau Protein Phosphorylated at Threonine 231 Correlates with Cognitive Decline in MCI Subjects. *Neurology* **2002**, *59*, 627–629, doi:10.1212/WNL.59.4.627.

102. Bird, M.; Nunn, P.; Lord, L. Formation of Glycine and Aminoacetone from L-Threonine by Rat Liver Mitochondria. *Biochim. Biophys. Acta BBA - Gen. Subj.* **1984**, *802*, 229–236, doi:10.1016/0304-4165(84)90166-1.

103. Cai, Z.; Yan, L.-J.; Li, K.; Quazi, S.H.; Zhao, B. Roles of AMP-Activated Protein Kinase in Alzheimer's Disease. *NeuroMolecular Med.* **2012**, *14*, 1–14, doi:10.1007/s12017-012-8173-2.

104. Ferguson, E.L.; Zimmerman, S.C.; Jiang, C.; Choi, M.; Meyers, T.J.; Hoffmann, T.J.; Gilsanz, P.; Oni-Orisan, A.; Wang, J.; Whitmer, R.A.; et al. Independent Associations of High-density Lipoprotein Cholesterol and Triglyceride Levels with Alzheimer's Disease and Related Dementias. *Alzheimers Dement.* **2025**, *21*, e14575, doi:10.1002/alz.14575.

105. Fan, L.-J.; Wang, F.-Y.; Zhao, J.-H.; Zhang, J.-J.; Li, Y.-A.; Tang, J.; Lin, T.; Wei, Q. From Physical Activity Patterns to Cognitive Status: Development and Validation of Novel Digital Biomarkers for Cognitive Assessment in Older Adults. *Int. J. Behav. Nutr. Phys. Act.* **2025**, *22*, 11, doi:10.1186/s12966-025-01706-x.

106. Jamshed, M.; Shahzad, A.; Riaz, F.; Kim, K. Exploring Inertial Sensor-Based Balance Biomarkers for Early Detection of Mild Cognitive Impairment. *Sci. Rep.* **2024**, *14*, 9829, doi:10.1038/s41598-024-59928-1.

107. Paolillo, E.W.; Saloner, R.; VandeBunte, A.; Lee, S.; Bennett, D.A.; Casaletto, K.B. Multimodal Lifestyle Engagement Patterns Support Cognitive Stability beyond Neuropathological Burden. *Alzheimers Res. Ther.* **2023**, *15*, 221, doi:10.1186/s13195-023-01365-9.

108. Pham, T.K.; Buczek, W.A.; Mead, R.J.; Shaw, P.J.; Collins, M.O. Proteomic Approaches to Study Cysteine Oxidation: Applications in Neurodegenerative Diseases. *Front. Mol. Neurosci.* **2021**, *14*, 678837, doi:10.3389/fnmol.2021.678837.

109. Alam, A.B.; Wu, A.; Power, M.C.; West, N.A.; Alonso, A. Associations of Serum Uric Acid with Incident Dementia and Cognitive Decline in the ARIC-NCS Cohort. *J. Neurol. Sci.* **2020**, *414*, 116866, doi:10.1016/j.jns.2020.116866.

110. Khan, A.A.; Quinn, T.J.; Hewitt, J.; Fan, Y.; Dawson, J. Serum Uric Acid Level and Association with Cognitive Impairment and Dementia: Systematic Review and Meta-Analysis. *AGE* **2016**, *38*, 16, doi:10.1007/s11357-016-9871-8.

111. Baumel, B.S.; Doraiswamy, P.M.; Sabbagh, M.; Wurtman, R. Potential Neuroregenerative and Neuroprotective Effects of Uridine/Choline-Enriched Multinutrient Dietary Intervention for Mild Cognitive Impairment: A Narrative Review. *Neurol. Ther.* **2021**, *10*, 43–60, doi:10.1007/s40120-020-00227-y.

112. Liu, P.; Yang, Q.; Yu, N.; Cao, Y.; Wang, X.; Wang, Z.; Qiu, W.-Y.; Ma, C. Phenylalanine Metabolism Is Dysregulated in Human Hippocampus with Alzheimer's Disease Related Pathological Changes. *J. Alzheimers Dis.* **2021**, *83*, 609–622, doi:10.3233/JAD-210461.

113. Chatterjee, P.; Cheong, Y.; Bhatnagar, A.; Goozee, K.; Wu, Y.; McKay, M.; Martins, I.J.; Lim, W.L.F.; Pedrini, S.; Tegg, M.; et al. Plasma Metabolites Associated with Biomarker Evidence of Neurodegeneration in Cognitively Normal Older Adults. *J. Neurochem.* **2021**, *159*, 389–402, doi:10.1111/jnc.15128.

114. Wang, L.; Wang, F.; Liu, J.; Zhang, Q.; Lei, P. Inverse Relationship between Baseline Serum Albumin Levels and Risk of Mild Cognitive Impairment in Elderly: A Seven-Year Retrospective Cohort Study. *Tohoku J. Exp. Med.* **2018**, *246*, 51–57, doi:10.1620/tjem.246.51.

115. Wolters, F.J.; Zonneveld, H.I.; Licher, S.; Cremers, L.G.M.; on behalf of the Heart Brain Connection Collaborative Research Group; Ikram, M.K.; Koudstaal, P.J.; Vernooij, M.W.; Ikram, M.A. Hemoglobin and Anemia in Relation to Dementia Risk and Accompanying Changes on Brain MRI. *Neurology* **2019**, *93*, doi:10.1212/WNL.0000000000008003.

116. He, X.-Y.; Yu, J.-T. Serum Clinical Laboratory Tests and Risk of Incident Dementia: A Prospective Cohort Study of 407,190 Individuals (P2-9.017). *Neurology* **2024**, *102*, 7013, doi:10.1212/WNL.0000000000208232.

117. Kim, J.W.; Byun, M.S.; Lee, J.H.; Yi, D.; Jeon, S.Y.; Sohn, B.K.; Lee, J.-Y.; Shin, S.A.; Kim, Y.K.; Kang, K.M.; et al. Serum Albumin and Beta-Amyloid Deposition in the Human Brain. *Neurology* **2020**, *95*, doi:10.1212/WNL.0000000000010005.

118. Karosas, T.; Wallace, T.C.; Li, M.; Pan, Y.; Agarwal, P.; Bennett, D.A.; Jacques, P.F.; Chung, M. Dietary Choline Intake and Risk of Alzheimer's Dementia in Older Adults. *J. Nutr.* **2025**, *155*, 2322–2332, doi:10.1016/j.tjnut.2025.05.015.

119. Aguree, S.; Zolnoori, M.; Atwood, T.P.; Owora, A. Association between Choline Supplementation and Alzheimer's Disease Risk: A Systematic Review Protocol. *Front. Aging Neurosci.* **2023**, *15*, 1242853, doi:10.3389/fnagi.2023.1242853.

120. Reiten, O.K.; Wilvang, M.A.; Mitchell, S.J.; Hu, Z.; Fang, E.F. Preclinical and Clinical Evidence of NAD+ Precursors in Health, Disease, and Ageing. *Mech. Ageing Dev.* **2021**, *199*, 111567, doi:10.1016/j.mad.2021.111567.

121. Chandran, S.; Binninger, D. Role of Oxidative Stress, Methionine Oxidation and Methionine Sulfoxide Reductases (MSR) in Alzheimer's Disease. *Antioxidants* **2023**, *13*, 21, doi:10.3390/antiox13010021.

122. Cristofano, A.; Sapere, N.; La Marca, G.; Angiolillo, A.; Vitale, M.; Corbi, G.; Scapagnini, G.; Intrieri, M.; Russo, C.; Corso, G.; et al. Serum Levels of Acyl-Carnitines along the Continuum from Normal to Alzheimer's Dementia. *PLOS ONE* **2016**, *11*, e0155694, doi:10.1371/journal.pone.0155694.

123. Ju, Y.H.; Bhalla, M.; Hyeon, S.J.; Oh, J.E.; Yoo, S.; Chae, U.; Kwon, J.; Koh, W.; Lim, J.; Park, Y.M.; et al. Astrocytic Urea Cycle Detoxifies Aβ-Derived Ammonia While Impairing Memory in Alzheimer's Disease. *Cell Metab.* **2022**, *34*, 1104-1120.e8, doi:10.1016/j.cmet.2022.05.011.

124. Fernandes, B.S.; Inam, M.E.; Enduru, N.; Quevedo, J.; Zhao, Z. The Kynurenine Pathway in Alzheimer's Disease: A Meta-Analysis of Central and Peripheral Levels. *Braz. J. Psychiatry* **2023**, doi:10.47626/1516-4446-2022-2962.

125. Xu, Y.; Zhao, M.; Han, Y.; Zhang, H. GABAergic Inhibitory Interneuron Deficits in Alzheimer's Disease: Implications for Treatment. *Front. Neurosci.* **2020**, *14*, 660, doi:10.3389/fnins.2020.00660.

126. Irino, Y.; Toh, R.; Nagao, M.; Mori, T.; Honjo, T.; Shinohara, M.; Tsuda, S.; Nakajima, H.; Satomi-Kobayashi, S.; Shinke, T.; et al. 2-Aminobutyric Acid Modulates Glutathione Homeostasis in the Myocardium. *Sci. Rep.* **2016**, *6*, 36749, doi:10.1038/srep36749.

127. Bonaparte, J.; Cook, M.; Dixon, S.; Wright, C.; Hill, W.; Wang, G. KYNURENINE METABOLITES INDUCE MICROGLIAL CELL SENESCENCE AND STIMULATE NEUROINFLAMMATION. *Innov. Aging* **2023**, *7*, 1073–1073, doi:10.1093/geroni/igad104.3447.

13 List of Training parameters:

Here is the list of training parameters to develop the Two Stage model except the categorical parameters such as Sex, Exposure Group.

Training Parameters	Count	Mean ± std
Height, cm	98	156.30 ± 9.07
Body weight, Kg	98	56.71 ± 10.42
AST	96	25.98 ± 7.46
ALT	98	20.12 ± 8.42
g-GTP	98	33.67 ± 37.62
CK	96	116.48 ± 59.57
T-BIL	98	0.96 ± 0.32
TP	98	7.68 ± 0.48
ALB	98	4.59 ± 0.29
GLU	98	102.79 ± 15.71
TG	98	104.94 ± 40.24
LDL-C	98	124.06 ± 30.24
HDL-C	98	73.40 ± 25.19
BUN	98	18.33 ± 5.88
UA	98	5.52 ± 1.36
CRE	98	0.89 ± 0.31
ALPif	98	77.52 ± 23.68
WBC	98	57.11 ± 12.06
RDW-SD	98	44.47 ± 2.89
RBC	98	440.98 ± 43.59
RDW-CV	98	12.91 ± 0.70
HGB	98	13.82 ± 1.42
PDW	98	12.92 ± 2.00
HCT	98	41.45 ± 3.99
MPV	98	10.49 ± 0.92
MCV	98	94.1 ± 3.87
P-LCR	98	29.41 ± 7.52
MCH	98	31.37 ± 1.71
PCT	98	0.22 ± 0.05
MCHC	98	33.33 ± 0.89
NRBC#	98	0.0031 ± 0.0173
PLT	98	20.98 ± 5.23
NRBC%	98	0.0061 ± 0.033
NEUT#	98	29.23 ± 7.62
NEUT%	98	51.31 ± 8.39
LYMPH#	98	20.93 ± 6.65
LYMPH%	98	36.48 ± 8.16
MONO#	98	4.80 ± 1.34

MONO%	98	8.49 ± 1.96
EO#	98	1.71 ± 1.29
EO%	98	2.96 ± 1.99
BASO#	98	0.45 ± 0.25
BASO%	98	0.78 ± 0.41
HbA1c	98	6.08 ± 0.51
L-Pro, nmol/mL(uM)	98	148.92 ± 54.31
D-Pro, nmol/mL(uM)	98	1.73 ± 1.50
Pro(D/(D+L)*100)	98	1.20 ± 0.99
L-Ser, nmol/mL(uM)	98	112.79 ± 20.77
D-Ser, nmol/mL(uM)	98	2.64 ± 0.87
Ser(D/(D+L)*100)	98	2.32 ± 0.79
L-Ala, nmol/mL(uM)	98	358.92 ± 73.29
D-Ala, nmol/mL(uM)	98	3.01 ± 2.26
Ala(D/(D+L)*100)	98	0.85 ± 0.64
L-Asn, nmol/mL(μM)	98	52.45 ± 8.26
D-Asn, nmol/mL(μM)	98	0.23± 0.15
Asn(D/(D+L)*100)	98	0.44 ± 0.29
L-Lys, nmol/mL(μM)	98	208.67 ± 29.87
L-Arg, nmol/mL（μM）	98	97.25 ± 18.39
L-His, nmol/mL(μM)	98	92.14 ± 11.26
L-Tyr, nmol/mL(μM)	98	59.47 ± 11.94
L-Gln, nmol/mL(μM)	98	593.40 ± 87.76
L-Ile, nmol/mL(μM)	98	49.06 ± 12.19
L-Thr, nmol/mL(μM)	98	116.55 ± 21.22
L-Orn, nmol/mL(μM)	98	59.86 ± 14.77
L-Val, nmol/mL(μM)	98	202.87 ± 40.05
L-Phe, nmol/mL(μM)	98	54.69 ± 9.61
L-Met, nmol/mL（μM）	98	25.85 ± 4.34
L-Glu, nmol/mL(μM)	98	36.85 ± 15.14
L-Asp, nmol/mL(μM)	98	3.60 ± 1.41
L-Leu, nmol/mL(μM)	98	107.20 ± 21.62
L-Cit, nmol/mL（μM）	98	39.88 ± 9.81
L-Trp, nmol/mL(μM)	98	56.62 ± 10.58
Gly, nmol/mL(μM)	98	285.01 ± 66.54
BMI	98	23.12 ± 3.15
Height(Reported)	94	157.51 ± 9.00
Weight(Reported)	94	56.95 ± 10.18
Last three day average after 1.5 or less(every minute)	83	302.74 ± 144.72
1MET or higher	96	572.78 ± 270.61
2MET or higher	96	203.40 ± 107.60
3MET or higher	96	63.62 ± 43.37
4MET or higher	96	16.26 ± 18.20
1METs～ 2 METs	83	427.23 ± 143.57

2METs～3 METs	83	161.67 ± 51.74
3METs～4 METs	83	54.77 ± 26.73
4METs～5 METs	83	14.21 ± 14.01
5METs～6 METs	83	3.22 ± 3.67
6METs～7 METs	83	1.01 ± 2.85
7METs～8 METs	83	0.22 ± 0.39
8METs or higher	83	0.15 ± 0.29
Total walking exercise(Ex)	83	1.70 ± 1.78
Activity(Ex)	83	2.83 ± 1.45
EX	83	4.52 ± 2.47
Number of steps	83	5453.16 ± 3353.64
Walking time (Minutes)	83	84.72 ± 43.74
Number of steps per minute	83	63.43 ± 16.95
EX fluctuation coefficient 1	83	0.35 ± 0.12
EX fluctuation coefficient 2	83	0.33 ± 0.21
EX fluctuation coefficient 3	83	4.37 ± 36.79
EX fluctuation coefficient 4	83	0.39 ± 0.23
EX fluctuation coefficient 5	83	0.36 ± 0.22
1METs or more fluctuation coefficient 5	79	2.75 ± 22.29
2METs or more fluctuation coefficient 5	83	1.44 ± 10.73
3METs or more fluctuation coefficient 5	83	0.73 ± 3.62
4METs or more fluctuation coefficient 5	82	0.63 ± 1.12
Shannon index p1	98	4.87 ± 0.65
Simpson index p1	98	0.91 ± 0.05
2-Aminobutyric acid	98	4.56 ± 1.32
4-Aminobutyric acid	98	0.03 ± 0.01
4-Hydroxyproline	98	3.62 ± 1.69
Acetylcarnitine	98	68.28 ± 22.46
Adenosine monophosphate	98	0.37 ± 0.24
Alanine	98	51.08 ± 10.69
Arginine	98	62.16 ± 17.28
Asparagine	98	3.33 ± 0.58
Aspartic acid	98	0.47 ± 0.15
Asymmetric dimethylarginine	98	0.28 ± 0.08
Carnitine	98	50.62 ± 11.67
Choline	98	7.34 ± 1.68
Citrulline	98	20.86 ± 6.72
Creatine	98	27.42 ± 10.39
Creatinine	98	62.70 ± 14.99
Cysteine	98	0.21 ± 0.07
Cystine	98	12.25 ± 4.12

Dimethylglycine	98	1.99 ± 0.85
Ergothioneine	98	1.38 ± 1.09
Glutamic acid	98	9.64 ± 3.38
Glutamine	98	66.14 ± 12.95
Glycine	98	4.65 ± 1.32
Glycolic acid	98	0.17 ± 0.04
Glyoxylic acid	98	0.20 ± 0.03
Histidine	98	40.58 ± 8.24
Homocysteine	98	0.05 ± 0.03
Hypoxanthine	98	0.55 ± 0.38
Isoleucine	98	13.22 ± 2.94
Kynurenine	98	2.18 ± 0.59
Lactic acid	98	5.40 ± 2.59
Leucine	98	17.86 ± 3.50
Lysine	98	77.00 ± 21.68
Methionine	98	10.79 ± 2.21
Methionine sulfoxide	98	0.22 ± 0.10
Niacinamide	98	0.45 ± 0.18
Ornithine	98	14.41 ± 4.80
Phenylalanine	98	327.78 ± 63.02
Proline	98	4.17 ± 1.23
Pyruvic acid	98	0.36 ± 0.08
Serine	98	14.14 ± 2.57
Serotonin	98	0.47 ± 0.35
Symmetric dimethylarginine	98	1.22 ± 0.43
Threonine	98	12.68 ± 2.65
Tryptophan	98	80.46 ± 15.42
Tyrosine	98	37.07 ± 7.71
Uracil	98	0.22 ± 0.07
Uric acid	98	1.83 ± 0.42
Uridine	98	1.25 ± 0.33
Valine	98	54.37 ± 8.66
Age	98	77.71 ± 2.99