INTERPRETABLE MACHINE LEARNING IN ALZHEIMER'S DISEASE DEMENTIA

INTERPRETABLE MACHINE LEARNING IN PATIENTS WITH ALZHEIMER'S DISEASE DEMENTIA

By MASON KADEM, HBSc, MSc

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Applied Science

McMaster University

MASTER OF APPLIED SCIENCE (2023)

Hamilton, Ontario, Canada (Biomedical Engineering)

TITLE:	Interpretable Machine Learning in Patients with Alzheimer’s Disease Dementia
AUTHOR:	Mason Kadem H.B.Sc., M.Sc. (Neuroscience), Western University, London, Canada
SUPERVISORS:	Dr. Michael Noseworthy and Dr. Thomas Doyle
NUMBER OF PAGES:	xvii, 127

Lay Abstract

Early identification of patients at the highest risk of Alzheimer's disease (AD) is crucial for possible pharmaceutical intervention. Existing prediction models have limitations, including inaccessible data and lack of interpretability. This research used a machine learning approach to identify patients at the highest risk of Alzheimer's disease and found that certain clinical features, such as specific executive function-related cognitive testing (i.e., task switching), combined with genetic predisposition, brain imaging, and demographics, were important contributors to AD risk. The models were able to reliably predict patient diagnosis and prognosis and were designed to be low-cost, non-invasive, clinically operable and easily accessible. The interpretable models provided an intuitive explanation of the decision process, making it a valuable tool for healthcare decision-making and planning.

Abstract

Alzheimer's disease (AD) is among the top 10 causes of global mortality, and dementia imposes a yearly $1 trillion USD economic burden. Of particular importance, women and minoritized groups are disproportionately affected by AD, with females having higher risk of developing AD compared to male cohorts. Differentiating mild cognitive impairment ($\text{MCI}_{\text{stable}}$) from early stage Alzheimer's disease (MCI_{AD}) is vital worldwide. Despite genetic markers, such as apo-lipoprotein-E (APOE), identification of patients before they develop early stages of MCI_{AD}, a critical period for possible pharmaceutical intervention, is not yet possible. Based on review of the literature three key limitations in existing AD-specific prediction models are apparent: 1) models developed by traditional statistics which overlook nonlinear relationships and complex interactions between features, 2) machine learning models are based on difficult to acquire, occasionally invasive, manually selected, and costly data, and 3) machine learning models often lack interpretability. Rapid, accurate, low-cost, easily accessible, non-invasive, interpretable and early clinical evaluation of AD is critical if an intervention is to have any hope at success. To support healthcare decision-making and planning, and potentially reduce the burden of AD, this research leverages the Alzheimer's Disease Neuroimaging Initiative (ADNI1/GO/2/3) database and a mathematical modelling approach based on supervised machine learning to identify 1) predictive markers of AD, and 2) patients at the highest risk of AD. Specifically, we implemented a supervised XGBoost classifier with diagnostic (Exp 1) and prognostic (Exp 2) objectives. In experiment 1 (n=441) classification of AD (n=72) was performed in comparison to healthy controls (n= 369), while experiment 2 (n=738) involved classification of MCI$_{\text{stable}}$ (n = 444) compared to MCI$_{\text{AD}}$(n = 294). In Experiment 1, machine learning tools identified three features (i.e., Everyday Cognition Questionnaire (Study partner) - Total, Alzheimers Disease Assessment Scale (13 items) and Delayed Total Recall) with ROC AUC scores consistently above 97%. Low performance on delayed recall alone appears to distinguish most AD patients. This finding is consistent with the pathophysiology of AD with individuals having problems storing new information into long-term memory. In experiment 2, the algorithm identified the major indicators of MCI-to-AD progression by integrating genetic, cognitive assessment, demographic and brain imaging to achieve ROC AUC scores consistently above 87%. This speaks to the multi-faceted nature of MCI progression and the utility of of comprehensive feature selection. These features are important because they are non-invasive and easily collected. As an important focus of this research, the interpretability of the ML models and their predictions were investigated. The interpretable model for both experiments maintained performance with their complex counterparts while improving their interpretability. The interpretable model provides an intuitive explanation of the decision process which are vital steps towards the clinical adoption of machine learning tools for AD evaluation. The models can reliably predict patient diagnosis (Exp 1) and prognosis (Exp 2). In summary, our work extends beyond the identification of high-risk factors for developing AD. We identified accessible clinical features, together with clinically operable decision routes, to reliably and rapidly predict patients at the highest risk of developing Alzheimer's disease. We addressed the aforementioned limitations by providing an intuitive explanation of the decision process among the high-risk noninvasive and accessible clinical features that lead to the patient's risk.

To my family ***B****,* ***M****,* ***E****,* ***J****, and to my mentors, past and present, who make me realize that it is not solely the advancement of technology or methods that will propel humanity forward, but rather it's the mentors themselves who inspire us to dream and achieve beyond our perceived limitations*

Acknowledgements

I extend my deepest gratitude to my wonderful supervisors, Dr. Thomas Doyle and Dr. Michael Noseworthy, for their invaluable guidance, mentorship, and unwavering support throughout the journey of this thesis. Their expertise and encouragement have been instrumental in making this project a success, and their willingness to take me on has been a true honor. I am also grateful to Dr. Dinesh Kumbhare for serving on my examination committee and providing valuable feedback and insights.

To my lab mates, past and present, I want to express my sincere thanks for your invaluable discussions. Your camaraderie, support, and valuable insights have been a source of inspiration and motivation throughout this journey, and I feel fortunate to have had the opportunity to work alongside such a talented and supportive group of individuals.

Table of Contents

List of Figures

List of Tables

Declaration of Academic Achievement

This thesis, by Mason Kadem, acknowledges contributions from Dr. Thomas E. Doyle and Dr. Michael Noseworthy. Mason Kadem led the study design, data preprocessing, analysis, experiments, and manuscript writing. Drs. Thomas E. Doyle and Michael Noseworthy contributed to the inception of the study, study design, provided resources, study design input, manuscript review, and guidance at all stages of the thesis. The thesis is original and contains no previously submitted or published material for any other degrees.

Chapter 1

Introduction

Alzheimer's disease (AD) is among the top 10 causes of global mortality, and dementia imposes a yearly $1 trillion USD economic burden [1]. Nearly 80% of dementia cases include adults living with AD [2]. Cognitive impairment manifests heterogeneously across patients with AD [3], and the differences at the individual level are a product of the various disease risk factors and progressive nature of the disease that consequently compromises brain integrity and function [4]. Improving our understanding of risk profiles and having the ability to predict patients at the highest risk of AD would have significant impacts on relieving global disease burden [1]. Well-establisehd early markers of AD exist, including genetic markers like apolipoprotein E (APOE) [5]; however, even with these early markers, we have yet to identify patients at the highest risk of developing AD [6]. Limitations in the capacity to conduct accurate risk prediction are highlighted by three main aspects of existing risk prediction models: 1) models developed by traditional statistics which overlook nonlinear relationships and complex interactions between features, 2) machine learning models are based on difficult to acquire, occasionally invasive, manually selected, and costly data, and 3) machine learning models often lack interpretability. Taken together, we are lacking accurate, feasible, understandable, and economical risk prediction models. The research herein leverages the Alzheimer's Disease Neuroimaging Initiative database and a mathematical modelling approach based on supervised machine learning to identify 1) predictive markers of AD, and 2) patients at the highest risk of AD. Furthermore, the thesis focuses on the interpretability of the machine learning models and their predictions to improve the clinical uptake of AD-specific machine learning models.

The follow sections introduce the pathophysiology of AD, thesis-specific machine learning overview and fundamentals of model development, interpretability, and limitations.

1.1 Pathophysiology of Alzheimer's disease

Half a million Canadians live with dementia and this number is expected to triple by the year 2050 [7]. Alzheimer's disease (AD) is the most prevalent type of dementia and some individuals with mild cognitive impairment eventually progress to AD. Age is the predominant risk factor for AD, and specifically adults over the age of 65 years are at heightened risk of AD onset. Several adaptations accompany aging and include (but are not limited to) changes in brain structure (i.e., build-up of beta amyloid plaques, and cortical thinning) [8], inflammatory profile (i.e., "inflammaging") [9], epigenetic dysregulation [10] and vascular dysfunction [11, 12]. Neuropsychiatric disorders like AD are complex, however, and several studies emphasize the importance of epigenetics in understanding AD etiology [13–15]. Epigenetic modifications (e.g., DNA methylation) occur with normal aging and disease, and specific AD epigenetic signatures [16] can impact memory and learning and ultimately affect cognition in

AD [17].

Despite AD presenting as overt changes in cognitive abilities, AD risk is multifactorial in nature, and risk factors include demographics, lifestyle and environmental factors, and genetic predisposition. Understanding risk factors in AD help with targeted interventions and also provide insight in identifying biomarkers that relate to risk of AD onset. One of the most widely established early biomarkers of AD is sign of neurodegeneration [18–20], or loss of brain cortical tissue (measured in volume or thickness) [21]. While normal aging presents with brain atrophy, AD presents with abnormal level of and differential regional brain atrophy [22, 23], with atrophy in the medial perirhinal cortex and entorhinal cortex seen in very early AD [24]. A recent study assessing progression of brain structural changes with AD over decades and how the trajectories differ from normal brain aging, found that the amygdala and hippocampus were the most severely impacted areas in AD [25]. Prior to conducting brain imaging tests, neuropsychological assessments of cognitive impairment (e.g., mini mental state examination (MMSE) [26] and the Montreal Cognitive Assessment (MoCA) [27]) may help identify cognitive impairment in AD, but it remains unknown whether changes in these cognitive scores and other biomarkers can be used as early detection of AD.

Although age is a prevailing risk factor for AD, the host of AD risk factors should be contextualized within ethnicity/race, level of education, and socioeconomic status. “Premature aging” can occur in individuals from low socioeconomic classes, and ultimately influencing higher risk of age-associated diseases[28] in these individuals. The notion of “brain resilience” [29] was higher in adults with higher brain intracranial volume, even if they had lower education levels [30]. Additionally, the way in which questions are asked in neuropsychological tests may influence AD assessment in individuals across different cultures [31]. Future work focusing on AD risk factors across members from low socio-economic status, racialized and other minorities groups, will help provide necessary insight on heterogeneity in disease risk profiles and potential targeted therapeutics in an aging population that is becoming more diverse [32].

1.2 Overview of Supervised machine learning

Supervised machine learning models are empirically derived mathematical equations $f(X)$, relating outcome Y and input X, that provide accurate individualized predictions y_i. Finding the optimal $f(X)$ goes beyond error minimization and requires the understanding of fundamental concepts such as the bias-variance trade off, generalizability, interpretability, feature selection, collinearity, and appropriate optimization and evaluation, which are further discussed in the literature review. Machine learning-based approaches can explore complex and nonlinear interactions among clinical features and find predictive biomarkers unknown to domain experts. We briefly formalize the supervised machine learning setup below.

Training data to the learner, denoted D, are provided as pairs of inputs (x_1, y_1) up to (x_n,y_n), where x is the input, or feature vector (e.g., considering clinical data, x_i^1 can refer to patient i's age in years, or blood biomarkers in pg/ml), and is a member of d-dimensional feature space R^d, y its label, and B the label space (e.g., 0,1 for a binary classification task detecting Normal [0] or disease [1]).

$$D = \ (x_1, y_1), \ldots, (x_n, y_n) \subseteq R^d \quad \times \ B, \ i.i.d \tag{1.2.1}$$

The objective of supervised machine learning is to find a predicting function or hypothesis h

$$h : R^d \to \ B, \ s.t. \ h(x_i \ ) \approx y_i \ \forall (x_i, y_i \ ) \in D_{train}; h(x_i \ ) \approx y_i \ \forall (x_i, y_i \ ) \notin D_{test} \quad (1.2.2)$$

To quantify a consistent model, we introduce a loss function, with the assumptions that h is determined with respect to an unknown data generating distribution Z, and true labelling function f.

$$L_{Z,f}(h) = \mathbb{P}_{x \sim Z}[h\,(x) \neq f(x)] \quad (1.2.3)$$

Given that the DD and f are unknown, we strategically pick (x,y) from our empirical distribution.

$$L_S(h) = \mathbb{P}_{(x,y) \sim D}[h\,(x) \neq y] \quad (1.2.4)$$

Given a training sample, the classifier evaluates the error of each $h \in H$, where H is our hypothesis space (set of all possible classifiers), and outputs a member of H that minimizes the loss, with the hope that h minimizes the empirical training loss with respect to our sample and the true data probability distribution as well, based on the weak law of large numbers.

$$h = \text{argmin}_{h \in \mathcal{H}} \frac{1}{|D_{\text{train}}|} \sum_{(\mathbf{x},y) \in D_{\text{train}}} \ell(\mathbf{x}, y|h) \quad (1.2.5)$$

$$\epsilon_{\text{test}} = \frac{1}{|D_{test}|} \sum_{(\mathbf{x},y)\in D_{\text{test}}} \ell(\mathbf{x}, y|h). \tag{1.2.6}$$

$$D_{\text{test}} \to +\infty \tag{1.2.7}$$

$$\epsilon_{\text{test}} \to \epsilon \tag{1.2.8}$$

1.2.1 Traditional Statistics vs Machine Learning

The relative efficacy of various machine learning and statistical approaches for disease diagnosis/prognosis remains an important discourse. A comparison of traditional statistical vs machine learning is provided, including model and algorithmic approaches, feature selection, performance metrics, interpretability, limitations and data requirements (Table 1.1). To summarize, traditional statistics focus on inference, fitting data-specific probability models through analytical solutions (e.g., linear regression), rely solely on domain expertise for selection of salient features, with metrics focused on accuracy; yet findings from traditional statistics approaches are easily interpreted, and perform well with smaller sample sizes under verified test assumptions. Traditional statistics also provide options to quantify the effect of interest and confidence in that effect, independent of noise or chance. Traditional statistics developed to capture linear relationships, can be adapted to handle nonlinear relationships but are limited in their capacity to handle complex non-linear relationships between features, especially many features and high dimensional data. The inter-predictor nonlinear relationships and complex interactions are consistent factors to consider with high dimensional data and traditional statistics approaches do not have capacity to adequately account for these factors.

In contrast, machine learning focuses on prediction, uses a general learning procedure that relies on the empirical capacity of the model and optimization techniques to improve an objective function, utilizes features extraction/selection tools to acquire relevant features, independent of domain expertise, is more difficult to interpret, and relies on larger data for algorithm development and optimization. Moreover, machine learning algorithms are specifically designed to handle non-linear relationships among many features and capture complex interactions in high-dimensional data even when traditional statistics models fail. While methods at times fall into a machine learning or statistical domain, more often they overlap (e.g., bootstrapping). The table below provides a comparison between traditional statistics and machine learning approaches, with the recognition that there is a significant overlap between the two fields.

Table 1.1: Traditional statistics vs Machine learning

Approach	**Traditional statistics**	**Machine learning**
Focus	Inference	Prediction
Models	Data-specific	Empirical capacity
Algorithms	Analytical solutions	Numerical optimization
Relevant Features	Domain expertise	Feature selection
Metrics	Accuracy	Varied
Interpretation	Easy	Difficult
Data	Smaller	Larger
Non-linear relationships	Limited	Yes
High dimensional data	Limited	Yes
Complex interactions	Limited	Yes

While modifying risk factors may lower dementia risk or delay onset [33], many AD-specific prediction models are based on more traditional statistical approaches [34–36] which have presented with limitations when interpreting patient specific risk factors as they can overlook complex interactions and nonlinear relationships among clinical features [37] [38–40, 33]. Advances in the performance of computational diagnostic models have been explored via the use of increasingly advanced machine learning methods. Machine learning tools can efficiently combine multiple sources of data, explore nonlinear relationships and complex interactions between features, and surpass traditional statistical approaches for diagnostic and prognostic prediction.

1.2.2 Features and their usage

Machine learning demands many observations, but not features (i.e., independent variables, covariates, predictors). Traditional statistics relies heavily on domain experts to select relevant features, machine learning does not. Feature selection reduces dimensionality by selecting relevant and removing irrelevant features, noise, redundancy, and collinearity. There are a plethora of feature selection techniques. Herein, we utilize the average feature importance according to our model, based on each feature's split-induced gain. The gain score estimates the contribution of each feature to the model's performance. The higher the gain score, the more important the feature is in predicting the target variable. Compared to other feature selection methods, the gain score considers both the presence of the feature in the model and its split quality. This makes it a more robust measure of feature importance than other metrics that just evaluate model feature frequency. However, most AD-specific machine learning models are based on manually selected, difficult to acquire, costly [41–44],, and invasive measures (e.g., cerebrospinal fluid analysis of β-amyloid (Aβ42) [34], Aβ-positron emission tomography [35, 36] which limits the availability of these biomarkers and their usage.

1.2.3 Interpretable machine learning

Despite the promise for machine learning to aide in expediting disease prognosis/diagnosis, clinical and translational benefits of machine learning are currently restricted by complex models that involve high processing costs, data/population heterogeneity, and generally lack interpretability which limits their clinical adoption. The complexity of high performing machine learning algorithm hinders the formation of clinically intuitive explanations of the decision process, thus impeding clinical adoption. Although explainability and interpretability are used interchangeably with machine learning research domain, herein, we differentiate the two terms. Specifically, in this thesis, explainability will refer to providing accessible explanations to end-users for complex models, whereas interpretable models will be used to indicate models that are understandable by design (e.g., how model processes information; decision trees, ML-based linear models). Together, these terms serve as crucial criteria for clinical adoption and are fundamental to data security and fairness in machine learning.

The motivation for interpretable machine learning can be illustrated with the following example: A patient may be diagnosed with cognitive impairment. The memory clinic may have a prognostic machine learning algorithm that helps identify patients at the highest risk for developing Alzhemier's disease dementia, based on the patient's clinical state (e.g., cognitive testing, genetics, imaging, demographics). The patient may then inquire about why and how the algorithm made such a prediction. When using interpretable machine learning, a prediction is generated, but an understanding of the machine learning and decision process also accompanies the prediction.

While it's difficult to understand the inner mechanics of complex machine learning models, there is a need to discover relationships between input and output data; however, complex machine learning applications to clinical problems remain challenging due to their lack of interpretability. There exists room for improvement in explaining the decision process in machine learning algorithms while maintaining performance of complex models. For example, the use of simpler models with fewer features can provide better interpretability (e.g., regularization, simple architectures). As alluded to above, feature importance can help understand the important features the model is using to make predictions. Some algorithms like decision trees are easily visualized, and decision trees can help identify important features but also their absolute thresholds and decision path that led to the final prediction. Further, gradient boosted algorithms use recursive decision trees, and offer great interpretability relative to black box approaches (e.g., neural networks). Using simulated data, Lundberg et al., (2020) varied the nonlinearity the data, and showed that an increase in nonlinearity is associated with a decrease in the accuracy and interpretability of machine learning based logistic regression models due to a mismatch between the model and data, resulting in an increase in % of weight attributed to irrelevant features, thus reducing interpretability overall [45]. To this end, non-linear models like the extreme gradient boosted ensembles can perform better while also remaining interpretable compared to machine learning models based on logistic regression, as the former places more emphasis on the input features and is perhaps a better representation of the data-generation process.

Finally, the clinical adoption and end-user trust of machine learning models requires intuitive explanation of the decision processes [46]; yet, complex models are abundant and interpretable risk prediction models for disease diagnosis and prognosis are limited [37, 47, 42, 48, 45], with a dire need for interpretable risk prediction machine learning models in AD [37, 47, 42]. Thus, rapid, accurate, low-cost, easily accessible, non-invasive, interpretable and early clinical evaluation of AD is critical at this time.

1.3 Gap Analysis Summary

Based on our search strategy (Fig. 2.1), we found three main gaps in current AD-specific risk prediction models: 1) models are based on traditional statistics model that miss nonlinear relationships and complex interactions between features, 2) while machine learning has been used to predict patients who may develop AD, the machine learning-based predictive models have included costly, manually selected, and invasive measures (e.g., cerebrospinal fluid analysis of β-amyloid (Aβ42) [34], Aβ-positron emission tomography [35, 36] 3)identifying high-risk indicators is useful for risk assessment, but determining threshold cut-off values for these factors are currently unknown which would offer clinical utility by defining informed decision routes (i.e., threshold values to distinguish high-risk from low-risk people), and would further aid in interpretability with understandable decision process and model architecture. Further, high-performing machine learning algorithms are complicated, making clinical adoption difficult. Improving machine learning interpretability while maintaining performance of complex models is challenging. Clinical adoption and end-user trust of machine learning models demand understandable explanation of decision processes [46]. Interpretable risk prediction models for AD diagnosis must therefore balance performance and interpretability. Taken together, we are lacking accurate, feasible, understandable, and economical risk prediction models that rely on easily accessible, non-invasive, and low-cost features.

1.4 Objectives and Thesis Question

To support healthcare decision-making and planning, answer the question: "Who is at the highest risk of Alzeheimer's Disease Dementia?" and potentially reduce the burden of AD, this thesis leverages the Alzheimer's Disease Neuroimaging Initiative database and a mathematical modelling approach based on supervised machine learning to go beyond identifying 1) high risk factors of AD and 2) patients at the highest risk of AD. The approaches herein provide an intuitive explanation of the key non-invasive features and decision processes which are vital steps towards the clinical adoption of machine learning tools for AD evaluation.

1.5 Evaluation overview

Data for model development included non-invasive biomarkers, imaging, genetic, cognitive testing, lifestyle and health history tabular data . Experiment 1 (n=441) classified controls (n= 369) vs AD (n = 72), while Experiment 2 (n=66) classified MCI_stable (n = 41) vs MCI_AD (n = 25) using an ADitional independent test set (n=43) with MCI_stable (n = 26) vs MCI_AD (n = 17). Each experiment had two models (complex and interpretable). For both experiments, we ranked the top 10 features according to their average feature importance accumulated from the XGBoost model, using random splits in a ratio of 7:3, over 100 repetitions with varied number seeds (0-99). The complex model performances, for both experiments, were evaluated with fivefold stratified cross-validation for 100 iterations. The interpretable models were created using a single 7:3 split and evaluated on the validation set (Exp 1; ADNI

3) and an independent test set (Exp 2; ADNI 2).

1.6 Thesis organization

The thesis herein is organized as follows:

Chapter 2 presents the pathophysiology of AD, followed by an overview of supervised machine learning, with a focus on methods used in the thesis. Finally, we highlight the fundamentals for model development, interpretability and current limitations in machine learning specific to the thesis.

Chapter 3 describe the data sources.

Chapter 4 presents the methodology for the machine learning process, application, and evaluation.

Chapter 5 is the results section.

Chapter 6 discusses and interprets the results.

Chapter 7 concludes the thesis and suggests next steps.

Chapter 2

Literature Review

The literature review will provide a brief overview of machine learning concepts followed by an overview of Alzheimer's disease dementia and its pathophysiology. Machine learning concepts are discussed that readers need to be familiar with to understand and appreciate the machine learning tools within the thesis. Particularly focused on machine learning in risk prediction, methods to mitigate overfitting, pitfalls, interpretability, generalizability and recommendations for data acquisition methods that improve model efficacy and reliability. The first three main parts of the literature review focus on 1) familiarizing the reader with key concepts in machine learning approaches, 2) highlighting common pitfalls and potential solutions when developing or evaluating these models, and 3) arguing for the need to develop more interpretable models.

2.1 Search strategy and data extraction

To assess the landscape of existing AD prediction models, we searched PubMed for "(Alzheimer's Disease Dementia) AND (prediction OR predict) AND (artificial intelligence OR machine learning) published between Jan. 1, 2012 and Nov.1, 2022. We evaluated only peer-reviewed English papers.

The considered papers were evaluated by the types of clinical features that were used, the approach (i.e., traditional statistics, ML, DL), and whether the models were interpretable. AI-assisted pipelines find relevant and filter irrelevant literature, by actively learning from the reviewers' judgements and reordering papers intelligently [49]. This approach cuts review time by 95% (Fig. 2.2). The machine learning model for active learning utilized Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction (i.e., determines word relevance based on word frequency) and a random forest classifier.

2.2 Pathophysiology of Alzheimer's Disease

Alzheimer's disease (AD) is the most prevalent type of dementia, with nearly half a million Canadians living with AD and this number is expected to nearly triple by the year 2050 [7]. Some individuals with mild cognitive impairment will progress to AD, and neuropsychological assessments of cognitive impairment, such as the mini mental state examination (MMSE) [26] and the Montreal Cognitive Assessment (MoCA) [27], have been proposed as a means to provide cut-off thresholds for identifying who is at risk of developing AD in adults with baseline mild cognitive impairment. From a clinical standpoint, use of MMSE on its own may be insensitive and insufficient

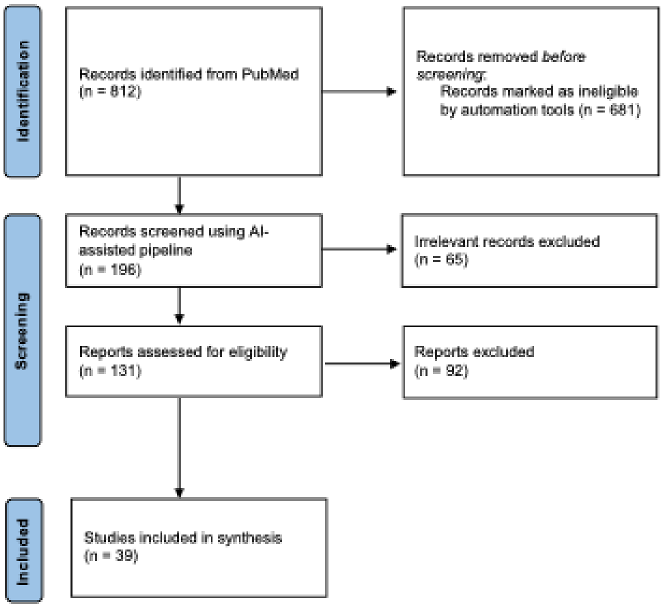

Identification
Records identified from PubMed
(n = 812)
Records removed *before screening*:
Records marked as ineligible by automation tools (n = 681)
Screening
Records screened using AI-assisted pipeline
(n = 196)
Irrelevant records excluded
(n = 65)
Reports assessed for eligibility
(n = 131)
Reports excluded
(n = 92)
Included
Studies included in synthesis
(n = 39)

Figure 2.1: PRISMA

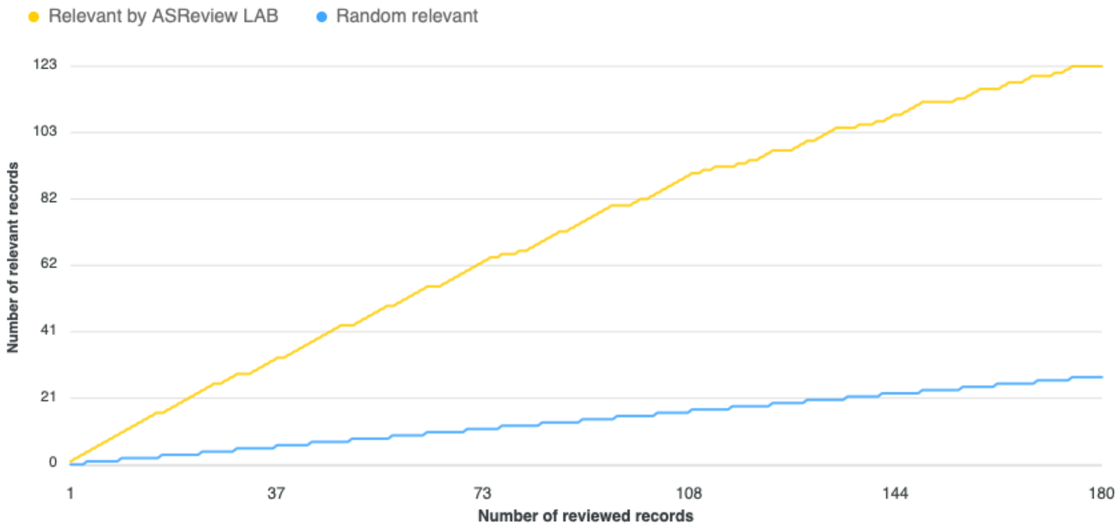


Figure 2.2: ASReview recall progress

at detecting subtle cognitive changes in the adults with mild cognitive impairment; however, it may be an informative means to track cognitive changes over time when combining MMSE results with other comprehensive cognitive assessments [50]. What remains unknown is whether changes in MMSE scores over time (rather than isolated baseline MMSE scores) can be combined with other cognitive metrics and/or biomarkers to improve early detection of AD.

Despite AD presenting as overt changes in cognitive abilities, it is important to recognize that AD risk is multi-factorial in nature, spanning different domains primarily including demographics, lifestyle and environmental factors, and genetic predisposition. More typical features of AD include cognitive impairment and changes in brain structure and function. AD presents in older age and specifically over the age of 65, with age being the most important risk factor for AD onset. Anatomical hallmarks of AD include β -amyloid plaques and tau protein aggregates, which accumulate with age [8]. Other major age-associated processes that are implicated in AD include inflammation, and epigenetic dysregulation [10]. Vascular dysfunction plays a significant role in AD progression, and extensive narratives on the importance of putative roles played by the vascular system in AD pathology are found in excellent reviews [11, 12].

Inflammation alongside aging has been coined as “inflammaging” and refers to low-grade systemic inflammation (without overt infection) [9], and associated with cortical thinning [51]. A recent analysis conducted on Swedish BioFINDER (Biomarkers For Identifying Neurodegenerative Disorders Early and Reliably) study showed that inflammatory profile changes in the cerebrospinal fluid correlate with amyloid and tau proteins, neurodegeneration and cognition in adults with AD [52]. A growing research area is use of anti-inflammatory interventions to target inflammatory biomarkers in efforts to mitigate negative impacts of heightened inflammation and the inflammatory profile within AD [10].

Two main variants of AD exist, which include a familial (i.e., genetic) early onset phenotype which include mutations in the amyloid precursor protein production. The second phenotype is the most common form of AD which is a sporadic late-onset variant (explaining over 95% of AD cases) where the apolipoprotein epsilon 4 gene (APOE4) gene seems to be the present as the most established genetic risk factor [53]. Recent genome-wide association studies (GWAS) have shown over 20 non-APOE-related loci with significant association in risk of AD onset [54, 55], which have been involved in the metabolic pathways (e.g., cholesterol and amyloid precursor protein), tau protein dysfunction, and protein trafficking. However, neuropsychiatric disorders like AD are complex, and several studies have implicated the importance of epigenetics in understanding AD etiology [13–15]. Epigenetic modifications occur with normal aging as well as in disease conditions and include histone and DNA modification (i.e., methylation). Epigenome-wide analysis studies have shown that AD has specific epigenetic signatures [16], which impact memory consolidation and learning processes and can lead to cognitive decline and AD [17]. Development of epigenetic drugs for AD prevention is an important advancement as these drugs could simultaneously regulate expression of several genes responsible for neuronal integrity; however further research is required to ensure gene expression does not lead to other harmful effects [10].

Although age imparts significant impact on AD development, it is imperative to contextualize the umbrella of AD risk factors within ethnicity/race, level of education, and socioeconomic status. Health disparities across minoritized and individuals from low socioeconomic classes, intersections between biological, socioeconomic and environmental factors present as "premature aging" and ultimately higher risk of age-associated diseases[28]. Analysis on the C-path online data repository determined that ethnicity may differentially affect how and when AD manifests across ethnicities, including how cognition changes with time, existence of co-morbidities, APOE4 gene status, and when AD presents itself over one's life course. Early onset AD is more common in individuals from African American, Alaskan, and Hawaiian ethnicities compared to Whites [56], and African American's have higher prevalence of APOE4 allele than other ethnicities [57]. Moreover, ethnoracial factors are important to consider in the context of biological factors. For example, a smaller change in cerebrospinal fluid tau protein markers and similar changes in white matter hyperintensities are associated with greater cognitive changes in African Americans (compared to Whites), which suggest that threshold cutoffs for these established biomarkers could lead to underdiagnosis of AD in African Americans [58]. Neuropsychological tests used to characterize cognitive function in AD and other dementia may be influenced by culture, including how questions are asked and in what language, and familiarity with test content [31]. The idea that higher education contributes to "cognitive reserve" is a working explanation for why higher education relates to delayed AD onset [59]. Higher education has been operationalized as "years of education", which is associated with lower AD diagnosis as shown by a Mendelian randomization study highlighting that genetic predisposition towards pursuing more years of education is associated with lower odds of AD [60]. However, educational attainment (as measured by years of education) was not associated with higher memory performance in Black patients with AD, but it did relate to improved memory performance in White patients with AD[61]. Taken together, the roles played by these factors in influencing AD outcome are important to understand against the backdrop of race and ethnicity; unfortunately, minoritized groups remain under-represented in research studies. Future work focusing on intersections between AD risk factors, and how they present in minoritized groups and members of low socioeconomic classes, will inform how we consider heterogenous risk profiles [32], and particularly as our aging population becomes more ethnically diverse.

Disease-modifying therapies are emerging as well as identifying biomarkers that can identify early signs of AD. Neurodegeneration is defined as the loss of brain cortical tissue (measured in volume or thickness), and is one of the early biomarkers used to identify early AD [21]. Normal aging is associated with brain atrophy; however, regional cortical atrophy appears to manifest differently between individuals with and without AD [22, 23]. With AD, there appears to be greater ventricular enlargement, sulcal widening, cortical thinning and hippocampal widening [62]. While hippocampal atrophy has been established as a hallmark of AD-related neurodegeneration [18–20], early stages of AD are associated with parahippocampal gyrus atrophy [63]. The parahippocampal gyrus is made up of the entorhinal cortex, perirhinal cortex and parahippocampal cortex in the medial temporal lobe. While the parahippocampal gyrus volume was not shown to be different between normal controls and adults with early AD, the medial perirhinal cortex and entorhinal cortex are atrophied in very early AD [24]. Negash et al. highlighted that intracranial volume and education were associated with “brain resilience” [30], described as the ability to withstand pathological changes to the brain without exhibiting overt clinical signs of the disease [29]. Moreover, even the adults with low education exhibited high resilience if they also had higher intracranial volume, indicating that intracranial volume is associated with brain resilience even with lower education [30]. A recent study by Planche et al. incorporated multiple large-scale MRI databases and whole-brain segmentation using deep neural networks to describe the first chronological progression of brain structural changes with AD over decades and how the trajectories differ from normal brain aging [25]. In the study by Planche et al., brain structural changes developed in hippocampus and amygdala, medial temporal gyrus, entorhinal cortex and parahippocampal cortex (as well as other temporal regions), with the amygdala and hippocampus (followed by entorhinal and parahippocampal cortices) as the most severely impacted areas in AD [25].

2.3 Overview of Supervised Machine Learning

Application of machine learning to the biomedical field involves more than just error minimization but requires understanding of key concepts from collinearity to appropriate evaluation. Herein, we introduce essential principles in machine learning to those interested in tackling research challenges in the biomedical domain. We particularly focus on key concepts and pitfalls that are important to consider when developing models that have practical implications, with an emphasis on clinical outcomes. For readers interested in the statistical framework underlying these approaches, we give a brief overview of the mathematical concepts related to methods to prevent overfitting. We discuss approaches to smaller datasets, and the movement from classical statistics to interpretable machine learning, and ways to reduce bias from inherent heterogeneity originating from patients and data acquisition methods. We recommend pre-processing and evaluation pipelines including open-source software and tools to expedite and simplify adoption of machine learning. Broad integration of machine learning in biomedical engineering will allow for faster modelling and improved experimental and simulated data quality and accuracy in research challenges related to the biomedical field.

Until recently, predictive capabilities were limited by the operator, as the programmer would have to input data and then design a rule-based program in order to produce a useful output. This rule-based program simulates human intellect in the same way a human domain expert might approach a task or detect patterns in their data (e.g., an electrocardiogram signal, ECG). Instead of being explicitly programmed, machine learning discovers patterns from data and solves challenges that are often beyond human abilities. Machine learning outputs a program that can replace the traditional programmer and generalize to new data.

Machine learning will continue to expedite research initiatives as computing power and digital data continue to grow. Traditional statistical approaches may be insufficient when applied to large and high dimensional data of [64, 65, 33]. By nature, humans have biases with regards to visualizing or understanding data, and this may lead to imposing or missing patterns in data [66]. In contrast, machine learning has less inherent operator bias and has the capacity to identify complex and nonlinear interactions between dependent and independent variables [65]. These approaches can generate new ways to characterize disease states and improve prognostication.

Further, machine learning algorithms can be deconstructed into representation, optimization and evaluation components. Representation can be numerical (e.g., support vector machines, neural networks), symbolic (e.g., decision trees), instance-based (e.g., Nearest-neighbor), or probabilistic (e.g., Naïve Bayes). There also different optimization algorithms with stochastic gradient descent and gradient boosting as a common examples. There are multiple ways to assess or evaluate machine learning models depending on whether the output variable is continuous (e.g., mean squared error) or discrete (e.g., log loss, see performance metrics section below for more details). We briefly introduce thesis-specific machine learning concepts including decision trees, boosting, gradient boosting, and ensembles.

2.3.1 Decision trees

Decision trees are hierarchically organized non-parametric binary algorithms that do not make any assumptions about the underlying distribution of the data. The decision tree model infers the class labels from the examples using a series of questions. Algorithmically, it selects the best feature based on impurity based-mathematical principles that define a good split (e.g., Gini impurity, information gain, entropy). Thereafter, the data are split based on the best features value, and the algorithm recursively constructs subtrees for each branch using the remaining features. Information gain is the ‘informativeness’ of a split, more informative splits come first, and is inversely related to entropy. Gini impurity is a criterion to minimize the probability of misclassification, which means that the node will be more homogeneous.

Bias and variance are critical factors to consider when evaluating decision tree performance. High bias in decision trees can lead to oversimplification and underfitting, often caused by a shallow tree depth. On the other hand, high variance can result in overfitting, which may occur when the tree is too deep and captures noise in the data. Balancing the trade-off between these factors, such as adjusting the depth of the tree, is essential for interpretability and improved generalization on unseen data. We can adjust hyperparameters that control the complexity and growth of the tree building process. We can avoid over fitting by reducing depth, not splitting using a feature if impurity does not decrease or only split based on a threshold (e.g., the minimum number of samples in a leaf node). In the subsequent sections, we will introduce the concepts of boosting and gradient boosting, culminating with a discussion of XGBoost. These techniques aim to maintain the relatively low bias associated with shallow decision trees while reducing variance, ultimately resulting in a more robust and accurate model.

```
Input: Data $D$, set of features $F$
Output: Decision tree $T$
Function DecisionTree($D$, $F$)
    if all instances in $D$ belong to the same class $c$ then
        return a leaf node with class $c$;
    if $F$ is empty then
        return a leaf node with the majority class in $D$;
    Choose the best feature $f$ to split the data $D$;
    Create a decision node for feature $f$;
    Partition the data $D$ into subsets $D_0$ and $D_1$ based on the values of $f$;
    $T_{left}$ = DecisionTree($D_0$, $F - \{f\}$);
    $T_{right}$ = DecisionTree($D_1$, $F - \{f\}$);
    Add $T_{left}$ and $T_{right}$ as left and right children of the decision node;
    return decision tree $T$;
```

Algorithm 1: Decision Tree Algorithm

2.3.2 Bootstrap

Bootstrapping is random sampling with replacement used in ensemble learning. The approach is used to estimate population parameters using a small sample size by averaging over mean and variance of many random sampled datasets.

```
Input: Data $D$, number of bootstrap samples $B$
Output: Set of $B$ bootstrap samples $D^*$
Function Bootstrap($D$, $B$)
    for $i = 1$ to $B$ do
        $D_i^* \leftarrow$ a random sample of size $n$ with replacement from $D$;
    return set of $B$ bootstrap samples $D^* = \{D_1^*, D_2^*, \dots, D_B^*\}$;
```

Algorithm 2: Bootstrapping Algorithm

2.3.3 Boosting

Boosting improves weak learners, often decision trees, by combining their outputs to create a more robust model. Each weak learner (i.e., decision stump) corrects the mistakes of the preceding one in the boosting process.

```
Input: Data $D$, set of weak learners $H$, number of iterations $T$
Output: Strong learner $F$
Function Boosting($D$, $H$, $T$)
    Initialize weights $w_i = 1/n$ for $i = 1, 2, \dots, n$;
    for $t = 1$ to $T$ do
        Fit weak learner $h_t \in H$ to data $D$ using weights $w$;
        Compute error $\epsilon_t = \sum_{i=1}^{n} w_i I(y_i \neq h_t(x_i))$;
        Compute $\alpha_t = \frac{1}{2} \ln(\frac{1-\epsilon_t}{\epsilon_t})$;
        Update weights $w_i \leftarrow w_i \exp(-\alpha_t y_i h_t(x_i))$ for $i = 1, 2, \dots, n$;
        Normalize weights $w_i \leftarrow w_i / \sum_{i=1}^{n} w_i$ for $i = 1, 2, \dots, n$;
    Define strong learner $F(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$;
    return strong learner $F$;
```

Algorithm 3: Boosting Algorithm

```
Input: Data matrix $X$, output vector $y$, number of bootstrap samples $B$
for $b \leftarrow 1$ to $B$ do
    sample_indices ← randomly sample $n$ rows with replacement from $X$
    $X_b \leftarrow X$[sample_indices]
    $y_b \leftarrow y$[sample_indices]
    fit model $f_b(X_b, y_b)$
$f_{avg} \leftarrow \frac{1}{B} \sum_{b=1}^{B} f_b$
```

Algorithm 4: Bootstrapping using matrix algebra

2.3.4 Gradient Boosting

Gradient boosting is a boosting algorithm that minimizes a loss function, and maintains low bias and reduces variance by adding weak learners sequentially, often decision stumps, to the existing strong model by optimizing the pseudo-residuals (i.e., negative gradients). A common loss function for classification tasks is the logistic loss. Gradient boosting is a generalization of gradient descent, adapted for functional space, whereas standard gradient descent operates in parameter space. While both gradient boosting and gradient descent minimize a loss function, in contrast with gradient descent which is used for optimizing parameters for a single model with gradients computed with respect to the model's paramters, gradient boosting is an ensemble method where the gradients are computed with respect to the predictions made by the model.

```
Input: Data $D$, loss function $L$, set of base models $M$, number of iterations $T$
Output: Boosted model $F$
Function GradientBoosting($D$, $L$, $M$, $T$)
    Initialize predictions $f_0(x) = 0$ for all $x \in D$;
    for $t = 1$ to $T$ do
        Compute the negative gradient $\delta_t(x_i) = -\frac{\partial L(y_i, f_{t-1}(x_i))}{\partial f_{t-1}(x_i)}$ for all $i = 1, 2, \ldots, n$;
        Fit base model $m_t \in M$ to the data $(x_i, \delta_t(x_i))$ for $i = 1, 2, \ldots, n$;
        Choose step size $\gamma_t$ via line search or fixed value;
        Update predictions $f_t(x) \leftarrow f_{t-1}(x) + \gamma_t m_t(x)$ for all $x \in D$;
    Define boosted model $F(x) = f_T(x)$;
    return boosted model $F$;
```

Algorithm 5: Gradient Boosting Algorithm

```
Input: Data matrix $X$, output vector $y$, loss function $L$, set of base models
       $M$, number of iterations $T$
Output: Boosted model $F$
Function GradientBoosting($D$, $L$, $M$, $T$)
    Initialize predictions $f_0(x) = 0$ for all $x$ in $X$
    for $t \leftarrow 1$ to $T$ do
        Compute negative gradient $\delta_t = -\frac{\partial L(y, f_t(X))}{\partial f_t(X)}$
        Fit base model $m_t$ to $(X, \delta_t)$
        Choose step size $\gamma_t$
        Update predictions $f_{t+1}(x) = f_t(x) + \gamma_t * m_t(X)$ ;
    Define boosted model $F(X) = f_T(X)$
return boosted model $F$;
```

Algorithm 6: Gradient boosting using matrix algebra

2.3.5 Ensemble learning

To reduce overfitting and improve model generalization, different machine learning models can be amalgamated. This can be done by adding models sequentially (i.e., boosting, e.g., XGBoost) or by bagging. Bagging involves training models in parallel on randomly selected subsets of the training data (bootstrapping; e.g., Random Forest).

2.3.6 Role of machine learning in medicine

Machine learning models can aide in decision making, monitoring patients, risk mitigation strategies, understanding the underlying mechanisms behind disease, hypothesis testing, participant recruitment and retention, patient monitoring, large data, translation to the clinic, and improving diagnostic and prognostic tools, which are all important in medicine [67, 68]. As an example, risk prediction can play an important role in predicting patients at the highest risk of developing a disease or a fatal outcome from a disease, which enables early intervention. Current risk scoring systems are biased [69–71] and based on more traditional statistical approaches, thereby limitating their utility and validity when used for interpreting patient specific risk factors as they can overlook complex interaction among features and non-linear relationships [64, 65, 33]. As an example, use of logistic regression for certain tasks can lead to poor external validation for high-risk subgroups and overestimate risk, which impairs the decision making processes. Machine learning on the other hand can discover operator-independent complex interactions among features unbeknownst to domain experts.

For patient-specific predictive tools, model assessment relies on both the discrimination ability (e.g., accuracy) and the individual risk predictions (i.e., calibration) [72]. Discrimination ability describes the ability for the algorithm to separate classes (e.g., groups), while calibration is the assessment of probabilities based on the actual risk in the population, or the correctness of predicted class probabilities. Machine learning models are calibrated when the probability estimate of a data point belonging to a class is very important (e.g., in medical risk) and the distribution of the predicted probability is matched to the expected distribution of probabilities for each class.

Further a model might perform well in discriminating patient risk status but have high miscalibration or poor capacity to estimate individual risk scores, which is required for developing patient-specific predictive tools. There is some evidence that more complex machine learning models may not surpass linear models such as logistic regression in prediction. Anita *et al.* showed preference for Logistic regression over optimized machine learning algorithms for prognostic and diagnostic prediction models [73], while others have showed use case preferences for logistic regression over more complex machine learning algorithms [74]. However, these studies may be limited in data. With enough data,paticularly nonlinear data, logistic regression cannot compete with extreme gradient boosting (XGBoost[39]), which is an implementation of the stochastic gradient boosting ensemble algorithm. We draw comparisons between machine learning algorithms and traditional statistical approaches in sections below.

2.3.7 Traditional statistics vs machine learning

The relative efficiency, or predictive accuracy, of various machine and statistical methods of learning for the prediction of patient outcomes is becoming increasingly important. Classical statistical methods focus on inference from fitting dataset specific probability models and relies heavily on the user's domain expertise. These models can verify assumptions and quantify the confidence of an effect, independent of noise or chance. Classical statistical methods can also measure the linear relationships between individual predictors, but they assume predictor independence and fail to take into consideration the inter-predictor nonlinear/complex interactions and systemic aspects. In contrast, machine learning uses a general learning procedure discussed above and is more useful in P>n datasets, where the number of participants, n, is less than the number of variables, P. Consequently, machine learning involves relying on the capacity of the model to learn from the observable data and generalize to new data. While methods generally, fall into one domain or the other, they can sometimes involve both machine learning and statistical methods (e.g., bootstrapping [75]).

Austin et al. (2021) used simulations from various data generating processes informed from empirical datasets, to deduce that unpenalized logistic regression (with no shrinkage, or feature selection) produces good performance, and, at times, is superior to more complex machine learning methods [76]. Given similar performance to more complex machine learning models, logistic regression approaches may be preferred as they have an interpretive advantage, including providing odds ratios that allow for the quantification of relative covariate contribution to the outcome. Lundberg et al., (2020) varied the nonlinearity in simulated data, and showed that an increase in nonlinearity is associated with a decrease in the accuracy and interpretability of machine learning based logistic regression models due to a mismatch between the model and data, resulting in an increase in % of weight attributed to irrelevant features, thus less interpretability [45]. To this end, low-bias non-linear models like the extreme gradient boosted ensembles can be better performing and more interpretable than high-bias models like logistic regression, as the former places more emphasis on the input features and is perhaps a better representation of the data-generation process. Further, this study highlights superior performance in boosted approaches. Boosting is an approach that combines weak learners (decision stumps) sequentially, so that each new tree corrects the errors of the previous ones.

2.3.8 Deep learning vs traditional machine learning

While deep learning is leading the AI revolution, traditional machine learning (e.g., gradient boosted ensembles) still outperform deep learning on heterogeneous tabular data. We provide a brief overview of deep learning and the convolutional neural network.

Historically, machine learning methods could not process raw data [77]. Instead, traditional machine learning techniques relied heavily on feature engineering and selection methods [78]. These methods required deep domain-expertise and 'hand-crafted' features to train a mathematical model (e.g., classifier) to produce useful outputs. The feature engineering methods have clear limitations, including operator-dependency and the potential for inadequate data differentiation within, and between, datasets [79]. Unlike traditional machine learning, which relies on feature selection, deep learning allows computers to discover the features that are most effective in differentiating the data using a general-purpose learning procedure [80]. Specifically, these models involve multiple steps (i.e., layers) to convert input data (e.g., pixel intensity of an image) into useful output (e.g., category classification).

Deep learning draws its inspiration from the biological neuron. The cell body receives input from the dendrite and produces output activations (after reaching a voltage threshold) to other neurons, transmitted via the axons to the axon terminal. An artificial neuron outputs a single value, in contrast with the biological neuron's time series of spikes. Every input influences every neuron in subsequent layers. Each of the inputs ($x1, x2, x3$) is connected to a node, with certain weights ($w1, w2, w3$). The output is generated by computing the weighted sum, then applying a non-linearity, to approximate complex decision boundaries. A bias is also introduced to constrain output to an appropriate range.

To train deep learning networks, model parameters (weights or connections between neurons) are identified that minimize the training error (the difference between the true data from estimated data). Best weights are determined as those that can

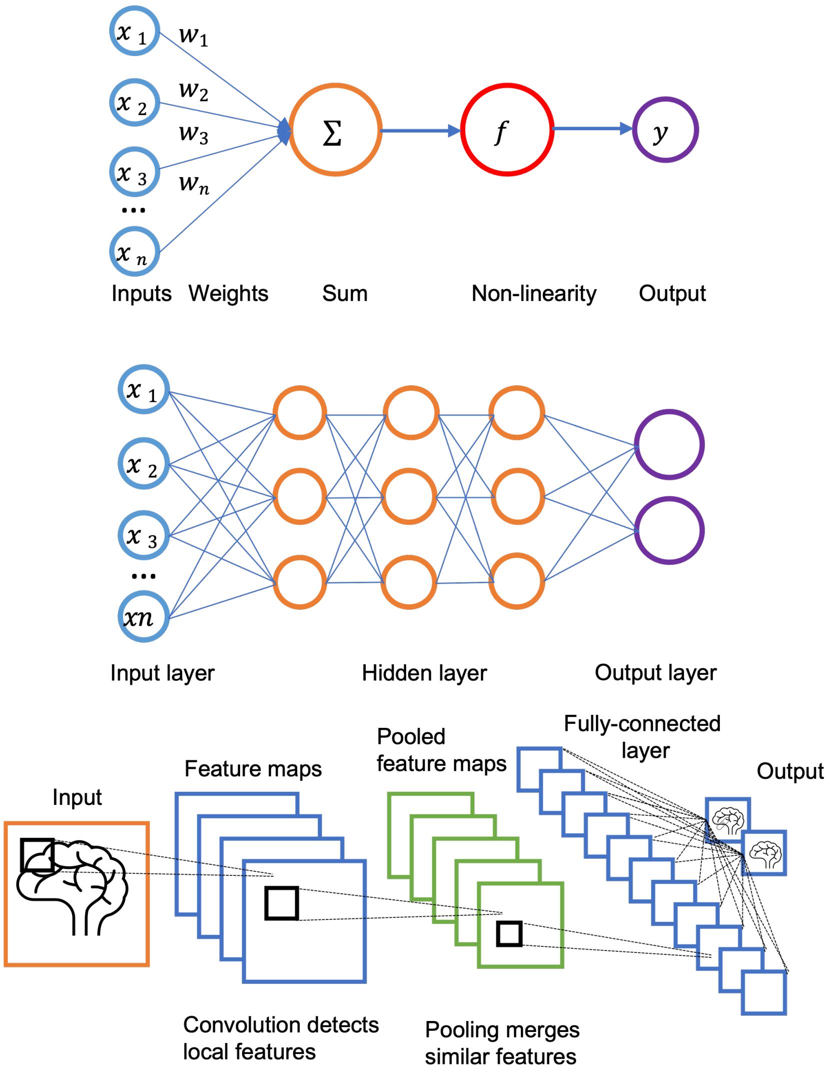


Figure 2.3: An overview of a simple convolutional neural network

correctly classify as many points as possible. A goal is to minimize an objective function (cost/loss function) that measures the error between the target and network output to maximize model performance. Stochastic gradient descent is the most popular method used to minimize the objective function by computing its gradient. The gradient of a loss function is imputed with respect to each weight, and then iterated backwards from the output to the input layer using the chain rule from calculus. The gradient for each weight, informs the potential error change if the weight is modified. The goal is to reach the minimum of the objective function (i.e., high performance) by adjusting the weight vector in the opposite direction to the gradient.

With the rise of computational power and digitized image data, deep convolutional neural networks have become the cutting-edge approach for image tasks. The deep learning approach can learn features from the input data in multiple convolutional layers within a deep architecture, with the term 'deep' referring to the depth of the layers (e.g., more than 5 layers). Image classification tasks using traditional artificial networks involve converting an image (pixel width x height x depth) into a 1-dimensional vector. The 1-dimensional vector is constructed by appending the pixels from each row of the image, whereby the top left region of pixels is represented at the beginning of the vector, while the end of the vector represents the bottom right region of the image. But this 1-dimensional representation of an image is not suited for image tasks as we lose spatial features of an image (i.e., locality) and results vary based on small shifts in individual pixels (translational invariance). To take advantage of spatial features of an image and become invariant to small shifts in individual pixels, deep convolutional neural networks use the process of convolution.

The convolutional neural network comprises of a convolution layer, nonlinearity, and a pooling layer; Additional convolutional/pooling layers can follow). In the convolutional layer, a matrix of weights (i.e., kernel/feature detector) is applied across the receptive fields (i.e., patch in input space that produces a feature of an image to detect features (e.g., horizontal/vertical lines). Element wise multiplication is then performed with the part of the image that the kernel is on, then sum up the values to get a single output. In other words, local regions in the input image are linked to artificial neurons and the convolutional layer computes the output of those artificial neurons. As the kernel traverses the image, multiple outputs are aggregated to form a feature/activation map. This allows for similar patterns to appear in different locations in the image. During the convolutional layer, weights are learned (via gradient descent and back propagation) to detect local features from the previous layer, while the pooling layer combines similar features (e.g., max value in a region), down-sampling, and reducing the size of the feature map.

With the input of raw data, deep learning, can facilitate the automatic discovery of features based on the best classification of the data [80]. As raw input (e.g., matrix of pixel intensities) passes through the multiple layers of the deep learning architecture to a more abstract level, the suppression of any irrelevant variations and the amplification of image characteristics important for classification occurs. For example, the first layer of a network may represent edges (or lack thereof) at unique positions or orientations. The second layer may detect recurring patterns by recognizing edge groupings, irrespective of any minor differences in edge positions. Third layers and beyond may aggregate patterns into larger amalgamations that begin to resemble a familiarity in the input image. Thus, the earlier layers are often more general, and as the layers deepen, they will have an increased specificity to the trained task [77].

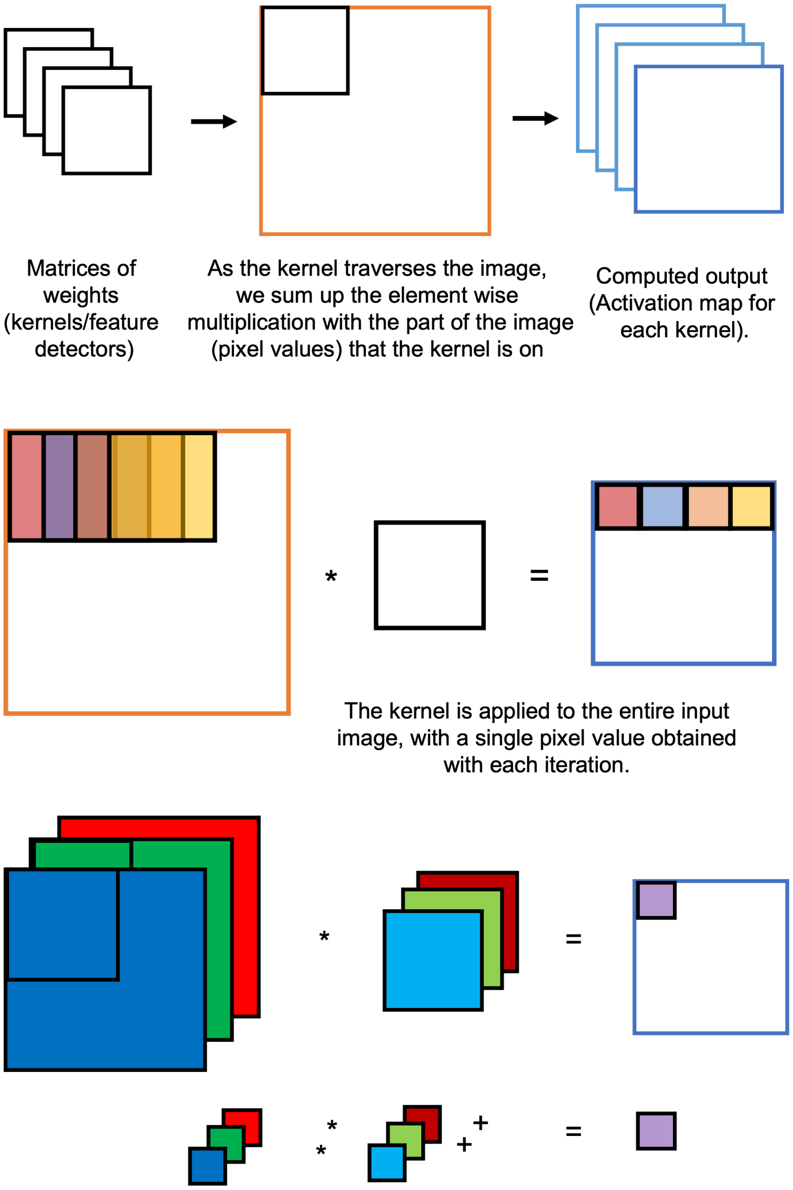

Matrices of weights (kernels/feature detectors)
As the kernel traverses the image, we sum up the element wise multiplication with the part of the image (pixel values) that the kernel is on
Computed output (Activation map for each kernel).
*
=
The kernel is applied to the entire input image, with a single pixel value obtained with each iteration.
*
=
*
*
+
+
=

Figure 2.4: The art of convolution

As a thought experiment, imagine employing a deep learning approach with the purpose of classifying medical images as belonging to either abnormal or normal categories (e.g., disease or non-disease states, respectively). Begin by collecting a large, labeled dataset of both abnormal and normal cases, input the data (e.g., medical images) and receive the output data as a vector of probability scores for each category (e.g., abnormal = 0.4, non-disease = 0.6). The goal is to have the largest probability for our desired category, but this is an unlikely outcome prior to training the model. In deep learning, mathematically, the function, f, generates outputs, y, from input data, x. f is approximated using a parameterized function, with parameter values that are learned during training. Adjusting these parameter values yield varying degrees of error between output scores and desired scores. A function is quantified that calculates the error between output scores and the desired scores with an aim to minimize this error function by allowing the algorithm to change its weights according to the error the weights produce. The weights of various features will require adjustments, depending on the classification probability of an image, estimated by the statistical properties of a set of training images. To reduce the error, the weights are adjusted along the direction of the gradient. This function is analogous to finding the steepest descent in a high-dimensional hilly environment.

2.3.9 Applications for smaller datasets

With the application of smaller datasets, transfer learning can be used on deep learning models that have been trained on larger training sets [77]. Given that images contain similar components (e.g., edges), pre-trained networks in one domain can classify images in a different domain, although, as the data moves further away from that of the original data source, the transferability of features decreases. A network trained on a completely different source material may outperform its original source material. Also, algorithms can be trained initially on a more general classification task, and then a more specific task [81]. Data augmentation methods create variations of the original image training set (e.g., flipping, cropping, rotating, scaling). These variations exponentially increase the size of the original dataset and allow the features to be generalizable and reduce overfitting [77]. The augmented training set does not equate to a non-augmented training set of the same size, as the augmented images are highly correlated and provide less learned information. While transfer learning and data augmentation attenuates the need for larger data sets, high performance deep learning models require large data sets. More recently, U-net based architecture and variations provide efficient and fast segmentation of images, without the requirement for large samples, and has been utilized in segmentation of medical images [82]. For small datasets in classical machine learning, it becomes crucial to use deep domain and statistical expertise to preselect input variables or resampling (e.g., boot strapping) to improve stability. Cost-sensitive algorithms can help in imbalanced data in lieu of resampling methods. Alternatively, generative models (e.g., CTGAN) can produce statistically similar data Further, semi-supervised approaches can improve performance on smaller datasets by using the limited data to label a larger unlabeled dataset to train.

2.3.10 Explainability and Interpretability

While explainability and interpretability is at times used interchangeably, herein, we differentiate explainability as providing explanations for complex models (e.g., Shapley values) without necessarily understanding the decision process. In contrast, interpretable models are understandable by design (e.g., decision trees, ML-based linear models). Further, as mentioned above, gradient boosted algorithms use recursive decision trees, and offer great interpretability relative to black box approaches (e.g., neural networks).

The motivation for explainable or interpretable AI can be illustrated with the following example: A patient may be diagnosed with cognitive impairment. The memory clinic may have a machine learning algorithm that indicate patients at the highest risk for developing Alzhemier's disease dementia, based on the patient's pre-intervention clinical state (e.g., baseline neuropsychological testing, blood biomarkers, imaging). But then the patient or end user may ask why did the algorithm make such a prediction? When using explainable AI, a prediction is generated, but an explanation also accompanies the prediction. While it's difficult to understand the inner mechanics of complex machine learning models, there is a need to discover relationships between input and output data; however, complex machine learning applications to clinic remain challenging due to their lack of explainability.For example, visualizing feature maps at each convolutional layer allows for correlation of the deep features extracted by the deep learning to the target object. Interpreting deep learning in a clinical or research task is still not a trivial and is operated like a blackbox with only a few algorithmic concepts of explainable AI [83]. While performance is an important metric, future applications of deep learning must strive to establish trust and implement methods to add the end-user to the training loop so as to provide feedback and interact with the provided explanations from the deep learning models in order to improve overall classification efficacy].

SHAP (Shapley ADitive exPlanations) [84] is a model-agnostic, game-theoretic approach to explaining machine learning outputs, which allocates the value of each feature for a specific prediction and has been successfully applied to clinic [48]. SHAP plots provide an intuitive explanation of what led to the patients' risk. The SHAP Python package can be found here: https://github.com/slundberg/shap. More recently, SHAP methods have been applied to tree based algorithms (e.g., random forests, gradient boosted) [45]. SHAP summary plots show feature's impact and importance. Each point represents a unique occurrence of the dataset (i.e., a single patient). Their location along the x-axis (i.e., SHAP value) shows that feature's influence on the model's output for that patient. Higher SHAP values are associated with a higher risk. A feature's importance is determined by its average absolute Shapley values and sorted along the y axis (a higher position equates to a greater importance). SHAP feature dependent plots show global and individual level variability. These charts demonstrate how one factor affects model predictions. Each point represents a patient, like in the prior summary plot. A point's x-axis location equates to the feature's value, the y-coordinate of a point defines its SHAP value. Nonlinear dynamics and interaction effects are depicted by vertical SHAP value dispersion for a single feature value. We can identify the most and least important feature values for the model based on their SHAP value influence. Those features with more influence had a broader range of SHAP values, whereas those with less significance had values close to zero. It's interesting to note that SHAP feature dependency graphs may also assist us in locating significant inflection points for the various features.

2.3.11 Limitations in machine learning

Machine learning can discover patterns unbeknownst to domain experts, but it can also learn undesired approaches that exploit confounding variables or artifacts in the data set, resulting in right classifications for the wrong reasons [85].

Deep learning performance has a linear relationship with the volume of the training data set, up to and beyond 300 million images [77]; thus, data acquisition is a barrier for efficient deep learning performance. By design, deep learning requires labeled and large-scale data acquisition. Extensive data sets allow deep learning algorithms to excel at task classification; however, there are many studies with small training data and/or which lack validation [79]. Besides large data sets, accurate and high-quality labelling will have a direct impact on the overall performance of a deep learning model. Consequently, the benefits of supervised learning are greater than that of an unsupervised approach for classification tasks.

Abnormal cases are less abundant, and more difficult to acquire data from, than normal cases, which limits deep learning models in their ability to classify anomalies as it creates imbalanced classes. If classes are imbalanced, the decision for learning algorithms will be biased towards the majority class (leading to overfitting) as the majority class is over-represented in the training data. For example, if abnormal cases represent only 20% of the training data, then a model, by predicting the majority class, would result in 80% accuracy without learning meaningful features. Combating class imbalance can be Addressed with cost-sensitive algorithms or sampling methods that rebalance the data by creating synthetic data. Data acquisition methods require the capacity to reach otherwise difficult to recruit populations and collect data from a spectrum of normal and abnormal cases, which will further help deep learning models to form accurate predictions. The acquisition of medical imaging data, with expert annotations, can prove to be a challenge, with diverse populations presenting a large variance in abnormal and normal cases, presenting an Additional barrier. Thus, the data requires high-quality labelling and measurements that depict patient anthropometrics, age, race, and ethnicity to help models cover more variability in abnormal case data.

2.4 Fundamental concepts for model development

2.4.1 Bias-variance tradeoff

In machine learning, the bias-variance trade off affects model accuracy and interpretation[54]. Models must be accurate, avoid bias (valid) and low variance (reliable, precise). The bias-variance trade off speaks to the inverse relationship where the need to decrease variance results in an increase in the bias, and vice versa . If a model has high bias, it means that it is underfitting the training dataset (model is too simple, overlooks important relationships in the data), and will show as low training and cross-validation accuracy . High bias can be resolved by increasing the number of features, or by decreasing regularization. High variance models overfit the data and show a large gap between training and validation accuracy, and can be resolved by collecting more data, reducing the number of features, or increasing regularization. In contrast to plotting the training and test accuracies as a function of training sample size in learning curves, we can use validation curves and plot accuracy as a function of a regularization parameter (e.g., C-statistic in logistic regression).

For example, to predict a cognitive variable (single value) y, from a predictor

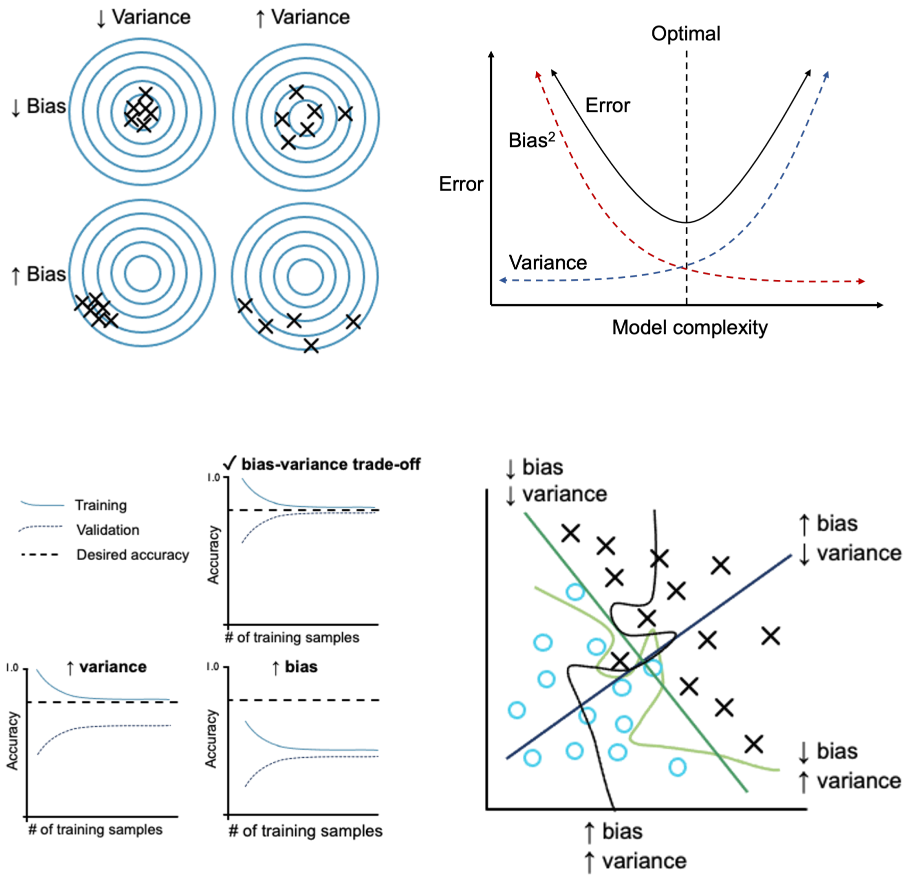


Figure 2.5: Illustrating the bias and variance trade off for model performance

variable X, requires determining a best estimate for y. To estimate y we need to introduce conditional expectation and variance.

Where y and X are random variables that represent our dependent variable, and predictor(s), respectively. The conditional expectation of y given X is a random variable that is a function of the random variable X, such that, for any set of possible values x, the expected value of Y is conditioned on the event that the value of X in the set x, equals the expected value of f(X) conditioned on the event of X in x.

$E[y|X]$ is the best possible approximation of y, based on the value of X. That is, knowing something about X, tells us something about y as these variables are not independent, and f(X) is our best guess. In particular (and by law of total expectation), the expected value of the expected value of y given X equals the expected value of y. In other words, the expectation of the conditional expectation is equal to the overall expectation of the random variable. $E[E[y \mid X] = E[y]$

The variance of a random variable, Y, is the expected value of the square of y minus the square of the expected value of y. Alternatively, the expected value of y minus the expected value of y, quantity squared.

Intuitively, y is random (e.g., a uniform value between 1-100), E[y] is fixed (an average with a value of 50), then the variance is, on average, the squared difference between the actual value of y(0-100) and its expected value (always 50). For a random sample of y , if we take each of the different values in our random sample, subtract 50 (average/expectation of y), square it and take the average (these values are never exactly the average, and will be off by a certain amount). The average squared amount that its off by is the variance (fixed number).

The conditional variance of y given X, equals (identical to the above with conditional expectations in place of the expectations) the expectation of y squared given X minus expectation of y given X, quantity squared. Alternately, the expectation of y minus expectation y given X, quantity squared, given X.

Simply, if we know the value of x, this gives us information or changes our understanding of the probability distribution of y, and that changed probability distribution of y, has its own expectation and variance based on X.

Back to our example or predicting cognitive scores (y) based on age (X). Our best guess is $E[y \mid X] = f(X)$. We will not get perfect predictions, as even with a given sample, two patients can have different cognitive scores and be the same age. Moreover, we do not know the underlying joint probability distribution of y and X, and all people that were and will be born in the population as we just have a messy sample. We create a model that says if the patient's X has a given value, then their cognitive score parameter will be this given value. Then we get new patients with X and now want to predict their cognitive scores based on past data. In this situation, how accurate do we expect our prediction to be? How do we quantify how good our function/model is at making predictions?

To quantify our model's ability to make accurate predictions, we can collect the deviations of each data point from the prediction (e.g., mean squared/absolute error). There are theoretical reasons for picking methods to quantify a model's ability to make predictions. For simplicity with this illustration, we will focus on mean squared error (MSE). The mean squared error (MSE) of f(x), is the expected value of the squared difference of the actual y-values, and the predicted values, quantity squared, conditioned on the associated X value equaling the little x, as we can have different observations for a single value of x.

The mean squared error decomposes into: 1) a (squared) bias term f(X)-E[f(x)], or the systematic/approximation error, or the bias in using our function f to approximate the true function (where high bias means the model is systematically off for x values, i.e., always low or high), 2) the variance of the statistical or machine learning process, or irreducible error (i.e., y cannot be perfectly predicted based on X, the random statistical fluctuations even for the most unbiased estimates), and 3) the variance in estimating our function Var (f(X)) (high variance means the model is hyper-responsive and changes drastically with a given input). The bias-variance trade-off speaks to the inverse relationship in the equation above, and specifically describes the need to increase variance in order to decrease the bias, and vice versa.

If we linearly interpolate the data, and consequently create a model that overreacts to the data, or overfits and thus has large variance, the f(X) is highly randomized in response to small changes in data, and thus contributes to a large error, even though the bias would be very small. At any point, there exists high variability, as the points can be anywhere, but the bias will be small (not systematically off). If we estimate based on a constant value, for example the average of all values (horizontal line), variance will be small and Adding data points will not change the average y value by much, but now with much higher and systemic bias, as each x value will be far off from the predication (i.e. large gap between E[f(x)] and the function f(X), and will increase error). An appropriate or consistent model has the bias and variance terms shrink towards 0, and error should converge to the irreducible error, as sample size, $n \to \infty$.

In practice, however, we have finite samples, and we choose between having a model that has higher bias or higher variance. This phenomenon depends on the type of model, Var(f(x)). Typically, linear machine learning models require more assumptions, leading to high bias but low variance as they are unable to capture relevant relationships. Non-linear algorithms show low bias but high variance (fewer assumptions), with quadratic and higher polynomials able to model increased complexity, thus having lower bias and higher variance. There is a trade-off between methods that are explanatory and give insight to the underlying mechanism, and methods with high predictive power[103]. Neural nets for example, are quite predictive but not as interpretable, whereas logistic regression is more interpretable. The relationship between predictability and interpretability can be described by a u-shaped curve, where with increasing complexity, an increase in predictive power and decrease in interpretability is observed, but at a certain point, predictive power will decrease.

2.4.2 Cross-validation

Cross-validation methods for internal validation are intended to assess the model's expected performance (by alternating independent test sets while using all available data) for future estimates of the target population, specifically the model's prediction accuracy (point estimate) and confidence intervals. In reality, cross-validation estimates are indicative of the average prediction error of models fit to a hypothetical training set taken out of the same sample [86]. In contrast to popular belief, cross-validation does not measure the prediction error (for both, data-splitting and bootstrapping methods) for the model fit to the training data.

In regular cross-validation we estimate the error of, and find the best hyperparameters for, the model on the same set of training and testing data, which may lead to high bias as the independence assumption is violated [86]. Further, a lower average prediction error is observed relative to the final model and lower and narrower, variance and confidence intervals, respectively. To estimate the unbiased error with low variance, we can use nested cross-validation as it produces an unbiased error of the estimate relative to the test dataset. In a nested cross-validation, albeit computationally expensive, we can run nested loops, an inner loop (feature/parameter selection) which functions similar to the normal cross-validation, but an Additional outer loop (to assess model performance) is run that withholds the test data. This can be repeated (e.g., 100 times) to improve statistical stability. More research needs to elucidate the efficacy of nested cross-validation over regular cross-validation. After internal validation, the model can then be trained on the sample in its entirety.

To assess whether these models generalize to the target population, we can use bootstrapping or the aforementioned nested/cross-validation methods as a form of internal validation. Bootstrapping (i.e., sampling with replacement) offers statistical stability, low prediction error score, low bias, and performs better than k-fold cross validation [87, 88], which preforms better than a single-split. Repeats (e.g., x100) of 10-fold cross validation are recommended for variables¿observations datasets. Further, it is important to contain pre-processing, feature selection, or any other processes in developing the model, within each bootstrap sample, or fold (see nested cross-validation above).

2.4.3 Generalizability

A lack of generalizability happens when the machine learning algorithm is overfit to the training data (analogous to a student memorizing the practice exam). To reduce overfitting, we can increase the volume of our dataset, reduce the complexity of our model (less variables), use regularization techniques, or ensembles methods. Moreover, we need to ensure that data leakage does not occur, in that the test data is kept away from the pre-processing and training stages (see section on cross-validation above). Sample data needs to represent the target population of interest. High-accuracy machine learning models trained on data from a single site (e.g., a single hospital or city) may not generalize to another clinical or academic site, because of overfitting and differences associated with, but not limited to, population characteristics (e.g., age, race, ethnicity) and non-medical dependent characteristics of an image (e.g., image processing/reconstruction, and data collection equipment). Data acquisition methods need to reach diverse cohorts, with variability in equipment and settings in which data are collected, thereby increasing generalizability of the models by covering more variability in the data [89]. Perhaps increasing generalizability of AI may rely on developing models based on diverse datasets.

One route to increase generalizability of models may require involvement of hospitals' Picture Archiving and Communication System (PACS) which houses and links patient data and history from diverse cohorts. While the PACS suffers from lack of open-source utility, hardware difficulties, and multi-site/hospital integration [90], integrating AI models into the PACS as well as developing AI models from PACS data would present an advantage to generalizability of these models as they would be developed on a range of diverse cohorts (i.e., socioeconomic status, race, ethnicity) with a spectrum of health disorders, accounting for patient history, hospital readmittance. Some countries, like Germany, have a PACS that's connected within the entire country, whereas in Canada, PACS systems are city-based, which significantly reduces data-sharing. Improving inter-connectedness of PACS technologies, and further access and sharing of data, would improve generalizability of AI models when applied to PACS as this would mean that model development would not rely solely on data from a single site.

Machine learning models trained on diverse data or protocols can still introduce bias. Machine learning-based algorithms aim to reduce bias from researchers but are still prone to incorporating and exacerbating bias from data itself [91]. There exists a large disparity with under-represented groups in terms of diagnosis, prescribing of pain medication, referrals, and treatment, which reduces the effectiveness of current treatment options [92]. These data are also training future AI algorithms, which can lead to more bias and deepen the divide for underrepresented groups [93]. With regards to the data, there exists heterogeneity in operator-dependent imaging modalities but also someone to interpret and label the findings.

These methodological aspects introduce bias and further contribute to reproducibility issues. A recent systematic review[94] found that none of the machine learning models are fit for clinical use for COVID-19. We encourage readers to follow the recommendations therein that include, but not limited to Addressing the lack of, external validation (i.e., in contrast with internal validation, where we test data from the same source, external validation speaks to testing data from a difference source) robustness analysis, reporting demographics, assessing significance, confidence intervals for performance, and generalizability. There also exists high heterogeneity in the results of machine learning algorithms that stem from the differential validation methods, or optimization. Data partitioning, feature selection and engineering methods, the reporting of different performance metrics, and all of this is exacerbated when methods are under reported. Finally, relative to other fields, machine learning in healthcare, compared to other domains, has less open source (less code/data availability) initiatives.

2.4.4 Feature engineering

The quality or robustness of the model depends on the sample size, and specifically sample amount and quality.

Sample size

Sufficient sample size is important for ascribing weights to covariates, the performance of the model, and to avoid overfitting. Sample size is dependent on signal to noise ratio, where lower signal to noise may require more samples. Finally, the complexity or the number of covariates to include in the model plays a role in sample size selection. The notion of 10 events per predictor parameter (one for each beta term in the modelling equation) is ill-advised [94]. The required sample size depends on more than just events per predictor parameters, but also depends on total participants, incidence rate in target population, and desired performance. The assumption here is that variables require only one beta term in the model equation when, more beta terms are actually required for modelling non-linear effects that are quite ubiquitous in the medical field, and for multi-category classification. For small datasets, it may be helpful to use deep domain expertise and/or feature selection methods that don't involve the labels to pre-select input variables. To calculate the minimum sample size that is required to minimize overfitting and provide accurate estimates, sample size formulas can be used, which vary with simulation approaches, or high dimensional data, or binary,continuous,multi-category outcomes. The accuracy of these models in the medical world is of high importance as these models may be used in predicting death and patient outcomes. We direct readers to a a recent paper by Reley at al. (2020) that discusses sample size for developing prediction models a priori, or for existing datasets, and how to implement sample size calculations using the R package pmsampsize.

Standardization

Some machine learning algorithms will require standardization of features, as performance issues arise when features exist at different scales (e.g., various cognitive test scores, brain volumes). For example, support vector machine and K-nearest neighbor algorithms depend on a distance measure, which becomes biased if features vary in their scales. It is also recommended to scale before data reduction techniques such as principal component analysis (PCA, see section on dimensionality reduction below). Generally, it is recommended to scale variables to the standard normal (i.e., scaling features to a mean of zero and a standard deviation of one) when algorithms exploit distances or similarities. Although, some algorithms may benefit from, but do not require, standardization (e.g., decision trees, naïve bayes, gradient boosted ensembles, linear models unless regularized).

Dimensionality reduction

Often, models can overfit high dimensional data, and it is strategic to reduce dimensionality (e.g., feature selection) to prevent overfitting. High-dimensional datasets result in overfitting to the training set, and consequently under perform in the test set. To retain most of the information, but reduce the dimensions of the data, we can use dimension reduction techniques such as PCA.

PCA is a method of summarizing data. For example, we can describe a patient by their age, by their cognitive testing scores, etc. We can amalgamate a list of characteristics for each patient, but many characteristics will measure related properties and so are redundant. If some characteristics are redundant then we can summarize each patient with fewer characteristics. This is what PCA does. PCA does not discard any characteristics, instead it creates synthetic characteristics that turn out to summarize patients well (i.e., linear combinations of age characteristics). PCA finds the best possible characteristics, by finding characteristics that show large variation across patients as possible (instead of using properties that are quite similar across patients). PCA is a technique to reduce dimension by taking linear combinations of the original variables. Each linear combination explains the most variance in the data it can. Each linear combination is uncorrelated with the others. To this end, variable x at time point y is what we should focus on for further analysis because that is where the most change is happening.

PCA can reduce the number of predictors by extracting linear combinations of variables making it useful for data visualization and data reduction for subsequent analyses. Each component (i.e., eigenvector) is a linear combination of all input variables and is orthogonal to all others, which allows for accessible figures. Similar to the fast Fourier transform (FFT) and independent component analysis (ICA), as these methods give data an alternative representation, which allows us to see the structure/patterns of the data and can be used as a form of filtering (e.g., by setting columns of PCA or ICA separation matrices that correspond to unwanted signals to zero). While FFT represents data in the frequency domain, ICA/PCA represent data in the statistical domain. Both ICA and PCA find a measure of independence, then decorrelate the data by maximizing this measure of independence. For PCA, the measure of independence is variance, which is used to discover axes which leads to ordering of independent gaussian noise sources. While ICA uses the non-Gaussian parts of a set of signals.

Feature selection

Machine learning demands many samples, but not features. Features are also referred to as independent variables, covariates, predictors, and their selection involves human based domain expertise (e.g., clinical input), careful selection methods and feature engineering. Feature selection is an approach for dimensionality reduction that aims to select relevant and eliminate irrelevant features, reduce noise, redundancy, and collinearity. Features can either be individually ranked by an algorithm according to their importance or the best features can be chosen using a specified measure (e.g., mutual information). Further, feature selection methods can be conceptually categorized into wrapper (e.g., stepwise selection), filter (e.g., correlation-based), and embedded methods (e.g., LASSO/Ridge). In contrast with wrappers, and embedded methods, filter methods are algorithm independent as they depend on the nature of the data itself. Wrapper methods find optimal features by evaluating a combination of features on a specific algorithm. Embedded methods incorporate feature selection in the model training process. Unsupervised approaches to data reduction techniques (e.g., PCA) are recommended in lieu of stepwise feature selection. Stepwise feature selection methods have reduced chances of finding the 'right' variables, high chances of overfitting, distorts confidence intervals and p values, and are ruined by collinearity, whereas these issues are not present with data reduction methods (e.g., PCA). Optimal data representation may best serve computation but may not be tailored for human understanding and use. Several feature engineering methods are used and include, but are not limited to, statistical summaries of the distributions of interest, and features picked based on deep domain expertise in the field (e.g., clinical input). Processing of the data to extract features can include applying methods such as wavelet transformations, data and noise reduction methods (e.g., principal and independent component analyses, log transformation, variable collapsing). To improve feature stability, we recommend bootstrapping feature selection in a nested cross-validation design, where feature and hyperparameter selection occurs in the inner loop and ensures no data leakage.

Collinearity

Ordinary models assume independent observations. However, observations with repeated measures, or subsets of measures tend to be correlated at each level, because lower-level observations share common groupings at higher levels (e.g., multiple measures of a patient). Failure to take these correlations into account may lead to biased estimates. In multiple regression for example, it is assumed the variables, X, are independent. If not, we can run into issues of co-linearity, that is, strong enough correlations or multiple correlations that can negatively impact regression estimates (namely the beta coefficients, errors, and significance).

The partial derivative is the change in Y for a 1-unit change in $X1$, while $X2$ is held constant, thus when observing $X1's$ unique impact on Y, we need to ensure that $X2$ is not interfering with the analysis. When collinearity exists between $X1$ and $X2$, our assumption for multiple regression is violated, thereby leading to inflated variances (and standard errors) of the regression coefficient, and the signs and magnitude of the weights may be inaccurate or non-significant with high R^2 values.

Variance inflation factor is indicative of the degree to which the standard errors are inflated due to collinearity and is mathematically the reciprocal of the tolerance. Tolerance is the % variance in the features that is not accounted for by the other features ($1 - R^2$) - values of <.10/.20 are problematic). Thus, VIF = 1/tolerance. For example, 1/.10 = 10, indicating that our standard error will be inflated by a factor of 10. The regression equation in matrix form, where y is a Nx1 data/response vector, X is a N×Q matrix, with Q columns/regressors, b is a vector of regression coefficients Qx1 tell us how strongly our response reflects the input of each column plus normally distributed noise, e. We can estimate the regression coefficients using ordinary least squares regression,

XTX is the gram matrix, on the diagonal is the sums of squares of each regressor, off diagonal is the inner product between first regressor and second regressor. This matrix is important as the diagonal elements tell us about the linear dependence (if zero = columns are linearly independent), or lack thereof, of our columns/regressors. This matrix also determines our variance-covariance matrix of our regression coefficients. The variance-covariance matrix of the ordinary least squares estimator, is the inverse of the gram matrix multiplied by the noise covariance. This matrix contains the variances of the regression coefficient estimates on the diagonal and the covariance between our regressors off the diagonal. If experiment with different values of collinearity, we see that regressors that have high co-linearity to other regressors in the gram matrix, will have higher estimation variance, than estimators that are independent (linear dependence closer to zero).

2.4.5 Performance metrics

A confusion matrix elucidates what occurred during classification (i.e., predictions compared to the ground-truth labels), from which we can derive other metrics of performance. Recall (i.e., also known as true positive rate, or sensitivity) indicates the proportion of participants who were correctly labeled (i.e., actual positives, identified as positives) and can be calculated as correct labels or true positive (TP) divided by total number of people with the label (TP + false negative (FN)). Specificity refers to the true negative rate or the proportion of actual negatives, identified as negatives (true negative/true negative + false positive; TN/TN+FP). Precision (i.e., 1-specificity) refers to the proportion of positive predictions that were actual positives). The false positive rate indicates the proportion of incorrectly labelled patients divided by all who do not have that label (i.e., all actual negatives). We might optimize for sensitivity for our model when our goal is to identify the most patients with the disease (i.e., less cases of disease are missed), or optimize for specificity when we want to reduce the positive rate. For example, accuracy (TP+TN/TP+FP+TN+FN) is the proportion of correct predictions, and an F1 score (2x precision x recall/precision+recall) of 1 indicates perfect recall and precision scores. R-squared indicates the goodness of fit of a regression-based model.

		Predicted		
		Positive	**Negative**	**Total**
Actual	**Positive**	True Positive (TP)	False Negative (FN)	$TP + FN$
	Negative	False Positive (FP)	True Negative (TN)	$FP + TN$
	Total	$TP + FP$	$FN + TN$	N

Table 2.1: Confusion matrix for binary classification

$$\begin{gathered}
Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \\
Precision = \frac{TP}{TP+FP} \\
Recall = \frac{TP}{TP+FN} \\
F1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN} \\
Sensitivity = Recall = \frac{TP}{TP+FN} \\
Specificity = \frac{TN}{FP+TN}
\end{gathered}$$

Using receiver operating characteristic (ROC) curves is a way to visually diagnosis or assess performance of a classifier and is a plot of the true positive rate (y-axis) against the false positive rate (x-axis). The diagonal identity line is indicative of random guessing and models falling above this line indicate above chance guessing. A perfect model would form an L shape in the top left corner of the graph (True positive rate = 1, False positive rate = 0). ROC area under the curve (ROC AUC, between 0 and 1) is a single value that measures the performance of the model and agrees with accuracy metrics [95]. Intuitively, AUC represents the probability that a randomly chosen disease subject is correctly rated with greater suspicion than a randomly chosen non-diseased patient. The concordance statistic (C-statistic) is another performance metric indicative of a covariate's predictive accuracy or to quantify a single model's predictive discrimination. In practice, it may not be feasible to acquire more data to minimize overfitting (i.e., hypersensitivity to training data heterogeneity). Even if one can acquire more data, this may not resolve overfitting issues (e.g., low signal to noise ratio). One way to evaluate models for potential overfitting is through learning curves. A learning curve is a graphical representation of a chosen performance metric as a function of training and validation samples. highlights the training and validation curves in the presence or absence of overfitting (i.e., presence of variance and bias). In other words, the learning curve plot illustrates the bias and variance of the model (Figure 2.5).

2.5 Results

Based on our search strategy (Fig. 2.1), we have identified three key limitations in existing AI-based AD prediction models: 1) models based on difficult to acquire, invasive and costly data: machine learning has been applied to predict patients that may end up developing AD [96, 41], but the machine learning-based predictive models have included costly [41–44], manually selected, and invasive measures (e.g., cerebrospinal fluid analysis of β-amyloid (Aβ42) [34], Aβ-positron emission tomography [35, 36] which limits the availability of these biomarkers and their usage, 2) models lacking clinical uptake: Further, while identifying high-risk factors is valuable to risk assessment, establishing threshold cut-off values for these risk factors provides clinical utility by outlining informed decision paths (i.e., threshold values to discriminate high vs. low-risk individuals), and 3) complex models: There is still room for improvement in explaining the machine learning decision process in complex machine learning algorithms while maintaining performance. The clinical adoption and end-user trust of

Table 2.2: AD-specific prediction models

Approach	n	Data	Cite	Interpretable?	Year	Metric	Score
TS	162	Multimodal	[97]	Yes	2021	c-index	95%
TS	3777	Cog tests	[98]	Yes	2013	aucroc	85%
TS	1606	Cog tests	[99]	Yes	2013	aucroc	89%
SML	1342	Genetic	[100]	No	2022	acc	92%
SML	168	Imaging	[101]	No	2022	aucroc	90.7%
SML	7703	Accessible	[102]	No	2022	aucroc	84.8%
SML	1048	Multimodal	[103]	Yes	2021	acc	93.95%
SML	678	Accessible	[104]	No	2022	acc	70%
DL	159	Imaging	[105]	No	2021	acc	93.8%
DL	536	Imaging	[106]	No	2021	acc	93.8%
DL	210	Eye-tracking	[107]	No	2022	aucroc	85%
DL	416	Imaging + Cog tests	[108]	No	2021	acc	84.82%

Note : TS - Traditional statistics, SML- Supervised Machine Learning, DL - Deep learning

machine learning models requires intuitive explanation of the decision processes [46]; Yet, interpretable risk prediction models for disease diagnosis are limited [37, 47, 42], as models for diagnosis must strike a balance between performance and interpretable risk prediction. The complexity of high performing machine learning algorithm hinders the formation of clinically intuitive explanations of the decision process, thus impeding clinical adoption.

2.6 Summary

In summary, rapid, accurate, low-cost, easily accessible, non-invasive and early clinical evaluation of AD is critical at this time. There is a need to identify novel clinical features in routinely collected clinical data that identify patients at the highest risk of AD. This will expedite the development of interpretable and explainable diagnostic risk prediction models that can predict long-term AD risk of patients, and potentially expedite research into the intervention initiatives. The current predictive markers highlight the multifaceted nature of risk prediction in AD and suggests the use of indices beyond the current indices to assess AD risk following mild cognitive symptoms, and highlight the importance of personalized, accurate, and rapid patient risk stratification. This scoping review suggest the importance of examining interactions of physical, cognitive, demographic and vascular features for treatment planning and assessing risk stratification for patients and many modifiable risk factors have been identified. Finally, there is a need for predictive models motivated for clinical uptake and to act as a tool in supporting clinical decisions.

Chapter 3

Data sources

Data for model development included retrospectively studied multi site (57+ US and Canadian sites) heterogeneous tabular data (i.e., brain imaging, genetic, neuropsychological testing, lifestyle and health history). Experiment 1 (n=441) classified controls (n= 369) vs AD (n = 72) (Table 3.1), while Experiment 2 (n=738) classified $\mathrm{MCI}_{\mathrm{stable}}$ (n = 444) vs $\mathrm{MCI}_{\mathrm{AD}}$(n = 294) (Table 3.2). ADNI research data (adni.loni.usc.edu) was obtained in accordance with the Declaration of Helsinki and Institutional Review Boards. Informed consent was obtained from all participants. All methods and measurements were performed in accordance with the relevant guidelines and regulations. Study data were de-identified and anonymized prior to transfer to our study database. Variables missing more than 25% of evaluations and were removed from the dataset. One variable was removed due to high collinearly (95% threshold). Model development relied only on baseline data in addition to follow up AD outcome data. Overall, 43 features were considered for experiment 1 (Figure 3.1), and 25 features (Figure 3.2) were considered for experiment 2. The referenced tables and violin plots highlight the distributions of all the features for healthy and disease groups across both experiments.

Table 3.1: Experiment 1 Data. Data were tested for normality using the Kolmogorov–Smirnov test. Continuous data are described by the median and quartiles, as all continuous variables did not fit a normal distribution (alpha = 0.05, and p <0.05). Categorical variables are described by percentages. The Mann-Whitney U test was used to compare continuous data, while the chi-square or Fisher exact tests were used to compare categorical variables.

	Grouped by Diagnosis			
	Overall	Clinically Normal	AD	P-Value
Demographics, n	441	369	72	
Age, median [Q1,Q3]	69.3 [65.8,74.8]	68.5 [65.6,73.5]	74.7 [69.0,79.6]	<0.001
Gender, n (%)				
Female	263 (59.6)	235 (63.7)	28 (38.9)	<0.001
Male	178 (40.4)	134 (36.3)	44 (61.1)	
Race, n (%)				
White/Asian	356 (83.2)	292 (82.0)	64 (88.9)	0.212
Black	72 (16.8)	64 (18.0)	8 (11.1)	
Ethnicity, n (%)				
Not Hisp/Latino	401 (90.9)	332 (90.0)	69 (95.8)	0.174
Hisp/Latino	40 (9.1)	37 (10.0)	3 (4.2)	
Marital Status, n (%)				
Not Married	115 (26.2)	106 (28.9)	9 (12.5)	0.006
Married	324 (73.8)	261 (71.1)	63 (87.5)	
Years of Education, median [Q1,Q3]	16.0 [15.0,18.0]	16.0 [16.0,18.0]	16.0 [14.0,17.2]	<0.001
Cognitive Measures				
Montreal Cognitive Assessment, median [Q1,Q3]	26.0 [23.0,28.0]	26.0 [24.0,28.0]	17.0 [14.0,20.0]	<0.001
Mini Mental State Exam, median [Q1,Q3]	29.0 [25.0,30.0]	29.0 [29.0,30.0]	23.0 [21.0,24.0]	<0.001
Delayed Total Recall, median [Q1,Q3]	12.0 [9.0,15.0]	13.0 [10.0,15.0]	1.0 [0.0,3.0]	<0.001
Everyday Cognition Questionnaire (Patient) - Total, median [Q1,Q3]	1.3 [1.1,1.6]	1.3 [1.1,1.4]	1.9 [1.4,2.4]	<0.001
Everyday Cognition Questionnaire (Patient) - Divided Attention, median [Q1,Q3]	1.2 [1.0,1.8]	1.2 [1.0,1.8]	2.0 [1.5,2.5]	<0.001
Everyday Cognition Questionnaire (Patient) - Planning, median [Q1,Q3]	1.0 [1.0,1.4]	1.0 [1.0,1.2]	1.6 [1.2,2.2]	<0.001
Everyday Cognition Questionnaire (Patient) - Visuospatial, median [Q1,Q3]	1.0 [1.0,1.3]	1.0 [1.0,1.2]	1.3 [1.0,2.2]	<0.001
Everyday Cognition Questionnaire (Patient) - Memory, median [Q1,Q3]	1.6 [1.2,2.0]	1.5 [1.2,1.9]	2.4 [1.9,3.1]	<0.001
Everyday Cognition Questionnaire (Patient) - Language, median [Q1,Q3]	1.2 [1.1,1.7]	1.2 [1.1,1.6]	1.9 [1.2,2.6]	<0.001
Everyday Cognition Questionnaire (Patient) - Organization, median [Q1,Q3]	1.2 [1.0,1.4]	1.0 [1.0,1.3]	1.7 [1.2,2.2]	<0.001
Everyday Cognition Questionnaire (Study partner) - Memory, median [Q1,Q3]	1.2 [1.0,1.8]	1.1 [1.0,1.4]	3.4 [3.1,3.8]	<0.001
Everyday Cognition Questionnaire (Study partner) - Language, median [Q1,Q3]	1.1 [1.0,1.3]	1.0 [1.0,1.1]	2.7 [2.0,3.1]	<0.001
Everyday Cognition Questionnaire (Study partner) - Divided Attention, median [Q1,Q3]	1.0 [1.0,1.5]	1.0 [1.0,1.2]	3.0 [2.2,3.7]	<0.001
Everyday Cognition Questionnaire (Study partner) - Visuospatial, median [Q1,Q3]	1.0 [1.0,1.2]	1.0 [1.0,1.0]	2.5 [1.8,3.2]	<0.001
Everyday Cognition Questionnaire (Study partner) - Total median [Q1,Q3]	1.1 [1.0,1.4]	1.1 [1.0,1.2]	2.7 [2.4,3.3]	<0.001
Everyday Cognition Questionnaire (Study partner) - Organization, median [Q1,Q3]	1.0 [1.0,1.3]	1.0 [1.0,1.0]	2.8 [2.0,3.7]	<0.001
Everyday Cognition Questionnaire (Study partner) - Planning, median [Q1,Q3]	1.0 [1.0,1.2]	1.0 [1.0,1.0]	2.3 [2.0,3.4]	<0.001
Rey Auditory Verbal Learning Test - Forgetting, median [Q1,Q3]	4.0 [2.0,6.0]	3.0 [1.0,5.0]	5.0 [4.0,6.0]	<0.001
Rey Auditory Verbal Learning Test - Percent Forgetting, median [Q1,Q3]	36.0 [15.4,69.2]	28.6 [13.3,50.0]	100.0 [100.0,100.0]	<0.001
Rey Auditory Verbal Learning Test - Immediate, median [Q1,Q3]	44.0 [34.0,54.0]	47.0 [39.0,56.0]	23.0 [19.0,27.0]	<0.001
Rey Auditory Verbal Learning Test - Learning, median [Q1,Q3]	4.0 [2.0,6.0]	3.0 [2.0,5.0]	5.0 [4.0,6.0]	<0.001
Modified Preclinical Alzheimer Cognitive Composite - Trails, median [Q1,Q3]	0.3 [-6.4,1.9]	1.3 [-0.5,2.6]	-13.3 [-15.8,-11.6]	<0.001
Modified Preclinical Alzheimer Cognitive Composite - Digit, median [Q1,Q3]	0.3 [-6.9,2.0]	0.9 [-0.7,2.7]	-16.5 [-17.8,-13.8]	<0.001
Alzheimers Disease Assessment Scale (13 items), median [Q1,Q3]	9.0 [5.7,14.0]	7.3 [5.0,11.0]	29.5 [24.2,35.3]	<0.001
Alzheimers Disease Assessment Scale (11 items), median [Q1,Q3]	6.0 [3.7,11.7]	4.3 [3.3,7.0]	17.2 [14.4,21.8]	<0.001
Task 4 of ADAS11, median [Q1,Q3]	3.0 [1.0,7.0]	2.0 [1.0,4.0]	9.0 [8.0,10.0]	<0.001
Trail Making Test Part B Time, median [Q1,Q3]	71.0 [54.2,98.0]	67.0 [52.0,86.0]	180.0 [100.5,300.0]	<0.001
Brain Imaging				
Intracranial Volume, median [Q1,Q3]	1444730.0 [1333600.0,1559175.0]	1438330.0 [1334740.0,1554250.0]	1450070.0 [1330005.0,1560957.5]	0.829
Whole Brain Volume, median [Q1,Q3]	1034000.0 [962302.0,1130290.0]	1049365.0 [971227.8,1135525.0]	978824.0 [883139.0,1062370.0]	<0.001
Ventricle Volume, median [Q1,Q3]	28216.8 [20749.7,44062.0]	25886.4 [19008.1,38448.6]	48664.9 [33408.2,62852.5]	<0.001
Middle Temporal Gyrus Volume, median [Q1,Q3]	20609.0 [18603.5,22929.0]	20962.0 [19260.8,23218.5]	18211.5 [16149.2,19730.8]	<0.001
Entorhinal Volume , median [Q1,Q3]	4015.0 [3590.5,4518.8]	4087.0 [3708.2,4579.2]	3222.5 [2559.0,3911.0]	<0.001
Fusiform Volume, median [Q1,Q3]	18305.0 [16710.0,20324.0]	19282.0 [17296.0,20920.0]	16219.0 [14809.0,17816.2]	<0.001
Hippocampus Volume, median [Q1,Q3]	7471.6 [6770.6,8067.4]	7610.2 [7050.5,8182.4]	5748.3 [5057.3,6355.1]	<0.001
Genetic				
APOE4, median [Q1,Q3]	0.0 [0.0,1.0]	0.0 [0.0,1.0]	1.0 [0.0,1.0]	<0.001

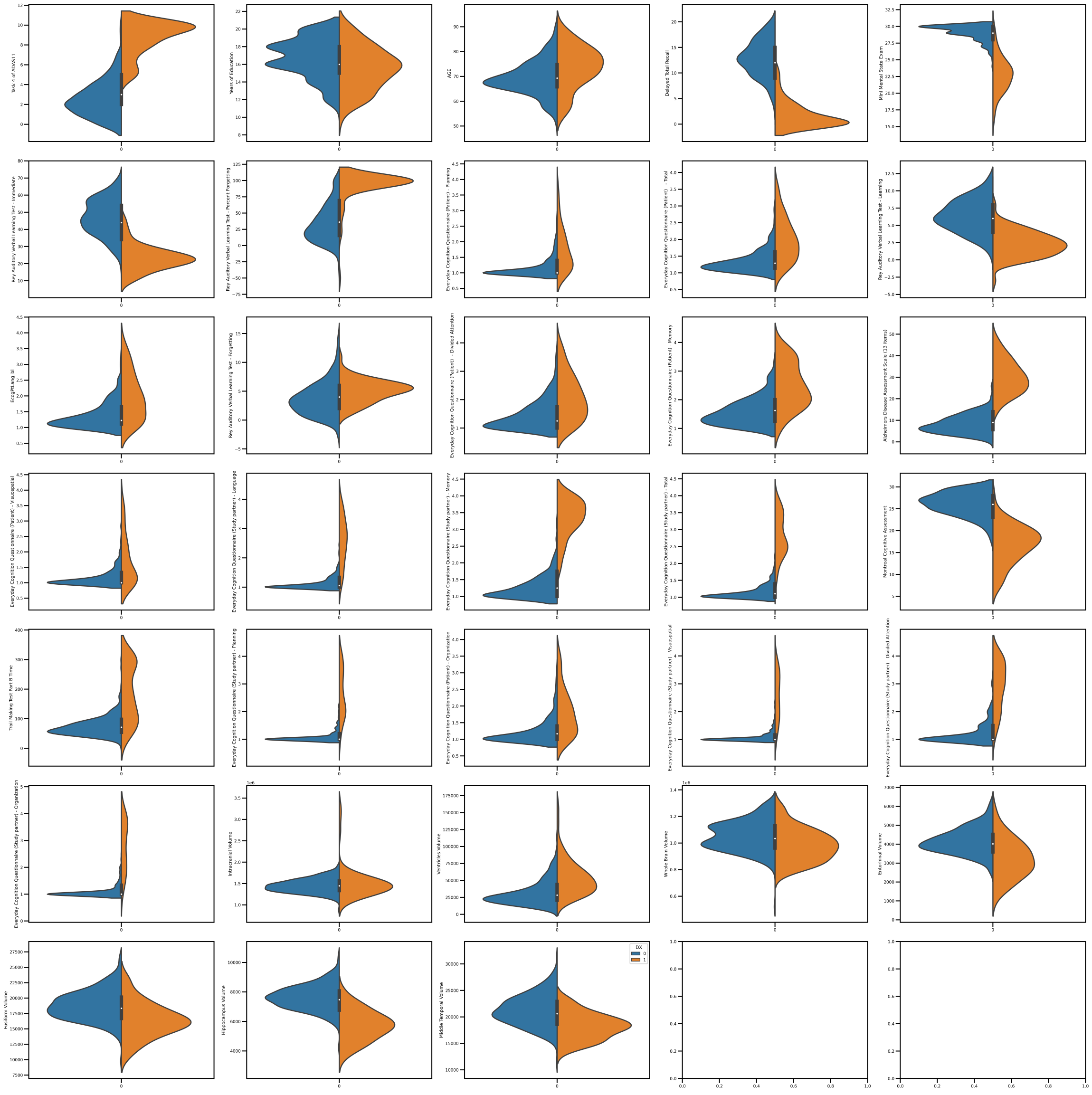
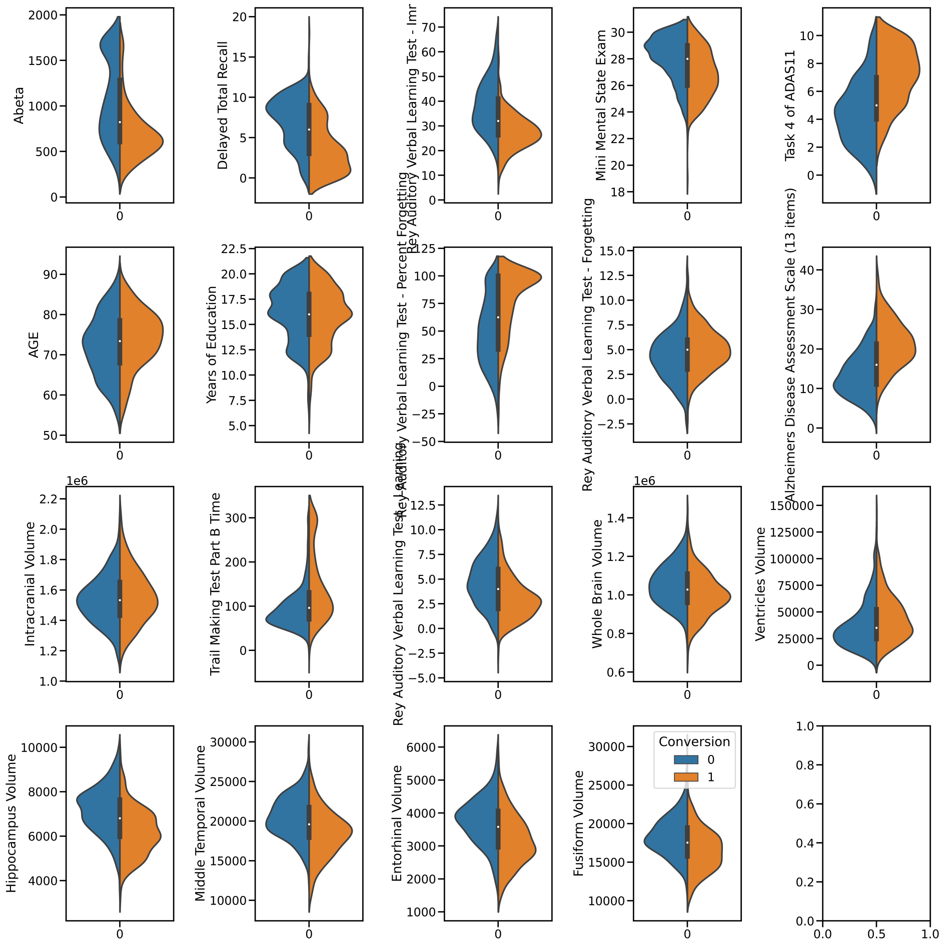


Figure 3.1: Violin plots highlighting distributions of numerical variables in experiment 1 hued on diagnosis (CN/AD)

Figure 3.2: Violin plots highlighting distributions of numerical variables in experiment 2 hued on conversion (stable/AD)



	Grouped by Conversion			
	Overall	MCI^{stable}	MCI^{AD}	P-Value
Demographics, n	738	444	294	
Age, median [Q1,Q3]	73.4 [67.9,78.5]	72.4 [66.5,77.9]	74.5 [69.6,79.1]	0.001
Gender, n (%)				
Female	298 (40.4)	177 (39.9)	121 (41.2)	0.784
Male	440 (59.6)	267 (60.1)	173 (58.8)	
Race, n (%)				
White/Asian/Other	710 (97.5)	425 (97.7)	285 (97.3)	0.901
Black	18 (2.5)	10 (2.3)	8 (2.7)	
Ethnicity, n (%)				
Not Hisp/Latino	713 (97.0)	428 (96.8)	285 (97.3)	0.905
Hisp/Latino	22 (3.0)	14 (3.2)	8 (2.7)	
Marital Status, n (%)				
Not Married	153 (20.8)	99 (22.4)	54 (18.4)	0.214
Married	582 (79.2)	342 (77.6)	240 (81.6)	
Years of Education, median [Q1,Q3]	16.0 [14.0,18.0]	16.0 [14.0,18.0]	16.0 [14.0,18.0]	0.085
Cognitive Measures				
Mini Mental State Exam, median [Q1,Q3]	28.0 [26.0,29.0]	28.0 [27.0,29.0]	27.0 [26.0,28.0]	<0.001
Delayed Total Recall, median [Q1,Q3]	6.0 [3.0,9.0]	8.0 [5.0,9.0]	3.0 [1.0,6.0]	<0.001
Rey Auditory Verbal Learning Test - Forgetting, median [Q1,Q3]	5.0 [3.0,6.0]	4.0 [3.0,6.0]	5.0 [4.0,6.0]	<0.001
Rey Auditory Verbal Learning Test - Percent Forgetting, median [Q1,Q3]	62.5 [33.3,100.0]	50.0 [25.0,73.3]	87.5 [60.0,100.0]	¡0.001
Rey Auditory Verbal Learning Test - Immediate, median [Q1,Q3]	32.0 [26.2,41.0]	37.0 [30.0,45.0]	27.0 [23.0,32.0]	<0.001
Rey Auditory Verbal Learning Test - Learning, median [Q1,Q3]	4.0 [2.0,6.0]	5.0 [3.0,7.0]	3.0 [2.0,4.0]	<0.001
Alzheimers Disease Assessment Scale (13 items), median [Q1,Q3]	16.0 [11.0,21.3]	13.0 [9.3,17.0]	21.0 [17.2,25.0]	<0.001
Task 4 of ADAS11, median [Q1,Q3]	5.0 [4.0,7.0]	4.0 [3.0,6.0]	7.0 [6.0,9.0]	<0.001
Trail Making Test Part B Time, median [Q1,Q3]	96.0 [70.2,131.0]	83.5 [65.0,112.8]	120.0 [85.0,194.8]	<0.001
Brain Imaging				
Intracranial Volume, median [Q1,Q3]	1533305.0 [1432292.5,1651465.0]	1527450.0 [1431647.5,1636420.0]	1542475.0 [1436612.5,1671352.5]	0.217
Whole Brain Volume, median [Q1,Q3]	1026515.0 [958550.5,1116612.5]	1074340.0 [969606.0,1139420.0]	970229.0 [904029.0,1075140.0]	0.007
Ventricle Volume, median [Q1,Q3]	42001.5 [28566.5,55888.8]	32882.0 [22479.0,52877.0]	48112.0 [39494.0,58946.0]	0.020
Middle Temporal Gyrus Volume, median [Q1,Q3]	18848.5 [16957.5,21570.0]	19393.5 [18706.0,23108.2]	17021.0 [15046.2,18222.5]	<0.001
Entorhinal Volume , median [Q1,Q3]	3311.5 [2776.0,3777.0]	3584.0 [3167.5,4004.5]	2891.5 [2454.8,3418.2]	0.001
Fusiform Volume, median [Q1,Q3]	16922.0 [15265.8,19518.5]	17826.0 [15784.5,20555.8]	15838.0 [13265.8,17248.5]	0.008
Hippocampus Volume, median [Q1,Q3]	6400.0 [5772.5,7235.5]	6630.0 [5998.0,7712.0]	5753.0 [4925.5,6941.0]	0.007
Genetic				
APOE4, median [Q1,Q3]	0.0 [0.0,1.0]	0.0 [0.0,1.0]	1.0 [0.0,1.0]	0.309

Table 3.2: Experiment 2 Data. Data were tested for normality using the Kolmogorov–Smirnov test. Continuous data are described by the median and quartiles, as all continuous variables did not fit a normal distribution (alpha = 0.05, and p <0.05). Categorical variables are described by percentages. The Mann-Whitney U test was used to compare continuous data, while the chi-square or Fisher exact tests were used to compare categorical variables.

Chapter 4

Machine Learning Process

Herein, we highlight the machine learning process (Fig. 4.1) for both experiments that begins with the model and feature selection and ends with model evaluation and interpretation.

We implemented a supervised extreme gradient boosting (XGBoost) classifier, a more regularized form of the stochastic gradient boosting ensemble algorithm, outperforming its predecessor in performance, scalability, and efficiency[109]. We theoretically (Table 4.1) and empirically ([45, 110] motivate our model selection. XGBoost combines decision trees sequentially, and each new tree in sequence corrects the errors of the previous tree to minimize the objective function in Eq. 4.0.1 . Because of XGBoost's iterative decision tree-based architecture, XGBoost can be highly interpretable.

Table 4.1: Motivating model selection

Models	**Scaling**	**Bias-Variance**	**Limitations**
Linear models	C(d)	↑ Bias, ↓ Variance	Plateaus in N >> M
Decision Trees	C(N)	↓ Bias, ↑ Variance	Deep trees lack generalization
KNN	C(1/Nd)	↑ Bias, ↓ Variance	Curse of dimensionality
Naïve Bayes	C(Nd)	↑ Bias, ↓ Variance	complex/non-linear interactions
Random forest	$C(N^2)$	↓ Bias, ↑ Variance	Homogeneous trees
XGBoost	C(N)	↑ Bias, ↓ Variance, boosting ↓ Bias	Heterogenous trees

The XGBoost algorithm optimizes an objective function:

$$Obj = Loss + \Omega \tag{4.0.1}$$

For binary classification tasks, the loss function (Loss) often uses the log loss function:

$$Loss = \sum_i [y_i \log(1 + e^{-y_{pred_i}}) + (1 - y_i) \log(1 + e^{y_{pred_i}})] \tag{4.0.2}$$

The regularization term (Ω) helps prevent overfitting:

$$\Omega = \gamma * T + 0.5 * \lambda * \sum_{j=1}^{T} w_j^2 \tag{4.0.3}$$

Using Taylor expansion and fixed-term simplifications, the approximate objective function at the t-th iteration is:

$$Obj(t) \approx \sum_{i=1}^{n} [g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \tag{4.0.4}$$

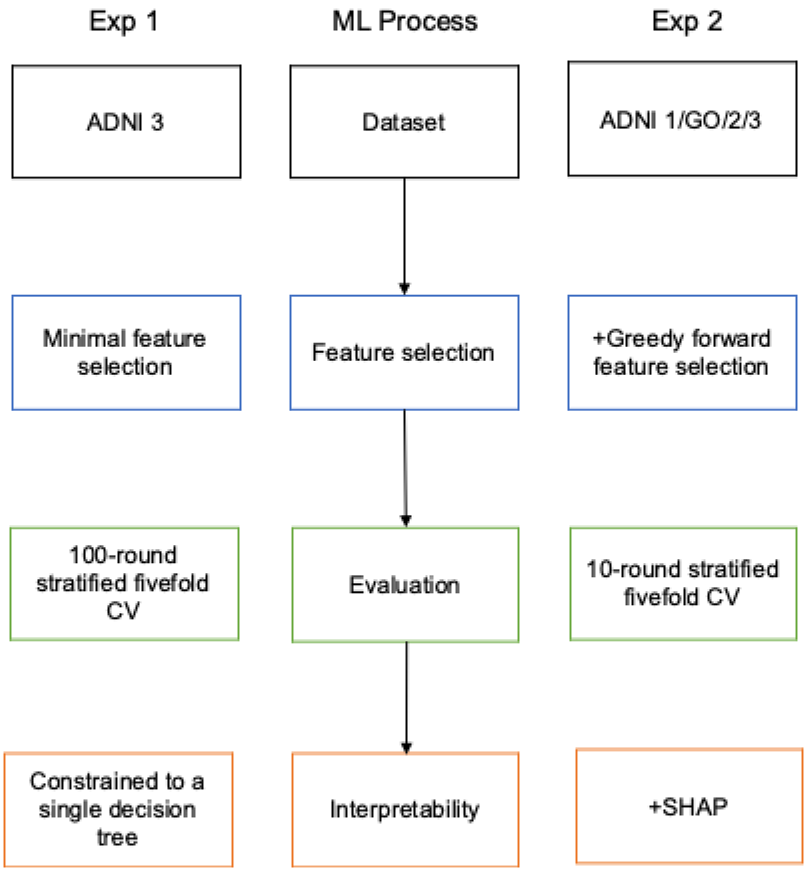
Gradients (g_i) and Hessians (h_i) of the loss function are calculated as:

$$g_i = \frac{\partial Loss}{\partial y_{pred_i}}, \quad h_i = \frac{\partial^2 Loss}{\partial y_{pred_i}^2} \tag{4.0.5}$$

The algorithm selects the best feature split using a greedy method, choosing the split with the highest gain:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{4.0.6}$$

Figure 4.1: Machine Learning Process


Exp 1
ML Process
Exp 2
ADNI 3
Dataset
ADNI 1/GO/2/3
Minimal feature selection
Feature selection
+Greedy forward feature selection
100-round stratified fivefold CV
Evaluation
10-round stratified fivefold CV
Constrained to a single decision tree
Interpretability
+SHAP

4.1 Feature selection

Feature selection allows us to reduce the dimensionality and complexity of the data and model, respectively. We begin the machine learning process by elucidating which features were important in AD risk. We first split the training dataset into 70% training and 30% test subsets. The average feature importance according to XGBoost was calculated based on each feature's proportionate contribution to the model, computed for each tree in the model. Specifically, when comparing proportionate contribution across features, a higher gain value is indicative of greater predictive performance and thus greater feature importance, and this process was repeated 100 times with different random number seeds (0-99) to ensure statistical stability. Thereafter, we selected the top 10 ranked features (Fig. 5.1).

To include the fewest features while maintaining performance, we initialized a baseline ROC-AUC score of 0, then trained, predicted and calculated the new ROC-AUC score using the XGBoost model with the top ranked feature only. Thereafter, we added both the first and second ranked feature, repeated the training, predicting, and calculating process to acquire a new ROC-AUC score. If the ROC-AUC score of the current selected features (e.g., top two features) showed improvement over the old score (e.g., top feature alone), the new feature was kept (i.e., second ranked feature), otherwise, the process stopped at the last added feature (i.e., top feature alone); Table 5.2). The models' performances were evaluated with fivefold stratified cross-validation for 100 iterations. This approach was utilized for both experiments. For experiment 2, we employed an additional greedy feature selection algorithm that evaluates the classifier's performance based on the best subset of features.

4.2 Model Evaluation

The simplified decision route XGBoost model (i.e., constrained to a single decision tree) was compared to that of the default XGBoost (i.e., trained with the default number of decision trees)(Figures 5.3 and 5.10). Fivefold stratified cross-validation for 100 iterations was used to decrease overly optimistic outcomes and improve statistical stability. To test model performance, we estimated the area under the curve (AUC) of the receiver operating characteristic (ROC) and precision-recall (PR) curves, as well as classification reports and confusion matrices. Plotting the true positive rate (TPR) against the false positive rate (FPR) at various thresholds (0-1) creates the ROC curves. This value indicates how well the model classifies patients who end up developing AD. Values closer to 1 indicate better classification performance. The PR curves show precision and recall for various thresholds (0 to 1). High recall and precision result in a high area under the curve score, with high precision indicating low false AD and high recall indicating low false AD. Each AUC value is computed as the mean over 100 iterations of fivefold stratified cross-validation.

4.3 Model interpretation

Understanding how a model makes a decision is important to establish trust for end users (e.g., clinicians). However, many complex models are not easily transparent. XGBoost constructs decision trees recursively from pseudo residuals and can develop decision trees that contribute most to a predictive model. If the performance of a decision tree remains high, lowering the model's complexity might yield a transparent decision path. To identify the decision path, we reduced the complexity of XGBoost by reducing the number of trees from 150 (default XGBoost) to 1 (Interpretable-XGBoost) by randomly splitting the available patients into training and validation datasets (70:30) and limiting the number of estimators to 1.

In experiment 2, we added SHAP explainability. SHAP (Shapley ADitive exPlanations) [84] is a model-agnostic, game-theoretic approach to explaining machine learning outputs, which allocates the value of each feature for a specific prediction and has been successfully applied to clinic [48]. SHAP plots provide an intuitive explanation of what led to the patients' risk. The SHAP Python package can be found here: https://github.com/slundberg/shap. More recently, SHAP methods have been applied to tree based algorithms (e.g., random forests, gradient boosted) [45]. SHAP summary plots show feature's impact and importance. Each point represents a unique occurrence of the dataset (i.e., a single patient). Their location along the x-axis (i.e., SHAP value) shows that feature's influence on the model's output for that patient. Higher SHAP values are associated with a higher risk. A feature's importance is determined by its average absolute Shapley values and sorted along the y axis (a higher position equates to a greater importance).

4.4 Integration test

Here we integrate our code with a public dataset to ensure that it processes the data correctly and produces the expected output.

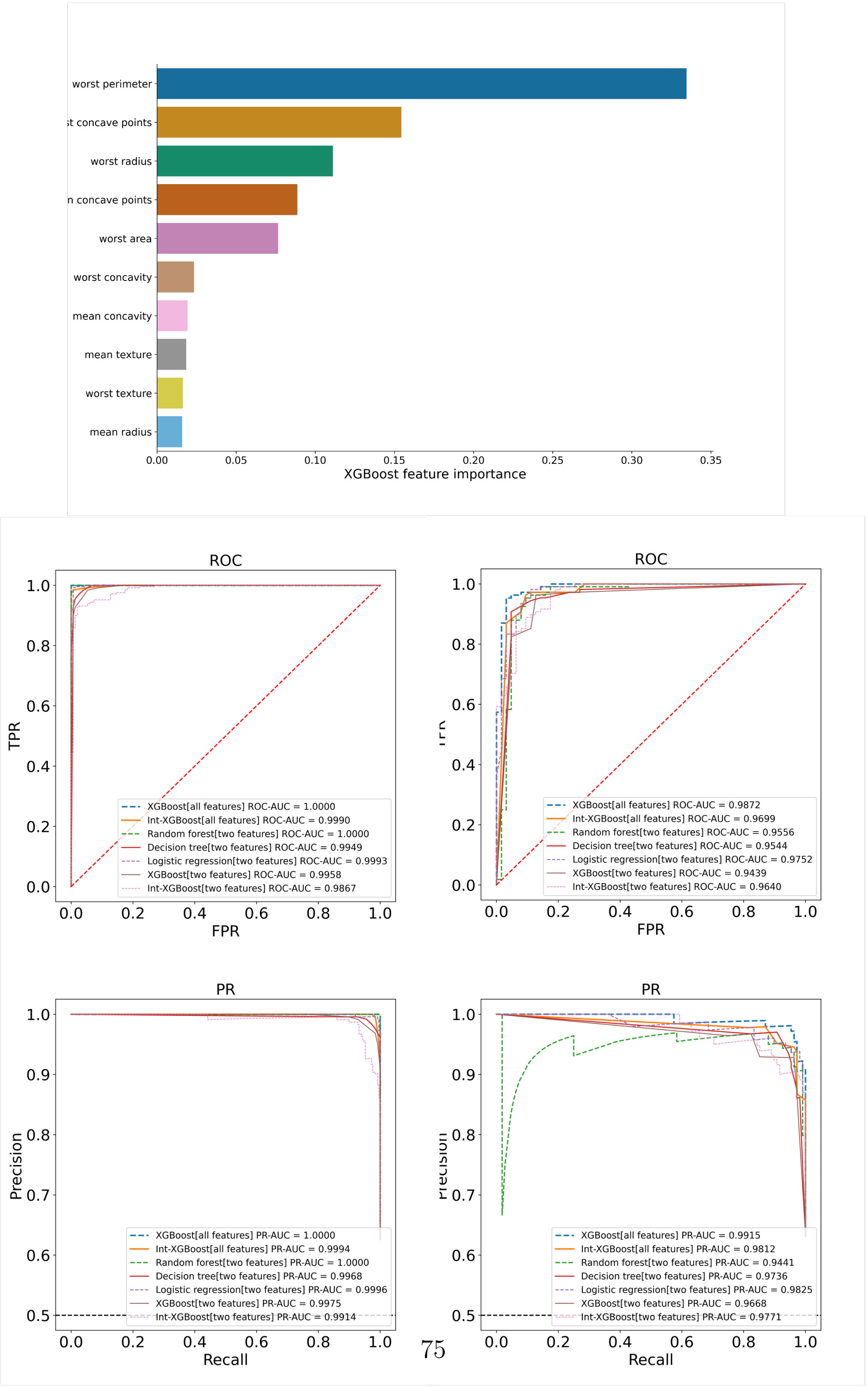


Figure 4.2: Integration test with public breast cancer data set

Chapter 5

Results

5.1 Experiment 1

Baseline patient characteristics and significant associations with developing AD

Compared to clinically normal controls, AD patients were older, male, married, had lower education, performed worse on all cognitive tests, were genetically predisposed, and had lower whole brain volume, middle temporal gyrus volume, entorhinal volume, hippocampus volume, fusiform volume and larger ventricle volume (Table **??**).

XGBoost classifiers' performance

We implemented a supervised XGBoost classifier with the objective of identifying high risk factors and patients at the highest risk of AD. Machine learning tools selected three features for experiment 1 (Table 5.2). Using the selected features, the XGBoost performance model achieved cross-validation accuracy of 98%(Table 5.2).

Table 5.1: Experiment 1 - XGBoost classification in differentiating between clinically normal and AD patients using 100-repeated fivefold cross-validation

Top Features	Training	Validation	Improvement
Everyday Cognition Questionnaire (Study partner) - Total	0.973 ± 0.007	0.948 ± 0.037	
Alzheimers Disease Assessment Scale (13 items)	0.997 ± 0.004	0.968 ± 0.029	.02
Delayed Total Recall	0.995 ± 0.004	0.977 ± 0.025	.009
Everyday Cognition Questionnaire (Study partner) - Memory	0.995 ± 0.004	0.974 ± 0.027	No improvement

Note : Mean % AUCROC scores +/- SD

The interpretable model maintained performance while improving on interpretability (Figures 5.3. The models can reliably predict patient diagnosis.

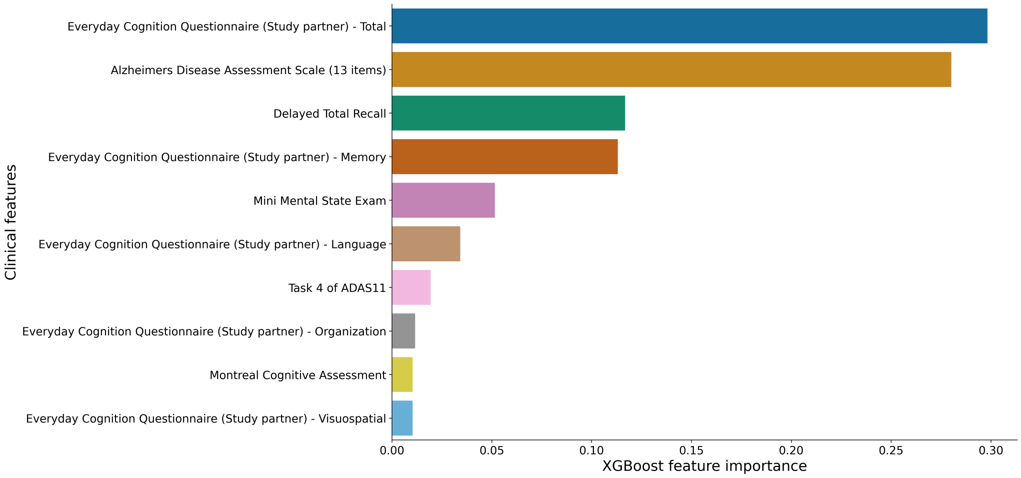


Figure 5.1: Exp 1: Top 10 ranked features by XGBoost according to their importance. Average feature importance accumulated from the XGBoost model, using random splits in a ratio of 7:3, over 100 repetitions with varied number seeds (0-99).

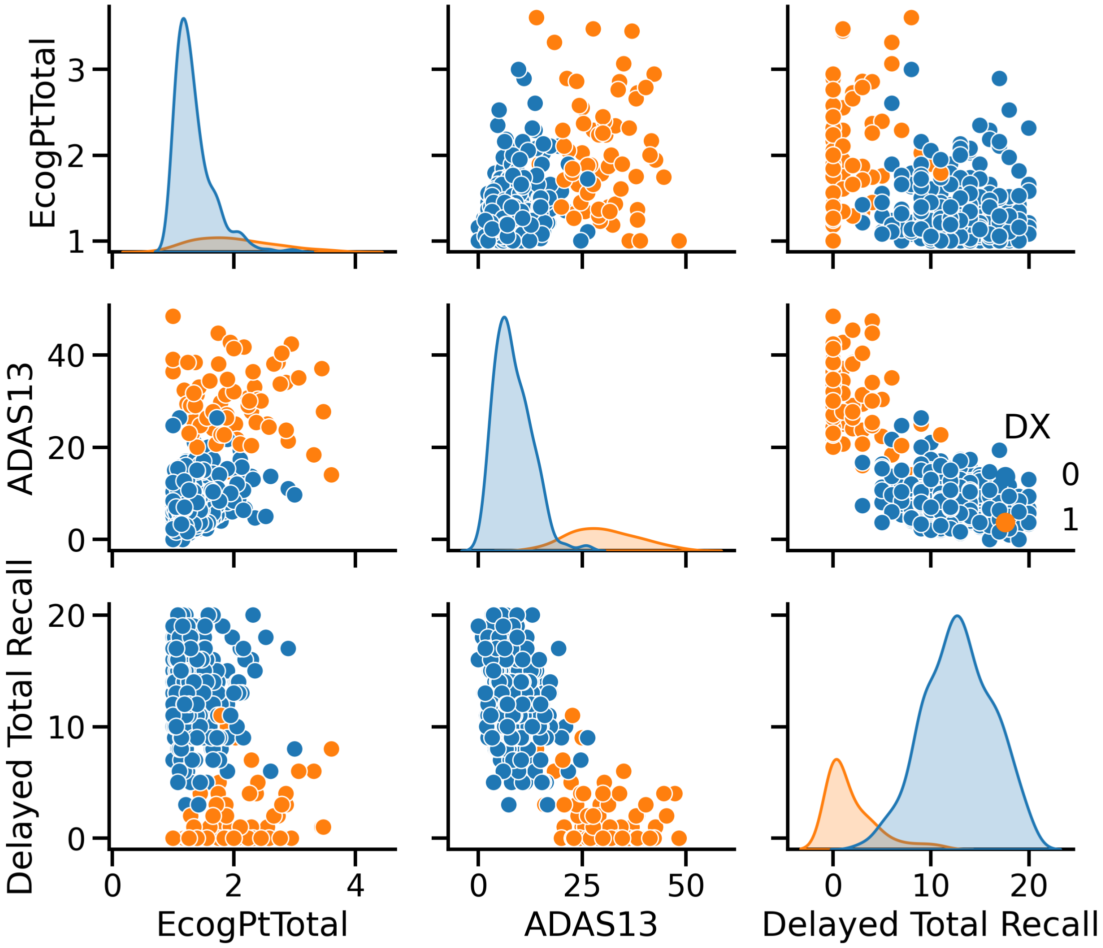


Figure 5.2: Pairplot for selected features in experiment 1 hued by diagnosis (0: Clinically Normal, 1: AD)

.

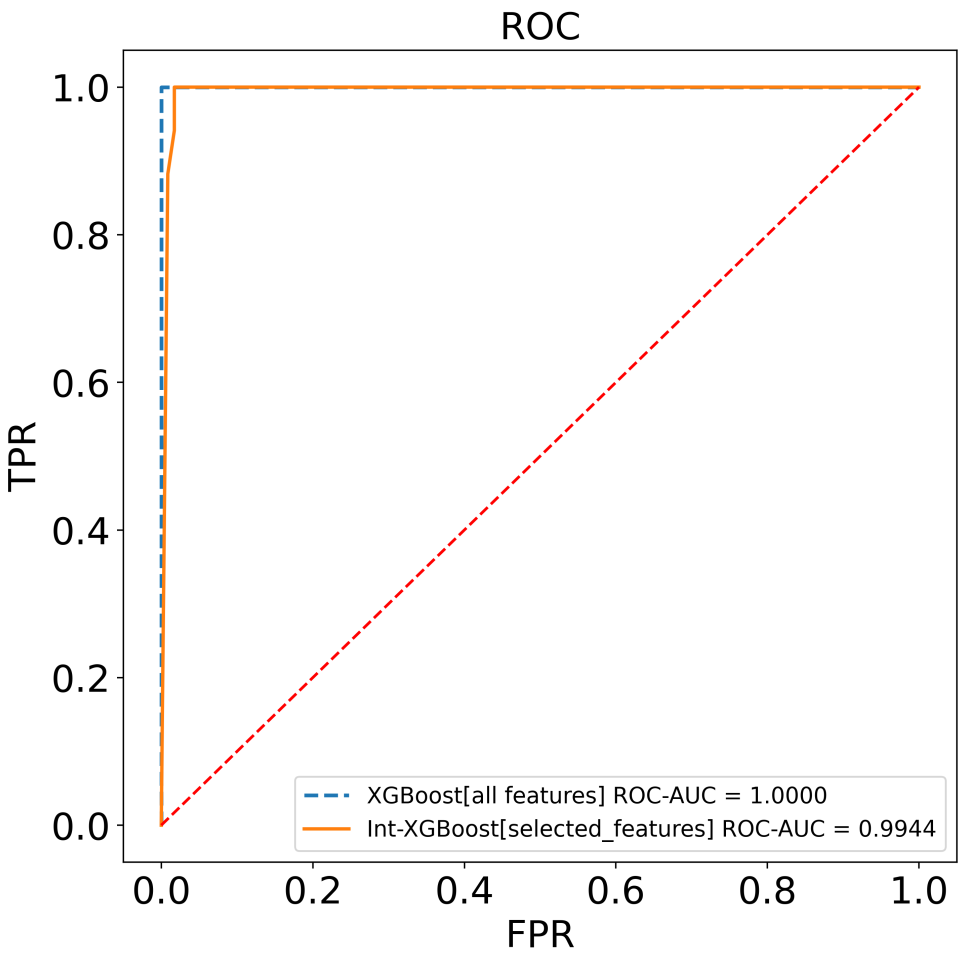


Figure 5.3: Area under the curve (0 - 1) of the receiver operating curves (ROC) for validation comparing XGBoost performance and interpretable models with all/selected features. The ROCs are created by plotting the true positive rate (TPR) against the false positive rate (FPR) for different thresholds (0 - 1). This value represents how good the model is at distinguishing between clinically normal and patients that end up developing AD, with values closer to 1 indicative of better classification performance.

5.1.1 Interpretability

Here we couple the identified, noninvasive and accessible clinical features, together with a decision route, to diagnose patients with Alzheimer's disease (Fig. 5.6)

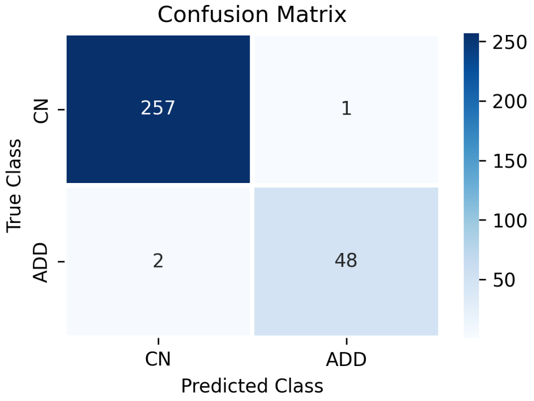


Figure 5.4: Confusion matrix for validation set

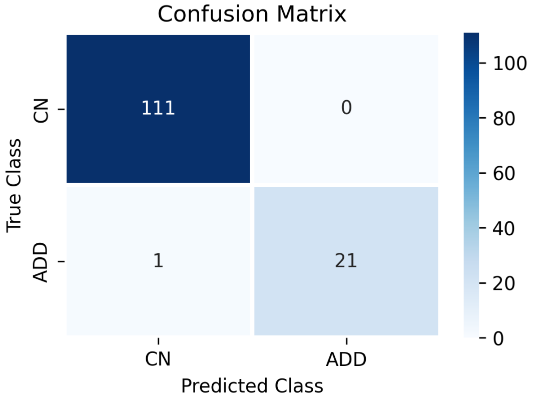


Figure 5.5: Confusion matrix for training set

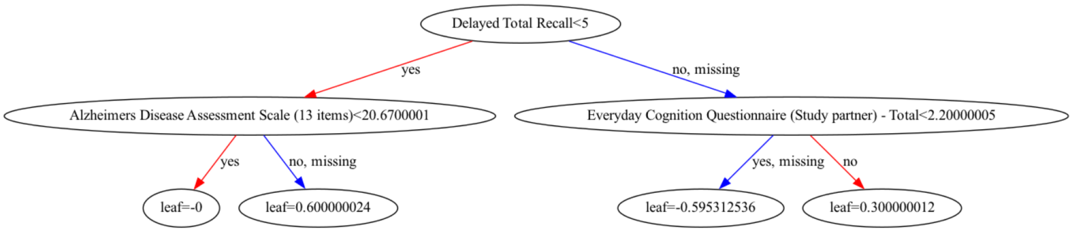

Delayed Total Recall<5
yes
no, missing
Alzheimers Disease Assessment Scale (13 items)<20.6700001
Everyday Cognition Questionnaire (Study partner) - Total<2.20000005
yes
no, missing
yes, missing
no
leaf=-0
leaf=0.600000024
leaf=-0.595312536
leaf=0.300000012

Figure 5.6: Data-driven decision route

5.2 Experiment 2

5.2.1 Baseline patient characteristics and significant associations with MCI-AD progression

In experiment 2 and compared to those that remained stable at mild cognitive impairment, those that converted to AD were older, scored higher on the Alzheimer's Disease Assessment Scale (13 items), had reduced whole brain, middle temporal gyrus, entorhinal, fusiform, and hippocampus volume. Moreover, stablized MCI patients performed better on the Rey Auditory Verbal Learning Test in the cagegories for learning, forgetting, and immediate (Table 3.2).

5.2.2 XGBoost classifiers' performance

We implemented a supervised XGBoost classifier with the objective of identifying high risk factors and patients at the highest risk of converting from MCI to AD. By combining genetic, cognitive evaluation, demographic, and brain imaging, the algorithm identified the primary indicators of MCI-to-AD progression with ROC AUC values over 87% (Table 5.3. The model can reliably predict patient prognosis. We've improved the performance limitations of the minimal feature selection method (Table **??** and Fig. 5.7 by comprehensively testing subsets of all features using a greedy sequential feature selection algorithm (Table 5.3 and Fig 5.8, 5.8) . We've also shown that we can improve interpretability while maintaining performance (Fig. 5.10).

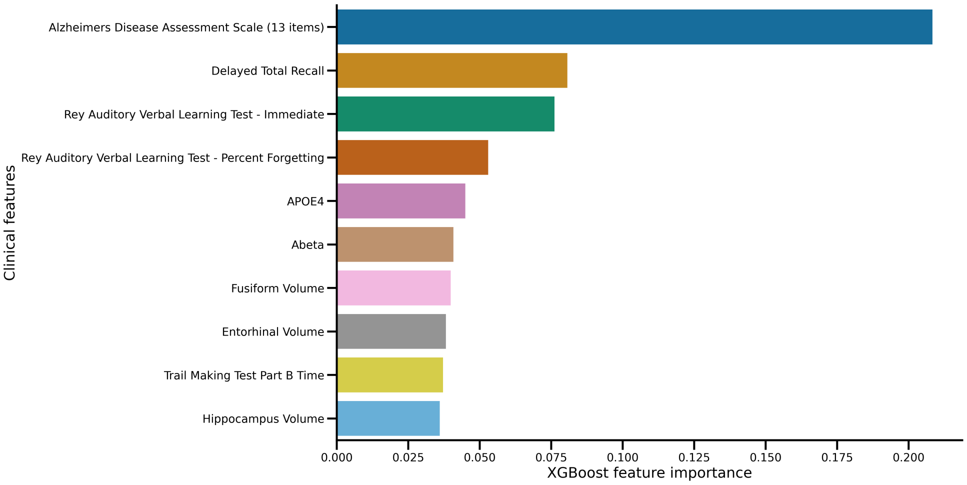


Figure 5.7: Exp 2: Top 10 ranked features by XGBoost according to their importance. Average feature importance accumulated from the XGBoost model, using random splits in a ratio of 7:3, over 100 repetitions with varied number seeds (0-99).

Table 5.2: Experiment 2 - XGBoost classification in differentiating between MCI_{stable} and MC_{AD} patients using 100-repeated fivefold cross-validation.

Top Features	**Training**	**Validation**	**Improvement**
Alzheimers Disease Assessment Scale (13 items)	0.754 ± 0.011	0.670 ± 0.035	
Delayed Total Recall	0.791 ± 0.010	0.719 ± 0.032	0.049
Rey Auditory Verbal Learning Test - Immediate	0.958 ± 0.007	0.707 ± 0.034	No improvement

Note : Mean % AUCROC scores +/- SD

5.2.3 Interpretability

In Figure 5.13, the model's prediction of higher risk is driven by a combination of factors: higher genetic predisposition, lower Verbal Learning - Immediate Test scores, lower hippocampus volume, and longer Trail Making Test Part B tim, followed by delayed total recall, years of education, ethnicity, and race. Furthermore, a longer Trail Making Test Part B time points to difficulties in cognitive processing, particularly in executive function tasks that require task switching.

Table 5.3: Experiment 2 - XGBoost classification in differentiating between $\mathrm{MCI}_{\mathrm{stable}}$ and $\mathrm{MC}_{\mathrm{AD}}$ patients using 10-repeated fivefold cross-validation and a greedy forward feature selection algorithm.

	Feature_subsets	AUC-ROC_score
1	Task 4 of ADAS11	0.788 ± 0.03
2	Delayed Total Recall, Task 4 of ADAS11	0.817 ± 0.03
3	Delayed Total Recall, Task 4 of ADAS11, APOE4	0.792 ± 0.02
4	Delayed Total Recall, Task 4 of ADAS11, APOE4, Hippocampus Volume	0.803 ± 0.03
5	Delayed Total Recall, Rey Auditory Verbal Learning Test - Immediate, Task 4 of ADAS11, APOE4, Hippocampus Volume	0.823 ± 0.03
6	Delayed Total Recall, Rey Auditory Verbal Learning Test - Immediate, Task 4 of ADAS11, APOE4, Trail Making Test Part B Time, Hippocampus Volume	0.838 ± 0.03
7	Delayed Total Recall, Rey Auditory Verbal Learning Test - Immediate, Task 4 of ADAS11, Rey Auditory Verbal Learning Test - Forgetting, APOE4, Trail Making Test Part B Time, Hippocampus Volume	0.854 ± 0.02
8	Delayed Total Recall, Rey Auditory Verbal Learning Test - Immediate, Task 4 of ADAS11, Years of Education, Rey Auditory Verbal Learning Test - Forgetting, APOE4, Trail Making Test Part B Time, Hippocampus Volume	0.860 ± 0.02
9	Delayed Total Recall, Rey Auditory Verbal Learning Test - Immediate, Task 4 of ADAS11, Years of Education, Rey Auditory Verbal Learning Test - Forgetting, APOE4, Trail Making Test Part B Time, Hippocampus Volume, Fusiform Volume	0.863 ± 0.02
10	Delayed Total Recall, Rey Auditory Verbal Learning Test - Immediate, Task 4 of ADAS11, Race, Years of Education, Rey Auditory Verbal Learning Test - Forgetting, APOE4, Trail Making Test Part B Time, Hippocampus Volume, Fusiform Volume	0.867 ± 0.03
11	Delayed Total Recall, Rey Auditory Verbal Learning Test - Immediate, Task 4 of ADAS11, Race, Years of Education, Ethnicity, Rey Auditory Verbal Learning Test - Forgetting, APOE4, Trail Making Test Part B Time, Hippocampus Volume, Fusiform Volume	0.870 ± 0.03

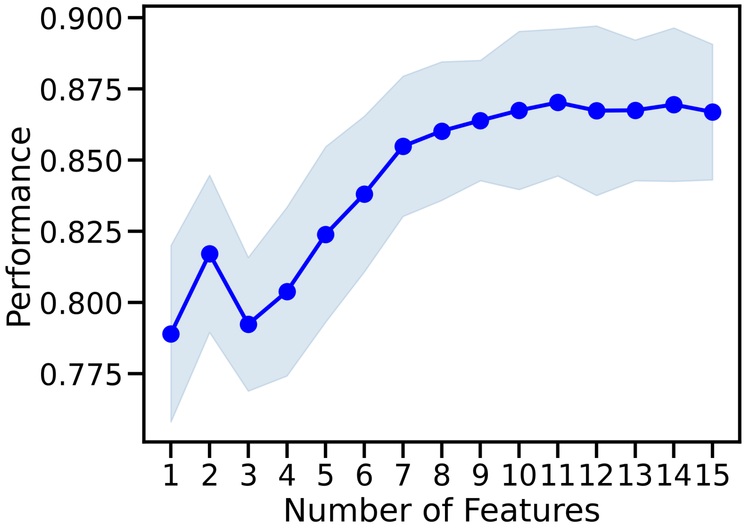


Figure 5.8: Performance as a function of number of features

.

Table 5.4: Exp 2 - Performance metrics of the interpretable model on validation set

	Precision	Recall	F1-score	Support
$\mathrm{MCI}_{\mathrm{stable}}$	89	95	92	134
$\mathrm{MCI}_{\mathrm{AD}}$	91	82	86	88
Accuracy			90	222
Macro avg	90	99	89	222
Weighted avg	90	90	90	222

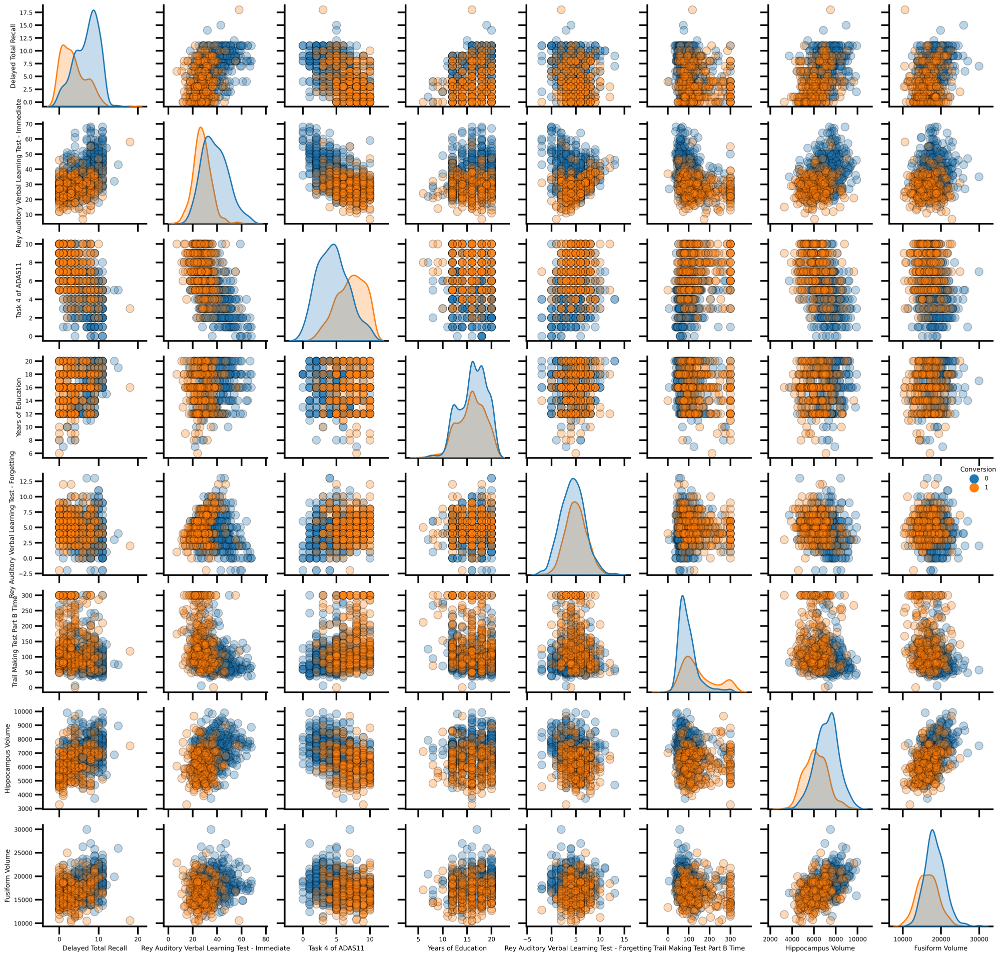


Figure 5.9: Pairplot for selected features in experiment 2 hued by prognosis (0: Stabilized, 1: Converted to AD), highlighting a more challenging classification task (i.e., less divisive features) compared to experiment 1

.

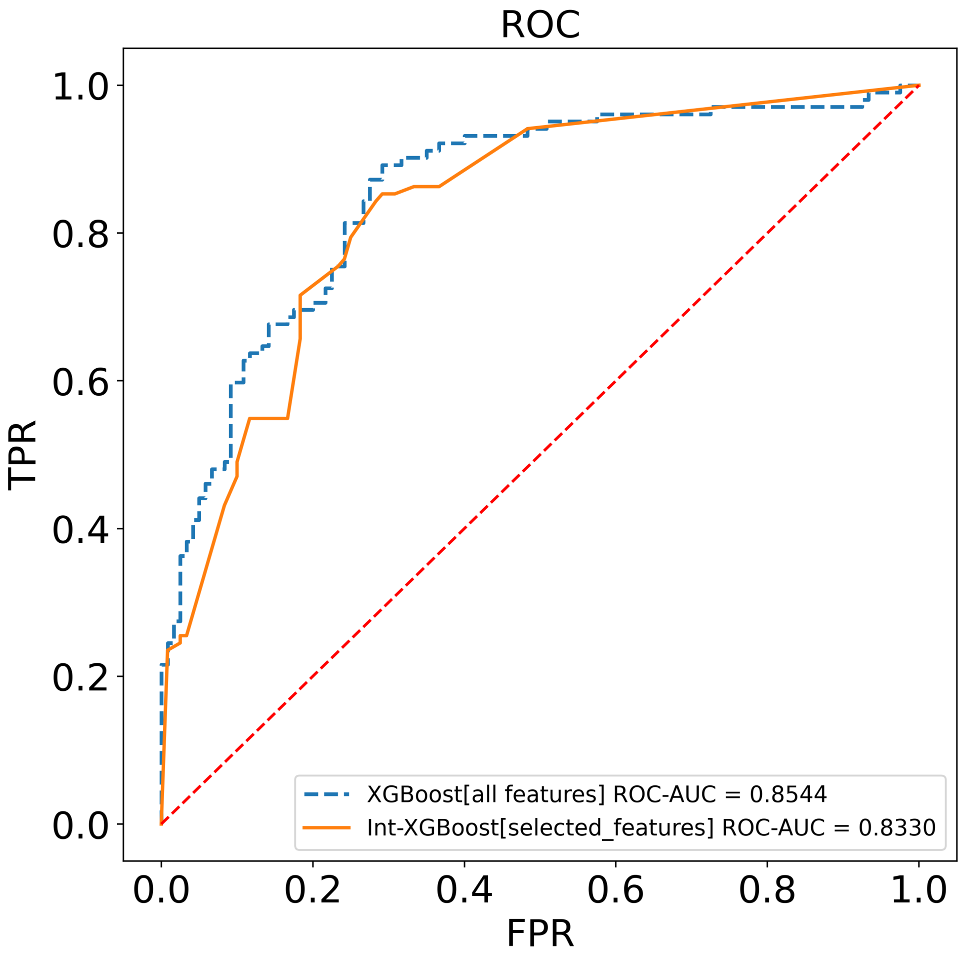


Figure 5.10: Area under the curve (0 - 1) of the receiver operating curves (ROC) for validation comparing XGBoost performance and interpretable models to that of other machine learning approaches with all/selected features. The ROCs are created by plotting the true positive rate (TPR) against the false positive rate (FPR) for different thresholds (0 - 1). This value represents how good the model is at distinguishing between clinically normal and patients that end up developing AD, with values closer to 1 indicative of better classification performance.

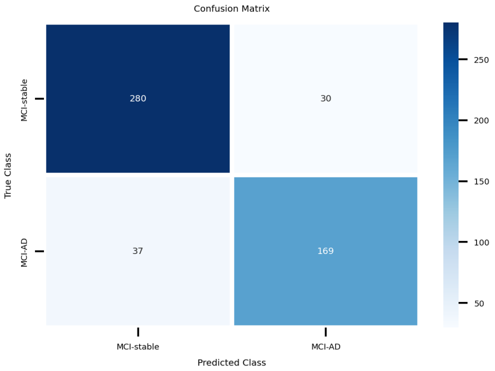


Figure 5.11: Confusion matrix for training set

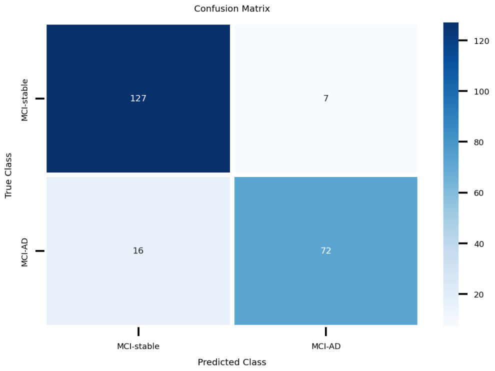


Figure 5.12: Confusion matrix for validation set

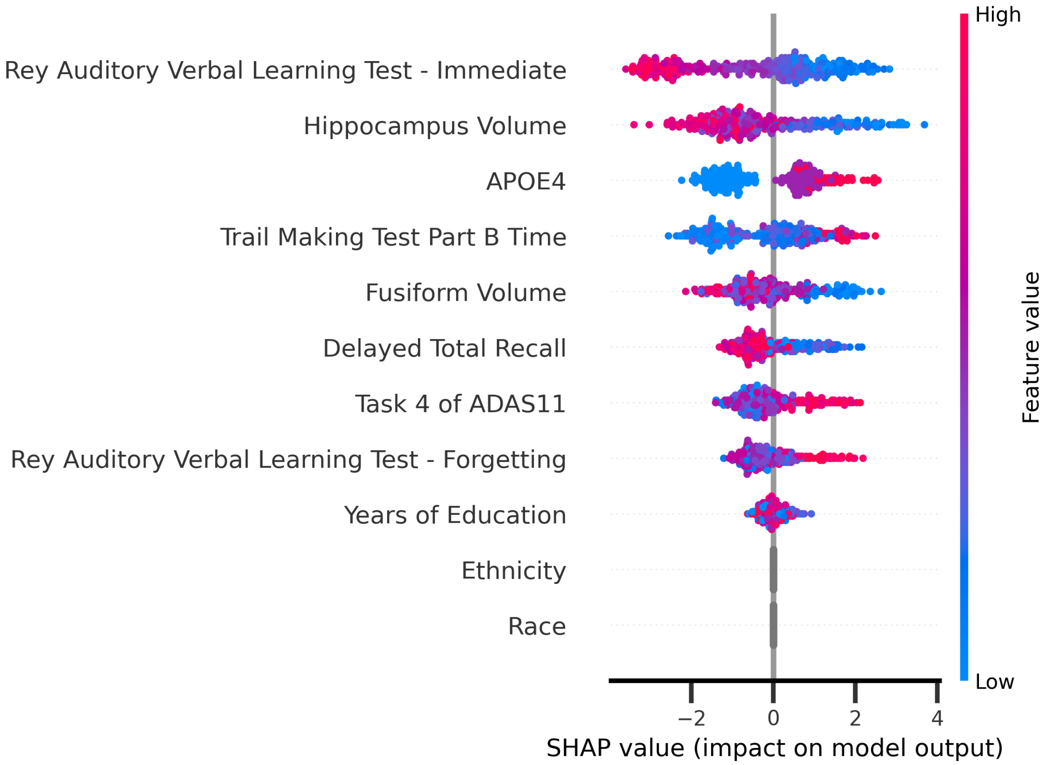


Figure 5.13: SHAP summary plots. SHAP summary plots show feature's impact and importance. Each point represents a unique occurrence of the dataset (i.e., a single patient). Their location along the x-axis (i.e., SHAP value) shows that feature's influence on the model's output for that patient. Higher SHAP values are associated with a higher risk. A feature's importance is determined by its average absolute Shapley values and sorted along the y axis (a higher position equates to a greater importance).

Chapter 6

Discussion

This work is significant in six ways. First, our work extends beyond the identification of high-risk factors for developing AD. It provides an intuitive explanation of the intricate interactions and non-linear relationships among the high-risk factors that lead to the patient's risk. Second, the most prominent predictors are easily accessible and readily acquired without the need of invasive procedures (e.g., lumbar punctures, PET), and can be collected easily. This is particularly useful during a pandemic when healthcare resources are limited. Third, the elucidation of the machine learning decision process and identification of absolute thresholds in the resulting decision tree may increase clinical adoption. Fourth, we further emphasize the ability of XGBoost to preserve performance while increasing its interpretability by constraining the number of decision trees to a single tree. Fifth, this work highlights the advantage of gradient boosted ensembles, in performance and explainability, over linear models, for heterogeneous tabular data and particularly in the context of predicting ADD risk. Finally, this paper proposes a simple decision path for identifying patients with the greatest risk of developing ADD, at the earliest possible time to inform symptom treatment, for memory clinic referral, and soon, interventions.

In experiment 1, the supervised extreme gradient boosting classifier performed well in predicting AD risk, selecting three features (i.e., Everyday Cognition Questionnaire (Study partner) - Total, Alzheimers Disease Assessment Scale (13 items) and Delayed Total Recall) with AUC ROC and AUC PR values consistently above 95% (Figure 5.3, Table 5.2. In experiment 2, the XGBoost classifier identified 11 features, including genetic, cognitive, demographic and brain imaging data, that reliably predict MCI-to-AD progression with ROC AUC scores consistently above 87(Table 5.3). Here, we expand on each clinical feature's capacity to assess AD risk by providing an absolute threshold. The interpretable models for both experiments maintained or improved performance (sensitivity and specificity above 90%) over their complex counterparts while improving their interpretability (Fig. 5.3 and Fig. 5.10). Predicting whether patients end up developing or converting to AD might prompt clinicians to pursue alternative patient-specific symptomatic treatment. Further, high-risk patient groups may be targeted for tighter surveillance and soon-preventive therapies, reducing healthcare system costs.

In the machine learning process, and out of the considered variables, Everyday Cognition Questionnaire (Study partner) - Total, Alzheimers Disease Assessment Scale (13 items) and Delayed Total Recall were the most important predictors of AD risk, followed by brain imaging and the Mini Mental State Exam. These features point to a multifaceted challenge stemming from cognitive health to a potential disturbance in the executive functions networks. The top key features suggest that the multi-modal acquisition (i.e., imaging, genetic, cognitive testing) of cognitive factors may play an important role in determining AD-risk.

6.1 Explaining the heterogeneity

AD patients were older, male, married, had lower education, performed worse on all cognitive tests, and were genetically predisposed (i.e., APOE4 variant). AD risk is multi-factorial and includes demographics, lifestyle, environmental factors and genetic predisposition. Age, especially beyond 65, is the greatest risk factor for AD. Higher education may postpone AD onset by increasing ”cognitive reserve”¡citestern cognitive 2012. Higher education, operationalized as years of schooling, is related with reduced AD diagnosis. Our demographics showed that there was not a difference in race across AD and clinically normal groups; however, we did not conduct analyses stratified by race categories. Educational attainment (measured by years of schooling) did not enhance memory function in Black AD patients, but it did in White AD patients[61]. These characteristics affect AD outcome, yet minorities are underrepresented in research studies, and future studies should focus on stratified and/or intentional analysis across race and ethnicity. Across AD and clinically normal groups, cognitive measures across several cognitive domains and assess via numerous tests (see Table 3.1) were different with the AD group performing poorly compared to the clinically normal group. The mini mental state examination (MMSE) and Montreal Cognitive Assessment (MoCA) have been suggested as cut-off thresholds for identifying adults with baseline mild cognitive impairment who are at risk of developing AD. MMSE alone may not be sensitive enough to identify minor cognitive changes in persons with mild cognitive impairment, but when combined with more comprehensive cognitive evaluations, it may be useful for tracking cognitive changes over time. Whether variations in MMSE scores over time may be used with other cognitive measurements and biomarkers to diagnose AD earlier is uncertain. However, the top two ranked features by the XGBoost model based on their importance were the Everyday Cognition Questionnaire and Alzheimers Disease Assessment Scale, with MMSE showing up as our 5th most important feature and MoCA as the 9th most important feature. The Alzheimers Disease Assessment Scale and delay-total recall differentiated the groups well (see Figure 6.6). Memory paradigms (e.g., delayed total recall) can track neurodegenerative disease trajectories [111].Taken together, delay-total recall and Alzheimers Disease Assessment Scale may be sensitive to neurodegenerative disease trajectories more so than MMSE or MoCA.

AD patients had lower whole brain volume, middle temporal gyrus volume, entorhinal volume, hippocampus volume, fusiform volume, and larger ventricle volume than clinically normal controls. Compared to stable mild cognitive impairment patients, those who converted to AD were older, scored higher on the Alzheimer's Disease Assessment Scale (13 items), and had smaller total brain, middle temporal gyrus, entorhinal, fusiform, and hippocampal volumes. Stabilized MCI patients fared better on the Rey Auditory Verbal Learning Test in learning, forgetting, and immediate categories. With AD, ventricular enlargement, sulcal widening, cortical thinning, and hippocampal broadening increase. Hippocampal shrinkage is a hallmark of AD-related neurodegeneration, although early stages of AD are associated with parahippocampal gyrus atrophy. Our findings align with previous work highlighting entorhinal, perirhinal, and parahippocampal cortex changes with early AD [25, 24].

Interestingly, the SHAP summary plot elucidated important markers for MCI-to-AD conversion (Fig. 5.13). SHAP is currently state of the art for model explainability and is a powerful tool that helps in interpreting the output of machine learning models by attributing feature importance to individual predictors. In 5.13, the higher genetic predisposition, lower Verbal Learning - Immediate Test scores, lower hippocampus volume, and longer Trail Making Test Part B time have collectively contributed to a higher risk prediction from the model. A higher genetic predisposition signifies an increased likelihood of developing the condition based on the individual's genetic makeup (i.e., APOE4). Lower scores on the Verbal Learning - Immediate Test indicate poorer cognitive performance in memory and learning tasks, which suggest potential cognitive decline. A reduced hippocampus volume is known to be associated with memory impairments and is often observed in individuals with neurodegenerative disorders. Lastly, a longer Trail Making Test Part B time implies difficulties in cognitive processing, especially in executive function tasks that include task switching. Bilinguals demonstrate reduced switching costs during task-switching exercises. Additionally, existing research implies that bilingualism may enhance cognitive reserve, as bilingual individuals usually display initial Alzheimer's disease symptoms around 5 years later than their monolingual counterparts [112]. Taken together, these modifiable factors provide valuable insights into the specific attributes that lead to a higher risk prediction, enabling a better understanding of the model's decision-making process and aiding in the development of soon targeted interventions.

6.2 Comparison with previous models

Most AD risk models use traditional statistical methods. More recently, machine learning has been applied to predict patients that may end up developing AD [96, 41], but the machine learning-based predictive models have included costly [41–44], manually selected, and invasive measures (e.g., cerebrospinal fluid analysis of

β-amyloid (Aβ42) [34], Aβ-positron emission tomography [35, 36] which limits the availability of these biomarkers and their usage. The most important predictors in our study were not invasive and can be obtained with little effort. Additionally, while identifying high-risk indicators is useful for risk assessment, setting threshold cut-off values for these factors provides clinical relevance by delineating informed decision routes (i.e., threshold values to distinguish high-risk from low-risk persons). Complex machine learning algorithms need better and more intuitive explanation of their decision processes while preserving performance [46]; Yet, interpretable risk prediction models for disease diagnosis and prognosis are limited [37, 47, 42]. Complex machine learning algorithms hinder clinically intuitive decision process explanations, preventing clinical uptake. Previous machine learning models for predicting AD risk are limited in their explainability and interpretability, but also in their approach to feature selection. Our complex models highlighted the performance capabilities of gradient boosted models, especially when conducting comprehensive feature selection (Exp 2), which allowed our interpretable model to surpass the performance of the latest AD-specific model ([104]) . The interpretable models were designed to support clinical uptake. XGBoost is more interpretable than ML-based logistic regression, especially with non-linear data [14]. Particularly when utilizing all features (data nonlinearity affects the performance of linear models like logistic regression, due to a mismatch between model and data) and when using metrics suited for imbalanced data and high-stakes classification like precision and recall (Fig. 5.10).

6.2.1 The role of multimodal machine learning in multiscale modelling of the brain

Integrating machine learning with multimodal data (e.g., tabular, imaging, time-series, text, simulation) will push medical AI beyond what humans can do [113]. The integration of machine learning in imaging and translational multiscale modelling enables 1) faster modelling/imaging as well as reduced computational burden [114], 2) enhanced data fidelity in both experimental and simulated datasets, and 3) optimization of methods involved in the modelling ranging from image segmentation [115] to preprocessing (e.g., imputation of missing values in tabular data [116]), supplementing (e.g., dealing with an imbalanced dataset [117]), analyzing (e.g., finding high risk factors) and interpreting (e.g., interpretable and explainable machine learning) data to investigate the pathophysiology [118]. In multi-scale modelling, machine learning can be used to improve the quality and accuracy of both experimental and simulation data [119]. In terms of data fidelity, machine learning may be used to 1) replace missing data, 2) streamline data with regards to minimizing data redundancy (e.g., PCA (See Dimensionality Reduction section below) to convert high-dimensional data to low-dimensional space), and/or 3) minimize noise in both experimental, image and simulation data, hence enhancing quality and accuracy. Recent methods that highlight how machine learning can improve data fidelity in physics-informed modelling include 1) a physics-informed neural network that improved resolution and reduced noise in fluid flow data[78], and 2) implementing machine learning when using 4D flow MRI trained using 3D modelling-generated synthetic 4D flow MRI data to increase spatial resolution [79].

Multiscale modelling approaches of the brain go beyond in vivo, in vitro and ex vivo experiments and provide non-invasive estimation of difficult-to-access parameters for identifying and understanding mechanisms across the brain's innervated structural hierarchy [120]. Multiscale modelling in the biomedical domain ranges from the quantum mechanical atomic scale to nanoscale (e.g., membranes), microscale (cellular damage), mesoscale (tissue heterogeneity), macroscale (tissue damage). The combination of 0-dimensional (0-D) to 3-dimensional (3-D) modelling to couple 1-D brain networks with 3-D morphology. The 0-D model is also called lumped parameter modelling and relates the of flow of fluids (e.g., brain blood flow related to vascular dementia) to the flow of electrons and use circuit elements to model flow and pressure across time. Lumped parameter models use "sources" and "sinks" to describe flow into and out of the brain's vascular beds. In the lumped parameter model, the electrical analog of blood flow is current. Viscous blood resistance and inertia are analogous to resistance and inductance, respectively, and vessel compliance is analogous to capacitance. Different formulations of these elements allow modelling from microvascular to macrovascular networks in cerebrovascular beds. In contrast with 0-D/lumped parameter models, whereby variables fluctuate only as a function of time and are represented by ordinary differential equations, 3-D model variables vary as a function of time and 3-D space and are represented by non-linear partial differential equations. 3-D models require solving Navier-Stokes equations (i.e., continuity and momentum), within a region of interest, and involve computational methods such as finite volume, element and difference methods or lattice Boltzmann method to obtain approximate numerical solutions[63]–[68]. Patient specific geometries of the region of interest are obtained from imaging modalities and can be used for developing computational models for estimating individualized metrics [121].

Despite the promise for machine learning to speed up, improve and automate processes in the clinical domain, the benefits of machine learning models are only as good as the data they are trained on. Thus, multi site representative samples that account for the heterogeneity from patients to imaging modalities and their algorithmic acquisition and construction of images is necessary to improve trust and clinical adoption. It is becoming more feasible to view the human architecture in more detail with increased computing, evolution of digital information, and advances in medical imaging, which will feedforward to computational modelling and machine learning methodologies. More advancements in broad integration of multimodal data and machine learning, are necessary before these technologies may be implemented in the clinical context. (see Kadem et al., (2022) for more details[122]).

6.3 Limitations

There is potential for improvement in this study, which will be addressed in future research. The approach herein is data dependent, and thus varies depending on different variable distributions. Despite the multi-site retrospectively studied data, external validation from a different continent is necessary to thoroughly assess the generalizability of the features and model, and thus this is a preliminary diagnostic and prognostic evaluation of AD patients. We look forward to incorporating additional data from additional sites to further improve the performance of the model. We compromised performance for interpretability and avoided overfitting by using a minimal number of clinical features as clinical settings favor interpretable models.

Chapter 7

Conclusion

7.1 Summation

Our work extends beyond the identification of high-risk factors for AD risk, providing an intuitive explanation of the decision process behind assessing risk. We identified data-driven decision routes with for key clinical features with diagnostic and prognostic performances above 97% and 87%, respectively. Finally, the most prominent predictors are non-invasive and easily collected.

In summary, we've identified accessible clinical features, together with clinically operable decision routes, to reliably and rapidly predict patients at the highest risk of developing Alzheimer's disease. We developed interpretable diagnostic and prognostic risk prediction models based on XGBoost with AUC ROC consistently above 0.87+ that can predict short-term AD risk of patients, enabling prevention and potentially expediting research into interventions. The key features highlight the multifaceted nature of risk prediction in AD and suggests the use of indices beyond the current markers to assess AD risk following mild cognitive symptoms. Of clinical relevance, these features explain some of the heterogeneity observed in short-term AD risk and serve to highlight the importance of personalized, accurate, and rapid patient risk stratification. Specifically, our findings suggest the importance of examining interactions of cognitive testing for treatment planning and assessing risk stratification for patients, and particularly older adults. Finally, our work provides a base for future studies focusing on explainable and interpretable predictive models motivated for clinical uptake and to act as a tool in supporting clinical decisions.

7.2 Future directions

In contrast with the revolution in language and computer vision, AI has yet to revolutionize medicine due to the heterogeneity of the population and clinical features. Current medical AI applications are confined to a particular disease and modality, but the rise in large biobanks and cheaper genome sequencing are powering multimodal medical AI models beyond human capabilities. Future studies should leverage larger and more diverse databanks and explainable multimodal AI approaches to help explain the heterogeneity in Alzhemier's Disease.

With the rise of computational power and digitized data, machine learning will continue to gain momentum in its applications to research and clinical settings. In this paper, we described key concepts in machine and deep learning, to familiarize students, clinicians, scientists, and engineers with fundamental principles of machine and deep learning. We identify important potential pitfalls that must be considered when building models with practical implications. Machine and deep learning will continue to 1) contribute to our understanding of medical abnormalities and 2) improve our capacity to create patient-specific diagnostic and prognostic tools for classifying and assessing outcomes following medical treatments, 3) help in treatment option selection and optimization, 4) minimize time, automate, and improve each part of the clinical processes and 5) adapt to heterogenous conditions.

To improve the generalizability of machine and deep learning and their uptake in clinical settings, we need multi-site, diverse datasets including data from underpresented groups that account for the heterogeneity in patients and imaging modalities, and operator interpretations. Implementing machine learning approaches in different research domains will also require adopting open science principles to improve reproducibility and reduce “blackbox” approaches.

Advances in machine and deep learning in recent and future years will require multidisciplinary collaborations between clinicians, scientists, industry and engineers to form personalized, robust, trustworthy, generalizable, explainable, reliable, reproducible, and safe diagnostic and predictive tools. Future work is required to determine whether machine learning-based approaches outperform traditional statistical approaches for risk assessment and to determine ways to increase uptake of reliable machine learning models across research domains.

Bibliography

[1] World Health Organization. *Global status report on the public health response to dementia*. World Health Organization, Geneva, 2021. ISBN 978-92-4-003324-5. URL https://apps.who.int/iris/handle/10665/344701. Section: xv, 251 p.

[2] Alzheimer's Association. 2019 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 15(3):321–387, 2019. ISSN 1552-5279. doi: 10.1016/j.jalz.2019.01.010. URL https://onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2019.01.010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.jalz.2019.01.010.

[3] Constantine G. Lyketsos, Oscar Lopez, Beverly Jones, Annette L. Fitzpatrick, John Breitner, and Steven DeKosky. Prevalence of neuropsychiatric symptoms in dementia and mild cognitive impairment: results from the cardiovascular health study. *JAMA*, 288(12):1475–1483, September 2002. ISSN 0098-7484. doi: 10.1001/jama.288.12.1475.

[4] Angela Guarino, Francesca Favieri, Ilaria Boncompagni, Francesca Agostini, Micaela Cantone, and Maria Casagrande. Executive Functions in Alzheimer

Disease: A Systematic Review. *Frontiers in Aging Neuroscience*, 10:437, January 2019. ISSN 1663-4365. doi: 10.3389/fnagi.2018.00437. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6341024/.

[5] Yu Yamazaki, Na Zhao, Thomas R. Caulfield, Chia-Chen Liu, and Guojun Bu. Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. *Nature Reviews Neurology*, 15(9):501–518, September 2019. ISSN 1759-4766. doi: 10.1038/s41582-019-0228-7. URL https://www.nature.com/articles/s41582-019-0228-7. Number: 9 Publisher: Nature Publishing Group.

[6] Sebastian Palmqvist, Pontus Tideman, Nicholas Cullen, Henrik Zetterberg, Kaj Blennow, Jeffery L. Dage, Erik Stomrud, Shorena Janelidze, Niklas Mattsson-Carlgren, and Oskar Hansson. Prediction of future Alzheimer's disease dementia using plasma phospho-tau combined with other accessible measures. *Nature Medicine*, 27(6):1034–1042, June 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01348-z. URL https://www.nature.com/articles/s41591-021-01348-z. Number: 6 Publisher: Nature Publishing Group.

[7] Navigating the Path Forward for Dementia in Canada: The Landmark Study Report #1. URL http://alzheimer.ca/en/research/reports-dementia/landmark-study-report-1-path-forward. Publication Title: Alzheimer Society of Canada.

[8] Julie Gonneaud, Eider M. Arenaza-Urquijo, Florence Mézenge, Brigitte Landeau, Malo Gaubert, Alexandre Bejanin, Robin de Flores, Miranka Wirth, Clémence Tomadesso, Géraldine Poisnel, Ahmed Abbas, Béatrice Desgranges, and Gaël Chételat. Increased florbetapir binding in the temporal neocortex

from age 20 to 60 years. *Neurology*, 89(24):2438–2446, December 2017. ISSN 1526-632X. doi: 10.1212/WNL.0000000000004733.

[9] Claudio Franceschi and Judith Campisi. Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, 69 Suppl 1: S4–9, June 2014. ISSN 1758-535X. doi: 10.1093/gerona/glu057.

[10] Yuko Hara, Nicholas McKeehan, and Howard M. Fillit. Translating the biology of aging into novel therapeutics for Alzheimer disease. *Neurology*, 92 (2):84–93, January 2019. ISSN 0028-3878, 1526-632X. doi: 10.1212/WNL.0000000000006745. URL `https://n.neurology.org/content/92/2/84`.

[11] Costantino Iadecola and Rebecca F. Gottesman. Cerebrovascular Alterations in Alzheimer Disease. *Circulation Research*, 123(4):406–408, August 2018. ISSN 1524-4571. doi: 10.1161/CIRCRESAHA.118.313400.

[12] Marta Cortes-Canteli and Costantino Iadecola. Alzheimer's Disease and Vascular Aging. *Journal of the American College of Cardiology*, 75(8):942–951, March 2020. ISSN 0735-1097. doi: 10.1016/j.jacc.2019.10.062. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8046164/`.

[13] Hamid M. Abdolmaleky, Sam Thiagalingam, and Marsha Wilcox. Genetics and epigenetics in major psychiatric disorders: dilemmas, achievements, applications, and future scope. *American Journal of Pharmacogenomics: Genomics-Related Research in Drug Development and Clinical Practice*, 5(3):149–160, 2005. ISSN 1175-2203. doi: 10.2165/00129785-200505030-00002.

[14] Ramón Cacabelos. The application of functional genomics to Alzheimer's disease. *Pharmacogenomics*, 4(5):597–621, September 2003. ISSN 1462-2416. doi: 10.1517/phgs.4.5.597.23795.

[15] J S Rao, V L Keleshian, S Klein, and S I Rapoport. Epigenetic modifications in frontal cortex from Alzheimer's disease and bipolar disorder patients. *Translational Psychiatry*, 2(7):e132, July 2012. ISSN 2158-3188. doi: 10.1038/tp.2012.55. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3410632/`.

[16] R. Delgado-Morales and M. Esteller. Opening up the DNA methylome of dementia. *Molecular Psychiatry*, 22(4):485–496, April 2017. ISSN 1476-5578. doi: 10.1038/mp.2016.242.

[17] Raúl Delgado-Morales, Roberto Carlos Agís-Balboa, Manel Esteller, and María Berdasco. Epigenetic mechanisms during ageing and neurogenesis as novel therapeutic avenues in human brain disorders. *Clinical Epigenetics*, 9:67, June 2017. ISSN 1868-7075. doi: 10.1186/s13148-017-0365-z. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5493012/`.

[18] L. Detoledo-Morrell, M. P. Sullivan, F. Morrell, R. S. Wilson, D. A. Bennett, and S. Spencer. Alzheimer's disease: in vivo detection of differential vulnerability of brain regions. *Neurobiology of Aging*, 18(5):463–468, 1997. ISSN 0197-4580. doi: 10.1016/s0197-4580(97)00114-0.

[19] Corina Pennanen, Miia Kivipelto, Susanna Tuomainen, Päivi Hartikainen, Tuomo Hänninen, Mikko P. Laakso, Merja Hallikainen, Matti Vanhanen, Aulikki Nissinen, Eeva-Liisa Helkala, Pauli Vainio, Ritva Vanninen, Kaarina Partanen, and Hilkka Soininen. Hippocampus and entorhinal cortex in mild

cognitive impairment and early AD. *Neurobiology of Aging*, 25(3):303–310, March 2004. ISSN 0197-4580. doi: 10.1016/S0197-4580(03)00084-8.

[20] H. Braak and E. Braak. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, 82(4):239–259, September 1991. ISSN 1432-0533. doi: 10.1007/BF00308809. URL `https://doi.org/10.1007/BF00308809`.

[21] Peter J. Nestor, Philip Scheltens, and John R. Hodges. Advances in the early detection of Alzheimer's disease. *Nature Medicine*, 10 Suppl:S34–41, July 2004. ISSN 1078-8956. doi: 10.1038/nrn1433.

[22] Akram Bakkour, John C. Morris, David A. Wolk, and Bradford C. Dickerson. The effects of aging and Alzheimer's disease on cerebral cortical anatomy: Specificity and differential relationships with cognition. *NeuroImage*, 76:332–344, August 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2013.02.059. URL `https://www.sciencedirect.com/science/article/pii/S1053811913002140`.

[23] Pierrick Coupé, José Vicente Manjón, Enrique Lanuza, and Gwenaelle Catheline. Lifespan Changes of the Human Brain In Alzheimer's Disease. *Scientific Reports*, 9(1):3998, March 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-39809-8. URL `https://www.nature.com/articles/s41598-019-39809-8`.

[24] Sabine Krumm, Sasa L. Kivisaari, Alphonse Probst, Andreas U. Monsch, Julia Reinhardt, Stephan Ulmer, Christoph Stippich, Reto W. Kressig, and Kirsten I. Taylor. Cortical thinning of parahippocampal subregions in very

early Alzheimer's disease. *Neurobiology of Aging*, 38:188–196, February 2016. ISSN 0197-4580. doi: 10.1016/j.neurobiolaging.2015.11.001. URL `https://www.sciencedirect.com/science/article/pii/S0197458015005539`.

[25] Vincent Planche, José V. Manjon, Boris Mansencal, Enrique Lanuza, Thomas Tourdias, Gwenaëlle Catheline, and Pierrick Coupé. Structural progression of Alzheimer's disease over decades: the MRI staging scheme. *Brain Communications*, 4(3):fcac109, June 2022. ISSN 2632-1297. doi: 10.1093/braincomms/fcac109. URL `https://doi.org/10.1093/braincomms/fcac109`.

[26] Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. "Mini-mental state". *Journal of Psychiatric Research*, 12(3):189–198, November 1975. ISSN 00223956. doi: 10.1016/0022-3956(75)90026-6. URL `https://linkinghub.elsevier.com/retrieve/pii/0022395675900266`.

[27] Daniel HJ Davis, Sam T. Creavin, Jennifer LY Yip, Anna H. Noel-Storr, Carol Brayne, and Sarah Cullum. Montreal Cognitive Assessment for the diagnosis of Alzheimer's disease and other dementias. *The Cochrane Database of Systematic Reviews*, 2015(10), October 2015. ISSN 1465-1858. doi: 10.1002/14651858.CD010775.pub2. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6682492/`. Publisher: John Wiley and Sons, Inc. and the Cochrane Library.

[28] Nicole Noren Hooten, Natasha L. Pacheco, Jessica T. Smith, and Michele K. Evans. The accelerated aging phenotype: The role of race and social determinants of health on aging. *Ageing Research Reviews*, 73:101536, January 2022. ISSN 1568-1637. doi: 10.1016/j.arr.2021.101536. URL `https:`

//www.sciencedirect.com/science/article/pii/S156816372100283X.

[29] Yaakov Stern. What is cognitive reserve? Theory and research application of the reserve concept. *Journal of the International Neuropsychological Society: JINS*, 8(3):448–460, March 2002. ISSN 1355-6177.

[30] Selam Negash, Sharon Xie, Christos Davatzikos, Christopher M. Clark, John Q. Trojanowski, Leslie M. Shaw, David A. Wolk, and Steven E. Arnold. Cognitive and functional resilience despite molecular evidence of Alzheimer's disease pathology. *Alzheimer's & Dementia*, 9(3):e89–e95, May 2013. ISSN 1552-5260. doi: 10.1016/j.jalz.2012.01.009. URL https://www.sciencedirect.com/science/article/pii/S1552526012000295.

[31] Mónica Rosselli, Idaly Vélez Uribe, Emily Ahne, and Layaly Shihadeh. Culture, Ethnicity, and Level of Education in Alzheimer's Disease. *Neurotherapeutics*, 19(1):26–54, January 2022. ISSN 1933-7213. doi: 10.1007/s13311-022-01193-z. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8960082/.

[32] Myron F. Weiner. Perspective on race and ethnicity in Alzheimer's disease research. *Alzheimer's & Dementia*, 4(4):233–238, July 2008. ISSN 1552-5260. doi: 10.1016/j.jalz.2007.10.016. URL https://www.sciencedirect.com/science/article/pii/S1552526007006358.

[33] Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, August 2001. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1009213726. publisher: Institute of Mathematical Statistics.

[34] Sebastian Palmqvist, Henrik Zetterberg, Kaj Blennow, Susanna Vestberg, Ulf Andreasson, David J. Brooks, Rikard Owenius, Douglas Hägerström, Per Wollmer, Lennart Minthon, and Oskar Hansson. Accuracy of brain amyloid detection in clinical practice using cerebrospinal fluid -amyloid 42: a cross-validation study against amyloid positron emission tomography. *JAMA neurology*, 71(10):1282–1289, October 2014. ISSN 2168-6157. doi: 10.1001/jamaneurol.2014.1358.

[35] Gil D. Rabinovici, Constantine Gatsonis, Charles Apgar, Kiran Chaudhary, Ilana Gareen, Lucy Hanna, James Hendrix, Bruce E. Hillner, Cynthia Olson, Orit H. Lesman-Segev, Justin Romanoff, Barry A. Siegel, Rachel A. Whitmer, and Maria C. Carrillo. Association of Amyloid Positron Emission Tomography With Subsequent Change in Clinical Management Among Medicare Beneficiaries With Mild Cognitive Impairment or Dementia. *JAMA*, 321(13):1286–1294, April 2019. ISSN 1538-3598. doi: 10.1001/jama.2019.2000.

[36] Niklas Mattsson, Sebastian Palmqvist, Erik Stomrud, Jacob Vogel, and Oskar Hansson. Staging -Amyloid Pathology With Amyloid Positron Emission Tomography. *JAMA Neurology*, 76(11):1319–1329, November 2019. ISSN 2168-6149. doi: 10.1001/jamaneurol.2019.2214. URL `https://doi.org/10.1001/jamaneurol.2019.2214`.

[37] María Vega García and José L. Aznarte. Shapley additive explanations for NO2 forecasting. 56:101039. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2019.101039. URL `https://www.sciencedirect.com/science/article/pii/S1574954119303498`.

[38] K. Walters, S. Hardoon, I. Petersen, S. Iliffe, R. Z. Omar, I. Nazareth, and G. Rait. Predicting dementia risk in primary care: development and validation of the Dementia Risk Score using routinely collected data. *BMC medicine*, 14: 6, January 2016. ISSN 1741-7015. doi: 10.1186/s12916-016-0549-y.

[39] Miia Kivipelto, Tiia Ngandu, Tiina Laatikainen, Bengt Winblad, Hilkka Soininen, and Jaakko Tuomilehto. Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study. *The Lancet. Neurology*, 5(9):735–741, September 2006. ISSN 1474-4422. doi: 10.1016/S1474-4422(06)70537-3.

[40] Kaarin J. Anstey, Nicolas Cherbuin, and Pushpani M. Herath. Development of a new method for assessing global risk of Alzheimer's disease for use in population health approaches to prevention. *Prevention Science: The Official Journal of the Society for Prevention Research*, 14(4):411–421, August 2013. ISSN 1573-6695. doi: 10.1007/s11121-012-0313-2.

[41] Lin Tang, Xiaojia Wu, Huan Liu, Faqi Wu, Rao Song, Wei Zhang, Dajing Guo, Junbang Feng, and Chuanming Li. Individualized Prediction of Early Alzheimer's Disease Based on Magnetic Resonance Imaging Radiomics, Clinical, and Laboratory Examinations: A 60-Month Follow-Up Study. *Journal of magnetic resonance imaging: JMRI*, 54(5):1647–1657, November 2021. ISSN 1522-2586. doi: 10.1002/jmri.27689.

[42] Enrico Pellegrini, Lucia Ballerini, Maria Del C. Valdes Hernandez, Francesca M. Chappell, Victor González-Castro, Devasuda Anblagan, Samuel Danso, Susana Muñoz-Maniega, Dominic Job, Cyril Pernet, Grant Mair, Tom J. MacGillivray,

Emanuele Trucco, and Joanna M. Wardlaw. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, 10:519–535, 2018. ISSN 2352-8729. doi: 10.1016/j.dadm.2018.07.004.

[43] Esther E. Bron, Stefan Klein, Janne M. Papma, Lize C. Jiskoot, Vikram Venkataghavan, Jara Linders, Pauline Aalten, Peter Paul De Deyn, Geert Jan Biessels, Jurgen A. H. R. Claassen, Huub A. M. Middelkoop, Marion Smits, Wiro J. Niessen, John C. van Swieten, Wiesje M. van der Flier, Inez H. G. B. Ramakers, Aad van der Lugt, Alzheimer's Disease Neuroimaging Initiative, and Parelsnoer Neurodegenerative Diseases study group. Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage. Clinical*, 31:102712, 2021. ISSN 2213-1582. doi: 10.1016/j.nicl.2021.102712.

[44] Silvia Basaia, Federica Agosta, Luca Wagner, Elisa Canu, Giuseppe Magnani, Roberto Santangelo, Massimo Filippi, and Alzheimer's Disease Neuroimaging Initiative. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage. Clinical*, 21:101645, 2019. ISSN 2213-1582. doi: 10.1016/j.nicl.2018.101645.

[45] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, 2(1):56–67, January 2020. ISSN 2522-5839. doi: 10.1038/s42256-019-0138-9.

[46] Patrick Hall and Navdeep Gill. *An introduction to machine learning interpretability.* O'Reilly Media, Incorporated, 2019.

[47] Esra Zihni, Vince Istvan Madai, Michelle Livne, Ivana Galinovic, Ahmed A. Khalil, Jochen B. Fiebach, and Dietmar Frey. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *PLOS ONE*, 15(4):e0231166, April 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0231166. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231166`. Publisher: Public Library of Science.

[48] Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.*, 2(10):749–760, October 2018. ISSN 2157-846X. doi: 10.1038/s41551-018-0304-0.

[49] Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje Willemsen, Yongchao Ma, Qixiang Fang, Sybren Hindriks, Lars Tummers, and Daniel L. Oberski. An open source machine learning framework for efficient and transparent systematic reviews. 3(2):125–133. ISSN 2522-5839. doi: 10.1038/s42256-020-00287-7. URL `https://www.nature.com/articles/s42256-020-00287-7`. Number: 2 Publisher: Nature Publishing Group.

[50] Ingrid Arevalo-Rodriguez, Nadja Smailagic, Marta Roqué i Figuls, Agustín

Ciapponi, Erick Sanchez-Perez, Antri Giannakou, Olga L Pedraza, Xavier Bonfill Cosp, and Sarah Cullum. Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). *The Cochrane Database of Systematic Reviews*, 2015(3): CD010783, March 2015. ISSN 1469-493X. doi: 10.1002/14651858.CD010783. pub2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6464748/.

[51] Fabian Corlier, George Hafzalla, Joshua Faskowitz, Lewis H. Kuller, James T. Becker, Oscar L. Lopez, Paul M. Thompson, and Meredith N. Braskie. Systemic inflammation as a predictor of brain aging: Contributions of physical activity, metabolic risk, and genetic risk. *NeuroImage*, 172:118–129, May 2018. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2017.12.027. URL https://www.sciencedirect.com/science/article/pii/S1053811917310492.

[52] Nicholas C. Cullen, A. nders Mälarstig, Erik Stomrud, Oskar Hansson, and Niklas Mattsson-Carlgren. Accelerated inflammatory aging in Alzheimer's disease and its relation to amyloid, tau, and cognition. *Scientific Reports*, 11(1): 1965, January 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-81705-7. URL https://www.nature.com/articles/s41598-021-81705-7.

[53] Henry W. Querfurth and Frank M. LaFerla. Alzheimer's disease. *The New England Journal of Medicine*, 362(4):329–344, January 2010. ISSN 1533-4406. doi: 10.1056/NEJMra0909142.

[54] Lars Bertram, Matthew B. McQueen, Kristina Mullin, Deborah Blacker, and

Rudolph E. Tanzi. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature Genetics*, 39(1):17–23, January 2007. ISSN 1061-4036. doi: 10.1038/ng1934.

[55] Lars Bertram and Rudolph E. Tanzi. Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses. *Nature Reviews Neuroscience*, 9(10):768–778, October 2008. ISSN 1471-0048. doi: 10.1038/nrn2494. URL `https://www.nature.com/articles/nrn2494`.

[56] Huei-Yang Chen and Peter K. Panegyres. The Role of Ethnicity in Alzheimer's Disease: Findings From The C-PATH Online Data Repository. *Journal of Alzheimer's Disease*, 51(2):515–523, January 2016. ISSN 1387-2877. doi: 10.3233/JAD-151089. URL `https://content.iospress.com/articles/journal-of-alzheimers-disease/jad151089`.

[57] 2021 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 17(3): 327–406, March 2021. ISSN 1552-5260. doi: 10.1002/alz.12328. URL `https://alz.journals.onlinelibrary.wiley.com/doi/10.1002/alz.12328`.

[58] Jennifer C. Howell, Kelly D. Watts, Monica W. Parker, Junjie Wu, Alexander Kollhoff, Thomas S. Wingo, Cornelya D. Dorbin, Deqiang Qiu, and William T. Hu. Race modifies the relationship between cognition and Alzheimer's disease cerebrospinal fluid biomarkers. *Alzheimer's Research & Therapy*, 9(1):88, December 2017. ISSN 1758-9193. doi: 10.1186/s13195-017-0315-1. URL `https://alzres.biomedcentral.com/articles/10.1186/s13195-017-0315-1`.

[59] Yaakov Stern. Cognitive reserve in ageing and Alzheimer's disease. *The Lancet.*

Neurology, 11(11):1006–1012, November 2012. ISSN 1474-4465. doi: 10.1016/S1474-4422(12)70191-6.

[60] Susanna C. Larsson, Matthew Traylor, Rainer Malik, Martin Dichgans, Stephen Burgess, Hugh S. Markus, and CoSTREAM Consortium, on behalf of the International Genomics of Alzheimer's Project. Modifiable pathways in Alzheimer's disease: Mendelian randomisation analysis. *BMJ (Clinical research ed.)*, 359: j5375, December 2017. ISSN 1756-1833. doi: 10.1136/bmj.j5375.

[61] Arash Rahmani, Babak Najand, Amanda Sonnega, Golnoush Akhlaghipour, Mario F. Mendez, Shervin Assari, and for the Alzheimer's Disease Neuroimaging Initiative. Intersectional Effects of Race and Educational Attainment on Memory Function of Middle-Aged and Older Adults With Alzheimer's Disease. *Journal of Racial and Ethnic Health Disparities*, December 2022. ISSN 2196-8837. doi: 10.1007/s40615-022-01499-w. URL `https://doi.org/10.1007/s40615-022-01499-w`.

[62] Yana Blinkouskaya and Johannes Weickenmeier. Brain Shape Changes Associated With Cerebral Atrophy in Healthy Aging and Alzheimer's Disease. *Frontiers in Mechanical Engineering*, 7, 2021. ISSN 2297-3079. URL `https://www.frontiersin.org/articles/10.3389/fmech.2021.705653`.

[63] C. Echávarri, P. Aalten, H. B. M. Uylings, H. I. L. Jacobs, P. J. Visser, E. H. B. M. Gronenschild, F. R. J. Verhey, and S. Burgmans. Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer's disease. *Brain Structure & Function*, 215(3):265–271, 2011. ISSN 1863-2653.

doi: 10.1007/s00429-010-0283-8. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041901/.

[64] R. Shouval, O. Bondi, H. Mishan, A. Shimoni, R. Unger, and A. Nagler. Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in SCT. *Bone Marrow Transplant.*, 49(3):332–337, March 2014. ISSN 1476-5365. doi: 10.1038/bmt.2013.146.

[65] Jack V. Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.*, 49(11):1225–1231, November 1996. ISSN 0895-4356. doi: 10.1016/S0895-4356(96)00002-9.

[66] Christopher J. Pannucci and Edwin G. Wilkins. Identifying and Avoiding Bias in Research. *Plast. Reconstr. Surg.*, 126(2):619, August 2010. doi: 10.1097/PRS.0b013e3181de24bc.

[67] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2):94, June 2019. doi: 10.7861/futurehosp.6-2-94.

[68] Mike May. Eight ways machine learning is assisting medicine - Nature Medicine. *Nature*, 27(1):2–3, January 2021. ISSN 1546-170X. doi: 10.1038/s41591-020-01197-2.

[69] Javier De Velasco Oriol, Edgar E. Vallejo, Karol Estrada, José Gerardo Taméz Peña, and The Alzheimer's Disease Neuroimaging Initiative. Benchmarking machine learning models for late-onset alzheimer's disease prediction

from genomic data. *BMC bioinformatics*, 20(1):709, December 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-3158-x.

[70] Hyun Kang. The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402, May 2013. doi: 10.4097/kjae.2013.64.5.402.

[71] Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *N. Engl. J. Med.*, 383(9):874–882, June 2020. ISSN 1533-4406. doi: 10.1056/NEJMms2004740.

[72] Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, and Ewout W. Steyerberg. Calibration: the Achilles heel of predictive analytics. *BMC Med.*, 17(1):1–7, December 2019. ISSN 1741-7015. doi: 10.1186/s12916-019-1466-7.

[73] Anita L. Lynam, John M. Dennis, Katharine R. Owen, Richard A. Oram, Angus G. Jones, Beverley M. Shields, and Lauric A. Ferrat. Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagn. Progn. Res.*, 4(1):1–10, December 2020. ISSN 2397-7523. doi: 10.1186/s41512-020-00075-2.

[74] Joshua J. Levy and A. James O'Malley. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. *BMC Med. Res. Method.*, 20(1):1–15, December 2020. ISSN 1471-2288. doi: 10.1186/s12874-020-01046-3.

[75] A. Ralph Henderson. The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. *Clin. Chim. Acta.*, 359(1-2):1–26, September 2005. ISSN 0009-8981. doi: 10.1016/j.cccn.2005.04.002.

[76] Peter C. Austin, Frank E. Harrell, Jr., and Ewout W. Steyerberg. Predictive performance of machine and statistical learning methods: Impact of data-generating processes on external validity in the ”large N, small p” setting. *Stat. Methods Med. Res.*, 30(6):1465–1483, June 2021. ISSN 1477-0334. doi: 10.1177/09622802211002867.

[77] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep Learning in Medical Image Analysis. *Annu. Rev. Biomed. Eng.*, 19:221, June 2017. doi: 10.1146/annurev-bioeng-071516-044442.

[78] Y. Bengio, A. Courville, and P. Vincent. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828. doi: 10.1109/tpami.2013.50.

[79] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Med. Image Anal.*, 42:60–88, December 2017. ISSN 1361-8423. doi: 10.1016/j.media.2017.07.005.

[80] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553):436–444, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539.

[81] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In *International Conference on Machine Learning*, pages 1857–1865. PMLR, July 2017. URL `https://proceedings.mlr.press/v70/kim17a.html`.

[82] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer, Cham, Switzerland, November 2015. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

[83] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digital Health*, 3(11):e745–e750, November 2021. ISSN 2589-7500. doi: 10.1016/S2589-7500(21)00208-9.

[84] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777. Curran Associates Inc., Red Hook, NY, USA, December 2017. ISBN 978-1-51086096-4. doi: 10.5555/3295222.3295230.

[85] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein,

and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.*, 2(8): 476–486, August 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-0212-3.

[86] Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it? *arXiv*, April 2021. doi: 10.48550/arXiv.2104.00673.

[87] Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, July 2010. URL `https://research-portal.uea.ac.uk/en/publications/on-over-fitting-in-model-selection-and-subsequent-selection-bias-`.

[88] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinf.*, 7(1):1–8, December 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-91.

[89] E. W. Steyerberg, F. E. Harrell, Jr., G. J. Borsboom, M. J. Eijkemans, Y. Vergouwe, and J. D. Habbema. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.*, 54(8): 774–781, August 2001. ISSN 0895-4356. doi: 10.1016/s0895-4356(01)00341-9.

[90] Mona Alhajeri and Syed Ghulam Sarwar Shah. Limitations in and Solutions for Improving the Functionality of Picture Archiving and Communication System: an Exploratory Study of PACS Professionals' Perspectives. *J. Digit. Imaging*, 32(1):54, February 2019. doi: 10.1007/s10278-018-0127-2.

[91] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z.

[92] Paulyne Lee, Maxine Le Saux, Rebecca Siegel, Monika Goyal, Chen Chen, Yan Ma, and Andrew C. Meltzer. Racial and ethnic disparities in the management of acute pain in US emergency departments: Meta-analysis and systematic review. *Am. J. Emerg. Med.*, 37(9):1770–1777, September 2019. ISSN 1532-8171. doi: 10.1016/j.ajem.2019.06.014.

[93] Richard Ribón Fletcher, Audace Nakeshimana, and Olusubomi Olubeko. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Front. Artif. Intell.*, 3, April 2021. ISSN 2624-8212. doi: 10.3389/frai.2020.561802.

[94] Richard D. Riley, Joie Ensor, Kym I. E. Snell, Frank E. Harrell, Glen P. Martin, Johannes B. Reitsma, Karel G. M. Moons, Gary Collins, and Maarten van Smeden. Calculating the sample size required for developing a clinical prediction model. *BMJ*, 368:m441, March 2020. ISSN 1756-1833. doi: 10.1136/bmj.m441.

[95] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, April 1982. ISSN 0033-8419. doi: 10.1148/radiology.143.1.7063747.

[96] Fahimeh Nezhadmoghadam, Antonio Martinez-Torteya, Victor Treviño, Emmanuel Martínez, Alejandro Santos, Jose Tamez-Peña, and null Alzheimer's

Disease Neuroimaging Initiative. Robust Discovery of Mild Cognitive Impairment Subtypes and Their Risk of Alzheimer's Disease Conversion Using Unsupervised Machine Learning and Gaussian Mixture Modeling. *Current Alzheimer Research*, 18(7):595–606, 2021. ISSN 1875-5828. doi: 10.2174/1567205018666210831145825.

[97] # Lin Tang, # Xiaojia Wu, Huan Liu, Faqi Wu, Rao Song, Wei Zhang, Dajing Guo, Junbang Feng, and Chuanming Li. Individualized Prediction of Early Alzheimer's Disease Based on Magnetic Resonance Imaging Radiomics, Clinical, and Laboratory Examinations: A 60-Month Follow-Up Study. *J. Magn. Reson. Imaging*, 54(5):1647–1657, November 2021. ISSN 1522-2586. doi: 10.1002/jmri.27689.

[98] Emilie Chary, Hélène Amieva, Karine Pérès, Jean-Marc Orgogozo, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Short- versus long-term prediction of dementia among subjects with low and high educational levels. *Alzheimer's & Dementia*, 9(5):562–571, September 2013. ISSN 1552-5260. doi: 10.1016/j.jalz.2012.05.2188.

[99] Steffen Wolfsgruber, Frank Jessen, Birgitt Wiese, Janine Stein, Horst Bickel, Edelgard Mösch, Siegfried Weyerer, Jochen Werle, Michael Pentzek, Angela Fuchs, Mirjam Köhler, Cadja Bachmann, Steffi G. Riedel-Heller, Martin Scherer, Wolfgang Maier, Michael Wagner, AgeCoDe Study Group, Wolfgang Maier, Martin Scherer, Heinz-Harald Abholz, Cadja Bachmann, Horst Bickel, Wolfgang Blank, Hendrik van den Bussche, Sandra Eifflaender-Gorfer, Marion Eisele, Annette Ernst, Angela Fuchs, Kathrin Heser, Frank Jessen, Hanna

Kaduszkiewicz, Teresa Kaufeler, Mirjam Köhler, Hans-Helmut König, Alexander Koppara, Carolin Lange, Hanna Leicht, Tobias Luck, Melanie Luppa, Manfred Mayer, Edelgard Mösch, Julia Olbrich, Michael Pentzek, Jana Prokein, Anna Schumacher, Steffi Riedel-Heller, Janine Stein, Susanne Steinmann, Franziska Tebarth, Michael Wagner, Klaus Weckbecker, Dagmar Weeg, Jochen Werle, Siegfried Weyerer, Birgitt Wiese, Steffen Wolfsgruber, and Thomas Zimmermann. The CERAD neuropsychological assessment battery total score detects and predicts Alzheimer disease dementia with high diagnostic accuracy. *Am. J. Geriatr. Psychiatry*, 22(10):1017–1028, October 2014. ISSN 1545-7214. doi: 10.1016/j.jagp.2012.08.021.

[100] Amy Xu, Valentina L. Kouznetsova, and Igor F. Tsigelny. Alzheimer's Disease Diagnostics Using miRNA Biomarkers and Machine Learning. *J. Alzheimers. Dis.*, 86(2):841–859, 2022. ISSN 1875-8908. doi: 10.3233/JAD-215502.

[101] J. Kim, S. C. Kim, D. Kang, D. K. Yon, and J. G. Kim. Classification of Alzheimer's disease stage using machine learning for left and right oxygenation difference signals in the prefrontal cortex: a patient-level, single-group, diagnostic interventional trial. *Eur. Rev. Med. Pharmacol. Sci.*, 26(21):7734–7741, November 2022. ISSN 2284-0729. doi: 10.26355/eurrev_202211_30122.

[102] Jia You, Ya-Ru Zhang, Hui-Fu Wang, Ming Yang, Jian-Feng Feng, Jin-Tai Yu, and Wei Cheng. Development of a novel dementia risk prediction model in the general population: A large, longitudinal, population-based machine-learning study. *eClinicalMedicine*, 53, November 2022. ISSN 2589-5370. doi: 10.1016/j.eclinm.2022.101665.

[103] Shaker El-Sappagh, Jose M. Alonso, S. M. Riazul Islam, Ahmad M. Sultan, and Kyung Sup Kwak. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Sci. Rep.*, 11 (2660):1–26, January 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-82098-3.

[104] # Ingrid Rye, # Alexandra Vik, # Marek Kocinski, Alexander S. Lundervold, and Astri J. Lundervold. Predicting conversion to Alzheimer's disease in individuals with Mild Cognitive Impairment using clinically transferable features. *Sci. Rep.*, 12(1):15566., September 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-18805-5.

[105] Xie Wei, Sun Huai-qiang, Chen Jia-wei, Zeng Yi, X. U. Xu, L. I. Zhen-lin, and Xia Chun-chao. A Preliminary Study of Applying Geometric Deep Learning in Brain Morphometry for Diagnosis of Alzheimer's Disease. *scdxxbyxb*, 52(2): 300–305, March 2021. ISSN 1672-173X. doi: 10.12182/20210360103.

[106] Xiaoran Zheng, Wei Zhang, Xing Wang, Renren Li, Meng Liu, Feiyang Xu, Yunxia Li, Jialin Zheng, and Zhiyu Nie. Extended Application of Digital Clock Drawing Test in the Evaluation of Alzheimer's Disease Based on Artificial Intelligence and the Neural Basis. *Curr. Alzheimer Res.*, 18(14):1127–1139, 2021. ISSN 1875-5828. doi: 10.2174/1567205018666211210150808.

[107] Jinglin Sun, Yu Liu, Hao Wu, Peiguang Jing, and Yong Ji. A novel deep learning approach for diagnosing Alzheimer's disease based on eye-tracking data. *Front. Hum. Neurosci.*, 16, September 2022. ISSN 1662-5161. doi: 10.3389/fnhum.2022.972773.

[108] Udit Singhania, Balakrushna Tripathy, Mohammad Kamrul Hasan, Noble C.

Anumbe, Dabiah Alboaneen, Fatima Rayan Awad Ahmed, Thowiba E. Ahmed, and Manasik M. Mohamed Nour. A Predictive and Preventive Model for Onset of Alzheimer's Disease. *Front. Public Health*, 9:751536., October 2021. ISSN 2296-2565. doi: 10.3389/fpubh.2021.751536.

[109] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. Association for Computing Machinery, New York, NY, USA, August 2016. ISBN 978-1-45034232-2. doi: 10.1145/2939672.2939785.

[110] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep Neural Networks and Tabular Data: A Survey. *arXiv*, October 2021. doi: 10.1109/TNNLS.2022.3229161.

[111] Cutter A. Lindbergh, Nicole Walker, Renaud La Joie, Sophia Weiner-Light, Adam M. Staffaroni, Kaitlin B. Casaletto, Fanny Elahi, Samantha M. Walters, Michelle You, Devyn Cotter, Breton Asken, Alexandra C. Apple, Elena Tsoy, John Neuhaus, Corrina Fonseca, Amy Wolf, Yann Cobigo, Howie Rosen, Joel H. Kramer, and Hillblom Aging Network. Worth the Wait: Delayed Recall after 1 Week Predicts Cognitive and Medial Temporal Lobe Trajectories in Older Adults. *J. Int. Neuropsychol. Soc.*, 27(4):382–388, April 2021. ISSN 1469-7661. doi: 10.1017/S1355617720001009.

[112] Víctor Costumero, Lidon Marin-Marin, Marco Calabria, Vicente Belloch, Joaquín Escudero, Miguel Baquero, Mireia Hernandez, Juan Ruiz de Miras, Albert Costa, Maria-Antònia Parcet, and César Ávila. A cross-sectional and

longitudinal study on the protective effect of bilingualism against dementia using brain atrophy and cognitive measures. *Alz. Res. Therapy*, 12(1):1–10, December 2020. ISSN 1758-9193. doi: 10.1186/s13195-020-0581-1.

[113] Shangran Qiu, Matthew I. Miller, Prajakta S. Joshi, Joyce C. Lee, Chonghua Xue, Yunruo Ni, Yuwei Wang, Ileana De Anda-Duran, Phillip H. Hwang, Justin A. Cramer, Brigid C. Dwyer, Honglin Hao, Michelle C. Kaku, Sachin Kedar, Peter H. Lee, Asim Z. Mian, Daniel L. Murman, Sarah O'Shea, Aaron B. Paul, Marie-Helene Saint-Hilaire, E. Alton Sartor, Aneeta R. Saxena, Ludy C. Shih, Juan E. Small, Maximilian J. Smith, Arun Swaminathan, Courtney E. Takahashi, Olga Taraschenko, Hui You, Jing Yuan, Yan Zhou, Shuhan Zhu, Michael L. Alosco, Jesse Mez, Thor D. Stein, Kathleen L. Poston, Rhoda Au, and Vijaya B. Kolachalama. Multimodal deep learning for Alzheimer's disease dementia assessment. *Nature Communications*, 13(1):3404, June 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-31037-5.

[114] Dianbo Liu, Ming Zheng, and Nestor Andres Sepulveda. Using Artificial Neural Network Condensation to Facilitate Adaptation of Machine Learning in Medical Settings by Reducing Computational Burden: Model Design and Evaluation Study. *JMIR Formative Research*, 5(12):e20767, December 2021. doi: 10.2196/20767.

[115] Hyunseok Seo, Masoud Badiei Khuzani, Varun Vasudevan, Charles Huang, Hongyi Ren, Ruoxiu Xiao, Xiao Jia, and Lei Xing. Machine Learning Techniques for Biomedical Image Segmentation: An Overview of Technical Aspects

and Introduction to State-of-Art Applications. *Med. Phys.*, 47(5):e148, June 2020. doi: 10.1002/mp.13649.

[116] Omar Boursalie, Reza Samavi, and Thomas E. Doyle. Evaluation methodology for deep learning imputation models. *Exp. Biol. Med.*, 247(22):1972–1987, September 2022. ISSN 1535-3702. doi: 10.1177/15353702221121602.

[117] Huijuan Lu, Yige Xu, Minchao Ye, Ke Yan, Zhigang Gao, and Qun Jin. Learning misclassification costs for imbalanced classification on gene expression data. *BMC Bioinf.*, 20(25):1–10, December 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-3255-x.

[118] Steven A. Niederer and Nic P. Smith. Using physiologically based models for clinical translation: predictive modelling, data interpretation or something inbetween? *J. Physiol.*, 594(23):6849–6863, December 2016. ISSN 0022-3751. doi: 10.1113/JP272003.

[119] Mark Alber, Adrian Buganza Tepole, William R. Cannon, Suvranu De, Salvador Dura-Bernal, Krishna Garikipati, George Karniadakis, William W. Lytton, Paris Perdikaris, Linda Petzold, and Ellen Kuhl. Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *npj Digital Med.*, 2(115): 1–11, November 2019. ISSN 2398-6352. doi: 10.1038/s41746-019-0193-y.

[120] Raj Prabhu and Mark Horstemeyer. *Multiscale Biomechanical Modeling of the Brain.* Elsevier, Academic Press, 2021. ISBN 978-0-12-818144-7. doi: 10.1016/C2018-0-03194-1.

[121] Hongtao Yu, George P. Huang, Zifeng Yang, and Bryan R. Ludwig. A multiscale computational modeling for cerebral blood flow with aneurysms and/or stenoses. *Int. J. Numer. Methods Biomed. Eng.*, 34(10):e3127, October 2018. ISSN 2040-7947. doi: 10.1002/cnm.3127.

[122] Mason Kadem, Louis Garber, Mohamed Abdelkhalek, Baraa K. Al-Khazraji, and Zahra Keshavarz-Motamed. Hemodynamic Modeling, Medical Imaging, and Machine Learning and Their Applications to Cardiovascular Interventions. *IEEE Rev. Biomed. Eng.*, 16:403–423, January 2022. ISSN 1941-1189. doi: 10.1109/RBME.2022.3142058.