XGBoost for Interpretable Alzheimer's Decision Support

Mason Kadem[1,2,5], **Michael Noseworthy**[1,3,4], **and Thomas Doyle**[1,3,4,6]
[1] School of Biomedical Engineering, McMaster University, Hamilton, Canada
[2] School of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Canada
[3] Department of Electrical and Computer Engineering, McMaster University, Hamilton, Canada
[4] Department of Radiology, McMaster University, Hamilton, Canada
[5] Population Health Research Institute, Hamilton, Canada
[6] Vector Institute for Artificial Intelligence, Toronto, Canada
kademm@mcmaster.ca, nosewor@mcmaster.ca, doylet@mcmaster.ca

Abstract

Despite their necessity in directing patient care worldwide, simple and accurate diagnostic tools for early Alzheimer's disease (AD) do not exist. To support healthcare decision-making and planning, this research leverages large, multi-site accessible data and state-of-the-art supervised machine learning (XGBoost) to enable rapid, accurate, low-cost, accessible, non-invasive, interpretable, and early clinical evaluation of AD. Machine learning was employed to combine three key features: Everyday Cognition Questionnaire, Alzheimers Disease Assessment Scale, and Delayed Total Recall, achieving area under the receiver operating characteristic curves scores consistently above 97%. The selected features are important because they are non-invasive and easily collected. Low performance on delayed recall alone appears to distinguish most AD patients, consistent with the pathophysiology of AD where individuals having problems storing new information into long-term memory. Distinguishing this research from existing literature was the focus of enhancing the model's interpretability while maintaining performance of more complex and opaque models. The interpretable model enables understanding of the decision process, vital for clinical adoption of machine learning tools in AD evaluation. In summary, we present a methodology which identified accessible and noninvasive features, each with their absolute thresholds, together with a clinically operable decision route, to accurately and rapidly detect, differentiate, and diagnose Alzheimer's disease patients.

Introduction

Alzheimer's disease (AD) is among the top 10 causes of global mortality, and dementia imposes a yearly $1 trillion USD economic burden (World Health Organization 2021). Nearly 80% of dementia cases include adults living with AD (Livingston et al. 2020). Cognitive impairment manifests heterogeneously across patients with AD (Lyketsos et al. 2002), 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 (Guarino et al. 2019). Improving our understanding of risk profiles and having the ability to differentiate and diagnose AD patients would have significant impacts on relieving global disease burden (World Health Organization 2021; van der Flier et al. 2023). Well-established early markers of AD exist, including genetic markers like apolipoprotein E (APOE) (Yamazaki et al. 2019) and modifiable risk factors (e.g., cognitive, physical, and social) that may lower dementia risk or delay onset (Breiman 2001); however, even with these early markers, we have yet to develop simple tools that can rapidly and accurately detect and diagnose early AD patients. (Palmqvist et al. 2021; van der Flier et al. 2023). Based on our literature review, validated using AI-assisted methodology (van de Schoot et al. 2021), we found three main gaps in current AD-specific prediction models:

1. Models are based on traditional statistics approaches that overlook nonlinear relationships and complex interactions between features (Breiman 2001; Kadem et al. 2022). 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.
2. While machine learning has been used to predict AD, the machine learning-based predictive models have included manually selected, difficult to acquire, costly (Tang et al. 2021; Pellegrini et al. 2018; Bron et al. 2021; Basaia et al. 2019), and invasive measures (e.g., cerebrospinal fluid analysis of β-amyloid (Aβ42) (Palmqvist et al. 2014), Aβ-positron emission tomography (Rabinovici et al. 2019; Mattsson et al. 2019) which limits the availability of these biomarkers and their usage. While identifying high-risk markers is useful for risk assessment, determining threshold cut-off values for these factors offers further clinical utility by defining informed decision routes, and would further aid in the understanding of the decision process.
3. Machine learning models often lack interpretability (Zihni et al. 2020; Pellegrini et al. 2018). The complexity of machine learning algorithm hinders the formation of clinically quantifiable explanations of the decision process, thus impeding clinical adoption. Clinical adoption and end-user trust of machine learning models demand understandable explanation of decision processes (Hall and Gill 2019). Although explainability and interpretability are often used interchangeably within the

machine learning research domain, herein, we differentiate the two terms. Specifically, in this paper, 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)(Rudin 2019). Prediction models must balance performance and interpretability, posing a challenge when improving machine learning interpretability while maintaining model performance.

Taken together, existing methods are generally lacking accurate, practical, understandable, and economical AD-specific diagnostic prediction models. Furthermore, such models should be based on easily accessible, non-invasive, and low-cost features. The research herein leverages the *Alzheimer's Disease Neuroimaging Initiative (ADNI) database and a mathematical modelling approach based on state of the art supervised machine learning to identify 1) predictive markers of AD, and 2) patients at the highest probability of having AD. Furthermore, the paper focuses on the interpretability of the machine learning model and its predictions to improve the clinical uptake of AD-specific machine learning models.

Methods

Data Sources

Data for model development (n=441) included retrospectively studied multi site (57+ US and Canadian sites) heterogeneous tabular data (i.e., brain imaging, genetic, neuropsychological testing, lifestyle and health history). ADNI research data (http://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 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 data on the confirmed presence or absence of AD. Overall, 43 features were considered. For more details on the methodology and features, please refer to (Kadem 2023).

All data were acquired from the Alzheimer's Disease Neuroimaging Initiative 3 database (ADNI 3) (http://adni.loni.usc.edu/). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessments can be combined to assess the progression of MCI and early AD.

Modelling

The XGBoost algorithm optimizes an objective function:

$$Obj = Loss + \Omega \tag{1}$$

where *Obj* is the objective function we aim to optimize, *Loss* is the prediction error calculated using a loss function, and Ω is the regularization term to prevent overfitting.

For our binary classification tasks, we used the log loss function:

$$Loss = \sum_i [y_i \log(1+e^{-y_{pred_i}}) + (1-y_i)\log(1+e^{y_{pred_i}})] \tag{2}$$

where *Loss* is the sum of log loss function for binary classification tasks, y_i is the true label for the i^{th} instance, and y_{pred_i} is the predicted label for the i^{th} instance.

The regularization term (Ω) helps prevent overfitting:

$$\Omega = \gamma * T + 0.5 * \lambda * \sum_{j=1}^{T} w_j^2 \tag{3}$$

where Ω is the regularization term, γ is a parameter that controls the complexity cost for each additional tree, T is the total number of trees, λ is the L2 regularization parameter, and w_j is the weight of the j-th leaf.

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

$$Obj(t) \approx \sum_{i=1}^{n} [g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

where $Obj(t)$ is the approximate objective function at the t-th iteration of the algorithm, with t representing the current iteration step, and T the total number of trees in the model at the t-th iteration, g_i and h_i are the gradient and hessian of the loss function with respect to y_{pred_i} for instance i, and $w_{q(x_i)}$ is the weight of the leaf corresponding to instance i.

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

$$g_i = \frac{\partial Loss}{\partial y_{pred_i}}, \quad h_i = \frac{\partial^2 Loss}{\partial y_{pred_i}^2} \tag{5}$$

where g_i is the gradient of the loss function with respect to y_{pred_i} for instance i, and h_i is the hessian (second derivative) of the loss function with respect to y_{pred_i} for instance i.

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

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \tag{6}$$

where *Gain* is the gain of a feature split, G_L and G_R are the sum of instance gradients in the left and right child, respectively, after the split, H_L and H_R are the sum of instance

hessians in the left and right child, respectively, after the split, λ is the L2 regularization term, and γ is the complexity cost parameter.

Machine Learning Process

We implemented a supervised XGboost classifier (Chen and Guestrin 2016), which is a more regularized form of the stochastic gradient boosting ensemble algorithm that reportedly improves performance, scalability, and efficiency when compared to its predecessor. We motivate selection of XGBoost using both theoretical and empirical evidence from literature (Lundberg et al. 2020; Borisov et al. 2022). 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. 1 . Because of its iterative decision tree-based architecture, XGBoost can provide a high level of interpretablility.

Feature Selection, Classification, and Evaluation Feature selection allows us to reduce the dimensionality and complexity of the data and model, respectively. We begin the machine learning process (Fig. 1) by elucidating which features were important in AD risk. We first split the training dataset into 70% training and 30% test subsets. The average XGboost-based feature importance was calculated based on each feature's gain (as defined in (Eq. 6) or proportionate split-induced contribution to the model, computed for each tree in the model. A higher gain value for a feature is indicative of greater performance (reduction in objective function) and thus greater feature importance. This process was repeated 100 times with different random number seeds (0-99) to stabilize features. Thereafter, we selected the top 10 ranked features.

To further reduce dimensionality and complexity while preserving the model's performance, we evaluated the ROC AUC performance of the top-ranked feature (N) through 100 iterations of fivefold stratified cross-validation (Table 1). We then included the next ranked feature (N + 1) and recalculated the mean ROC-AUC. If the validation score with this additional feature surpassed the previous score, we kept the additional feature. Otherwise, the process stopped at the last added feature, resulting in a minimal feature set. To improve interpretability while maintaining model performance, this minimal feature set was then used within a boosted interpretability XGBoost model.

Boosting Interpretability

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. While XGBoost is known for its interpretability, its regularization, ensemble and subsampling process can limit its transparency. XGBoost constructs decision trees recursively from pseudo residuals and can identify decision trees that contribute most to a predictive model. If the performance of a model remains high, lowering the model's complexity might yield a transparent decision path. To boost the interpretability of XGBoost and offer a transparent decision path, we restricted the number of estimators to 1, ensured all data were utilized by eliminating the stochastic sampling of features and instances, and removed regularization. This ensured our model's decisions were solely driven by the data and easier to understand.

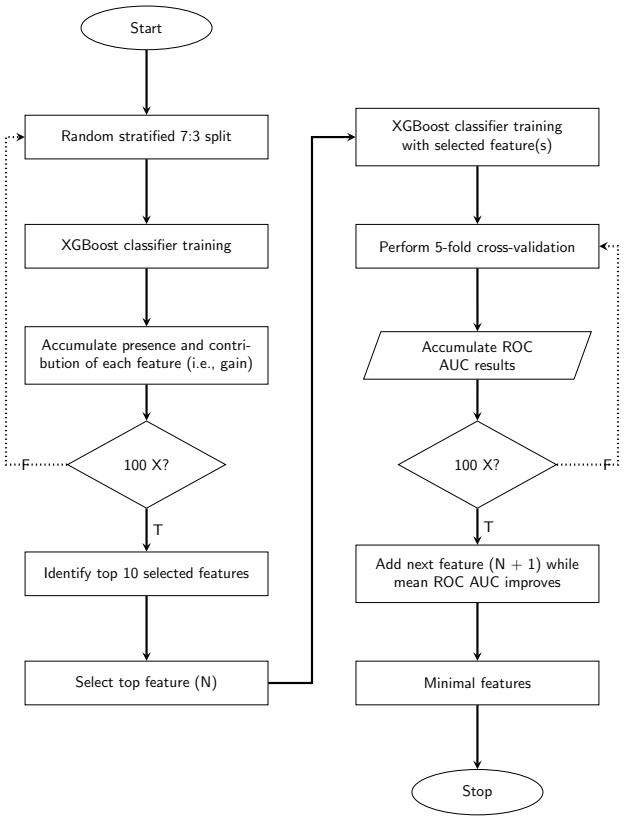

Figure 1: Machine Learning Process Flow Diagram

Results

Baseline Characteristics

Compared to clinically normal controls, AD patients were more likely to be 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.

XGBoost Classifiers' Performance

The objective of the XGboost classifier was to identify diagnostic markers of AD and patients at the highest risk of AD. Our two stage feature selection and evaluation process selected three features with ROC AUC performance consistently above 97% (Table 1). The model can reliably detect, differentiate and diagnose patients with AD.

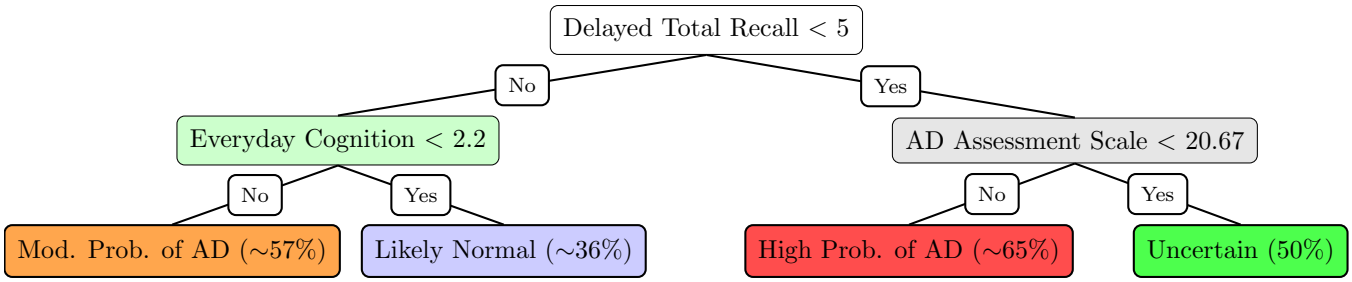

Figure 2: Boosting the interpretability of XGBoost. By restricting the number of estimators to 1, utilizing all data without stochastic sampling, and removing regularization, our model's decisions are solely driven by the data, offering a transparent and actionable path for high stakes clinical decision-making. The prior feature selection using the full ensemble of boosted trees captured complex interactions and nonlinear relationships, enabling identification of a highly predictive minimal feature set for the boosted interpretable XGBoost model.

Features	**AUC Train %**	**AUC Valid %**
ECog-T	97.3 ± 0.7	94.8 ± 3.7
ECog-T, ADA	99.7 ± 0.4	96.8 ± 2.9
ECog-T, ADA, DTR	99.5 ± 0.4	97.7 ± 2.5

Table 1: XGBoost classification performance for AD diagnosis averaged over 100 iterations of fivefold stratified cross validation.

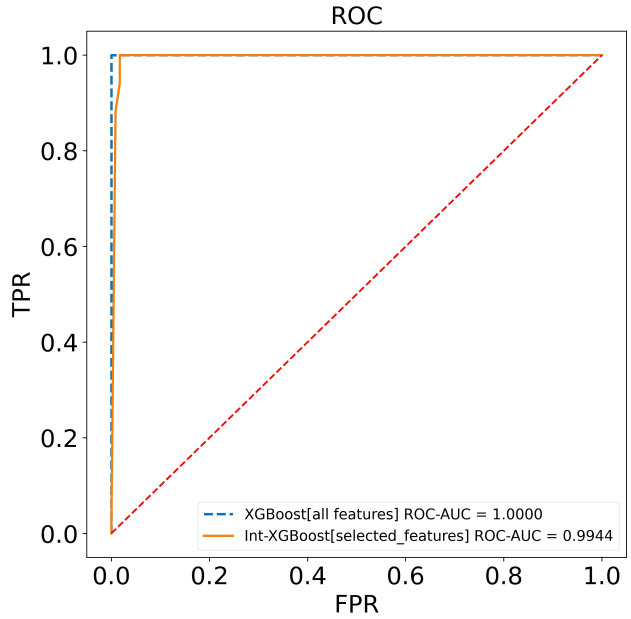

Figure 3: Area under the curve (0 - 1) of the receiver operating curves (ROC) for validation, comparing XGBoost and the boosted interpretable model/decision path with selected features. Plotting the true positive rate (TPR) against the false positive rate (FPR) at various thresholds (0-1) creates the ROC curves. This value indicates how well the model classifies patients who end up developing AD. Values closer to 1 indicate better classification performance.

Boosted Interpretability XGBoost and Data-driven Decision Path

Figure 2 depicts the boosted interpretability model and decision path. The decision path begins with Delayed Total Recall (< 5), which splits into two paths. If the answer is 'Yes', the AD Assessment Scale (≤ 20) is checked. A 'Yes' indicates an 'Uncertain' diagnosis with a 50% AD probability (log-odds = 0), while a 'No' suggests a high probability of AD (65%, log-odds = 0.60). If the answer is 'No', the model considers Everyday Cognition (< 2.2). A 'Yes' suggests a 'Likely Normal' diagnosis with a 36% AD probability (log-odds = -0.5953), and a 'No' indicates a moderate AD probability of 57% (log-odds = 0.30).

Figure 3 compares the performance between the boosted interpretability XGBoost model (Int-XGBoost) and the full XGBoost ensemble (default parameters with `col_sample` and `subsample` set to 0.9 to reduce overfitting (Chen and Guestrin 2016). Int-XGboost maintained performance while improving interpretability. Both Figures 2 and 3 were based on a random stratified 7:3 split of the dataset.

Discussion

Below we discuss six contributions of this work that address the aforementioned limitations in existing AD-specific prediction models.

1. Our work extends beyond the traditional statistical approaches in identifying high-risk factors, by leveraging XGBoost to capture complex interactions and non-linear relationships among features.
2. The most prominent factors are easily acquired remotely without the need of invasive or costly procedures (e.g., lumbar punctures, PET).
3. Establishing absolute thresholds greatly enhance the utility of these risk factors.
4. The elucidation of the machine learning decision process and the coupling of absolute thresholds and decision path may increase clinical utility, broader adoption, and improve trust.
5. We further highlight the ability of XGBoost to preserve

performance while increasing its interpretability by constraining its estimators (Yan et al. 2020).

6. This work highlights the advantage of gradient boosted ensembles, in performance and interpretability for heterogeneous tabular data and particularly in the context of predicting AD risk.

The XGBoost 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 values consistently above 97% (Figure 1, surpassing the performance of the latest AD-specific model (Rye et al. 2022). Our findings support the idea that memory paradigms, such as delayed total recall, can serve as effective measures to track neurodegenerative disease trajectories (Lindbergh et al. 2021). Moreover, these features may be sensitive to neurodegenerative disease trajectories more so than Mini-Mental State Examination or Montreal Cognitive Assessment. The interpretable model maintained or improved performance over their complex counterparts while improving their interpretability (Fig. 3). Detecting AD early in patients might prompt clinicians to pursue earlier care, referrals, symptomatic treatment, and soon interventions. Further, the simple, accessible, accurate and expedited detection, differentiation and diagnosis of AD reduces healthcare system costs. Taken together, this combination of diagnostic markers and their absolute thresholds 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.

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 data, external validation from a different continent is necessary to thoroughly assess the generalizability of the features and model. We look forward to incorporating additional data from additional sites to further improve the performance of the model. We compromised performance for interpretability by using a minimal number of clinical features as clinical settings favor interpretable models.

Conclusion

Our work extends beyond traditional statistical approaches by leveraging XGBoost to capture complex interactions and non-linear relationships in identifying high-risk factors. We identified accessible, cost-effective, and noninvasive clinical features with ROC AUC scores consistently above 97%. By establishing absolute thresholds for these features, coupled with a transparent decision path, we enhance interpretability, clinical utility and trust.

Acknowledgements

The authors would like to acknowledge the financial support of the Canadian Department of National Defence Innovation for Defence Excellence and Security (DND IDEaS) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

References

Basaia, S.; Agosta, F.; Wagner, L.; Canu, E.; Magnani, G.; Santangelo, R.; Filippi, M.; and Alzheimer's Disease Neuroimaging Initiative. 2019. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage. Clinical*, 21: 101645.

Borisov, V.; Leemann, T.; Seßler, K.; Haug, J.; Pawelczyk, M.; and Kasneci, G. 2022. Deep Neural Networks and Tabular Data: A Survey. *IEEE Trans. Neural Networks Learn. Syst.*, 1–21.

Breiman, L. 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3): 199–231. Publisher: Institute of Mathematical Statistics.

Bron, E. E.; Klein, S.; Papma, J. M.; Jiskoot, L. C.; Venkatraghavan, V.; Linders, J.; Aalten, P.; De Deyn, P. P.; Biessels, G. J.; Claassen, J. A. H. R.; Middelkoop, H. A. M.; Smits, M.; Niessen, W. J.; van Swieten, J. C.; van der Flier, W. M.; Ramakers, I. H. G. B.; van der Lugt, A.; Alzheimer's Disease Neuroimaging Initiative; and Parelsnoer Neurodegenerative Diseases study group. 2021. Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage. Clinical*, 31: 102712.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-45034232-2.

Guarino, A.; Favieri, F.; Boncompagni, I.; Agostini, F.; Cantone, M.; and Casagrande, M. 2019. Executive Functions in Alzheimer Disease: A Systematic Review. *Frontiers in Aging Neuroscience*, 10: 437.

Hall, P.; and Gill, N. 2019. *An introduction to machine learning interpretability*. O'Reilly Media, Incorporated.

Kadem, M. 2023. *Interpretable Machine Learning in Alzheimer's Disease Dementia*. Master's thesis, McMaster University.

Kadem, M.; Garber, L.; Abdelkhalek, M.; Al-Khazraji, B. K.; and Keshavarz-Motamed, Z. 2022. Hemodynamic Modeling, Medical Imaging, and Machine Learning and Their Applications to Cardiovascular Interventions. *IEEE Rev. Biomed. Eng.*, 16: 403–423.

Lindbergh, C. A.; Walker, N.; La Joie, R.; Weiner-Light, S.; Staffaroni, A. M.; Casaletto, K. B.; Elahi, F.; Walters, S. M.; You, M.; Cotter, D.; Asken, B.; Apple, A. C.; Tsoy, E.; Neuhaus, J.; Fonseca, C.; Wolf, A.; Cobigo, Y.; Rosen, H.; Kramer, J. H.; and Network, H. A. 2021. 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.

Livingston, G.; Huntley, J.; Sommerlad, A.; Ames, D.; Ballard, C.; Banerjee, S.; Brayne, C.; Burns, A.; Cohen-Mansfield, J.; Cooper, C.; Costafreda, S. G.; Dias, A.; Fox, N.; Gitlin, L. N.; Howard, R.; Kales, H. C.; Kivimäki, M.; Larson, E. B.; Ogunniyi, A.; Orgeta, V.; Ritchie, K.; Rockwood, K.; Sampson, E. L.; Samus, Q.; Schneider, L. S.; Selbæk, G.; Teri, L.; and Mukadam, N. 2020. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet (London, England)*, 396(10248): 413–446.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S.-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, 2(1): 56–67.

Lyketsos, C. G.; Lopez, O.; Jones, B.; Fitzpatrick, A. L.; Breitner, J.; and DeKosky, S. 2002. Prevalence of neuropsychiatric symptoms in dementia and mild cognitive impairment: results from the cardiovascular health study. *JAMA*, 288(12): 1475–1483.

Mattsson, N.; Palmqvist, S.; Stomrud, E.; Vogel, J.; and Hansson, O. 2019. Stagingβ-Amyloid Pathology With Amyloid Positron Emission Tomography. *JAMA Neurology*, 76(11): 1319–1329.

Palmqvist, S.; Tideman, P.; Cullen, N.; Zetterberg, H.; Blennow, K.; Dage, J. L.; Stomrud, E.; Janelidze, S.; Mattsson-Carlgren, N.; and Hansson, O. 2021. Prediction of future Alzheimer's disease dementia using plasma phospho-tau combined with other accessible measures. *Nature Medicine*, 27(6): 1034–1042. Number: 6 Publisher: Nature Publishing Group.

Palmqvist, S.; Zetterberg, H.; Blennow, K.; Vestberg, S.; Andreasson, U.; Brooks, D. J.; Owenius, R.; Hägerström, D.; Wollmer, P.; Minthon, L.; and Hansson, O. 2014. Accuracy of brain amyloid detection in clinical practice using cerebrospinal fluid β-amyloid 42: a cross-validation study against amyloid positron emission tomography. *JAMA neurology*, 71(10): 1282–1289.

Pellegrini, E.; Ballerini, L.; Hernandez, M. D. C. V.; Chappell, F. M.; González-Castro, V.; Anblagan, D.; Danso, S.; Muñoz-Maniega, S.; Job, D.; Pernet, C.; Mair, G.; MacGillivray, T. J.; Trucco, E.; and Wardlaw, J. M. 2018. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, 10: 519–535.

Rabinovici, G. D.; Gatsonis, C.; Apgar, C.; Chaudhary, K.; Gareen, I.; Hanna, L.; Hendrix, J.; Hillner, B. E.; Olson, C.; Lesman-Segev, O. H.; Romanoff, J.; Siegel, B. A.; Whitmer, R. A.; and Carrillo, M. C. 2019. 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.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5): 206–215.

Rye, I.; Vik, A.; Kocinski, M.; Lundervold, A. S.; and Lundervold, A. J. 2022. Predicting conversion to Alzheimer's disease in individuals with Mild Cognitive Impairment using clinically transferable features. *Sci. Rep.*, 12(15566): 1–11.

Tang, L.; Wu, X.; Liu, H.; Wu, F.; Song, R.; Zhang, W.; Guo, D.; Feng, J.; and Li, C. 2021. 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.

van de Schoot, R.; de Bruin, J.; Schram, R.; Zahedi, P.; de Boer, J.; Weijdema, F.; Kramer, B.; Huijts, M.; Hoogerwerf, M.; Ferdinands, G.; Harkema, A.; Willemsen, J.; Ma, Y.; Fang, Q.; Hindriks, S.; Tummers, L.; and Oberski, D. L. 2021. An open source machine learning framework for efficient and transparent systematic reviews. 3(2): 125–133. Number: 2 Publisher: Nature Publishing Group.

van der Flier, W. M.; de Vugt, M. E.; Smets, E. M. A.; Blom, M.; and Teunissen, C. E. 2023. Towards a future where Alzheimer's disease pathology is stopped before the onset of dementia. *Nat. Aging*, 3(5): 494–505.

World Health Organization. 2021. *Global status report on the public health response to dementia*. Geneva: World Health Organization. ISBN 978-92-4-003324-5. Section: xv, 251 p.

Yamazaki, Y.; Zhao, N.; Caulfield, T. R.; Liu, C.-C.; and Bu, G. 2019. Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. *Nature Reviews Neurology*, 15(9): 501–518. Number: 9 Publisher: Nature Publishing Group.

Yan, L.; Zhang, H.-T.; Goncalves, J.; Xiao, Y.; Wang, M.; Guo, Y.; Sun, C.; Tang, X.; Jing, L.; Zhang, M.; Huang, X.; Xiao, Y.; Cao, H.; Chen, Y.; Ren, T.; Wang, F.; Xiao, Y.; Huang, S.; Tan, X.; Huang, N.; Jiao, B.; Cheng, C.; Zhang, Y.; Luo, A.; Mombaerts, L.; Jin, J.; Cao, Z.; Li, S.; Xu, H.; and Yuan, Y. 2020. An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.*, 2(5): 283–288.

Zihni, E.; Madai, V. I.; Livne, M.; Galinovic, I.; Khalil, A. A.; Fiebach, J. B.; and Frey, D. 2020. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *PLOS ONE*, 15(4): e0231166. Publisher: Public Library of Science.