

AI-based differential diagnosis of dementia etiologies on multimodal data

Chonghua Xue^{1,2,*}, Sahana S. Kowshik^{1,3,*}, Diala Lteif^{1,4}, Shreyas Puducheri¹, Varuna H. Jasodanand¹, Olivia T. Zhou¹, Anika S. Walia¹, Osman B. Guney^{1,2}, J. Diana Zhang^{1,5}, Serena T. Pham⁶, Artem Kaliaev⁶, V. Carlota Andreu-Arasa⁶†, Brigid C. Dwyer⁷†, Chad W. Farris⁶†, Honglin Hao⁸†, Sachin Kedar⁹†, Asim Z. Mian⁶†, Daniel L. Murman¹⁰†, Sarah A. O’Shea¹¹†, Aaron B. Paul¹²†, Saurabh Rohatgi¹²†, Marie-Helene Saint-Hilaire⁷†, Emmett A. Sartor⁷†, Bindu N. Setty⁶†, Juan E. Small¹³†, Arun Swaminathan¹⁴†, Olga Taraschenko¹⁰†, Jing Yuan⁸†, Yan Zhou⁸†, Shuhan Zhu¹⁵†, Cody Karjadi¹⁶, Ting Fang Alvin Ang^{16,17}, Sarah A. Bargal¹⁸, Bryan A. Plummer⁴, Kathleen L. Poston¹⁹, Meysam Ahangaran¹, Rhoda Au^{1,7,16,17,20,21} & Vijaya B. Kolachalama^{1,3,4,20,‡}

¹*Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA*

²*Department of Electrical & Computer Engineering, Boston University, MA, USA*

³*Faculty of Computing & Data Sciences, Boston University, MA, USA*

⁴*Department of Computer Science, Boston University, MA, USA*

⁵*School of Chemistry, University of New South Wales, Sydney, Australia*

⁶*Department of Radiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA*

⁷*Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA*

⁸*Department of Neurology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China*

⁹*Departments of Neurology & Ophthalmology, Emory University School of Medicine, Atlanta, GA, USA*

¹⁰*Department of Neurological Sciences, University of Nebraska Medical Center, Omaha, NE, USA*

¹¹*Department of Neurology, Columbia University Irving Medical Center, New York, NY, USA*

¹²*Department of Radiology, Massachusetts General Hospital, Boston, MA, USA*

¹³*Department of Radiology, Lahey Hospital & Medical Center, Burlington, MA, USA*

¹⁴*Department of Neurology, SSM Health, Madison, WI, USA*

¹⁵*Department of Neurology, Brigham & Women’s Hospital, Boston, MA, USA*

¹⁶*The Framingham Heart Study, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA*

¹⁷*Department of Anatomy and Neurobiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA*

¹⁸*Department of Computer Science, Georgetown University, Washington DC, USA*

¹⁹*Department of Neurology, Stanford University, Palo Alto, CA, USA*

²⁰*Boston University Alzheimer’s Disease Research Center, Boston, MA, USA*

²¹*Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA*

* These authors contributed equally to this work

† Listed in alphabetical order

‡ Corresponding author: Vijaya B. Kolachalama, PhD; Email: vkola@bu.edu; ORCID: <https://orcid.org/0000-0002-5312-8644>

1 **Abstract**

2 Differential diagnosis of dementia remains a challenge in neurology due to symptom overlap across etiolo-
3 gies, yet it is crucial for formulating early, personalized management strategies. Here, we present an AI
4 model that harnesses a broad array of data, including demographics, individual and family medical history,
5 medication use, neuropsychological assessments, functional evaluations, and multimodal neuroimaging, to
6 identify the etiologies contributing to dementia in individuals. The study, drawing on 51,269 participants
7 across 9 independent, geographically diverse datasets, facilitated the identification of 10 distinct dementia
8 etiologies. It aligns diagnoses with similar management strategies, ensuring robust predictions even with
9 incomplete data. Our model achieved a micro-averaged area under the receiver operating characteristic
10 curve (AUROC) of 0.94 in classifying individuals with normal cognition, mild cognitive impairment and
11 dementia. Also, the micro-averaged AUROC was 0.96 in differentiating the dementia etiologies. Our model
12 demonstrated proficiency in addressing mixed dementia cases, with a mean AUROC of 0.78 for two co-
13 occurring pathologies. In a randomly selected subset of 100 cases, the AUROC of neurologist assessments
14 augmented by our AI model exceeded neurologist-only evaluations by 26.25%. Furthermore, our model
15 predictions aligned with biomarker evidence and its associations with different proteinopathies were sub-
16 stantiated through postmortem findings. Our framework has the potential to be integrated as a screening
17 tool for dementia in various clinical settings and drug trials, with promising implications for person-level
18 management.

1 Dementia is one of the most pressing health challenges of our time. With nearly 10 million new
2 cases reported annually, this syndrome, characterized by a progressive decline in cognitive function severe
3 enough to impede daily life activities, continues to present considerable clinical and socioeconomic chal-
4 lenges. In 2017, the World Health Organization's global action plan highlighted the need for prompt and
5 precise diagnosis of dementia as a pivotal strategic objective in response to the growing number of dementia
6 worldwide.^{1,2} As such, diagnostic precision in the varied landscape of dementia remains a critical, yet unmet
7 need, particularly as the global population ages and the demand for more accurate participant screening in
8 drug trials increases.³ This challenge primarily stems from the overlapping clinical presentation of different
9 dementia types, which is further complicated by the heterogeneity in findings on magnetic resonance imag-
10 ing (MRI) scans.^{4,5} The necessity for improvements in the field becomes ever more pressing considering the
11 projected shortage of specialists including neurologists, neuropsychologists and geriatric care providers,⁶⁻⁸
12 emphasizing the urgency to innovate and evolve our diagnostic tools.

13 Accurate differential diagnosis of dementia is pivotal for prescribing targeted therapeutic interven-
14 tions, enhancing treatment efficacy and slowing symptom progression. While Alzheimer's disease (AD) is
15 a leading cause, other forms such as vascular dementia (VD), Lewy body dementia (LBD), and frontotem-
16 poral dementia (FTD) are also prevalent.⁹⁻¹¹ These etiologies can often coexist, as marked by symptom
17 overlap and variable symptom intensity, which further complicate the diagnostic process.¹² Importantly,
18 diagnostic errors are prevalent among older adults, particularly those with comorbid conditions.¹³ These
19 misdiagnoses can translate into inappropriate medication use and adverse health outcomes.¹⁴ For example,
20 while patients with early-stage AD may be candidates for anti-amyloid therapies,¹⁵⁻¹⁷ the coexistence of
21 pathology from other etiologies, such as vascular dementia, can increase the risk of amyloid-related imag-
22 ing abnormalities.¹⁸ This highlights the critical need for accurately assessing the full spectrum of etiological
23 factors contributing to dementia to inform appropriate therapeutic strategies and optimize patient care.¹⁹

24 The imperative for scalable diagnostic tools in AD and related dementias is becoming increasingly
25 urgent, given the significant challenges in accessing gold-standard testing. Recent regulatory approvals have
26 facilitated the transition of cerebrospinal fluid (CSF) and positron emission tomography (PET) biomarkers
27 from research environments to clinical settings. While promising, the clinical integration of accurate blood-
28 based biomarkers remains an area of active research.²⁰⁻²² Despite these advancements, accessibility to these
29 diagnostic tools is still constrained, not only in remote and economically developing regions but also in
30 urban healthcare centers, as exemplified by prolonged waiting periods for specialist consultations.²³ This
31 challenge is compounded by a global shortage of specialists, such as behavioral neurologists and neuropsy-
32 chologists, leading to an overreliance on cognitive assessments that may not be culturally appropriate due to
33 the lack of formal training programs in neuropsychology in many parts of the world.^{24,25} Although conven-
34 tional methods like clinical evaluations, neuropsychological testing, and MRI remain central to antemortem
35 differential dementia diagnosis, their effectiveness relies on a diminishing pool of specialist clinicians. This
36 underscores an urgent need for healthcare systems to evolve and adapt to the rapidly changing dynamics of
37 dementia diagnosis and treatment.

38 Machine learning (ML) has the potential to enhance the accuracy and efficiency of dementia diagnosis.²⁶⁻²⁸
39 Previous ML methods have largely focused on leveraging neuroimaging data to distinguish cognitively nor-
40 mal (NC) individuals from those with mild cognitive impairment (MCI) and dementia (DE), with AD being
41 the main etiology given its ubiquity in dementia diagnosis.^{29,30} A few studies have attempted to discern
42 neuroimaging signatures unique to AD by contrasting them with other dementia types³¹⁻⁴⁰ However, this
43 primary emphasis on AD can have limited practical implications given the prevalence and co-occurrence

44 of other etiologies. In addition, a focus on imaging data alone can be insufficient in providing a holistic
45 understanding of an individual's neurological condition. Recently, we proposed a novel approach to stratify
46 individuals based on cognitive status and discern likely AD cases from non-AD dementia types by incor-
47 porating imaging with non-imaging data such as demographics, medical histories, and neuropsychological
48 assessments.³⁹ These investigations have begun to illuminate the complex matrix of factors contributing to
49 dementia. However, for ML models to be adopted into clinical practice, they must be able to accommodate
50 the intricacies of mixed etiologies, as well as the inclusion or exclusion of different data modalities that may
51 or may not be available. Therefore, the development of AI methodologies capable of harnessing multimodal
52 data facilitates the accurate quantification of diverse dementia etiologies, irrespective of clinical resources,
53 thereby aligning treatment strategies with individual patient profiles.

54 In this study, we propose a multimodal machine learning framework that harnesses a diverse array
55 of data, including demographics, personal and family medical history, medication use, neuropsycholog-
56 ical assessments, functional evaluations, and multimodal neuroimaging to perform differential dementia
57 diagnosis. Our model, designed to mirror real-world scenarios, aligns diagnoses with similar management
58 strategies and outputs probabilities for each etiology. This approach is intended to mimic clinical reasoning
59 and aid practitioners in dementia screening and treatment planning. The model's robustness is demonstrated
60 through validation on independent, geographically diverse datasets. In comparative analyses, we found
61 that AI-augmented clinician assessments achieved superior diagnostic accuracy compared to clinician-only
62 assessments. By validating our model against gold-standard biomarker and postmortem data for different
63 etiologies, we further emphasize our model's ability to dissect the intricate pathophysiology underlying de-
64 mentia. Our algorithmic framework demonstrates the potential to enhance dementia screening in various
65 clinical settings, illustrating AI's capacity to improve healthcare outcomes.

1 Results

Glossary 1

Acronym	Description
NC	Normal cognition
MCI	Mild cognitive impairment
DE	Dementia
AD	Alzheimer's disease
LBD	Lewy body dementia including dementia with Lewy bodies and Parkinson's disease dementia
VD	Vascular dementia, vascular brain injury, and vascular dementia including stroke
PRD	Prion disease including Creutzfeldt-Jakob disease
FTD	Frontotemporal lobar degeneration and its variants, including primary progressive aphasia, corticobasal degeneration and progressive supranuclear palsy, and with or without amyotrophic lateral sclerosis
NPH	Normal pressure hydrocephalus
SEF	Systemic and environmental factors including infectious diseases (HIV included), metabolic, substance abuse / alcohol, medications, systemic disease, and delirium
PSY	Psychiatric conditions including schizophrenia, depression, bipolar disorder, anxiety, and post-traumatic stress disorder
TBI	Moderate/severe traumatic brain injury, repetitive head injury, and chronic traumatic encephalopathy
ODE	Other dementia conditions including neoplasms, Down syndrome, multiple systems atrophy, Huntington's disease, seizures, etc.

2

3 Leveraging the power of multimodal data obtained from various cohorts (Tables 1 & S1 - S6), our
4 model adopts a nuanced approach to differential dementia diagnosis (Fig. 1). Our framework assigns in-
5 dividuals to one or more of thirteen diagnostic categories (refer to Glossary 1), which were meticulously
6 defined through consensus among a team of expert neurologists. This practical categorization is designed
7 with clinical management pathways in mind, thereby echoing real-world scenarios. For instance, we have
8 grouped dementia with Lewy bodies and Parkinson's disease dementia under the comprehensive category
9 of Lewy body dementia (LBD). This classification stems from an understanding that the care for these con-
10 ditions often follows a similar path, typically overseen by a multidisciplinary team of movement disorder
11 specialists. In the context of vascular dementia (VD), we included individuals who exhibited symptoms of a
12 stroke, possible or probable VD, or vascular brain injury. This encompassed cases with symptomatic stroke,
13 cystic infarct in cognitive networks, extensive white matter hyperintensity, and/or executive dysfunction as
14 the primary contributors to the observed cognitive impairment. The inclusion criteria were based on the ex-
15 pectation that such patients would typically receive care from clinicians specializing in stroke and vascular
16 diseases. Likewise, we have considered various psychiatric conditions, such as schizophrenia, depression,
17 bipolar disorders, anxiety, and post-traumatic stress disorder, under one category (PSY), acknowledging that
18 their management predominantly falls within the realm of psychiatric care providers. By aligning diagnostic
19 categories with clinical care pathways, our model serves not only to classify an individual's condition but

20 also to direct appropriate clinical management strategies.

21 **Model performance on NC, MCI and DE** We first sought to evaluate the performance of the model
22 on test cases comprising individuals along the cognitive spectrum of NC, MCI and DE. The receiver op-
23 erating characteristic (ROC) and precision-recall (PR) curves reflected strong model performance across
24 different averaging methods (Figs. 2a & 2b). In the test set, comprising the NACC dataset unused in train-
25 ing, the Alzheimer’s Disease Neuroimaging Initiative (ADNI) and the Framingham Heart Study (FHS) data,
26 our model demonstrated robust classification abilities for NC, MCI, and DE, achieving a micro-averaged
27 AUROC of 0.94 and a micro-averaged AUPR of 0.90. Additionally, the macro-averaged metrics showed
28 an AUROC of 0.93 and an AUPR value of 0.84. The weighted-average AUROC and AUPR values further
29 demonstrated the model’s efficacy, standing at 0.94 and 0.87, respectively. Detailed model performance met-
30 rics across the three test cohorts are provided in Table S7. We also evaluated our model’s effectiveness by
31 benchmarking it against a baseline machine learning algorithm, CatBoost,⁴¹ using identical case sets. This
32 comparison was executed over two feature subsets, revealing that our model and CatBoost exhibited simi-
33 lar performances on the NACC dataset. Conversely, on the ADNI and FHS datasets, our model surpassed
34 CatBoost, achieving higher AUROC and AUPR scores across all diagnostic categories with improvements
35 ranging from 0.02 to 0.21 for AUROC and 0.03 to 0.17 for AUPR, as detailed in Table S8. This compar-
36 ison highlights the improved generalizability of our model over traditional machine learning approaches in
37 diagnostic tasks.

38 **Model performance on incomplete data** To evaluate the model’s resilience to incomplete data, we
39 artificially introduced varying levels of data missingness in the NACC cohort and assessed the impact on
40 its predictive performance by selectively removing portions of the data to simulate different constraints.
41 As depicted in the chord diagram (Fig. 2c), even when confronted with missing features, whether it be
42 MRIs, UPDRS, GDS, NPI-Q, FAQ, NP tests or other parameters, our model consistently produced reliable
43 scores. This reinforces not only its predictive stability, but also its potential applicability in various clinical
44 scenarios where complete datasets are generally unattainable. Examples of this are found in our results on
45 ADNI and FHS, which we used as external testing datasets (Tables S4 & S5). The ADNI cohort exhibited
46 approximately 69% missing data compared to NACC, yet model predictions achieved a weighted-average
47 AUROC of 0.91 and AUPR of 0.86 for NC, MCI, and DE categories. Similarly, with 94% fewer features
48 than NACC, the model’s performance on FHS data also resulted in weighted-average AUROC and AUPR
49 scores of 0.68 and 0.53 for NC, MCI, and DE categories, respectively.

50 **Model alignment with prodromal disease** We aimed to evaluate the model’s efficacy in identifying
51 MCI individuals with an etiological diagnosis of AD, comparing the model’s predicted AD probabilities,
52 $P(AD)$, between MCI cases with AD as either a primary or contributing cause of their impairment and an
53 etiological diagnosis of AD. Despite the fact that our model was trained on identifying AD dementia rather
54 than prodromal AD, we found that it consistently assigned higher $P(AD)$ to cases with MCI due to AD,
55 compared to those with MCI due to other factors (Fig. 2d & Table S9). These results highlight our model’s
56 clinical relevance in facilitating early disease detection and aiding clinicians in making informed decisions.
57 Specifically, our findings support a therapeutic strategy of preemptive intervention in the AD continuum.

58 **Model alignment with clinical dementia ratings** We conducted a comparison between the model's
59 predicted DE probability scores, $P(DE)$, and the clinical dementia ratings (CDR) available for all partici-
60 pants in the NACC testing, and ADNI cohorts (Figs. 2d & 2e, Table S10). Despite not incorporating CDR as
61 input during model training, our predictions exhibited a strong correlation with CDR scores. In our analysis
62 of the NACC dataset, we observed that $P(DE)$ progressively increased with higher CDR scores, with statis-
63 tically significant differences manifest across the spectrum of cognitive impairment ($p < 0.0001$). However,
64 this pattern did not hold between CDR scores of 2.0 and 3.0, where no significant statistical difference was
65 discerned. In the ADNI dataset, we found a statistically significant demarcation ($p < 0.0001$) in $P(DE)$
66 between the baseline CDR rating and higher gradations. This points to the model's sensitivity to incremental
67 impairment in clinical dementia assessments. In the FHS dataset (Fig. 2f), which substitutes a consensus
68 panel's diagnostic categorization (normal, impaired, and dementia) for CDR scores, a marked statistical sig-
69 nificance ($p < 0.0001$) was evident in $P(DE)$ across these diagnostic strata, with the exception of normal
70 versus impaired. This indicates a challenge for the model in distinguishing the early stages of cognitive
71 decline when relying on a limited set of features. Such limitations are likely due to the community-based
72 nature of the FHS cohort and the specificities of consensus panel ratings at FHS (Table S4). Collectively,
73 these findings illuminate the model's robust capacity to delineate differential cognitive states, showcasing
74 its potential as a tool for identifying levels of cognitive impairment across datasets.

75 **Evaluation of single and co-occurring dementias** We evaluated our model's diagnostic ability
76 across ten distinct dementia etiologies. The ROC and PR curves in (Fig. 3a-b) reflect strong model per-
77 formance on the model's overall assessment on identifying dementia etiologies across different averaging
78 methods, attaining micro-averaged AUROC and AUPR values of 0.96 and 0.70, respectively. In macro-
79 averaged terms, the AUROC and AUPR stood at 0.91 and 0.36. Moreover, the weighted-average values
80 for AUROC and AUPR were 0.94 and 0.73, respectively. The model's performance, characterized by
81 high micro-averaged and weighted-average AUROC and AUPR scores, underscores its diagnostic accu-
82 racy across a broad spectrum of dementia etiologies. While the lower macro-average AUPR scores indicate
83 that our model may perform better on certain diagnoses relative to others, the weighted-average scores, ad-
84 justing for the prevalence of each dementia type, support the model's effectiveness in a real-world clinical
85 setting, where some dementia types are more common than others.

86 To further assess the model performance on co-occurring dementias, we adopted a maximum variance
87 threshold of 0.01 for AUROC calculations.⁴² This selection aimed to balance the sensitivity and specificity
88 of the model, enabling it to discern subtle diagnostic differences. This resulted in a minimum positive sam-
89 ple size of 25. In instances where two dementias co-occurred (Fig. 3c), the model's AUROC scores varied
90 from 0.63 to 0.97, reflecting a spectrum of diagnostic accuracy, with the LBD and PSY combination achiev-
91 ing the highest AUROC. AUPR scores ranged from 0.08 to 0.60, again with the conjunction of LBD and
92 PSY recording the highest AUPR value. In the case of AD occurring with two other etiologies (VD & PSY),
93 the AUROC score was 0.73 and the AUPR was 0.48. While our model demonstrated robust diagnostic dis-
94 crimination, as evidenced by high AUROC values, the variability in AUPR scores may reflect challenges in
95 consistently identifying less prevalent or more complex dementia etiologies within the dataset. Importantly,
96 a similar pattern was found in subsequent analyses of expert neurologists' performance for conditions such
97 as SEF and TBI (Table S15). Additional performance metrics and visualizations that illustrate our model's
98 ability to assess single and co-occurring dementias are presented in the Supplement (Table S7 & Fig. S1).

99 **Model validation with biomarkers** Model-predicted probabilities for AD, FTD, and LBD were aligned
100 with the presence of respective biomarkers, as demonstrated in the raincloud plots in Fig. 4 & Table S11.
101 For AD, $P(AD)$ correlated with $A\beta$, tau, and FDG PET biomarkers across the NACC and ADNI cohorts,
102 indicating statistically significant differences between biomarker-negative and positive groups ($p < 0.0001$).
103 Notably, $P(AD)$ was consistently higher in $A\beta$, tau, and FDG PET positive groups, demonstrating that our
104 framework's diagnostic process aligns well with the current amyloid, tau, and neurodegeneration (ATN)
105 criteria for AD diagnosis.⁴³ Within the NACC cohort, FTD probabilities, $P(FTD)$, were significantly as-
106 sociated with MRI and FDG PET biomarkers, with the biomarker positive groups having higher $P(FTD)$.
107 This result corroborates the capability of our model to detect FTD in alignment with observed patterns
108 of fronto-temporal hypometabolism and atrophy.⁴⁴ Finally, LBD probabilities, $P(LBD)$, also displayed a
109 clear differentiation when analyzed in relation to DaTscan evidence for LBD,⁴⁵ with the DaTscan positive
110 group exhibiting higher probabilities of LBD. Taken together, these findings validate the model's effec-
111 tiveness in capturing the pathophysiological underpinnings of prevalent dementia types in addition to the
112 clinical syndrome, offering etiology-specific probability scores that closely match respective biomarker pro-
113 files. This alignment not only substantiates the model's predictive validity, but also highlights its relevance
114 to contemporary clinical practice as its mechanism for differential diagnosis of dementia reflects established
115 biomarker criteria.

116 **Model validation with neuropathological evidence** In cases with postmortem data (Table S12),
117 we validated our model's etiology-specific probability scores against neuropathological markers of common
118 dementia types (Fig. 5 & Table S13). The composite violin and box plots indicate that, with increasing
119 pathological severity, there is a corresponding elevation in the model-predicted likelihood of the etiology.
120 The first three plots (Figs. 5a-c) compare AD probabilities against three key AD pathological markers with
121 progressive stages: Thal phases of $A\beta$ plaques, Braak stages of neurofibrillary degeneration, and Con-
122 sortium to Establish a Registry for Alzheimer's Disease (CERAD) density scores of neocortical neuritic
123 plaques, denoted by A1-A3, B1-B3 and C1-C3, respectively. Each demonstrated an upward shift in the
124 median probability of AD and an expansion of the interquartile range as the stages advanced, with statistical
125 significance ($p < 0.001$ for Thal stage and $p < 0.0001$ for Braak and CERAD stages, respectively). We fur-
126 ther evaluated our model's predicted probabilities against cerebral amyloid angiopathy (CAA), a common
127 pathological finding in AD confirmed postmortem cases. Similarly, we observed that our model predicted
128 significantly higher AD probabilities in individuals with mild, moderate or severe CAA relative to those
129 without CAA. Collectively, these plots illustrate a clear trend where advancing stages of AD-related pathol-
130 ogy are associated with increased $P(AD)$. Finally, significant differences were observed in $P(FTD)$ and
131 $P(VD)$ based on their respective pathological markers: $P(FTD)$ differed significantly between cases with
132 and without TDP-43 pathology ($p < 0.01$) and tauopathy ($p < 0.05$), $P(VD)$ varied between cases with
133 and without old microinfarcts ($p < 0.001$) and arteriolosclerosis ($p < 0.001$) (Figs. 5e-h). The results are
134 consistent with the well-documented association between TDP-43 protein aggregation and its prevalence in
135 FTD.^{46,47} Additionally, the clear linkage between cerebrovascular pathologies and the incidence of VD is
136 reinforced by our data. Crucially, these outcomes highlight the capability of our AI-driven framework to
137 align model-generated probability scores with a range of neuropathological states beyond AD, supporting
138 its potential utility in the evaluation of broader neurodegenerative diseases.

139 **AI-augmented clinician assessments** We aimed to assess whether our AI framework can compare to,
140 and significantly enhance differential diagnosis of dementia performed by expert clinicians. To this end,
141 we compared our model predicted probabilities with clinicians' diagnoses, which were made in the form of

142 confidence scores (0 to 100 scale). Neurologists reviewed 100 randomly selected cases, including various
143 dementia subtypes, with comprehensive data including demographics, medical history, neuropsychological
144 tests, and multi-sequence MRI scans. We observed that, in instances where the diagnosis was confirmed
145 (true positives), the neurologists' confidence scores across NC, MCI, DE, AD, LBD, VD, FTD, NPH, and
146 PSY were higher in comparison to cases deemed non-diagnostic (true negatives) ($p < 0.01$) (Fig. S2a).
147 In contrast, for the same 100 cases, our model's predicted probabilities on true positive cases for all cat-
148 egories other than ODE were higher than the predicted probabilities for true negative cases ($p < 0.01$),
149 indicating an enhanced ability for our model to detect true positives across more conditions (Fig. S2a). We
150 then analyzed pairwise Pearson correlation coefficients to assess inter-rater agreement for each diagnostic
151 category, both among neurologists' confidence scores, and between the neurologists' confidence scores and
152 our model's predicted probabilities (Fig. S3a). Among clinicians' assessments, we found the most robust,
153 consistent associations within the NC and DE groups, followed by modest associations between assessments
154 of MCI, AD, LBD, VD, FTD and PSY. In contrast, PRD, NPH, SEF, TBI and ODE demonstrated the least
155 consistency between neurologists' assessments. This analysis shed light on dementia types that are rela-
156 tively more challenging to diagnose, as evidenced by the variability in diagnostic confidence among expert
157 clinicians. When comparing neurologists' confidence scores with our model's predicted probabilities, we
158 found that the assessments provided by our model were generally consistent with those provided by the
159 neurologists for NC, MCI, DE, AD, and LBD, as indicated by Pearson correlation coefficients that exceeded
160 0.7 (Fig. S3b). Associations were modest for VD, FTD, PSY, where mean Pearson correlation coefficients
161 were approximately 0.5, while associations were less consistent for PRD, NPH, SEF, TBI, and ODE. The
162 lower correlations observed here reflect the complex nature of these conditions, compounded by a lack of
163 necessary features to tease out their unique signatures.

164 To determine whether our model could augment the assessments provided by neurologists, we com-
165 puted AI-assisted neurologist confidence scores, which was defined as the mean of the neurologists' con-
166 fidence scores and our model's predicted probabilities. We then compared the diagnostic performance of
167 individual neurologist assessments with that of AI-augmented neurologist assessments (Figs. 6a-b & Ta-
168 bles S14 & S15). We consistently found significant increases in AUROC and AUPR for all etiologies
169 ($p < 0.05$). There was a mean percent increase in AUROC of 26.25% and a mean percent increase in AUPR
170 of 73.23% across all categories. The greatest improvement in diagnostic performance was for PRD and TBI,
171 where there was a percent increase in mean AUROC of 73% and 72%, respectively, and a percent increase
172 in mean AUPR of 242% and 257%, respectively. In a separate assessment, neuroradiologists evaluated a
173 randomly selected set of 70 clinically diagnosed dementia cases, and were provided with multi-sequence
174 MRIs, as well as demographic information. For these 70 cases, we assessed the diagnostic performance of
175 radiologists and AI-augmented radiologists, which was defined as the mean of the radiologists' confidence
176 scores and our model's probabilities (Figs. 6c-d & Tables S14 & S15). Across various dementia etiologies,
177 we observed an average increase of 17.22% in AUROC and 42.17% in AUPR. A significant enhancement in
178 AUROC ($p < 0.05$) was noted across all etiologies, with PRD showing the highest mean AUROC improve-
179 ment at 68%. AUPR also displayed improvements, most markedly in PRD, where the mean AUPR surged
180 by 190%.

1 Discussion

2 We present an AI model designed for differential dementia diagnosis by processing a range of multimodal
3 data. Unlike our previous work,^{39,48} our model addresses the clinical challenge of distinguishing between
4 various dementia etiologies, including but not limited to AD, VD, and LBD. Such differentiation is crucial
5 for the precise identification of the multi-factorial nature of dementia, which is linked to the optimization
6 of personalized therapeutic interventions and patient management strategies. The model's robustness was
7 established through its training and validation across a diverse set of independent cohorts. Additionally,
8 our model predictions on various etiologies were corroborated by their validation on a subset of cases for
9 which biomarker and postmortem data were available. In a randomly selected subset of cases, our model's
10 predictions, when combined with neurologist assessments, outperformed the assessments conducted by neu-
11 rologists alone. These results underscore our model's potential in enhancing the efficacy of diagnosing
12 dementia-related disorders.

13 Our model is designed to address the complex nature of mixed dementias by providing probability
14 scores for each contributing etiology. This approach is significant as it enables clinicians to systematically
15 prioritize possible drivers of cognitive impairment based on available data. The model effectively captures
16 the multi-factorial and overlapping characteristics of various dementia types, offering a clear framework to
17 guide clinical decision-making. For example, misdiagnoses in the initial stages of dementia are frequent,
18 often due to symptom misattribution to psychiatric disorders, a situation further complicated by the pres-
19 ence of multiple co-pathologies.^{49,50} Specifically, LBD has historically been difficult to diagnose as early
20 symptoms often resemble those of AD and PSY. The co-occurrence of LBD and AD further complicates
21 diagnosis and tends to be missed entirely until post-mortem evaluation.⁵¹ Our model demonstrated no-
22 table performance, particularly in identifying the AD and LBD combination, highlighting its capability in
23 detecting mixed dementias that are commonly recognized only through postmortem analysis.^{4,52,53} This ca-
24 pability is crucial, given that a significant portion of dementia cases are linked to modifiable risk factors.⁵⁴
25 The insights provided by our model could therefore inform early intervention strategies, potentially altering
26 the disease course and enhancing patient outcomes. Notably, our model represents a significant step forward
27 in the field, surpassing previous machine learning approaches in detecting mixed dementia, thereby offering
28 a valuable tool for refining diagnostic accuracy in clinical practice.

29 Powered by a transformer architecture as the backbone, the utility of our modeling framework is
30 founded on its robust processing of diverse input types and its adept handling of incomplete datasets. These
31 properties are essential for clinicians requiring immediate and accurate diagnostic information in environ-
32 ments with variable data availability. For example, when a general practitioner records clinical observations
33 and cognitive test results for an elderly person with possible cognitive decline, our model can calculate a
34 probability score indicative of MCI or DE. This function facilitates early medical intervention and more
35 informed decisions regarding specialist referrals. At a specialized memory clinic, the addition of exten-
36 sive neuroimaging data and in-depth neuropsychological battery to the model may increase the precision of
37 the diagnosis, which, in turn, enhances the formulation of individual management strategies with a revised
38 probability score. Such capacity to tailor its output to the scope of input data exemplifies our modeling
39 framework's role in different healthcare settings, including those where swift and resource-efficient diag-
40 nosis is paramount. The generation of specific, quantifiable probability scores by the model augments its
41 utility, establishing it as a useful component in the healthcare delivery process. Displaying diagnostic ac-
42 curacy using varied training data — ranging from demographic information to clinical signs, neuroimaging
43 findings, and neurological test results — the model's versatility facilitates its adaptation to varied clinical

44 operations without necessitating a fundamental overhaul of existing workflows. Consequently, our model
45 fosters a seamless transition across the different levels of dementia care, enabling general practitioners to
46 perform preliminary cognitive screenings and specialists to conduct thorough examinations. Its inclusive
47 functionality assures an accessible and comprehensive tool ensuring fail-safe operation in early detection,
48 continuous monitoring, and the fine-tuning of differential diagnoses, thereby elevating the standard of de-
49 mentia care.

50 While our study has the potential to advance the field of differential dementia diagnosis, it does have
51 certain limitations that warrant consideration. Our model was developed and validated on 9 distinct cohorts
52 but its full generalizability across diverse populations and clinical settings remains to be determined. Mov-
53 ing forward, we see potential in evaluating the model's efficacy across the care continuum, encompassing
54 primary care facilities, geriatric and general neurology practices, family medicine, and specialized clinics
55 in tertiary medical centers. Furthermore, AI models like ours possess the capability to enhance the patient
56 screening procedures for clinical trial recruitment.⁵⁵ Our study's datasets primarily consist of AD cases,
57 and while AD is the most common type of dementia, this could potentially skew our model towards im-
58 proved recognition of this specific subtype, introducing a bias. Although we incorporated various dementia
59 etiologies, the imbalanced representation might affect the model's generalizability and sensitivity towards
60 less frequent types. It is important to note that, beyond data imbalance, certain diseases were inherently
61 more challenging to diagnose given the provided feature set, as exemplified by the lower performance met-
62 rics of expert neurologists in conditions such as SEF and TBI (Tables S14, S15). Additionally, we chose
63 to amalgamate mild, moderate, and severe dementia cases into a single category. We acknowledge that
64 this categorization method might not completely reflect the nuanced individual staging practiced in specific
65 healthcare settings, where varying degrees of dementia severity carry distinct implications for treatment and
66 management strategies. Our focus was primarily on differential diagnosis rather than disease staging, which
67 motivated this decision. Future enhancements to our model could potentially include disease staging as an
68 additional dimension, thereby augmenting its granularity and relevance. Finally, our study does not fully ad-
69 dress the considerable heterogeneity inherent in AD, which is characterized by diverse clinical presentations
70 and pathological features.^{56,57}

71 The evidence collected from this study signals a convergence between advanced computational meth-
72 ods and the nuanced task of differential dementia diagnosis, crucial for scenarios with scarce resources
73 and the multifaceted realm of mixed dementia, a condition frequently encountered yet diagnostically com-
74 plex. Our model efficiently integrates multimodal data, showing strong performance across diverse settings.
75 Future validations, encompassing a wider demographic and geographical expanse, will be pivotal to substan-
76 tiate the model's robustness and enhance its diagnostic utility in dementia care. Our pragmatic investigation
77 accentuates the potential of neural networks to refine the granularity of diagnostic evaluations in neurocog-
78 nitive disorders.

1 Methods

2 **Study population** We collected demographics, personal and family history, laboratory results, findings
3 from the physical/neurological exams, medications, neuropsychological tests, and functional assessments as
4 well as multi-sequence magnetic resonance imaging (MRI) scans from 9 distinct cohorts, totaling 51,269
5 participants. There were 19,849 participants with normal cognition (NC), 9,357 participants with mild
6 cognitive impairment (MCI), and 22,063 participants with dementia (DE). We further identified 10 primary
7 and contributing causes of dementia: 17,346 participants with Alzheimer's disease (AD), 2,003 partici-
8 pants with dementia with Lewy bodies and Parkinson's disease dementia (LBD), 2,032 participants with
9 vascular brain injury or vascular dementia including stroke (VD), 114 participants with Prion disease in-
10 cluding Creutzfeldt-Jakob disease (PRD), 3,076 participants with frontotemporal lobar degeneration and its
11 variants, which includes corticobasal degeneration (CBD) and progressive supranuclear palsy (PSP), and
12 with or without amyotrophic lateral sclerosis (FTD), 138 participants with normal pressure hydrocephalus
13 (NPH), 808 participants suffering from dementia due to infections, metabolic disorders, substance abuse
14 including alcohol, medications, delirium and systemic disease - a category termed as systemic and external
15 factors (SEF), 2,700 participants suffering from psychiatric diseases including schizophrenia, depression,
16 bipolar disorder, anxiety, and post-traumatic stress disorder (PSY), 265 participants with dementia due to
17 traumatic brain injury (TBI), and 1,234 participants with dementia due to other causes which include neo-
18 plasms, multiple systems atrophy, essential tremor, Huntington's disease, Down syndrome, and seizures
19 (ODE).

20 The cohorts include the National Alzheimer's Coordinating Center (NACC) dataset ($n = 45,349$),⁵⁸
21 the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset ($n = 2,404$),⁵⁹ the frontotemporal lo-
22 bar degeneration neuroimaging initiative (NIFD) dataset ($n = 253$),⁶⁰ the Parkinson's Progression Marker
23 Initiative (PPMI) dataset ($n = 198$),⁶¹ the Australian Imaging, Biomarker and Lifestyle Flagship Study
24 of Ageing (AIBL) dataset ($n = 661$),⁶² the Open Access Series of Imaging Studies-3 (OASIS) dataset
25 ($n = 491$),⁶³ the 4 Repeat Tauopathy Neuroimaging Initiative (4RTNI) dataset ($n = 80$),⁶⁴ and three
26 in-house datasets maintained by the Lewy Body Dementia Center for Excellence at Stanford University
27 (LBDSU) ($n = 182$),⁶⁵ and the Framingham Heart Study (FHS) ($n = 1,651$).⁶⁶ Since its inception in 1948,
28 FHS has been dedicated to identifying factors contributing to cardiovascular disease, monitoring multiple
29 generations from Framingham, Massachusetts. Over time, the study has pinpointed major cardiovascular
30 disease risk factors and explored their effects, while also investigating risk factors for conditions like de-
31 mentia and analyzing the relationship between physical traits and genetics. Additional details on the study
32 population are presented in Tables 1 & S1.

33 **Inclusion and exclusion criterion** Individuals from each cohort were eligible for study inclusion if
34 they were diagnosed with normal cognition (NC), mild cognitive impairment (MCI), or dementia (DE). We
35 used the National Alzheimer's Coordinating Center (NACC) dataset,⁵⁸ which is based on the Uniform Data
36 Set (UDS) 3.0 dictionary,⁶⁷ as the baseline for our study. To ensure data consistency, we organized the data
37 from the other cohorts according to the UDS dictionary. For individuals from the NACC cohort who had
38 multiple clinical visits, we initially prioritized the visits at which the person received the diagnostic label
39 of dementia. We then selected the visit with the most data features available prioritizing the availability of
40 neuroimaging information. If multiple visits met all the above criteria, we chose the most recent visit among
41 them. This approach maximized the sample sizes of dementia cases, as well as ensured that each individual
42 had the latest record included in the study while maximizing the utilization of available neuroimaging and

43 non-imaging data. We included participants from the 4RTNI dataset⁶⁴ with frontotemporal lobar degenera-
44 tion (FTD)-related disorders like progressive supranuclear palsy (PSP) or corticobasal syndrome (CBS). For
45 other cohorts (NIFD,⁶⁰ PPMI,⁶¹ LBDSU,⁶⁵ AIBL,⁶² ADNI,⁵⁹ and OASIS⁶³), participants were included if
46 they had at least one MRI scan within 6 months of an officially documented diagnosis. From the FHS,⁶⁶ we
47 utilized data from the Original Cohort (Gen 1) enrolled in 1948, and the Offspring Cohort (Gen 2) enrolled
48 in 1971. For these participants, we selected available data including demographics, history, clinical exam
49 scores, neuropsychological test scores, and MRI within 6 months of the date of diagnosis. We did not ex-
50 clude cases based on the absence of features (including imaging) or diagnostic labels. Instead, we employed
51 our innovative model training approach to address missing features or labels (See below).

52 **Data processing and training strategy** Various non-imaging features (n=391) corresponding to sub-
53 ject demographics, medical history, laboratory results, medications, neuropsychological tests, and functional
54 assessments were included in our study. We combined data from 4RTNI, AIBL, LBDSU, NACC, NIFD,
55 OASIS, and PPMI to train the model. We used a portion of the NACC dataset for internal testing, while the
56 ADNI and FHS cohorts served for external validation (Tables 1, S1–S5). We used a series of steps such as
57 standardizing the data across all cohorts and formatting the features into numerical or categorical variables
58 before using them for model training. We used stratified sampling at the person-level to create the train-
59 ing, validation, and testing splits. As we pooled the data from multiple cohorts, we encountered challenges
60 related to missing features and labels. To address these issues and enhance the robustness of our model
61 against data unavailability, we incorporated several strategies such as random feature masking and masking
62 of missing labels (see below).

63 **MRI processing** Our investigation harnessed the potential of multi-sequence magnetic resonance imag-
64 ing (MRI) volumetric scans sourced from diverse cohorts (Table S6). The majority of these scans encom-
65 passed T1-weighted (T1w), T2-weighted (T2w), diffusion-weighted (DWI), susceptibility-weighted (SWI),
66 and fluid-attenuated inversion recovery (FLAIR) sequences. The collected imaging data were stored in the
67 NIFTI file format, categorized by participant and the date of their visit. The MRI scans underwent a series of
68 pre-processing steps involving skull stripping, linear registration to the MNI space, and intensity normaliza-
69 tion. Skull stripping was performed using SynthStrip,⁶⁸ a computational tool designed for extracting brain
70 voxels from various image types. Then, the MRI scans were registered using FSL’s ‘flirt’ tool for linear reg-
71 istration of whole brain images,⁶⁹ based on the MNI152 atlas.⁷⁰ Prior to linear registration to the MNI space,
72 we utilized the ‘fslorient2std’ function within FSL to standardize the orientation across all scans to match
73 the MNI template’s axis order. As a result, the registered scans followed the dimensions of the MNI152
74 template, which are $182 \times 218 \times 182$. Finally, all MRI scans underwent intensity normalization to the range
75 $[0,1]$ to increase the homogeneity of the data. To ensure the purity of the dataset, we excluded calibration,
76 localizer, and 2D scans from the downloaded data before initiating model training.

77 **Backbone architecture** Our modeling framework harnesses the power of the transformer architecture
78 to interpret and process a vast array of diagnostic parameters, including person-level demographics, medical
79 history, neuroimaging, functional assessments, and neuropsychological test scores. Each of these distinct
80 features is initially transformed into a fixed-length vector using a modality-specific strategy, forming the
81 initial layer of input for the transformer model. Following this, the transformer acts to aggregate these
82 vector inputs, decoding them into a series of predictions. A distinguishing strength of this framework lies

83 in its integration of the transformer’s masking mechanism,^{71,72} strategically deployed to emulate missing
84 features. This capability enhances the model’s robustness and predictive power, allowing it to adeptly handle
85 real-world scenarios characterized by incomplete data.

86 **Multimodal data embeddings** Transformers use a uniform representation for all input tokens, typi-
87 cally in the form of fixed-length vectors. However, the inherent complexity of medical data, with its variety
88 of modalities, poses a challenge to this requirement. Therefore, medical data needs to be adapted into a
89 unified embedding that our transformer model can process. The data we accessed falls into three primary
90 categories: numerical data, categorical data, and imaging data. Each category requires a specific method of
91 embedding. Numerical data typically encompasses those data types where values are defined in an ordinal
92 manner that holds distinct real-world implications. For instance, chronological age fits into this category, as
93 it serves as an indicator of the aging process. To project numerical data into the input space of the trans-
94 former, we employed a single linear layer to ensure an appropriate preservation of the structure inherent to
95 the original data space. Categorical data encompasses those inputs that can be divided into distinct cate-
96 gories yet lack any implicit order or priority. An example of this is gender, which can be categorized as
97 ‘male’ or ‘female’. We utilized a lookup table to translate categorical inputs into corresponding embeddings.
98 It is noteworthy that this approach is akin to a linear transformation when the data is one-hot vectorized,
99 but is computationally efficient, particularly when dealing with a vast number of categories. Imaging data,
100 which includes MRI scans in medical applications, can be seen as a special case of numerical data. How-
101 ever, due to their high dimensionality and complexity, it is difficult to compress raw imaging data into a
102 significantly lower-dimensionality vector using a linear transformation, while still retaining essential infor-
103 mation. We leveraged the advanced capabilities of modern deep learning architectures to extract meaningful
104 imaging embeddings (see below). Once these embeddings were generated, they were treated as numerical
105 data, undergoing linear projection into vectors of suitable length, thus enabling their integration with other
106 inputs to the transformer.

107 **Imaging feature extraction** We harnessed the Swin UNETR (Fig. S4),^{73,74} a three-dimensional (3D)
108 transformer-based architecture, to extract embeddings from a multitude of brain MRI scans, encompassing
109 various sequences including T1-weighted (T1w), T2-weighted (T2w), diffusion-weighted (DWI), susceptibility-
110 weighted (SWI), and fluid-attenuated inversion recovery (FLAIR) imaging sequences. The Swin UNETR
111 model consists of a Swin Transformer encoder, designed to operate on 3D patches, seamlessly connected
112 to a convolutional neural network (CNN)-based decoder through multi-resolution skip connections. Com-
113 mencing with an input volume $X \in \mathbb{R}^{H \times W \times D}$, the encoder segmented X into a sequence of 3D tokens
114 with dimensions $\frac{H}{H'} \times \frac{W}{W'} \times \frac{D}{D'}$, and projected them into a C -dimensional space via an embedding layer.
115 It employed a patch size of $2 \times 2 \times 2$ with a feature dimension of $2 \times 2 \times 2 \times 1$ and an embedding space
116 dimension of $C = 48$. The Swin UNETR encoder was subsequently interconnected with a CNN-based
117 decoder at various resolutions through skip connections, collectively forming a ‘U-shaped’ network. This
118 decoder amalgamated the encoder’s outputs at different resolutions, conducted upsampling via deconvolu-
119 tions, ultimately generating a reconstruction of the initial input volume. The pre-trained weights were the
120 product of self-supervised pre-training of the Swin UNETR encoder, primarily conducted on 3D volumes
121 encompassing the chest, abdomen, and head/neck.^{73,74}

122 The process of obtaining imaging embeddings began with several transformations applied to the MRI
123 scans. These transformations included resampling the scans to standardized pixel dimensions, foreground

124 cropping, and spatial resizing, resulting in the creation of sub-volumes with dimensions of $128 \times 128 \times$
125 128. Subsequently, these sub-volumes were input into the Swin UNETR model, which in turn extracted
126 encoder outputs sized at $768 \times 4 \times 4 \times 4$. These extracted embeddings underwent downsampling via a
127 learnable embedding module, consisting of four convolutional blocks, to align with the input token size of
128 the downstream transformer. As a result, the MRI scans were effectively embedded into one-dimensional
129 vectors, each of size 256. These vectors were then combined with non-imaging features and directed into
130 the downstream transformer for further processing. The entire process utilized a dataset comprising 8,155
131 MRI volumes, which were allocated for model training, validation, and testing (Table S6).

132 **Random feature masking** To enhance the robustness of the backbone transformer in handling data
133 incompleteness, we leveraged the masking mechanism^{71,72} to emulate arbitrary missing features during
134 training. The masking mechanism, when paired with the attention mechanism, effectively halts the informa-
135 tion flow from a given set of input tokens, ensuring that certain features are concealed during prediction. A
136 practical challenge arises when considering the potential combinations of input features, which increase ex-
137 ponentially. With hundreds of features in play, capturing every potential combination is intractable. Inspired
138 by the definition of Shapley values, we deployed an efficient strategy for feature dropout. Given a sample
139 with feature set S , S is randomly permuted as σ ; simultaneously, an integer i is selected independently from
140 the range $[1, |S|]$. Subsequent to this, the features $\sigma_{i+1}, \sigma_{i+2}, \dots, \sigma_{|S|}$ are masked out from the backbone
141 transformer. It's noteworthy that the dropout process was applied afresh across different training batches or
142 epochs to ensure that the model gets exposed to a diverse array of missing information even within a single
143 sample.

144 **Handling missing labels** The backbone transformer was trained by amalgamating data from multiple
145 different cohorts, each focused on distinct etiologies, which introduced the challenge of missing labels
146 in the dataset. While most conventional approaches involve discarding records with incomplete output
147 labels during training, we chose a more inclusive strategy to maximize the utility of the available data.
148 Our approach framed the task as a multi-label classification problem, introducing thirteen separate binary
149 heads, one for each target label. With this design, for every training sample, we generated a binary mask
150 indicating the absence of each label. We then masked the loss associated with samples lacking specific
151 labels before backpropagation. This method ensured optimal utilization of the dataset, irrespective of label
152 availability. The primary advantage of this approach lies in its adaptability. By implementing this label-
153 masking strategy, our model can be evaluated against datasets with varying degrees of label availability,
154 granting us the flexibility to address a wide spectrum of real-world scenarios.

155 **Loss function** Our backbone model was trained by minimizing the loss function (\mathcal{L}) composed of two
156 loss terms: “Focal Loss (FL)”⁷⁵ (\mathcal{L}_{FL}) and “Ranking Loss (RL)” (\mathcal{L}_{RL}), along with the standard L2 regu-
157 larization term. FL is a variant of standard cross-entropy loss that addresses the issue of class imbalance. It
158 assigns low weight to easy (well-classified) instances and employs a balance parameter. This loss function
159 was used for each of the diagnostic categories (a total of 13, see Glossary 1). Therefore, our \mathcal{L}_{FL} term was:

$$\mathcal{L}_{\text{FL}} = \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{13} -y_{k,i} \alpha_i (1 - p_{k,i})^\gamma \log(p_{k,i}) - (1 - y_{k,i}) (1 - \alpha_i) (p_{k,i})^\gamma \log(1 - p_{k,i}),$$

160 where N was the batch size (i.e., $N = 128$), and other parameters and variables were as defined. The
161 focusing parameter γ was set to 2, which had been reported to work well in most of the experiments in the
162 original paper.⁷⁵ Moreover, $\alpha_i \in [0, 1]$ was the balancing parameter that influenced the weights of positive
163 and negative instances. It was set as the square of the complement of the fraction of samples labeled as 1,
164 varying for each i due to the differing level of class imbalance across diagnostic categories (refer to Table 1).
165 The FL term did not take inter-class relationships into account. To address these relationships in our overall
166 loss function, we also incorporated the RL term that induced loss if the sigmoid outputs for diagnostic
167 categories labeled as 0 were not lower than those labeled as 1 by a predefined margin of ϵ , for any training
168 sample k . We defined the RL term for any pair of diagnostic categories i and j , as follows:

$$\mathcal{L}_{\text{RL}}^{(i,j)}(\mathbf{p}_k, \mathbf{y}_k) = \max(0, (p_{k,i} - p_{k,j})(y_{k,j} - y_{k,i}) + \epsilon),$$

169 Overall, the RL term was:

$$\mathcal{L}_{\text{RL}} = \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{13} \sum_{j=i+1}^{13} \mathcal{L}_{\text{RL}}^{(i,j)}(\mathbf{p}_k, \mathbf{y}_k).$$

170 Combining all terms, our overall loss function (\mathcal{L}) was:

$$\mathcal{L} = \mathcal{L}_{\text{FL}} + \lambda \mathcal{L}_{\text{RL}} + \beta \|\mathbf{w}\|^2,$$

171 where λ and β were the weights that controlled the importance of \mathcal{L}_{RL} and the L2 regularization terms,
172 respectively. The training was done using the mini-batch strategy with the AdamW optimizer,⁷⁶ an improved
173 version of the Adam optimizer,⁷⁷ with a learning rate of 0.001 for a total of 256 epochs. Additionally, we
174 utilized a cosine learning rate scheduler with warm restarts,⁷⁸ initiating the first restart after 64 epochs and
175 extending the restart period by a factor of 2 for each subsequent restart. The values of ϵ , λ , and β were
176 determined to be $\epsilon = 0.25$, $\lambda = 0.005$, and $\beta = 0.0005$, respectively, based on an evaluation of the overall
177 model performance on the validation set. During training, the model performance was evaluated on the
178 validation set at the end of each epoch, and the model with the highest performance was selected.

179 **Traditional machine learning models** To assess our model's ability to classify NC, MCI and DE
180 cases, we compared its performance with the CatBoost model, a tree-based classification framework.^{39,41}
181 Given the variability in available features across the test cohorts, we divided the data into two feature subsets,
182 as detailed in Tables S2, S4 and S5. This stratification enabled a comparison with CatBoost, offering insights
183 into our model's performance using a range of parameters. The first feature subset consisted of variables
184 common across all cohorts, including demographics, MMSE, and Boston Naming Test scores. The second
185 subset expanded on this by incorporating additional neuropsychological measures found in the NACC and
186 ADNI cohorts, such as trail making tests A and B, logical memory IIA delayed recall, MoCA scores, and
187 digit span forward and backward tests. We trained separate CatBoost models for each feature set but applied
188 our model to both subsets without retraining, allowing for a consistent evaluation across different feature
189 configurations.

190 **Biomarker validation** The predicted probabilities of the model for various etiologies were cross-validated
191 with established gold-standard biomarkers pertinent to each respective etiology. Both the NACC and ADNI
192 test cohorts were used in AD biomarker analyses, while only NACC data were used for FTD and LBD due
193 to biomarker availability. In the NACC dataset, binary UDS variables were used to define positivity for

194 amyloid β ($A\beta$), tau and fluorodeoxyglucose F18 (FDG) PET biomarkers for AD due to varying PET pro-
195 cessing methods across centers. Binary UDS variables were also used to define FDG and MRI evidence for
196 FTD, and dopamine transporter scan (DATscan) as evidence for LBD. In ADNI, the University of Califor-
197 nia, Berkeley (UCB) $A\beta$ PET processing pipeline yields Freesurfer-defined cortical summary and reference
198 regions, as well as centiloids (CL). A cut-off value of 20 CL was chosen to define positivity.⁷⁹ For tau, The
199 UCB processing pipeline yields standardized uptake value ratios (SUVr) in Freesurfer-defined regions. A
200 meta-temporal region of interest (ROI) was constructed following established standards.⁸⁰ A Gaussian mix-
201 ture model (GMM) with two components identified 1.74 SUVr as the optimal threshold to separate the two
202 distributions, where values greater than 1.74 indicated tau PET positivity. Finally, the UCB FDG PET pro-
203 cessing pipeline yields a meta-ROI, on which a GMM with two components identified 1.21 SUVr as the best
204 threshold, where values less than 1.21 indicating positivity for neurodegeneration. Information regarding
205 the PET processing protocols can be found in the summaries of UCB amyloid, tau, and FDG PET methods
206 available on the LONI Image Data Archive website.⁸¹

207 **Neuropathologic validation** The model's predictive capacity for various dementia etiologies was sub-
208 stantiated through alignment with neuropathological evaluations sourced from the NACC, FHS and ADNI
209 cohorts (Table S12). We included participants who conformed to the study's inclusion criteria, had un-
210 dergone MRI scans no more than three years prior to death, and for whom neuropathological data were
211 available. Standardization of data was conducted in accordance with the Neuropathology Data Form Ver-
212 sion 10 protocols from the National Institute on Aging.⁸² We pinpointed neuropathological indicators that
213 influence the pathological signature of each dementia etiology, such as arteriolosclerosis, the presence of
214 neurofibrillary tangles and amyloid plaques, cerebral amyloid angiopathy (CAA), and markers of tauopa-
215 thy. These indicators were carefully chosen to reflect the complex pathological terrain that defines each
216 form of dementia. To examine the Thal phase for amyloid plaques (A score), subjects were categorized into
217 two groups: one encompassing Phase 0, indicative of no amyloid plaque presence, and a composite group
218 merging Phases 1-5, reflecting varying degrees of amyloid pathology. The model's predictive performance
219 was then compared across these groupings. For the Braak stage of neurofibrillary degeneration (B score),
220 we consolidated stages I-VI into a single collective, representing the presence of AD-type neurofibrillary
221 pathology, whereas stage 0 was designated for cases devoid of AD-type neurofibrillary degeneration. With
222 respect to the density of neocortical neuritic plaques, assessed by the (CERAD or C score), individuals
223 without neuritic plaques constituted one group, while those with any manifestation of neuritic plaques —
224 sparse, moderate, or frequent (C1-C3) — were aggregated into a separate group for comparative analysis
225 of the model's predictive outcomes. To evaluate model alignment with the severity of CAA, subjects were
226 classified into two groups: one representing the absence of CAA, and another encapsulating all stages of
227 CAA severity, ranging from mild to severe. Furthermore, to evaluate the model's concordance with non-
228 AD pathologies, we analyzed the association between the model-generated probabilities of FTD with the
229 presence of TDP-43 pathology and tauopathy, and VD with the presence of old microinfarcts and arteri-
230 olosclerosis.

231 **AI-augmented clinician assessments** We aimed to ascertain if our model could bolster the diagnostic
232 prowess of clinicians specializing in dementia care and diagnosis. To this end, a group of 12 neurologists
233 and 7 neuroradiologists were invited to participate in diagnostic tasks on a subset of NACC cases (see 'Data
234 processing and training strategy'). Neurologists were presented with 100 cases, which included 15 cases
235 each of NC and MCI, and 7 cases for each of the dementia etiologies. The data encompassed person-level
236 demographics, medical history, social history, neuropsychological tests, functional assessments, and multi-

237 sequence MRI scans where possible (i.e., T1-weighted, T2-weighted, FLAIR, DWI and SWI sequences).
238 They were asked to provide their diagnostic impressions, as well as a confidence score ranging from 0 to
239 100 for the diagnosis of each of the 13 labels. These confidence scores quantitatively reflect the clinician's
240 certainty in their diagnosis, with higher scores indicating greater certainty. This scoring system facilitated
241 a quantitative comparison between the clinicians' diagnostic certainty and the predictive probabilities gen-
242 erated by our model. Similarly, neuroradiologists were provided with the same multi-sequence MRI scans
243 used by our model, along with information on age, gender, race, and education status from 70 clinically
244 diagnosed DE cases. They were also tasked with providing diagnostic impressions, as well as confidence
245 scores concerning the origin of dementia (Refer to Glossary 1). To evaluate the potential enhancement of
246 clinical judgments by our model, we calculated AI-augmented confidence scores by averaging the clinicians'
247 confidence scores with our model's predicted probabilities. We then assessed the diagnostic accuracy of the
248 clinicians' original and AI-augmented confidence scores using AUROC and AUPR metrics. The specifics
249 of the case samples and questionnaires provided to the neurologists and neuroradiologists are detailed in the
250 Supplementary Information.

251 **Statistical analysis** We used one-way ANOVA and the χ^2 test for continuous and categorical variables,
252 respectively to assess the overall differences in the population characteristics between the diagnostic groups
253 across the study cohorts. We applied the Kruskal-Wallis H-test for independent samples and subsequently
254 conducted post-hoc Dunn's testing with Bonferroni correction to evaluate the relationship between clinical
255 dementia rating (CDR) scores and the model-predicted probabilities, as well as between neuropathologic
256 scores and the model-predicted probabilities. We used the two-sample Kolmogorov-Smirnov (K-S) test to
257 compare model predicted AD probabilities, $P(AD)$, between MCI cases with an etiological diagnosis of
258 AD and MCI cases without one. We opted for non-parametric tests because the Shapiro-Wilk test indi-
259 cated significant deviations from normality. In order to assess whether the model's predicted probabilities
260 for AD, FTD and LBD were significantly higher for their respective biomarker positive cases compared to
261 biomarker negative ones, a one-sided Mann-Whitney U test was conducted when the Shapiro-Wilk test indi-
262 cated significant deviations from normality. ADNI's $A\beta$ groups did not significantly deviate from normality
263 and were therefore compared using the one-sided independent samples t-test. To compare model predic-
264 tions with expert-driven assessments, we used the Brunner Munzel test to identify statistically significant
265 increases in the mean disease probability scores between the levels of scoring categories. We conducted
266 a Shapiro-Wilk test on the distributions of the true negative and true positive cases for each etiology. The
267 Brunner-Munzel test was then used to compare the expert and model confidence scores for the true negative
268 and true positive cases for each etiology. To evaluate the inter-rater reliability of label-specific confidence
269 scores, we performed pairwise Pearson correlation analyses between clinicians' scores and those generated
270 by the model.⁸³ We calculated the average correlation coefficient across pairs and determined its 95% con-
271 fidence interval. In addition, we estimated the mean Pearson correlation coefficient between the confidence
272 score of neurologists and the model's score for each diagnostic label using a bootstrapping approach. Pair-
273 wise statistical comparisons of AI-augmented clinician diagnostic performance (AUROC and AUPR) and
274 clinicians only diagnostic performance were performed with the one-sided Wilcoxon signed rank test. All
275 statistical analyses were conducted at a significance level of 0.05.

276 **Performance metrics** We generated receiver operating characteristic (ROC) and precision-recall (PR)
277 curves from predictions on both the NACC test data and other datasets. From each ROC and PR curve,
278 we further derived the area under the curve values (AUC and AUPR, respectively). Further, we computed
279 micro-, macro- and weighted-average AUC and AUPR values. Of note, the micro-average approach con-

280 solidates true positives, true negatives, false positives, and false negatives from all classes into a unified
281 curve, providing a global performance metric. In contrast, the macro-average calculates individual ROC/PR
282 curves for each class before computing their unweighted mean, disregarding potential class imbalances. The
283 weighted-average, while similar in approach to macro-averaging, assigns a weight to each class's ROC/PR
284 curve proportionate to its representation in the dataset, thereby acknowledging class prevalence. We also
285 evaluated the model's accuracy, sensitivity, specificity, and Matthews correlation coefficient, with the latter
286 being a balanced measure of quality for classes of varying sizes in a binary classifier.

287 **Computational hardware and software** All MRI and non-imaging data were processed on a work-
288 station equipped with an Intel i9 14-core 3.3 GHz processor and 4 NVIDIA RTX 2080Ti GPUs. Our soft-
289 ware development utilized Python (version 3.11.7) and the models were developed using PyTorch (version
290 2.1.0). We used several other Python libraries to support data analysis, including pandas (version 1.5.3),
291 scipy (version 1.10.1), tensorboardX (version 2.6.2), torchvision (version 0.15), and scikit-learn (version
292 1.2.2). Training the model on a single Quadro RTX8000 GPU on a shared computing cluster had an average
293 runtime of 7 minutes per epoch, while the inference task took less than a minute per instance. All clinicians
294 reviewed MRIs using 3D Slicer (version 4.10.2) and logged their findings in REDCap (version 11.1.3).

295 **Data and code availability** Data from ADNI, AIBL, NACC, NIFD, OASIS, PPMI and 4RTNI can be
296 downloaded from publicly available resources. Data from FHS and LBDSU can be obtained upon request,
297 subject to institutional approval. Details on our model can be found on the Kolachalama Laboratory's
298 GitHub page (<https://github.com/vkola-lab>).

1 Acknowledgements

2 This project was supported by grants from the Karen Toffler Charitable Trust (VBK), National Institute
3 on Aging's Artificial Intelligence and Technology Collaboratories (P30-AG073014, VBK), the American
4 Heart Association (20SFRN35460031, VBK & RA), Gates Ventures (RA & VBK), the Michael J. Fox
5 Foundation (KLP), and the National Institutes of Health (R01-HL159620 [VBK], R21-CA253498 [VBK],
6 R43-DK134273 [VBK], RF1-AG062109 [RA & VBK], U19-AG068753 [RA], P20-GM130447 [OT], K23-
7 NS075097 [KLP], P50-AG047366 [KLP], and R01-NS115114 [KLP]). We acknowledge grant support from
8 Boston University, CTSI 1UL1TR001430, for the REDCap Survey. We acknowledge the efforts of several
9 individuals from the ADNI, AIBL, FHS, LBDSU, NACC, NIFD, OASIS, PPMI, and 4RTNI for providing
10 access to data. Finally, we thank Drs. Shangran Qiu, Joyce C. Lee, Courtney E. Takahashi, Andrew M.
11 Stern and Jesse B. Mez for several useful discussions.

12 The NACC database is funded by NIA grant U24-AG072122. The ADNI database is funded by NIA
13 grant U01-AG024904. More details are outlined in the supplement.

14 Contributions

15 C.X. and S.S.K. contributed equally to this work. S.S.K., D.L., S.P., V.H.J., O.T.Z., A.S.W., A.K., C.K.,
16 and T.F.A.A. performed data collection. C.X. and S.S.K. designed and developed the machine learning
17 framework. C.X., S.S.K., D.L., S.P., V.H.J., O.B.G., and M.A. performed model training and validation.
18 S.S.K., S.P., V.H.J., and M.A. performed statistical analysis. C.X., S.S.K., D.L., S.P., V.H.J., O.T.Z., A.S.W.,
19 O.B.G., J.D.Z., S.T.P. and M.A. generated the figures and tables. V.C.A.A., B.C.D., C.W.F., H.H., S.K.,
20 A.Z.M., D.L.M., S.O., A.B.P., S.R., M-H.S-H., E.A.S., B.N.S., J.E.S., A.S., O.T., J.Y., Y.Z. and S.Z. are
21 practicing clinicians who reviewed the cases. S.A.B. and B.A.P. provided guidance on the modeling frame-
22 work. K.L.P. and R.A. provided access to data. V.B.K. wrote the manuscript. All authors reviewed, edited
23 and approved the manuscript. V.B.K. conceived, designed and directed the study.

24 Ethics declarations

25 V.B.K. is on the scientific advisory board for Altoida Inc., and serves as a consultant to AstraZeneca. S.K.
26 serves as consultant to AstraZeneca. C.W.F. is a consultant to Boston Imaging Core Lab. K.L.P. is a mem-
27 ber of the scientific advisory boards for Curasen, Biohaven, and Neuron23, receiving consulting fees and
28 stock options, and for Amprion, receiving stock options. R.A. is a scientific advisor to Signant Health and
29 NovoNordisk. She also serves as a consultant to Davos Alzheimer's Collaborative. The remaining authors
30 declare no competing interests.

1 References

- 3 1. Organization, W. H. *et al.* *Global Status Report on the Public Health Response to Dementia: Web Annex*
4 *Methodology for Producing Global Dementia Cost Estimates* (World Health Organization, 2022).
- 5 2. Cahill, S. Who's global action plan on the public health response to dementia: some challenges and
6 opportunities. *Aging & Mental Health* **24**, 197–199 (2019).
- 7 3. Gauthier, S. *et al.* Why has therapy development for dementia failed in the last two decades?
8 *Alzheimer's & Dementia* **12**, 60–64 (2016).
- 9 4. Schneider, J. A., Arvanitakis, Z., Bang, W. & Bennett, D. A. Mixed brain pathologies account for most
10 dementia cases in community-dwelling older persons. *Neurology* **69**, 2197–2204 (2007).
- 11 5. Habes, M. *et al.* Disentangling heterogeneity in alzheimer's disease and related dementias using data-
12 driven methods. *Biological psychiatry* **88**, 70–82 (2020).
- 13 6. Dall, T. M. *et al.* Supply and demand analysis of the current and future us neurology workforce.
14 *Neurology* **81**, 470–478 (2013).
- 15 7. Burton, A. How do we fix the shortage of neurologists? *The Lancet Neurology* **17**, 502–503 (2018).
- 16 8. Lester, P. E., Dharmarajan, T. S. & Weinstein, E. The looming geriatrician shortage: Ramifications and
17 solutions. *J Aging Health* **32**, 1052–1062 (2020). Epub 2019 Oct 4.
- 18 9. Hayden, K. M. *et al.* Vascular risk factors for incident alzheimer disease and vascular dementia: the
19 cache county study. *Alzheimer Disease & Associated Disorders* **20**, 93–100 (2006).
- 20 10. Kane, J. P. *et al.* Clinical prevalence of lewy body dementia. *Alzheimer's research & therapy* **10**, 1–8
21 (2018).
- 22 11. Onyike, C. U. & Diehl-Schmid, J. The epidemiology of frontotemporal dementia. *International review*
23 *of psychiatry* **25**, 130–137 (2013).
- 24 12. Verdi, S., Marquand, A. F., Schott, J. M. & Cole, J. H. Beyond the average patient: how neuroimaging
25 models can address heterogeneity in dementia. *Brain* **144**, 2946–2953 (2021).
- 26 13. Skinner, T. R., Scott, I. A. & Martin, J. H. Diagnostic errors in older patients: a systematic review of
27 incidence and potential causes in seven prevalent diseases. *International journal of general medicine*
28 137–146 (2016).
- 29 14. Gaugler, J. E. *et al.* Characteristics of patients misdiagnosed with alzheimer's disease and their medi-
30 cation use: an analysis of the nacc-uds database. *BMC geriatrics* **13**, 1–10 (2013).
- 31 15. Cummings, J. *et al.* Lecanemab: Appropriate use recommendations. *Journal of Prevention of*
32 *Alzheimer's Disease* **10**, 362–377 (2023). Author manuscript; available in PMC 2024 Jan 1.
- 33 16. Sevigny, J. *et al.* The antibody aducanumab reduces abeta plaques in alzheimer's disease. *Nature* **537**,
34 50–56 (2016).
- 35 17. van Dyck, C. H. *et al.* Lecanemab in early alzheimer's disease. *New England Journal of Medicine* **388**,
36 9–21 (2023).
- 37 18. Hampel, H. *et al.* Amyloid-related imaging abnormalities (aria): radiological, biological and clinical
38 characteristics. *Brain* **146**, 4414–4424 (2023).

- 39 19. Knopman, D. S. *et al.* Practice parameter: Diagnosis of dementia (an evidence-based review). *Neurology* **56**, 1143–1153 (2001).
40
- 41 20. Kandiah, N. *et al.* Current and future trends in biomarkers for the early detection of alzheimer’s disease
42 in asia: expert opinion. *Journal of Alzheimer’s disease reports* **6**, 699–710 (2022).
- 43 21. Thijssen, E. H. & Rabinovici, G. D. Rapid progress toward reliable blood tests for alzheimer disease.
44 *JAMA Neurology* **78**, 143–145 (2021).
- 45 22. Teunissen, C. E. *et al.* Blood-based biomarkers for alzheimer’s disease: towards clinical implementa-
46 tion. *Lancet Neurology* **21**, 66–77 (2022).
- 47 23. Liddy, C., Drosinis, P., Joschko, J. & Keely, E. Improving access to specialist care for an aging popula-
48 tion. *Gerontology and Geriatric Medicine* **2**, 2333721416677195 (2016).
- 49 24. Crombie, A. *et al.* Rural general practitioner confidence in diagnosing and managing dementia: A two-
50 stage, mixed methods study of dementia-specific training. *Australian Journal of Rural Health* (2024).
- 51 25. Ferri, C. P. & Jacob, K. Dementia in low-income and middle-income countries: different realities
52 mandate tailored solutions. *PLoS medicine* **14**, e1002271 (2017).
- 53 26. Martin, S. A., Townend, F. J., Barkhof, F. & Cole, J. H. Interpretable machine learning for dementia: A
54 systematic review. *Alzheimer’s & Dementia* **19**, 2135–2149 (2023).
- 55 27. Myszczyńska, M. A. *et al.* Applications of machine learning to diagnosis and treatment of neurodegenera-
56 tive diseases. *Nature Reviews Neurology* **16**, 440–456 (2020).
- 57 28. Borchert, R. J. *et al.* Artificial intelligence for diagnostic and prognostic neuroimaging in dementia: A
58 systematic review. *Alzheimer’s & Dementia* (2023). Online ahead of print.
- 59 29. Ahmed, M. R., Mahmood, A. N., Huq, M. A., Funk, P. & Mafi, A. Neuroimaging and machine learn-
60 ing for dementia diagnosis: Recent advancements and future prospects. *IEEE Reviews in Biomedical*
61 *Engineering* **12**, 19–33 (2019).
- 62 30. Bron, E. E. *et al.* Ten years of image analysis and machine learning competitions in dementia. *Neu-
63 roImage* **253** (2022).
- 64 31. Vemuri, P. *et al.* Antemortem differential diagnosis of dementia pathology using structural mri:
65 Differential-stand. *NeuroImage* **55**, 522–531 (2011).
- 66 32. Zheng, Y., Zhang, Y., Zhang, Y., Wang, Y. & Zheng, B. Machine learning-based framework for dif-
67 ferential diagnosis between vascular dementia and alzheimer’s disease using structural mri features.
68 *Frontiers in Neurology* **10** (2019).
- 69 33. Kim, J. *et al.* Machine learning based hierarchical classification of frontotemporal dementia and
70 alzheimer’s disease. *NeuroImage: Clinical* **23** (2019).
- 71 34. Castellazzi, G. *et al.* A machine learning approach for the differential diagnosis of alzheimer and
72 vascular dementia fed by mri selected features. *Frontiers in Neuroinformatics* **14** (2020).
- 73 35. Burgos, N. *et al.* Machine learning for classification and prediction of brain diseases: recent advances
74 and upcoming challenges. *Current Opinion in Neurology* **33**, 439–450 (2020).

- 75 36. Nemoto, K. *et al.* Differentiating dementia with lewy bodies and alzheimer's disease by deep learning
76 to structural mri. *Journal of Neuroimaging* **31**, 579–587 (2021).
- 77 37. Chagué, P. *et al.* Radiological classification of dementia from anatomical mri assisted by machine
78 learning-derived maps. *Journal of Neuroradiology* **48**, 412–418 (2021).
- 79 38. Hu, J. *et al.* Deep learning-based classification and voxel-based visualization of frontotemporal demen-
80 tia and alzheimer's disease. *Frontiers in Neuroscience* **14** (2021).
- 81 39. Qiu, S., Miller, M., Joshi, P. *et al.* Multimodal deep learning for alzheimer's disease dementia assess-
82 ment. *Nature Communications* **13**, 3404 (2022).
- 83 40. Moguilner, S. *et al.* Visual deep learning of unprocessed neuroimaging characterises dementia subtypes
84 and generalises across non-stereotypic samples. *EBioMedicine* **90**, 104540 (2023).
- 85 41. Dorogush, A. V., Ershov, V. & Gulin, A. Catboost: gradient boosting with categorical features support.
86 *arXiv* (2018).
- 87 42. Cortes, C. & Mohri, M. Confidence intervals for the area under the roc curve. In Saul, L., Weiss, Y. &
88 Bottou, L. (eds.) *Advances in Neural Information Processing Systems*, vol. 17 (MIT Press, 2004).
- 89 43. Jack, C. R. J. *et al.* A/t/n: An unbiased descriptive classification scheme for alzheimer dis-
90 ease biomarkers. *Neurology* **87**, 539–547 (2016). URL [https://doi.org/10.1212/WNL.](https://doi.org/10.1212/WNL.0000000000002923)
91 0000000000002923.
- 92 44. Foster, N. L. *et al.* Fdg-pet improves accuracy in distinguishing frontotemporal dementia and
93 alzheimer's disease. *Brain* **130**, 2616–2635 (2007). URL [https://doi.org/10.1093/brain/](https://doi.org/10.1093/brain/awm177)
94 [awm177](https://doi.org/10.1093/brain/awm177).
- 95 45. McCleery, J. *et al.* Dopamine transporter imaging for the diagnosis of dementia with lewy bodies.
96 *Cochrane Database of Systematic Reviews* **2015**, CD010633 (2015).
- 97 46. Jo, M. *et al.* The role of tdp-43 propagation in neurodegenerative diseases: integrating insights from
98 clinical and experimental studies. *Experimental & Molecular Medicine* **52**, 1652–1662 (2020). Epub
99 2020 Oct 13.
- 100 47. Cairns, N. J. *et al.* Tdp-43 in familial and sporadic frontotemporal lobar degeneration with ubiquitin
101 inclusions. *The American Journal of Pathology* **171**, 227–240 (2007).
- 102 48. Qiu, S. *et al.* Development and validation of an interpretable deep learning framework for alzheimer's
103 disease classification. *Brain* **143**, 1920–1933 (2020).
- 104 49. Maia da Silva, M. N. *et al.* Frontotemporal dementia and late-onset bipolar disorder: the many direc-
105 tions of a busy road. *Frontiers in Psychiatry* **12**, 768722 (2021).
- 106 50. Arshad, F. & Alladi, S. The most difficult question in a cognitive disorders clinic. *JAMA neurology*
107 (2024).
- 108 51. Chatterjee, A. *et al.* Clinico-pathological comparison of patients with autopsy-confirmed alzheimer's
109 disease, dementia with lewy bodies, and mixed pathology. *Alzheimer's & Dementia: Diagnosis, As-*
110 *essment & Disease Monitoring* **13**, e12189 (2021).
- 111 52. Armstrong, R. A., Lantos, P. L. & Cairns, N. J. Overlap between neurodegenerative disorders. *Neu-*
112 *ropathology* **25**, 111–124 (2005). 15875904.

- 113 53. Rahimi, J. & Kovacs, G. G. Prevalence of mixed pathologies in the aging brain. *Alzheimer's Research*
114 *& Therapy* **6**, 82 (2014).
- 115 54. Livingston, G. *et al.* Dementia prevention, intervention, and care: 2020 report of the lancet commission.
116 *The Lancet* **396**, 413–446 (2020).
- 117 55. Miller, M. I., Shih, L. C. & Kolachalama, V. B. Machine learning in clinical trials: A primer with
118 applications to neurology. *Neurotherapeutics* **20**, 1066–1080 (2023). Epub 2023 May 30.
- 119 56. Ferreira, D., Nordberg, A. & Westman, E. Biological subtypes of alzheimer disease: a systematic
120 review and meta-analysis. *Neurology* **94**, 436–448 (2020).
- 121 57. Vogel, J. W. *et al.* Four distinct trajectories of tau deposition identified in alzheimer's disease. *Nature*
122 *medicine* **27**, 871–881 (2021).
- 123 58. Beekly, D. L. *et al.* The national alzheimer's coordinating center (nacc) database: an alzheimer disease
124 database. *Alzheimer Disease & Associated Disorders* **18**, 270–277 (2004).
- 125 59. Mueller, S. G. *et al.* Ways toward an early diagnosis in alzheimer's disease: The alzheimer's disease
126 neuroimaging initiative (adni). *Alzheimer's & Dementia* **1**, 55–66 (2005).
- 127 60. Boxer, A. L. *et al.* Frontotemporal degeneration, the next therapeutic frontier: Molecules and animal
128 models for frontotemporal degeneration drug development. *Alzheimer's & Dementia* **9**, 176–188
129 (2013).
- 130 61. Marek, K. *et al.* The parkinson progression marker initiative (ppmi). *Progress in Neurobiology* **95**,
131 629–635 (2011). Biological Markers for Neurodegenerative Diseases.
- 132 62. Ellis, K., Ames, D., Martins, R., Hudson, P. & Masters, C. The australian biomarkers lifestyle and
133 imaging flagship study of ageing. *Acta Neuropsychiatrica* **18**, 285–285 (2006).
- 134 63. Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C. & Buckner, R. L. Open Access Series
135 of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults. *Journal of*
136 *Cognitive Neuroscience* **22**, 2677–2684 (2010).
- 137 64. Dutt, S. *et al.* Progression of brain atrophy in psp and cbs over 6 months and 1 year. *Neurology* **87**,
138 2016–2025 (2016).
- 139 65. Linortner, P. *et al.* White matter hyperintensities related to parkinson's disease executive function.
140 *Movement Disorders Clinical Practice* **7**, 629–638 (2020).
- 141 66. Yang, J. *et al.* Establishing cognitive baseline in three generations: Framingham heart study.
142 *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **15** (2023).
- 143 67. Beekly, D. L. *et al.* The national alzheimer's coordinating center (nacc) database: the uniform data set.
144 *Alzheimer Disease & Associated Disorders* **21**, 249–258 (2007).
- 145 68. Hoopes, A., Mora, J. S., Dalca, A. V., Fischl, B. & Hoffmann, M. Synthstrip: skull-stripping for any
146 brain image. *NeuroImage* **260**, 119474 (2022).
- 147 69. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate
148 linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).

- 149 70. Fonov, V. S., Evans, A. C., McKinstry, R. C., Almlí, C. R. & Collins, D. Unbiased nonlinear average
150 age-appropriate brain templates from birth to adulthood. *NeuroImage* **47**, S102 (2009).
- 151 71. Vaswani, A. *et al.* Attention is all you need. In *Advances in neural information processing systems*,
152 5998–6008 (2017).
- 153 72. Kenton, J. D. M.-W. C. & Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for
154 language understanding. In *Proceedings of NAACL-HLT*, 4171–4186 (2019).
- 155 73. Hatamizadeh, A. *et al.* Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri
156 images. In Crimi, A. & Bakas, S. (eds.) *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic*
157 *Brain Injuries*, 272–284 (Springer International Publishing, Cham, 2022).
- 158 74. Tang, Y. *et al.* Self-supervised pre-training of swin transformers for 3d medical image analysis. In
159 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20730–20740
160 (2022).
- 161 75. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *2017*
162 *IEEE International Conference on Computer Vision (ICCV)*, 2999–3007 (2017).
- 163 76. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on*
164 *Learning Representations* (2019).
- 165 77. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv* (2014). 1412.6980.
- 166 78. Loshchilov, I. & Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International*
167 *Conference on Learning Representations* (2017).
- 168 79. Royse, S. K. *et al.* Validation of amyloid pet positivity thresholds in centiloids: a multisite pet study
169 approach. *Alzheimer's research & therapy* **13**, 99 (2021).
- 170 80. Villemagne, V. L. *et al.* Centaur: toward a universal scale and masks for standardizing tau imaging
171 studies. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **15**, e12454 (2023).
- 172 81. LONI. Image Data Archive (IDA). URL <https://ida.loni.usc.edu/login.jsp>.
- 173 82. National alzheimer's coordinating center. neuropathology data form version 10, january 2014. URL
174 <https://nacccdata.org/data-collection/forms-documentation/np-10>.
- 175 83. de Raadt, A., Warrens, M. J., Bosker, R. J. & Kiers, H. A. A comparison of reliability coefficients for
176 ordinal rating scales. *Journal of Classification* 1–25 (2021).

Dataset (group)	Age mean \pm std	Male gender (percentage)	Education in years mean \pm std	Race (White; Black; Asian; American Indian; Pacific; Multi-race)	CDR mean \pm std
NACC					
NC [n = 17242]	71.25 \pm 11.16	6009, 34.85%	15.83 \pm 2.98 [^]	(13266, 2541, 528, 109, 10, 575) [^]	0.05 \pm 0.15
MCI [n = 7582]	73.72 \pm 9.81	3615, 47.68%	15.16 \pm 3.45 [^]	(5708, 1185, 231, 53, 5, 276) [^]	0.45 \pm 0.18
AD [n = 16131]	76.0 \pm 10.31	7234, 44.85%	14.52 \pm 3.74 [^]	(13161, 1702, 354, 92, 10, 458) [^]	1.2 \pm 0.73
LBD [n = 1913]	75.01 \pm 8.55	1365, 71.35%	15.12 \pm 3.63 [^]	(1659, 128, 39, 17, 0, 37) [^]	1.29 \pm 0.78
VD [n = 1919]	80.32 \pm 8.76	947, 49.35%	14.15 \pm 4.22 [^]	(1394, 332, 67, 2, 1, 68) [^]	1.22 \pm 0.74
PRD [n = 114]	60.07 \pm 10.36	62, 54.39%	14.8 \pm 3.33 [^]	(93, 5, 5, 0, 1, 1) [^]	1.95 \pm 0.95
FTD [n = 2898]	65.86 \pm 9.36	1603, 55.31%	15.45 \pm 3.09 [^]	(2664, 69, 73, 4, 5, 39) [^]	1.2 \pm 0.83
NPH [n = 138]	79.1 \pm 9.24	69, 50.0%	15.0 \pm 3.28 [^]	(119, 10, 4, 0, 0, 4) [^]	1.18 \pm 0.71
SEF [n = 808]	76.3 \pm 11.15	413, 51.11%	14.6 \pm 3.77 [^]	(646, 95, 15, 5, 2, 31) [^]	1.11 \pm 0.7
PSY [n = 2700]	73.74 \pm 10.78	1102, 40.81%	14.13 \pm 4.12 [^]	(2163, 238, 59, 14, 5, 87) [^]	1.1 \pm 0.64
TBI [n = 265]	72.87 \pm 11.23	192, 72.45%	14.42 \pm 4.13 [^]	(212, 27, 3, 2, 1, 11) [^]	1.11 \pm 0.69
ODE [n = 1234]	72.94 \pm 12.14	654, 53.0%	14.5 \pm 3.78 [^]	(1046, 93, 28, 5, 4, 36) [^]	1.2 \pm 0.76
<i>p-value</i>	<1.0e-200	<1.0e-200	<1.0e-200	8.341e-145	<1.0e-200
NIFD					
NC [n = 124]	63.21 \pm 7.27	56, 45.16%	17.48 \pm 1.87 [^]	(89, 0, 0, 0, 0, 3) [^]	0.03 \pm 0.12 [^]
FTD [n = 129]	63.66 \pm 7.33	75, 58.14%	16.18 \pm 3.29 [^]	(109, 1, 1, 0, 0, 4) [^]	0.82 \pm 0.54 [^]
<i>p-value</i>	6.266e-01	5.246e-02	2.606e-04	6.531e-01	4.333e-28
PPMI					
NC [n = 171]	62.74 \pm 10.12	109, 63.74%	15.82 \pm 2.93	(163, 3, 2, 0, 0, 1) [^]	N.A.
MCI [n = 27]	68.04 \pm 7.32	22, 81.48%	15.52 \pm 3.08	(24, 1, 1, 0, 0, 1)	N.A.
<i>p-value</i>	1.006e-02	1.115e-01	6.194e-01	2.910e-01	N.A.
AIBL					
NC [n = 480]	72.45 \pm 6.22	203, 42.29%	N.A.	N.A.	0.03 \pm 0.12
MCI [n = 102]	74.73 \pm 7.11	53, 51.96%	N.A.	N.A.	0.47 \pm 0.14
AD [n = 79]	73.34 \pm 7.77	33, 41.77%	N.A.	N.A.	0.93 \pm 0.54
<i>p-value</i>	5.521e-03	1.887e-01	N.A.	N.A.	4.542e-158
OASIS					
NC [n = 424]	71.34 \pm 9.43	164, 38.68%	15.79 \pm 2.62 [^]	(53, 18, 1, 0, 0, 0) [^]	0.0 \pm 0.02
MCI [n = 27]	75.04 \pm 7.25	14, 51.85%	15.19 \pm 2.76	(4, 1, 0, 0, 0, 0) [^]	0.52 \pm 0.09
AD [n = 32]	77.44 \pm 7.42	20, 62.5%	15.19 \pm 2.8	(8, 1, 0, 0, 0, 0) [^]	0.86 \pm 0.44
LBD [n = 4]	74.75 \pm 5.67	4, 100.0%	16.0 \pm 2.83	N.A.	1.0 \pm 0.0
FTD [n = 4]	64.25 \pm 8.61	3, 75.0%	16.5 \pm 2.96	(4, 0, 0, 0, 0, 0)	1.25 \pm 0.75
<i>p-value</i>	7.789e-04	3.239e-03	5.507e-01	8.735e-01	2.855e-169
LBDSU					
NC [n = 134]	68.77 \pm 7.62	61, 45.52%	17.27 \pm 2.47 [^]	N.A.	N.A.
MCI [n = 35]	70.16 \pm 8.41	26, 74.29%	16.6 \pm 2.58	N.A.	N.A.
LBD [n = 13]	73.42 \pm 7.81	8, 61.54%	16.77 \pm 2.15	N.A.	N.A.
<i>p-value</i>	1.033e-01	7.863e-03	3.243e-01	N.A.	N.A.
4RTNI					
NC [n = 12]	68.08 \pm 4.92	5, 41.67%	15.45 \pm 2.57 [^]	(12, 0, 0, 0, 0, 0)	0.0 \pm 0.0
MCI [n = 31]	67.61 \pm 7.0	11, 35.48%	16.68 \pm 4.02	(25, 1, 2, 0, 1, 1) [^]	0.55 \pm 0.15
FTD [n = 37]	69.14 \pm 7.43	20, 54.05%	16.46 \pm 4.21	(31, 1, 0, 0, 1, 2) [^]	1.27 \pm 0.55
<i>p-value</i>	6.691e-01	2.992e-01	6.843e-01	7.620e-01	5.700e-16
ADNI					
NC [n = 868]	72.7 \pm 6.57	383, 44.12%	16.51 \pm 2.52	(730, 92, 28, 2, 0, 12) [^]	0.0 \pm 0.04 [^]
MCI [n = 1119]	72.77 \pm 7.65	648, 57.91%	15.97 \pm 2.75	(1023, 56, 17, 2, 2, 13) [^]	0.5 \pm 0.06
AD [n = 417]	74.99 \pm 7.78	232, 55.64%	15.25 \pm 2.92	(383, 20, 10, 0, 0, 4)	0.77 \pm 0.27
<i>p-value</i>	8.911e-08	3.090e-09	2.869e-14	2.828e-05	<1.0e-200
FHS					
NC [n = 394]	74.9 \pm 10.22 [^]	206, 52.28%	N.A.	(394, 0, 0, 0, 0, 0)	0.0 \pm 0.0
MCI [n = 434]	79.92 \pm 8.8 [^]	203, 46.77%	N.A.	(434, 0, 0, 0, 0, 0)	0.49 \pm 0.07
AD [n = 687]	82.99 \pm 7.87 [^]	211, 30.71%	N.A.	(687, 0, 0, 0, 0, 0)	2.04 \pm 0.88
LBD [n = 73]	79.34 \pm 9.37 [^]	46, 63.01%	N.A.	(73, 0, 0, 0, 0, 0)	1.84 \pm 0.84
VD [n = 113]	81.74 \pm 7.3 [^]	48, 42.48%	N.A.	(113, 0, 0, 0, 0, 0)	1.85 \pm 0.8
FTD [n = 8]	85.67 \pm 5.91 [^]	4, 50.0%	N.A.	(8, 0, 0, 0, 0, 0)	2.0 \pm 0.87
<i>p-value</i>	1.316e-31	7.905e-14	N.A.	1.0	<1.0e-200

Table 1: **Study population.** Nine independent datasets were used for this study, including ADNI, NACC, NIFD, PPMI, OASIS, LBDSU, 4RTNI, and FHS. Data from NACC, NIFD, PPMI, OASIS, LBDSU, and 4RTNI were used for model training. Data from ADNI, FHS, and a held-out set from NACC were used for model testing. The p-value for each dataset indicates the statistical significance of inter-group differences per column. We used one-way ANOVA and χ^2 tests for continuous and categorical variables, respectively. Please refer to Glossary 1 for more information on the acronyms. Here N.A. denotes not available. The symbol [^] indicates that data was not available for some subjects.

* Due to the absence of CDR scores in the FHS dataset, we used the following definition: 0.0 - normal cognition, 0.5 - cognitive impairment, 1.0 - mild dementia, 2.0 - moderate dementia, 3.0 - severe dementia.

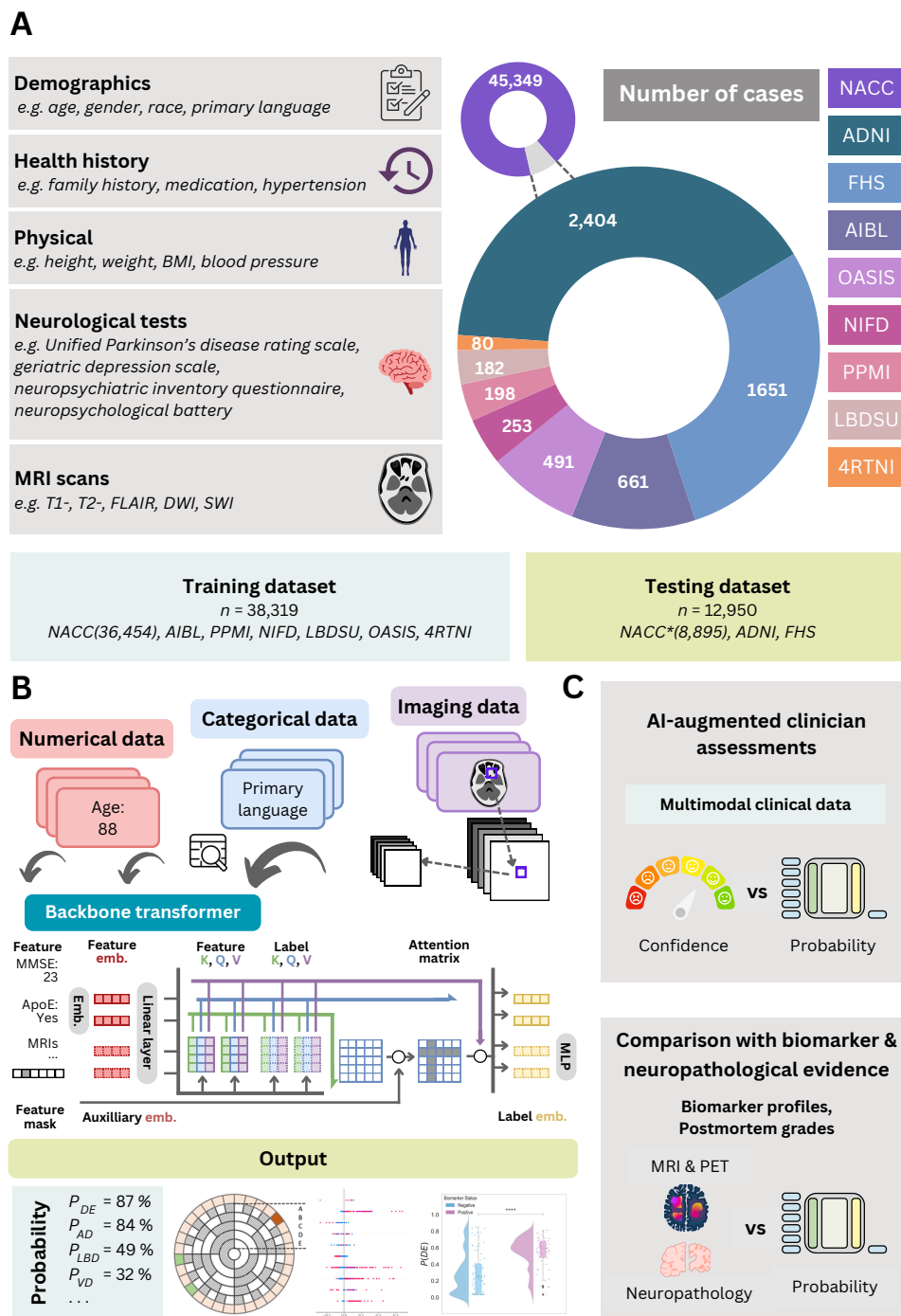


Figure 1: Data, model architecture and modeling strategy. (a) Our model for differential dementia diagnosis was developed using diverse data modalities, including individual-level demographics, health history, neurological testing, physical/neurological exams, and multi-sequence MRI scans. These data sources whenever available were aggregated from nine independent cohorts: 4RTNI, ADNI, AIBL, FHS, LBDSU, NACC, NIFD, OASIS, and PPMI (Tables 1 & S1). For model training, we merged data from NACC, AIBL, PPMI, NIFD, LBDSU, OASIS and 4RTNI. We employed a subset of the NACC dataset for internal testing. For external validation, we utilized the ADNI and FHS cohorts. (b) A transformer served as the scaffold for the model. Each feature was processed into a fixed-length vector using a modality-specific embedding strategy and fed into the transformer as input. A linear layer was used to connect the transformer with the output prediction layer. (c) A subset of the NACC dataset was randomly chosen to conduct a comparative analysis between neurologists' performance augmented with the AI model and their performance without AI assistance. Similarly, we carried out comparative evaluations with practicing neuroradiologists, who were provided with a randomly selected sample of confirmed dementia cases from the NACC testing cohort, to assess the impact of AI augmentation on their diagnostic performance. For both these evaluations, the model and clinicians had access to the same set of multimodal data. Finally, we assessed the model's predictions by comparing them with biomarker profiles and pathology grades available from the NACC, ADNI, and FHS cohorts.

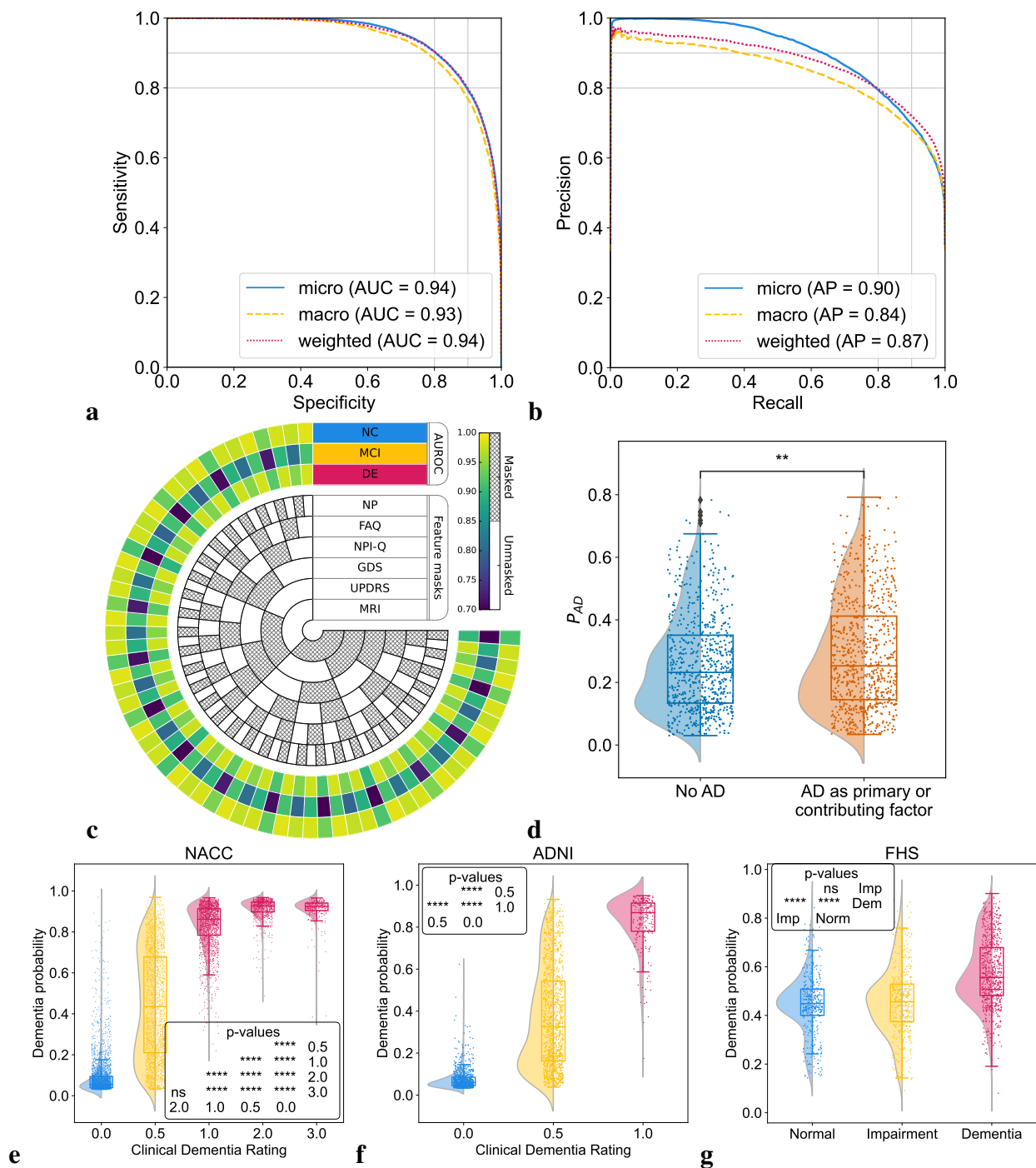


Figure 2: Model performance on individuals along the cognitive spectrum. (a,b) Receiver operating characteristic (ROC) and precision-recall (PR) curves, with their respective micro-average, macro-average, and weighted-average calculations based on the labels for NC, MCI, and DE. These averaging techniques consolidated the model's performance across the spectrum of cognitive states. Cases from the NACC testing, ADNI and FHS were used. (c) Chord diagram indicating varied levels of model performance in the presence of missing data. The inner concentric circles represent various scenarios in which particular test information was either omitted (masked) or included (unmasked). The three outer concentric rings depict the model's performance as measured by the area under the receiver operating characteristic curve (AUROC) for the NC, MCI and DE labels. (d, e, f) Raincloud plots with violin and box diagrams are shown to denote the distribution of clinical dementia rating scores (x-axis) versus model-predicted probability of dementia (y-axis), on the NACC, ADNI and FHS cohorts, respectively. (g) Raincloud plots are used to demonstrate the model's ability to distinguish between MCI cases in the NACC cohort where AD was a factor for cognitive impairment and those attributed to non-AD etiologies. For plots (d-g), significance levels are denoted as 'ns' (not significant) for $p \geq 0.05$; * for $p < 0.05$; ** for $p < 0.01$; *** for $p < 0.001$; and **** for $p < 0.0001$ based on Kruskal-Wallis H-test for independent samples followed by post-hoc Dunn's testing with Bonferroni correction.

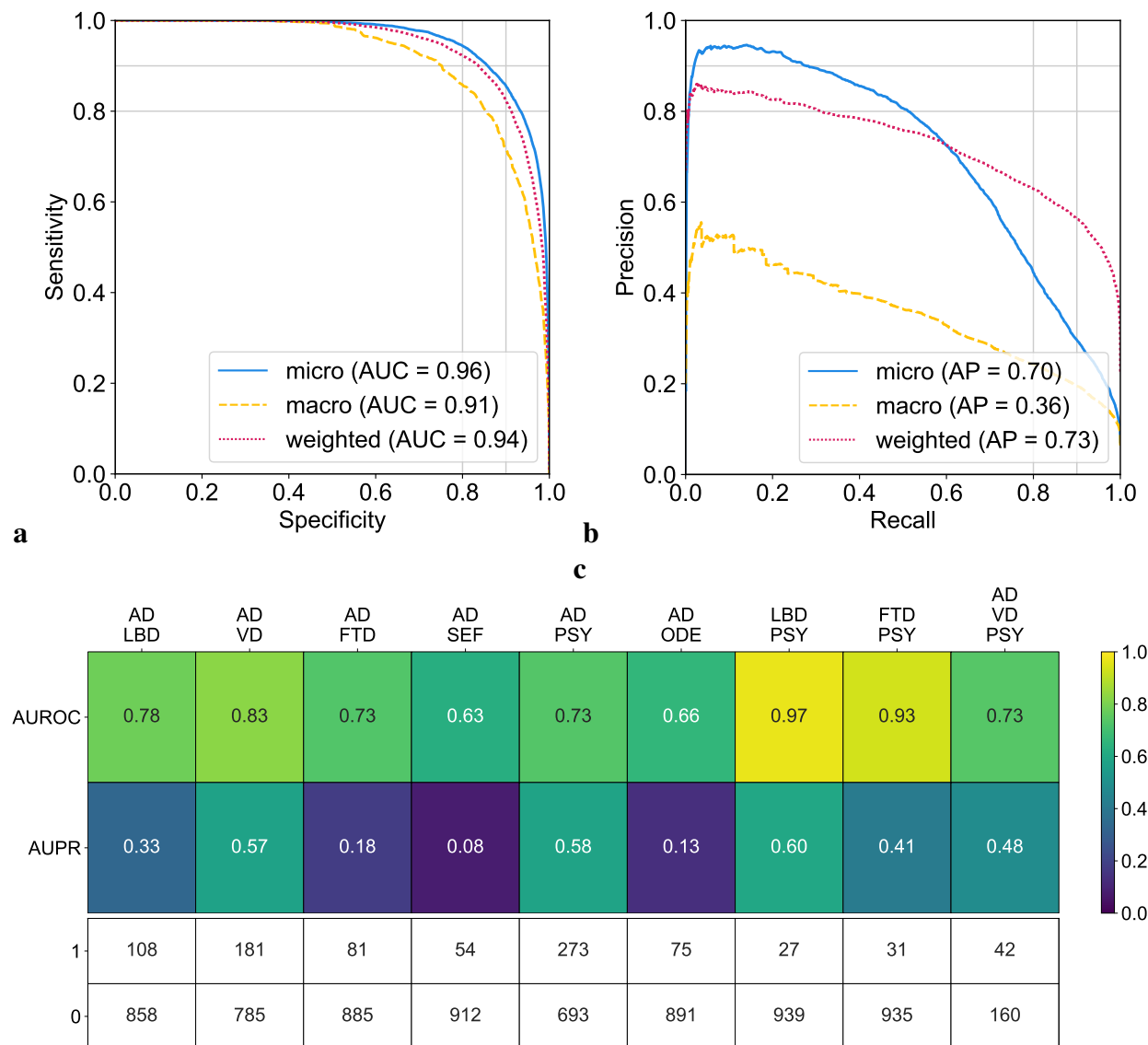


Figure 3: Model assessment on single and co-occurring dementias. (a, b) Receiver operating characteristic (ROC) and precision-recall (PR) curves are provided, utilizing micro-average, macro-average, and weighted-average methods across all the dementia diagnostic labels. These averages were computed to synthesize the performance metrics across all dementia etiologies. (c) Heatmaps are used to depict the model's performance on co-occurring dementias. We considered all combinations where two or more etiologies co-occurred from the NACC testing cohort, provided there were at least 25 positive samples. This ensured that the maximum variance of the AUROC calculation over all possible continuous distributions was upper bounded by 0.01.⁴² The first row shows the AUROC values and the second row shows the AUPR values. The table also displays the sample sizes for each case, with 1 representing a positive case and 0 indicating a negative sample.

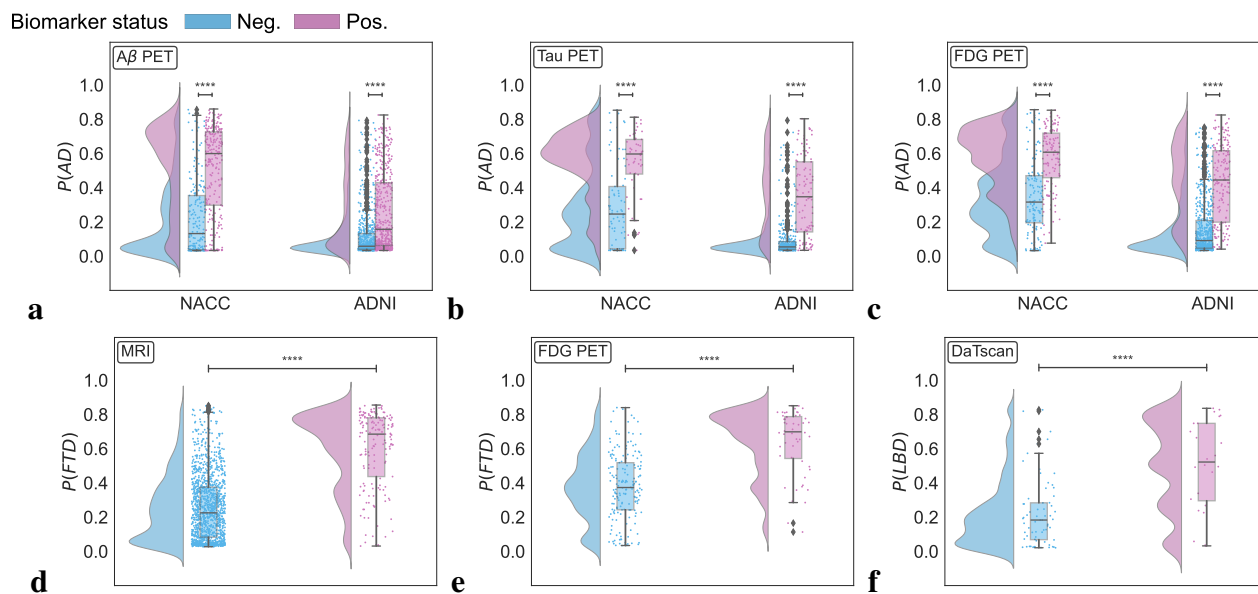


Figure 4: Biomarker-level validation. Raincloud plots representing model probabilities for dementia etiologies across their respective biomarker negative (blue) and positive groups (pink). (a-c) Model predicted probabilities for Alzheimer's disease ($P(AD)$) were analyzed in relation to amyloid β ($A\beta$), tau, and fluorodeoxyglucose (FDG) PET biomarkers. Differences between $A\beta$ negative and positive groups regarding $P(AD)$ were evaluated using a one-sided Mann-Whitney U test for the NACC cohort and a one-sided t-test for ADNI. Similar analyses for tau and FDG PET biomarkers were conducted using one-sided Mann-Whitney U tests, with **** denoting $p < 0.0001$. (d-e) For frontotemporal lobar degeneration ($P(FTD)$), probabilities were assessed across MRI and FDG PET biomarker groups in the NACC cohort, using a one-sided Mann-Whitney U test, marked by **** for $p < 0.0001$. (f) Lewy body dementia ($P(LBD)$) probabilities were analyzed between DaTscan negative and positive groups using a one-sided Mann-Whitney U test, with **** indicating $p < 0.0001$.

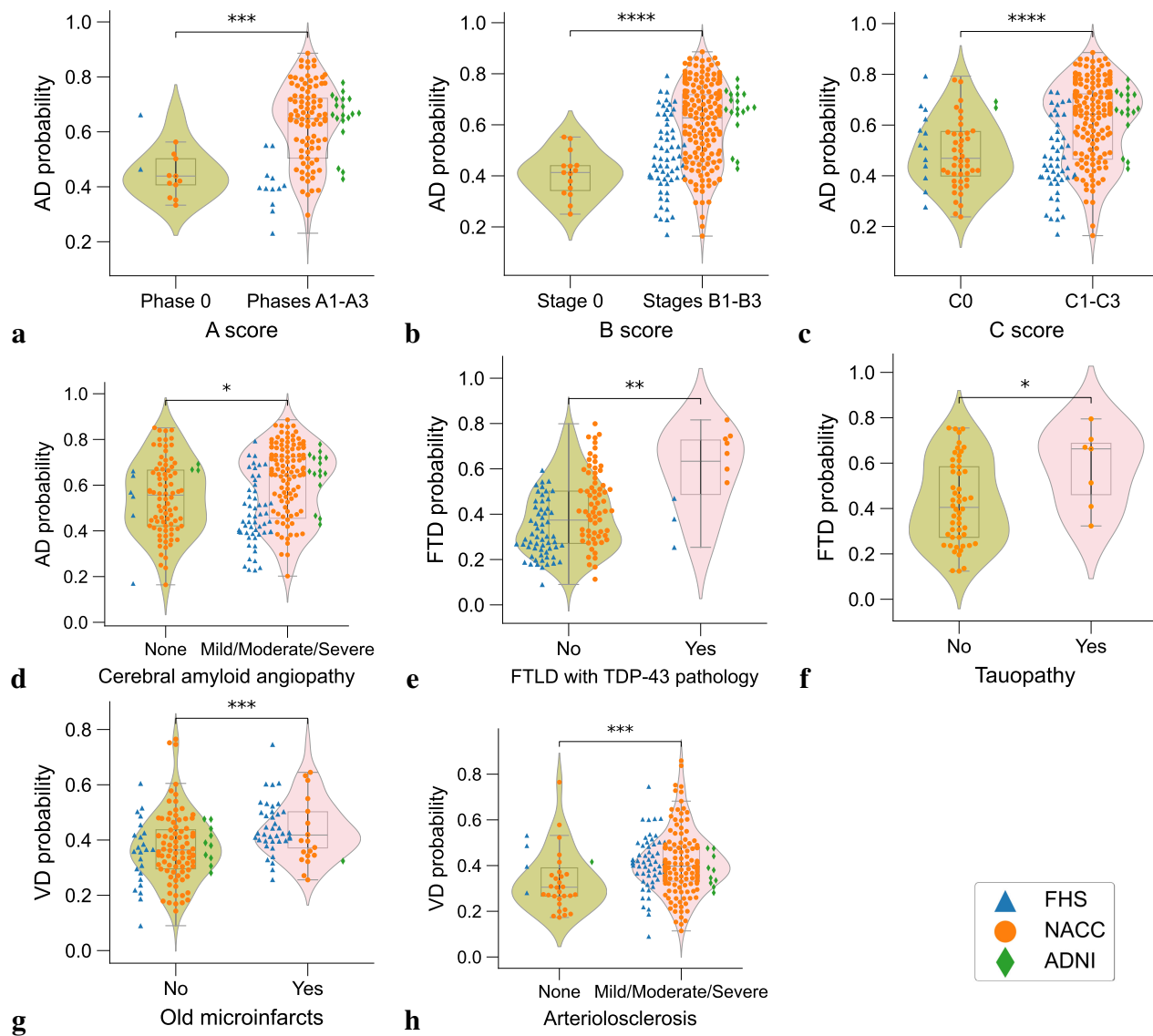


Figure 5: Neuropathological validation. Array of violin plots with integrated box plots, delineating the probability distributions as predicted by the model for different neuropathological grades. The analysis encompasses data from three distinct cohorts: the Framingham Heart Study (FHS), the National Alzheimer’s Coordinating Center (NACC), and the Alzheimer’s Disease Neuroimaging Initiative (ADNI), each denoted by unique markers (triangles, circles, and diamonds, respectively). Statistical significance is encoded using asterisks, determined by Dunn-Bonferroni post-hoc test: one asterisk (*) for $p < 0.05$; two asterisks (**) for $p < 0.01$, three asterisks (***) for $p < 0.001$, and four asterisks (****) for $p < 0.0001$, reflecting increasing levels of statistical significance. Table S13 presents more details on the statistics.

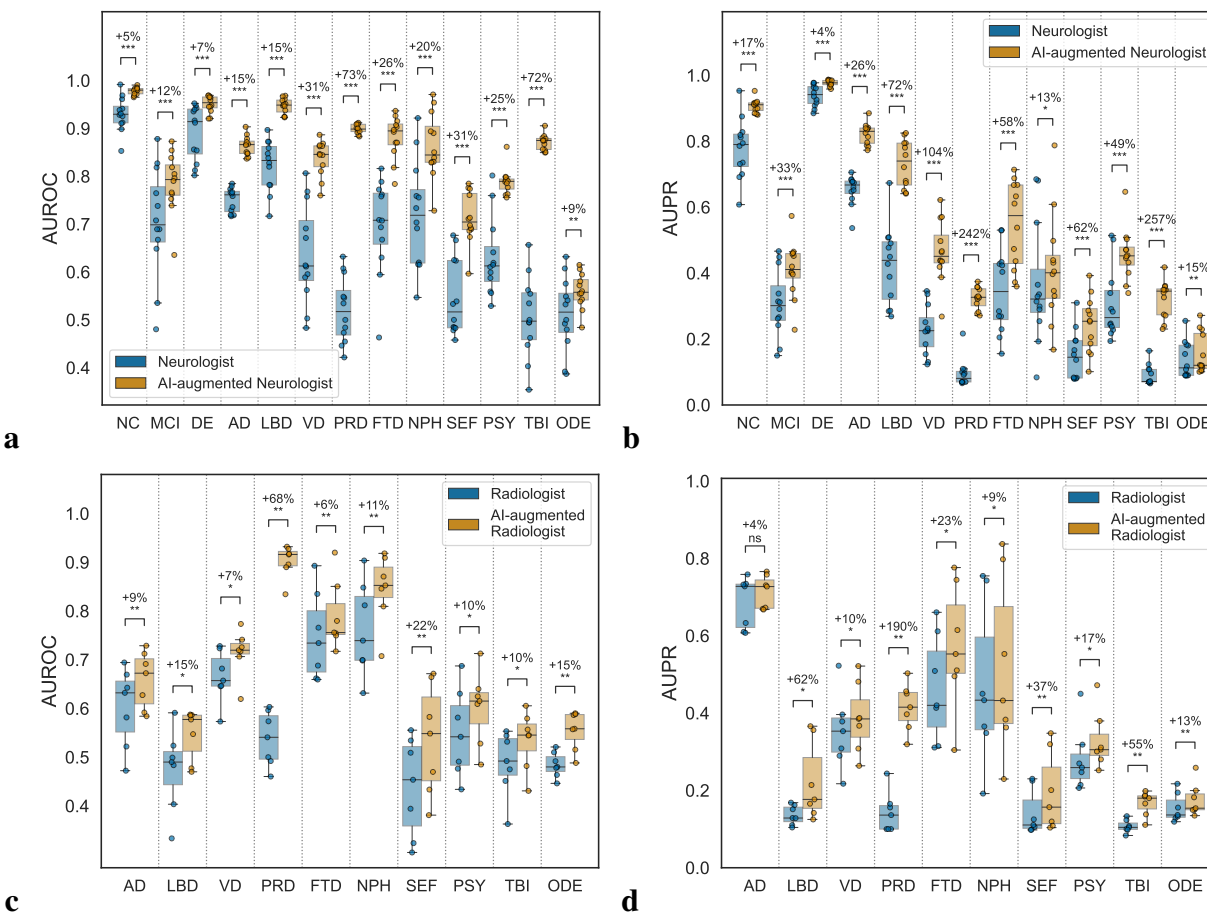


Figure 6: AI-augmented clinician assessments. Comparison between the performance of the assessments provided by practicing clinicians versus model-assisted clinicians is shown. (a-b) For the analysis, neurologists were given 100 randomly selected cases encompassing individual-level demographics, health history, neurological tests, physical as well as neurological examinations, and multi-sequence MRI scans. The neurologists were then tasked with assigning confidence scores for NC, MCI, DE, and the 10 dementia etiologies: AD, LBD, VD, PRD, FTD, NPH, SEF, PSY, TBI, and ODE (see Glossary 1). The boxplots show (a) AUROC and (b) AUPR for individual neurologist and model-assisted neurologist performance (defined as the mean between model and neurologist confidence scores). Pairwise statistical comparisons were conducted using the Wilcoxon signed-rank test and significance levels are denoted as: ns (not significant) for $p \geq 0.05$; * for $p < 0.05$; ** for $p < 0.01$; *** for $p < 0.001$; and **** for $p < 0.0001$. The percent increase in mean performance for each etiology is also presented above each statistical annotation. (c-d) Similarly, in a separate analysis, radiologists were given 70 randomly selected cases with a confirmed dementia diagnosis encompassing individual-level demographics and multi-sequence MRI scans. The radiologists were tasked with assigning confidence scores for the 10 dementia etiologies, and the boxplots show (c) AUROC and (d) AUPR for individual radiologist and model-assisted radiologist performance for the 10 etiologies. Statistical annotations and percent increase in mean performance with respect to each etiology are shown in a similar fashion.