AI-based differential diagnosis of dementia etiologies on multimodal data

Chonghua Xue[1,2,*], Sahana S. Kowshik[1,3,*], Diala Lteif[1,4], Shreyas Puducheri[1], Olivia T. Zhou[1], Anika S. Walia[1], Osman B. Guney[1,2], J. Diana Zhang[1,5], Serena T. Pham[6], Artem Kaliaev[6], V. Carlota Andreu-Arasa[6]†, Brigid C. Dwyer[7]†, Chad W. Farris[6]†, Honglin Hao[8]†, Sachin Kedar[9]†, Asim Z. Mian[6]†, Daniel L. Murman[10]†, Sarah A. O'Shea[11]†, Aaron B. Paul[12]†, Saurabh Rohatgi[12]†, Marie-Helene Saint-Hilaire[7]†, Emmett A. Sartor[7]†, Bindu N. Setty[6]†, Juan E. Small[13]†, Arun Swaminathan[14]†, Olga Taraschenko[10]†, Jing Yuan[8]†, Yan Zhou[8]†, Shuhan Zhu[15]†, Cody Karjadi[16], Ting Fang Alvin Ang[16,17], Sarah A. Bargal[18], Bryan A. Plummer[4], Kathleen L. Poston[19], Meysam Ahangaran[1], 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

Abstract

Differential diagnosis of dementia, with its overlapping symptomatology, remains a significant challenge in neurology. Here we present an algorithmic framework employing state-of-the-art techniques such as transformers as well as self-supervised frameworks and harnessing a broad array of data including demographics, person-level and family medical history, medication use, neuropsychological exams, functional evaluations, and multimodal neuroimaging to identify the etiologies contributing to dementia in individuals. The study utilized 9 independent, geographically diverse datasets, including the National Alzheimer's Coordinating Center with $45,349$ participants, the Alzheimer's Disease Neuroimaging Initiative encompassing $1,821$ participants, and the Frontotemporal Lobar Degeneration Neuroimaging Initiative comprising 253 participants. Additionally, the Parkinson's Progression Marker Initiative with 198 participants, the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing cohort including 661 participants, the Open Access Series of Imaging Studies dataset with 491 participants, and the 4 Repeat Tauopathy Neuroimaging Initiative comprising 80 participants were used. The study also included two in-house datasets: one from the Lewy Body Dementia Center for Excellence at Stanford University with 182 participants, and another from the Framingham Heart Study including $1,651$ individuals. Our model traverses the intricate spectrum of dementia by mirroring real-world clinical settings, aligning diagnoses with similar management strategies, and delivering robust predictions, even in the face of incomplete data. On the testing cohort, our model achieved a micro-averaged area under the receiver operating characteristic curve (AUROC) of 0.93, and a micro-averaged area under precision-recall curve (AUPR) of 0.87, in classifying individuals with normal cognition, mild cognitive impairment and dementia. Also, the micro-averaged AUROC was 0.95 and micro-averaged AUPR was 0.68 in differentiating 10 distinct dementia etiologies, defined through a consensus among a team of neurologists. One key strength lies in our model's capability to address mixed dementias, a prevalent challenge in clinical practice, and the incorporation of interpretability techniques further unveiled vital disease-specific patterns. On a randomly selected subset ($n = 100$), our model differentiated true positive and true negative cases across 12 out of 13 categories ($p < 0.01$), as opposed to the neurologists' expertise in identifying 9 out of these 13 categories ($p < 0.01$). Furthermore, the model's correlations with different proteinopathies were substantiated through postmortem analyses. This included a significant association with the global Alzheimer's disease neuropathologic change (ADNC) score ($p < 0.001$), and notable correlations with TDP-pathology, the presence of old microinfarcts, arteriosclerosis, and Prion disease (all with $p < 0.05$). Our framework has the potential to be integrated as a screening tool for dementia in various clinical settings and drug trials, with promising implications for person-level management.

Research in context

Systematic review: Previous studies have demonstrated that models utilizing multimodal data can differentiate individuals across the dementia spectrum, identifying those with normal cognition (NC), mild cognitive impairment (MCI), and dementia (DE). Some studies have also ventured beyond this tripartite classification, aiming to differentiate Alzheimer's disease (AD) from other forms of non-AD dementia. Majority of these investigations have approached the task as a binary classification, primarily focusing on the distinction between AD and other dementia types. Also, limited studies have effectively tackled the intricate challenge of diagnosing mixed dementia, which is a common and complex issue encountered in clinical practice.

Methods and findings: Employing multimodal data from 9 distinct cohorts, encompassing $50,686$ participants, we developed an algorithmic framework that leverages transformers and self-supervised learning to facilitate differential dementia diagnoses. This model adeptly classifies individuals into 13 curated diagnostic categories, each tailored to reflect real-world clinical needs. These categories comprehensively cover the cognitive spectrum, ranging from NC, MCI to DE, and extend to 10 distinct dementia types. Our model demonstrates the capability to accurately diagnose dementia, even with incomplete data, and efficiently manage cases involving multiple co-occurring dementia conditions, a common occurrence in clinical practice. It has shown commendable performance, surpassing expert clinical assessments, and its predictions have been corroborated by postmortem data, particularly in relation to various proteinopathies.

Interpretation: Our work provides a robust and adaptable framework for comprehensive dementia screening for drug trials and in various clinical settings, ranging from primary care to memory clinics.

Dementia is one of the most pressing health challenges of our time. With nearly 10 million new cases reported annually, this syndrome, characterized by a progressive decline in cognitive function severe enough to impede daily life activities, continues to present considerable clinical and socio-economic challenges. In 2017, the World Health Organization's global action plan on the public health response to dementia highlighted the prompt and precise diagnosis of dementia as a pivotal strategic objective.[1] As such, diagnostic precision in the varied landscape of dementia remains an unmet need, even as the demands for such precision escalate with the aging global population and the imperative for better participant screening in drug trials. The clinical presentation of different dementia forms often overlaps, further compounded by the heterogeneity in findings on magnetic resonance imaging (MRI) scans. The necessity for improvements in the field becomes ever more pressing considering the projected shortage of specialists including neurologists, neuropsychologists and geriatric care providers,[2–4] emphasizing the urgency to innovate and evolve our diagnostic tools. While Alzheimer's disease (AD) is a leading cause, other forms such as vascular dementia (VD), Lewy body dementia (LBD), and frontotemporal dementia (FTD) are also prevalent, and often co-occur in most individuals. The significant intersection of symptoms among these disorders and other dementia etiologies, amplified by the varying intensity of symptom manifestation, frequently complicates the process of differential diagnosis.[5]

The imperative for enhanced diagnostic accuracy in AD and related neurodegenerative disorders is becoming increasingly critical in the context of significant advancements in medical diagnostics. While recent regulatory approvals have facilitated the transition of cerebrospinal fluid (CSF) biomarkers and advanced imaging techniques like positron emission tomography (PET) from research environments to broader clinical applications, the development and potential clinical integration of blood-based biomarkers remain an area of active research.[6,7] However, accessibility to these diagnostic tools remains constrained, not only in remote and economically developing regions but also in urban healthcare centers, as exemplified by prolonged waiting periods for specialist consultations. This challenge is compounded by a global shortage of specialists, such as behavioral neurologists and neuropsychologists, leading to overreliance on cognitive assessments that may not be culturally appropriate due to the lack of formal training programs in neuropsychology in many parts of the world. The recent advent of disease-modifying therapies marks a significant shift in the treatment landscape for AD,[8,9] further highlighting the necessity of early and accurate risk stratification in both primary care and general neurology settings. While conventional methods like clinical evaluations, neuropsychological testing, and MRI remain central to antemortem differential dementia diagnosis, their effectiveness heavily relies on the diminishing pool of specialist clinicians. This situation underscores an urgent need for the healthcare system, particularly in primary care, to evolve and adapt to the rapidly changing dynamics of dementia diagnosis and treatment.

Machine learning holds promising potential for enhancing the accuracy and efficiency of dementia diagnosis.[10–12] While most previous research leveraging these advanced frameworks has concentrated on neuroimaging data,[13,14] few have ventured to combine imaging with non-imaging data,[15] such as demographics, medical histories, and neuropsychological assessments to distinguish cognitively normal (NC) individuals from those with mild cognitive impairment (MCI) and dementia. AD being the most prevalent type of dementia, has naturally been the primary focus of much of this research. Few studies have also attempted to discern neuroimaging signatures unique to AD by contrasting them with other dementia etiologies.[16–22] Recently, we proposed a nuanced approach to stratify individuals based on cognitive status and discern likely AD cases from other non-AD dementia types.[23] These investigations, coupled with other studies,[24–26] have begun to illuminate the complex matrix of factors contributing to dementia. Moreover, the challenge of differential diagnosis in dementia is further compounded by the inherent limitations in existing dementia risk scoring systems.[27–29] Research indicates that individualized assessments based on these

generic scores frequently lead to significant error rates, thus diminishing their effectiveness in accurately identifying individuals suitable for targeted dementia prevention strategies.[30] These observations highlight the critical need for innovative solutions in differential dementia diagnosis, emphasizing the integration of diverse data sets to surpass the limitations of traditional risk assessment methods.

In this study, we propose a multimodal machine learning framework, harnessing a diverse array of data, including demographics, personal and family medical history, medication use, neuropsychological tests, functional evaluations, and multimodal neuroimaging to perform differential dementia diagnosis. The model incorporates state-of-the-art techniques such as transformers and self-supervised learning, enabling it to navigate the spectrum of dementia conditions with improved performance. Our model for differential dementia diagnosis reflects real-world scenarios, with diagnostic categories designed for clinical relevance, aligning diagnoses with similar management strategies to aid neurologists and other practitioners in screening and treatment planning. The model's robustness is demonstrated through rigorous validation on independent, geographically diverse datasets, achieving parity with expert clinical diagnoses in comparative analysis. By employing advanced interpretability methods, our model elucidates disease-specific patterns critical for differential diagnosis, contributing significantly to our understanding of heterogeneous dementia phenotypes. The fidelity of these patterns was confirmed by postmortem data, underscoring our model's ability to dissect the intricate pathophysiology of dementia. Our algorithmic framework opens new avenues for dementia screening in various clinical settings, with significant implications for person-level management. This study underscores the potential of AI-driven tools in healthcare, paving the way for improved diagnostic accuracy, efficient resource utilization, and better outcomes.

Results

Glossary 1

Acronym	Description
NC	Normal cognition
MCI	Mild cognitive impairment
DE	Dementia
AD	Alzheimer's disease including Down syndrome
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, multiple systems atrophy, Huntington's disease, seizures, etc.

Leveraging the power of routinely gathered clinical data, our model provides a nuanced approach to differential dementia diagnosis. This framework assigns individuals to one or more of the thirteen diagnostic categories (refer to Glossary 1), which were meticulously defined through consensus among a team of expert neurologists. This practical and intuitive categorization is designed with clinical management pathways in mind, thereby echoing real-world scenarios. For instance, we have grouped dementia with Lewy bodies and Parkinson's disease dementia under the comprehensive category of Lewy body dementia (LBD). This classification stems from an understanding that the care for these conditions often follows a similar path, typically overseen by a multidisciplinary team of movement disorder specialists. In the context of vascular dementia (VD), we included individuals who exhibited symptoms of a stroke, possible or probable VD, or vascular brain injury. This encompassed cases with symptomatic stroke, cystic infarct in cognitive networks, extensive white matter hyperintensity, and/or executive dysfunction, where these conditions were identified as the primary contributors to the observed cognitive impairment. The inclusion criteria were based on the expectation that such patients would typically receive care from clinicians specializing in stroke and vascular diseases. Likewise, we have considered various psychiatric conditions - such as schizophrenia, depression, bipolar disorders, anxiety, and post-traumatic stress disorder - under one category (PSY), acknowledging that their management predominantly falls within the realm of psychiatric care providers. By aligning diagnostic categories with practical clinical care pathways, our model serves not only to demystify

an individual's condition but also to direct efficient and appropriate clinical management strategies. The result is a scientifically robust tool that is also intuitively aligned with the real-world scenarios encountered in clinical care.

Our model's development and testing leveraged a robust variety of data, encompassing thousands of individuals from population-level studies, community, clinical as well as research cohorts (Table 1, Table S1 & Fig. 1). This broad, diverse set of data serves as a testament to our study's core strength, and the ensuing results speak to the model's strong accuracy and generalizability. The model accurately classified cognitive status across the full spectrum of clinical diagnosis tasks. Across three categories - normal cognition (NC), mild cognitive impairment (MCI), and dementia (DE) - our model achieved high performance as shown via accuracy (Acc.), sensitivity (Se.), specificity (Sp.), Matthews correlation coefficient (MCC), as well as the area under the receiver operating characteristic (AUC) and area under precision-recall (AP) curves (Fig. 2a). Further, the model delivered striking results across AD, LBD, VD, PRD, FTD, NPH, SEF, PSY, TBI, and ODE, reaffirming the model's high performance and reliability in diagnosing various forms of dementia. With these compelling metrics, we present robust empirical evidence of our model's ability to tackle differential dementia diagnosis efficiently and effectively. Also, the utility of our model is notably demonstrated in its robustness and adaptability to various datasets acquired via different protocols at multiple institutions (Fig. 2a). An example of this is found in our tests on data from the National Alzheimer's Coordinating Center (NACC) cohort that was not used during model training. The model continued to exhibit its robust performance on external cohorts, including the Alzheimer's Disease Neuroimaging Initiative (ADNI), and the Framingham Heart Study (FHS), substantiating its consistency across diverse datasets. Despite the fact that the ADNI and FHS datasets comprised of a limited set of factors (Table S2), the model effectively accomplished all classification tasks. The model's ability to deliver reliable predictions with constrained data sets suggests that its performance, already robust under conditions of incomplete information, could be further optimized with access to more comprehensive datasets. To further evaluate the model's resilience to incomplete data, we artificially introduced varying levels of data missingness in the NACC cohort and assessed the impact on its predictive performance. By selectively removing portions of the data from the NACC cohort, we aimed to test our model's predictive performance under various constraints. As depicted in the chord diagram (Fig. 2b), even when confronted with missing data elements — whether it be MRI results, UPDRS, GDS, NPI-Q, FAQ, NP tests or other parameters — our model consistently produced high, reliable scores. This reinforces not only its predictive stability but also its potential applicability in diverse clinical scenarios where complete datasets might be unattainable.

We applied two-dimensional t-distributed stochastic neighbor embedding (tSNE) to evaluate the distribution of our model's predictions within the cognitive spectrum encompassing NC, MCI and DE as depicted in Fig. 2c. The visualization distinctly segregates the cognitive states using color codes: a densely populated blue pathway denotes NC, indicating a homogeneous assemblage of individuals with normal cognitive health. The points coded in yellow for MCI exhibit a sparser distribution, symbolizing a transitional and more varied cohort positioned between NC and DE. The pink stream earmarks DE, forming a dense aggregation that likely corresponds to individuals diagnosed with dementia. This graphical representation proves critical for distilling complex, multi-dimensional datasets into a comprehensible two-dimensional space, thus providing a graphical synopsis of how individual-level multimodal data processed via our modeling framework can demarcate these cognitive conditions.

Shapley analysis [31] was employed to determine feature importance on the NACC test set. The process began by categorizing cases according to their labels, focusing exclusively on those with correct predictions

for Shapley value calculations. This approach yielded feature-specific Shapley values for the chosen cases, illuminating each feature's role in influencing the model's final decision on the NC, MCI and DE cases, respectively (Figs. 2d-f). These features were arranged based on their mean impact, quantified by the average magnitude of their respective Shapley values. Additionally, the computation of Shapley values incorporated considerations for data missingness; Shapley values of the features that were unavailable were assigned a constant value of zero. The Shapley value distribution for each feature indicates the magnitude and direction of the feature's impact on the model's prediction, with higher absolute Shapley values signifying greater influence. For NC, key features included cognitive status based on neuropsychological exam, MoCA and MMSE scores, memory-related tasks (like naming vegetables and animals), and daily functionality questions. For MCI, similar cognitive and memory-related features were found to be impacting, in addition to questions about daily living activities and independence levels. For DE, the model placed high importance on cognitive status based on neuropsychological exams, the difficulty with daily tasks, and medication usage for AD symptoms. Across all conditions, the plots show that certain features (e.g., cognitive status based on neuropsychological exam) consistently have a higher impact on the model's predictions, as indicated by larger Shapley values. The distribution of Shapley values for each feature — spread across a spectrum from high to low impact — provides insights into the heterogeneity of feature influence on the prediction of each cognitive state. The consistency of cognitive assessments in predicting across all cognitive states suggests their fundamental role in the model's decision-making process. Conversely, the variability in the impact of daily functionality and memory-related tasks underscores the nuanced differences in their relevance to each cognitive state. Overall, the Shapley value plots offered a detailed and quantifiable visualization of how each feature contributes to the model's predictions, which could be crucial for understanding and improving the model's interpretability and accuracy in a clinical setting.

The receiver operating characteristic (ROC) and precision-recall (PR) curves reflected strong model performance across different averaging methods (Figs. 2g-h). On the test cohort, comprising the NACC dataset (unused in training) along with ADNI and FHS data, our model demonstrated strong classification abilities for NC, MCI, and DE, achieving a micro-averaged AUROC of 0.93 and a micro-averaged AUPR of 0.87. Additionally, the macro-averaged metrics showed an AUROC of 0.91 and a AUPR value of 0.80. The weighted-average AUROC and AUPR values further underscored the model's efficacy, standing at 0.92 and 0.84, respectively. Of note, the micro-average approach consolidates true positives, true negatives, false positives, and false negatives from all classes into a unified curve, providing a global performance metric. In contrast, the macro-average calculates individual ROC/PR curves for each class before computing their unweighted mean, disregarding potential class imbalances. The weighted-average, while similar in approach to macro-averaging, assigns a weight to each class's ROC/PR curve proportionate to its representation in the dataset, thereby acknowledging class prevalence. The slight variations observed in the AUROC and AUPR values between the averaging methods might reflect the influence of class distribution and prevalence in the performance metric calculations. Overall, the micro-averaging method performed slightly better, which could be due to the fact that it takes the class imbalance into account by weighting the performance by the number of samples in each class.

We conducted a comparison between the model's predicted probability scores and the clinical dementia ratings (CDR) available for all participants in the NACC testing, and ADNI cohorts (Figs. 2i & 2j). Remarkably, despite not incorporating CDR scores as input during model training, our predictions exhibited a strong correlation with CDR scores. In the analysis of the NACC dataset, the model's predictions exhibited increasing heterogeneity as a function of ascending Clinical Dementia Rating (CDR), with statistically significant divergences manifest across the spectrum of cognitive impairment ($p < 0.0001$). Notably, this pattern did not hold between CDR scores of 2.0 and 3.0, where no significant statistical difference was dis-

cerned. The ADNI dataset revealed a statistically significant demarcation ($p < 0.0001$) in model-predicted probabilities between the baseline CDR rating and higher gradations. This underscores the model's sensitivity to incremental exacerbations in clinical dementia assessments. For the FHS dataset (Fig. 2k), which substitutes a consensus panel's diagnostic categorization (normal, impaired, and dementia) for CDR scores, a marked statistical significance ($p < 0.0001$) was evident in the model's predicted probabilities across these diagnostic strata, with the exception of the distinction between normal cognition and impairment. This suggests a nuanced challenge for the model in discriminating between the initial stages of cognitive decline. Collectively, these findings illuminate the model's robust capacity to delineate differential cognitive states, showcasing its potential as a discerning tool for identifying levels of cognitive impairment across diverse clinical datasets.

Our methodology for conducting differential diagnosis led to the simultaneous prediction of probabilities associated with multiple potential causes of dementia, as visually represented in Fig. 3. To delve deeper into the intricate nature of real-world clinical scenarios, where persons frequently exhibit symptoms suggesting the coexistence of multiple dementia types, we created a comprehensive visualization,[32] which depicts probabilities that reflect co-occurring conditions. These plots provided a valuable tool for understanding the complex landscape of dementia diagnoses, where a substantial number of cases involve a combination of contributing factors. The box-and-whisker plot on the joint probabilities is particularly informative in illustrating the range and variability of the model's predictions. For instance, the uppermost row corresponds to non-dementia cases. Here, the model's predictions are expressed as joint probabilities — the product of individual probabilities for each condition, symbolized as $p(AD = 0) \times p(PSY = 0) \times p(ODE = 0) \times p(SEF = 0) \times p(LBD = 0) \times p(VD = 0) \times p(PRD = 0) \times p(FTD = 0) \times p(NPH = 0) \times p(TBI = 0)$. The bottom row highlights cases that include the combination of AD, PSY, ODE and SEF, where the model's predictions are expressed as $p(AD = 1) \times p(PSY = 1) \times p(ODE = 1) \times p(SEF = 1) \times p(LBD = 0) \times p(VD = 0) \times p(PRD = 0) \times p(FTD = 0) \times p(NPH = 0) \times p(TBI = 0)$. This level of quantification underscores the model's capacity to gauge the intricacies of dementia presentations that are multifactorial in nature, highlighting its utility in complex clinical assessments. Taken together, these findings demonstrate that our model effectively identified instances where multiple etiologies were at play, reflecting its ability to navigate the challenging task of recognizing mixed causes of dementia. This accomplishment underscores the significant contribution of our research, particularly in addressing the diagnostic complexities encountered in clinical practice when individuals present with overlapping symptoms indicative of multiple dementia types.

Similar to the Shapley value illustrations shown in Figs. 2d-f, we observed that Shapley value analyses revealed distinct feature importance across various dementia types (Figs. 4a-j). For instance, features such as 'cognitive status based on neuropsychological exams', 'level of independence', and 'difficulty in performing daily activities' were influential across several dementia categories. Such consistency of certain features across various dementia types underscores their potential as universal markers of neurodegeneration. Moreover, the analysis identified several features that are emblematic of well-known etiologies underlying dementia, as well as factors that are widely recognized as influential in driving the progression of specific diseases. In predicting AD, factors like the presence of ApoE alleles and the use of FDA-approved medication for AD symptoms emerged as influential. For LBD, a prior diagnosis of Parkinson's disease, the presence of gait disorders, and speech difficulties were among the key predictors. VD prediction was driven by the history of stroke, the impact of cerebrovascular disease on cognitive impairment, and the occurrence of multiple infarcts. Presence of depression or dysphoria in the last month as well as GDS score were significant for predicting psychiatric conditions (PSY). These plots collectively demonstrate the model's reliance on a mix of features, reflecting the multifaceted nature of dementia disorders and the complexity of their

prediction. Additionally, the ROC and PR curves reflected strong model performance on the model's overall assessment on identifying dementia etiologies across different averaging methods (Figs. 4k-l). Our model attained impressive results, with micro-averaged values of AUROC and AUPR at 0.95 and 0.68, respectively. In macro-averaged terms, the AUROC and AUPR stood at 0.91 and 0.34. Moreover, the weighted-average values for AUROC and AUPR were 0.94 and 0.72, respectively.

Etiology-specific model probability scores revealed significant correlations with neuropathological evidence found in common dementia types. The composite violin and box plots in Fig. 5 illustrate the distribution and probability distributions and median tendencies for each cohort indicating that with increasing neuropathological severity, there is a corresponding elevation in the likelihood of neurodegeneration according to the model. The first three plots (Figs. 5a-c) compare baseline stages of Thal phase for Aβ plaques (Thal), Braak stage for neurofibrillary degeneration (Braak) and density of neocortical neuritic plaques (CERAD) against progressive Thal, Braak and CERAD stages (A1-A3, B1-B3 and C1-C3, respectively). Each demonstrated an upward shift in the median probability of AD and an expansion of the interquartile range as the stages advance, with statistical significance ($p < 0.001$ for Thal stage and $p < 0.0001$ for Braak and CERAD stages, respectively). Also, we rejected the null hypothesis of there being no significant differences in model-predicted AD probabilities between semi-quantitative scores of Thal, Braak and CERAD scores (Fig. S2). Furthermore, by contrasting individuals without AD against those with varying degrees of NIA-AA Alzheimer's disease neuropathologic change (ADNC), which unifies the Thal, Braak and CERAD scores (Fig. 5d), we observed a similar shift towards higher AD probabilities in the latter group ($p < 0.001$). Collectively, these plots illustrate a clear trend where advancing stages of AD-related neurodegeneration are associated with increased probabilities of AD. Finally, we rejected the null hypothesis of there being no significant differences in FTD probability between the presence or absence of TDP-43 pathology ($p < 0.05$), VD probability between the presence or absence of old microinfarcts ($p < 0.05$) and arteriosclerosis ($p < 0.05$), as well as PRD probability between the presence or absence of Prion disease ($p < 0.05$)(Figs. 5e-h). The results are consistent with the well-documented association between TDP-43 protein aggregation and its prevalence in FTD as well as other neurodegenerative diseases.[33,34] Additionally, the clear linkage between cerebrovascular pathologies and the incidence of VD is reinforced by our data. Crucially, these outcomes highlight the capability of our AI-driven framework to align model-generated probability scores with a range of neuropathological states beyond AD, supporting its potential utility in broader neurodegenerative disease research.

The incorporation of independent confidence scores for diagnostic tasks improved the model's interpretability, and facilitated comparison of model performance with the clinicians. Neurologists reviewed 100 randomly selected cases, including various dementia subtypes, with comprehensive data including demographics, medical history, neuropsychological tests, and multi-sequence MRI scans. The Shapiro-Wilk test revealed non-normal distributions in the confidence scores, both from the model and the experts. Consequently, the Brunner-Munzel test was applied to compare the differences between these sets of confidence scores. Pearson correlation analysis was used to measure the interrater reliability among the confidence scores assigned by different evaluators. Notably, in instances where the diagnosis was confirmed (true positives), the neurologists' confidence scores across categories such as NC, MCI, DE, AD, LBD, VD, FTD, NPH, and PSY were statistically significant in comparison to cases deemed non-diagnostic (true negatives) ($p < 0.01$) (Fig. 6a). Similarly, the model's probabilities in true positive cases across an extended range of conditions — including those aforementioned plus PRD, SEF, and TBI — were statistically significant when contrasted with true negatives ($p < 0.01$). However, for certain etiologies, including PRD, SEF, TBI, and ODE, there was no apparent statistical distinction in the confidence ratings provided by neurologists between true positive and true negative cases. This absence of statistical significance was also observed in the model's

predicted probabilities for ODE between true positive and true negative cases. Several potential explanations may elucidate this inconsistency. Despite providing neurologists with comprehensive information, including demographics, medical history, medication usage, neuropsychological test results, functional assessments, and multimodal neuroimaging data for all cases, this information might have been insufficient to confirm the impact of these etiologies. Additionally, the model's limitations could be attributed to a smaller training dataset specific to certain etiologies and a potential deficiency in the breadth of features needed for the identification of etiology-specific contributions, especially concerning ODE. Pearson correlation coefficients offered a comprehensive overview of the interrelationships among the neurologists and the model across a spectrum of neurological conditions (Fig. 6b). NC was distinguished by the highest correlation coefficient between the model-predicted probabilities and mean neurologist confidence scores (0.92 ± 0.02), indicating robust, consistent associations in this group. DE showed notable correlations (0.89 ± 0.02). In contrast, MCI, AD, LBD, VD, FTD and PSY exhibited modest correlations, highlighting the potential heterogeneity within these disorders. The lower correlations observed on certain etiologies (PRD, NPH, SEF, TBI and ODE) underscore the diverse and complex nature of these conditions compounded by the lack of extensive features to tease out their unique signatures. These observations underscore the model's potential in complementing neurologist expertise, yet also highlights the complexities and limitations in accurately diagnosing more diverse and less represented conditions.

In a separate assessment, neuroradiologists evaluated a randomly selected set of 70 clinically diagnosed dementia cases, concentrating on MRI findings and demographic information. They provided ratings for atrophy and pathological changes within specific brain regions, including the temporal lobes, frontal lobe, insula, limbic systems, fusiform gyrus, and overall brain considerations (Fig. S3). The calculated pairwise Pearson correlation coefficients, representing interrater reliability among seven neuroradiologists, revealed a moderate overall agreement with a mean coefficient of (0.39 ± 0.02). Radiologists were mostly in agreement on the whole brain-level assessments related to the presence of hyperintensities (0.67 ± 0.03) and identifying prior infarcts (0.59 ± 0.08). Within the temporal lobe, the highest concordance was found in ratings of the anterior temporal lobe (0.68 ± 0.09), while the posterior temporal lobe showed the least agreement (0.18 ± 0.17). An interesting pattern emerged, indicating a trend towards greater agreement in assessments of the right-sided brain regions compared to the left-sided ones. We stratified cases based on expert consensus on atrophy and pathological changes and compared the distributions of etiology-specific model probability scores between the groups (Fig. S4). We found that the presence of infarcts, widely recognized as a diagnostic marker for VD,[35] was associated with an increase in model probabilities of VD. The presence of infarcts was also associated with elevated model probabilities for TBI. This association can be attributed to common secondary injuries in TBI, such as cranial hypoperfusion, subsequent ischemia, post-traumatic cerebral infarctions, and encephalomalacia.[36–40] Furthermore, atrophy in the anterior cingulate gyrus was linked with higher model probabilities for FTD, supporting its recognized predictive value for behavioral variant FTD.[41–43] Overall, these observations suggest the model's ability to mirror expert radiological evaluations, particularly in recognizing key brain changes indicative of different dementia types.

Discussion

The key contributions of our study are highlighted as follows. By leveraging a transformer architecture as the backbone and using principles of self-supervised learning, we developed an algorithmic framework that can process flexible combinations of multimodal data and perform differential diagnosis of dementia. This approach allows it to effectively navigate the spectrum of cognitive states, assigning probability scores to normal cognition (NC), mild cognitive impairment (MCI), and dementia (DE), as well as to the specific etiologies underlying DE. Unlike our previous work,[15,23] a unique capability of our model is its adeptness in quantifying co-existing dementia conditions within individuals. The model's robustness was further established through its training and validation across a diverse set of independent cohorts, demonstrating its broad applicability. Additionally, our model can identify key features associated with various cognitive states and dementia etiologies, providing insights that align with established pathologies. In a comparative evaluation on a randomly selected case subset, our model's proficiency in distinguishing true positive and true negative cases notably surpassed that of neurologist-level assessments. These results underscore our model's significant potential in enhancing the precision and efficacy of diagnosing dementia-related disorders.

Our modeling framework encompasses a diverse set of elements, each with unique strengths. For instance, the backbone transformer architecture, serving as the foundation, excels at learning long-range dependencies in data, a crucial asset for tasks like medical diagnosis and prediction. Furthermore, our model integrates feature-specific embedding modules, facilitating the absorption of richer representations from various data types, including nominal, ordinal, numerical features, and complex imaging data. Notably, our framework's capability to handle diverse MRI sequences is an advantage. SwinUNETR, a fusion of U-shaped network design with Swin transformer encoder and CNN-based decoder connected through skip connections, efficiently generates image embeddings. These embeddings, when integrated into our backbone transformer, empower the processing of multimodal data for differential dementia diagnosis. Moreover, the backbone transformer framework has the potential for extension to incorporate text data from electronic health records and emerging digital data types, such as voice recordings as well as sensor data from wearable technologies.

Accommodating the intricacies of dementia, our model assigns distinct probabilities to each potential diagnosis. We note this as a key contribution in our work because these values present a nuanced framework for clinicians to create a rank-ordered list of plausible etiologies. In essence, the model processes available multimodal data and outputs probabilities that outline the person's primary condition and account for the multifactorial and often overlapping nature of dementias. The visualizations that we generated were instrumental in quantifying the variability and distribution of the model's predictions, facilitating an understanding of the likelihood of various diagnostic intersections. Effectively, our model acknowledges and characterizes the complexity of mixed dementias, a clinical scenario frequently encountered in practice, and often confirmed via postmortem evidence.[44–46] Such capability fosters a detailed understanding of the person's condition, but also operationalizes the process of differential diagnosis and could facilitate clinical uptake of software-based assistive tools.

The utility of our modeling framework is founded on its robust processing of diverse input types and its adept handling of incomplete datasets, properties that are essential for clinicians requiring immediate and accurate diagnostic information in environments with variable data availability. For example, when a general practitioner records clinical observations and cognitive test results for an elderly person with possible cognitive decline, our model can calculate a probability score indicative of MCI or DE. This function facili-

tates early medical intervention and more informed decisions regarding specialist referrals. At a specialized memory clinic, the addition of extensive neuroimaging data and in-depth neuropsychological battery to the model may increase the precision of the diagnosis, which, in turn, enhances the formulation of individual management strategies with a revised probability score. Such capacity to tailor its output to the scope of input data exemplifies our modeling framework's role in different healthcare settings, including those where swift and resource-efficient diagnosis is paramount. The generation of specific, quantifiable probability scores by the model augments its utility, establishing it as a useful component in the healthcare delivery process. Displaying diagnostic accuracy using varied training data — ranging from demographic information to clinical signs, neuroimaging findings, and neurological test results — the model's versatility facilitates its adaptation to varied clinical operations without necessitating a fundamental overhaul of existing workflows. Consequently, our model fosters a seamless transition across the different levels of dementia care, enabling general practitioners to perform preliminary cognitive screenings and specialists to conduct thorough examinations. Its inclusive functionality assures an accessible and comprehensive tool ensuring fail-safe operation in early detection, continuous monitoring, and the fine-tuning of differential diagnoses, thereby elevating the standard of dementia care.

Shapley value analysis served as an indispensable interpretive mechanism in our study, elucidating the specific variables' influence on our deep learning model's predictions and bridging computational forecasts with clinical implications. By quantifying the influence of individual factors on the outcomes for NC, MCI, and DE as well as for the dementia etiologies, Shapley values not only bolster the transparency of our models but also reinforce the validity of our approach by aligning with established medical evidence. Such corroboration with recognized diagnostic standards was crucial for embedding machine learning into healthcare, ensuring trust in its predictive capabilities and foundational logic. Additionally, through model-specific associations with postmortem data, our study robustly validated the alignment of our model with dementia-related neurodegeneration. In essence, our model's capacity to link probability predictions with semi-quantitative postmortem scores paves the way for the integration of deep learning methodologies with well-established clinical evidence.

While our study has the potential to advance the field of differential dementia diagnosis, it does have certain limitations that warrant consideration. Our model was developed and validated on multiple cohorts from numerous studies, and its full generalizability across diverse populations and clinical settings remains to be determined. Moving forward, we see potential in evaluating the model's efficacy across the care continuum, encompassing primary care facilities, geriatric and general neurology practices, family medicine, and specialized clinics in tertiary medical centers. The datasets used in our study predominantly feature AD cases, which could potentially introduce a bias towards better recognition of this particular dementia subtype. Although we incorporated various dementia etiologies, the imbalanced representation might affect the model's generalizability and sensitivity towards less frequent types. Also, we chose to amalgamate mild, moderate, and severe dementia cases into a single category. We acknowledge that this categorization method might not completely reflect the nuanced individual staging practiced in specific healthcare settings, where varying degrees of dementia severity carry distinct implications for treatment and management strategies. Our focus was primarily on differential diagnosis rather than disease staging, which motivated this decision. Future enhancements to our model could potentially include disease staging as an additional dimension, thereby augmenting its granularity and relevance.

The evidence collected from this study signals a convergence between advanced computational methods and the nuanced task of differential diagnosis in dementia, crucial for scenarios with scarce resources and

the multifaceted realm of mixed dementia, a condition frequently encountered yet diagnostically complex. Our adaptable model efficiently integrates multimodal data, showing strong performance across diverse settings. Future validations, encompassing a wider demographic and geographical expanse, will be pivotal to substantiate the model's robustness and enhance its diagnostic utility in dementia care. Our pragmatic investigation accentuates the potential of neural networks to refine the granularity of diagnostic evaluations in neurocognitive disorders.

Methods

Study population We collected demographics, personal and family history, laboratory results, findings from the physical/neurological exams, medications, neuropsychological tests, and functional assessments as well as multi-sequence magnetic resonance imaging (MRI) scans from 9 distinct cohorts, totaling $50,686$ participants. There were $19,462$ participants with normal cognition (NC), $9,209$ participants with mild cognitive impairment (MCI), and $22,015$ participants with dementia (DE). We further identified 10 primary and contributing causes of dementia: $17,298$ participants with Alzheimer's disease (AD), $2,003$ participants with dementia with Lewy bodies and Parkinson's disease dementia (LBD), $2,032$ participants with vascular brain injury or vascular dementia including stroke (VD), 114 participants with Prion disease including Creutzfeldt-Jakob disease (PRD), $3,076$ participants with frontotemporal lobar degeneration and its variants, which includes corticobasal degeneration (CBD) and progressive supranuclear palsy (PSP), and with or without amyotrophic lateral sclerosis (FTD), 138 participants with normal pressure hydrocephalus (NPH), 808 participants suffering from dementia due to infections, metabolic disorders, substance abuse including alcohol, medications, delirium and systemic disease - a category termed as systemic and external factors (SEF), $2,700$ participants suffering from psychiatric diseases including schizophrenia, depression, bipolar disorder, anxiety, and post-traumatic stress disorder (PSY), 265 participants with dementia due to traumatic brain injury (TBI), and $1,234$ participants with dementia due to other causes which include neoplasms, multiple systems atrophy, essential tremor, Huntington's disease, and seizures (ODE).

The cohorts include the National Alzheimer's Coordinating Center (NACC) dataset ($n = 45,349$),[47] the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset ($n = 1,821$),[48] the frontotemporal lobar degeneration neuroimaging initiative (NIFD) dataset ($n = 253$),[49] the Parkinson's Progression Marker Initiative (PPMI) dataset ($n = 198$),[50] the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL) dataset ($n = 661$),[51] the Open Access Series of Imaging Studies-3 (OASIS) dataset ($n = 491$),[52] the 4 Repeat Tauopathy Neuroimaging Initiative (4RTNI) dataset ($n = 80$),[53] and three in-house datasets maintained by the Lewy Body Dementia Center for Excellence at Stanford University (LBDSU) ($n = 182$),[54] and the Framingham Heart Study (FHS) ($n = 1,651$).[55] Since its inception in 1948, FHS has been dedicated to identifying factors contributing to cardiovascular disease, monitoring multiple generations from Framingham, Massachusetts. Over time, the study has pinpointed major cardiovascular disease risk factors and explored their effects, while also investigating risk factors for conditions like dementia and analyzing the relationship between physical traits and genetics. Additional details on the study population are presented in Tables 1 & S1.

Inclusion and exclusion criterion Individuals from each cohort were eligible for study inclusion if they were diagnosed with normal cognition (NC), mild cognitive impairment (MCI), or dementia (DE). We used the National Alzheimer's Coordinating Center (NACC) dataset,[47] which is based on the Uniform Data Set (UDS) 3.0 dictionary,[56] as the baseline for our study. To ensure data consistency, we organized the data from the other cohorts according to the UDS dictionary. For individuals from the NACC cohort who had multiple clinical visits, we initially prioritized the visits at which the person received the diagnostic label of dementia. We then selected the visit with the most data features available prioritizing the availability of neuroimaging information. If multiple visits met all the above criteria, we chose the most recent visit among them. This approach maximized the sample sizes of dementia cases, as well as ensured that each individual had the latest record included in the study while maximizing the utilization of available neuroimaging and non-imaging data. We included participants from the 4RTNI dataset[53] with frontotemporal lobar degenera-

tion (FTD)-related disorders like progressive supranuclear palsy (PSP) or corticobasal syndrome (CBS). For other cohorts (NIFD,[49] PPMI,[50] LBDSU,[54] AIBL,[51] ADNI,[48] and OASIS[52]), participants were included if they had at least one MRI scan within 6 months of an officially documented diagnosis. From the FHS,[55] we utilized data from the Original Cohort (Gen 1) enrolled in 1948, and the Offspring Cohort (Gen 2) enrolled in 1971. For these participants, we selected available data including demographics, history, clinical exam scores, neuropsychological test scores, and MRI within 6 months of the date of diagnosis. We did not exclude cases based on the absence of features (including imaging) or diagnostic labels. Instead, we employed our innovative model training approach to address missing features or labels (See below).

Data processing and training strategy Various non-imaging features (n=391) corresponding to subject demographics, medical history, laboratory results, medications, neuropsychological tests, and functional assessments were included in our study. We combined data from 4RTNI, AIBL, LBDSU, NACC, NIFD, OASIS, and PPMI to train the model. We used a portion of the NACC dataset for internal testing, while the ADNI, and FHS cohorts served for external validation (Tables 1, S1, S2, & S3). We used a series of steps such as standardizing the data across all cohorts and formatting the features into numerical or categorical variables before using them for model training. We used stratified sampling at the person-level to create the training, validation, and testing splits. As we pooled the data from multiple cohorts, we encountered challenges related to missing features and labels. To address these issues and enhance the robustness of our model against data unavailability, we incorporated several strategies such as random feature masking and masking of missing labels (see below).

Our investigation harnessed the potential of multi-sequence magnetic resonance imaging (MRI) volumetric scans sourced from diverse cohorts (Table S3). The majority of these scans encompassed a range of sequences, including T1-weighted, T2-weighted, diffusion-weighted imaging (DWI), susceptibility-weighted imaging (SWI), and fluid attenuated inversion recovery (FLAIR). The collected imaging data were stored in the NIFTI file format, categorized by participant and the date of their visit. The MRI scans underwent a singular pre-processing step, which involved skull stripping using SynthStrip,[57] a computational tool designed for extracting brain voxels from various image types. No registration procedures were applied to the resulting scans. To ensure the purity of the dataset, we excluded calibration, localizer, and 2D scans from the downloaded data before initiating model training.

Backbone architecture Our modeling framework harnesses the power of the transformer architecture to interpret and process a vast array of diagnostic parameters, including person-level demographics, medical history, neuroimaging, functional assessments, and neuropsychological test scores. Each of these distinct features is initially transformed into a fixed-length vector using a modality-specific strategy, forming the initial layer of input for the transformer model. Following this, the transformer acts to aggregate these vector inputs, decoding them into a series of predictions. A distinguishing strength of this framework lies in its integration of the transformer's masking mechanism,[58,59] strategically deployed to emulate missing features. This capability enhances the model's robustness and predictive power, allowing it to adeptly handle real-world scenarios characterized by incomplete data.

Multimodal data embeddings Transformers use a uniform representation for all input tokens, typically in the form of fixed-length vectors. However, the inherent complexity of medical data, with its variety

of modalities, poses a challenge to this requirement. Therefore, medical data needs to be adapted into a unified embedding that our transformer model can process. The data we accessed falls into three primary categories: numerical data, categorical data, and imaging data. Each category requires a specific method of embedding. Numerical data typically encompasses those data types where values are defined in an ordinal manner that holds distinct real-world implications. For instance, chronological age fits into this category, as it serves as an indicator of the aging process. To project numerical data into the input space of the transformer, we employed a single linear layer to ensure an appropriate preservation of the structure inherent to the original data space. Categorical data encompasses those inputs that can be divided into distinct categories yet lack any implicit order or priority. An example of this is gender, which can be categorized as 'male'or 'female'. We utilized a lookup table to translate categorical inputs into corresponding embeddings. It's noteworthy that this approach is akin to a linear transformation when the data is one-hot vectorized, but is computationally efficient, particularly when dealing with a vast number of categories. Imaging data, which includes MRI scans in medical applications, can be seen as a special case of numerical data. However, due to their high dimensionality and complexity, it is difficult to compress raw imaging data into a significantly lower-dimensionality vector using a linear transformation, while still retaining essential information. We leveraged the advanced capabilities of modern deep learning architectures to extract meaningful imaging embeddings (see below). Once these embeddings were generated, they were treated as numerical data, undergoing linear projection into vectors of suitable length, thus enabling their integration with other inputs to the transformer.

Imaging feature extraction We harnessed the Swin UNETR (Fig. S1),[60,61] a three-dimensional (3D) transformer-based architecture, to extract embeddings from a multitude of brain MRI scans, encompassing various sequences including T1-weighted (T1w), T2-weighted (T2w), diffusion-weighted (DWI), susceptibility-weighted (SWI), and fluid-attenuated inversion recovery (FLAIR) imaging sequences. The Swin UNETR model consists of a Swin Transformer encoder, designed to operate on 3D patches, seamlessly connected to a convolutional neural network (CNN)-based decoder through multi-resolution skip connections. Commencing with an input volume $X \in \mathbb{R}^{H \times W \times D}$, the encoder segmented X into a sequence of 3D tokens 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. 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 dimension of $C = 48$. The Swin UNETR encoder was subsequently interconnected with a CNN-based decoder at various resolutions through skip connections, collectively forming a 'U-shaped' network. This decoder amalgamated the encoder's outputs at different resolutions, conducted upsampling via deconvolutions, ultimately generating a reconstruction of the initial input volume. The pre-trained weights were the product of self-supervised pre-training of the Swin UNETR encoder, primarily conducted on 3D volumes encompassing the chest, abdomen, and head/neck.[60,61]

The process of obtaining imaging embeddings began with several transformations applied to the MRI scans. These transformations included resampling the scans to standardized pixel dimensions, foreground cropping, and spatial resizing, resulting in the creation of sub-volumes with dimensions of $128 \times 128 \times 128$. Subsequently, these sub-volumes were input into the Swin UNETR model, which in turn extracted encoder outputs sized at $768 \times 4 \times 4 \times 4$. These extracted embeddings underwent downsampling via a learnable embedding module, consisting of four convolutional blocks, to align with the input token size of the downstream transformer. As a result, the MRI scans were effectively embedded into one-dimensional vectors, each of size 128. These vectors were then combined with non-imaging features and directed into the downstream transformer for further processing. The entire process utilized a dataset comprising $11,438$ MRI volumes, which were allocated for model training, validation, and testing (Table S3).

Random feature masking To enhance the robustness of the backbone transformer in handling data incompleteness, we leveraged the masking mechanism [58,59] to emulate arbitrary missing features during training. The masking mechanism, when paired with the attention mechanism, effectively halts the information flow from a given set for input tokens, ensuring that certain features are concealed during prediction. A practical challenge arises when considering the potential combinations of input features, which increase exponentially. With hundreds of features in play, capturing every potential combination is intractable. Inspired by the definition of Shapley values, we deployed an efficient strategy for feature dropout. Given a sample with feature set S, S is randomly permuted as σ; simultaneously, an integer i is selected independently from the range $[1, |S|]$. Subsequent to this, the features $\sigma_{i+1}, \sigma_{i+2}, \ldots, \sigma_{|S|}$ are masked out from the backbone transformer. It's noteworthy that the dropout process was applied afresh across different training batches or epochs to ensure that the model gets exposed to a diverse array of missing information even within a single sample.

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

Loss function Our model was trained by minimizing the loss function ($\mathcal{L}$) composed of two loss terms: "Focal Loss (FL)" [62] ($\mathcal{L}_{\text{FL}}$) and "Ranking Loss (RL)" ($\mathcal{L}_{\text{RL}}$), along with the standard L2 regularization term. FL is a variant of standard cross-entropy loss that addresses the issue of class imbalance. It assigns low weight to easy (well-classified) instances and employs a balance parameter. This loss function was used for each of the diagnostic categories (a total of 13, see Glossary 1). Therefore, our $\mathcal{L}_{\text{FL}}$ term was:

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

where N was the batch size (i.e., $N = 64$), and other parameters and variables were as defined. The focusing parameter γ was set to 2, which had been reported to work well in most of the experiments in the original paper.[62] Moreover, $\alpha_i \in [0, 1]$ was the balancing parameter that influenced the weights of positive and negative instances. It was was set as the square of the complement of the fraction of samples labeled as 1, varying for each i due to the differing level of class imbalance across diagnostic categories (refer to Table 1). The FL term did not take inter-class relationships into account. To address these relationships in our overall loss function, we also incorporated the RL term that induced loss if the sigmoid outputs for diagnostic categories labeled as 0 were not lower than those labeled as 1 by a predefined margin of ϵ, for any training sample k. We defined the RL term for any pair of diagnostic categories i and j, as follows:

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

Overall, the RL term was:

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

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

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

where λ and β were the weights that controlled the importance of $\mathcal{L}_{\text{RL}}$ and the L2 regularization terms, respectively. The training was done using the mini-batch strategy with the AdamW optimizer,[63] an improved version of the Adam optimizer,[64] with a learning rate of 0.0005 for a total of 256 epochs. Additionally, we utilized a cosine learning rate scheduler with warm restarts,[65] initiating the first restart after 64 batches and extending the restart period by a factor of 2 for each subsequent restart. The values of ϵ, λ, and β were determined to be $\epsilon = 0.25$, $\lambda = 0.025$, and $\beta = 0.01$, respectively, based on an evaluation of the overall model performance on the validation set. During training, the model performance was evaluated on the validation set at the end of each epoch, and the model with the highest performance was selected.

Interpretability analysis The primary goal of interpretability analysis is to demystify machine learning models by providing clear insights into how various features influence predictions. At the heart of this field lies the Shapley value,[31] originally a game theory concept, now repurposed to evaluate feature significance in machine learning models. In this context, each instance is considered a unique 'game', where features act as players contributing to the outcome. The model's output is analogous to the game's payoff, with the Shapley value quantifying each feature's contribution towards this outcome. Calculating exact Shapley values necessitates evaluating the model with every possible combination of missing features. Given the extensive size of our input features, this process becomes computationally prohibitive. To address this challenge, we employed the permutation sampling method for Shapley value estimation.[66] This approach involves randomly sampling permutations of the features and approximating the contribution of each feature to the prediction by averaging its marginal contributions across these permutations. It significantly reduces computational load while still providing a reasonable estimate of the Shapley values.

Head-to-head expert validation We evaluated our model's predictive power against the diagnostic acumen of clinicians who are directly involved in dementia diagnosis and care. A group of neurologists and neuroradiologists were invited to participate in diagnostic tasks using a select subset of NACC cases (see 'Data processing and training strategy'). Neurologists were presented with 100 cases – 15 cases each of NC and MCI, and 7 cases each of the dementia subtypes. The data encompassed person-level demographics, medical history, social history, neuropsychological tests, functional assessments, and multi-sequence MRI scans where possible (i.e., T1-weighted, T2-weighted, FLAIR, DWI and SWI sequences). They were asked to provide their diagnostic impressions. Similarly, neuroradiologists were provided with the same multi-sequence MRI scans, along with details on age, gender, race, and education status from 70 clinically diagnosed DE cases. They were tasked with providing diagnostic impressions concerning the origin of dementia (Refer to Glossary 1). Additionally, neuroradiologists completed a REDCap questionnaire to rate atrophy and pathological changes in each brain sub-region on a scale from 0 to 3, with higher scores indicating more severe degeneration, and the presence or absence of specific disease markers such as infarcts. Case samples and questionnaires supplied to the neurologists and neuroradiologists can be found in the Supplementary Information.

Neuropathologic validation The model's predictive capacity for various dementia etiologies was substantiated through alignment with neuropathological evaluations sourced from the NACC, FHS and ADNI cohorts (Table S4). We included participants who conformed to the study's inclusion criteria, had undergone MRI scans no more than three years prior to death, and for whom neuropathological data were available. Standardization of data was conducted in accordance with the Neuropathology Data Form Version 10 protocols from the National Institute on Aging.[67] We pinpointed neuropathological indicators that influence the pathological signature of each dementia subtype, such as arteriolosclerosis, the presence of neurofibrillary tangles and amyloid plaques, cerebral amyloid angiopathy, and markers of tauopathy. These indicators were carefully chosen to reflect the complex pathological terrain that defines each form of dementia. To examine the Thal phase for amyloid plaques (A score), subjects were categorized into two groups: one encompassing Phase 0, indicative of no amyloid plaque presence, and a composite group merging Phases 1-5, reflecting varying degrees of amyloid pathology. The model's predictive performance was then compared across these groupings. For the Braak stage of neurofibrillary degeneration (B score), we consolidated stages I-VI into a single collective, representing the presence of AD-type neurofibrillary pathology, whereas stage 0 was designated for cases devoid of AD-type neurofibrillary degeneration. With respect to the density of neocortical neuritic plaques, assessed by the (CERAD or C score), individuals without neuritic plaques constituted one group, while those with any manifestation of neuritic plaques—sparse, moderate, or frequent (C1-C3)—were aggregated into a separate group for comparative analysis of the model's predictive outcomes. The Thal, Braak, and CERAD scores were integrated into a composite ABC score, delineating the National Institute on Aging-Alzheimer's Association (NIA-AA) criteria for AD neuropathological change (ADNC).[68] This summation resulted in two distinct groups for analysis: one encapsulating cases with no neuropathological evidence of Alzheimer's disease, and another amalgamating cases classified across the spectrum of low, intermediate, and high ADNC. Furthermore, to evaluate the model's concordance with non-AD pathologies, we analyzed the association between the model-generated probabilities and the presence or absence of TDP-43 pathology, old microinfarcts, arteriosclerosis, and Prion disease.

Statistical analysis We used one-way ANOVA and the χ^2 test for continuous and categorical variables, respectively to assess the overall differences in the population characteristics between the diagnostic groups across the study cohorts. We applied the Kruskal-Wallis H-test for independent samples and subsequently conducted post-hoc Dunn's testing with Bonferroni correction to evaluate the relationship between clinical dementia rating scores and the model-predicted probabilities, as well as the relationship between neuropathologic scores and the model-predicted probabilities. We opted for non-parametric tests because the Shapiro-Wilk test indicated significant deviations from normality. For comparing model predictions with expert-driven assessments, we used the Brunner Munzel test to identify statistically significant increases in the mean disease probability scores between the levels of scoring categories. We conducted a Shapiro-Wilk test on the distributions of the true negative and true positive cases for each etiology. The Brunner-Munzel test was then used to compare the expert and model confidence scores for the true negative and true positive cases for each etiology. To evaluate the interrater reliability of label-specific confidence scores, we performed pairwise Pearson correlation analyses between clinicians' scores and those generated by the model.[69] We calculated the average correlation coefficient across pairs and determined its 95% confidence interval. In addition, we assessed the correlation between the aggregated confidence score of neurologists and the model's score for each diagnostic label. Using a bootstrapping approach with resampling, we created $1,000$ iterations of the consensus score from the pool of individual neurologist scores. We then calculated the Pearson correlation for each iteration against the model's scores, from which we derived the average correlation coefficient and its 95% confidence interval. All statistical analyses were conducted at a significance level of 0.05.

Performance metrics We generated receiver operating characteristic (ROC) and precision-recall (PR) curves from predictions on both the NACC test data and other datasets. From each ROC and PR curve, we further derived the area under the curve values. Also, we evaluated the model's accuracy, sensitivity, specificity, and Matthews correlation coefficient, with the latter being a balanced measure of quality for classes of varying sizes in a binary classifier. We also gauged inter-expert consensus using Cohen's kappa (κ), which measures the degree to which two experts agree on a diagnosis. For each subgroup task, we computed the average pairwise κ as a comprehensive measure of agreement between the expert clinicians.

Computational hardware and software All MRI and non-imaging data were processed on a workstation equipped with an Intel i9 14-core 3.3 GHz processor and 4 NVIDIA RTX 2080Ti GPUs. Our software development utilized Python (version 3.7.7) and the models were developed using PyTorch (version 1.5.1). We used several other Python libraries to support data analysis, including pandas (version 1.0.3), scipy (version 1.3.1), tensorboardX (version 1.9), torchvision (version 0.6), and scikit-learn (version 0.22.1). Training the model on a single Quadro RTX8000 GPU on a shared computing cluster had an average runtime of 7 minutes per epoch, while the inference task took less than a minute per instance. All clinicians reviewed MRIs using 3D Slicer (version 4.10.2) and logged their findings in REDCap (version 11.1.3).

Data and code availability Data from ADNI, AIBL, NACC, NIFD, OASIS, PPMI and 4RTNI can be downloaded from publicly available resources. Data from FHS and LBDSU can be obtained upon request, subject to institutional approval. The Python scripts used in this study can be found on the Kolachalama Laboratory's GitHub page (`https://github.com/vkola-lab`).

Acknowledgements

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

Contributions

C.X. and S.S.K. contributed equally to this work. S.S.K., D.L., S.P., O.T.Z., A.S.W., A.K., C.K., and T.F.A.A. performed data collection. C.X. and S.S.K. designed and developed the machine learning framework. C.X., S.S.K., D.L., and O.B.G. performed model training and validation. S.S.K., S.P. and M.A. performed statistical analysis. C.X., S.S.K., D.L., S.P., O.T.Z., A.S.W., 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., 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 practicing clinicians who reviewed the cases. S.A.B. and B.A.P. provided guidance on the modeling framework. K.L.P. and R.A. provided access to data. V.B.K. wrote the manuscript. All authors reviewed and approved the manuscript. V.B.K. conceived, designed and directed the study.

Ethics declarations

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

References

1. Cahill, S. Who's global action plan on the public health response to dementia: some challenges and opportunities. *Aging & Mental Health* **24**, 197–199 (2019).

2. Dall, T. M. *et al.* Supply and demand analysis of the current and future us neurology workforce. *Neurology* **81**, 470–478 (2013).

3. Burton, A. How do we fix the shortage of neurologists? *The Lancet Neurology* **17**, 502–503 (2018).

4. Lester, P. E., Dharmarajan, T. S. & Weinstein, E. The looming geriatrician shortage: Ramifications and solutions. *J Aging Health* **32**, 1052–1062 (2020). Epub 2019 Oct 4.

5. Knopman, D. S. *et al.* Practice parameter: Diagnosis of dementia (an evidence-based review). *Neurology* **56**, 1143–1153 (2001).

6. Thijssen, E. H. & Rabinovici, G. D. Rapid progress toward reliable blood tests for alzheimer disease. *JAMA Neurology* **78**, 143–145 (2021).

7. Teunissen, C. E. *et al.* Blood-based biomarkers for alzheimer's disease: towards clinical implementation. *Lancet Neurology* **21**, 66–77 (2022). URL `https://doi.org/10.1016/S1474-4422(21)00361-6`.

8. Sevigny, J. *et al.* The antibody aducanumab reduces abeta plaques in alzheimer's disease. *Nature* **537**, 50–56 (2016).

9. van Dyck, C. H. *et al.* Lecanemab in early alzheimer's disease. *New England Journal of Medicine* **388**, 9–21 (2023).

10. Martin, S. A., Townend, F. J., Barkhof, F. & Cole, J. H. Interpretable machine learning for dementia: A systematic review. *Alzheimer's & Dementia* **19**, 2135–2149 (2023).

11. Myszczynska, M. A. *et al.* Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology* **16**, 440–456 (2020).

12. Borchert, R. J. *et al.* Artificial intelligence for diagnostic and prognostic neuroimaging in dementia: A systematic review. *Alzheimer's & Dementia* (2023). Online ahead of print.

13. Ahmed, M. R., Mahmood, A. N., Huq, M. A., Funk, P. & Mafi, A. Neuroimaging and machine learning for dementia diagnosis: Recent advancements and future prospects. *IEEE Reviews in Biomedical Engineering* **12**, 19–33 (2019).

14. Bron, E. E. *et al.* Ten years of image analysis and machine learning competitions in dementia. *NeuroImage* **253** (2022).

15. Qiu, S. *et al.* Development and validation of an interpretable deep learning framework for alzheimer's disease classification. *Brain* **143**, 1920–1933 (2020).

16. Nemoto, K. *et al.* Differentiating dementia with lewy bodies and alzheimer's disease by deep learning to structural mri. *Journal of Neuroimaging* **31**, 579–587 (2021).

17. Zheng, Y., Zhang, Y., Zhang, Y., Wang, Y. & Zheng, B. Machine learning-based framework for differential diagnosis between vascular dementia and alzheimer's disease using structural mri features. *Frontiers in Neurology* **10** (2019).

18. Castellazzi, G. *et al.* A machine learning approach for the differential diagnosis of alzheimer and vascular dementia fed by mri selected features. *Frontiers in Neuroinformatics* **14** (2020).

19. Hu, J. *et al.* Deep learning-based classification and voxel-based visualization of frontotemporal dementia and alzheimer's disease. *Frontiers in Neuroscience* **14** (2021).

20. Kim, J. *et al.* Machine learning based hierarchical classification of frontotemporal dementia and alzheimer's disease. *NeuroImage: Clinical* **23** (2019).

21. Nguyen, H.-D. *et al.* Multimodal deep learning for alzheimer's disease dementia assessment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022*, 55–65 (2022).

22. Moguilner, S. *et al.* Visual deep learning of unprocessed neuroimaging characterises dementia subtypes and generalises across non-stereotypic samples. *EBioMedicine* **90**, 104540 (2023).

23. Qiu, S., Miller, M., Joshi, P. *et al.* Multimodal deep learning for alzheimer's disease dementia assessment. *Nature Communications* **13**, 3404 (2022).

24. Vemuri, P. *et al.* Antemortem differential diagnosis of dementia pathology using structural mri: Differential-stand. *NeuroImage* **55**, 522–531 (2011).

25. Chagué, P. *et al.* Radiological classification of dementia from anatomical mri assisted by machine learning-derived maps. *Journal of Neuroradiology* **48**, 412–418 (2021).

26. Burgos, N. *et al.* Machine learning for classification and prediction of brain diseases: recent advances and upcoming challenges. *Current Opinion in Neurology* **33**, 439–450 (2020).

27. Barnes, D. E. *et al.* Development and validation of a brief dementia screening indicator for primary care. *Alzheimers Dement* **10**, 656–665.e1 (2014). Author manuscript; available in PMC 2015 Nov 1. Published in final edited form as: Alzheimers Dement. 2014 Nov; 10(6): 656–665.e1. Published online 2014 Feb 1.

28. Anstey, K. J. *et al.* A self-report risk index to predict occurrence of dementia in three independent cohorts of older adults: the anu-adri. *PLoS One* **9**, e86141 (2014). ECollection 2014.

29. Sindi, S. *et al.* The caide dementia risk score app: The development of an evidence-based mobile application to predict the risk of dementia. *Alzheimers Dement (Amst)* **1**, 328–333 (2015). ECollection 2015 Sep.

30. Kivimäki, M. *et al.* Estimating dementia risk using multifactorial prediction models. *JAMA Netw Open* **6**, e2318132 (2023).

31. Shapley, L. S. A value for n-person games. *Contributions to the Theory of Games* **2**, 307–317 (1953).

32. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics* **20**, 1983–1992 (2014).

33. Jo, M. *et al.* The role of tdp-43 propagation in neurodegenerative diseases: integrating insights from clinical and experimental studies. *Experimental Molecular Medicine* **52**, 1652–1662 (2020). Epub 2020 Oct 13.

34. Cairns, N. J. *et al.* Tdp-43 in familial and sporadic frontotemporal lobar degeneration with ubiquitin inclusions. *The American Journal of Pathology* **171**, 227–240 (2007).

35. O'Brien, J. T. & Thomas, A. Vascular dementia. *The Lancet* **386**, 1698–1706 (2015). URL https://www.sciencedirect.com/science/article/pii/S0140673615004638.

36. Le, T. H. & Gean, A. D. Neuroimaging of traumatic brain injury. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine* **76**, 145–162 (2009).

37. Mirvis, S., Wolf, A., Numaguchi, Y., Corradino, G. & Joslyn, J. Posttraumatic cerebral infarction diagnosed by ct: prevalence, origin, and outcome. *American Journal of Neuroradiology* **11**, 355–360 (1990).

38. Tian, H.-L. *et al.* Risk factors for posttraumatic cerebral infarction in patients with moderate or severe head trauma. *Neurosurgical Review* **31**, 431–437 (2008).

39. Haber, M. *et al.* Vascular abnormalities within normal appearing tissue in chronic traumatic brain injury. *Journal of Neurotrauma* **35**, 2250–2258 (2018).

40. Latronico, N. *et al.* Impact of a posttraumatic cerebral infarction on outcome in patients with TBI: The Italian Multicenter cohort INCEPT study. *Crit Care* **24**, 33 (2020). URL https://link.springer.com/article/10.1186/s13054-020-2746-5.

41. Hornberger, M. *et al.* In vivo and post-mortem memory circuit integrity in frontotemporal dementia and alzheimer's disease. *Brain : a journal of neurology* (2012).

42. Rabinovici, G. *et al.* Distinct mri atrophy patterns in autopsy-proven alzheimer's disease and frontotemporal lobar degeneration. *American Journal of Alzheimer's Disease & Other Dementias®* **22**, 474–488 (2008).

43. Chu, M. *et al.* Investigating the roles of anterior cingulate in behavioral variant frontotemporal dementia: A pet/mri study. *Journal of Alzheimer's Disease* **84**, 1771–1779 (2021).

44. Armstrong, R. A., Lantos, P. L. & Cairns, N. J. Overlap between neurodegenerative disorders. *Neuropathology* **25**, 111–124 (2005). 15875904.

45. Schneider, J. A., Arvanitakis, Z., Bang, W. & Bennett, D. A. Mixed brain pathologies account for most dementia cases in community-dwelling older persons. *Neurology* **69**, 2197–2204 (2007).

46. Rahimi, J. & Kovacs, G. G. Prevalence of mixed pathologies in the aging brain. *Alzheimer's Research & Therapy* **6**, 82 (2014).

47. Beekly, D. L. *et al.* The national alzheimer's coordinating center (nacc) database: an alzheimer disease database. *Alzheimer Disease & Associated Disorders* **18**, 270–277 (2004).

48. Mueller, S. G. *et al.* Ways toward an early diagnosis in alzheimer's disease: The alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia* **1**, 55–66 (2005).

49. Boxer, A. L. *et al.* Frontotemporal degeneration, the next therapeutic frontier: Molecules and animal models for frontotemporal degeneration drug development. *Alzheimer's & Dementia* **9**, 176–188 (2013).

50. Marek, K. *et al.* The parkinson progression marker initiative (ppmi). *Progress in Neurobiology* **95**, 629–635 (2011). Biological Markers for Neurodegenerative Diseases.

51. Ellis, K., Ames, D., Martins, R., Hudson, P. & Masters, C. The australian biomarkers lifestyle and imaging flagship study of ageing. *Acta Neuropsychiatrica* **18**, 285–285 (2006).

52. Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C. & Buckner, R. L. Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults. *Journal of Cognitive Neuroscience* **22**, 2677–2684 (2010).

53. Dutt, S. *et al.* Progression of brain atrophy in psp and cbs over 6 months and 1 year. *Neurology* **87**, 2016–2025 (2016).

54. Linortner, P. *et al.* White matter hyperintensities related to parkinson's disease executive function. *Movement Disorders Clinical Practice* **7**, 629–638 (2020).

55. Yang, J. *et al.* Establishing cognitive baseline in three generations: Framingham heart study. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **15** (2023).

56. Beekly, D. L. *et al.* The national alzheimer's coordinating center (nacc) database: the uniform data set. *Alzheimer Disease & Associated Disorders* **21**, 249–258 (2007).

57. Hoopes, A., Mora, J. S., Dalca, A. V., Fischl, B. & Hoffmann, M. Synthstrip: skull-stripping for any brain image. *NeuroImage* **260**, 119474 (2022). URL `https://www.sciencedirect.com/science/article/pii/S1053811922005900`.

58. Vaswani, A. *et al.* Attention is all you need. In *Advances in neural information processing systems*, 5998–6008 (2017).

59. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

60. Hatamizadeh, A. *et al.* Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. arxiv 2022. *arXiv preprint arXiv:2201.01266* .

61. Tang, Y. *et al.* Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20730–20740 (2022).

62. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007 (2017).

63. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (2019). URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

64. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv* (2014). `1412.6980`.

65. Loshchilov, I. & Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations* (2017). URL `https://openreview.net/forum?id=Skq89Scxx`.

66. Mitchell, R., Cooper, J., Frank, E. & Holmes, G. Sampling permutations for shapley value estimation. *Journal of Machine Learning Research* **23**, 1–46 (2022).

67. National alzheimer's coordinating center. neuropathology data form version 10, January 2014. URL `https://naccdata.org/data-collection/forms-documentation/np-10`.

68. Montine, T. J. *et al.* National institute on aging-alzheimer's association guidelines for the neuropathologic assessment of alzheimer's disease: A practical approach. *Acta Neuropathologica* **123**, 1–11 (2012).

69. de Raadt, W. M. B. R. e. a., A. A comparison of reliability coefficients for ordinal rating scales. *Journal of Classification* **38**, 519–543 (2021). URL https://link.springer.com/article/10.1007/s00357-021-09386-5.

Dataset (group)	Age mean ± std	Male gender (percentage)	Education in years mean ± std	Race (White; Black; Asian; American Indian; Pacific; Multi-race)	CDR mean ± std
NACC					
NC [n = 17242]	71.25 ± 11.16	6009, 34.85%	15.83 ± 2.98ˆ	(13266, 2541, 528, 109, 10, 575)ˆ	0.05 ± 0.15
MCI [n = 7582]	73.72 ± 9.81	3615, 47.68%	15.16 ± 3.45ˆ	(5708, 1185, 231, 53, 5, 276)ˆ	0.45 ± 0.18
AD [n = 16131]	76.0 ± 10.31	7234, 44.85%	14.52 ± 3.74ˆ	(13161, 1702, 354, 92, 10, 458)ˆ	1.2 ± 0.73
LBD [n = 1913]	75.01 ± 8.55	1365, 71.35%	15.12 ± 3.63ˆ	(1659, 128, 39, 17, 0, 37)ˆ	1.29 ± 0.78
VD [n = 1919]	80.32 ± 8.76	947, 49.35%	14.15 ± 4.22ˆ	(1394, 332, 67, 2, 1, 68)ˆ	1.22 ± 0.74
PRD [n = 114]	60.07 ± 10.36	62, 54.39%	14.8 ± 3.33ˆ	(93, 5, 5, 0, 1, 1)ˆ	1.95 ± 0.95
FTD [n = 2898]	65.86 ± 9.36	1603, 55.31%	15.45 ± 3.09ˆ	(2664, 69, 73, 4, 5, 39)ˆ	1.2 ± 0.83
NPH [n = 138]	79.1 ± 9.24	69, 50.0%	15.0 ± 3.28ˆ	(119, 10, 4, 0, 0, 4)ˆ	1.18 ± 0.71
SEF [n = 808]	76.3 ± 11.15	413, 51.11%	14.6 ± 3.77ˆ	(646, 95, 15, 5, 2, 31)ˆ	1.11 ± 0.7
PSY [n = 2700]	73.74 ± 10.78	1102, 40.81%	14.13 ± 4.12ˆ	(2163, 238, 59, 14, 5, 87)ˆ	1.1 ± 0.64
TBI [n = 265]	72.87 ± 11.23	192, 72.45%	14.42 ± 4.13ˆ	(212, 27, 3, 2, 1, 11)ˆ	1.11 ± 0.69
ODE [n = 1234]	72.94 ± 12.14	654, 53.0%	14.5 ± 3.78ˆ	(1046, 93, 28, 5, 4, 36)ˆ	1.2 ± 0.76
p-value	<1.0e-200	<1.0e-200	<1.0e-200	8.341e-145	<1.0e-200
NIFD					
NC [n = 124]	63.21 ± 7.27	56, 45.16%	17.48 ± 1.87ˆ	(89, 0, 0, 0, 0, 3)ˆ	0.03 ± 0.12ˆ
FTD [n = 129]	63.66 ± 7.33	75, 58.14%	16.18 ± 3.29ˆ	(109, 1, 1, 0, 0, 4)ˆ	0.82 ± 0.54ˆ
p-value	6.266e-01	5.246e-02	2.606e-04	6.531e-01	4.333e-28
PPMI					
NC [n = 171]	62.74 ± 10.12	109, 63.74%	15.82 ± 2.93	(163, 3, 2, 0, 0, 1)ˆ	N.A.
MCI [n = 27]	68.04 ± 7.32	22, 81.48%	15.52 ± 3.08	(24, 1, 1, 0, 0, 1)	N.A.
p-value	1.006e-02	1.115e-01	6.194e-01	2.910e-01	N.A.
AIBL					
NC [n = 480]	72.45 ± 6.22	203, 42.29%	N.A.	N.A.	0.03 ± 0.12
MCI [n = 102]	74.73 ± 7.11	53, 51.96%	N.A.	N.A.	0.47 ± 0.14
AD [n = 79]	73.34 ± 7.77	33, 41.77%	N.A.	N.A.	0.93 ± 0.54
p-value	5.521e-03	1.887e-01	N.A.	N.A.	4.542e-158
OASIS					
NC [n = 424]	71.34 ± 9.43	164, 38.68%	15.79 ± 2.62ˆ	(53, 18, 1, 0, 0, 0)ˆ	0.0 ± 0.02
MCI [n = 27]	75.04 ± 7.25	14, 51.85%	15.19 ± 2.76	(4, 1, 0, 0, 0, 0)ˆ	0.52 ± 0.09
AD [n = 32]	77.44 ± 7.42	20, 62.5%	15.19 ± 2.8	(8, 1, 0, 0, 0, 0)ˆ	0.86 ± 0.44
LBD [n = 4]	74.75 ± 5.67	4, 100.0%	16.0 ± 2.83	N.A.	1.0 ± 0.0
FTD [n = 4]	64.25 ± 8.61	3, 75.0%	16.5 ± 2.96	(4, 0, 0, 0, 0, 0)	1.25 ± 0.75
p-value	7.789e-04	3.239e-03	5.507e-01	8.735e-01	2.855e-169
LBDSU					
NC [n = 134]	68.77 ± 7.62	61, 45.52%	17.27 ± 2.47ˆ	N.A.	N.A.
MCI [n = 35]	70.16 ± 8.41	26, 74.29%	16.6 ± 2.58	N.A.	N.A.
LBD [n = 13]	73.42 ± 7.81	8, 61.54%	16.77 ± 2.15	N.A.	N.A.
p-value	1.033e-01	7.863e-03	3.243e-01	N.A.	N.A.
4RTNI					
NC [n = 12]	68.08 ± 4.92	5, 41.67%	15.45 ± 2.57ˆ	(12, 0, 0, 0, 0, 0)	0.0 ± 0.0
MCI [n = 31]	67.61 ± 7.0	11, 35.48%	16.68 ± 4.02	(25, 1, 2, 0, 1, 1)ˆ	0.55 ± 0.15
FTD [n = 37]	69.14 ± 7.43	20, 54.05%	16.46 ± 4.21	(31, 1, 0, 0, 1, 2)ˆ	1.27 ± 0.55
p-value	6.691e-01	2.992e-01	6.843e-01	7.620e-01	5.700e-16
ADNI					
NC [n = 481]	74.26 ± 6.0	235, 48.86%	16.34 ± 2.67	(432, 36, 8, 1, 0, 3)ˆ	0.0 ± 0.0ˆ
MCI [n = 971]	72.84 ± 7.71	572, 58.91%	15.94 ± 2.81	(903, 34, 15, 2, 2, 12)ˆ	0.5 ± 0.04
AD [n = 369]	74.91 ± 7.84	203, 55.01%	15.18 ± 2.97	(343, 15, 7, 0, 0, 4)	0.77 ± 0.26
p-value	2.565e-06	1.364e-03	1.872e-08	1.132e-01	<1.0e-200
FHS					*
NC [n = 394]	74.9 ± 10.22ˆ	206, 52.28%	N.A.	(394, 0, 0, 0, 0, 0)	0.0 ± 0.0
MCI [n = 434]	79.92 ± 8.8ˆ	203, 46.77%	N.A.	(434, 0, 0, 0, 0, 0)	0.49 ± 0.07
AD [n = 687]	82.99 ± 7.87ˆ	211, 30.71%	N.A.	(687, 0, 0, 0, 0, 0)	2.04 ± 0.88
LBD [n = 73]	79.34 ± 9.37ˆ	46, 63.01%	N.A.	(73, 0, 0, 0, 0, 0)	1.84 ± 0.84
VD [n = 113]	81.74 ± 7.3ˆ	48, 42.48%	N.A.	(113, 0, 0, 0, 0, 0)	1.85 ± 0.8
FTD [n = 8]	85.67 ± 5.91ˆ	4, 50.0%	N.A.	(8, 0, 0, 0, 0, 0)	2.0 ± 0.87
p-value	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 ˆ 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.

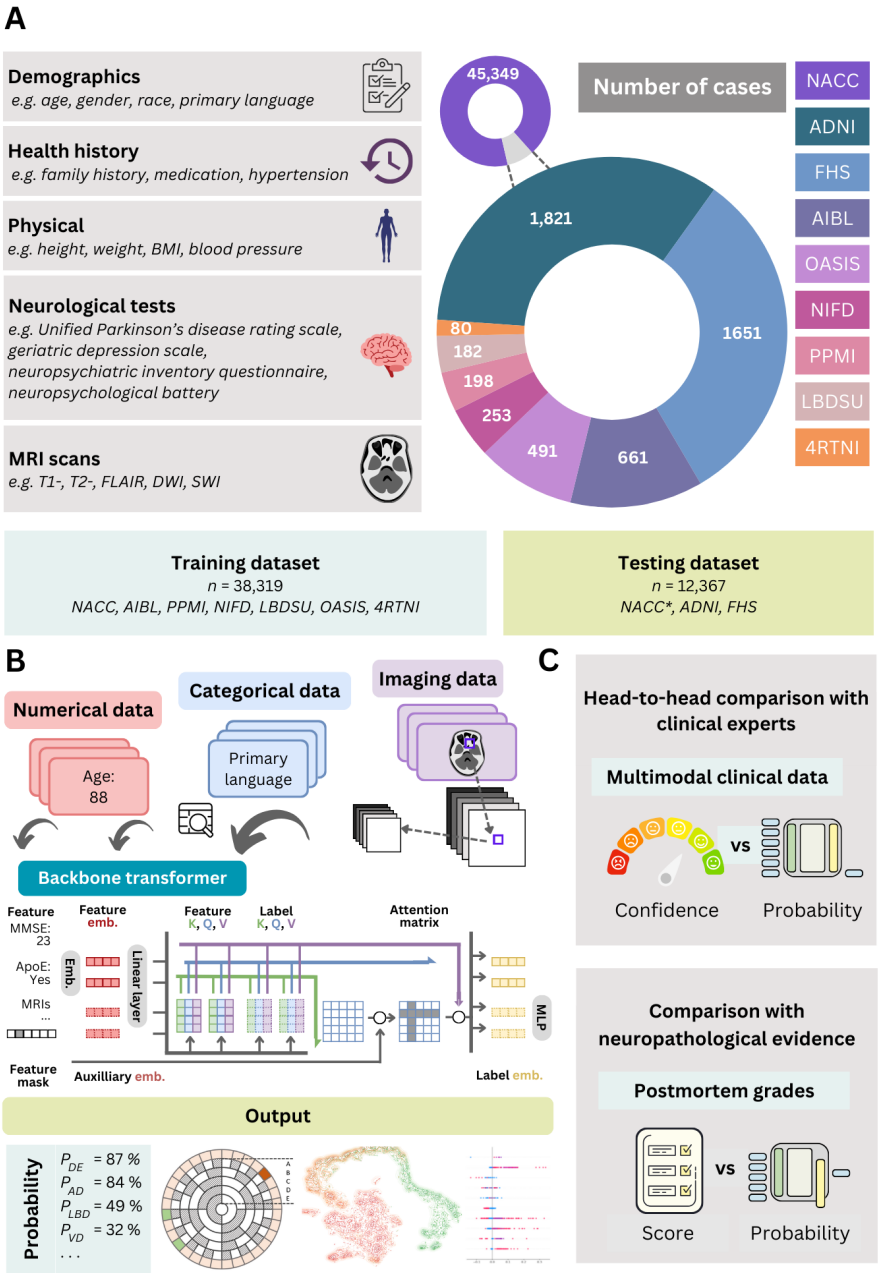

Figure 1: **Data, model architecture and modeling strategy.** (a) Our model for differential dementia diagnosis was developed using diverse data modalities, including individual-level demographics, health history, neurological testing, physical/neurological exams, and multi-sequence MRI scans. These 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 distinct portion of the NACC dataset was randomly selected to enable a comparative analysis of the model's performance against practicing neurologists. Furthermore, we conducted a direct comparison between the model and a team of practicing neuroradiologists using a random sample of cases with confirmed dementia from the NACC testing cohort. 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 pathology grades available from the NACC, ADNI, and FHS cohorts.

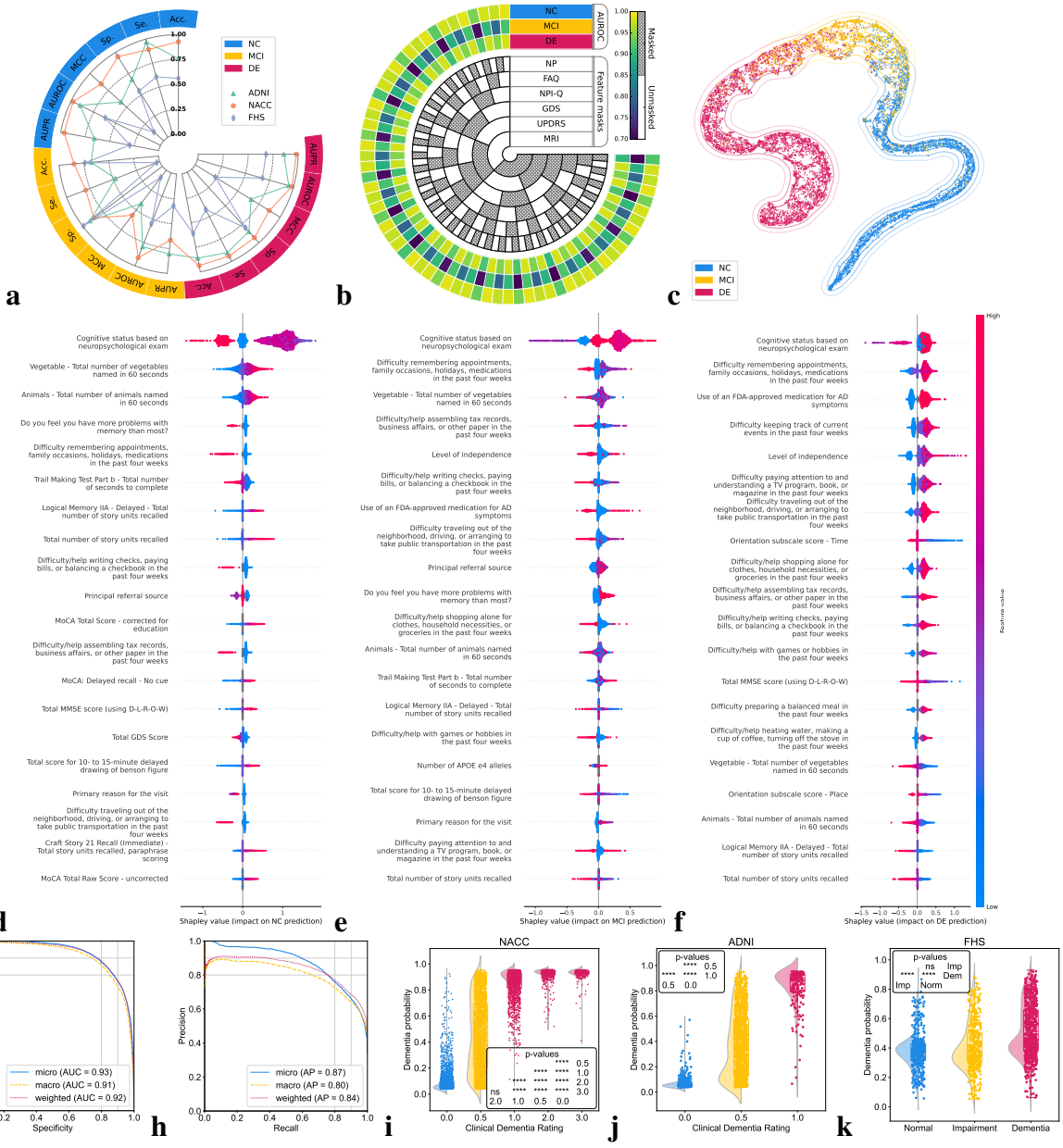

Figure 2: **Model performance on individuals along the cognitive spectrum.** (a) Radar plot illustrating the performance of the model on individuals with normal cognition (NC), mild cognitive impairment (MCI), and dementia (DE) is shown. We present a range of metrics including mean values along with their standard deviations, for model accuracy, sensitivity, specificity, precision, area under the receiver operating characteristic curve, area under the precision-recall curve, F1-score, and Matthews correlation coefficient. (b) 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. (c) Two-dimensional t-distributed stochastic neighbor embeddings obtained from the penultimate layer of the model are shown. The legend in the lower-left corner indicates the color coding representing NC, MCI and DE, respectively. (d, e, f) Beeswarm plots visualize Shapley values for subjects classified as NC, MCI, and DE, respectively. (g, h) The receiver operating characteristic (ROC) and precision-recall (PR) curves are presented, 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. (i, j, k) Raincloud plots with overlying violin and box diagrams are shown to denote the distribution of clinical dementia rating scores (horizontal axis) versus model-predicted probability of dementia (vertical axis), on the NACC, ADNI and FHS cohorts, respectively. For (a,g,h), cases from the NACC testing, ADNI and FHS were used. 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.

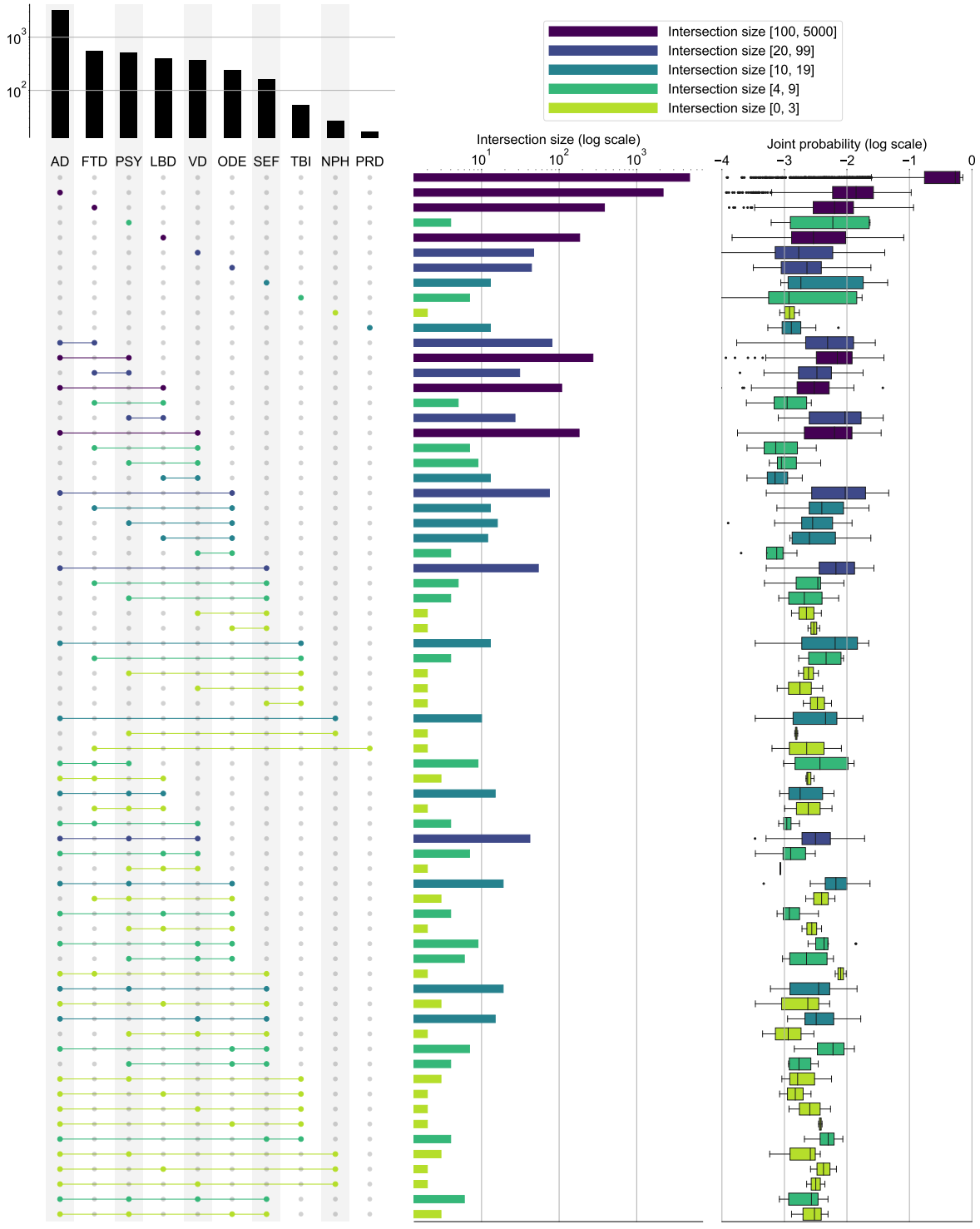

Figure 3: **Model assessment on mixed dementias.** Visualization of the distribution and model-predicted probabilities of various etiological categories found within the NACC testing dataset. The left segment enumerates both single and co-occurring diagnostic categories, offering a tally of each condition's frequency within the dataset. In the center, a logarithmic scale is used to delineate the overlap among these categories, shedding light on their relative commonality and the extent of their coexistence. This method grants a refined perspective on the prevalence of comorbid conditions. Additionally, the legend in the upper right interprets the counts within the central panel, providing a reference for the logarithmic data representation. The panel on the far right features a box-and-whisker plot, delineating the spread and central tendency of the model's predicted probabilities for each combination of diagnostic categories.

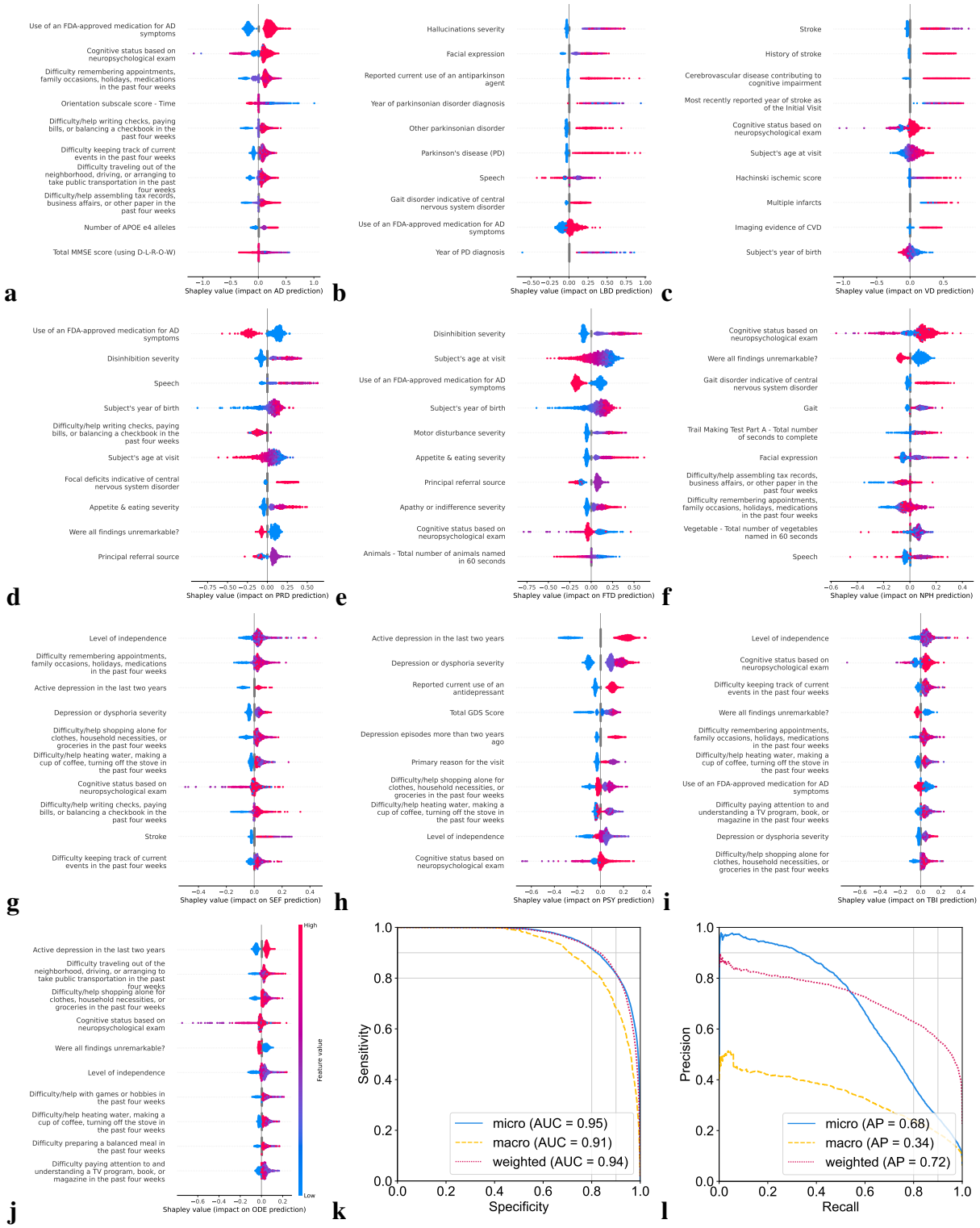

Figure 4: **Model assessment on dementia etiologies.** (a-j) Beeswarm plots illustrating Shapley values for subjects classified with AD, LBD, VD, PRD, FTD, NPH, SEF, PSY, TBI, and ODE are depicted. Adjacent to plot (j) is a colorbar that delineates the range of feature values for these conditions. (k, l) 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 the dementia etiologies.

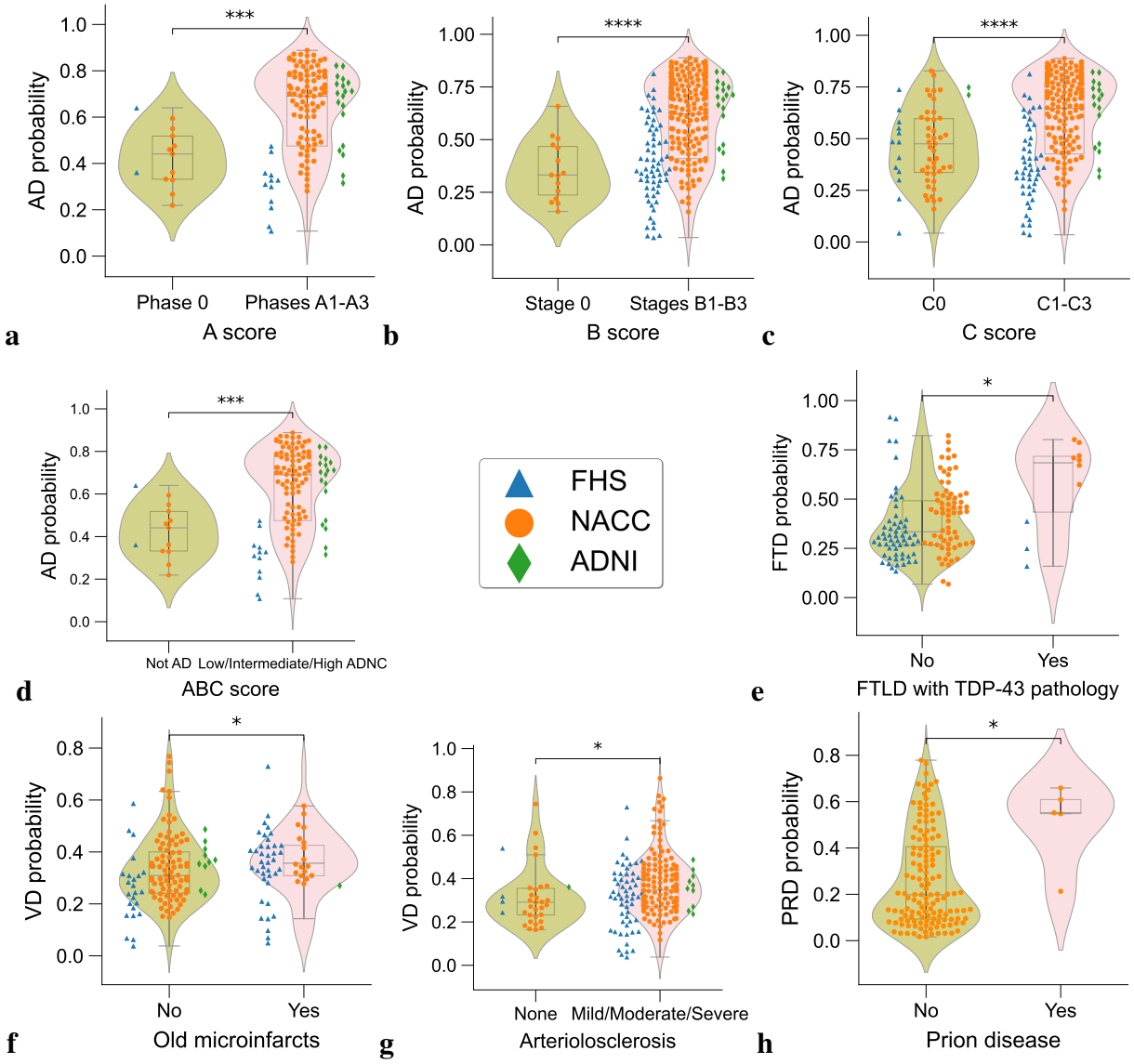

Figure 5: **Neuropathological validation.** Array of violin plots with integrated box plots, delineating the probability distributions as predicted by the model for different neuropathological grades. The analysis encompasses data from three distinct cohorts: the Framingham Heart Study (FHS), the National 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 a p-value less than 0.05; two asterisks (**) for p-values less than 0.01; three asterisks (***) for p-values less than 0.001; and four asterisks (****) for p-values less than 0.0001, reflecting increasing levels of statistical significance.

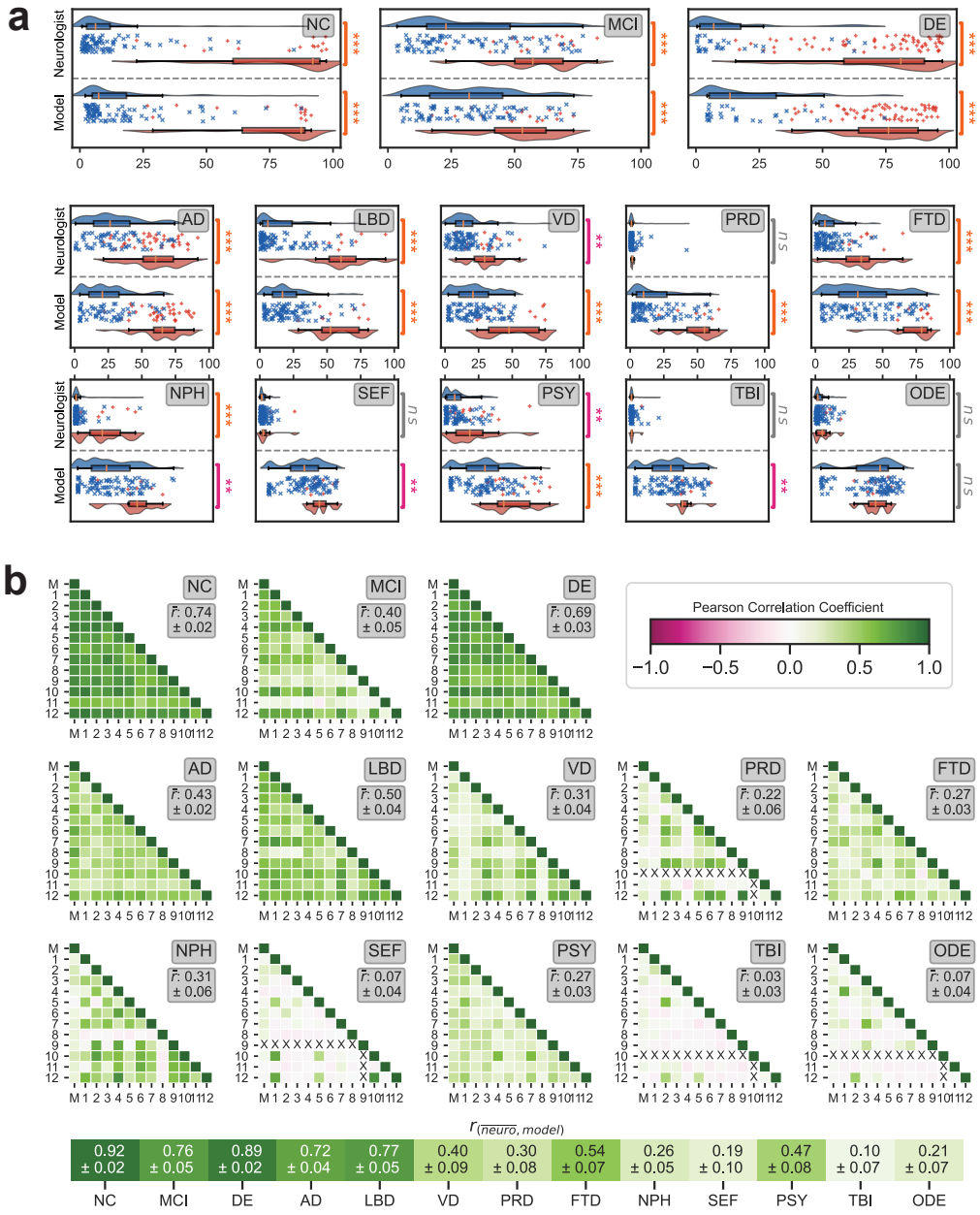

Figure 6: **Neurologist-level validation.** (a) Comparison between model-predicted probability scores and the assessments provided by practicing neurologists is shown. For the analysis, neurologists were given 100 randomly selected cases encompassing individual-level demographics, health history, neurological 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). In the visual representation, the boxplot in blue indicates the distribution of confidence scores for true negative cases, while the boxplot in red signifies true positive cases. The symbol '+' represents true positive cases, and 'x' denotes true negative cases. 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 $p < 0.0001$. These levels were determined using pairwise comparisons via the Brunner-Munzel test. (b) The figure presents the Pearson correlation coefficient across different diagnostic categories, comparing assessments from the neurologists ($n = 12$) and the model, marked as 'M'. Each diagnostic category from NC to ODE includes a matrix reflecting correlation coefficient values between individual neurologists and the model. Shades of green signify positive correlation, indicating agreement between the model and neurologists, whereas magenta shades suggest negative correlations, indicating potential discrepancies in assessments. The mean pairwise Pearson correlation coefficient for each etiology is presented along with a 95% confidence interval. The symbol 'X' denotes rater pairs where the Pearson correlation was not calculable, due to one or both raters giving label-specific confidence scores with no variance. The heatmap at the bottom shows the Pearson correlation coefficients between model probabilities and mean neurologist confidence scores for each label.