Received: 3 April 2023 | Revised: 19 July 2023 | Accepted: 23 July 2023

DOI: 10.1002/alz.13441

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

REVIEW ARTICLE

Artificial intelligence for dementia research methods optimization

Magda Bucholc[1] | Charlotte James[2] | Ahmad Al Khleifat[3] | AmanPreet Badhwar[4,5,6] | Natasha Clarke[4] | Amir Dehsarvi[7] | Christopher R. Madan[8] | Sarah J. Marzi[9,10] | Cameron Shand[11] | Brian M. Schilder[9,10] | Stefano Tamburin[12] | Hanz M. Tantiangco[13] | The Deep Dementia Phenotyping (DEMON) Network[1] | Ilianna Lourida[14] | David J. Llewellyn[14,15] | Janice M. Ranson[14]

[1]Cognitive Analytics Research Lab, School of Computing, Engineering & Intelligent Systems, Ulster University, Derry, UK

[2]NIHR Bristol Biomedical Research Centre, University Hospitals Bristol and Weston NHS Foundation Trust and University of Bristol, Bristol, UK

[3]Department of Basic and Clinical Neuroscience, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK

[4]Multiomics Investigation of Neurodegenerative Diseases (MIND) Lab, Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal, Montréal, Quebec, Canada

[5]Institut de génie biomédical, Université de Montréal, Montréal, Quebec, Canada

[6]Département de Pharmacologie et Physiologie, Université de Montréal, Montréal, Quebec, Canada

[7]Aberdeen Biomedical Imaging Centre, School of Medicine, Medical Sciences, and Nutrition, University of Aberdeen, Aberdeen, UK

[8]School of Psychology, University of Nottingham, Nottingham, UK

[9]UK Dementia Research Institute, Imperial College London, London, UK

[10]Department of Brain Sciences, Imperial College London, London, UK

[11]Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK

[12]Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy

[13]Information School, University of Sheffield, Sheffield, UK

[14]University of Exeter Medical School, Exeter, UK

[15]The Alan Turing Institute, London, UK

Correspondence
Magda Bucholc, Cognitive Analytics Research Lab, School of Computing, Engineering & Intelligent Systems, Ulster University, BT48 7JL, Derry, UK.
E-mail: m.bucholc@ulster.ac.uk

Funding information
Alan Turing Institute/Engineering and Physical Sciences Research Council, Grant/Award Number: EP/N510129/1; Medical Research Council, Grant/Award Number: MR/X005674/1; National Institute for Health

Abstract

Artificial intelligence (AI) and machine learning (ML) approaches are increasingly being used in dementia research. However, several methodological challenges exist that may limit the insights we can obtain from high-dimensional data and our ability to translate these findings into improved patient outcomes. To improve reproducibility and replicability, researchers should make their well-documented code and modeling pipelines openly available. Data should also be shared where appropriate. To enhance the acceptability of models and AI-enabled systems to users, researchers should

Magda Bucholc and Charlotte James are joint first authors.

David J. Llewellyn and Janice M. Ranson are joint senior authors.

Research, Applied Research Collaboration South West Peninsula, National Health and Medical Research Council, National Institute on Aging/National Institutes of Health, Grant/Award Number: RF1AG055654; Economic and Social Research Council, Grant/Award Number: ES/W010240/1; Motor Neurone Disease Association Fellowship, Grant/Award Number: Al Khleifat/Oct21/975-799; ALS Association Milton Safenowitz Research Fellowship, Grant/Award Number: 22-PDF-609; UKRI Future Leaders Fellowship, Grant/Award Number: MR/S03546X/1; E-DADS project, Grant/Award Number: EU JPND; EU Horizon 2020, Grant/Award Number: 666992; Alzheimer's Research UK; EU (SEUPB) INTERREG, Grant/Award Number: ERDF/SEUPB; HSC R&D, Grant/Award Number: COM/5750/23; National Institute for Health and Care Research Bristol Biomedical Research Centre; Fonds de recherche du Québec Santé - Chercheur boursiers Junior 1; Canadian Consortium for Neurodegeneration in Aging and the Courtois Foundation; NIHR Maudsley Biomedical Research Centre; Edmond and Lily Safra Early Career Fellowship Program; UK Dementia Research Institute; The Darby Rimmer Foundation

prioritize interpretable methods that provide insights into how decisions are generated. Models should be developed using multiple, diverse datasets to improve robustness, generalizability, and reduce potentially harmful bias. To improve clarity and reproducibility, researchers should adhere to reporting guidelines that are co-produced with multiple stakeholders. If these methodological challenges are overcome, AI and ML hold enormous promise for changing the landscape of dementia research and care.

KEYWORDS

artificial intelligence, classification, clinical utility, deep learning, dementia, generalizability, interpretability, machine learning, methods optimization, regression, replicability, semi-supervised learning, supervised learning, transferability, unsupervised learning

Highlights

- Machine learning (ML) can improve diagnosis, prevention, and management of dementia.
- Inadequate reporting of ML procedures affects reproduction/replication of results.
- ML models built on unrepresentative datasets do not generalize to new datasets.
- Obligatory metrics for certain model structures and use cases have not been defined.
- Interpretability and trust in ML predictions are barriers to clinical translation.

1 | INTRODUCTION

Dementia is an age-related condition with increasing global prevalence and an annual global cost estimated at ≈ US$1 trillion.[1] The timely detection of dementia is crucial in enabling effective disease management and providing optimal health care.[2] However, the complexity of underlying pathologies combined with considerable clinical heterogeneity present unique challenges to the development of effective treatments and early diagnostic tools for dementia. In recent years, developments in high-performance computing and machine learning (ML) algorithms have shown promise in improving dementia detection, monitoring, and management.[3–5] ML is a subset of artificial intelligence (AI) focused on the training of algorithms to perform tasks by learning patterns from data. The use of ML methods has enabled analysis of large volumes of high-dimensional data, integration of various data sources (i.e., clinical, imaging, genetic), and identification of new disease associations and disease subtypes not previously discovered with traditional statistical approaches.[6] The application of ML algorithms has enabled the development of more flexible and scalable models that can advance our understanding of complex disease pathways with minimal human intervention. In this review, we provide a summary of ML applications in dementia research, focusing on Alzheimer's disease (AD) and related dementias. Our review focuses on the ML techniques being applied, the data they integrate, and their intended use, and highlights major opportunities and challenges in translating ML technologies from research to clinical practice.

This review is the final article in a special issue on "Artificial Intelligence for Alzheimer's Disease and Related Dementias" published in *Alzheimer's & Dementia*. This series of eight articles provides a comprehensive overview of current applications of AI to dementia, and future opportunities for innovation to accelerate research. Each review focuses on a different area of dementia research, including experimental models,[7] drug discovery and trials optimization,[8] genetics and omics,[9] biomarkers,[10] neuroimaging,[11] prevention,[12] applied models and digital health,[13] and finally, this article on methods optimization.

2 | TYPES OF ML TECHNIQUES USED IN DEMENTIA RESEARCH

ML methods that have been applied in dementia research can be divided broadly into two categories: (1) traditional ML and (2) deep learning (DL). While traditional ML approaches require several sequential steps to identify relationships in data, specifically data pre-processing, feature extraction, and stable feature selection, DL techniques are inspired by the way biological nervous systems process information and hence, can learn directly from the input without the need for human intervention. Both traditional ML and DL algorithms can fall into one of four learning types: supervised, unsupervised, semi-supervised, and reinforcement learning. In this section, we briefly discuss each type of learning technique and the scope of their applicability to dementia research, except the use of reinforcement learning for dementia, which remains a relatively unexplored research area.[14,15]

2.1 | Traditional machine learning

2.1.1 | Supervised learning

Most ML approaches use supervised learning, which is a subcategory of ML that uses labeled data to learn a target function that best maps input variables to an output.[16,17] Supervised learning is commonly separated into classification and regression, each suited for specific types of problems with distinct output types.

Classification approaches

Classification models seek to determine which of a set of predefined groups/categories an instance belongs to, given a set of labeled examples. In the context of dementia, classification approaches have been developed for disease detection, prognosis, and management.[3,5] Many of them have been derived from logistic regression,[18–20] random forest (RFclass),[20,21] naïve Bayes,[22,23] K-nearest neighbor (KNNclass),[23,24] decision tree,[20] and support vector machine (SVM)[20,23–25] algorithms.

Comparing these approaches, the most appropriate model is problem-dependent, and will be influenced by the type of data; the data collection procedure and its underlying distribution; the classification task itself (e.g., multi-class dementia status prediction); the training and optimization procedure which, in most cases, involves human intervention; and a host of other factors.[24] This is further complicated by various methods of model evaluation which, depending on the relative importance of different evaluation metrics, can lead to differing conclusions on which approach is optimal. Such evaluation metrics include: accuracy (the number of correct predictions made by a model in relation to the total number of predictions made), sensitivity (the ability to predict the condition when the condition is present), specificity (the ability to predict the absence of the condition when the condition is not present), and the classifier discriminative power (as estimated from the area under the receiver operating characteristic [ROC] curve). In fact, different studies using different data sets have failed to generate one model that performed best in all applications.[24] For example, while RFclass has shown advantages with non-linearly correlated data,[21,26] SVM has demonstrated additional utility when there is a small number of samples and high number of features.[20,27]

Classification models in dementia have been built using many different amounts and types of input data.[28,29] Single-modality ML frameworks have been developed using cognitive and functional assessments (CFA),[29] magnetic resonance imaging (MRI),[30] positron emission tomography (PET),[31] cerebrospinal fluid (CSF) biomarkers,[32] and genomic data,[33] while more complex ML models use a combination of data inputs.[34] Evidence shows that classification approaches incorporating multiple data modalities generally lead to improvements in model performance.[24,35] However, the individual contributions of data types to the overall performance of multi-modality ML frameworks are often not assessed, raising the question about the trade-off between performance and cost effectiveness or efficiency of the proposed solutions.

RESEARCH IN CONTEXT

1. **Systematic review**: The development of machine learning (ML) models for dementia diagnostics, prevention, and monitoring is well documented, and their potential to transform clinical practice, experimental medicine, and clinical trial design has been highlighted in numerous studies. However, few models have been deployed clinically. While researchers search for the best ML solution, increased attention needs to be given to methodological challenges related to their development and adoption in clinical practice.
2. **Interpretation**: The implementation of ML models in clinical settings is currently a high-risk proposition due to their over-reliance on a single (often unrepresentative) data source, limited external validation, and an insufficient understanding of both the mechanisms driving model predictions and clinical utility.
3. **Future directions**: To overcome barriers to clinical translation, researchers need to ensure ML models are interpretable, externally validated, and assessed for risk of bias. The prediction modeling pipeline should be made openly available to facilitate replicability.

Regression approaches

Akin to classification, regression aims to learn the relationship between a dependent variable and several independent variables. Several regression approaches have been applied in dementia studies.[36–38] The simplest and most frequently used form is linear regression in which a continuous dependent variable is regressed onto independent variables, a process during which coefficients in a linear model are estimated. Linear regression has been used nearly universally across different topics in dementia research, including to predict disease outcomes, estimate time to dementia, and identify biomarkers and subtype dementia phenotypes.[39,40] These studies span varied modalities including clinical and phenotypic information,[41] molecular measurements in peripheral tissues,[42] neuroimaging,[43] as well as multiple types of omic analyses in *post mortem* brains.[44]

Although linear regression is the most widely used regression model in dementia, the construction of a linear regression incorporating a large number of predictor variables often results in poor generalization performance. To ease this problem, different penalization functions have been proposed, each imposing different constraints. In the dementia context, penalized approaches have been shown to produce more stable results for correlated data and data for which the number of predictors is much larger than the sample size.[45] Ridge regression performed especially well in the presence of high collinearity in linguistic data.[46] Lasso regression has shown some success in addressing high-dimensional AD data, especially in the context of genetic risk detection,[47] biomarker discovery,[48] and analysis of

neuroimaging-based endophenotypes.[49] Last, elastic net regression, which effectively combines lasso and ridge regression, has been shown to be a good compromise for variable selection and reduction of overfitting, while allowing for fast computational solutions and scaling to even more features than typical lasso regression. Elastic net regression has been used to study functional brain connectivity networks in the AD brain,[50] derive epigenetic biomarkers of Parkinson's disease (PD),[51] and classify AD and frontotemporal dementia (FTD) based on anatomical and functional imaging data.[52]

Linear regression models are empirical models that only describe the observed data, without a true understanding of the underlying mechanism that generates the data. Non-linear regression approaches, in which the function capturing the relationship between dependent and independent variables is more complex, are typically based on the underlying mechanisms that generate the data and, therefore, produce predictions that can be more reliable than linear models. In dementia research, linear regression models have often been ineffective in capturing non-linear relationships between biomarkers (e.g., neuroimaging data) and cognitive measures, in particular when a small number of observations and a large number of features were used for model training.[36] On the other hand, non-parametric kernel-based methods, a non-linear approach, have achieved relatively robust estimates of the regression function.[38] A possible explanation is that non-linear models are more powerful and better capture the complex relationships between model input and output. As such, non-linear models have been successfully implemented to identify potential descriptors for the decline of cognition using both single modality and multimodality data.[38] Moreover, non-parametric methods for the development of predictive models, including support vector regression (SVR), KNN regression (kNNreg), and RF regression (RFreg) have been used to differentiate between stages of dementia severity and improve risk prediction of AD.[24]

Supervised ML approaches are one of the most commonly implemented methods in dementia research.[16,17] However, both classification and regression often require a large amount of labeled data (especially when relationships are complex). This makes them challenging to apply when the number of cases is small, for example, when investigating rare dementia subtypes.

2.1.2 | Unsupervised learning

Unlike supervised algorithms, unsupervised algorithms search for previously unknown patterns within unlabeled data sets. As such, they have particular utility in dementia studies in which the labels (e.g., clinical diagnoses) are either unavailable[53] or uncertain.[54]

Unsupervised learning comprises a wide variety of approaches, of which mixed-effects models,[55] item response theory,[56] Gaussian processes,[57] kernel density estimation,[58] and mixture models[59,60] are a few examples. Broadly, these models have been applied to identify disease trajectories or subtypes[55,60,61] or produce a progression risk score.[56] Disease progression models have been used to model cognitive trajectories, predict decline, and provide pathophysiological insights.[62] One such model, the event-based model,[59] builds upon the hypothetical cascade model to obtain a sequence of events that describe one or more subtypes of disease progression.[60,63]

The largest category of unsupervised learning methods is cluster analysis, in which the aim is to find distinct groups within data, contingent on a suitable measure of similarity. Traditional clustering approaches, such as hierarchical clustering and density-based spatial clustering of applications with noise (DBSCAN), have been used to identify groups with, for example, different rates of atrophy[64] or CSF biomarker profiles.[65] These methods do not, however, account for the temporal component of dementia[66] and are difficult to evaluate due to the lack of any ground truth.[67]

As with supervised algorithms, the increasing richness of biological data across multiple modalities (e.g., imaging, biomarkers, clinical, and genetics) provides further opportunities for unsupervised learning, with the potential to uncover complex relationships and elucidate the underlying pathophysiological mechanisms.[66,68] This, and continuous methodological improvements for incorporating multi-modal data in a single model, increases the utility of unsupervised learning in extracting patterns from data. Recent applications include identifying a differential treatment response between data-driven subgroups[61] and reducing heterogeneity in clinical trials.[56,68] Nonetheless, the increased difficulty of model validation in unsupervised learning necessitates further work before clinical adoption becomes an option.

2.1.3 | Semi-supervised learning

Semi-supervised learning falls between supervised and unsupervised ML, using both labeled and unlabeled data. It is typically used in scenarios in which there is a large amount of data available, yet only a small proportion of samples have been labeled.

Semi-supervised algorithms use the information from the unlabeled data points to improve the performance of a model trained on the small amount of labeled data.[69] Therefore, these approaches are most useful in applications in which labeled data are limited. Different types of semi-supervised algorithms have been used for the classification of AD and mild cognitive impairment (MCI) with datasets of different modalities, including brain imaging,[30,70-72] among others.[73]

There are many examples from dementia research that demonstrate the superiority of semi-supervised algorithms for diagnosis or prognosis, relative to supervised algorithms based upon more limited data. Batmanghelich et al.[71] presented a framework for dimensionality reduction that showed a semi-supervised algorithm outperformed supervised learning methods, for both classifier accuracy and area under the receiver operating characteristic curve (AUC). Filipovych and Davatzikos[72] confirmed that in some scenarios, for example, in the absence of long-term follow-up evaluations, semi-supervised techniques may be preferable to identify individuals with progressive disorders, such as those at risk of conversion from MCI to AD. The high performance of semi-supervised algorithms in predicting MCI to AD conversion was also demonstrated in Moradi et al.[30] An et al.[70] developed a semi-supervised feature selection framework for

diagnostic purposes using both imaging and genetic data, achieving superior performance (as defined by AUC) in different dementia prediction classification tasks compared to an SVM model using only labeled data. Furthermore, experimental results of semi-supervised distance metric learning with label propagation (SRF-LP) showed superior accuracy compared to standard supervised learning algorithms, including RF, SVM, and AdaBoost, with an increase in the performance gap when the number of training samples was small.[73]

Given the increasing amount of data available and the inherent uncertainty around labels (i.e., clinical diagnosis) in dementia research, semi-supervised learning provides the opportunity to combine prediction of clinically relevant features with the utility of unsupervised learning, making the most of the available data. New techniques for semi-supervised learning are being developed, primarily centered around extending deep neural networks.[69] These methods remain underused in dementia research, though with researchers showing an increasing interest in DL algorithms, this is likely to change in the future.

2.2 | Deep learning

Deep learning is a subset of ML inspired by the structure and information processing of biological neurons, which are organized into stacked layers to form a deep neural network. DL does not require any human-designed rules to operate; the main advantage of DL over traditional ML approaches is that the time-consuming steps of pre-processing and feature engineering of datasets are minimal and less critical because DL models are able to obtain new representations of data (e.g., combinations of biomarkers, DNA sequence motifs) via multiple non-linear transformations.[74] DL, unlike traditional ML algorithms, can produce extremely high-level data representations from enormous amounts of raw data. Success of DL in domains such as computer vision and natural language processing (NLP), an increasing availability of large datasets, and improvements in computational power have resulted in DL algorithms becoming more popular with dementia researchers in recent years.

2.2.1 | Supervised learning

Supervised DL applications in dementia research have involved convolutional neural networks (CNNs),[75,76] deep belief networks (DBNs),[77] graph convolutional networks (GCNs),[78] and recurrent neural networks (RNNs).[79,80]

CNNs are typically applied to medical imaging data because they can capture multi-scale spatial information.[81] However, recent studies extended the use of CNNs by integrating multi-modal data.[76] Spasov et al.[76] built a CNN model combining MRI, neuropsychological, demographic, and apolipoprotein E data that achieved an AUC of 0.93 when identifying subjects with MCI that converted to AD (vs. stable MCI) over 3 years and an AUC of 1 when differentiating between AD patients and control subjects. These results demonstrate that DL models incorporating multiple data modalities may become vital to fully use the wealth of information available for dementia research. Furthermore, CNN-related models have recently found increasing application in NLP tasks for AD detection, including sentence classification, search query retrieval, and semantic parsing.[82]

RNNs are typically used on data that is sequential or time dependent in nature because their hidden state component (i.e., "memory cell") allows previous inputs to influence a given output. For example, Alam et al.[79] applied a long short-term memory (LSTM)–based RNN to predict onset of physical agitation episodes in patients with dementia, using motion sequences obtained from smartwatches. The LSTM–RNN model achieved a higher F1-score (0.85) than traditional ML methods, such as KNNclass (0.69), SVM (0.67), and AdaBoost (0.71), highlighting the potential of using such models in sensing-based behavior inference. In speech-based AD detection, DL model architectures including RNN, LSTM, gated recurrent unit, and bidirectional LSTM, have been used to extract timing information from audio data.[82]

DBNs have been used to explore dementia-related factors using genetic data[77] while GCNs have been applied primarily to neuroimaging data to capture brain network information for dementia classification.[78]

2.2.2 | Unsupervised learning

Unsupervised DL algorithms, such as autoencoders and variational autoencoders, allow for the automatic learning of data representations without the need for labeled samples. In dementia research, these methods have been effectively used for tasks such as anomaly detection, feature extraction, and patient clustering.[83] Bertini et al.[84] used a special type of autoencoder, auDeep, that allowed for unsupervised feature extraction from audio data to classify AD patients using their spontaneous spoken English. In Ithapu et al.,[85] an autoencoder was developed to take, as input, multi-modal imaging markers (fluorodeoxyglucose PET [FDG PET], florbetapir PET, and structural MRI), for predicting future decline to AD. The output of the model was a novel trial enrichment criterion, known as the random denoising AE marker (rDAm), for identifying patients that are most likely to progress from MCI to AD. The authors suggested that the use of the rDAm model could significantly improve our ability to design cost-effective AD trials, with smaller sample sizes and sufficient statistical power.

2.2.3 | Semi-supervised learning

Semi-supervised DL is less frequently used in dementia research compared to supervised or unsupervised DL techniques. So far, it has been applied to distinguish dementia from non-dementia patients using clock drawing data[86] and to determine the AD severity based on neuropsychological assessments.[87]

There are several key challenges that limit the translation of DL methods into clinical practice. DL algorithms require a large amount of data. Data processing can be time consuming and costly, and there is

also a lack of sufficient high-quality data, especially in multi-modality studies.[88] DL models are often referred to as "black-box" models due to non-linear feature transformations that emerge from multiple hidden layers, thereby reducing their interpretability compared to other ML approaches. This lack of interpretability can lead to a lack of trust in the models, which can act as a barrier to implementation (see Section 4.2 for more detail).

3 | GOALS OF THE STUDIES IMPLEMENTING MACHINE LEARNING APPROACHES FOR DEMENTIA RESEARCH

In dementia research, the goal of using ML approaches is typically to enhance clinical practice; identify new drugs and genetic targets; and aid the design of, and recruitment to, clinical trials. Due to the complex nature of dementia and the increasing size and complexity of datasets available traditional statistical methods are often insufficient. We discuss here areas of dementia research in which ML approaches have the potential to be transformative.

3.1 | Machine learning in clinical practice

In the context of clinical practice, ML has been applied to assessment of dementia risk, clinical diagnosis, prognosis, and care.[3] Of these, diagnostic studies have primarily focused on differentiating between stages of cognitive impairment or dementia subtypes.[24,89] Many have used neuroimaging data as input, mainly T1-weighted MRI and/or FDG-PET. These data have been used to develop ML models for identifying the severity of cognitive impairment (e.g., classify normal controls vs. MCI vs. dementia/AD) and differentiating between dementia subtypes (e.g., AD vs. FTD).[90] More recently, studies have emerged that use a multi-modal framework to integrate heterogeneous data types such as demographic, neuropsychological, clinical, genetic, CSF, or other omics data.[24] Novel ML approaches, such as a kernel-based SVM classifier with a truncated singular value decomposition dimensionality reduction technique, have emerged to more effectively handle the heterogeneity of such data types and the additional complexity introduced by considering their multi-scale interactions.[91] Multi-modal data integration can be a very useful strategy for early detection of dementia status or susceptibility, more robustly identifying disease targets, and identifying causal links among different biomarkers, symptoms, and clinical subtypes.

In the absence of established disease-modifying treatments for AD and other neurodegenerative diseases, the bulk of ML prognostic studies have focused on predicting the conversion from MCI to dementia using MRI,[92] electroencephalography (EEG),[93] magnetoencephalography (MEG),[93] neuropsychological measures,[94] genetic data,[95] or combinations of modality types.[96] A recent systematic review of studies predicting MCI conversion to dementia included results of 234 experiments from 111 articles.[97] The authors found that, despite some methodological issues, incorporating domain-targeted cognitive measures and 18F-FDG PET data into the model results in its superior predictive performance over models built without these data types. Furthermore, the addition of other feature types does not significantly improve performance of ML models compared to using cognitive or FDG-PET features alone. Similar observations have been made by Bucholc et al.[24] They found that classifiers built using CFA were the ones that performed consistently better than models based on other types of data, and that incorporating multi-modality features (e.g., cognitive and MRI or CSF data) into the predictive model provided only a small performance increase. In fact, considering all studies, it appears that the improvement one gains by including other data types along with cognitive measures is often not significant.[24,98] This is somehow encouraging given the fact that cognitive measures can be easily collected as part of clinical routine, at a low cost. However, if specialist data such as neuroimaging and genetics becomes less expensive and more practical to collect in the future, classification approaches incorporating multiple data modalities, which can improve predictions even by a small margin, may become clinically advantageous.

Combining NLP and ML techniques also has the potential to greatly enhance research in the dementia field. The input generated from large language models has demonstrated impressive performance on many NLP tasks.[99] Numerous studies have suggested that analysis of acoustic features extracted from speech audio and linguistic features derived from written texts or speech transcripts may lead to the discovery of novel, non-invasive biomarkers of cognitive impairment given that subtle changes in language can be observed long before clinical diagnosis of dementia.[100] In particular, the detection of dementia from spontaneous speech has become a prominent topic in recent years due to the fact that it constitutes a time-effective, cost-effective, and non-invasive procedure.[99]

Given the literature on ML approaches in dementia research, it is fair to say that a number of different algorithms have been shown to detect dementia and its prodromal phase with relatively high predictive accuracy, but their performance significantly varies when other performance metrics are considered.[24] In some cases, the high accuracy might have been arbitrarily increased by using a dataset with a large proportion of people without dementia.[101] From the clinical point of view, no single metric captures all the desirable properties of a model and therefore, it is important to have thorough understanding of, distinctions between, and uses and misuses of each of these metrics, especially in the context of clinical utility. For instance, Model 1 (e.g., used for automatic dementia screening) could identify ≈10% of the population with probable dementia, generating a very high case load for clinicians to screen, review, and test. This would have high sensitivity (i.e., proportion of patients with actual dementia identified) but a low positive predictive value (PPV; i.e., a proportion of those identified as having dementia in routinely collected data sets that are true dementia cases). Conversely, Model 2 could identify only 1% of the population as probable dementia, requiring clinical review. Even though this has lower sensitivity, it would be more efficient in having a higher PPV. Hence, it is essential to provide all the necessary information

about a ML model, as well as the reference standard and the data used to develop it, to characterize a diagnostic process adequately.

Apart from distinctive differences in ML metrics for performance comparison, other issues need to be addressed before diagnostic, prognostic, and risk prediction algorithms can be routinely applied in settings such as memory clinics. These include reproducibility, model validation, data leakage, generalizability of the model to other settings, and model interpretability. Recently, a framework using a transfer learning paradigm with ensemble learning algorithms (using multiple methods in tandem to make a consensus prediction) has been proposed for risk prediction of dementia at both population and individual levels.[102] Compared to a baseline model, the target model, using a parameter-transfer learning approach (training on larger, less task-specific datasets and then fine-tuning on smaller, more specific datasets) to update the decision boundaries of the baseline model, achieved better performance across all the performance metrics, including an increase in sensitivity of 19.1%, specificity of 2.7%, accuracy of 16.9%, and AUC of 11%. This shows the potential for transfer learning to overcome some of the big challenges of dementia research, such as regulatory challenges associated with data aggregation, management, privacy, and informed consent for the collection, use, and sharing of data. In the future, some of the larger datasets used for dementia research could serve as source data for the development of ML models that smaller studies could use transfer learning to build upon.

Finally, the limitations of ML systems need to be clearly communicated to clinical end users before a new generation of clinical decision support systems (CDSSs) designed to exploit the potentials of data-driven decision making are adopted and routinely used in clinical practice. Although a few studies have reported on the implementation of ML methodologies for determining the severity of dementia[24,103] and differential diagnosis of dementia in memory clinics,[104] CDSSs for dementia diagnosis, prognosis, and management at the point of care are still not routinely used.

3.2 | Machine learning in experimental medicine

The theoretical number of unique small molecules is more than an order of magnitude greater than the number of atoms in the observable universe.[105] Therefore, optimizing the design of synthetic molecules for a desired therapeutic outcome is a computationally intractable task to perform by exhaustive brute force calculations. ML has the potential to much more efficiently search this complex space to design drugs that best approximate viable candidate molecules.[106] DL in particular has proven to be adept at these tasks as it is able to learn new representations of multi-modal data (e.g., neuroimaging, genomics, and clinical records) using non-linear mappings.[107] As a consequence, there has been an explosion of applications to pharmacology and its related fields, chemoinformatics, and structural biology.[108] Example use cases include: predicting 3D structure and function of small molecules from chemical formulas,[109] protein–protein or protein–drug interaction modeling, clinical outcome or biomarker prediction,[110,111] and personalized genome–drug interactions (i.e., pharmacogenomics). In addition to the design of new drugs, ML has also been applied to the repurposing of existing therapies developed for other kinds of treatment. Of particular relevance to dementia, a recent study by Dias et al.[112] used the IBM Watson for Drug Discovery online tool, incorporating an NLP algorithm, to extract lists of gene–disease and drug–disease relationships from millions of published research articles related to the medical sciences. They then combined these relationships with gene co-expression data from human brain samples and created an extended knowledge network that revealed previously unknown relationships between different psychiatric and neurological disorders and hundreds of drugs. As a result, several drug candidates were identified to reposition as therapies for AD ($n = 25$), PD ($n = 1$), and dementia ($n = 1$; see Table S4 of original publication for details).[112] Furthermore, potential associations between the pathological stage of AD and genes using a ML-based Drug Repurposing In AD (DRIAD) framework were evaluated in Rodriguez et a.l[113] Here, DRIAD was applied to lists of genes that were differentially expressed after exposing neuronal cells to a test panel of 80 clinically approved drugs, generating a ranking of possible repurposing candidates that, after additional validation in relevant in vitro and in vivo AD model systems, could be evaluated in a clinical trial.

Outside the domain of drug discovery, ML has been applied in disease genomics, for example, to predict the functional consequences of mutations in both protein-coding and non-coding genomic sequences.[114,115] These kinds of sequence-informed ML approaches hold several considerable advantages over explicit rule-based models, including the ability to then conduct in silico mutagenesis to probe the effects of all possible mutations.[116] These are powerful tools for functional impact prediction (e.g., gene expression, chromatin modification). However, when it comes to predicting disease status from genomic data, ML-based approaches have yet to show substantial performance increases over simpler additive methods like polygenic risk scores in autoimmune disease or AD.[33] This could be due to several factors including lack of sufficient sample sizes, the use of genotype arrays instead of whole-genome sequencing, the use of highly processed input data (e.g., disease-associated variants identified through simple linear models), or the use of suboptimal ML architectures for this problem.

Another challenge is that sequence prediction models have not typically addressed the tissue- and cell-type–specific nature of mutational effects. Recent advances in single-cell transcriptomics, epigenomics, and proteomics have permitted the accumulation of single-cell atlases in model organisms of dementia-related diseases, primary cells from living patients (e.g., blood),[117] *post mortem* samples (e.g., brain tissue),[118] and patient-derived induced pluripotent stem cells (iPSCs).[119] Some dedicated databases have emerged for integrating and hosting AD-related single-cell datasets.[120] These datasets can subsequently be used to train DL models to learn latent representation of cell-type–specific responses to various genetic and chemical perturbations, as has been done in cancer cell lines previously,[121] as well as natural genomic variation.[122,123] Therefore, the evaluation of cell-type–specific effects of dementia-associated mutations is becoming increasingly feasible at scale.

3.3 | Machine learning in clinical trials

A focal point of research using ML for clinical trials is the development of algorithms that possess the capability to identify patients who will develop dementia in the future. This can enhance the precision of patient selection for clinical trials and aid in the monitoring of disease progression. Additionally, ML techniques can be used to analyze vast amounts of data derived from clinical trials, such as electronic medical records and imaging data, to discern patterns and potential new treatment options.[124]

Several studies have investigated the role of ML in clinical trial design.[124,125] Ezzati and Lipton[126] developed a ML framework incorporating the KNNclass algorithm, to identify individuals who were more likely to show cognitive decline during the follow-up and used this subgroup of participants for analysis of treatment effects. Their results indicated the ML model could provide ≈17% and ≈25% improvement in predictions of cognitive decline at 12 months and 24 months follow-up respectively, and hence, could be used to improve the power of clinical trials. Reith et al.[127] used baseline clinical information and CNN-extracted PET features to predict changes in quantitative biomarkers of brain pathology with gradient-boosted decision trees. The use of ML helped them identify a cohort with the fastest amyloid deposition, at a two to four times higher rate than random selection or other common selection methods used for patient recruitment. The potential of different ML approaches to assist in recruiting patients at risk of dementia has also been shown in other studies.[85,128] Hane et al.[128] showed that the incorporation of clinical notes into ML frameworks can aid model accuracy and can therefore be routinely used to identify individuals for interventions, such as disease management programs and screening for clinical trials. Another study applied a new ML approach, Subtype and Stage Inference (SuStaIn), to routinely acquired MRI scans from patients with dementia to identify different subtypes of dementia early in the disease process.[63] The algorithm was able to determine three different subtypes of AD, which broadly matched those found in *post mortems* of brain tissue.

It is increasingly recognized that dementias are preceded by a pre-symptomatic or prodromal period of varying duration, during which the underlying disease process unfolds. This highlights opportunities to slow disease progression during different pre-symptomatic phases of the disease, when it is more likely that pathological changes can be slowed, arrested, or even reversed. Selecting study participants at high risk for dementia or MCI is therefore essential to design cost-effective prevention trials. The use of ML in clinical trials for dementia has the potential to increase the success, generalizability, and efficiency of these trials. These technologies can assist with patient recruitment and cohort composition, improve patient retention and protocol adherence, help process and manage large quantities of data from sources such as wearables and other smart devices, identify drug targets and candidate molecule generation, and discover subgroup effects. Ultimately, this can lead to the offering of new treatments to the right population at a faster pace.[124,125]

Despite the increasing number of ML models that have been developed for dementia research, not many have been used for patient–trial matching and recruitment before the start of a clinical trial or patient monitoring during the trial. There are still challenges to be addressed, including ensuring the safety and privacy of patient data, and addressing concerns regarding the interpretability of ML-generated results. Further research is necessary to develop robust and validated ML models that can be widely adopted within the clinical trial process for dementia.

4 | REPRODUCIBILITY, REPLICABILITY, INTERPRETABILITY, AND CLINICAL APPLICABILITY ISSUES IN DEMENTIA RESEARCH

4.1 | Reproducibility and replicability

As we develop novel computational models to understand dementia, issues of reproducibility and replicability become increasingly important. While reproducibility involves obtaining the same results from the same model and data, replicability is based on applying the same model to independent datasets and observing generalizability of the findings.

Reproducibility is a long-standing issue in scientific research, particularly experimental medicine, as results can vary due to both controllable (e.g., following identical protocols) and uncontrollable (e.g., system stochasticity) factors. The use of computational models therefore significantly improves reproducibility: an algorithm trained to perform a task will always give the same results when applied to the same data. However, for research to be fully reproducible, both data and code need to be made available. A recent review of the use of ML for modeling progression to AD found that, while 75% of studies used publicly available data, only 7% shared their implementation code.[129] If guidelines such as the Turing Way and Findability, Accessibility, Interoperability, and Reuse (FAIR) principles are routinely followed, the increasing use of ML in dementia research will allow results to be easily reproduced.[130]

In contrast to reproducibility, replicability is much harder to achieve: an algorithm optimized to, for example, predict incident dementia in one cohort is not guaranteed to perform well in a completely different cohort. This is particularly true for algorithms, such as neural networks, that are susceptible to overfitting during training. One approach to improving replicability is to use data from multiple sources. There exist many large-scale studies that are easily accessible to dementia researchers, for example, Alzheimer's Disease Neuroimaging Initiative (ADNI)[131], Longitudinal Aging Study in India (LASI),[132] National Alzheimer's Coordinating Center (NACC),[133] and UK Biobank.[134] In combination, these studies provide an opportunity to improve model generalization; however, simply combining the data would be an enormous task due to differences in, for example, the reporting of variables or diagnostic criteria.

Sample size has been known to play an important role in the generalizability of ML models.[135] Larger datasets often provide a more comprehensive representation of the target population, encompassing a wider range of variations and diversity. This helps the model identify the patterns and relationships that are highly relevant to

unseen data, ultimately improving the model's ability to generalize effectively. In dementia research, studies using ML techniques have been shown to exhibit substantial variability in sample size. In the systematic review of 92 studies integrating interpretable ML methods for dementia prediction, Martin et al.[136] showed that sample size across all included studies ranged from $11 \leq n \leq 95{,}202$, with larger datasets providing increased confidence and better estimations of the model performance on unseen data. This is consistent with findings by Javeed et al.[137] They assessed the performance of ML-based diagnostic systems for dementia across various dataset sizes and showed that ML models using imaging data exhibit improved accuracy when applied to larger datasets. Although there are studies with relatively small samples in which ML techniques, like SVM ($n = 137$), deep polynomial network combined with SVM ($n = 103$), and linear discriminant analysis ($n = 223$) have been used with high accuracy in classification problems (i.e., 88%, 96.3%, and 89%, respectively),[138–140] there is ongoing controversy surrounding the impact of different combinations of data types, data validation techniques, and the type of classification tasks (e.g., healthy vs. AD, MCI vs. AD) on the performance of these models. For instance, although Zheng et al.[139] obtained a classification accuracy of 96.3% in discriminating between healthy controls and AD patients using a small neuroimaging dataset ($n = 103$), their algorithm was not tested against unseen data, and the size of their training/validation sets used in 10-fold cross-validation was relatively small ($n = \approx 90$ for the training sets, and $n = 10$ for validation sets). Furthermore, ML models used to discriminate between AD and healthy controls have, on average, shown higher accuracy compared to those designed for differentiating MCI versus healthy control or AD, or MCI converters versus MCI non-converters to AD, with the same problems of the sample size and repeated use of the same data.[141] Small datasets provide less reliable estimates of the underlying data distribution and do not capture more subtle patterns present in the data. Consequently, ML models built using small datasets demonstrate reduced stability, leading to a less precise indication of the model's performance in real-world scenarios. Larger datasets provide a broader context, reducing the likelihood of overfitting and improving the model's ability to generalize to new data. It is, however, important to note that simply increasing the dataset size does not guarantee perfect generalizability. Other factors such as data quality, appropriate model architecture, and representation bias also play important roles.

In particular, the heavy reliance on studies using data from a single source, such as clinic-based cohorts (ADNI, NACC), although promoting cross-comparisons of results, may inflate the estimates of accuracy and impose limitations on the generalizability of the methods used. For instance, the original design of ADNI aimed to characterize a clinical trial population that primarily consisted of older age groups, with more advanced pathology.[142] These characteristics deviate from what is typically observed in the broader population. Similarly, those enrolled in the NACC Alzheimer's Disease Research Centers' (ADRC) cohorts are not random volunteers and therefore, are not representative of a wider population. Their educational attainment and income levels are higher than the national average, and ≈50% of ADRC participants have a family history of dementia.[143] The community-based cohort study of the UK Biobank, widely used for dementia research, also lacks natural data representations, limiting our ability to develop ML models that can be used effectively and fairly in clinical practice.[144] Despite the fact that datasets from memory clinics or hospitals provide the advantage of tailoring the research protocols to meet specific research needs, ensure balanced group sizes, and maintain consistency across multiple timepoints, these studies often encounter limitations in participant recruitment, which can consequently result in smaller sample sizes.[145] This in turn may lead to the inability of the ML model to capture more fine-grained patterns and discriminative features present in the wider population.

State-of-the-art ML methods such as transfer learning and multi-task learning, have the potential to improve replicability and generalizability of results. For example, transfer learning could be used to train an algorithm on data from more than one study, providing some variables in both data sets are the same, by finding a mapping between the data domains. Such an approach could help improve generalizability of the results and reduce overfitting to the cohort used for training, thus increasing the chance of replicability in further cohorts. Successful applications of transfer learning in dementia have been demonstrated by Danso et al.[102]

While transfer learning implies a sequentially shared representation, multi-task learning is related to a shared representation that is developed concurrently across different tasks. So far, few studies implemented the multi-task ML approach in dementia research.[146] One example is the use of a deep feedforward neural network (DFNN) approach based on multi-task learning (i.e., using multiple loss functions) to simultaneously detect AD and determine its progression stage.[146] Khoei et al.[146] showed that the proposed model can accurately classify and predict AD and its stages using both binary and multi-class classification of AD (three and four different class labels), over 5 and 10 years.

4.2 | Interpretability of machine learning models

Alongside reproducibility and replicability, the interpretability of ML algorithms is an important consideration for many applications, particularly in cases in which the algorithm is making life-changing decisions. This is clearly the case for dementia research, and health-care applications in general. For a clinical decision maker or patient to trust an algorithm, an understanding of how it makes its predictions is pertinent; lack of transparency represents a barrier to translation from research to clinical practice.[147,148]

The interpretability of an algorithm can be divided into two classes: global interpretability and local interpretability. Global interpretability refers to how the components of an algorithm, such as features and weights, combine to make decisions. In contrast, local interpretability is associated with individual decisions; how did the algorithm make its decision about a specific sample or patient.[149] Global interpretability is hard to achieve due to the often-complex nature of ML algorithms. However, it is possible to interrogate algorithms at a modular level to

understand how varying one feature or weight impacts the decisions being made.

One of the most common methods used in dementia research to interpret algorithmic decisions is (permutation) feature importance, which does exactly this: it informs the user how much each feature contributes to the decisions being made.[150] One benefit of feature importance is that it is model agnostic and so can be used in combination with many ML algorithms. Model-specific approaches to interpretability are more often used alongside DL algorithms.[151,152] For example, Qiu et al.[153] constructed "disease probability maps" to visualize how a fully convolutional network determined AD status from MRI images.

Common ML algorithms used in dementia research, such as decision trees, SVM, and Bayesian networks, are inherently interpretable; they learn a set of rules or relationships between variables with a known structure that can be interrogated by the user. In contrast, "black-box" models, including anything involving a neural network, are harder to interpret because the mathematical relationships between variables are learned from the data and are inaccessible to the user. State-of-the-art methods, such as local interpretable model-agnostic explanations (LIME)[154] and Shapley values,[155] can be used to assess how a "black-box" model came to a certain decision. However, while methods such as these exist and produce encouraging results, they are currently not routinely used in dementia research.

It may be questioned whether methods for interrogating "black-box" models are necessary for dementia research. There are many examples in the literature in which the performance of different ML algorithms has been compared. For the task of dementia diagnosis, interpretable algorithms often perform as well as, if not better than, "black-box" models.[156,157] However, which algorithm performs best depends on the data, the application, and how performance is measured[4,158]; interpretable algorithms can perform better in terms of predictive power[159] while "black-box" models often achieve higher discriminative accuracy.[158]

This boost in accuracy may make "black-box" models seem attractive, but this is at the cost of interpretability. Recent reviews have found that while "black-box" models are being used in dementia research, researchers still favor interpretable models.[3,17] In the future, if methods to interpret "black-box" models are routinely implemented, and the use of ML becomes more common in clinical practice, this pattern may shift.

4.3 | Clinical applicability issues

There are currently limited examples of ML models for dementia diagnosis, prognosis, and care being successfully deployed into clinical practice. Issues associated with dataset collection, such as dataset shift (i.e., the scenario in which the joint distribution of inputs and outputs differs between the training and test sets), omitted variable bias, unintended discriminatory bias, and repeated use of limited datasets, result in impairing the ML model's ability to generalize to new populations. Furthermore, technical issues related to difficulties in extracting patient data in a harmonized, machine-readable format; the lack of understanding of the mechanistic basis of model predictions; differences between clinical settings (e.g., including differences in equipment, coding definitions, and computer systems, such as electronic health records); and variations in local clinical practices further complicate the clinical adoption of ML solutions in dementia. Upon deployment, differences in the distribution of true patients and healthy individuals compared to that seen during training can also impact model performance. For example, a well-calibrated model will identify healthy individuals at a rate equal to that seen in the training data. If this is artificially high because proportionally fewer true patients volunteer for research, the model will predict a lower prevalence of disease.[160]

The impact of these issues is demonstrated in the neuroimaging field, in which the mean squared error of a CNN trained to predict regional brain atrophy using ADNI data increased from 0.31 to 0.41 when tested on memory clinic data.[161] Performance improved when the training set included a wider range of scanner images and protocol types, suggesting a route to improved clinical application through increasing heterogeneity of training data.[161] Similar approaches, training on heterogeneous memory clinic data,[162] or using a transfer learning approach,[102] have demonstrated improvements in generalizability. Better understanding of the genetic architecture of dementia has also been increased through transethnic genome-wide association studies (GWAS), enhancing research previously limited to European-centric populations with more diverse populations.[163]

Studies that trained models on cognitive data face similar issues. Promisingly, an ML-based iPad application to assess global cognition has demonstrated good performance detecting MCI and mild AD, with AUCs of 0.81 and 0.88 respectively, using two linguistically and culturally distinct datasets.[164] Detecting dementia through speech using ML and NLP is a growing field; however, studies have largely focused on English-speaking participants, limiting generalizability to other languages.[165] Recent studies have found that training on multilingual datasets improves performance (F-score = 0.85) compared to training on an English corpus alone (F-score = 0.80).[166]

Parallel issues that may limit clinical applicability are cost and availability of resources, for example, relying on expensive imaging data. Multiomics approaches may increase this problem, and so the use of complementary and unique methods to find the most informative features will be key.[167] Models based on widely available and accessible data, such as wearable devices, may aid clinical translatability.[168] Collaboration between researchers and stakeholders will also improve clinical translation, such as working with clinicians to develop ML solutions for "real-world settings."[169] A lack of trust in algorithmic decision making may also limit translation, compounded by opaque models with low interpretability; trust can be improved with transparent, explainable approaches,[169] and reporting an algorithm's confidence in its decision.[170] In addition, performance metrics should aim to capture real clinical applicability and be easily understandable to clinicians. Given a lack of data on outcomes, it is challenging to predict the impact of ML methods for dementia, but models should be rigorously tested prior to being deployed.[161] Clinical trials for ML interventions

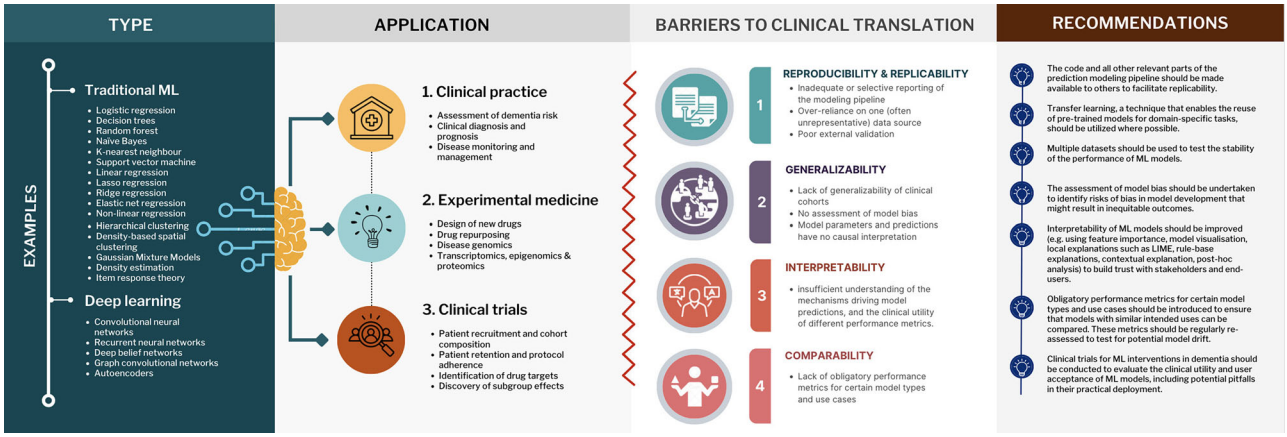

FIGURE 1 Applications of machine learning in dementia research and barriers to clinical translation. ML, machine learning

in dementia (similar to those designed for cancer research[171]) could evaluate the clinical utility of ML models given their inherent opacity and "black-box" nature, as well as address the challenges related to generalizability and interpretability of ML models.

5 | DISCUSSION

5.1 | Summary of existing methods and their limitations

It is clear that the use of ML in dementia research is becoming increasingly common (Figure 1). Supervised learning algorithms are most frequently implemented: classification is commonly applied to predicting the risk of progression to dementia and differential diagnosis of dementia, while regression is used for detecting cognitive changes, evaluating dementia severity, estimating time to dementia, and identifying risk factors.[172] In situations in which labeled data are not available or insufficient, unsupervised and semi-supervised approaches are being used with promising results.[61,73] Furthermore, the application of DL models for early detection and classification of dementia has gained considerable attention in recent years, mostly due to their improved performance over traditional ML approaches.[76] However, the successful implementation of ML models in clinical settings is currently a high-risk proposition due to their over-reliance on one data source; limited external validation; insufficient understanding of the mechanistic basis of model predictions and clinical utility of different performance metrics; and potential bias caused by missing data or inappropriate use of methods for missing data imputations. Other sources of bias, including measurement bias, evaluation bias, sampling bias, and algorithm bias also impair the ability of ML models to generalize to new populations, and can lead to bias against specific groups or individuals. Sampling bias occurs when the sample used to build a ML model is not representative of the target population from which it is drawn. In other words, the individuals or elements in the sample are not randomly selected and do not accurately reflect the characteristics of the entire population. This can affect the model's ability to adapt accurately to new, previously unseen data.[173] For example, the lack of generalizability of the findings of large datasets (e.g., ADNI and similar highly selected clinical cohorts) to other populations, such as community-based samples, may critically limit their ability to produce ML models readily transferable to other settings.[174] Measurement bias refers to the systematic error or inaccuracy in the measurement of certain variables or outcomes related to dementia. For example, while NACC ARDCs use standard criteria and procedures, there might still be some variation in selection and diagnosis factors across different centers.[175,176] Evaluation bias, on the other hand, occurs during the analysis and interpretation of data. It refers to the bias introduced by the researcher's judgment, or the analytical methods used, leading to distorted or misleading conclusions. Evaluation bias can arise due to subjective choices made during data analysis, selective reporting of results, or the influence of prior expectations on the interpretation of findings. In dementia research, evaluation bias occurs when researchers selectively report only the results that support their hypothesis, for example, by assessing the model's performance solely based on its accuracy, without specifying other performance metrics.[177] Finally, algorithm bias refers to the presence of systematic and unfair inaccuracies or limitations in ML algorithms used to analyze the data. It may occur when researchers unknowingly include certain features in the dataset during the development of the ML model (e.g., dementia diagnosis) that are correlated with other features (e.g., ethnicity or socioeconomic status). As a result, the ML model may inadvertently learn and reinforce existing biases present in the data. For instance, if the dataset is biased toward higher income individuals from a particular ethnic group, the model might disproportionately associate higher income with a lower risk of dementia. As a consequence of algorithm bias, the model's predictions could be less accurate and potentially discriminatory when applied to individuals from underrepresented or different ethnic and socioeconomic backgrounds. As noted in Gianattasio et al.[178] bias rooted in unrepresentative datasets used for training and poor model calibration can lead to racial bias in model application.

The "black-box" nature of many ML models, in particular DL models in which input data can undergo complex transformations over

many layers, means that they have no explicit declarative knowledge representation and hence, provide predictions without any accompanying justification. Other ML methods may be able to list dependencies between the target prediction output and the input features but those relationships are too complex to understand or verify.

The availability of data for ML algorithm development and validation in dementia is also a significant barrier to their adoption and routine use in clinical practice. The majority of ML models require large amounts of data for training and testing to ensure generalizability beyond the training set. In addition, different data modalities, such as cognitive scores, behavioral measures, neuroimaging, or genetic data, provide distinct information about the underlying mechanisms of the disease and the patient's condition. Understanding the unique characteristics and challenges associated with each data modality is therefore essential for building effective and accurate models for dementia research. For example, each data modality requires a specific representation for analysis. Cognitive scores may be numerical or categorical, while neuroimages consist of spatial data with complex patterns. The ML model's ability to handle and effectively represent the input data can influence its performance. Furthermore, different data types require specific feature extraction techniques. Cognitive scores may require simple statistical features, while neuroimaging data often necessitates more complex methods like CNNs to capture relevant patterns. Different data modalities may also capture different aspects of the disease and its progression. For example, neuroimages can reveal structural changes in the brain associated with AD and cognitive abnormalities, while genetic data may highlight underlying genetic risk factors. The complexity and non-linearity of relationships between data modalities can substantially impact the performance of a ML model. Incorporating multiple data modalities into a ML model can lead to a more comprehensive and accurate understanding of dementia. However, challenges may arise due to data heterogeneity, missing data, and the need for sophisticated data integration techniques. Successful use of these diverse data types requires careful data preprocessing, feature selection, and model architecture design, considering the unique characteristics and challenges associated with each modality.

Finally, the objective comparison of models across dementia studies is challenging as obligatory performance metrics for certain model structures and use cases have not been defined. All these barriers to ML adoption require attention at each stage of model development, validation, and use, to enhance stakeholder trust (including regulatory agencies, researchers, clinicians, and industry partners) in the process and results (Figure 1).

5.2 | Recommendations for future research

While the use of ML in dementia research presents many opportunities for the future, for algorithms to be translated to clinical practice, issues surrounding reproducibility, replicability, generalizability, interpretability, comparability, and trust in decisions need to be addressed.

5.2.1 | Reproducibility and replicability

Issue

Inadequate or selective reporting of the modeling pipeline affects replication and reproduction of research results.

Recommendation

The code and all other relevant parts of the prediction modeling pipeline should be made available to others to facilitate replicability. This includes details, such as (1) the nature of the applied algorithms; (2) the seeds used for selecting the partitions of the dataset for training, validation, and testing; (3) the process of handling missing data (if applicable); and (4) descriptions of data pre-processing and validation procedures including class imbalance handling (if applicable), hyperparameter selection, optimization, and thresholding.

Issue

The majority of dementia studies implementing ML approaches assume that once the model performs well on the data for the specific domain or problem it is tasked with, the model can be successfully applied to new data. However, this assumption does not hold in many cases. Because the model is constructed for a specific problem/dataset, it has to be retrained and tuned for any new problem/dataset. This not only affects the time- and cost-effectiveness of the proposed solution but also may lead to the changes in model performance, often caused by discrepancies in data distribution or feature space.

Recommendation

Transfer learning, which allows re-using pre-trained models for domain-specific tasks, represents a time-saving alternative. The application of transfer learning frameworks in dementia research already shows some promising results, including the improved accuracy and prevention of overfitting when dealing with relatively small datasets.[179]

5.2.2 | Generalizability

Issue

Certain disease stages and patient characteristics are less likely to be well represented in datasets produced for clinical research. The lack of generalizability of clinical cohorts to other populations limits their ability to produce ML models that can adapt to other datasets. Furthermore, developing ML models using retrospective datasets does not always translate well to prospective applications.

Recommendation

To ensure generalizability of ML models beyond the training data set, developers should use multiple datasets to test the stability of model performance. In addition, the assessment of model bias should be undertaken to identify risks of bias in model development that might result in inequitable outcomes. This should include a

transparent, easily accessible mechanism to monitor the impact of algorithmic bias on users subgroups post-deployment.

Issue

The overwhelming majority of ML models used in dementia are purely associative, that is, they focus on predicting outcomes based on a predefined set of variables. Neither their parameters nor their predictions have a causal interpretation and hence, they should not be used to identify causal relationships. Because they are not constructed to understand causality between input data and output predictions, they cannot generalize well when changes in input data occur or when there are multiple possible causes of patient symptoms (e.g., differential diagnosis of dementia). The associative nature of ML algorithms for dementia places constraints on their performance and can lead to suboptimal diagnoses.

Recommendation

Causal inference, that is, understanding how diagnosis is obtained and clearly defining the desired output, can be used to produce more robust and generalizable ML models.[180] In particular, meta-modeling and meta-learning can be used to overcome the current limitation of ML in performing causal discovery.[181]

5.2.3 | Interpretability

Issue

Interpretability and trust in algorithmic decisions are barriers for research being translated to a clinical setting. The "black-box" nature of many ML algorithms means the connection between features and predictions is obscured, and it is unclear how they make decisions.

Recommendation

Methods such as LIME and Shapley values can be used to improve interpretability and trust in models. These methods allow clinicians to interrogate the decision-making process of ML models, thus increasing transparency. Furthermore, clinical trials for ML interventions in dementia should be conducted to evaluate the clinical utility and user acceptance of ML models, including potential pitfalls in their practical deployment.

5.2.4 | Comparability

Issue

Depending on the intended use case of a ML model, different metrics give different interpretations of the model performance.

Recommendation

Obligatory performance metrics for certain model types and use cases can be introduced to ensure that models with similar intended uses can be compared.[182] These metrics should be regularly re-assessed to test for potential model drift.

6 | CONCLUSIONS

Overcoming the barriers that exist between applications of ML in dementia research and the translation of models to clinical practice will require a paradigm shift in the way researchers design, implement, and evaluate ML models. To address the challenges of reproducibility and replicability, any code integral to the modeling pipeline needs to be openly available and well documented, to enable the external validity of research results to be assessed. Data should also be made available where possible. To ensure models are interpretable, which in turn will assist with clinical acceptability, researchers need to routinely implement methods that enable clinical users to interrogate ML models and understand how they make their decisions. To improve the generalizability of research findings, during model development researchers should not rely on a single source of data. Instead, models should be developed using multiple, heterogeneous datasets to reduce the risk of bias in model predictions. The comparability of ML models will be improved when researchers adhere to reporting guidelines. There exists a clear opportunity for researchers, clinicians, and other stakeholders to work together to develop guidelines for ML models that are: reproducible, replicable, interpretable, comparable, generalizable, trustworthy, and transferable to clinical practice. If these challenges are overcome, then ML holds promise to change the future landscape of dementia research and care.

ACKNOWLEDGMENTS

With thanks to the Deep Dementia Phenotyping (DEMON) Network State of the Science symposium participants (in alphabetical order): Peter Bagshaw, Robin Borchert, Magda Bucholc, James Duce, Charlotte James, David Llewellyn, Donald Lyall, Sarah Marzi, Danielle Newby, Neil Oxtoby, Janice Ranson, Tim Rittman, Nathan Skene, Eugene Tang, Michele Veldsman, Laura Winchester, Zhi Yao. This paper was the product of a DEMON Network state of the science symposium entitled "Harnessing Data Science and AI in Dementia Research" funded by Alzheimer's Research UK. J.M.R. and D.J.L. are supported by Alzheimer's Research UK and the Alan Turing Institute/Engineering and Physical Sciences Research Council (EP/N510129/1). D.J.L. also receives funding from the Medical Research Council (MR/X005674/1), National Institute for Health Research (NIHR) Applied Research Collaboration South West Peninsula, National Health and Medical Research Council (NHMRC), and National Institute on Aging/National Institutes of Health (RF1AG055654). M.B. is supported by Alzheimer's Research UK, Economic and Social Research Council (ES/W010240/1), EU (SEUPB) INTERREG (ERDF/SEUPB), and HSC R&D (COM/5750/23). This work was additionally supported by Alzheimer's Research UK (C.J.), National Institute for Health and Care Research Bristol Biomedical Research Centre (C.J.), Fonds de recherche du Québec Santé—Chercheur boursiers Junior 1 (A.B.), Canadian Consortium for Neurodegeneration in Aging and the Courtois Foundation (A.B., N.C.), the Motor Neurone Disease Association Fellowship (Al Khleifat/Oct21/975-799) (A.A.K.), ALS Association Milton Safenowitz Research Fellowship (22-PDF-609) (A.A.K.), NIHR Maudsley Biomedical Research Centre (A.A.K.), the

Darby Rimmer Foundation (A.A.K.), UKRI Future Leaders Fellowship (MR/S03546X/1) (C.S.), E-DADS project (EU JPND) (C.S.), EuroPOND project (EU Horizon 2020, no. 666992) (C.S.). S.J.M. is funded by the Edmond and Lily Safra Early Career Fellowship Program and the UK Dementia Research Institute, which receives its funding from UK DRI Ltd., funded by the UK Medical Research Council, Alzheimer's Society, and Alzheimer's Research UK.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest. Author disclosures are available in the supporting information.

REFERENCES

1. Patterson C. World Alzheimer report 2018. 2018.
2. Robinson L, Tang E, Taylor J. Dementia: Timely diagnosis and early intervention. *BMJ*. 2015;350.
3. Pellegrini E, Ballerini L, Maria del C Valdes H, et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. *Alzheimers Dement*. 2018;10:519-535.
4. Ahmed MR, Zhang Y, Feng Z, Lo B, Inan OT, Liao H. Neuroimaging and machine learning for dementia diagnosis: Recent advancements and future prospects. *IEEE Rev Biomed Eng*. 2018;12:19-33.
5. Enshaeifar S, Zoha A, Markides A, et al. Health management and pattern analysis of daily living activities of people with dementia using in-home sensors and machine learning techniques. *PLoS One*. 2018;13(5):e0195605.
6. Mitelpunkt A, Galili T, Kozlovski T, et al. Novel Alzheimer's disease subtypes identified using a data and knowledge driven strategy. *Sci Rep*. 2020;10(1):1-13.
7. Marzi SJ, Schilder B, Nott A, et al. Artificial intelligence for neurodegenerative experimentalmodels. *Alzheimers Dement*. Accepted.
8. Doherty T, Yao Z, Al Khleifat A, et al. Artificial intelligence for dementia drug discovery and trials optimization. *Alzheimers Dement*. Submitted.
9. Bettencourt C, Skene N, Bandres-Ciga S, et al. Artificial intelligence for dementia genetics and omics. *Alzheimers Dement*. Submitted.
10. Winchester LM, Harshfield EL, Shi L, et al. Artificial intelligence for alzheimer's disease and associated dementia biomarkers. *Alzheimers Dement*. Submitted.
11. Borchert RJ, Azevedo T, Badhwar AP, et al. Artificial intelligence for diagnostic and prognostic neuroimaging in dementia: A systematic review. *Alzheimers Dement. 2023. Portico*. doi:10.1002/alz.13412
12. Newby D, Orgeta V, Marshall CR, et al. Artificial intelligence for dementia prevention. *Alzheimers Dement*. Submitted.
13. Lyall DM, Kormilitzin A, Lancaster C, et al. Artificial intelligence for dementia—Applied models and digital health. *Alzheimers Dement*. 2023. Portico. doi:10.1002/alz.13391
14. Hui HYH, Ran AR, Dai JJ, Cheung CY. Deep Reinforcement Learning-Based Retinal Imaging in Alzheimer's Disease: Potential and Perspectives. *J Alzheimers Dis*. 2023;94(1):39-50.
15. Fikry M, Mairittha N, Inoue S. Modelling Reminder System for Dementia by Reinforcement Learning. In Sensor-and Video-Based Activity and Behavior Computing: Proceedings of 3rd International Conference on Activity and Behavior Computing (ABC 2021). 2022. Singapore: Springer Nature Singapore.
16. Kumar S, Oh I, Schindler S, Lai AM, Payne PRO, Gupta A. Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: A systematic literature review. *JAMIA Open*. 2021;4(3):ooab052.
17. Grueso S, Viejo-Sobera R. Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: A systematic review. *Alzheimers Res Ther*. 2021;13(1):1-162.
18. Barnes DE, Covinsky KE, Whitmer RA, Kuller LH, Lopez OL, Yaffe K. Dementia risk indices: A framework for identifying individuals with a high dementia risk. *Alzheimers Dement*. 2010;6(2):138.
19. Chary E, Amieva H, Pérès K, Orgogozo J, Dartigues J, Jacqmin-Gadda H. Short-versus long-term prediction of dementia among subjects with low and high educational levels. *Alzheimers Dement*. 2013;9(5):562-571.
20. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonça A. Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*. 2011;4(1):1-14.
21. Lebedev AV, Westman E, Van Westen G, et al. Random forest ensembles for detection and prediction of alzheimer's disease with a good between-cohort robustness. *NeuroImage: Clinical*. 2014;6:115-125.
22. Wei W, Visweswaran S, Cooper GF. The application of naive bayes model averaging to predict Alzheimer's disease from genome-wide data. *J Am Med Inform Assoc*. 2011;18(4):370-375.
23. Kruthika KR, Maheshappa HD. Alzheimer's Disease Neuroimaging Initiative. Multistage classifier-based approach for Alzheimer's disease prediction and retrieval. *Inf Med Unlock*. 2019;14:34-42.
24. Bucholc M, Ding X, Wang H, et al. A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. *Expert Syst Appl*. 2019;130:157-171.
25. Sharma A, Kaur S, Memon N, Fathima AJ, Ray S, Bhatt MW. Alzheimer's patients detection using support vector machine (SVM) with quantitative analysis. *Neurosci Inform*. 2021;1(3):100012.
26. Moore PJ, Lyons TJ, Gallacher J, Alzheimer's Disease Neuroimaging Initiative. Random forest prediction of Alzheimer's disease using pairwise selection from time series data. *PLoS One*. 2019;14(2):e0211558.
27. Lehmann C, Koenig T, Jelic V, et al. Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). *J Neurosci Methods*. 2007;161(2):342-350.
28. James C, Ranson JM, Everson R, Llewellyn DJ. Performance of machine learning algorithms for predicting progression to dementia in memory clinic patients. *JAMA Netw Open*. 2021;4(12):e2136553.
29. Cui Y, Liu B, Luo S, et al. Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. *PLoS One*. 2011;6(7):e21896.
30. Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J. Alzheimer's Disease Neuroimaging Initiative. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage*. 2015;104:398-412.
31. Grimmer T, Wutz C, Alexopoulos P, et al. Visual versus fully automated analyses of 18F-FDG and amyloid PET for prediction of dementia due to Alzheimer disease in mild cognitive impairment. *J Nucl Med*. 2016;57(2):204-207.
32. Forlenza OV, Radanovic M, Talib LL, et al. Cerebrospinal fluid biomarkers in Alzheimer's disease: Diagnostic accuracy and prediction of dementia. *Alzheimers Dement*. 2015;1(4):455-463.
33. De Velasco Oriol J, Vallejo EE, Estrada K, Tamez Pena JG. Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data. *BMC Bioinf*. 2019;20.
34. Korolev IO, Symonds LL, Bozoki AC. Alzheimer's Disease Neuroimaging Initiative. Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. *PLoS One*. 2016;11(2):e0138866.

35. Prakash M, Abdelaziz M, Zhang L, Strange BA, Tohka J. Alzheimer's Disease Neuroimaging Initiative. Quantitative longitudinal predictions of Alzheimer's disease by multi-modal predictive learning. *J Alzheimer's Dis*. 2020(Preprint):1-14.
36. Duchesne S, Caroli A, Geroldi C, Collins DL, Frisoni GB. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. *Neuroimage*. 2009;47(4):1363-1370.
37. Duchesne S, Caroli A, Geroldi C, Frisoni GB, Collins DL. Predicting clinical variable from MRI features: application to MMSE in MCI. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2005 Oct 26 (pp. 392-399). Springer, Berlin, Heidelberg.
38. Youssofzadeh V, McGuinness B, Maguire LP, Wong-Lin K. Multi-kernel learning with dartel improves combined MRI-PET classification of Alzheimer's disease in AIBL data: group and individual analyses. *Front Hum Neurosci*. 2017;11:380.
39. Mofrad SA, Lundervold A, Lundervold AS. Alzheimer's Disease Neuroimaging Initiative. A predictive framework based on brain volume trajectories enabling early detection of Alzheimer's disease. *Comput Med Imaging Graphics*. 2021;90:101910.
40. Growdon ME, Schultz AP, Dagley AS, et al. Odor identification and Alzheimer disease biomarkers in clinically normal elderly. *Neurology*. 2015;84(21):2153-2160.
41. Wolters FJ, Zonneveld HI, Hofman A, van der Lugt A, Koudstaal PJ, Vernooij MW. Heart-brain connection collaborative research group. cerebral perfusion and the risk of dementia. *Circulation*. 2017;136:719-728.
42. Mattsson N, Zetterberg H, Janelidze S, et al. Plasma tau in Alzheimer disease. *Neurology*. 2016;87(17):1827-1835.
43. Lee S, Viqar F, Zimmerman ME, et al. White matter hyperintensities are a core feature of Alzheimer's disease: Evidence from the dominantly inherited Alzheimer network. *Ann Neurol*. 2016;79(6):929-939.
44. Irwin DJ, Grossman M, Weintraub D, et al. Neuropathological and genetic correlates of survival and dementia onset in synucleinopathies: A retrospective analysis. *Lancet Neurol*. 2017;16(1):55-65.
45. Asanomi Y, Shigemizu D, Akiyama S, Sakurai T, Ozaki K, Ochiya T, Niida S. Dementia subtype prediction models constructed by penalized regression methods for multiclass classification using serum microRNA expression data. *Sci Rep*. 2021;11(1):20947.
46. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health*. 2022;1(12):e0000168.
47. Yang T, Wang J, Sun Q, et al. Detecting genetic risk factors for Alzheimer's disease in whole genome sequence data via lasso screening. 2015:985-989.
48. Dayon L, Guiraud SP, Corthésy J, et al. One-carbon metabolism, cognitive impairment and CSF measures of Alzheimer pathology: Homocysteine and beyond. *Alzheimers Res Ther*. 2017;9(1):1-11.
49. Silver M, Janousova E, Hua X, Thompson PM, Montana G. Alzheimer's Disease Neuroimaging Initiative. Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. *Neuroimage*. 2012;63(3):1681-1694.
50. Teipel SJ, Grothe MJ, Metzger CD, et al. Robust detection of impaired resting state functional connectivity networks in Alzheimer's disease using elastic net regularized regression. *Front Aging Neurosci*. 2017;8:318.
51. Kaut O, Schmitt I, Tost J, et al. Epigenome-wide DNA methylation analysis in siblings and monozygotic twins discordant for sporadic parkinson's disease revealed different epigenetic patterns in peripheral blood mononuclear cells. *Neurogenetics*. 2017;18(1):7-22.
52. Bouts MJ, Möller C, Hafkemeijer A, et al. Single subject classification of Alzheimer's disease and behavioral variant frontotemporal dementia using anatomical, diffusion tensor, and resting-state functional magnetic resonance imaging. *J Alzheimer's Dis*. 2018;62(4):1827-1839.
53. Cleret de Langavant L, Bayen E, Bachoud-Lévi A, Yaffe K. Approximating dementia prevalence in population-based surveys of aging worldwide: An unsupervised machine learning approach. *Alzheimers Dement*. 2020;6(1):e12074.
54. Ranson JM, Kuźma E, Hamilton W, Muniz-Terrera G, Langa KM, Llewellyn DJ. Predictors of dementia misclassification when using brief cognitive assessments. *Neurology: Clinical Practice*. 2019;9(2):109-117.
55. Raket LL. Statistical disease progression modeling in Alzheimer disease. *Front Big Data*. 2020;3.
56. Leoutsakos J, Gross AL, Jones RN, Albert MS, Breitner J. 'Alzheimer's progression score': Development of a biomarker summary outcome for AD prevention trials. *J Prev Alzheimers Dis*. 2016;3(4):229.
57. Lorenzi M, Filippone M, Frisoni GB, Alexander DC, Ourselin S. Alzheimer's Disease Neuroimaging Initiative. Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease. *Neuroimage*. 2019;190:56-68.
58. Firth NC, Primativo S, Brotherhood E, et al. Sequences of cognitive decline in typical Alzheimer's disease and posterior cortical atrophy estimated using a novel event-based model of disease progression. *Alzheimers Dement*. 2020;16(7):965-973.
59. Fonteijn HM, Modat M, Clarkson MJ, et al. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *Neuroimage*. 2012;60(3):1880-1889.
60. Young AL, Oxtoby NP, Daga P, et al. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain*. 2014;137(9):2564-2577.
61. Eshaghi A, Young AL, Wijeratne PA, et al. Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. *Nat Commun*. 2021;12(1):1-12.
62. Vogel JW, Young AL, Oxtoby NP, et al. Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med*. 2021;27(5):871-881.
63. Young AL, Marinescu RV, Oxtoby NP, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nat Commun*. 2018;9(1):1-16.
64. Whitwell JL, Przybelski SA, Weigand SD, et al. Distinct anatomical subtypes of the behavioural variant of frontotemporal dementia: A cluster analysis study. *Brain*. 2009;132(11):2932-2946.
65. Toschi N, Lista S, Baldacci F, et al. Biomarker-guided clustering of Alzheimer's disease clinical syndromes. *Neurobiol Aging*. 2019;83:42-53.
66. Oxtoby NP, Alexander DC. Imaging plus X: Multimodal models of neurodegenerative disease. *Curr Opin Neurol*. 2017;30(4):371.
67. Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM. Perona I. An extensive comparative study of cluster validity indices. *Pattern Recognit*. 2013;46(1):243-256.
68. Khanna S, Domingo-Fernández D, Ayyappan A, Emon MA, Hofmann-Apitius M, Fröhlich H. Using multi-scale genetic, neuroimaging and clinical data for predicting alzheimer's disease and reconstruction of relevant biological mechanisms. *Sci Rep*. 2018;8(1):1-13.
69. Van Engelen JE, Hoos HH. A survey on semi-supervised learning. *Mach Learning*. 2020;109(2):373-440.
70. An L, Adeli E, Liu M, Zhang J, Shen D. Semi-supervised hierarchical multimodal feature and sample selection for alzheimer's disease diagnosis. 2016:79-87.
71. Batmanghelich KN, Dong HY, Pohl KM, Taskar B, Davatzikos C. Disease classification and prediction via semi-supervised dimensionality reduction. 2011:1086-1090.
72. Filipovych R, Davatzikos C. Alzheimer's Disease Neuroimaging Initiative. Semi-supervised pattern classification of medical

images: Application to mild cognitive impairment (MCI). *Neuroimage*. 2011;55(3):1109-1119.
73. Teramoto R. Prediction of Alzheimer's diagnosis using semi-supervised distance metric learning with label propagation. *Comput Biol Chem*. 2008;32(6):438-441.
74. Bengio Y, LeCun Y. Scaling learning algorithms towards AI. *Large-Scale Kernel Machines*. 2007;34(5):1-41.
75. Morabito FC, Campolo M, Ieracitano C, et al. Deep convolutional neural networks for classification of mild cognitive impaired and Alzheimer's disease patients from scalp EEG recordings. 2016:1-6.
76. Spasov S, Passamonti L, Duggento A, Lio P, Toschi N. Alzheimer's Disease Neuroimaging Initiative. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *Neuroimage*. 2019;189:276-287.
77. Mahendran N, Vincent P M DR. Deep belief network-based approach for detecting Alzheimer's disease using the multi-omics data. *Comput Struct Biotechnol J*. 2023;21:1651-1660.
78. Li W, Zhao J, Shen C, Zhang J, et al. Regional Brain Fusion: Graph Convolutional Network for Alzheimer's Disease Prediction and Analysis. *Front Neuroinform*. 2022;16:886365.
79. Alam R, Anderson M, Bankole A, Lach J. Inferring physical agitation in dementia using smartwatch and sequential behavior models. 2018:170-173.
80. Li H, Fan Y. Early prediction of Alzheimer's disease dementia based on baseline hippocampal MRI and 1-year follow-up cognitive measures using deep recurrent neural networks. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) 2019 Apr 8 (pp. 368-371). IEEE.
81. Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: A survey. *Evol Intell*. 2021:1-22.
82. Yang Q, Li X, Ding X, Xu F, Ling Z. Deep learning-based speech analysis for Alzheimer's disease detection: A literature review. *Alzheimers Res Ther*. 2022;14(1):1-6.
83. Baur C, Denner S, Wiestler B, Navab N, Albarqouni S. Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. *Med Image Anal*. 2021;69:101952.
84. Bertini F, Allevi D, Lutero G, et al. An automatic Alzheimer's disease classifier based on spontaneous spoken English. *Comput Speech Lang*. 2022;72:101298.
85. Ithapu VK, Singh V, Okonkwo OC, et al. Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment. *Alzheimer Dementia*. 2015;11(12):1489-1499.
86. Bandyopadhyay S, Wittmayer J, Libon DJ, Tighe P, Price C, Rashidi P. Explainable semi-supervised deep learning shows that dementia is associated with small, avocado-shaped clocks with irregularly placed hands. *Sci Rep*. 2023;13(1):7384.
87. Wang Y, Gu X, Hou W, Zhao M, Sun L, Guo C. Dual Semi-Supervised Learning for Classification of Alzheimer's Disease and Mild Cognitive Impairment Based on Neuropsychological Data. *Brain Sci*. 2023;13(2):306.
88. Ebrahimighahnavieh MA, Luo S, Chiong R. Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. *Comput Methods Programs Biomed*. 2020;187:105242.
89. Klöppel S, Stonnington CM, Chu C, et al. Automatic classification of MR scans in Alzheimer's disease. *Brain*. 2008;131(3):681-689.
90. Burgos N, Colliot O. Machine learning for classification and prediction of brain diseases: Recent advances and upcoming challenges. *Curr Opin Neurol*. 2020;33(4):439-450.
91. Gupta Y, Lama RK, Kwon G, et al. Prediction and classification of Alzheimer's disease based on combined features from apolipoprotein-E genotype, cerebrospinal fluid, MR, and FDG-PET imaging biomarkers. *Front Comput Neurosci*. 2019;13:72.
92. Bron EE, Klein S, Reinke A, Papma JM, Maier-Hein L, Alexander DC, Oxtoby NP. Ten years of image analysis and machine learning competitions in dementia. *Neuroimage*. 2022;253:119083.
93. Yang S, Bornot JMS, Wong-Lin K, Prasad G. M/EEG-Based Bio-Markers to Predict the MCI and Alzheimer's Disease: A Review From the ML Perspective. *IEEE Trans Biomed Eng*. 2019;66(10):2924-2935.
94. Bucholc M, Titarenko S, Ding X, Canavan C, Chen T. A hybrid machine learning approach for prediction of conversion from mild cognitive impairment to dementia. *Expert Syst Appl*. 2023:119541.
95. Lin RH, Wang CC, Tung CW. A machine learning classifier for predicting stable MCI patients using gene biomarkers. *Int J Environ Res Public Health*. 2022;19(8):4839.
96. Lee G, Nho K, Kang B, Sohn KA, Kim D. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Sci Rep*. 2019;9(1):1-2.
97. Ansart M, Epelbaum S, Bassignana G, et al. Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. *Med Image Anal*. 2020:101848.
98. Samper-Gonzalez J, Burgos N, Bottani S, Habert MO, Evgeniou T, Epelbaum S, Colliot O. Reproducible evaluation of methods for predicting progression to Alzheimer's disease from clinical and neuroimaging data. In Medical Imaging 2019: Image Processing 2019 Mar 15 (Vol. 10949, pp. 221-233). SPIE.
99. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health*. 2022;1(12):e0000168.
100. Fraser KC, Fors KL, Kokkinakis D. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Comput Speech Lang*. 2019;53:121-139.
101. Vonk JM, Greving JP, Gudnason V, Launer LJ, Geerlings MI. Dementia risk in the general population: large-scale external validation of prediction models in the AGES-Reykjavik study. *Eur J Epidemiol*. 2021;36(10):1025-1041.
102. Danso SO, Zeng Z, Muniz-Terrera G, Ritchie CW. Developing an explainable machine learning-based personalised dementia risk prediction model: A transfer learning approach with ensemble learning algorithms. *Front Big Data*. 2021;4:21.
103. Antila K, Lötjönen J, Thurfjell L, et al. The PredictAD project: Development of novel biomarkers and analysis software for early diagnosis of the Alzheimer's disease. *Interface Focus*. 2013;3(2):20120072.
104. Bruun M, Frederiksen KS, Rhodius-Meester HF, et al. Impact of a clinical decision support tool on prediction of progression in early-stage dementia: a prospective validation study. *Alzheimer Re s Therapy*. 2019;11(1):1-1.
105. Gorse AD. Diversity in medicinal chemistry space. *Curr Top Med Chem*. 2006;6(1):3-18.
106. David L, Arús-Pous J, Karlsson J, et al. Applications of deep-learning in exploiting large-scale and heterogeneous compound data in industrial pharmaceutical research. *Front Pharmacol*. 2019;10:1303.
107. Lipinski CF, Maltarollo VG, Oliveira PR, da Silva AB, Honorio KM. Advances and perspectives in applying deep learning for drug design and discovery. *Frontiers in Robotics and AI*. 2019;6:108.
108. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347-1358.
109. Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*. 2018;4(2):268-276.
110. Vatansever S, Schlessinger A, Wacker D, et al. Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: State-of-the-arts and future directions. *Med Res Rev*. 2021;41(3):1427-1473.
111. Leung MK, Delong A, Alipanahi B, Frey BJ. Machine learning in genomic medicine: A review of computational problems and data sets. *Proc IEEE*. 2015;104(1):176-197.

112. Dias TL, Schuch V, Beltrão-Braga PCB, et al. Drug repositioning for psychiatric and neurological disorders through a network medicine approach. *Translational Psychiatry*. 2020;10(1):1-10.
113. Rodriguez S, Hug C, Todorov P, et al. Machine learning identifies candidates for drug repurposing in Alzheimer's disease. *Nat Commun*. 2021;12(1):1-3.
114. Sundaram L, Gao H, Padigepati SR, et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet*. 2018;50(8):1161-1170.
115. Dey KK, Van de Geijn B, Kim SS, Hormozdiari F, Kelley DR, Price AL. Evaluating the informativeness of deep learning annotations for human complex diseases. *Nat Commun*. 2020;11(1):1-9.
116. Breiman L. Statistical modeling: The two cultures. *Quality Control and Applied Statistics*. 2003;48(1):81-82.
117. Navarro E, Udine E, de Paiva Lopes K, et al. Discordant transcriptional signatures of mitochondrial genes in Parkinson's disease human myeloid cells. *Biorxiv*. 2020.
118. Grubman A, Chew G, Ouyang JF, et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat Neurosci*. 2019;22(12):2087-2097.
119. Anderson NC, Chen P, Meganathan K, et al. Balancing serendipity and reproducibility: Pluripotent stem cells as experimental systems for intellectual and developmental disorders. *Stem Cell Rep*. 2021.
120. Jiang J, Wang C, Qi R, Fu H, Ma Q. scREAD: A single-cell RNA-seq database for Alzheimer's disease. *iScience*. 2020;23(11):101769.
121. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. *Nat Methods*. 2019;16(8):715-721.
122. Li H, Guan Y. Fast decoding cell type–specific transcription factor binding landscape at single-nucleotide resolution. *Genome Res*. 2021;31(4):721-731.
123. Quang D, Xie X. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*. 2019;166:40-47.
124. Health TL. Guiding better design and reporting of AI-intervention trials. *The Lancet. Digital Health*. 2020;2(10):e493.
125. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. *Trends Pharmacol Sci*. 2019;40(8):577-591.
126. Ezzati A, Lipton RB. Alzheimer's Disease Neuroimaging Initiative. Machine learning predictive models can improve efficacy of clinical trials for Alzheimer's disease. *J Alzheimer's Dis*. 2020;74(1):55-63.
127. Reith FH, Mormino EC, Zaharchuk G. Predicting future amyloid biomarkers in dementia patients with machine learning to improve clinical trial patient selection. *Alzheimer Dementia*. 2021;7(1):e12212.
128. Hane CA, Nori VS, Crown WH, Sanghavi DM, Bleicher P. Predicting onset of dementia using clinical notes and machine learning: case-control study. *JMIR Medical Informatics*. 2020;8(6):e17819.
129. Kumar S, Oh I, Schindler S, Lai AM, Payne PRO, Gupta A. Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review. *JAMIA Open*. 2021;4(3):ooab052.
130. Wilkinson MD, Dumontier M, Aalbersberg IJ, t al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3(1):1-9.
131. Mueller SG, Weiner MW, Thal LJ, et al. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin N Am*. 2005;15(4):869.
132. Arokiasamy P, Bloom D, Lee J, Feeney K, Ozolins M. Longitudinal aging study in india: Vision, design, implementation, and preliminary findings. In: *Aging in asia: Findings from new and emerging data initiatives*. National Academies Press (US); 2012.
133. Beekly DL, Ramos EM, Lee WW, et al. The National Alzheimer's Coordinating Center (NACC) database: the uniform data set. *Alzheimer Dis Assoc Disord*. 2007;21(3):249-258.
134. Matthews PM, Sudlow C. The UK biobank. *Brain*. 2015;138(12):3463-3465.
135. Fukunaga K, Hayes RR. Effects of sample size in classifier design. *IEEE Trans Pattern Anal Mach Intell*. 1989;11(8):873-885.
136. Martin SA, Townend FJ, Barkhof F, Cole JH. Interpretable machine learning for dementia: A systematic review. *Alzheimer Dementia*. 2023;19(5): 2135-2149.
137. Javeed A, Dallora AL, Berglund JS, Ali A, Ali L, Anderberg P. Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. *J Med Syst*. 2023;47(1):17.
138. Archana M, Ramakrishnan S. Detection of Alzheimer disease in MR images using structure tensor. *In Annu Int Conf IEEE Eng Med Biol Soc*. 2014: 1043-1046.
139. Zheng X, Shi J, Li Y, Liu X, Zhang Q. Multi-modality stacked deep polynomial network based feature learning for Alzheimer's disease diagnosis. *In 2016 IEEE 13th international symposium on biomedical imaging (ISBI)*. 2016. 851-854.
140. McEvoy LK, Fennema-Notestine C, Roddey JC, et al. Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment. *Radiology*. 2009;251(1):195-205.
141. Pellegrini E, Ballerini L, Hernandez MD, et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. *Alzheimer Dementia*. 2018;10:519-535.
142. Mueller SG, Weiner MW, Thal LJ, et al. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*. 2005;15(4):869-877.
143. Bucholc M, McClean PL, Bauermeister S, et al. Association of the use of hearing aids with the conversion from mild cognitive impairment to dementia and progression of dementia: a longitudinal retrospective study. *Alzheimer Dementia*. 2021;7(1):e12122.
144. You J, Zhang YR, Wang HF, et al. Development of a novel dementia risk prediction model in the general population: A large, longitudinal, population-based machine-learning study. *EClinicalMedicine*. 2022;53.
145. Durongbhan P, Zhao Y, Chen L, et al. A dementia classification framework using frequency and time-frequency features based on EEG signals. *IEEE Trans Neural Syst Rehabil Eng*. 2019;27(5):826-835.
146. Khoei TT, Ahajjam MA, Hu WC, Kaabouch N. A Deep Learning Multi-Task Approach for the Detection of Alzheimer's Disease in a Longitudinal Study. In 2022 IEEE International Conference on Electro Information Technology (eIT) 2022 (pp. 315-319). IEEE.
147. Benrimoh D, Israel S, Fratila R, et al. ML and AI safety, effectiveness and explainability in healthcare. *Front Big Data*. 2021;4.
148. Bruckert S, Finzel B, Schmid U. The next generation of medical decision support: A roadmap toward transparent expert companions. *Front Artif Intell Appl*. 2020;3:75.
149. Molnar C. *Interpretable machine learning*. Lulu. com; 2020.
150. Gómez-Ramírez J, Ávila-Villanueva M, Fernández-Blázquez MÁ. Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. *Sci Rep*. 2020;10(1):1-15.
151. Cole J, Wood D, Booth T. Visual attention as a model for interpretable neuroimage classification in dementia: Doctor AI: Making computers explain their decisions. *Alzheimer Dementia*. 2020;16:e037351.
152. Dyrba M, Pallath AH, Marzban EN. Comparison of cnn visualization methods to aid model interpretability for detecting Alzheimer's disease. In: *Bildverarbeitung für die medizin 2020*. Springer; 2020:307-312.
153. Qiu S, Joshi PS, Miller MI, et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain*. 2020;143(6):1920-1933.
154. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" explaining the predictions of any classifier. 2016:1135-1144.

155. Shapley LS. *17. A value for n-person games*. Princeton University Press; 2016.
156. Spooner A, Chen E, Sowmya A, et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Sci Rep*. 2020;10(1):1-10.
157. Bansal D, Chhikara R, Khanna K, Gupta P. Comparative analysis of various machine learning algorithms for detecting dementia. *Procedia Computer Science*. 2018;132:1497-1502.
158. Williams JA, Weakley A, Cook DJ, Schmitter-Edgecombe M. Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. 2013.
159. Chen R, Herskovits EH. Machine-learning techniques for building a diagnostic model for very mild dementia. *Neuroimage*. 2010;52(1):234-244.
160. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*. 2018;286(3):800-809.
161. Mårtensson G, Ferreira D, Granberg T, et al. The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study. *Med Image Anal*. 2020;66:101714.
162. Klöppel S, Peter J, Ludl A, et al. Applying automated MR-based diagnostic methods to the memory clinic: A prospective study. *J Alzheimer's Dis*. 2015;47(4):939-954.
163. Jun GR, Chung J, Mez J, et al. Transethnic genome-wide scan identifies novel Alzheimer's disease loci. *Alzheimers Dement*. 2017;13(7):727-738.
164. Kalafatis C, Modarres MH, Apostolou P, et al. Validity and Cultural Generalisability of a 5-Minute AI-Based, Computerised Cognitive Assessment in Mild Cognitive Impairment and Alzheimer's *Dementia. Front Psychiatry*. 2021;12:706695.
165. Clarke N, Foltz P, Garrard P. How to do things with (thousands of) words: Computational approaches to discourse analysis in Alzheimer's disease. *Cortex*. 2020;129:446-463.
166. Fraser KC, Fors KL, Kokkinakis D. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Comput Speech Lang*. 2019;53:121-139.
167. Badhwar A, McFall GP, Sapkota S, et al. A multiomics approach to heterogeneity in Alzheimer's disease: Focused review and roadmap. *Brain*. 2020;143(5):1315-1331.
168. Graham SA, Lee EE, Jeste DV, et al. Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. *Psychiatry Res*. 2020;284:112732.
169. Chandler C, Foltz PW, Cohen AS, et al. Machine learning for ambulatory applications of neuropsychological testing. *Intell Based Med*. 2020;1:100006.
170. El-Sappagh S, Alonso JM, Islam SR, Sultan AM, Kwak KS. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Sci Rep*. 2021;11(1):1-26.
171. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJ, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. *JAMA Netw Open*. 2022;5(9):e2233946.
172. Nation DA, Ho JK, Dutt S, Han SD, Lai MHC; Alzheimer's Disease Neuroimaging Initiative. Neuropsychological Decline Improves Prediction of Dementia Beyond Alzheimer's Disease Biomarker and Mild Cognitive Impairment Diagnoses. *J Alzheimers Dis*. 2019;69(4):1171-1182.
173. Wen J, Thibeau-Sutre E, Diaz-Melo M, et al. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. *Med Image Anal*. 2020;63:101694.
174. Gianattasio KZ, Bennett EE, Wei J, et al. Generalizability of findings from a clinical sample to a community-based sample: A comparison of ADNI and ARIC. *Alzheimers Dement*. 2021;17(8):1265-1276.
175. Steenland K, Macneil J, Bartell S, Lah J. Analyses of diagnostic patterns at 30 Alzheimer's disease centers in the US. *Neuroepidemiology*. 2010;35(1):19-27.
176. Dodge HH, Zhu J, Woltjer R, et al. Risk of incident clinical diagnosis of Alzheimer's disease–type dementia attributable to pathology-confirmed vascular disease. *Alzheimers Dement*. 2017;13:613-623.
177. Takahashi M, Idani T, Fukuda H, Kawashima R, Kitamura M. Application of machine learning method to diagnosis of alzheimer diseases using SPECT brain image. In SICE Annual Conference. 2005.
178. Gianattasio KZ, Ciarleglio A, Power MC. Development of Algorithmic Dementia Ascertainment for Racial/Ethnic Disparities Research in the US Health and Retirement Study. *Epidemiology*. 2020;31(1):126-133.
179. Mehmood A, Yang S, Feng Z, et al. A transfer learning approach for early diagnosis of Alzheimer's disease on MRI images. *Neuroscience*. 2021;460:43-52.
180. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun*. 2020;11(1):1-9.
181. Lecca P. Machine learning for causal inference in biological networks: Perspectives of this challenge. *Front bioinform*. 2021;1:746712.
182. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res*. 2016;18(12):e323.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bucholc M, James C, Khleifat AA, et al. Artificial intelligence for dementia research methods optimization. *Alzheimer's Dement*. 2023;19:5934–5951. https://doi.org/10.1002/alz.13441