Emerging Topics in Life Sciences (2021) **5** 765–777
https://doi.org/10.1042/ETLS20210249

Review Article

Applied machine learning in Alzheimer's disease research: omics, imaging, and clinical data

Ziyi Li[1], Xiaoqian Jiang[2], Yizhuo Wang[1] and Yejin Kim[2]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A.; [2]School of Biomedical Informatics, The University of Texas Health Science Center, Houston, TX, U.S.A.

Correspondence: Ziyi Li (ZLi16@mdanderson.org)

Alzheimer's disease (AD) remains a devastating neurodegenerative disease with few preventive or curative treatments available. Modern technology developments of high-throughput omics platforms and imaging equipment provide unprecedented opportunities to study the etiology and progression of this disease. Meanwhile, the vast amount of data from various modalities, such as genetics, proteomics, transcriptomics, and imaging, as well as clinical features impose great challenges in data integration and analysis. Machine learning (ML) methods offer novel techniques to address high dimensional data, integrate data from different sources, model the etiological and clinical heterogeneity, and discover new biomarkers. These directions have the potential to help us better manage the disease progression and develop novel treatment strategies. This mini-review paper summarizes different ML methods that have been applied to study AD using single-platform or multi-modal data. We review the current state of ML applications for five key directions of AD research: disease classification, drug repurposing, subtyping, progression prediction, and biomarker discovery. This summary provides insights about the current research status of ML-based AD research and highlights potential directions for future research.

Introduction

Alzheimer's disease (AD) impacted more than five million Americans in 2020, which has imposed a huge psychological and economic burden on patients, their families, and society [1]. For many years, only symptomatic treatments were available, as no drug existed to effectively stop or alter disease progression [2]. Recently, the first therapeutic drug, Aduhelm, was approved by the US Food and Drug Administration bringing new hope to those suffering from AD [3]. However, this drug was only shown to effectively reduce amyloid plaque, a protein surrogate for disease outcome [4]. Additional evidence on the actual treatment is still needed to confirm drug efficacy.

As a neurodegenerative disease, the complex nature of AD has been recognized since the very first report on the topic [5,6]. There are at least three layers of complexity to work through while seeking to fully understanding the disease. First, AD is heterogeneous, both etiologically and clinically [7,8]. Many past efforts have tried to delineate the number of AD distinct subtypes, but there is still no consensus [9]. Knowledge of potential subtyping may bring opportunities to identify subject-specific mediation and treatment approaches [10]. Second, AD is a progressive disease with a long prodromal phase [11]. Previous findings show that the disease etiology may start years or even decades before symptom onset [12]. Early diagnosis is especially desirable to manage disease progression [13]. Third, multi-faceted factors are involved in the disease. Numerous studies have recognized that no single genetic or environmental factor has enough accuracy to predict the onset of AD in a clinical setting [14,15].

Recent AD research uses novel technology or multi-modal data to understand the disease from various aspects including genomics, transcriptomics, metabolites, imaging, and clinical features [16,17]. These explorations have transformed our understanding of AD and provided new opportunities to

Received: 9 September 2021
Revised: 5 November 2021
Accepted: 17 November 2021

Version of Record published:
9 December 2021

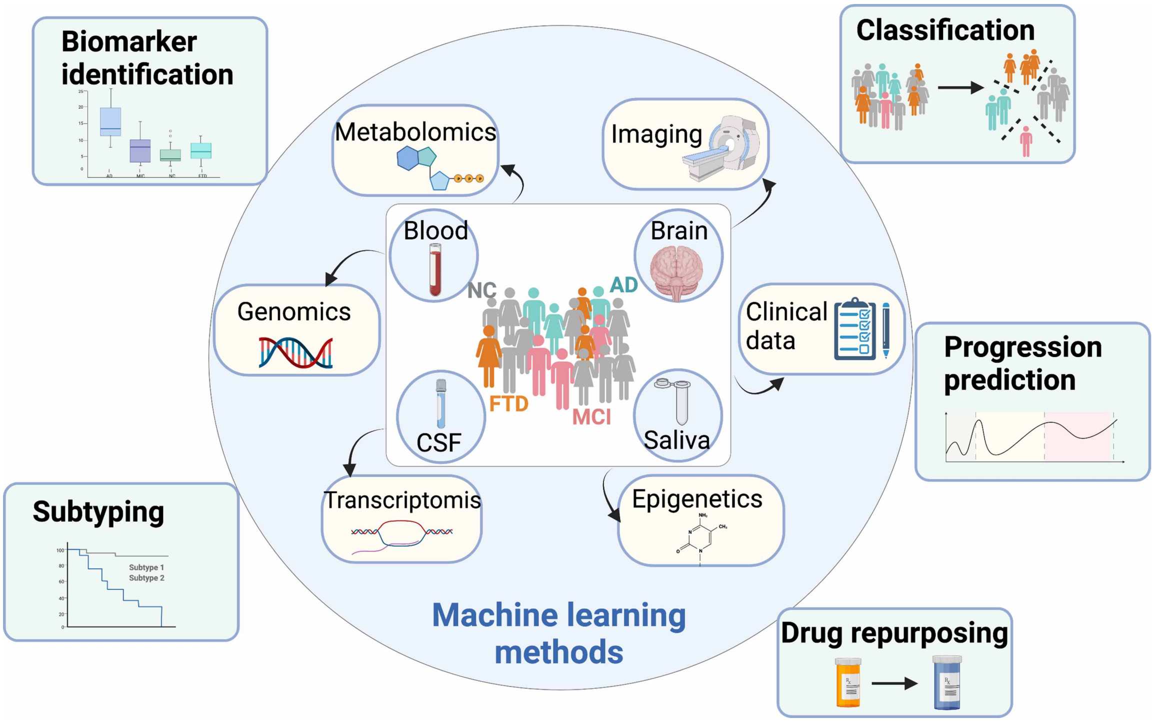

Figure 1. A summary of topics covered in this mini review.

improve our ability to manage disease progression and identify potential treatments. Meanwhile, these data are usually high in volume and data complexity, imposing challenges to data integration and analysis that traditional computation tools may not be able to fully address [18–20].

Machine learning (ML) methods have grown rapidly over recent decades and have been applied widely in the context of precision medicine [21]. The latest developments in deep learning (DL) methods further increase the ability and accuracy of analyzing large-scale complicated data [22,23]. Currently, ML methods have been explored and used in many health-related applications, as reported in cancers [24,25], cardiovascular disease [26], HIV/AIDS [27], and other health-related areas. Reviews are also available for application of ML methods using specific datatypes, for example, single cell RNA-seq [28,29], medical imaging [30–32], and multi-omic data integration [33,34]. In contrast, the application of ML methods in AD are still in their embryonic stage. Due to the complicated nature of AD, however, ML methods have the potential to further improve our understanding of the disease.

This mini-review provides a focused discussion of ML applications to AD using data from one or multi-platforms. To the best of our knowledge, this is the first systematic review of ML methods in AD research covering a wide range of applications. Specifically, we consider the following five aspects of applications (summarized in Figure 1): (i) disease classification, (ii) subtyping, (iii) prediction of disease progression, (iv) biomarker discovery, and (v) drug repurposing. The papers reviewed are also summarized in Supplementary Material S1. We aim to summarize the applied ML methods for each aspect concisely and provide readers a quick head-start in the related direction.

Backgrounds

Machine learning

ML is one of the most common subsets of artificial intelligence concerned with how computers tackle complex learning tasks from past data [35]. As a burgeoning interdisciplinary field, ML is born at the intersection of statistics, which explores general concepts of inference, and computer science, which develops faster programming algorithms [36]. The key difference between conventional statistical methodologies and ML is that the latter draws inference and allows decisions to be made from examples rather than programming explicitly with rules [37].

Based on the nature of desired outcomes provided to the learning algorithms, ML can be supervised [38], unsupervised [39], or reinforced [40]. In supervised learning, the algorithm is presented with a labeled training

Table 1 Summary of publicly available data resources for AD research

Data type	Source
Omics	AD Knowledge Portal (e.g. AMP-AD [41], M2OVE-AD [42], Psych-AD [43], ROSMAP [44], National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) of GSE5281 [45], GSE36980 [46]
Nominated target and drug repurposing	Agora [47], AD Atlas [48], DRIAD [49]
Preclinical efficacy data	AlzPED [50]
Real-world patient data (neuropsychological tests and imaging)	ADNI [18], National Alzheimer's Coordinating Center or NACC [51], OASIS [52,53], DementiaBank [54], CCC [55], TADPOLE [56]
Knowledge repository	AMP-AD [41], AlzForum [57]

set and aims to find a pattern mapping the input data to the output data. In unsupervised learning, the training set is unlabeled and only contains the input data. In other words, the algorithm does not predict output data, but rather aims to find unique structures in the input data. In reinforcement learning, the algorithm is trained in a dynamic environment and is being taught by a rewards program. It gets 'rewarded' for correct decisions and 'punished' for wrong decisions. In this way, the algorithm learns from experience instead of data.

Machines generally take more examples than humans to learn the same task, as machines lack common sense. On the other hand, machines can process a massive wealth of data. In this respect, ML algorithms advance in taking in tens of thousands of clinical data stored in electronic health records (EHRs) and hundreds of millions of genomic data, as well as imaging data generated from the laboratory experiments [36].

Available data resources for AD research

The huge economic and psychological burden, as well as the serious damage done by AD cannot be ignored. Over decades, research communities have collected data from AD or asymptotic AD patients to investigate disease-associated factors. In Table 1, we provide a summary of available data resources for AD research. This is not an exhaustive collection, but a list of some commonly used data sources to provide some references for researchers new to the field. To facilitate diverse therapeutic target identification, various omics data platforms, including AMP-AD, M2OVE-AD, Psych-AD, and ROSMAP, provide transcriptomic and genomic variants obtained from animal models and human cohorts. Agora is a data platform where AD researchers nominate such therapeutic targets. Drug repurposing data resources, such as AD Atlas and DRIAD, provide online tool to identify repurposable FDA-approved drugs via network analysis and enrichment analysis. Potential drugs identified by genomic approach are to be tested using *in vivo* or *in vitro* models and AlzPED is the database that stores the *in vivo* efficacy of drug candidates. From clinical perspective, several consortiums, such as ADNI and NACC, have collected real-world data of cognitive normal, prodromal, and AD patients with imaging, blood-based biomarkers, and neuropsychological tests.

ML tasks in AD research

Disease classification

AD patients usually have a long prodromal phase when effective treatment strategies may be applied to delay or alter the onset of the symptoms. To effectively and accurately identify AD patients or the subjects who are at high risk of developing AD, a series of studies have been conducted to classify AD from mild cognitive impairment (MCI) and from healthy controls.

Imaging data, including magnetic resonance imaging (MRI), positron emission tomography (PET) and electroencephalography (EEG), are the most commonly used data types for classification studies. Early works to classify AD patients from normal controls mainly adopt traditional ML methods, such as support vector machine [58], multi-layer perceptron [58,59], autoencoder [60], and convolutional neural network [61,62]. These methods generally can achieve a classification accuracy ~0.9. A few modified versions of traditional methods, such as Bayesian Gaussian process logistic regression [63,64] and elastic net regularized logistic regression [65], have also demonstrated favorable performance with classification accuracies ~0.95. A review paper by Khan and Usman [66] provides a summary of 11 papers using ML for early diagnosis of AD, including some of the papers presented here.

With the booming development of DL, researchers also have started to apply related techniques to classifying imaging data for AD diagnosis. For example, DeepAD [67] used the Inception architecture on MRI data to achieve a prediction accuracy of more than 0.98. The Inception architecture was originally built by Google, and it can learn the non-linear function by changing how convolutional layers are connected [68]. Along the same line, Hon and Khan [69] adapted two additional popular DL architectures, VEGG16 and InceptionV4, to MRI data. Additionally, they innovatively used transfer learning to greatly reduce the required training size to ∼10% of that in Szegedy et al. [68], while still achieving a comparable performance. These models are typically designed for 2D data only. Currently many studies work on extending CNNs to 3D data which is more common in neuroimaging (MRI, PET) [70].

ML methods also have been used for analyzing neuropsychological data, such as the acoustic, semantic, and syntactic elements of speech records. When extracted features are available, methods have been applied to classify AD patients from controls with an accuracy of ∼0.80; these methods include decision tree [55], support vector machine [71], and random forest [72]. When raw language text is of interest, ML methods like decision tree and bagging have been applied, but they only achieved an accuracy of ∼0.83 [73,74]. Conversely, DL models showed their advantage by obtaining more than 0.90 accuracy in the same settings. Example applications include deep-deep neural network language model [55,75] and convolutional neural network-long short-term memory model [76]. Lyu [77] provided a more detailed review for the application of ML methods in neuropsychological data from AD patients.

Another advantage of ML and DL methods is the ability to integrate data from multiple platforms for disease classification. Previous works that used multi-modal data include combining different platforms of imaging data, e.g. MRI and PET by stacked autoencoder [63,78], MRI and FDG-PET by deep neural network [79]; combining imaging data and patient features, e.g. MRI and cerebrospinal fluid markers by support vector machine [80], MRI and clinical features by local weighted learning [81] and by XGBoost [81,82], MRI and neuropsychological data as well as biomarkers by multi-task deep neural network [83]; and combining data from multiple omics platforms, e.g. gene expression and DNA methylation data by deep neural network [84].

A review of the above classification studies reveals all of the above classification studies reveals that integrating multiple data sources does not necessarily provide higher accuracy than using a single data platform. Studies using imaging data, however, do tend to have better prediction performance than those without imaging data. Likewise, DL models tend to produce higher accuracy. Also of note, classifying MCI patients from controls or AD from MCI are harder problems than classifying AD from controls, which always has lower accuracy (0.76 ∼ 0.87).

Drug repurposing

Current computational AD drug repurposing has been studied from various perspectives: transcriptome, network pharmacology, and treatment effects in real-world patient observation [85].

Drug-induced gene expression

The transcriptomic-based strategy to drug repurposing compares drug-induced gene expression with AD gene expression [86–88]. This approach captures integrated molecular changes in AD pathology. Such methods focus on the genetic signature of drugs and disease to investigate the association between drug-induced perturbation and the disease [89–91]. Williams et al. [87] first applied the genetic signature approach to discover drugs that oppose disease-associated genes in neuronal cells. Rodriguez et al. [92] extended the work to the disease's genetic signatures from various disease stages (e.g. Braak stage) and calculated the association between a gene expression on drug-induced perturbation of neuronal cells and molecular changes in the brains of AD patients at different stages.

Network pharmacology

The network-based approach represents drugs' multi-target capacity in a human interaction network and aims to estimate proximity between disease modules and drugs [93–97]. This tactic can facilitate drug repurposing by helping to identify targets and drug–target interaction prediction.

For target identification, several efforts integrate multi-omics data (e.g. metabolites, proteins, epigenetic modification, and GWAS catalog) by integrating multiple biological interactions [14,98,99]. Another line of studies uses a broader set of data associated with drugs (e.g. side effect, pharmacological hierarchy) and leverages knowledge graph representation to identify AD-related genes [100,101]. In particular, several platforms curate the multi-modal and comprehensive interactions from public data collections and experimental data generated by multiple consortiums [60,101,102].

Prior studies on drug–target interaction prediction have aimed to identify hidden interactions among drugs and proteins (i.e. putative disease target). By identifying hidden drug–target interaction, it is possible to identify existing drugs that may have new indications for AD. Hidden drug–target interaction can be revealed by finding new drug–target binding (off-target) or by integrating multi-modal interactions (on-target) [103]. The off-target approach uses biochemical properties (structural, ligand-based molecular docking) or biophysical properties (3D conformation) to predict drug–target binding [85,104]. The on-target approach uses protein–protein interactions or drug–drug similarity to estimate network proximity between entities. Predicting drug–target interactions can be facilitated by a large biological knowledgebase with multi-modal interactions (e.g. drug–target, target–target, drug–disease, drug–side effect, drug–drug, gene–functional ontology, drug–functional ontology) and use graph ML (e.g. network proximity, graph neural representation) [93,101,105–107].

Population-based treatment effect

This approach, based on real-world patient data, leverages large-scale patient datasets to obtain off-the-label drug efficacy via counterfactual inference [108,109]. Real-world patient data includes administrative data (EHRs, insurance claim data), clinical observational data, and clinical trials data. Several statistics and ML methodologies are applied to the patient data, such as potential treatment outcome models (or target trial) [110], and meta-analysis [108]. The techniques from causal inference, such as propensity score matching, have also been used in potential outcomes or target trial approaches [109]. The merit of the population-based approach is that it captures different drug responses in heterogeneous populations.

Subtyping of AD

The purpose of AD/ADRD patient subtyping is to use computational approach to mine big healthcare data to identify clinically homogeneous group of patients based on their characteristics and biological markers, considering existing biomedical knowledge and clues derived from data. AD is complicated by different etiologies and a large variability of patient characteristics. Depending on comorbidity, genotype, race, and gender, different patients exhibit different degradation pathways. There is no one-size-fits-all solution to model the complexity of the patient population. AD subtyping with ML is a current research topic, and various researchers have applied different methods to gain a better understanding of the complex and heterogeneous patient population.

Alexander et al. [111] used UK primary care EHRs from the CALIBER resource to identify and characterize clinically meaningful clusters of patients using unsupervised learning approaches of multiple correspondence analysis (MCA) and K-means. Vizcarra et al. [112] validated previously published ML algorithms using convolutional neural networks (CNNs) and to determine if pathological heterogeneity may alter algorithm-derived measures using 40 cases from the Goizueta Emory Alzheimer's Disease Center brain bank, which displays an array of pathological diagnoses (including AD with and without Lewy body disease (LBD) and/or TDP-43-positive inclusions) and evaluated their levels of Aβ pathologies. Shehzad et al. [113] used individualized symptom profiles from the pooled data (clinical data from 717 people from three sources: (1) a memory clinic, (2) long-term care, and (3) an open-label trial of donepezil in vascular and mixed dementia) to train various ML models to predict dementia severity (MCI, mild dementia, moderate dementia, or severe dementia). Tsao et al. [114] combined a predictive multi-task ML method (cFSGL) with a novel ML-based multivariate morphometric surface map of the hippocampus (mTBM) to predict future cognitive scores (Alzheimer's Disease Assessment Scale cognitive scores 6, 12, 24, 36, and 48 months from baseline) of patients. Giang, Nguyen, and Tran [115] proposed a fast-multiple kernel learning framework, referred to as fMKL-DR, to optimize equations to calculate matrix chain multiplication and reduce dimensions in data space to stratify AD patients into different phases. Mar et al. [116] validated random forest models by using them to identify depressive and psychotic clusters according to their presence in the EHRs of all patients diagnosed with dementia.

Various AD subtyping studies have used different data modalities, including imaging data, clinical records, clinical notes, cognitive scores, and genetic profiles, to offer partial evidence of patient stratification, but none of these have provided deterministic characterization or a biomarker that allows researchers to separate patient populations (e.g. into fast and slow progressors). Such challenges have prevented the development of targeted clinical trials and hamper personalized health care.

Prediction of disease progression

Predicting disease progression has two unique tasks. The first is to identify MCI or normal patients who are at higher risk of converting to AD. The second is to predict longitudinal AD-related scores. The first task is similar to the classification problem discussed in Section 2.1. However, the methods in that section usually took advantage of longitudinal observations and had a special focus on disease progression. Some studies only used baseline information to predict the MCI-to-AD conversion in the future and the ML methods they adopted include support vector machine with linear kernel [117], multi-task neural network classifier [83], logistic regression [118] and multi-kernel learning [119]. When longitudinal measurements, such as lab results and cognitive tests, were available, the problem of predicting disease progression was more complicated. A few recent works addressed this problem using advanced DL methods, including conditional restricted Boltzmann machine [120], recurrent neural network [121], and ensemble model based on stacked convolutional neural network and bidirectional long short-term memory network [122].

For the second task, a pioneering work in 2016 applied nonlinear supervised sparse regression-based random forest on the MRI data from the ADNI to predict a variety of longitudinal AD clinical scores [123]. Another recent work incorporated the multi-modal data from MRI, PET, and FDG-PET with support vector machines and predicted rates of decline in patients' global cognition and memory [124].

Due to the complexity of the longitudinal data and the problem, most of the methods reviewed in this section were developed using multi-source data. For example, combining imaging information (MRI, PET) and cerebrospinal fluid markers [117,118,125]. Clinical biomarkers (e.g. lab results, neuropsychological) have also been analyzed together with imaging data [83]. Although statistical methods are not the focus of the current review, researchers also have adopted more complicated statistical modeling to accommodate the longitudinal observations and predict patients' progression [126].

Biomarker discovery

Identifying novel biomarkers to distinguish AD from MCI or normal controls is highly associated with classification and progression prediction. In fact, the first step of establishing a classification or prediction model with a large dataset is usually to select a set of informative biomarkers. For example, Challis et al. [64] selected features with the largest absolute Kendall tau correlation coefficients versus the class label. When constructing the Gaussian process logistic regression, they also applied automatic relevance determination parameterizations to down weight the contribution of less relevant features. In the study using multi-source datasets, it is even more important to perform feature selection so that a parsimonious model can be established. An example of such a study is Zhang and Shen [127]. The authors used a multi-task feature selection that selects the common subset of relevant features from each modality. Then the joint set of selected features were pooled together for later steps of classification analysis. Similarly, Park and Park [84] applied differential analysis on gene expression and DNA methylation datasets, respectively, and selected the top differentially expressed signals as the features. This feature selection procedure facilitates the establishment of deep neural network models to classify AD and normal controls. However, linking these selected features to 'true' biomarkers that can be used for clinical utility is still a hard problem and needs further exploration.

Motivated by these limitations, some studies also aimed to identify optimal combinations of existing biomarkers for disease progression [128–130]. Most of these methods used straight-forward models such as linear regressions or logistic regressions. Some other methods used more complicated tools. For example, Szalkai et al. [131] used association rule mining and Karaglani et al. [132] used an automatic ML pipeline of SVMs to identify optimal combinations of biomarkers. Some of these biomarker identification methods have been reviewed by Chang et al. [133]. Badhwar et al. [14] also provided reviews on the biomarker-related methods to identify imaging, metabolomics and genomics biomarkers. Those studies mainly used straight-forward methods such as logistic regression and SVM to evaluate different biomarker candidates [134,135].

Challenges and opportunities in AD research using ML

Heterogeneity

Individuals diagnosed with AD usually demonstrate a high level of heterogeneity in clinical trajectory, symptoms, as well as neurodegenerative biomarkers. Such heterogeneity is highly associated with the etiology of the disease, and thus it is important to take such heterogeneity into consideration for each analytical task. Some existing works have already recognized this and focused on a sub-group of AD patients with more homogenized clinical features [136–140]. The existing datasets, such as ADNI and ROSMAP, provided data for relatively large AD sample groups. However, when focusing on subgroups with specific clinical characteristics, the sample size diminishes. Compared with distinguishing AD from controls, the tasks of comparing MCI versus control or MCI versus AD are more difficult and need more samples to properly train ML methods. Moreover, the current data from different sources tend to have incompatible formats and various qualities. There is a need to collect more data to address these questions, especially more high-quality and integrated data.

In addition to increasing study sample size, another way to address heterogeneity is to use multi-omics or multi-modal data. Different sources of biomarkers (e.g. imaging features, clinical measurements, omics) provide information from different aspects. Many existing studies have already incorporated data from multiple imaging platforms or imaging data with clinical features [82,124]. However, limited studies have used multi-omics data or combined multi-omics with imaging and clinical features [14]. The power of DL in handling high-volume and multi-source data has been demonstrated in many other scenarios [141,142]. With the accumulation of data and the increasing number of available biomarker sources, incorporating multi-source data with DL models to further improve classification and predictive ability could be a promising future direction.

Early diagnosis and identification of alterable factors

As a progressive disease with a long prodromal phase, early diagnosis and subtyping can guide clinical decision making and improve prognostic outcomes. Limited by the available resources, the majority of existing studies focus on studying AD versus controls, while only a few study control versus MCI or MCI versus AD. Admittedly, the latter two are harder problems, but still there is significant room to improve the accuracy of the existing models for distinguishing MCI versus controls or MCI versus AD for increased clinical utility [127].

In addition, identifying alterable factors is an important task that may substantially influence AD treatment and prognosis. While existing studies demonstrate that ML models using imaging data generally have higher diagnosis accuracy than using other data sources, the omics data types (e.g. genomics, metabolomics, as well epigenetics) actually provide a better chance to identify alterable gene/pathway markers or metabolite targets [84,143]. This could be another promising direction for ML and DL to play an important role in future research.

Integrating multi-source data

Researchers integrate different sources of data when targeting different outcomes. For example, the prediction and classification studies integrate data from multiple imaging platforms [63,78] or integrate imaging data with clinical or omics data [81,82]. The drug repurposing studies more often use data from multiple omics modalities to depict the relevant biological processes [14,98,99]. Assembling the multi-source data expands the view from a single platform and may generate a more comprehensive understanding of the research objective.

However, many challenges exist for such data integration. First, combining data from different sources requires unifying the data format and controlling the data quality. To address these needs, a series of data harmonization guidelines and tools have been developed [144–146]. Most of these tools are developed in recent years and still need further validation to understand their performance. Second, more analytical methods are needed to provide rigorous data integration and analysis. A general data integration method usually requires extensive feature selection from each platform to reduce the computational burden in the data integrating step [80,127]. Usually, the feature selection is performed in an *ad hoc* way and may become a hurdle for having reproducible findings. The high volume of data also poses challenges for data storage and data sharing. Some recent reviews in this area provided more in-depth discussions [14,20].

Last but not least, incorporating sparse or incomplete datasets is another challenge often present in addressing multi-modal data analysis. For example, how to train machine learning modeling using data with missing or censored values? How to integrate high-dimensional imaging or omic data that were not collected from all of the participants. Although some efforts have been made to provide a solution [147,148], more methods are still needed to have a consensus answer across the research communities.

Accurate identification of AD from real-world patient data

Several subtyping or drug repurposing studies have utilized administrative real-world patient data such as EHRs or insurance claim data. In spite of the EHRs' advantages of extensively covering a population's long medical history in breadth, EHRs are mainly collected for billing, not for scholarly studies, and therefore diagnosis billing codes in EHRs are sometimes incomplete and lack details, which consequently make accurate AD

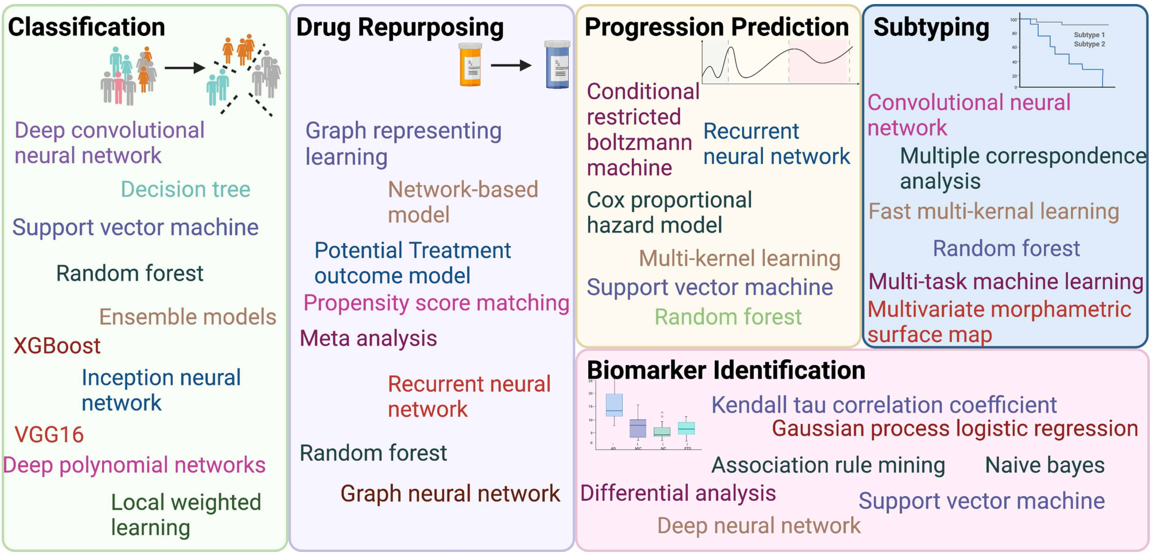

Figure 2. Machine learning methods applied to Alzheimer's disease research.

identification difficult compared with observational studies with neuropsychological or imaging-based diagnosis. Differential diagnosis with related dementia (e.g. vascular dementia, Lewy body dementia, frontotemporal dementia) is also practically challenging due to heterogeneity of AD. Identifying computable AD phenotypes based on multi-modal information (e.g. co-occurrence with medication or imaging procedure, clinical notes) could help better detection of AD.

Reproducibility of the results

Almost all of the existing studies applied the proposed methods to only one or two datasets. It is quite difficult to compare performance (e.g. accuracy, sensitivity, specificity) from studies using different datasets. Moreover, the methods tailored for existing datasets may not perform well in real clinical settings or heterogeneous populations [149]. Therefore, there is a need to systematically benchmark existing methods or future proposals and to collect more datasets that contain diverse populations from real clinical settings.

Lastly, more than half of the methodology works we reviewed in this paper did not provide their software in publicly available repositories. We believe promoting software availability is crucial in reproducible research. At the same time, a user-friendly software allows more researchers and clinicians to apply the methods in practice.

Conclusion

This review provides a concise summary of ML methods applied in AD research. Figure 2 summarizes the ML methods used for different AD research areas. Written for an interdisciplinary audience, the goal of the paper is to provide up-to-date information with recent advances, useful overviews and emerging trends for the readers. Over the past decade, the field quickly embraces the power of ML for complex data analysis and integration. There is also an increasing trend using deep learning techniques to mine the high-volume and high-complicated data in AD research.

Summary

- Massive amounts of data from different platforms for AD research pose great challenges for data analysis.
- A wide variety of ML and DL methods have been applied to classify and subtype patients, predict progression, identify biomarkers, and explore drug repurposing.

- Further efforts are needed to expand the current datasets, incorporate heterogeneity into the analysis, develop methods for addressing issues in multi-source data integration, and apply the findings in real clinical practice.

Competing Interests

The authors declare that there are no competing interests associated with the manuscript.

Author Contributions

Z.L., Y.K., and X.J. conceived the structure. All the authors drafted the manuscript and approved the final version.

Abbreviations

AD, Alzheimer's disease; CNNs, convolutional neural networks; DL, deep learning; EHRs, electronic health records; MCI, mild cognitive impairment; ML, machine learning; MRI, magnetic resonance imaging; PET, positron emission tomography.

References

1 CDC Aging Report: https://www.cdc.gov/aging/aginginfo/alzheimers.htm

2 Yiannopoulou, K.G. and Papageorgiou, S.G. (2013) Current and future treatments for Alzheimer's disease. *Ther. Adv. Neurol. Disord.* **6**, 19–33 https://doi.org/10.1177/1756285612461679

3 Tanzi, R.E. (2021) FDA approval of aduhelm paves a new path for Alzheimer's disease. *ACS Chem. Neurosci.* **12**, 2714–2715 https://doi.org/10.1021/acschemneuro.1c00394

4 Karlawish, J. and Grill, J.D. (2021) The approval of Aduhelm risks eroding public trust in Alzheimer research and the FDA. *Nat. Rev. Neurol.* **17**, 523–524 https://doi.org/10.1038/s41582-021-00540-6

5 O'connor, S.D., Prusiner, S. and Dychtwald, K. (2010) The age of Alzheimer's. *Age* **6**, 39 http://tony-silva.com/eslefl/miscstudent/downloadpagearticles/ageofalzheimers-nyt.pdf

6 Ferraro, A. and Jervis, G.A. (1941) Alzheimer's disease. *Psychiatr. Q.* **15**, 3–16 https://doi.org/10.1007/BF01613948

7 Lam, B., Masellis, M., Freedman, M., Stuss, D.T. and Black, S.E. (2013) Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome. *Alzheimers Res. Ther.* **5**, 1 https://doi.org/10.1186/alzrt155

8 Cummings, J.L. (2000) Cognitive and behavioral heterogeneity in Alzheimer's disease: seeking the neurobiological basis. *Neurobiol. Aging* **21**, 845–861 https://doi.org/10.1016/S0197-4580(00)00183-4

9 Jack, Jr, C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., Haeberlein, S.B. et al. (2018) NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement.* **14**, 535–562 https://doi.org/10.1016/j.jalz.2018.02.018

10 Mohanty, R., Mårtensson, G., Poulakis, K., Muehlboeck, J.S., Rodriguez-Vieitez, E., Chiotis, K. et al. (2020) Comparison of subtyping methods for neuroimaging studies in Alzheimer's disease: a call for harmonization. *Brain Commun.* **2**, fcaa192 https://doi.org/10.1093/braincomms/fcaa192

11 Amieva, H., Le Goff, M., Millet, X., Orgogozo, J.M., Pérès, K., Barberger-Gateau, P. et al. (2008) Prodromal Alzheimer's disease: successive emergence of the clinical symptoms. *Ann. Neurol.* **64**, 492–498 https://doi.org/10.1002/ana.21509

12 Wilson, R.S., Leurgans, S.E., Boyle, P.A. and Bennett, D.A. (2011) Cognitive decline in prodromal Alzheimer disease and mild cognitive impairment. *Arch. Neurol.* **68**, 351–356 https://doi.org/10.1001/archneurol.2011.31

13 Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W. et al. (2005) Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). *Alzheimers Dement.* **1**, 55–66 https://doi.org/10.1016/j.jalz.2005.06.003

14 Badhwar, A., McFall, G.P., Sapkota, S., Black, S.E., Chertkow, H., Duchesne, S. et al. (2020) A multiomics approach to heterogeneity in Alzheimer's disease: focused review and roadmap. *Brain* **143**, 1315–1331 https://doi.org/10.1093/brain/awz384

15 Nativio, R., Lan, Y., Donahue, G., Sidoli, S., Berson, A., Srinivasan, A.R. et al. (2020) An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease. *Nat. Genet.* **52**, 1024–1035 https://doi.org/10.1038/s41588-020-0696-0

16 Wang, M., Li, A., Sekiya, M., Beckmann, N.D., Quan, X., Schrode, N. et al. (2021) Transformative network modeling of multi-omics data reveals detailed circuits, key regulators, and potential therapeutics for Alzheimer's disease. *Neuron* **109**, 257–72.e14 https://doi.org/10.1016/j.neuron.2020.11.002

17 Clark, C., Dayon, L., Masoodi, M., Bowman, G.L. and Popp, J. (2021) An integrative multi-omics approach reveals new central nervous system pathway alterations in Alzheimer's disease. *Alzheimers Res. Ther.* **13**, 71 https://doi.org/10.1186/s13195-021-00814-7

18 Jack, Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D. et al. (2008) The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* **27**, 685–691 https://doi.org/10.1002/jmri.21049

19 Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J. et al. (2010) Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology* **74**, 201–209 https://doi.org/10.1212/WNL.0b013e3181cb3e25

20 Hasin, Y., Seldin, M. and Lusis, A. (2017) Multi-omics approaches to disease. *Genome Biol.* **18**, 83 https://doi.org/10.1186/s13059-017-1215-1

21 Jordan, M.I. and Mitchell, T.M. (2015) Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 https://doi.org/10.1126/science.aaa8415

22 Yan, L.C., Yoshua, B. and Geoffrey, H. (2015) Deep learning. *Nature* **521**, 436–444 https://doi.org/10.1038/nature14539

23 Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*, MIT Press, 800 p

24 Cruz, J.A. and Wishart, D.S. (2006) Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* **2**, 59–77 https://doi.org/10.1177/117693510600200030

25 Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I. (2015) Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 8–17 https://doi.org/10.1016/j.csbj.2014.11.005

26 Al'Aref, S.J., Anchouche, K., Singh, G., Slomka, P.J., Kolli, K.K., Kumar, A. et al. (2019) Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur. Heart J.* **40**, 1975–1986 https://doi.org/10.1093/eurheartj/ehy404

27 Bisaso, K.R., Anguzu, G.T., Karungi, S.A., Kiragga, A. and Castelnuovo, B. (2017) A survey of machine learning applications in HIV clinical research and care. *Comput. Biol. Med.* **91**, 366–371 https://doi.org/10.1016/j.compbiomed.2017.11.001

28 Petegrosso, R., Li, Z. and Kuang, R. (2020) Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Brief. Bioinform.* **21**, 1209–1223 https://doi.org/10.1093/bib/bbz063

29 Oller-Moreno, S., Kloiber, K., Machart, P. and Bonn, S. (2021) Algorithmic advances in machine learning for single-cell expression analysis. *Curr. Opin. Syst. Biol.* **25**, 27–33 https://doi.org/10.1016/j.coisb.2021.02.002

30 Fu, G.-S., Levin-Schwartz, Y., Lin, Q.-H. and Zhang, D. (2019) Machine learning for medical imaging. *J. Healthc. Eng.* **2019**, 9874591 https://doi.org/10.1155/2019/9874591

31 Shen, D., Wu, G., Zhang, D., Suzuki, K., Wang, F. and Yan, P. (2015) Machine learning in medical imaging. *Comput. Med. Imaging Graph.* **41**, 1–2 https://doi.org/10.1016/j.compmedimag.2015.02.001

32 Wernick, M.N., Yang, Y., Brankov, J.G., Yourganov, G. and Strother, S.C. (2010) Machine learning in medical imaging. *IEEE Signal. Process Mag.* **27**, 25–38 https://doi.org/10.1109/MSP.2010.936730

33 Nicora, G., Vitali, F., Dagliati, A., Geifman, N. and Bellazzi, R. (2020) Integrated multi-Omics analyses in oncology: a review of machine learning methods and tools. *Front. Oncol.* **10**, 1030 https://doi.org/10.3389/fonc.2020.01030

34 Reel, P.S., Reel, S., Pearson, E., Trucco, E. and Jefferson, E. (2021) Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol. Adv.* **49**, 107739 https://doi.org/10.1016/j.biotechadv.2021.107739

35 Mitchell, T.M. (1997) Machine Learning. 414 p

36 Rajkomar, A., Dean, J. and Kohane, I. (2019) Machine learning in medicine. *N. Engl. J. Med.* **380**, 1347–1358 https://doi.org/10.1056/NEJMra1814259

37 Duda, R.O., Hart, P.E. and Stork, D.G. (2012) *Pattern Classification*, John Wiley & Sons, New Jersey, 680 p

38 Vapnik, V.N. (1995) The Nature of Statistical Learning Theory

39 Hinton, G.E. and Sejnowski, T.J. (1999) *Unsupervised Learning: Foundations of Neural Computation*, MIT Press, New York, 398 p

40 Burnetas, A.N. and Katehakis, M.N. (1997) Optimal adaptive policies for markov decision processes. *Math. Operat. Res.* **22**, 222–255 https://doi.org/10.1287/moor.22.1.222

41 Hodes, R.J. and Buckholtz, N. (2016) Accelerating medicines partnership: Alzheimer's disease (AMP-AD) knowledge portal aids Alzheimer's drug discovery through open data sharing. *Expert Opin. Ther Targets* **20**, 389–391 https://doi.org/10.1517/14728222.2016.1135132

42 M2OVE-AD: https://adknowledgeportal.synapse.org/Explore/Programs/DetailsPage?Program=M2OVE-AD

43 Greenwood, A.K., Montgomery, K.S., Kauer, N., Woo, K.H., Leanza, Z.J., Poehlman, W.L. et al. (2020) The AD knowledge portal: a repository for multi-omic data on Alzheimer's disease and aging. *Curr. Protoc. Hum. Genet.* **108**, e105 https://doi.org/10.1002/cphg.105

44 Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J.Z. et al. (2019) Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 https://doi.org/10.1038/s41586-019-1195-2

45 Liang, W.S., Dunckley, T., Beach, T.G., Grover, A., Mastroeni, D., Walker, D.G. et al. (2007) Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiol. Genomics* **28**, 311–322 https://doi.org/10.1152/physiolgenomics.00208.2006

46 Hokama, M., Oka, S., Leon, J., Ninomiya, T., Honda, H., Sasaki, K. et al. (2014) Altered expression of diabetes-related genes in Alzheimer's disease brains: the hisayama study. *Cereb. Cortex* **24**, 2476–2488 https://doi.org/10.1093/cercor/bht101

47 Agora: https://agora.ampadportal.org/genes

48 Atlas A: https://adatlas.org

49 DRIAD: https://labsyspharm.shinyapps.io/DRIAD/

50 AlzPED. https://alzped.nia.nih.gov

51 National Alzheimer's Coordinating Center. https://naccdata.org/

52 Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C. and Buckner, R.L. (2010) Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *J. Cogn Neurosci.* **22**, 2677–2684 https://doi.org/10.1162/jocn.2009.21407

53 Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C. and Buckner, R.L. (2007) Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn Neurosci.* **19**, 1498–1507 https://doi.org/10.1162/jocn.2007.19.9.1498

54 Boller, F. and Becker, J. (2005) *Dementiabank Database Guide*, University of Pittsburgh

55 Guinn, C.I. and Habash, A. (2012) *Language analysis of speakers with dementia of the Alzheimer's type. 2012 AAAI Fall Symposium Series: aaai.org; 2012*

56 Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W. et al. (2018) TADPOLE challenge: prediction of longitudinal evolution in Alzheimer's disease. *arXiv [q-bioPE]*

57 ALZFORUM

58 Zhang, D., Wang, Y., Zhou, L., Yuan, H. and Shen, D. (2011) Alzheimer's disease neuroimaging I. multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* **55**, 856–867 https://doi.org/10.1016/j.neuroimage.2011.01.008

59 Munteanu, C.R., Fernandez-Lozano, C., Mato Abad, V., Pita Fernández, S., Álvarez-Linera, J., Hernández-Tamames, J.A. et al. (2015) Classification of mild cognitive impairment and Alzheimer's disease with machine-learning techniques using 1H magnetic resonance spectroscopy data. *Expert Syst. Appl.* **42**, 6205–6214 https://doi.org/10.1016/j.eswa.2015.03.011

60 Liu, H., Wang, L., Lv, M., Pei, R., Li, P., Pei, Z. et al. (2014) Alzplatform: an Alzheimer's disease domain-specific chemogenomics knowledgebase for polypharmacology and target identification research. *J. Chem. Inform. Model.* **54**, 1050–1060 https://doi.org/10.1021/ci500004h
61 Gupta, A., Ayhan, M. and Maida, A. (2013) Natural Image Bases to Represent Neuroimaging Data. In *Proceedings of the 30th International Conference on Machine Learning* (Dasgupta, S. and McAllester, D., eds), pp. 987–994, PMLR, Atlanta, Georgia
62 Payan, A. and Montana, G. (2015) Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *arXiv [csCV]*
63 Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R. and Feng, D. (2014) *Early diagnosis of Alzheimer's disease with deep learning. 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pp. 1015–1018
64 Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S. and Cercignani, M. (2015) Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage* **112**, 232–243 https://doi.org/10.1016/j.neuroimage.2015.02.037
65 Casanova, R., Barnard, R.T., Gaussoin, S.A., Saldana, S., Hayden, K.M., Manson, J.E. et al. (2018) Using high-dimensional machine learning methods to estimate an anatomical risk factor for Alzheimer's disease across imaging databases. *Neuroimage* **183**, 401–411 https://doi.org/10.1016/j.neuroimage.2018.08.040
66 Khan, A. and Usman, M. (2015) *Early Diagnosis of Alzheimer's Disease using Machine Learning Techniques - A Review Paper. Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*
67 Sarraf, S., DeSouza, D.D., Anderson, J., Tofighi, G., for the Alzheimer's Disease Neuroimaging Initiative (2016) DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *BioRxiv* https://doi.org/10.1101/070441
68 Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A.A. (2017) *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. Thirty-First AAAI Conference on Artificial Intelligence; 2017/2/12: aaai.org*
69 Hon, M. and Khan, N.M. (2017) *Towards Alzheimer's disease classification through transfer learning. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*
70 Feng, C., Elazab, A., Yang, P., Wang, T., Zhou, F., Hu, H. et al. (2019) Deep learning framework for Alzheimer's disease diagnosis via 3D-CNN and FSBi-LSTM. *IEEE Access* **7**, 63605–63618 https://doi.org/10.1109/ACCESS.2019.2913847
71 Orimaye, S.O., Wong, J.S.-M. and Golden, K.J. (2014) *Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality, aclweb.org*, pp. 78–87
72 Yancheva, M. and Rudzicz, F. (2016) *Vector-space topic models for detecting Alzheimer's disease. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*
73 Liu, L., Zhao, S., Chen, H. and Wang, A. (2020) A new machine learning method for identifying Alzheimer's disease. *Simul. Model. Pract. Theory* **99**, 102023 https://doi.org/10.1016/j.simpat.2019.102023
74 Rentoumi, V., Raoufian, L., Ahmed, S., de Jager, C.A. and Garrard, P. (2014) Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *J. Alzheimers Dis.* **42**, S3–17 https://doi.org/10.3233/JAD-140555
75 Orimaye, S.O., Wong, J.S.M. and Fernandez, J.S.G. (2016) Deep-deep neural network language models for predicting mild cognitive impairment. BAI@IJCAI
76 Karlekar, S., Niu, T. and Bansal, M. (2018) *Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies, Volume 2 (Short Papers)*
77 Shi Lyu, G. (2018) *A Review of Alzheimer's Disease Classification Using Neuropsychological Data and Machine Learning. 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*
78 Shi, J., Zheng, X., Li, Y., Zhang, Q. and Ying, S. (2018) Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. *IEEE J. Biomed. Health Inform.* **22**, 173–183 https://doi.org/10.1109/JBHI.2017.2655720
79 Lu, D., Popuri, K., Ding, G.W., Balachandar, R. and Beg, M.F. (2018) Alzheimer's disease neuroimaging I. multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. *Sci. Rep.* **8**, 5697 https://doi.org/10.1038/s41598-018-22871-z
80 Westman, E., Muehlboeck, J.S. and Simmons, A. (2012) Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. *Neuroimage* **62**, 229–238 https://doi.org/10.1016/j.neuroimage.2012.04.056
81 Escudero, J., Ifeachor, E., Zajicek, J.P., Green, C., Shearer, J., Pearson, S. et al. (2013) Machine learning-based method for personalized and cost-effective detection of Alzheimer's disease. *IEEE Trans. Biomed. Eng.* **60**, 164–168 https://doi.org/10.1109/TBME.2012.2212278
82 Bloch, L. and Friedrich, C. (2021) *Developing a Machine Learning Workflow to Explain Black-box Models for Alzheimer's Disease Classification. Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*
83 Spasov, S., Passamonti, L., Duggento, A., Liò, P. and Toschi, N. (2019) Alzheimer's disease neuroimaging I. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *Neuroimage* **189**, 276–287 https://doi.org/10.1016/j.neuroimage.2019.01.031
84 Park, C., Ha, J. and Park, S. (2020) Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. *Expert Syst. Appl.* **140**, 112873 https://doi.org/10.1016/j.eswa.2019.112873
85 Paranjpe, M.D., Taubes, A. and Sirota, M. (2019) Insights into computational drug repurposing for neurodegenerative disease. *Trends Pharmacol. Sci.* **40**, 565–576 https://doi.org/10.1016/j.tips.2019.06.003
86 Siavelis, J.C., Bourdakou, M.M., Athanasiadis, E.I., Spyrou, G.M. and Nikita, K.S. (2016) Bioinformatics methods in drug repurposing for Alzheimer's disease. *Brief. Bioinform.* **17** https://doi.org/10.1093/bib/bbv048
87 Williams, G., Gatt, A., Clarke, E., Corcoran, J., Doherty, P., Chambers, D. et al. (2019) Drug repurposing for Alzheimer's disease based on transcriptional profiling of human iPSC-derived cortical neurons. *Transl. Psychiatry* **9**, 220 https://doi.org/10.1038/s41398-019-0555-x
88 Sirota, M., Dudley, J.T., Kim, J., Chiang, A.P., Morgan, A.A., Sweet-Cordero, A. et al. (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* **3**, 96ra77 https://doi.org/10.1126/scitranslmed.3001318
89 Lamb, J. (2006) The connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 https://doi.org/10.1126/science.1132939

90 Duan, Q., Reid, S.P., Clark, N.R., Wang, Z., Fernandez, N.F., Rouillard, A.D. et al. (2016) L1000CDS: LINCS L1000 characteristic direction signatures search engine. *NPJ Syst. Biol. Appl.* **2**, 16015 https://doi.org/10.1038/npjsba.2016.15

91 Wang, Z., Lachmann, A., Keenan, A.B. and Ma'ayan, A. (2018) L1000FWD: fireworks visualization of drug-induced transcriptomic signatures. *Bioinformatics* **34**, 2150–2152 https://doi.org/10.1093/bioinformatics/bty060

92 Rodriguez, S., Hug, C., Todorov, P., Moret, N., Boswell, S.A., Evans, K. et al. (2021) Machine learning identifies candidates for drug repurposing in Alzheimer's disease. *Nat. Commun.* **12**, 1–13 https://doi.org/10.1038/s41467-020-20314-w

93 Fang, J., Pieper, A.A., Nussinov, R., Lee, G., Bekris, L., Leverenz, J.B. et al. (2020) Harnessing endophenotypes and network medicine for Alzheimer's drug repurposing. *Med. Res. Rev.* **40**, 2386–2426 https://doi.org/10.1002/med.21709

94 Pham, T.-H., Qiu, Y., Zeng, J., Xie, L. and Zhang, P. (2020) A deep learning framework for high-throughput mechanism-driven phenotype compound screening. *bioRxiv*

95 Huang, L., Li, F., Sheng, J., Xia, X., Ma, J., Zhan, M. et al. (2014) Drugcomboranker: drug combination discovery based on target network analysis. *Bioinformatics* **30**, i228–i236 https://doi.org/10.1093/bioinformatics/btu278

96 Huang, L., Brunell, D., Stephan, C., Mancuso, J., Yu, X., He, B. et al. (2019) Driver network as a biomarker: systematic integration and network modeling of multi-omics data to derive driver signaling pathways for drug combination prediction. *Bioinformatics* **35**, 3709–3717 https://doi.org/10.1093/bioinformatics/btz109

97 Chen, Y. and Xu, R. (2019) Context-sensitive network analysis identifies food metabolites associated with Alzheimer's disease: an exploratory study. *BMC Med. Genomics* **12**, 133–142 https://doi.org/10.1186/s12920-019-0583-7

98 Zhang, M., Schmitt-Ulms, G., Sato, C., Xi, Z., Zhang, Y., Zhou, Y. et al. (2016) Drug repositioning for Alzheimer's disease based on systematic 'omics' data mining. *PLoS ONE* **11**, e0168812 https://doi.org/10.1371/journal.pone.0168812

99 Sancesario, G.M. and Bernardini, S. (2018) Alzheimer's disease in the omics era. *Clin. Biochem.* **59**, 9–16 https://doi.org/10.1016/j.clinbiochem.2018.06.011

100 Nguyen, T.-P., Priami, C. and Caberlotto, L. (2015) Novel drug target identification for the treatment of dementia using multi-relational association mining. *Sci. Rep.* **5**, 1–13 https://doi.org/10.1038/srep11104

101 Zhou, Y., Fang, J., Bekris, L.M., Kim, Y.H., Pieper, A.A., Leverenz, J.B. et al. (2021) AlzGPS: a genome-wide positioning systems platform to catalyze multi-omics for Alzheimer's drug discovery. *Alzheimers Res. Ther.* **13**, 1–13 https://doi.org/10.1186/s13195-020-00736-w

102 Sügis, E., Dauvillier, J., Leontjeva, A., Adler, P., Hindie, V., Moncion, T. et al. (2019) HENA, heterogeneous network-based data set for Alzheimer's disease. *Sci. Data* **6**, 1–18 https://doi.org/10.1038/s41597-019-0152-0

103 Gaudelet, T., Day, B., Jamasb, A.R., Soman, J., Regep, C., Liu, G. et al. (2021) Utilizing graph machine learning within drug discovery and development. *Brief. Bioinform.* **22**, bbab159 https://doi.org/10.1093/bib/bbab159

104 Xie, H., Wen, H., Qin, M., Xia, J., Zhang, D., Liu, L. et al. (2016) In silico drug repositioning for the treatment of Alzheimer's disease using molecular docking and gene expression data. *RSC Adv.* **6**, 98080–98090 https://doi.org/10.1039/C6RA21941A

105 Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R. and Cheng, F. (2019) deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* **35**, 5191–5198 https://doi.org/10.1093/bioinformatics/btz418

106 Issa, N.T., Kruger, J., Wathieu, H., Raja, R., Byers, S.W. and Dakshanamurthy, S. (2016) DrugGenEx-Net: a novel computational platform for systems pharmacology and gene expression-based drug repurposing. *BMC Bioinformatics* **17**, 202 https://doi.org/10.1186/s12859-016-1065-y

107 Nam, Y., Kim, M., Chang, H.-S. and Shin, H. (2019) Drug repurposing with network reinforcement. *BMC Bioinformatics* **20**, 1–10 https://doi.org/10.1186/s12859-018-2565-8

108 Geifman, N., Brinton, R.D., Kennedy, R.E., Schneider, L.S. and Butte, A.J. (2017) Evidence for benefit of statins to modify cognitive decline and risk in Alzheimer's disease. *Alzheimers Res. Ther.* **9**, 10 https://doi.org/10.1186/s13195-017-0237-y

109 Zissimopoulos, J.M., Barthold, D., Brinton, R.D. and Joyce, G. (2017) Sex and race differences in the association between statin use and the incidence of Alzheimer disease. *JAMA Neurol.* **74**, 225–232 https://doi.org/10.1001/jamaneurol.2016.3783

110 Hernán, M.A. and Robins, J.M. (2016) Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* **183**, 758–764 https://doi.org/10.1093/aje/kwv254

111 Alexander, N., Alexander, D.C., Barkhof, F. and Denaxas, S. (2020) Using unsupervised learning to identify clinical subtypes of Alzheimer's disease in electronic health records. *Stud. Health Technol. Inform.* **270**, 499–503 https://doi.org/10.3233/SHTI200210

112 Vizcarra, J.C., Gearing, M., Keiser, M.J., Glass, J.D., Dugger, B.N. and Gutman, D.A. (2020) Validation of machine learning models to detect amyloid pathologies across institutions. *Acta Neuropathol. Commun.* **8**, 59 https://doi.org/10.1186/s40478-020-00927-4

113 Shehzad, A., Rockwood, K., Stanley, J., Dunn, T. and Howlett, S.E. (2020) Use of patient-reported symptoms from an online symptom tracking tool for dementia severity staging: development and validation of a machine learning approach. *J. Med. Internet Res.* **22**, e20840 https://doi.org/10.2196/20840

114 Tsao, S., Gajawelli, N., Zhou, J., Shi, J., Ye, J., Wang, Y. et al. (2017) Feature selective temporal prediction of Alzheimer's disease progression using hippocampus surface morphometry. *Brain Behav.* **7**, e00733 https://doi.org/10.1002/brb3.733

115 Giang, T.-T., Nguyen, T.-P. and Tran, D.-H. (2020) Stratifying patients using fast multiple kernel learning framework: case studies of Alzheimer's disease and cancers. *BMC Med. Inform. Decis. Mak.* **20**, 108 https://doi.org/10.1186/s12911-020-01140-y

116 Mar, J., Gorostiza, A., Arrospide, A., Larrañaga, I., Alberdi, A., Cernuda, C. et al. (2021) Estimation of the epidemiology of dementia and associated neuropsychiatric symptoms by applying machine learning to real-world data. *Rev. Psiquiatr. Salud. Ment. (Engl. Ed.)* https://doi.org/10.1016/j.rpsm.2021.03.001

117 Frölich, L., Peters, O., Lewczuk, P., Gruber, O., Teipel, S.J., Gertz, H.J. et al. (2017) Incremental value of biomarker combinations to predict progression of mild cognitive impairment to Alzheimer's dementia. *Alzheimers Res. Ther.* **9**, 84 https://doi.org/10.1186/s13195-017-0301-7

118 Dickerson, B. (2013) Biomarker-based prediction of progression in MCI: comparison of AD signature and hippocampal volume with spinal fluid amyloid-β and tau. *Front. Aging Neurosci.* **5**, 55 https://doi.org/10.3389/fnagi.2013.00055

119 Hinrichs, C., Singh, V., Xu, G. and Johnson, S.C. (2011) Alzheimers disease neuroimaging I. predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *Neuroimage* **55**, 574–589 https://doi.org/10.1016/j.neuroimage.2010.10.081

120 Fisher, C.K., Smith, A.M. and Walsh, J.R. and Coalition Against Major Diseases; Abbott, Alliance for Aging Research, Alzheimer's Association, Alzheimer's Foundation of America, AstraZeneca Pharmaceuticals LP, Bristol-Myers Squibb Company, Critical Path Institute, CHDI Foundation, Inc., Eli

Lilly and Company, F. Hoffmann-La Roche Ltd, Forest Research Institute, Genentech, Inc., GlaxoSmithKline, Johnson & Johnson, National Health Council, Novartis Pharmaceuticals Corporation, Parkinson's Action Network, Parkinson's Disease Foundation, Pfizer, Inc., sanofi-aventis. Collaborating Organizations: Clinical Data Interchange Standards Consortium (CDISC), Ephibian, Metrum Institute (2019) Machine learning for comprehensive forecasting of Alzheimer's disease progression. *Sci. Rep.* **9**, 13622 https://doi.org/10.1038/s41598-019-49656-2

121 Lee, G., Nho, K., Kang, B. and Sohn, K.-A. (2019) Kim D, for Alzheimer's disease neuroimaging I. predicting Alzheimer's disease progression using multi-modal deep learning approach. *Sci. Rep.* **9**, 1952 https://doi.org/10.1038/s41598-018-37769-z

122 El-Sappagh, S., Abuhmed, T., Riazul Islam, S.M. and Kwak, K.S. (2020) Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. *Neurocomputing* **412**, 197–215 https://doi.org/10.1016/j.neucom.2020.05.087

123 Huang, L., Jin, Y., Gao, Y., Thung, K.-H. and Shen, D. (2016) Alzheimer's disease neuroimaging I. longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. *Neurobiol. Aging* **46**, 180–191 https://doi.org/10.1016/j.neurobiolaging.2016.07.005

124 Franzmeier, N., Koutsouleris, N., Benzinger, T., Goate, A., Karch, C.M., Fagan, A.M. et al. (2020) Predicting sporadic Alzheimer's disease progression via inherited Alzheimer's disease-informed machine-learning. *Alzheimers Dement.* **16**, 501–511 https://doi.org/10.1002/alz.12032

125 Nie, L., Zhang, L., Meng, L., Song, X., Chang, X. and Li, X. (2017) Modeling disease progression via multisource multitask learners: a case study With Alzheimer's disease. *IEEE Trans. Neural Netw. Learn. Syst.* **28**, 1508–1519 https://doi.org/10.1109/TNNLS.2016.2520964

126 Liu, W., Zhang, B., Zhang, Z. and Zhou, X.-H. (2013) Joint modeling of transitional patterns of Alzheimer's disease. *PLoS ONE* **8**, e75487 https://doi.org/10.1371/journal.pone.0075487

127 Zhang, D. and Shen, D. (2012) Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage* **59**, 895–907 https://doi.org/10.1016/j.neuroimage.2011.09.069

128 Wang, G., Zhou, Y., Huang, F.-J., Tang, H.-D., Xu, X.-H., Liu, J.-J. et al. (2014) Plasma metabolite profiles of Alzheimer's disease and mild cognitive impairment. *J. Proteome Res.* **13**, 2649–2658 https://doi.org/10.1021/pr5000895

129 Mapstone, M., Lin, F., Nalls, M.A., Cheema, A.K., Singleton, A.B., Fiandaca, M.S. et al. (2017) What success can teach us about failure: the plasma metabolome of older adults with superior memory and lessons for Alzheimer's disease. *Neurobiol. Aging* **51**, 148–155 https://doi.org/10.1016/j.neurobiolaging.2016.11.007

130 de Leeuw, F.A., Peeters, C.F.W., Kester, M.I., Harms, A.C., Struys, E.A., Hankemeier, T. et al. (2017) Blood-based metabolic signatures in Alzheimer's disease. *Alzheimers Dement.* **8**, 196–207 https://doi.org/10.1016/j.dadm.2017.07.006

131 Szalkai, B., Grolmusz, V.K. and Grolmusz, V.I. (2017) Coalition against major D. identifying combinatorial biomarkers by association rule mining in the CAMD Alzheimer's database. *Arch. Gerontol. Geriatr.* **73**, 300–307 https://doi.org/10.1016/j.archger.2017.08.006

132 Karaglani, M., Gourlia, K., Tsamardinos, I. and Chatzaki, E. (2020) Accurate blood-Based diagnostic biosignatures for Alzheimer's disease via automated machine learning. *J. Clin. Med. Res.* **9**, 3016 https://doi.org/10.3390/jcm9093016

133 Chang, C.-H., Lin, C.-H. and Lane, H.-Y. (2021) Machine learning and novel biomarkers for the diagnosis of Alzheimer's disease. *Int. J. Mol. Sci.* **22**, 2761 https://doi.org/10.3390/ijms22052761

134 Figueira, J., Jonsson, P., Adolfsson, A.N., Adolfsson, R., Nyberg, L. and Öhman, A. (2016) NMR analysis of the human saliva metabolome distinguishes dementia patients from matched controls. *Mol. BioSyst.* **12**, 2562–2571 https://doi.org/10.1039/C6MB00233A

135 Huan, T., Tran, T., Zheng, J., Sapkota, S., MacDonald, S.W., Camicioli, R. et al. (2018) Metabolomics analyses of saliva detect novel biomarkers of Alzheimer's disease. *J. Alzheimers Dis.* **65**, 1401–1416 https://doi.org/10.3233/JAD-180711

136 De Velasco Oriol, J., Vallejo, E.E., Estrada, K. and Taméz Peña, J.G. (2019) Disease neuroimaging initiative TAs. benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data. *BMC Bioinformatics* **20**, 709 https://doi.org/10.1186/s12859-019-3158-x

137 Bahado-Singh, R.O., Vishweswaraiah, S., Aydas, B., Yilmaz, A., Metpally, R.P., Carey, D.J. et al. (2021) Artificial intelligence and leukocyte epigenomics: evaluation and prediction of late-onset Alzheimer's disease. *PLoS ONE* **16**, e0248375 https://doi.org/10.1371/journal.pone.0248375

138 Kim, Y., Jiang, X., Giancardo, L., Pena, D., Bukhbinder, A.S., Amran, A.Y. et al. (2020) Multimodal phenotyping of Alzheimer's disease with longitudinal magnetic resonance imaging and cognitive function data. *Sci. Rep.* **10**, 1–10 https://doi.org/10.1038/s41598-019-56847-4

139 Qin, L., Xu, Q., Li, Z., Chen, L., Li, Y., Yang, N. et al. (2020) Ethnicity-specific and overlapping alterations of brain hydroxymethylome in Alzheimer's disease. *Hum. Mol. Genet.* **29**, 149–158 https://doi.org/10.1093/hmg/ddz273

140 Ling, Y., Upadhyaya, P., Chen, L., Jiang, X. and Kim, Y. (2021) Heterogeneous treatment effect estimation using machine learning for healthcare application: tutorial and benchmark. *arXiv*

141 Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S. and Lew, M.S. (2016) Deep learning for visual understanding: a review. *Neurocomputing* **187**, 27–48 https://doi.org/10.1016/j.neucom.2015.09.116

142 Mater, A.C. and Coote, M.L. (2019) Deep learning in chemistry. *J. Chem. Inf. Model.* **59**, 2545–2559 https://doi.org/10.1021/acs.jcim.9b00266

143 Kim, D.H., Gim, J.-A., Yoon, D., Kim, S. and Kim, H.-S. (2017) Metabolomics and mitochondrial dysfunction in Alzheimer's disease. *Genes Genomics* **39**, 295–300 https://doi.org/10.1007/s13258-016-0494-3

144 Doiron, D., Burton, P., Marcon, Y., Gaye, A., Wolffenbuttel, B.H., Perola, M. et al. (2013) Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerg. Themes Epidemiol.* **10**, 1–8 https://doi.org/10.1186/1742-7622-10-12

145 Fortier, I., Doiron, D., Burton, P. and Raina, P. (2011) Invited commentary: consolidating data harmonization—how to obtain quality and applicability? *Am. J. Epidemiol.* **174**, 261–264 https://doi.org/10.1093/aje/kwr194

146 Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O. et al. (2016) Inter-site and inter-scanner diffusion MRI data harmonization. *NeuroImage* **135**, 311–323 https://doi.org/10.1016/j.neuroimage.2016.04.041

147 Liu, X., Zhu, X., Li, M., Wang, L., Tang, C., Yin, J. et al. (2018) Late fusion incomplete multi-view clustering. *IEEE Trans Pattern Anal. Mach. Intell.* **41**, 2410–2423 https://doi.org/10.1109/TPAMI.2018.2879108

148 Wen, J., Yan, K., Zhang, Z., Xu, Y., Wang, J., Fei, L. et al. (2020) *Adaptive Graph Completion Based Incomplete Multi-View Clustering*, IEEE Transactions on Multimedia

149 Brayne, C. and Davis, D. (2012) Making Alzheimer's and dementia research fit for populations. *Lancet* **380**, 1441–1443 https://doi.org/10.1016/S0140-6736(12)61803-0