An overview of data integration in neuroscience with focus on Alzheimer's Disease

Rosanna Turrisi[*], Margherita Squillario[*], Giulia Abate, Daniela Uberti, Annalisa Barla

***Abstract*—This work represents the first attempt to provide an overview of how to face data integration as the result of a dialogue between neuroscientists and computer scientists. Indeed, data integration is fundamental for studying complex multifactorial diseases, such as the neurodegenerative diseases. This work aims at warning the readers of common pitfalls and critical issues in both medical and data science fields. In this context, we define a road map for data scientists when they first approach the issue of data integration in the biomedical domain, highlighting the challenges that inevitably emerge when dealing with heterogeneous, large-scale and noisy data and proposing possible solutions. Here, we discuss data collection and statistical analysis usually seen as parallel and independent processes, as cross-disciplinary activities. Finally, we provide an exemplary application of data integration to address Alzheimer's Disease (AD), which is the most common multifactorial form of dementia worldwide. We critically discuss the largest and most widely used datasets in AD, and demonstrate how the emergence of machine learning and deep learning methods has had a significant impact on disease's knowledge particularly in the perspective of an early AD diagnosis.**

***Index Terms*—Multimodal data integration; Machine and Deep learning; Multidisciplinary; Neurodegenerative diseases; Alzheimer's Disease.**

I. INTRODUCTION

ONE of the main challenges in neuroscience is the early diagnosis of neurodegenerative diseases (NDs), that are multifactorial diseases particularly difficult to diagnose in their early stages when symptoms are not still evident. In fact, pathological brain changes can take decades before symptoms appear. Due to their complexity, an earlier clinical diagnosis of NDs and, in turn, their management requires the integration of multiple data sources, such as genetic characteristics, clinical conditions, and environmental factors (e.g., education level, lifestyle).

In the last decades, researchers aiming at developing early diagnosis methods of NDs and, in particular, of Alzheimer's Disease (AD) [1], the most prevalent form of dementia have focused on integrating different types of data (e.g., as medical images, genetics, cognitive tests, cerebrospinal fluid (CSF), blood biomarkers). Several efforts have been made towards this direction in terms of collecting more datasets and developing data driven methods. This resulted in numerous databases obtained through different data collection modalities, and in a rich literature on diagnosis algorithms based on data integration [2]–[6]. Nonetheless, the large majority of studies is still far from the real applicability in clinical practice. To tackle this problem effectively, it is essential to foster a strong partnership between medical and data experts. By embracing a multidisciplinary approach, we can give equal importance to both computational and biological components and make a tangible difference in healthcare.

Based on this approach, this manuscript emerges from a dialogue between biomedical and data scientists. Its objective is to offer readers useful insights, address significant challenges that biostatisticians may face when handling NDs multimodal datasets, and enable the creation of a reproducible and reliable Machine Learning (ML) pipeline for NDs.

The reader will be guided through a road map uncovering the characteristics of the optimal multimodal databases, the intrinsic issues of collecting such a corpus especially in the clinical field, and the possible strategies to overcome the limits of real-world datasets. We then address the challenges of fusing multimodal data and we illustrate the approaches to integrate it in a statistical model, reporting their pros and cons.

In the following sections, more attention will be paid to the specific characteristics and initial reasons behind the creation of the main existing AD databases, comparing them from the data scientist's point of view.

Finally, we will explore literature studies which applied statistical methods to diagnose AD or novel risk factors that would allow to set up preventive strategies to delay the disease onset. We emphasize that this is not a systematic review but the main aim is to enlighten how the state-of-the-art evolved during time and compare the performance of the models introduced in the general framework.

The rest of the paper is structured as follows. In Section II, we explore data integration in terms of both data collection and methods design, enlightening the most important characteristics they should have and the challenges they present. This section should provide a based-knowledge for

Manuscript received September 9, 2022. R. Turrisi was supported by a research fellowship funded by the DECIPHER-ASL – Bando PRIN 2017 grant (2017SNW5MB - Ministry of University and Research, Italy).

R. Turrisi is with the University of Genova, Genova, 16146, Italy (corresponding author, phone: 0103536981; e-mail: rosanna.turrisi@edu.unige.it).

M. Squillario is with IRCCS Ospedale Policlinico San Martino, LISCOMPLab, 16132, Genoa, Italy (e-mail: margherita.squillario@hsanmartino.it).

G. Abate is with the University of Brescia, Brescia, 25121, Italy (e-mail: giulia.abate@unibs.it).

D. Uberti is with the University of Brescia, Brescia, 25121, Italy (e-mail: daniela.uberti@unibs.it).

A. Barla.is with the University of Genova, Genova, 16146, Italy (e-mail: annalisa.barla@unige.it).

[*] These authors contributed equally

researchers who intend to approach multi-modal analysis by machine and Deep Learning (DL) [7].

Section III aims at describing well-established datasets and AI-based algorithms on data integration in the context of AD. Finally, we highlight useful tips to road the map from theory to practice in Section IV.

II. The challenges of data integration

As data integration approaches became popular only recently, many issues have not been addressed yet in terms of both data acquisition and methods design. In the following, we delve into the collection and the use of multimodal data. Specifically, we focus on the most common data in the medical context, including, besides the demographic and clinical characteristics of patients, variables/parameters related to imaging (i.e., Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET) with or without tracers), and omics data (i.e., next generation sequencing data (whole exome sequencing (WES), exome, RNA-Seq, Chip-Seq), proteomics, metabolomics).

A. Multimodal data collection

Although several multi-dimensional cohorts are today available, the majority of them result from an *a posteriori* integration of existing datasets and, as a consequence, such databases lack some characteristics that are crucial for the development of advanced methods. Indeed, the biggest problem of ML is that algorithms are sensitive to the amount and quality of data. In the following we present the most common data issues and the properties of the ideal dataset.

Data quality

While the algorithm's data hunger is well known as the enemy number one of AI, the problem of data quality is often underestimated and not fully addressed. For instance, Nagle et al. [8] state that only 3% of companies' data achieves a standard in data quality. If this is a widespread issue in the Artificial Intelligence (AI) world, it is even more pronounced in the medical domain and neuroscience [9], [10]. As reported in [11] data quality metrics are typically developed ad hoc for specific problems. In [12], authors propose a quality assessment system for medical contexts consisting of three phases. First, the raw dataset is evaluated through metadata extraction, descriptive statistics and data annotation. Then, the data quality control step looks for inconsistencies, missing values, outliers and duplicates. Finally, data standardization must be performed. Complete reviews of data quality assessment methods have been addressed in just a few and very specific health related applications such as The Human Brain Initiative [13] and Public Health Information [14].

Another relevant issue is linked to the lack of universal standards for health data formats and interoperability. In clinical practice the disruptive absence of such standards affects the communication and the exchange of critical data among entities, also within the same clinical facility. In the research process, it prevents the implementation of reliable and statistically robust predictive models of complex pathologies. Finally, in healthcare management, it increases costs while also affecting patient safety and privacy[15], [16]. Although data collection is usually a task carried out by the hospital management and IT, we would like to emphasize how data analysts should play a relevant role in actively taking part in the process and requiring data interoperability.

Data incompleteness: heterogeneity and missingness

Real-world data is usually represented by heterogeneous and missing data, that is samples are represented by different data types where some observations are missing. The most common issue is, typically, the heterogeneity of the corpus. Differently from research databases that attempt to have completed data in all research oriented variables, Health Care institutions, whose purpose is mainly focused to document clinical care for a certain disease, fail in the completeness of data.

An obvious example is the asymmetry between groups, consisting in having different numbers of subjects per group. This is an intrinsic issue in the biomedical field as, for instance, when dealing with rare diseases the number of patients will always be considerably small compared to the healthy population. When dealing with imbalanced groups, specific techniques are required during both training and testing phases of the algorithm. At the first, re-sampling strategies (i.e., under- or over-sampling) are necessary to balance the dataset for training [17], [18], while robust metrics (e.g., F1-score, sensitivity and specificity, Brier score) are fundamental for a reliable evaluation of the model [19], [20].

Another common case of data incompleteness consists in the longitudinal incoherence: data collection protocols typically change over time and some exams can be later excluded as unnecessary, resulting in heterogeneous modalities over time.

Also, for some subjects few or many modalities may not be present if data comes from different hospitals or due to the restricted eligibility rules on recruiting patients. Indeed, for healthy subjects some measures that require invasive and/or expensive techniques (e.g., CSF, Imaging) are not performed or acquired only if necessary. This brings heterogeneity between groups.

Further, it is very unlikely to collect a real-world dataset where all observations are consistently present for all considered subjects. This may be due to data corruption during the acquisition procedure, collection procedures failing to record data or human negligence. Such issues inevitably lead to incomplete and sparse data matrices, where applying ML algorithms is unfeasible as these methods do not naturally deal with missing portions of datasets.One solution may consist of restricting the usage to the subset of subjects for which the same data modalities are accessible. Clearly, this would limit the amount of data and the robustness of the statistical model trained on it. Alternatively, ad-hoc methods may be adopted to deal with heterogeneous data and imputing strategies may be employed to deal with missing values. Recently, researchers

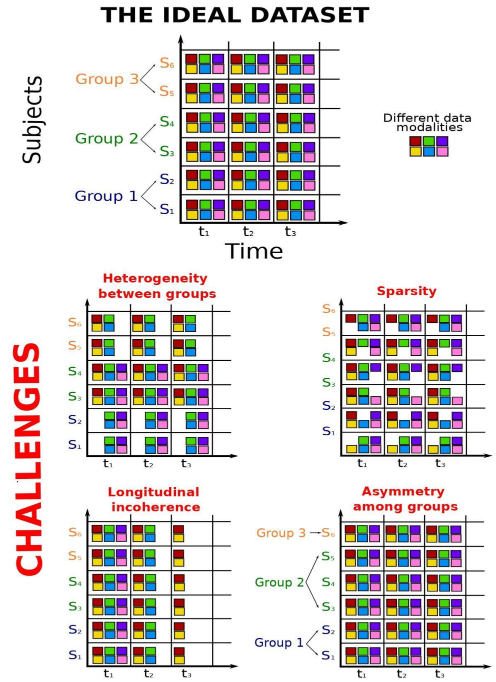

Fig. 1.The graph at the top shows an example of the ideal dataset, whereas the four graphs at the bottom show the challenges of data collection: heterogeneity, sparsity, longitudinal incoherence and asymmetry.

emphasized the importance of using such strategies to estimate the unobtainable observations while making use of the available information (see [21] for an exhaustive literature analysis from 2006 to 2017). For instance, in [22] the authors use k-nearest neighbors [23] and a similarity measure based on cross-entropy to impute missing values.

Very recently, many studies developed genomic data imputation based on deep learning and, more specifically, on auto-encoders [24]–[27].

Ideal properties of a dataset

In brief, the main characteristics of the ideal dataset can be summarized in:

- homogeneity (i.e. same type of data for all groups and subjects)
- density (i.e., not missing data)
- longitudinally coherence (i.e., same data modalities at baseline and follow up)
- symmetry among groups (i.e., similar number of subjects per group).

In Fig. 1, we represent different data modalities with colored blocks and we picture the ideal dataset that should present all the blocks for every subject at each time step. For comparison, we also show how a dataset appears when one of the mentioned-above properties is missing.

B. *Multimodal data modeling*

Having access to high-quality data lays the foundation for the design of efficient AI algorithms and to avoid the so-called “garbage in, garbage out problem”. However, developing avantgarde integrative methods require asking many questions:

- How to manage completely different data? At which model level is it better to integrate the information (bottom-intermediate-top) from both computational and clinical viewpoints?
- What is the difference between machine and deep learning integrative methods and which one is more suitable?
- How to overcome the computational costs?
- Once we have a data integration-based model, how can we interpret data? How to evaluate a model in a meaningful way for the clinician?

In the following, we attempt to answer all these questions and provide the reader with a general guideline. Note that here we focus on classification/regression models, as these represent the most challenging attempts of data integration. Nonetheless, multimodal data can also be exploited to insert prior knowledge in the model or for regression analysis, as we will see in Section III.B.2.

Heterogeneous data management and integration level

In a prediction model, we can distinguish two phases. The feature extraction step takes the raw data as input and extracts meaningful features, while the classification phase uses the extracted features to make a prediction. This implies that we mainly have three objects: the raw data, the extracted features and the output. The data integration can be performed by concatenating one of these objects for all available modalities (Fig. 2). Specifically, the first approach we can adopt occurs at a bottom level (Fig. 2a) and consists in the concatenation of the raw data modalities to feed the model that will perform both the feature extraction and classification steps. Note that this approach is rarely used, as multimodal data may have different dimensions and structure (e.g. scalar, vectors, matrices or tensors). Also they may fall into substantially different numerical ranges and therefore this integration type may lead to unreliable statistical results.

Fig. 2b shows an intermediate approach in which features are differently extracted from all modalities and their concatenation is used as input for a classifier. Here concatenation may be done leveraging common approaches such as ETL (Extract, Transform, Load) techniques when dealing with structured data in data warehouses systems [28] or semantic integration based on ontologies when handling unstructured and semi-structured data [29]. Particularly in the latter case, data standardization plays a critical role to ensure the data conforms to a common set of criteria such as consistency, accuracy and shared meaning across the different available sources [30]. Finally, the top-level integration (Fig.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3268729

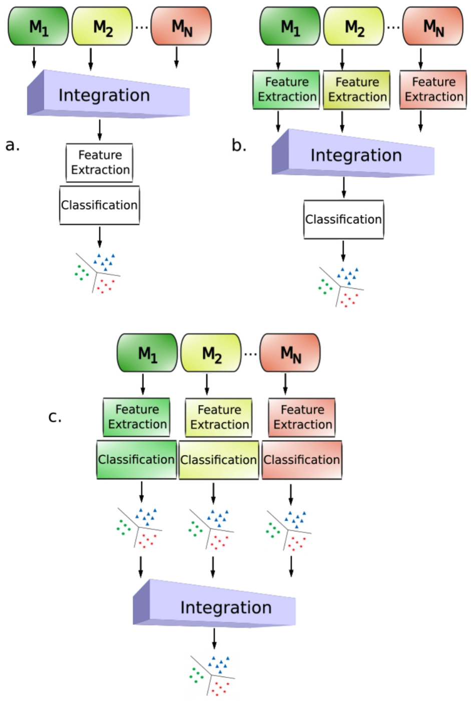

Fig. 2. The three main levels of data show how to integrate **N** modalities (**M**). a) Bottom-level integration. b) Intermediate-level integration. c) Top-level integration.

2c) performs a decision-level combination by voting or averaging the monomodal model outputs.

The best choice depends on the relation among the data modalities and on the study purpose. For example, if the different modalities share some information the bottom-level integration may extract sparse features, avoiding redundant information. On the contrary, if the data modalities differ considerably, a feature extraction strategy specific for each modality may be successful. For instance, genomics data are typically high dimensional (e.g. ~from 10000 to 100000) and require to find a sparse representation, while it is well-known that Convolutional Neural Network (CNN) [31] optimally performs the feature extraction from imaging data. Hence, in this case an intermediate- or top-level integration may result to be more suitable. Further, the strategy to integrate data should also be strictly related to the feasibility in the clinical application. When integrating multiple modalities we should be aware if the adopted method requires the use of multimodal data only during training or also in testing. Indeed, in medical practice, having simultaneous access to multiple modalities (e.g., data obtained by invasive and costly techniques) and, consequently, using them as input for a statistical model may be unfeasible. As a consequence, depending on which modalities we are dealing with, methods making use of both single- and multimodal data (e.g., integration at top level) can be more suitable. In this regard, the third type of approach is advantageous with respect to the intermediate-level integration.

TABLE I

Shallow machine learning	Deep learning
Data pre-processing needed	It can take the raw input
Feature extractors and classifiers are separately learned	Feature extractors and classifiers can be jointly learned
The extracted features are known	The extracted features are optimal w.r.t. the task to perform
More easy to be interpreted	Difficult to be interpreted
When using multiple feature extractors, they can be run in parallel	Computationally expensive

Comparison between shallow ML and DL in terms of pre-processing, data modalities handling, feature extraction, interpretability and computational cost.

Shallow Machine Learning Vs Deep Learning

Note that all described integrative approaches can be accomplished by learning the feature extractor and the classifier separately or simultaneously. While the first strategy can be performed both by leveraging machine and deep learning, the latter one relies on deep neural networks. Indeed, while traditional (or shallow) ML algorithms rely on the feature choice made by the expert through feature engineering or prior knowledge [23], DL - without any prior knowledge - can automatically extract features based on data measures or to maximize task performance (e.g., classification) [7]. Table I summarizes the key differences between shallow ML and DL. The second approach is typically more effective when studying complex diseases, where the feature choice may be wrong, biased or limited. A clear example is the case of medical imaging for which DL architectures result to be state-of-the-art in several tasks, such as brain tumor segmentation [32]–[35], lesion identification in multiple sclerosis [36]–[38], and electroencephalography (EEG) signal modeling [39], [40].

Further, and most importantly, in traditional ML the feature extraction and the classification task are performed independently . Consequently, ML can treat multiple data as a whole unit (Fig.2a) or independent units (each feature extractor only depends on its modality; Fig. 2b and 2c). Instead, deep neural networks also allow joint training of several feature extractors (e.g., one for each modality) and the classifier (Fig. 2b). The advantage of this approach is that the feature extractors differ based on the data modality but they are trained to minimize a classification loss that depends on all modalities. This provides a feature extractor specific to a single modality that, however, incorporates information coming from all data.

Although these approaches may be more effective in capturing meaningful information from multimodal data and better model the phenomena, DL techniques have two main drawbacks: they are computationally expensive and, often, lack interpretability.

These issues have been often underestimated by the research community as in some fields, such as face recognition, the data dimensionality is manageable and the performance accuracy is the main model driving criterion. This does not apply to the medical field where the data is high dimensional and the interpretability of the model is a *sine qua non* for its real use in the healthcare world [41].

Computational cost

Among the approaches to reduce the computational cost, recent advances in optimization, numerical linear algebra and random projections were exploited to maximize ML methods efficiency [42]. For DL methods, one of the most popular strategies is the network pruning that consists in removing parameters from an existing network in order to obtain a smaller model with the same accuracy. Although this technique was introduced in the late 80s [43], [44], it only recently attracted the interests of researchers [45],[46], as a consequence of exponential growth of available data and the rise of bigger and high memory-requiring models. For more details, we refer the reader to the survey by Blalock et al. [47].

Alternatively, several studies are based on knowledge distillation [48], where a small model (student) attempts to replicate the output of a larger model (teacher). However, these methods allow to reduce the computational cost only retrospectively, after a bigger network has been trained.

Interpretability and evaluation metrics

A similar problem affects the model interpretability that is more often performed after the model training, rather than incorporated in the algorithm itself. This is more evident in DL as, contrary to ML models in which the selected features are known, it extracts high-level and abstract features. Due to the difficulties of designing complex models that are also interpretable, the majority of studies aim at providing tools to interpret the learning paradigm and the model results. The main one tries to estimate the input variables importance by measuring the increase in the model's prediction error after perturbing one or a group of variables [49]. The more the prediction changes, the more the model will be dependent on that variable. This technique can be particularly useful, for instance, to find the genes mostly involved in the disease diagnosis or to individuate the part of the medical image that is more important for the diagnosis.

Alongside interpretability, the goodness of a statistical model is also evaluated by means of quantitative measures that estimate its performance. Accuracy is typically the first - and often the only - metric used by computer scientists to assess a predictive model. Nonetheless, we emphasize that in order to lay the foundation of an interdisciplinary collaboration, the model should be also evaluated following criteria that take into account the peculiarity of the biomedical problems. For instance, a common request is to estimate sensitivity, specificity, F1-score or Matthews Correlation Coefficient to better understand the phenomenon in terms of type-I and type-II error types [50].

III. DATA INTEGRATION IN ALZHEIMER'S DISEASE

In this section, we focus on data integration in Alzheimer's disease as this is, among the multifactorial neurodegenerative diseases, the most widespread and studied disorder [51].

Alzheimer's disease is indeed the most common form of dementia, currently affecting more than 30 million people in the world [52]. It is characterized by progressive neurodegeneration, leading to decline in cognitive and functional capabilities, affecting everyday activities, eventually causing death [53]. The delay of diagnosis, the lack of effective therapies, and the associated chronic disability render this disease a socio-economic calamity. In the past, the appearance of dementia symptoms marked the AD onset, and diagnosis was only confirmed postmortem by verifying the presence of beta amyloid aggregation and tau protein hyperphosphorylation [54]. However, thanks to scientific advancements in AD knowledge, it is currently well recognized that AD exists as a clinical continuum [55]:

1. a pre-symptomatic stage where pathological molecular changes, such as accumulation of the neurotoxic beta amyloid peptide, and neuronal dysfunction occur at brain level
2. a very early stage characterized by mild cognitive symptoms (identified as mild cognitive impairment (MCI) syndrome) that could be confused with aged-related physiological cognitive deterioration
3. the early-stage where AD cognitive symptoms might be recognized during a long diagnosis workflow
4. the late stage with overt dementia.

Many characteristics of AD render it especially defiant: developmental of disease occurs insidiously over the course of years or decades, the causes of disease and factors related to its severity are likely multifactorial, and a considerable phenotypic heterogeneity (ranging from typical memory loss, to canonical atypical clinical symptoms as such visual/spatial, language, motor or executive functions impairment) exist [56]. Although a great effort has been made to identify potential druggable targets, only recently, after a timespan of 20 years, the Food and Drug Administration (FDA) has approved a new drug (Aducanumab) whose true efficacy in modifying the progression of AD is yet to be confirmed [57], [58]. The long preclinical phase of AD gives hope that early intervention may be the right approach to prevent, slow or even stop the disease. In this regard researches have been mainly focusing on two broad themes:

1. identifying novel biomarkers or risk factors (e.g., apolipoprotein, cardiovascular risk factors) to diagnose AD from occurrence and testing efficacy of interventions, such as physical activity or diet, to delay the disease onset;

2. tracking AD progression using imaging, cerebrospinal fluid, and blood biomarkers (e.g., Pittsburgh compound B).

These challenging fields of investigation can be tackled thanks to recent technological advancements that have empowered us to generate, collect and manage massive amounts of data [59]. In this context, longitudinal data collection and sharing initiatives could accelerate the identification of the key factors triggering AD risk and progression [60], [61]. We believe that AD research will be largely impacted by data-driven models if we are able to successfully share and integrate this large-scale data across different organizations, groups and countries. Below, we detail the main multimodal data cohorts (see also Table II) and the state-of-the-art ML algorithms (see also Fig. 3) based on data integration that focuses on AD.

A. Datasets

Collecting multimodal, longitudinal data involving multicenter allowed us to work on a very wide set of subjects and to improve the chances to better understand the mechanisms of AD.

One of the most used publicly available dataset is the North American Alzheimer's Disease Neuroimaging Initiative (ADNI) [62]. The primary aim of ADNI is to discover, optimize, standardize, and validate clinical trial measures and biomarkers used in AD research. Indeed, ADNI is a longitudinal multicenter study that aims to develop clinical, genetic and biomedical biomarkers for AD early detection. Up to now, ADNI has experienced four different phases: ADNI1, ADNI/GO, ADNI2 and ADNI 3, including over 2000 subjects affected by different degrees of cognitive impairment.
The data types collected in the four ADNI initiatives include: (*i*) MRI (structural, diffusion weighted imaging, perfusion, and resting state sequences), (*ii*) amyloid and tau PET imaging using different specific tracers, (*iv*) CSF for Aβ, tau, phosphorylated tau (AKA phospho tau), and other proteins, (*vi*) genetic data, (*vii*) autopsy data to determine the relationship of these biomarkers to baseline clinical status and cognitive decline. The ADNI dataset has been extensively exploited within the AD Big Data DREAM Challenge at White House, a pioneering initiative launched to advance the global effort for diagnosis techniques and identifying new AD biomarkers through open source data [63].

Several other efforts for sharing AD-related data were launched afterwards, each of them with different purposes, which we briefly discuss in the remainder of this section.

For example, the Mayo Clinical Study of Ageing (MCSA) [64], designed for a population-based prospective study of cognitive healthy aging, MCI and dementia, enrolled nearly 2700 subjects through an evaluation of their medical history from a population living in Minnesota of the United States and reported only clinical characteristics, including dementia assessed by phone interviews.

The overarching aim of the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) [65] was to discover which biomarkers, cognitive characteristics, and health and lifestyle factors can be implicated in the subsequent development of symptomatic AD. One of the peculiarities of AIBL is its focus on lifestyle and long term monitoring of patients. From late 2006 to mid-2008 the AIBL study assembled and assessed an Inception cohort of 1112 individuals with the intention of conducting re-assessments every 18 months to determine the extent to which their baseline cognitive profile, demographic factors, Aß-amyloid brain load, blood and CSF biomarkers, genetic and lifestyle factors could predict their future cognitive function and clinical status with respect to the development of AD . The evolution of the Inception cohort at baseline, 18, 36, 54, 72, 90, 108, and 126 months follow-up, was enriched with recruitment of 1,247 new participants to compensate for attrition (Enrichment cohort), yielding the current database of 2,359 participants with 8,592 person-contact years.

The Alzheimer Cohorts Consortium (ACC) [66] is composed of 9 cohorts selected based on predetermined criteria. Across the cohorts there are more than 70,000 individuals of whom around 6300 have developed dementia to date. Specifically, cohorts included in ACC should be designed as prospective, population-based, have in-person examinations, with at least 15 years of available follow-up, and include at least 2,000 participants at baseline. Further, most of the cohorts should have data available on genotype, cardiovascular factors and acquisition of brain MRI.

The European initiative Amyloid imaging to prevent Alzheimer's disease (AMYPAD) [67] is aimed to better clarify the etiopathological value of PET-imaging of β-amyloid in order to: i) improve the diagnostic workflow and management of individuals with suspected AD; ii) stratify AD risk and select homogeneous groups for therapeutic intervention strategies; and iii) better quantify variables indicative of treatment effects. AMYPAD studied the onset, dynamics, and clinical relevance of brain β-amyloid in the clinical continuum from normal aging to preclinical and prodromal AD in nearly 4000 subjects in close collaboration with European Prevention of Alzheimer's Dementia (EPAD) [68].

Several datasets presented above may be exploited to answer one given question (diagnostic confirmation, prognostic, predictive, lifestyle influence) even if they provide different biological and biomedical parameters (GWAS, CSF biomarkers, MRI, PET, blood biomarkers, and lifestyle). Ideally, it would be optimal to simultaneously employ all the available datasets but, as they were originally conceived to answer multiple biomedical questions within heterogeneous domains, using them in unified and integrated investigation is highly demanding and requires ad hoc methods. Further the datasets may be exploited for tasks that are different but related to the original one. For this reason, several consortia have been recently raised to drive the dataset choice across the AD data landscape, such as the Heterogeneous Network-based

TABLE II

Dataset	URL
ADNI	http://adni.loni.usc.edu/
MCSA	https://www.mayo.edu/research/centers-programs/alzheimers-disease-research-center/research-activities/mayo-clinic-study-aging/for-researchers/data-sharing-resources
AIBL	https://aibl.csiro.au/
AMYPAD	https://amypad.eu/

Alzheimer disease cohort, consortia and initiative useful in the search of the optimal dataset.

TABLE III

Tools	URL
ALZFORUM databases	https://www.alzforum.org/databases
GAAIN	http://gaain.org
AData (Viewer)	https://adata.scai.fraunhofer.de
HENA	https://github.com/esugis/HENA

Tools for navigating the AD landscape of databases.

HENA [69] was created to integrate AD-related variables collected from well-validated clinical recruitments, as well as an innovative experimental and computational datasets created by the Age brain SYBRIO consortium.

dataset for Alzheimer's disease (HENA) and the AData(Viewer). AData(Viewer) [70] is an interactive web portal that helps researchers to select their optimal AD data cohorts exploring data within nine major AD clinical cohort studies. These datasets are quite heterogeneous considering for example the number of subjects enrolled by diagnosis, the availability of relevant biomarkers for AD, the demographic composition and also for the predominance in the cohorts of the whites/Caucasians. Therefore, the advantage of AData (viewer) is to have a metadata-based approach that allows studies to be classified according to the characteristics of the relative patient-data in each cohort. Table II reports the digital link to the above cited datasets, and Table III provides additional digital tools such as (ALZFORUM [71], GAAIN [72], AData and HENA) useful for the reader for browsing the landscape of AD databases. The aim of these tools is to aid biostatisticians in accomplishing a task efficiently, quickly, and more accurately.

B. Data integration methods

In the context of Alzheimer's Disease, we can distinguish two main types of multimodal methods that pursue two distinct scopes. The first type aims at inferring risk factors involved in the development and degeneration of the disease. Within this category we focus on the Genome-Wide Association Study (GWAS) [73], consisting in identifying the genes whose mutations are associated with AD. The second one integrates different data modalities to predict the disorder or its decline. The approaches relying on the last group are typically classification algorithms trained in a supervised way. In the following, we provide an overview of the most relevant studies in both domains.

B.1 GWAS-based Multimodal Analysis

In the context of AD studies, GWAS is used for the identification of genetic risk factors, fundamental to determine who is at a higher pathological risk and, therefore, for developing new prevention and treatment strategies. Tests for association devise independence between input measures, i.e., single nucleotide polymorphisms (SNPs) and the output, i.e. a phenotype of interest, which could be dichotomous (*affected*, *unaffected*) or quantitative (*fluid biomarker levels, rate of longitudinal change on imaging metrics, etc.*). As the total number of variables usually outnumbers the amount of available samples, it is a common procedure to encode all available prior knowledge to boost the statistical robustness of the results. This is done by incorporating, for instance, biological knowledge as gene modules or pathways from verified sources such as Gene Ontology [74] or the Kyoto Encyclopedia of Genes and Genomes [75].

Logistic and linear regression, Principal Component Analysis (PCA) and multiple hypothesis testing with Bonferroni correction represent the most employed statistical methods in AD genetic risk factors studies [76]-[79]. Studies based on GWA data have confirmed the strong influence of APOE [80], [81] among many other genes involved specific functional categories, such as *immune response*, *lipid metabolism* and *synaptic functioning* [82].

In the last decades, several studies showed that the analysis of GWA data alone can be improved by exploring additional data, such as imaging data. Improvements in both genotyping underlying GWA and brain imaging have boosted new approaches to study the influence of genetic variations on both the function and the structure of the brain [83]-[85]. This effort has led to the birth of a new research field named *imaging genetics* where genetic variations are evaluated using imaging measures as continuous phenotypes or quantitative traits (QTs) which have increased statistical power and thus decreased sample size requirements [86], [87]. Several SNPs and other polymorphisms in many genes, including APOE, have been related to neuroimaging measures in MCI and AD and also in nondemented carriers [88]. However, the complexity of these disorders remains to be unveiled because effectively relating high density SNP data to large scale image data is still a goal to be achieved. This happens because whole brain imaging studies usually find associations with few genetic variables, while GWAS coupled with imaging usually consider a low number of imaging variables [85], [89], [90]. These operations of feature reduction allow the identification of strong but few QTs-genetic variations associations that represent only a few pieces of the larger and more complex puzzle which are diseases such as AD or MCI. In imaging genetics, new methods able to overcome power limitations and multiple comparison issues need to be conceived and this can be achieved only by considering multidisciplinary teams, methods and multimodal datasets in order to overcome the limitations of the analysis of monomodal datasets.

B.2 Multimodal Data Integration for disease prediction

In the following, we report the most cited or recent methodological studies on AD, from 2010 to 2023. We did not include similar papers, as our goal is to show the progress of methods over time and to compare the performance of different approaches. All considered publications make use of the ADNI dataset, as it is the standard *de facto* dataset for modeling several aspects of AD pathology. The remainder of this section showcases first the publications where shallow learning approaches were employed and then illustrates those using more advanced deep learning methods.

Shallow machine learning

The first attempts to integrate multiple data to perform a classification task were based on Single-Kernel methods, such as Support Vector Machine (SVM) [91].

In [92], the authors concatenate features extracted from PET and MRI images with CSF, APOE genotype and cognitive information to perform a binary classification via SVM. They reach 70% and 82% on MCI and AD classification, respectively. The major limitation of these works is that the same kernel is applied to different modalities.

Moving a step forward, [93] employs different Gaussian kernels with mixed L_{21}norm penalty on the kernel weights to enforce group sparsity among different feature modalities (CSF and MRI). This method can distinguish between control and AD subjects with 87% accuracy. [94] proposes a Multi-Kernel Learning (MKL) for the binary classification, considering two imaging modalities (PET and MRI) and, for each, computing several voxel-wise features and kernel functions (i.e., linear, quadratic and Gaussian) resulting in 24 kernel matrices par modality. Further, the authors take into account non-imaging modalities, such as CSF assays, NeuroPsychological Status Exam (NPSE) scores and APOE genotype, giving three kernels per modality. Finally, MKL classifier is trained to: i) distinguish between CN and AD; ii) predict the conversion from MCI to AD. This integration method accomplishes the first task with 92% accuracy, while it fails at the second task.

A similar approach [95] adopts multiple kernels for the PET, MRI and CSF modalities, and combines them to train a SVM classifier. The AD (MCI) classification is performed with 93% (77%) of accuracy.

All these intermediate-level integration studies showed that a multi-kernel approach outperforms the single-kernel SVM classification. We emphasize that in all the described methods the monomodal feature extractors and the classifier are learned separately. Recently, [96], [97] proposed ensemble classifiers based on SVM (eSVM). Both studies rely on the top-level integration in which the monomodal SVM classifiers are averaged. Specifically, [96] employs features extracted from ROIs of different imaging modalities, i.e. MRIs, Diffusion Tensor Imaging (DTI), and PET. They show that the combination of all imaging modalities yields an accuracy of 98% in AD diagnosis, outperforming both monomodal classifiers and 2-modalities based classifiers. Similar conclusions can be retrieved from [97] in which the combined use of MRIs and Transcriptomic data achieves 95% of accuracy in AD/CN classification, while the use of single modalities (MRI/Transcriptomic) reach 93% and 86% of accuracy, respectively. Differently, they showed that MRI alone (64% accuracy) can outperform multimodal data integration (56% accuracy) in CN/MCI classification. However, they emphasize that none of the considered sets of features lead to acceptable performance in the challenging task of discriminating between MCI and CN subjects.

Deep learning

To treat multimodal data many authors focused on DL methods [7].

For instance, in [98], the authors look for a shared representation of PET and MRI images by using the Deep Boltzmann Machine (DBM). This is then used as input for hierarchical classifiers, reaching 94% (85%) of accuracy in the AD (MCI) classification.

Lian et al. [99] fuse 1.5T and 3T T1-weighted MRI images by proposing a variant of a CNN, named Hierarchical Fully Convolutional Network (H-FCN). They also perform network pruning to reduce the computational cost. This method diagnoses AD with 90% of accuracy and predicts the conversion from MCI to AD with 80% of accuracy. Both [98] and [99] studies rely on the second-level integration in which the feature extraction and the classification steps are jointly performed.

Venugopalan and co-authors [100] employ Denoising Auto-Encoders (DAE) to extract features from clinical and genetic data, and a 3 dimensional-CNN for imaging data. The authors compare the performance of bottom-, intermediate- (in which feature extractors and classifiers are learned separately), and top-level integration, showing that the intermediate one provides the most effective performance. Contrary to all previous studies, their algorithm is not limited to the binary classification but it also performs a multi-class (NC/MCI/AD) classification with 85% accuracy.

Finally, it is worthwhile to mention The Multimodal Longitudinal Data Integration (MildInt) work [101], that differs from the previous studies as it takes into account longitudinal data. This approach is composed of two phases: 1) extracting fixed-size features from different modalities, represented by time series; 2) integrating the extracted features and learning a classifier to make the final decision. This corresponds to an intermediate-integration approach, in which the feature extractors and the classifier are trained separately. Specifically, the extractors are based on Recurrent Neural Network (RNN) models to capture the data time dependency, while the classifier is learned by using a logistic regression function. The authors adopt the method to distinguish between

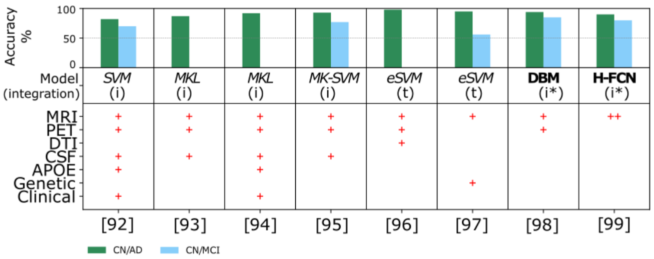

Fig. 3. The schema reports current and recent literature on data integration approaches for the study of AD. For each contribution, from bottom to top, we report the reference, which data modalities were used, the ML and integration methods employed, and the accuracy performance for two binary classification tasks: CN vs MCI and CN vs AD. Shallow learning methods are reported in *italic*, whereas deep learning methods are reported in **bold**. Integration type may be intermediate (i)/(i*) or top (t).

MCI-converter and -non converter and show that it provides a performance improvement over the monomodal approaches, in which CSF, MRI or Cognitive Performance data only is used. MildInt is publicly available as a Python package[2] and it can be employed as a preconstructed architecture in other data integration contexts.

Fig. 3 summarizes those papers that tackle the problem of classifying CN/AD or CN/MCI. For each column from bottom to top, the citation reference, data type considered in the study, learning and integration method and achieved accuracy for the two classification tasks. Methods falling into the shallow learning category are reported in *italic,* those in the deep learning category are reported in bold. Integration is either happening at the top level (t) or intermediate where feature extraction and classification are performed in either (i) two separate steps, or (i*) a joint training phase. We want to emphasize that, due to the absence of test benchmarking and the use of different modalities, it is not possible to make a direct comparison between studies and select the best strategy. This strictly depends on the task at hand and on the available data. Nonetheless, we here report some possible conclusions:

- multi-kernel strategies always outperform single-kernel algorithm suggesting the needs for treating each modality differently;
- DL models outperform shallow ML in the discrimination between CN and MCI;
- top-level integration reaches excellent AD diagnosis performance when the feature extraction phase is based on a hybrid human-AI approach where data-driven methods support human decision;
- MRI is the most employed data type.

Finally, we note that in this section we did not consider the integration across-datasets but we focused on the integration of different data modalities. The reason is that, as emphasized in the previous section, datasets may have been designed from distant scopes and their integration may not be suitable. However, for some specific tasks, their joint use may be convenient.

IV. Conclusion

This work serves as a useful foundation for establishing research based on data integration, where data collection and statistical analysis are intertwined and mutually affect one another instead of being treated as independent phases. As such, two main ingredients of data integration must be considered: (i) how to select and use available datasets in the light of the task at hand and the characteristics that an ideal multimodal dataset should fulfill; (ii) how and at which level to integrate different data modalities in statistical models to meet both the computational and clinical requirements.

First, a multimodal dataset can be considered optimal from the perspective of the computer scientists community, when data is homogeneous (i.e., the same modalities are available for all groups and subjects), dense (i.e., there is no missing data), longitudinal coherent (i.e., same modalities are present at baseline and follow up), and symmetric among groups (i.e., different groups include a similar number of subjects). However, we observed that in practice current AD datasets do not satisfy these principles due to both intrinsic issues, such as constraints of ethical, economical or temporal nature, and fallacious data collection practices (e.g., the use of different protocols across hospitals). Therefore, one of the results of our work is a list of techniques to overcome these shortcomings (e.g., standardized protocols, estimation of missing values). Moreover, we remark that the choice of the dataset to employ in statistical methods cannot be based only on computational reasons (e.g., the largest and mostly homogeneous dataset) but it should strictly depend on the biomedical question**.**

Second, our contribution is also to provide a hands-on guide on how to treat different data, describing three possible approaches of data integration (i.e., bottom-, intermediate- and top- level) discussing their benefits and drawbacks in terms of feasibility, computational cost, and model interpretability.

In [100] the authors claim that, when treating heterogeneous data type, the intermediate-level approach provides the best performance in terms of efficiency and model accuracy. Nonetheless, it is fundamental to be aware that this approach may not be the best one from the clinical point of view. Indeed, this method requires all modalities to be available also during the testing phase: a doctor could make a diagnosis only when information from all modalities are present. On the contrary, other techniques, such as the top-level integration, employ multimodal data for learning the model but they can provide an outcome also when a single-modality is available. Hence, methods that are less efficient from the data science perspective may have some advantages in medical practice.

This study explores the use of multiple data modalities for investigating NDs. While current ML models are not necessarily yet prepared for implementation and deployment in medical centers, our research indicates that data integration methods can significantly improve the effectiveness of standard monomodal approaches.

[2] https://github.com/goeastagent/MildInt

Our findings are not limited to AD, as all reviewed methods may be transferred to other NDs, suggesting that data integration combined with advanced machine learning methods may be key for a future in which digital-assisted diagnosis will support clinicians towards a timely and accurate diagnosis of NDs, ultimately leading to better patient prognosis.

References

[1] R. Sims, M. Hill, and J. Williams, "The multiplex model of the genetics of Alzheimer's disease," Nat. Neurosci., vol. 23, no. 3, Art. no. 3, Mar. 2020, doi: 10.1038/s41593-020-0599-5.

[2] I. Garali et al., "A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia," Brief. Bioinform., vol. 19, no. 6, pp. 1356–1369, Nov. 2018, doi: 10.1093/bib/bbx060.

[3] J. Chen et al., "Integration of Multimodal Data for Deciphering Brain Disorders," Annu. Rev. Biomed. Data Sci., vol. 4, no. 1, pp. 43–56, 2021, doi: 10.1146/annurev-biodatasci-092820-020354.

[4] M. E. Martone, A. Gupta, and M. H. Ellisman, "e-Neuroscience: challenges and triumphs in integrating distributed data from molecules to brains," Nat. Neurosci., vol. 7, no. 5, Art. no. 5, May 2004, doi: 10.1038/nn1229.

[5] E. G. Baxi et al., "Answer ALS, a large-scale resource for sporadic and familial ALS combining clinical and multi-omics data from induced pluripotent cell lines," Nat. Neurosci., vol. 25, no. 2, Art. no. 2, Feb. 2022, doi: 10.1038/s41593-021-01006-0.

[6] M. B. Makarious et al., "Multi-modality machine learning predicting Parkinson's disease," Npj Park. Dis., vol. 8, no. 1, Art. no. 1, Apr. 2022, doi: 10.1038/s41531-022-00288-w.

[7] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[8] T. Nagle, T. C. Redman, and D. Sammon, "Only 3% of companies' data meets basic quality standards," Harv. Bus. Rev., p. 5, 2017.

[9] Z. Obermeyer and E. J. Emanuel, "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine," N. Engl. J. Med., vol. 375, no. 13, pp. 1216–1219, Sep. 2016, doi: 10.1056/NEJMp1606181.

[10] A. Kolossa and B. Kopp, "Data quality over data quantity in computational cognitive neuroscience," NeuroImage, vol. 172, pp. 775–785, May 2018, doi: 10.1016/j.neuroimage.2018.01.005.

[11] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," Commun. ACM, vol. 45, no. 4, pp. 211–218, Apr. 2002, doi: 10.1145/505248.506010.

[12] V. C. Pezoulas et al., "Medical data quality assessment: On the development of an automated framework for medical data curation," Comput. Biol. Med., vol. 107, pp. 270–283, Apr. 2019, doi: 10.1016/j.compbiomed.2019.03.001.

[13] G. M. Shepherd et al., "The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data," Trends Neurosci., vol. 21, no. 11, pp. 460–468, Nov. 1998, doi: 10.1016/S0166-2236(98)01300-9.

[14] H. Chen, D. Hailey, N. Wang, and P. Yu, "A Review of Data Quality Assessment Methods for Public Health Information Systems," Int. J. Environ. Res. Public. Health, vol. 11, no. 5, pp. 5170–5207, May 2014, doi: 10.3390/ijerph110505170.

[15] P. Aspden, Institute of Medicine (U.S.), and Committee on Data Standards for Patient Safety, Patient safety: achieving a new standard for care. Washington, D.C.: National Academies Press, 2004. Accessed: Sep. 01, 2021. [Online]. Available: http://public.eblib.com/choice/publicfullrecord.aspx?p=3376726

[16] P. Brooks, "Standards and Interoperability in Healthcare Information Systems: Current Status, Problems, and Research Issues," p. 8, 2010.

[17] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," Expert Syst. Appl., vol. 73, pp. 220–239, May 2017, doi: 10.1016/j.eswa.2016.12.035.

[18] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[19] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," Pattern Recognit. Lett., vol. 30, no. 1, pp. 27–38, Jan. 2009, doi: 10.1016/j.patrec.2008.08.010.

[20] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: a review," Int. J. Pattern Recognit. Artif. Intell., vol. 23, no. 04, pp. 687–719, Jun. 2009, doi: 10.1142/S0218001409007326.

[21] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," Artif. Intell. Rev., vol. 53, no. 2, pp. 1487–1509, Feb. 2020, doi: 10.1007/s10462-019-09709-4.

[22] E. Tavazzi, S. Daberdaku, R. Vasta, A. Calvo, A. Chiò, and B. Di Camillo, "Exploiting mutual information for the imputation of static and dynamic mixed-type clinical data with an adaptive k-nearest neighbours approach," BMC Med. Inform. Decis. Mak., vol. 20, no. 5, p. 174, Aug. 2020, doi: 10.1186/s12911-020-01166-2.

[23] T. Hastie, R. Tibshirani, and T. Friedman, The Elements of Statistical Learning. 2009.

[24] Y. L. Qiu, H. Zheng, and O. Gevaert, "A deep learning framework for imputing missing values in genomic data." bioRxiv, p. 406066, Sep. 03, 2018. doi: 10.1101/406066.

[25] J. Chen and X. Shi, "Sparse Convolutional Denoising Autoencoders for Genotype Imputation," Genes, vol. 10, no. 9, Art. no. 9, Sep. 2019, doi: 10.3390/genes10090652.

[26] Y. L. Qiu, H. Zheng, and O. Gevaert, "Genomic data imputation with variational auto-encoders," GigaScience, vol. 9, no. 8, p. giaa082, Aug. 2020, doi: 10.1093/gigascience/giaa082.

[27] T. Islam et al., A Deep Learning Method to Impute Missing Values and Compress Genome-wide Polymorphism Data in Rice. 2021, p. 109. doi: 10.5220/0010233901010109.

[28] Q. Yao, Y. Tian, P.-F. Li, L.-L. Tian, Y.-M. Qian, and J.-S. Li, "Design and Development of a Medical Big Data Processing System Based on Hadoop," J. Med. Syst., vol. 39, no. 3, p. 23, Feb. 2015, doi: 10.1007/s10916-015-0220-8.

[29] R. Lenz, M. Beyer, and K. A. Kuhn, "Semantic integration in healthcare networks," Int. J. Med. Inf., vol. 76, no. 2, pp. 201–207, Feb. 2007, doi: 10.1016/j.ijmedinf.2006.05.008.

[30] M. Jayaratne et al., "A data integration platform for patient-centered e-healthcare and clinical decision support," Future Gener. Comput. Syst., vol. 92, pp. 996–1008, Mar. 2019, doi: 10.1016/j.future.2018.07.061.

[31] Y. LeCun, Y. Bengio, and T. B. Laboratories, "Convolutional Networks for Images, Speech, and Time-Series," p. 14, 1995.

[32] M. Havaei et al., "Brain tumor segmentation with Deep Neural Networks," Med. Image Anal., vol. 35, pp. 18–31, Jan. 2017, doi: 10.1016/j.media.2016.05.004.

[33] S. Nema, A. Dudhane, S. Murala, and S. Naidu, "RescueNet: An unpaired GAN for brain tumor segmentation," Biomed. Signal Process. Control, vol. 55, p. 101641, Jan. 2020, doi: 10.1016/j.bspc.2019.101641.

[34] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images," IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1240–1251, May 2016, doi: 10.1109/TMI.2016.2538465.

[35] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating FCNNs and CRFs for brain tumor segmentation," Med. Image Anal., vol. 43, pp. 98–111, Jan. 2018, doi: 10.1016/j.media.2017.10.002.

[36] S. Cerri et al., "A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis," NeuroImage, vol. 225, p. 117471, Jan. 2021, doi: 10.1016/j.neuroimage.2020.117471.

[37] M. Sadeghibakhi, H. Pourreza, and H. Mahyar, "Multiple Sclerosis Lesions Segmentation Using Attention-Based CNNs in FLAIR Images," IEEE J. Transl. Eng. Health Med., vol. 10, pp. 1–11, 2022, doi: 10.1109/JTEHM.2022.3172025.

[38] A. Shoeibi et al., "Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review," Comput. Biol. Med., vol. 136, p. 104697, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104697.

[39] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," Hum. Brain Mapp., vol. 38, no. 11, pp. 5391–5420, 2017, doi: 10.1002/hbm.23730.

[40] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," J. Neural Eng., vol. 16, no. 5, p. 051001, Aug. 2019, doi: 10.1088/1741-2552/ab260c.

[41] C. Meng, L. Trinh, N. Xu, J. Enouen, and Y. Liu, "Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset," Sci. Rep., vol. 12, no. 1, Art. no. 1, May 2022, doi: 10.1038/s41598-022-11012-2.

[42] G. Meanti, L. Carratino, L. Rosasco, and A. Rudi, "Kernel Methods Through the Roof: Handling Billions of Points Efficiently," in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 14410–14422. Accessed: Apr. 29, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/a59afb1b7d82ec353921a55c579ee26d-Abstract.html

[43] E. D. Karnin, "A simple procedure for pruning back-propagation trained neural networks," IEEE Trans. Neural Netw., vol. 1, no. 2, pp. 239–242, Jun. 1990, doi: 10.1109/72.80236.

[44] J. SIETSMA, "Neural net pruning-why and how," Proc. Int. Conf. Neural Netw. San Diego CA 1988, vol. 1, pp. 325–333, 1988.

[45] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the Value of Network Pruning," ArXiv181005270 Cs Stat, Mar. 2019, Accessed: Jul. 12, 2021. [Online]. Available: http://arxiv.org/abs/1810.05270

[46] P. Molchanov, A. Mallya, S. Tyree, I. Frosio, and J. Kautz, "Importance Estimation for Neural Network Pruning," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11264–11272. Accessed: Jul. 12, 2021. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Molchanov_Importance_Estimation_for_Neural_Network_Pruning_CVPR_2019_paper.html

[47] D. Blalock, J. J. G. Ortiz, J. Frankle, and J. Guttag, "What is the State of Neural Network Pruning?," ArXiv200303033 Cs Stat, Mar. 2020, Accessed: Jul. 12, 2021. [Online]. Available: http://arxiv.org/abs/2003.03033

[48] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," ArXiv150302531 Cs Stat, Mar. 2015, Accessed: Jul. 12, 2021. [Online]. Available: http://arxiv.org/abs/1503.02531

[49] A. Fisher, C. Rudin, and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," p. 81, 2019.

[50] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, p. 6, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[51] "2020 Alzheimer's disease facts and figures," Alzheimers Dement., vol. 16, no. 3, pp. 391–460, 2020, doi: 10.1002/alz.12068.

[52] Y.-T. Wu et al., "The changing prevalence and incidence of dementia over time - current evidence," Nat. Rev. Neurol., vol. 13, no. 6, pp. 327–339, Jun. 2017, doi: 10.1038/nrneurol.2017.63.

[53] B. Dubois et al., "Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria," Alzheimers Dement., vol. 12, no. 3, pp. 292–323, 2016, doi: 10.1016/j.jalz.2016.02.002.

[54] H. Hampel et al., "Blood-based biomarkers for Alzheimer disease: mapping the road to the clinic," Nat. Rev. Neurol., vol. 14, no. 11, pp. 639–652, Nov. 2018, doi: 10.1038/s41582-018-0079-7.

[55] C. R. Jack Jr. et al., "NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease," Alzheimers Dement., vol. 14, no. 4, pp. 535–562, 2018, doi: 10.1016/j.jalz.2018.02.018.

[56] V. van der Velpen et al., "Systemic and central nervous system metabolic alterations in Alzheimer's disease," Alzheimers Res. Ther., vol. 11, no. 1, p. 93, Nov. 2019, doi: 10.1186/s13195-019-0551-7.

[57] C. for D. E. and Research, "FDA's Decision to Approve New Treatment for Alzheimer's Disease," FDA, Jan. 2022, Accessed: Mar. 10, 2023. [Online]. Available: https://www.fda.gov/drugs/news-events-human-drugs/fdas-decision-approve-new-treatment-alzheimers-disease

[58] L. Schneider, "A resurrection of aducanumab for Alzheimer's disease," Lancet Neurol., vol. 19, no. 2, pp. 111–112, Feb. 2020, doi: 10.1016/S1474-4422(19)30480-6.

[59] H. Geerts et al., "Big data to smart data in Alzheimer's disease: The brain health modeling initiative to foster actionable knowledge," Alzheimers Dement., vol. 12, no. 9, pp. 1014–1021, 2016, doi: 10.1016/j.jalz.2016.04.008.

[60] G. Abate et al., "A Conformation Variant of p53 Combined with Machine Learning Identifies Alzheimer Disease in Preclinical and Prodromal Stages," J. Pers. Med., vol. 11, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/jpm11010014.

[61] J. H. Park et al., "Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data," Npj Digit. Med., vol. 3, no. 1, pp. 1–7, Mar. 2020, doi: 10.1038/s41746-020-0256-0.

[62] S. G. Mueller et al., "The Alzheimer's Disease Neuroimaging Initiative," Neuroimaging Clin. N. Am., vol. 15, no. 4, pp. 869–xii, Nov. 2005, doi: 10.1016/j.nic.2005.09.008.

[63] G. I. Allen et al., "Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease," Alzheimers Dement. J. Alzheimers Assoc., vol. 12, no. 6, pp. 645–653, Jun. 2016, doi: 10.1016/j.jalz.2016.02.006.

[64] R. O. Roberts et al., "The Mayo Clinic Study of Aging: Design and Sampling, Participation, Baseline Measures and Sample Characteristics," Neuroepidemiology, vol. 30, no. 1, pp. 58–69, Mar. 2008, doi: 10.1159/000115751.

[65] K. A. Ellis et al., "The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease," Int. Psychogeriatr., vol. 21, no. 4, pp. 672–687, Aug. 2009, doi: 10.1017/S1041610209009405.

[66] L. B. Chibnik et al., "Trends in the incidence of dementia: design and methods in the Alzheimer Cohorts Consortium," Eur. J. Epidemiol., vol. 32, no. 10, pp. 931–938, Oct. 2017, doi: 10.1007/s10654-017-0320-5.

[67] "AMYPAD," Amyloid imaging to prevent Alzheimer's disease, 2016. https://amypad.eu/ (accessed Sep. 30, 2021).

[68] C. W. Ritchie, J. L. Molinuevo, L. Truyen, A. Satlin, S. Van der Geyten, and S. Lovestone, "Development of interventions for the secondary prevention of Alzheimer's dementia: the European Prevention of Alzheimer's Dementia (EPAD) project," Lancet Psychiatry, vol. 3, no. 2, pp. 179–186, Feb. 2016, doi: 10.1016/S2215-0366(15)00454-X.

[69] E. Sügis et al., "HENA, heterogeneous network-based data set for Alzheimer's disease," Sci. Data, vol. 6, no. 1, p. 151, Aug. 2019, doi: 10.1038/s41597-019-0152-0.

[70] "ADataViewer," 2020. https://adata.scai.fraunhofer.de/ (accessed Sep. 30, 2021).

[71] J. Kinoshita and T. Clark, "Alzforum," in Neuroinformatics, C. J. Crasto and S. H. Koslow, Eds. Totowa, NJ: Humana Press, 2007, pp. 365–381. doi: 10.1007/978-1-59745-520-6_19.

[72] A. W. Toga, S. C. Neu, P. Bhatt, K. L. Crawford, and N. Ashish, "The Global Alzheimer's Association Interactive Network," Alzheimers Dement., vol. 12, no. 1, pp. 49–54, 2016, doi: 10.1016/j.jalz.2015.06.1896.

[73] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five Years of GWAS Discovery," Am. J. Hum. Genet., vol. 90, no. 1, pp. 7–24, Jan. 2012, doi: 10.1016/j.ajhg.2011.11.029.

[74] Gene Ontology Consortium, "The Gene Ontology (GO) database and informatics resource," Nucleic Acids Res., vol. 32, no. suppl_1, pp. D258–D261, Jan. 2004, doi: 10.1093/nar/gkh036.

[75] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," Nucleic Acids Res., vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.

[76] C. Antúnez et al., "The membrane-spanning 4-domains, subfamily A (MS4A) gene cluster contains a common variant associated with Alzheimer's disease," Genome Med., vol. 3, no. 5, p. 33, May 2011, doi: 10.1186/gm249.

[77] D. Harold et al., "Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease, and shows evidence for additional susceptibility genes," Nat. Genet., vol. 41, no. 10, pp. 1088–1093, Oct. 2009, doi: 10.1038/ng.440.

[78] P. Hollingworth et al., "Common variants in ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease," Nat. Genet., vol. 43, no. 5, pp. 429–435, May 2011, doi: 10.1038/ng.803.

[79] A. C. Naj et al., "Common variants in MS4A4/MS4A6E, CD2uAP, CD33, and EPHA1 are associated with late-onset Alzheimer's disease," Nat. Genet., vol. 43, no. 5, pp. 436–441, May 2011, doi: 10.1038/ng.801.

[80] L. Bertram, M. B. McQueen, K. Mullin, D. Blacker, and R. E. Tanzi, "Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database," Nat. Genet., vol. 39, no. 1, pp. 17–23, Jan. 2007, doi: 10.1038/ng1934.

[81] M. Squillario, G. Abate, F. Tomasi, V. Tozzo, A. Barla, and D. Uberti, "A telescope GWAS analysis strategy, based on SNPs-genes-pathways ensemble and on multivariate algorithms, to characterize late onset Alzheimer's disease," Sci. Rep., vol. 10, p. 12063, Jul. 2020, doi: 10.1038/s41598-020-67699-8.

[82] W. Shao, D. Peng, and X. Wang, "Genetics of Alzheimer's disease: From pathogenesis to clinical usage," J. Clin. Neurosci., vol. 45, pp. 1–8, Nov. 2017, doi: 10.1016/j.jocn.2017.06.074.

[83] D. C. Glahn, T. Paus, and P. M. Thompson, "Imaging genomics: Mapping the influence of genetics on brain structure and function," Hum. Brain Mapp., vol. 28, no. 6, pp. 461–463, Apr. 2007, doi: 10.1002/hbm.20416.

[84] J. R. Harrison et al., "Imaging Alzheimer's genetic risk using diffusion MRI: A systematic review," NeuroImage Clin., vol. 27, p. 102359, Jul. 2020, doi: 10.1016/j.nicl.2020.102359.

[85] M. Klein et al., “Brain imaging genetics in ADHD and beyond – mapping pathways from gene to disorder at different levels of complexity,” Neurosci. Biobehav. Rev., vol. 80, pp. 115–155, Sep. 2017, doi: 10.1016/j.neubiorev.2017.01.013.

[86] S. G. Potkin et al., “Genome-wide Strategies for Discovering Genetic Influences on Cognition and Cognitive Disorders: Methodological Considerations,” Cognit. Neuropsychiatry, vol. 14, no. 4–5, pp. 391–418, 2009, doi: 10.1080/13546800903059829.

[87] Z. Zhu et al., “Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets,” Nat. Genet., vol. 48, no. 5, pp. 481–487, May 2016, doi: 10.1038/ng.3538.

[88] J.-Q. Li et al., “GWAS-Linked Loci and Neuroimaging Measures in Alzheimer’s Disease,” Mol. Neurobiol., vol. 54, no. 1, pp. 146–153, Jan. 2017, doi: 10.1007/s12035-015-9669-1.

[89] X. Meng et al., “Multivariate genome wide association and network analysis of subcortical imaging phenotypes in Alzheimer’s disease,” BMC Genomics, vol. 21, no. Suppl 11, 2020, doi: 10.1186/s12864-020-07282-7.

[90] A. Wiberg et al., “Handedness, language areas and neuropsychiatric diseases: insights from brain imaging and genetics,” Brain, vol. 142, no. 10, pp. 2938–2947, Oct. 2019, doi: 10.1093/brain/awz257.

[91] N. Cristianini, J. Shawe-Taylor, and D. of C. S. R. H. J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, 2000.

[92] O. Kohannim et al., “Boosting power for clinical trials using classifiers based on multiple biomarkers,” Neurobiol. Aging, vol. 31, no. 8, pp. 1429–1442, Aug. 2010, doi: 10.1016/j.neurobiolaging.2010.04.022.

[93] F. Liu, L. Zhou, C. Shen, and J. Yin, “Multiple Kernel Learning in the Primal for Multimodal Alzheimer’s Disease Classification,” IEEE J. Biomed. Health Inform., vol. 18, no. 3, pp. 984–990, May 2014, doi: 10.1109/JBHI.2013.2285378.

[94] C. Hinrichs, V. Singh, G. Xu, and S. C. Johnson, “Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population,” NeuroImage, vol. 55, no. 2, pp. 574–589, Mar. 2011, doi: 10.1016/j.neuroimage.2010.10.081.

[95] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial Landmark Detection by Deep Multi-task Learning,” in Computer Vision – ECCV 2014, Cham, 2014, pp. 94–108. doi: 10.1007/978-3-319-10599-4_7.

[96] D. Agostinho, F. Caramelo, A. P. Moreira, I. Santana, A. Abrunhosa, and M. Castelo-Branco, “Combined Structural MR and Diffusion Tensor Imaging Classify the Presence of Alzheimer’s Disease With the Same Performance as MR Combined With Amyloid Positron Emission Tomography: A Data Integration Approach,” Front. Neurosci., vol. 15, 2022, Accessed: Mar. 08, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.638175

[97] L. Maddalena, I. Granata, M. Giordano, M. R. Manzo, M. R. Guarracino, and A. D. N. Initiative (ADNI), “Classifying Alzheimer’s Disease using MRIs and Transcriptomic Data,” presented at the 9th International Conference on Bioimaging, Mar. 2023, pp. 70–79. Accessed: Mar. 08, 2023. [Online]. Available: https://www.scitepress.org/Link.aspx?doi=10.5220/0010902900003123

[98] H.-I. Suk, S.-W. Lee, and D. Shen, “Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis,” NeuroImage, vol. 101, pp. 569–582, Nov. 2014, doi: 10.1016/j.neuroimage.2014.06.077.

[99] C. Lian, M. Liu, J. Zhang, and D. Shen, “Hierarchical Fully Convolutional Network for Joint Atrophy Localization and Alzheimer’s Disease Diagnosis Using Structural MRI,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 4, pp. 880–893, Apr. 2020, doi: 10.1109/TPAMI.2018.2889096.

[100] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, “Multimodal deep learning models for early detection of Alzheimer’s disease stage,” Sci. Rep., vol. 11, no. 1, Art. no. 1, Feb. 2021, doi: 10.1038/s41598-020-74399-w.

[101] G. Lee, B. Kang, K. Nho, K.-A. Sohn, and D. Kim, “MildInt: Deep Learning-Based Multimodal Longitudinal Data Integration Framework,” Front. Genet., vol. 10, p. 617, 2019, doi: 10.3389/fgene.2019.00617.