Ensemble deep learning for Alzheimer's disease characterization and estimation

ABSTRACT

Alzheimer's Disease (AD) is the most common form of dementia, characterized by a continual deterioration of cognitive abilities in elderly people. Neuroimaging data, e.g., magnetic resonance imaging and positron emission tomography, enable the identification of structural and functional changes caused by AD in the brain. Diagnosing AD is critical in medical settings, as it supports early intervention, treatment planning, and contributes to expanding our knowledge of the AD dynamics in the brain. Lately, ensemble deep learning has become popular for enhancing the performance and reliability of AD diagnosis. These models combine several deep neural networks to increase the prediction's robustness. This work revisits key developments of ensemble deep learning, connecting their design – the type of ensemble, its heterogeneity, and data modalities – with their application to AD diagnosis using neuroimaging and genetic data. Trends and challenges are discussed thoroughly to assess where our knowledge in this area stands.

Introduction

Alzheimer's Disease (AD) is a persistent neurological disorder characterized by a decrease in cognitive abilities with no proven cure[1], usually affecting people in their mid-60s[2]. In 2019, AD was listed in seventh place for its global contribution to deaths in 2020-21[3]. The global disease liability of AD is expected to hit $2 trillion by 2030, posing an utmost urgency for advances supporting its early detection[4]. Late-life depression (LLD) and learning disabilities present significant symptoms for the elderly population and society as a whole and can contribute to negative consequences, including compromised physical health, cognitive decline, heightened risk of dementia, and increased mortality, thereby pernicious affecting the AD patient[5]. Mild Cognitive Impairment (MCI) is a transitional stage between cognitively normal (CN) older individuals and those with AD[6]. It describes people with slight signs of brain dysfunction who can nevertheless carry out their daily activities.

The exact causes of AD are not fully understood. However, research has identified several factors that may contribute to its occurrence. One of the primary reasons for AD is the accumulation of amyloid beta ($A\beta$) plaques in the brain, which subsequently leads to structural and functional changes[7,8]. $A\beta$ plaques induce the formation of neurofibrillary tangles, constructed of the protein tau and manifesting as twisted protein fibers. Tangles interfere with the internal structure of nerve cells, impairing their communication and adversely affecting their overall health. In addition, the accumulation of $A\beta$ plaques and neurofibrillary tangles disrupt the intricate network of connections between neurons. Over a longer period of time, this disruption of electrical signals affects essential processes such as learning, memory, and cognitive ability.

Early AD detection is a grand challenge, as it often progresses slowly and its symptoms may be subtle at early stages. In clinical settings, early detection of AD is crucial, as it allows for timely intervention and management of the disease, reducing healthcare costs, potentially slowing its progression, and improving the quality of life for patients. A range of tests and assessments are available to identify cognitive impairment and assess the likelihood of developing AD, including neuropsychological tests, brain imaging, and genetic testing. In essence, healthcare professionals (e.g., doctors and radiologists) use brain imaging scans to detect early signs of AD in individuals at risk of developing the condition and identify structural and functional changes in the brain caused by AD. Among different brain imaging scans, Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI) are the most widely used modalities in this area due to their high accuracy and ability to provide comprehensive information about the brain's structure and functioning. PET scans are advantageous in evaluating the functional aspects of the brain, for instance, glucose metabolism. Moreover, PET scans can give data about the presence and distribution of $A\beta$ plaques in the brain. On the other hand, structural MRI (sMRI) scans can provide a detailed look at the brain's anatomy, which can be used to assess any structural changes related to AD. Radiologists can use MRI scans to measure the reduction in the size of certain brain regions, such as the hippocampus and cerebral cortex, which are commonly affected in AD. These structural changes observed in MRI scans can help verify the diagnosis and track the development of the disease over time. Unfortunately, analyzing brain scans by healthcare professionals such as doctors and radiologists is time-consuming and imposes a significant financial burden on healthcare systems.

To overcome these issues, a plethora of studies have suggested using AI tools, particularly machine learning (ML) models, to tackle data-based AD diagnosis[9,10]. Several ML models[11] such as random forest (RF), support vector machines (SVMs), and naive Bayes (NB) have been employed in the detection and diagnosis of AD at an early stage. However, these models undergo several limitations such as their requirement of developing good predictors/features for modeling tasks or an appropriate

tuning process of the model's parameters[12]. Deep learning (DL) architectures can extract highly abstract and representative features from raw data, resulting in a better generalization capability. Several researchers have attempted to diagnose AD using slice-based and voxel-based DL models. Based on the systematic reviews, voxel-based DL networks include 3D AlexNet, 3D ResNet, and patch-based models. Slice-based DL networks include ResNet, autoencoders, recurrent neural networks, and graph convolutional neural networks. However, DL architectures often require large quantities of high-quality training data for achieving good generalization capabilities, especially when dealing with highly complex problems like AD staging. Due to the difficulty of data acquisition and quality annotation in the medical field, the limited availability of data is often considered a major hindrance to AD classification using these highly-parametric models[13].

In scenarios with low availability of data, ensemble learning (EL) is often considered due to several benefits derived from the combination of the predictions of multiple models[14]. In order to successfully capture complex patterns in the training data while also avoiding too detailed and noisy features, single models typically fail to find the perfect balance. This can lead to poor generalization when applied to unseen data. The ML community has paid significant attention to the design of EL methods to mitigate the aforementioned challenges of single models. Ensemble methods combine the predictions of multiple individual models (known as *base models* or *base learners*), each trained on a different subset or perspective of the data[15]. By doing so, they harness the collective wisdom of these models, effectively reducing the risk of overfitting. Combining the diverse and unstable base models, the ensemble model is more stable and performs, in general, better than any of its constituent models.

As a result, EL is widely employed in several fields such as cancer classification[16–18], sequence analysis[19,20], proteomics data analysis[21], Ribonucleic acid (RNA) structure prediction[22], and among other biomedical applications. The intersection of EL with deep neural networks is known as ensemble deep learning (EDL). EL approaches help to address several key challenges of DL such as model selection problem (i.e., what is the best model for a certain classification problem?), using an ensemble of many different models rather than just one may reduce the likelihood of selecting a particularly undesirable performing model, vast amounts of data or lack of appropriate data (often encountered in AD with medical image data), overly complex decision boundaries, data availability acquired from many different sources that may give complementary information, class imbalance problem, and noisy data issues[23]. Dietterich identifies three main reasons for employing ensemble learning: the first being statistical, the second computational, and the third representational[14]. Moreover, EDL models combine the strengths of DL and EL to yield a model with improved generalization performance. Researchers are increasingly interested in EDL frameworks due to their ability to combine a wide variety of models and several feature representation methods.[24]. Traditional ML approaches and DL methods for AD detection have been extensively investigated and reviewed[10]; Nevertheless, exploring an EDL architecture/framework for the specific objective of AD detection remains relatively unexplored. This review article explores the foundations and latest developments in EDL, organizing contributions reported so far in this area into coherent categories. The ultimate goal of this categorization is to establish a referential guide to encourage research efforts in the rapidly growing field of EDL for AD. To this end, we provide a solid rationale for employing EDL for AD detection and subsequently examine existing EDL applications in this domain. Finally, we delve into the primary hurdles encountered in this area and propose potential research directions – emphasizing explainability, causality, and uncertainty estimation – that can inspire researchers to pursue further research and development endeavors to enhance AD diagnosis in clinical practice.

Fundamentals of Ensemble Deep Learning (EDL)

The ML community has paid significant attention to EL, which combines the outputs of several ML base models to improve their generalization performance[25]. Classical EL builds upon combining conventional ML models[26]. Despite its historical path of success, the main limitation of classical EL is the nature of conventional ML models, whose performance depends on hand-crafted features that are frequently challenging to build and insufficiently expressive[25]. DL aims to use hierarchical architectures to learn high-level abstractions from data. Research has shown the superiority of deep architectures (mostly deep neural networks (DNNs)) over shallow architectures across a variety of tasks[27,28]. The key distinction between DNNs and conventional ML methods is the automation of feature extractors. DNNs have proven to be effective at identifying complex structures in high-dimensional data and are thus successfully applied to complex tasks in diverse domains[28].

The success of DL has given rise to a new era in EDL research. The usual approach for developing an ensemble deep model is to employ DNNs into the classical EL framework, i.e., replace conventional ML models with DL models[29]. A few fundamental challenges that EDL addresses include small sample sizes, unequal class distribution, high complexity, and noisy and heterogeneous data produced from multiple domains. Deep models have high variance because they are very flexible, and there is a possibility that they might face local loss minima problems while being trained. To address these issues, researchers have shown that integrating the results of numerous DL models yields better generalization performance than a single DL model[30]. EDL methods are suitable in practice despite their increased complexity due to several reasons:

- **Improved robustness:** the combination of multiple deep learning models may capture a wider range of patterns and reduce model-specific errors, leading to more reliable and robust predictions.

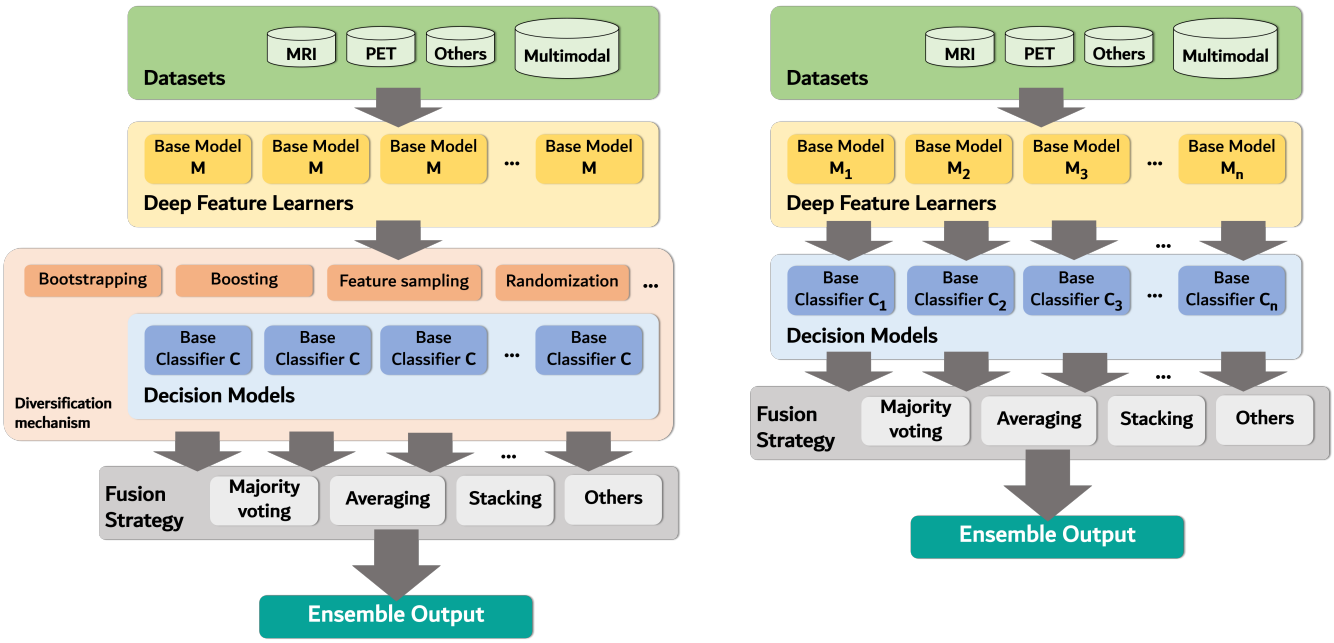
- **Straightforward measurement of uncertainty:** By measuring the level of disagreement between their base models, ensembles provide an estimation of the uncertainty associated to their predictions, which is critical in high-risk applications where decisions are made based on the output of these models (as in AD diagnosis).
- **Effective management of class imbalance:** By aggregating multiple models, one can ensure that the minority class in severely imbalanced datasets is well represented in the ensemble, leading to improved generalization performance and lower class bias. Techniques such as bagging can also assist in balancing class distributions within the ensemble.
- **Easy incorporation of new information:** To incorporate new knowledge, ensembles can be updated by retraining or adding new models. This flexibility makes ensembles useful in dynamic setups where data are subject to exogenous factors that imprint variability on them.
- **Reduced variance:** Ensembles reduce variance and provide reliable outcomes, which is important in situations where deep learning models exhibit high variance and produce unstable or inconsistent predictions.
- **Complex data patterns:** Ensembles are effective at capturing complex and non-linear relationships in data.

Moreover, diversity in DL models can be achieved in many different ways such as varying network architectures and layer configurations; different weight initialization schemes; different activation functions; regularization techniques (i.e., dropout, weight regularization, batch normalization); hyper-parameter settings (i.e., learning rate, solvers, mini-batch sizes); or randomization, among others. Improving the robustness and generalization performance of deep neural networks relies heavily on achieving diversity in the networks. This can be done by using a variety of techniques that are specific to the problem, the data, and the desired outcomes. Often, a combination of these methods is used to maximize diversity and successfully address complex and dynamic data patterns.

In general, key ensemble strategies can be categorized into homogeneous and heterogeneous ensemble approaches[31]. In order to develop an ensemble model, a pool of learners is strategically generated and aggregated together. An ensemble model composed of base learners (trained over different subsets of data) from the same family is known as an homogeneous ensemble model (Fig. 1 (a)). In contrast, an heterogeneous ensemble model is composed of base learners (trained over the same data) from different families (Fig.1 (b)). Diversity amongst the base learners of homogeneous models can be induced by using techniques such as bootstrapping the samples or sampling the feature space so that each base learner can be trained over different training sets. By contrast, heterogeneous ensemble models possess intrinsic diversity due to the base learners belonging to different families. An homogeneous approach might be subject to the same algorithmic biases and constraints. This may restrict their capacity to capture a wide range of patterns within the dataset. Therefore, when it is unclear which model is most appropriate for a given scenario, heterogeneous ensembles are often very helpful in practice. In contrast, heterogeneous ensembles pose greater implementation and fine-tuning challenges compared to homogeneous ensembles, which offer faster development and deployment. The No-Free-Lunch theorem[32] states that there is no model performing best across all possible collections of data. Selecting between homogeneous and heterogeneous ensembles should be determined by considering the particular problem and dataset at hand, available resources, base learners, and other functional and non-functional constraints of the scenario under consideration. In horizontal, vertical voting of deep ensembles[33], a series of classifiers are trained on the intermediate feature representation in vertical voting. The horizontal voting network is trained on a selectively stable epoch range, and the predictions of the top-level feature representation of the selected epochs are assembled. Temporal ensembles[34] use different augmentation of input features, regularisation, and training epochs to generate the ensembles, whereas ensembles for image classification[35] and disease prediction[36] train multiple networks. Training multiple DL architectures in an ensemble requires optimizing millions or billions of parameters, leading to high computational complexity that might not be feasible.

To improve the feasibility of deep ensemble models, the heterogeneous ensemble wherein traditional models are used in tandem with the DL architectures enjoys the benefits of lower complexity and the ensemble approach. For example, a heterogeneous ensemble for classification of text[37] employs NB, SVM, RF, and convolutional neural networks (CNN) models. Heterogeneous deep network fusion[38] uses different data, model, and decision fusion perspectives to generate an ensemble. To reduce the computation cost, implicit ensemble models have also been developed wherein a single model is trained to achieve a performance similar to that of multiple models. Here, random deactivation of the neurons and the layers is performed while training the network. Dropout[39] randomly deactivates the neurons, DropConnect[40] randomly drops the connection links, and Stochastic depth[41] randomly drops the residual blocks for generating the implicit ensemble predictions. Swapout[42] generalizes DropOut and Stochastic depth. Other implicit ensemble approaches include knowledge distillation[43] and Gradual (regularised) DropIn[44]. Usually, such strategies are, to a significant extent, computationally more affordable than training several DL models.

The idea of stacking[45] is to learn multiple models and combine them by training a meta-model. Unlike homogeneous approaches (e.g., bagging and boosting), which directly aggregate the outputs of several base models to achieve the final prediction, stacking is a special kind of ensemble learning strategy that combines several base learners (level-0 models) via



(a) Homogeneous ensemble **(b)** Heterogeneous ensemble

Figure 1. Schematic diagram illustrating different Ensemble Deep Learning strategies.

training a meta-model (level-1 model). In addition, stacking employs several base algorithms on the same data to generate models with unique perspectives on the input data. The meta-model can encompass any ML method, with the objective of acquiring knowledge on the most effective way to amalgamate the outputs generated by the underlying models. As a result, stacking ensembles can be homogeneous or heterogeneous, depending on whether the level-1 model belongs to the same family than the level-0 models.

Alzheimer's disease applications of ensemble deep learning

This section provides a categorization of AD-based EDL methods. The section provides insights into the various EDL approaches implemented by the various research articles. Tabulation is performed based on the data accessing approach into the model, i.e., slice-based or voxel-based. Slice-based approaches deal with models whose input data approach is a 2D slice instead of an entire 3D MRI scan. Similarly, voxel-based approaches consist of models in which the entire 3D neuroimage is adopted either directly in the models or instead, features are extracted from 3D scans. In delineating the data formats pertinent to the 3D voxel-based methodology versus the 2D slice-based approach, it is noteworthy that the former predominantly access neuroscans in the Neuroimaging Informatics Technology Initiative (NIfTI) format, denoted as ".nii", as well as in the Digital Imaging and Communications in Medicine (Dicom) format, represented as ".dcm". On the contrary, the latter relies on the extraction of slices from NIfTI scans, which are typically rendered in standard picture formats such as ".png" and ".jpg". Continuing the trajectory of feature extraction, in the realm of 3D data feature learning, DL models employ 3D convolutional architectures, incorporating specialized layers tailored for this purpose. Conversely, in the domain of 2D slice data, deep learners rely on models underpinned by 2D convolutional layers. These models may include pretrained models like ALexNet, ResNet, and similar exemplars. Figure 2 visualizes the feature learning process from slice-based and voxel-based approaches, highlighting their differences and similarities.

The slice-based approach primarily emphasizes the analysis of individual image slices, potentially providing a more granular assessment of regional variations within the brain. This approach may be particularly useful in scenarios where localized abnormalities hold clinical significance. On the other hand, voxel-based approaches consider information at a finer spatial resolution, capturing nuances within each voxel. This method enables a comprehensive examination of the entire brain volume, potentially revealing global patterns and relationships that may not be as readily apparent in a slice- based analysis. The choice between slice-based and voxel-based EDL approaches depends on several crucial factors. Firstly, it is contingent on the availability and nature of the neuroimaging data. A voxel-based approach may be preferable if the dataset primarily comprises volumetric scans. Secondly, computational resources play a significant role; voxel-based analyses require higher computational

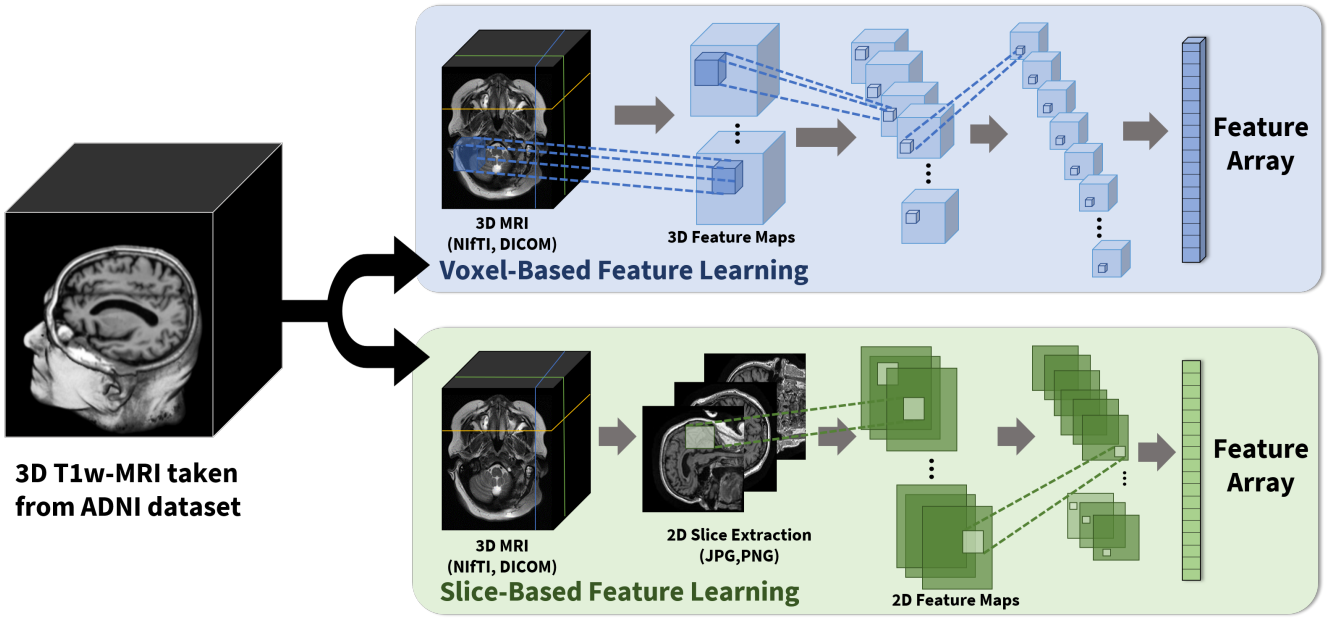


Figure 2. Feature Learning approaches from 3D MRI scans.

effort due to the increased data scales. In contrast, slice-based approaches are in general more computationally efficient. Lastly, the selection should align with the research objectives. A slice-based analysis offers detailed in- sights into specific brain regions, while a voxel-based approach provides a comprehensive view of the entire brain volume. Considering these factors, we aim to make an informed choice that best aligns with the unique characteristics and goals of the study, ultimately advancing on the understanding of AD. The main drawback of using slice-based DLs on 3D MRI scans is that they are unable to recognize the dependencies between voxels in MRI volumes. When 3D MRI scans are transformed into 2D image slices, there is a loss of data, particularly features related to the sizes and shapes of brain regions, which may cause a reduction in detail, potentially causing confusion about the image[46]. In contrast, the complexity of 3D DLs (voxel-based approaches) makes them difficult to train, as they require a large number of parameters, leading to the risk of overfitting[46]. This issue holds when constructing EDL models for 3D MRI scan data based on slice-based DL based models; nevertheless, the increased robustness of EDL models may counteract the aforementioned issue to an extent.

Slice-based Homogeneous EDL approach for AD detection

Single Modality

In the larger landscape of AD detection, innovative methodologies continue to surface, each unveiling novel dimensions for refined classification. By embracing diverse data modalities and strategic ensemble approaches, the journey toward effective diagnostic tools for AD progresses. As evidenced by Tanveer et al.[47], integrating VGG-16-based models within a heterogeneous framework paves the way for efficient AD detection. Their model's emphasis on transfer learning ameliorates computational constraints without compromising performance metrics. In pursuing improved diagnostic tools, Maji et al.[48] delve into single-modality approaches, unraveling avenues for enhanced classification using MRI data. Their investigation traverses different MRI planes, utilizing an ensemble of plane-specific features for accurate classification. Furthermore, Pan et al.[12] introduce a two-stage ensemble approach with a homogeneous EDL framework. In their research, a CNN algorithm is used to different 2D MRI slices, including sagittal, coronal, and transverse orientations. This leads to the creation of three classifier ensembles, with optimal base models chosen through validation. The subsequent integration of these ensembles through a majority voting mechanism yields a robust ensemble classifier. Another investigation by Razzak et al.[49] capitalizes on the strengths of DenseNet and its variants to implement PartialNet. This ensemble network layered in a hierarchical pipeline, leverages MRI inputs for AD detection. Integrating diverse depths and deep supervision within PartialNet contributes to its potential advantages. An inventive strategy surfaces in the form of slice-wise ensembling, as proposed by Wang et al.[50]; their methodology harnesses multiple VGG-19-based models to yield preliminary predictions for distinct slices. These predictions converge within an AlexNet-based model, culminating in a refined final prediction. Although their performance metrics show promise, the study's scope is constrained by using a limited dataset.

In a distinct avenue, the exploration of AD detection extends to PET imaging. Zheng et al.[51] advance an ensemble technique

employing the AlexNet algorithm. By leveraging 62 anatomical brain slices through automated anatomical labeling (AAL) cortical parcellation, they fine-tune a pretrained AlexNet to capture pertinent features. Converting DL and anatomical data presents a pathway to enhanced AD detection through PET imaging.

Multimodality

Venturing into the domain of AD detection, researchers have embarked on a journey of innovation through multimodal methodologies. A fusion of diverse data sources—genetic markers, MRI, and PET scans—unveils a promising path for enhanced classification accuracy. In pursuing a holistic understanding, Zeng et al.[52] pioneer a distinctive MRI and single nucleotide polymorphism (SNP)-based multimodal strategy. This approach delves into the intricate landscape of genome biomarker prediction by intertwining genetic insights with neuroimaging data. Similarly, in[53], authors bridge MRI and 18FDG PET scans, harnessing the synergy of these modalities. Their method, involving Wavelet-transform-driven slice fusion, culminates in feature learning and classification fusion via CNN and Random vector functional link (RVFL) classifiers. While potential thrives within, the challenge of dataset constraints and model complexity emerges. Meanwhile, Zhang et al.[54] pivot toward a modified ResNet-50 model, orchestrating feature harmonization between MRI and PET scans. The algorithm dynamically allocates fusion ratios by introducing an attention model, navigating the intricate interplay between depth and feature space. This landscape, propelled by interwoven data and forward-looking models, echoes a chorus of potential. The march toward comprehensive AD detection gains momentum through genetic markers, neuroimaging data, and cutting-edge architectures. A leap into the frontier of modern models beckons with Tang et al.[55]. Their venture into Vision Transformer (ViT) ensembles marks a decisive move, unifying insights extracted from MRI and PET datasets.

Homogeneous ensemble techniques are trained using a single type of base classifier, which may lead to a bias toward certain aspects of the dataset[56]. In order to ensure that the classification error converges to its asymptotic value, a homogeneous ensemble employs a large number of classifiers. As a result, classifiers require in general a large amount of memory for their persistence, whereas inference consumes significant computing power for every test case.

Slice-based Heterogeneous EDL approach for AD detection

Single Modality

In AD detection, diverse strategies harness MRI data for early diagnosis. Choi et al.[57] present an ensemble of AlexNet, GoogleNet, and VGG16 models, employing three-plane projections. Complexity emerges, yet computational concerns persist due to undefined slice selection and a probability-based classification approach. Ji et al.[58] adopt a three-model ensemble, CONVNet, with ResNet, NasNet, and MobileNet. Each pretrained CNN-based model receives 20 slices. Commendable accuracy for three-class classifications is achieved, yet challenges in slice selection clarity arise. Authors in[59] proposed a multimodal ensemble with GAN, VGG-16, and ResNet-50. Non-consecutive 2D slices pose information loss, as only the coronal plane is employed for ensemble learning.

Leveraging transfer learning in[60], implementing five efficient CNN models—VGG19, Inception-ResNetv2, ResNet152v2, EfficientNetB5, and EfficientNetB6—alongside a custom model. A weighted average ensemble of all six classifiers enhances performance. Another work explores transfer learning through multiple DenseNet variants[61]. The majority of voting combines insights from three MRI planes. Jabason et al.[62] merge DenseNet-201 and ResNet50 for each view, leveraging pretrained weights for commendable performance. Khanna[63] combines CNN, LSTM, and MobileNet for single-modal data. Using tri-transfer learning, Yang et al.[64] fuses SVM, Softmax, and DNN-based classifiers.

Other than T1-W MRI modality, some works incorporate other data types such as Functional MRI (fMRI) and Magnetoencephalography (MEG). Sethuraman et al.[65] tackle fMRI data with customized AlexNet and Inception-v2 ensembles. Traditional classifiers intersect their decision-making, though classification challenges arise. Following the trend, authors in[59] merge MRI and MEG data, employing deep transfer modeling and ensemble classification with an AlexNet-based feature extractor. Performance metrics surface, though limited dataset size poses considerations.

Multimodality

In the quest to amplify AD detection, diverse strategies converge in exploring multimodal approaches, seamlessly merging different data sources for improved accuracy. Ying et al.[66] pave the way with a CNN-multilayer perceptron ensemble model, integrating MRI and SNP information. The intrinsic synergy of this multimodal approach augments feature scale by introducing complementary insights. Challenges emerge from a relatively constrained dataset for training and testing. Ismail et al.[67] further the narrative with MultiAz-Net, an ensemble of three DNN models intricately weaving information from PET and MRI fused images. Their architecture's threefold journey—image fusion, feature extraction, and classification—unfolds precisely. To optimize model layers, a Multi-Objective Grasshopper Optimization Algorithm is introduced. Strides toward enhanced AD detection resonate in this intricate landscape of multimodal methodologies. By harmonizing distinct data types, these endeavors bring us closer to more accurate diagnostic tools for AD.

Heterogeneous ensembles harness the benefits of different base models that successfully learn distinctive characteristics of the training data to provide in general greater generalization performance than their homogeneous counterparts. On the negative side, the development of a heterogeneous ensemble poses various challenges that must be carefully addressed, including the selection of diverse and complementary base models, the determination of an optimal set of weights to combine the outputs of base models, and the identification and selection of an optimal subset of classifiers from the ensemble.

Slice-based Stacking EDL approach for AD detection

Multimodality

The methodologies discussed, designed to harness the potential of multimodal data, present a tapestry of strategies to enhance accuracy. Yang et al.[68] delve into the fusion of MRI and PET scans, orchestrating an ensemble through a VGG-16 based model. This integration unfolds using two distinct data sources, Alzheimer's Disease Neuroimaging Initiative (ADNI) and Japanese-ADNI (J-ADNI). Despite this approach's promise, the challenge of a constrained feature set surfaces, leading to an accuracy of 82.5%. Nevertheless, a direct comparison of performance proves challenging due to the substantial disparities in both the models employed and the dataset characteristics. Despite this, it is worth noting that achieving the stated accuracy is relatively modest when contrasted with alternative slice-based models that have achieved superior levels of accuracy. Similar intent, Fang et al.[69] opt for an ensemble route, uniting pre-trained networks—GoogleNet, ResNet, and DenseNet—across MRI and PET domains. The integration of Adaboost as the final prediction classifier offers potential yet introduces heightened computational demands due to the utilization of multiple deep models.

Voxel-based Homogeneous EDL approach for AD detection

Single Modality

Within the realm of AD detection, using EDL algorithms for single-modality MRI analysis unveils a series of diverse strategies to enhance diagnostic accuracy. Ahmed et al.[13] navigate the intricacies of ensemble approaches, employing CNNs for multiple datasets— Genetic and Rare Diseases (GARD) dataset for training and ADNI for testing. Through patch-based classification, authors extract 32*32 patches from various regions of interest (ROI) within sMRI. While commendable, this strategy's full potential is yet to be unlocked through whole-brain computation. Embarking on a journey to comprehend progressive mental deterioration, in[70], authors introduce a DL approach involving ensemble learning methods and deep neural networks. Using CNN with dual region input culminates in understanding brain region correlations leading to AD. AdaBoost further enhances the saliency of region pairs towards AD correlation.

Suk et al.[71] harnesses the power of deep ensemble learning of sparse regression models, weaving multiple sparse regression models trained with different regularization parameters into the framework. The integration of CNN yields non-linear weights, enhancing AD vs. CN, MCI vs. CN, and progressive MCI vs. stable MCI classification. Wang et al.[72] curate MRIs through gradient drift and grad-warping, removing extraneous tissues before alignment with a standard template. Through 3D-DenseNet models and probability-based fusion techniques, their ensemble approach diagnoses AD and MCI, highlighting the potential of this fusion method. Similarly, Ruiz et al.[73] explore 3D densely connected neural networks, ensembling them to predict early MCI, late MCI, and AD. Simplicity emerges as a key to success, achieved through training with lower parameters that amplify gradient and data flow, resulting in an accuracy of 83.33%. However, the attained accuracy is relatively low when compared to the other single-modal homogeneous approach, thereby leading to the scope for improvement.

A homogeneous ensemble model, proposed in[11], with 11 autoencoders designed for various template-based preprocessed images. The model attains an impressive accuracy of 95%, yet limitations stem from the relatively small dataset and random selection of MRI scans, suppressing the full potential of ensemble methodology. Exploring a novel direction, Chen et al.[74] introduce a modified 3D pretrained network-based model, integrating deep feature extraction and ensembling based on relevant brain regions. While innovative, this approach's computational cost reaches a higher threshold. Malik et al.[75] pioneer an EDL model leveraging volume-based analysis through T1-Weighted MRI scans from the ADNI dataset. Deep RVFL networks, coupled with a graph embedding technique, spearhead feature learning and classification. This approach, however, may find limitations in large multidimensional datasets. Ganaie et al.[76] echo this pursuit, presenting an ensemble deep RVFL network coupled with the LUPI technique for feature classification.

Beyond the intricacies of EL, Colbaugh et al.[77] embrace transfer learning via stacked autoencoders, refining predictions to optimize the use of target data. This approach's effectiveness in blood-borne microRNA-based AD diagnosis underscores its potential. The journey into Voxel-based methods introduces the potential for high-dimensional 3D information extraction from brain scans. However, this comes hand in hand with high computation loads, raising challenges in handling local information intricacies within these scans.

Multimodality

In an endeavor to elevate AD diagnosis, the exploration of multimodal EDL approaches emerges. A cascade of strategies, interwoven through deep neural networks, reflects the pursuit of improved diagnostic accuracy.

Lu et al.[78] embrace a multimodal ensembling approach, meticulously extracting patches from both MRI and PET. The model's training navigates the fusion of insights from both modalities through multiscale feature extraction. The ensemble of six deep neural networks facilitates classification training, ultimately channeling into an ensemble probability-based classifier for decisive calculation. Despite its promise, reconciling classification probability output with source MRI and PET scans poses nonlinear transformation challenges.

In a similar bid, Zhang et al.[79] present a unique multi-modal cross-attention AD diagnostic (MCAD) paradigm. This model leverages sMRI, FDG-PET, and CSF biomarkers to enhance AD diagnosis. Modality interactions are effectively captured by integrating cascaded dilated convolutions and CSF encoders, refining the diagnostic process. Diversifying strategies, El et al.[80] craft an ensemble involving CNN and Bi-LSTM for multimodal heterogeneous AD classification. Incorporating MRI and PET as inputs, this approach underscores joint learning through CNN and Bi-LSTM layers. While classification and regression are executed using reduced features, the model's computational demands are heightened by including Bi-LSTM.

Voxel-based Stacking EDL approach for AD detection

Single Modality

Another heterogeneous model for voxel-based ensemble learning is proposed in[81]. The authors implement two sparse AE followed by Deep Belief Networks as a feature extractor. The probability-based classifier is used as the final decision-making tool. However, the approach is less likely to miss an AD diagnosis, which costs relatively more than misdiagnoses in primary care settings.

Multimodality

A collection of deep belief networks is subsequently constructed and proposed in[82], with the ultimate forecast being selected using a voting mechanism. In this study, two deep learning architectures are developed and evaluated, along with four distinct voting systems. The findings demonstrate the effectiveness of the proposed classification framework, which leverages unsupervised learning to compute discriminative features.

In various contexts demanding automated learning models, it is customary to encounter data sourced from diverse origins, often offering complementary insights. The sophisticated integration of such insights exhibits the potential to enhance the precision of classification decisions, as opposed to relying solely on individual data sources. In the diagnosis of AD, a researcher leverages inputs such as EEG readings (capturing one-dimensional time series data) and neuroscans from MRI, or PET scans (yielding two-dimensional spatial data). Additionally, considerations encompass the levels of specific chemicals in the cerebrospinal fluid, alongside demographic factors like age, gender, and educational attainment of the subject, which manifest as scalar and categorical values. Attempting to amalgamate these disparate features into a singular training set for a classifier proves unwieldy. Hence, the recourse often lies in employing an EDL approach. In this approach, each distinct modality undergoes independent training with dedicated homogeneous or heterogeneous learners. Subsequently, the verdicts rendered by each learning model can be harmonized utilizing any of the combination methodologies discussed previously (namely, bagging, boosting or stacking).

Challenges and future research directions

While the field has reached a certain level of maturity, our critical analysis has uncovered several research areas that still require further exploration and attention from the scientific community. In light of these findings, we present a comprehensive outline of these challenges and potential research directions that offer promising avenues for effectively addressing them (refer to Figure 3 for a visual overview).

Integration of medical knowledge-driven features with data-driven features

Incorporating medical knowledge-based features such as genetic, environmental, and behavioral variables into brain-driven features extracted from neuroimaging data for the detection of AD is a current research focus. The aim is to develop more accurate detection frameworks to identify clinically homogeneous groups of AD patients and to improve our understanding of the underlying disease processes[85,86]. ML has been used in a number of studies to blend medical knowledge-based features, brain imaging, neuropsychological tests, and other biomarkers in order to predict the stages of AD[87,88]. The integration of these two types of features can enhance AD research and diagnosis in several ways:

- Improved diagnostic accuracy: Medical knowledge-driven features can provide a strong foundation for understanding AD's clinical manifestations and diagnostic criteria. By incorporating data-driven features, such as neuroimaging biomarkers or genetic risk factors, the accuracy of diagnostic models can be enhanced, leading to more precise and early detection of the disease.

Table 1. Classification of the EDL-based approaches for AD detection reviewed in this study.

Ensemble Type/ Input Approach	DL Architectures	Modalities Used		
		Single Modality		Multi Modality
		MRI	Others	
Homogeneous/Slice-based	VGG-19	[50]	-	-
	CNN Based	[12]	-	[52]
	CNN+RVFL	-	-	[53]
	PartialNet	[49, 83]	-	-
	VGG-16	[47]	-	-
	Vision Transformer	-	-	[55]
	AlexNet based	-	[51]	-
	ResNet-50	[48]	-	[54]
Homogeneous/Voxel-based	3D CNN Based	[13, 70, 71]	-	-
	3D DenseNet	[72, 73]	-	-
	Autoencoder based CNN	[11]	[77]	-
	3D ResNet-10	[74]	-	-
	Deep NN	-	-	[78]
	Dilated CNN	-	-	[79]
	CNN+LSTM	-	-	[80]
	edRVFL	[75, 76]	-	-
Heterogeneous/Slice-based	VGG-16+AlexNet+GoogleNet	[57]	-	-
	ResNet-50+NasNet+MobileNet	[58]	-	-
	VGG-16+GAN+ResNet-50	[84]	-	-
	VGG19+Inception- ResNetv2+ ResNet152v2+EfficientNetB5+EfficientNetB6	[60]	-	-
	CNN+MLP	-	-	[66]
	DenseNet-121+DenseNet-161+DenseNet-169	[61]	-	-
	DenseNet-201+ResNet-50	[62]	-	-
	MobileNet+LSTM	[63]	-	-
	DNN+SVM+Softmax	[64]	-	-
	AlexNet+Inception-v3+ResNet-18	-	-	[67]
	AlexNet+Inception-v2	-	[65]	-
	AlexNet+LDA+SVM	-	[59]	-
Hetrogeneous(Stacking)/Slice-based	VGG-16	-	-	[68]
	GoogleNet+ResNet-50+DenseNet	-	-	[69]
Hetrogeneous(Stacking)/Voxel-based	3D CNN Based	-	[81]	-
	DBN	-	-	[82]

Figure 3. Graphical representation of challenges and future directions of Ensemble Deep Learning for AD diagnosis.

- Enhanced risk prediction: Integrating medical knowledge-driven features with data-driven features allows for the development of comprehensive risk prediction models. By considering both established risk factors (e.g., age, family history) and emerging biomarkers identified through data analysis, clinicians can better assess an individual's risk of developing AD.
- Mechanistic insights: AD is a complex disorder with multiple underlying mechanisms. Integrating medical knowledge-driven features with data-driven features can help uncover novel disease mechanisms, identify potential therapeutic targets, and guide the development of new treatment approaches.
- Longitudinal monitoring: Combining medical knowledge-driven features with data-driven features enables disease progression and treatment response monitoring over time. By integrating clinical assessments, neuroimaging data, and other biomarkers, researchers and clinicians can track changes in patients' cognitive abilities, brain structure, and biochemical profiles, providing valuable insights into the effectiveness of interventions.

To realize the above integration, several research areas in ML and AI should receive increasing attention from the community, including physics-informed neural networks, the exploitation of knowledge graphs and semantic databases for ML, multimodal data fusion, and in general, all learning mechanisms that allow incorporating expert knowledge during data modeling. A current research focus is incorporating medical knowledge-based features such as genetic, environmental, and behavioral variables into brain-driven features extracted from neuroimaging data for detecting AD. The aim is to develop more accurate detection frameworks to identify clinically homogeneous groups of AD patients and to improve our understanding of the underlying disease processes. ML has been used in a number of studies to blend medical knowledge-based features, brain imaging, neuropsychological tests, and other biomarkers to predict the stages of AD.

Multimodal Data Analysis

This second challenge hinges on effectively incorporating new data modalities into AD characterization, using multiple data sources for this purpose. As the understanding of AD expands, it becomes increasingly crucial to integrate diverse data types

beyond traditional clinical assessments and neuroimaging. New data modalities, such as neuroimaging biomarkers and omics data (genomics, proteomics, metabolomics), offer valuable insights into the disease progression and underlying mechanisms. Incorporating these data modalities via EDL is relevant for AD characterization to complement feature analysis and improve generalization. However, challenges remain around data quality, computational cost, standardization, and the need for robust analytical frameworks to leverage the potential of these emerging data sources.

EDL techniques can be employed to analyze these data modalities effectively. EDL combines the power of EL and DL, enabling the study of diverse data modalities in AD research. Here are some examples of emerging data modalities and how they can be approached through the utilization of EDL:

- Neuroimaging Data: Different neuroimaging modalities like MRI, PET, and fMRI offer detailed information on brain structure, metabolism, and functional connectivity. DL models that use ensemble techniques are predominantly built on MRI data, as these scans are more accessible than PET or fMRI scans. PET, however, measures glucose metabolism to provide metabolic information. fMRI reveals alterations in brain activity that are associated with the onset of AD's early symptoms. Incorporating all this information together will improve the diagnosis accuracy of AD. EDL can be utilized for integrating structural, metabolic, and functional data by training multiple DL models to extract relevant features from different imaging modalities. The ensemble of models can capture complementary aspects of AD-related changes, leading to improved disease characterization and prediction.

- Omics Data: Integrating multiple omics data modalities, such as genomics, proteomics, and metabolomics, provides a holistic AD view. Genomic study investigates the association between genetic variations and gene expression patterns associated with AD risk and progression. Proteomics is the comprehensive investigation of protein structure, functionality, and information transmission within the cellular milieu of an organism. Metabolomics are small chemicals that are produced during metabolism which include amino acids, fatty acids, and carbohydrates. Integrating genomics, proteomics, and metabolomics data using EDL will help to learn from each modality, collecting information that performs better and improving AD prediction.

In the extent of AD diagnosis, the practice of using 2D slice data over 3D voxel data within EDL brings forth a spectrum of opportunities and hurdles. On the positive side, the 3D voxel data allows for the seamless integration of contextual information and heightens sensitivity to subtle structural changes. Nonetheless, it also introduces a set of challenges including escalated computational requisites, intricacies in data preprocessing, and the need to effectively integrate data from various imaging modalities. The path forward entails a focus on refined fusion methodologies, leveraging transfer learning and pretraining, and optimizing both hardware and software components. Successfully navigating the intricacies of 3D voxel data, particularly in the context of multimodality, emerges as a pivotal avenue for forthcoming research in EDL for AD. This endeavor holds promise for markedly amplifying the precision and dependability of diagnostic models.

Retina is an anatomical outgrowth of the brain with many of the same characteristics, such as an embryologic origin, a complicated neurotransmitter pathophysiology, a precise neural cell layer, blood vessels, microvasculature, and microglia. Researchers have investigated the visual changes in AD patients, such as altered contrast sensitivity, aberrant color vision, and visual field defects, among other prognostic factors. In the mouse model, researchers have also found protein aggregation involving A and tau proteins in the retina. Structural changes have been also identified including decreased choroidal thickness, increased macular thickness and volume, and thicker retinal nerve fiber layer (RNFL). The use of retinal biomarkers in conjunction with neuroimaging modalities using EDL for their fusion and modeling can aid in the early identification and monitoring of AD.

Leveraging EDL techniques allows researchers to effectively analyze these new data modalities in AD characterization. The ensemble nature of EDL enables the capture of complementary information, robustness against noise, and improved generalization. However, it is crucial to note that the application of EDL to AD characterization necessitates careful model selection, hyperparameter tuning, and validation to ensure reliable results and interpretability of ensemble predictions.

Computational Efficiency

Training an ensemble of independent models with homogeneous or heterogeneous techniques is a reasonably easy way to benefit from parallel processing. However, even though it is simple to parallelize, the amount of computing required during training time is expensive if the individual models are huge deep architectures and the datasets are very huge. Therefore, applying computationally expensive complex EDL models for diagnosing AD may not be feasible. Knowledge distillation has become increasingly popular in this area to transfer domain-specific knowledge and/or knowledge captured by complex ML models to simpler models that are easier to understand and eventually trust[43,89]. It is an ML technique used to transfer knowledge from a complex or deep model to a simpler, computationally lighter, and/or more interpretable model. Therefore, to overcome the problems with AD detection, there is a huge possibility of designing appropriate EDL-based architectures. We

refer to ensemble designs that are "incremental", so that new data can be learned efficiently and diversification between the knowledge captured by the models in the ensemble should be controlled in an intelligent way to avoid constructing ensembles by "brute force" (i.e., let bagging or sampling induce diversity at random, without any control). This latter strategy makes sense when the learners can learn fast (e.g., trees in an RF), but with DNNs, the ensembling technique has to diversify knowledge more intelligently[90].

Model Explainability

DL models have shown outstanding performance in various tasks[91]. However, explaining their actions is difficult because of their hierarchical non-linearity and black-box behavior. Deep models are challenging to interpret, which raises serious questions regarding their reliability in crucial prediction tasks like AD detection[92]. The mixture of numerous deep models exacerbates this problem. To increase model transparency during both the design and evaluation processes, interpretability and explainability methodologies should be adopted[93]. Therefore, it is highly desirable to develop EDL frameworks that are easy to understand and meet the requirements for a trustworthy actionability of their outcomes in clinical practice[94].

In AD, it is important to analyze which parts of the brain are more likely to be affected by the disease in addition to making the disease's diagnosis. Model interpretability is, therefore, crucial as we go from predictive to preventative aspects, increasing the utility and trust of the clinical practitioner in the model. Further along this line, model interpretability can accelerate manual annotation processes by easing the inspection of visual predictors within medical data, concentrating annotation efforts invested by medical experts on those instances characterized by the highest uncertainty and less precise explanations. Interaction between diagnostic models relying on EDL and practitioners can exploit the inherent capability of these models to aggregate the uncertainty and explanations of their compounding DL models.

Novel Architectures and Training Strategies

Determining an optimal single or hybrid architecture is essential for attaining the best performance in a given area and application[95]. Deep learning has made great progress in learning from enormous amounts of data. However, deep neural networks exhibit poor generalization and inconsistent performance over small datasets. The primary feature of ensemble approaches is stability. A range of EDL techniques that specifically addressed the challenges of small datasets by utilizing this crucial property allowed for the employment of deep learning with ensemble techniques in AD detection. With modern ensemble techniques and a broad spectrum of DL models, there is tremendous potential to design novel architectures that are suitable to cope with issues associated with small datasets for AD detection. Learning how to adjust the parameters in DNN architectures is a developing field in computer science. Plenty of parameters in DNNs need to be updated. An increase in the number of hidden nodes also increases the likelihood that the algorithm will get stuck in the local optimum. Novel training algorithms can extract features and reduce the loss of information to overcome both the local optima and the curse of dimensionality issues[96]. Existing medical data is not completely utilized by machine learning mainly because it resides in data silos, and access to this data is restricted due to privacy issues. However, EDL will not reach its full potential without access to enough data. Federated learning[97] can offer a solution to the high demands of EDL for AD disease characterization and estimation, allowing several remote models in different hospitals and medical centers to leverage each other's modeled knowledge while guaranteeing the privacy of locally captured medical data. To this, Federated Learning comprises different algorithms to collect, distribute, aggregate, and update locally learned models, sharing information about the model's parameters themselves while keeping data localized and private, and avoiding sharing them during the process.

Furthermore, DNNs, which use gradient-based training procedures, possess several additional difficulties: getting stuck in local minima, having millions or billions of tunable parameters, and slow training convergence. To address these issues, there are multiple approaches such as randomization-based neural nets, forward-forward methods for neural training and deep ensemble decision trees, often known as deep forest (gcForest)[98], which pose new challenges in terms of their architectural optimization, epistemic uncertainty and interpretability. Besides a lower training latency, pretrained features alongside these alternative models can also offer reduced inductive biases when compared to a neural network trained from scratch over a new dataset. Exploring the tradeoffs between complexity, performance and explainability of different EDL flavors, including the more interpretable ensembles of decision trees, is an uncharted avenue for future research.

Region of interest-based analysis and association with cognitive decline

Over the last few years, EDL models have been extensively utilized to distinguish between AD patients and CN. It is imperative to take into account any other medically relevant aspects. One such area of focus is the prediction of conversion from mild cognitive impairment (MCI) to AD, differentiating between stable and progressive MCI. Notably, the U.S. Food and Drug Administration (FDA) approved Aducanumab (Aduhelm) as a treatment for AD. Aduhelm is a treatment that can be given to individuals with mild dementia or MCI; it reduces the amyloid plaques in the brain. However, it is vital to determine the fitness of Aduhelm for MCI patients, as it is particularly recommended for those at risk of dementia caused by AD (termed as progressive MCI). Thus, accurately identifying MCI patients who are at risk of progressing to dementia through the use of EDL

models could prove to be of great value in clinical settings. In addition, an encouraging research direction is using EDL models to detect people with Subjective Memory Complaints (SMC) in the early stages. Recognizing the early signs of cognitive decline plays a pivotal role in early detection, predictive value, research cohort selection, and intervention opportunities for AD. In addition to AD diagnosis, EDL models can also be employed in other aspects of AD research, such as the analysis of white matter hyperintensities (WMH)[99]. The presence and severity of WMH are associated with an increased risk of developing AD or experiencing cognitive decline. EDL models aim to improve the accuracy and reliability of AD diagnosis by leveraging multiple DL architectures and combining their predictions. The ROI-based approach allows for a targeted investigation of specific brain regions implicated in AD pathology. Despite their potential, there is still room for further exploration and optimization of ensemble methods for ROI-based analysis in AD diagnosis. Future research in this area can delve into selecting relevant and most affected brain regions. It can unveil variations in the affected brain regions based on the disease progression. Therefore, future research efforts can be directed toward case studies investigating the impact of different brain regions and shedding light on the intricate relationship between brain atrophies and demographic factors. For example, accumulation of $A\beta$ plaques in the brain has been hypothesized as an early damaging event in the pathogenesis of AD. The deposition of plaques and neurofibrillary tangles mostly occurs in specific brain regions, including the hippocampus, entorhinal cortex, amygdala, and basal forebrain. These regions are known to play a significant role in memory, learning, and emotional behaviors. Region of interest (ROI) based analysis can assist in the prediction of conversion from mild cognitive impairment (MCI) to AD, differentiating between stable and progressive MCI. ROIs are identified in a voxel-based study by using brain atlases like the Harvard-Oxford Atlas and the Anatomical Labeling (AAL) Atlas using the FreeSurfer toolbox. In the slice-based study, 2D slices can be extracted from a 3D volume image and segmented into white matter, gray matter, and CSF for region-based analysis. Using statistical analysis, such as p-value, the most significant ROIs can be detected for AD staging. The most significant ROIs can be ensembled for predicting MCI to AD using EDL. In this effort, rather than using multimodal information, experiments are performed on sMRI only but with more significant information.

In the extent of AD diagnosis, the practice of using 2D slice data over 3D voxel data within EDL brings forth a spectrum of opportunities and hurdles. On the positive side, 3D voxel data allow for the seamless integration of contextual information and heightens sensitivity to subtle structural changes. Nonetheless, it also poses a set of challenges, including escalated computational requirements, intricacies in data preprocessing, and the need for effectively integrating data from various imaging modalities. The path forward entails a focus on refined fusion methodologies such as cross-modal attention mechanisms or deep feature fusion architectures, thereby effectively combining information from multiple modalities, leveraging transfer learning and pretraining, and optimizing both hardware and software components. Successfully navigating the complexities of 3D voxel data, particularly in the context of multimodality, emerges as a pivotal avenue for forthcoming research in EDL for AD. This endeavor holds promise for markedly amplifying the precision and dependability of diagnostic models.

Concluding remarks

Alzheimer's disease is one of the mainsprings of fatality, particularly in developing nations. Because the precise prognosis of Alzheimer's disease in clinics is complex, using a computer-based diagnosis approach in conjunction with medical specialists has much to encourage it to identify Alzheimer's disease. Ensemble deep learning has received a lot of emphasis in recent years for this goal. In this study, we have shown how ensemble learning has aided in developing AD prediction systems. This paper began with a description of Alzheimer's disease and its prodrome, followed by an overview of the contemporary diagnostic standards and associated biomarkers such as MRI, PET, and fMRI. Merging multiple neuroimaging modalities can improve the prediction of Alzheimer's disease. It can be used in association with other parameters, such as Neuropsychic tests and genetic data, to furnish a meticulous assessment of the disease.

Regarding model complexity, slice-based approaches are more economical than voxel-based approaches due to their potential to handle 2D neuroscans with low computational power compared to 3D scans. This work has covered a wide range of deep neural models. Regarding the classification approach, CNN-based pre-trained models have been employed most frequently, with higher-reported performance metrics in this field. An efficient multi-modal longitudinal method is suggested as the ultimate goal for an AD prediction system. However, low dataset issues and the model characteristic subject can be resolved using ensemble deep learning-based approaches for a tradeoff with computational power.

We hope that our review has sparked interest in ensemble deep learning techniques for AD detection, settling a landmark to assess where we currently stand and suggesting fresh research directions to coherently advance in years to come.

References

1. Huang, Y. *et al.* A machine learning approach to brain epigenetic analysis reveals kinases associated with Alzheimer's disease. *Nat. Commun.* **12**, 1–12 (2021).

2. DeTure, M. A. & Dickson, D. W. The neuropathological diagnosis of Alzheimer's disease. *Mol. Neurodegener.* **14**, 1–18 (2019).
3. Gaugler, J. *et al.* 2022 Alzheimer's disease facts and figures. *Alzheimers & Dementia* **18**, 700–789 (2022).
4. Lee, G., Nho, K., Kang, B., Sohn, K.-A. & Kim, D. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Sci. reports* **9**, 1–12 (2019).
5. Seitz-Holland, J. *et al.* Major depression, physical health and molecular senescence markers abnormalities. *Nat. Mental Heal.* **1**, 200–209 (2023).
6. Scheltens, P. Mild cognitive impairment—amyloid and beyond. *Nat. Rev. Neurol.* **9**, 493–495 (2013).
7. Yu, B., Shan, Y. & Ding, J. A literature review of MRI techniques used to detect amyloid-beta plaques in Alzheimer's disease patients. *Annals Palliat. Medicine* **10**, 10062–10074 (2021).
8. Bao, W., Xie, F., Zuo, C., Guan, Y. & Huang, Y. H. PET neuroimaging of Alzheimer's disease: radiotracers and their utility in clinical research. *Front. Aging Neurosci.* **13**, 624330 (2021).
9. Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A. & Davatzikos, C. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* **155**, 530–548 (2017).
10. Ebrahimighahnavieh, M. A., Luo, S. & Chiong, R. Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. *Comput. Methods Programs Biomed.* **187**, 105242 (2020).
11. Hedayati, R., Khedmati, M. & Taghipour-Gorjikolaie, M. Deep feature extraction method based on ensemble of convolutional auto encoders: Application to Alzheimer's disease diagnosis. *Biomed. Signal Process. Control.* **66**, 102397 (2021).
12. Pan, D., Zeng, A., Jia, L., Huang, Y., Frizzell, T. & Song, X. Early detection of Alzheimer's disease using magnetic resonance imaging: a novel approach combining convolutional neural networks and ensemble learning. *Front. Neurosci.* **14**, 259 (2020).
13. Ahmed, S., Kim, B. C., Lee, K. H., Jung, H. Y. & Alzheimer's Disease Neuroimaging Initiative. Ensemble of ROI-based convolutional neural network classifiers for staging the Alzheimer disease spectrum from magnetic resonance imaging. *PLoS One* **15**, e0242712 (2020).
14. Dietterich, T. G. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, 1–15 (Springer, 2000).
15. Dong, X., Yu, Z., Cao, W., Shi, Y. & Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **14**, 241–258 (2020).
16. Grewal, J. *et al.* Application of a neural network whole transcriptome–based pan-cancer method for diagnosis of primary and metastatic cancers. *JAMA Netw. Open* **2**, e192597–e192597 (2019).
17. Karim, M. R., Rahman, A., Jares, J. B., Decker, S. & Beyan, O. A snapshot neural ensemble method for cancer-type prediction based on copy number variations. *Neural Comput. Appl.* **32**, 15281–15299 (2020).
18. Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S. & Sidow, A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat. Commun.* **9**, 1–14 (2018).
19. Bartoszewicz, J. M., Seidel, A., Rentzsch, R. & Renard, B. Y. Deepac: predicting pathogenic potential of novel dna with reverse-complement neural networks. *Bioinformatics* **36**, 81–89 (2020).
20. Cao, Z., Pan, X., Yang, Y., Huang, Y. & Shen, H.-B. The lnclocator: a subcellular localization predictor for long non-coding rnas based on a stacked ensemble classifier. *Bioinformatics* **34**, 2185–2194 (2018).
21. Zohora, F.T., Rahman, M.Z., Tran, N.H., Xin, L., Shan, B., & Li, M. Deepiso: A deep learning model for peptide feature detection from lc-ms map. *Sci. Reports* **9**, 1–13 (2019).
22. Singh, J., Hanson, J., Paliwal, K. & Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.* **10**, 1–13 (2019).
23. Polikar, R. Ensemble learning. *Ensemble machine learning: Methods applications* 1–34 (2012).
24. Cao, Y., Geddes, T. A., Yang, J. Y. H. & Yang, P. Ensemble deep learning in bioinformatics. *Nat. Mach. Intell.* **2**, 500–508 (2020).
25. Yang, Y., Lv, H. & Chen, N. A survey on ensemble learning under the era of deep learning. *Artif. Intell. Rev.* **56**, 5545–5589 (2023).

26. Guo, C., Liu, M. & Lu, M. A dynamic ensemble learning algorithm based on k-means for icu mortality prediction. *Appl. Soft Comput.* **103**, 107166 (2021).

27. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis Mach. Intell.* **35**, 1798 – 1828 (2013).

28. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436–444 (2015).

29. Ganaie, M. A., Hu, M., Malik, A., Tanveer, M. & Suganthan, P. N. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **115**, 105151 (2022).

30. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

31. Matloob, F. *et al.* Software defect prediction using ensemble learning: A systematic literature review. *IEEE Access* **9**, 98754–98771 (2021).

32. Wolpert, D. H. The lack of a priori distinctions between learning algorithms. *Neural computation* **8**, 1341–1390 (1996).

33. Xie, J., Xu, B. & Chuang, Z. Horizontal and vertical ensemble with deep representation for classification. *arXiv preprint arXiv:1306.2759* (2013).

34. Laine, S. & Aila, T. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242* (2016).

35. Ciregan, D., Meier, U. & Schmidhuber, J. Multi-column deep neural networks for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3642–3649 (IEEE, 2012).

36. Grassmann, F. *et al.* A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology* **125**, 1410–1420 (2018).

37. Kilimci, Z. H. & Akyokus, S. Deep learning-and word embedding-based heterogeneous classifier ensembles for text classification. *Complexity* **2018** (2018).

38. Tabik, S., Alvear-Sandoval, R.F., Ruiz, M.M., Sancho-Gómez, J., Figueiras-Vidal, A.R. & Herrera, F. MNIST-NET10: a heterogeneous deep networks fusion based on the degree of certainty to reach 0.1% error rate. ensembles overview and proposal. *Inf. Fusion* (2020).

39. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., Mele, B. & Altarelli, G. Dropout: a simple way to prevent neural networks from overfitting. *The J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

40. Wan, L., Zeiler, M., Zhang, S., Cun, Y. L. & Fergus, R. Regularization of Neural Networks using DropConnect. In Dasgupta, S. & McAllester, D. (eds.) *Proceedings of the 30th International Conference on Machine Learning*, vol. 28-3 of *Proceedings of Machine Learning Research*, 1058–1066 (PMLR, Atlanta, Georgia, USA, 2013).

41. Huang, G., Sun, Y., Liu, Z., Sedra, D. & Weinberger, K. Q. Deep networks with stochastic depth. In *European Conference on Computer Vision*, 646–661 (Springer, 2016).

42. Singh, S., Hoiem, D. & Forsyth, D. Swapout: Learning an ensemble of deep architectures. Tech. Rep. (2016).

43. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

44. Smith, L. N., Hand, E. M. & Doster, T. Gradual dropin of layers to train very deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4763–4771 (2016).

45. Wolpert, D. H. Stacked generalization. *Neural Networks* **5**, 241–259 (1992).

46. Ebrahimi, A., Luo, S., Chiong, R., Initiative, A. D. N. *et al.* Deep sequence modelling for alzheimer's disease detection using mri. *Comput. Biol. Medicine* **134**, 104537 (2021).

47. Tanveer, M, Rashid, A.H., Ganaie, MA, Reza, M, Razzak, Imran & Hua, Kai−Lung. Classification of Alzheimer's disease using ensemble of deep neural networks trained through transfer learning. *IEEE J. Biomed. Heal. Informatics* (2021).

48. Maji, K., Sharma, R., Verma, S. & Goel, T. Rvfl classifier based ensemble deep learning for early diagnosis of Alzheimer's Disease. In *Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part III*, 616–626 (Springer, 2023).

49. Razzak, I. *et al.* Mutliresolutional ensemble partialnet for Alzheimer detection using magnetic resonance imaging data. *Int. J. Intell. Syst.* (2022).

50. Wang, R., Li, H., Lan, R., Luo, S. & Luo, X. Hierarchical ensemble learning for Alzheimer's disease classification. In *2018 7th International Conference on Digital Home (ICDH)*, 224–229 (IEEE, 2018).

51. Zheng, C., Xia, Y., Chen, Y., Yin, X. & Zhang, Y. Early diagnosis of Alzheimer's disease by ensemble deep learning using FDG-PET. In *International Conference on Intelligent Science and Big Data Engineering*, 614–622 (Springer, 2018).

52. Zeng, A. *et al.* Discovery of genetic biomarkers for Alzheimer's disease using adaptive convolutional neural networks ensemble and genome-wide association studies. *Interdiscip. Sci. Comput. Life Sci.* **13**, 787–800 (2021).

53. Sharma, R. *et al.* Conv-ervfl: Convolutional neural network based ensemble RVFL classifier for Alzheimer's disease diagnosis. *IEEE J. Biomed. Heal. Informatics* (2022).

54. Zhang, T. & Shi, M. Multi-modal neuroimaging feature fusion for diagnosis of Alzheimer's disease. *J. Neurosci. Methods* **341**, 108795 (2020).

55. Tang, C. *et al.* Csagp: Detecting Alzheimer's disease from multimodal images via dual-transformer with cross-attention and graph pooling. *J. King Saud Univ. Inf. Sci.* 101618 (2023).

56. Haque, M. N., Noman, N., Berretta, R. & Moscato, P. Heterogeneous ensemble combination search using genetic algorithm for class imbalanced data classification. *PloS one* **11**, e0146116 (2016).

57. Choi, J. Y. & Lee, B. Combining of multiple deep networks via ensemble generalization loss, based on MRI images, for Alzheimer's disease classification. *IEEE Signal Process. Lett.* **27**, 206–210 (2020).

58. Ji, H., Liu, Z., Yan, W. Q. & Klette, R. Early diagnosis of Alzheimer's disease using deep learning. In *Proceedings of the 2nd International Conference on Control and Computer Vision*, 87–91 (2019).

59. Giovannetti, A. *et al.* Deep-MEG: spatiotemporal CNN features and multiband ensemble classification for predicting the early signs of Alzheimer's disease with magnetoencephalography. *Neural Comput. Appl.* 1–17 (2021).

60. Sadat, S. U. *et al.* Alzheimer's disease detection and classification using transfer learning technique and ensemble on convolutional neural networks. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 1478–1481 (2021).

61. Islam, J. & Zhang, Y. An ensemble of deep convolutional neural networks for Alzheimer's disease detection and classification. *arXiv preprint arXiv:1712.01675* (2017).

62. Jabason, E., Ahmad, M. O. & Swamy, M. Classification of Alzheimer's disease from MRI data using an ensemble of hybrid deep convolutional neural networks. In *2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS)*, 481–484 (IEEE, 2019).

63. Rajesh Khanna, M. Multi-level classification of Alzheimer disease using dcnn and ensemble deep learning techniques. *Signal, Image Video Process.* 1–9 (2023).

64. Yang, Y., Li, X., Wang, P., Xia, Y. & Ye, Q. Multi-source transfer learning via ensemble approach for initial diagnosis of Alzheimer's disease. *IEEE J. Transl. Eng. Heal. Medicine* **8**, 1–10 (2020).

65. Sethuraman, S. K. *et al.* Predicting Alzheimer's disease using deep neuro-functional networks with resting-state fMRI. *Electronics* **12**, 1031 (2023).

66. Ying, Q. *et al.* Multi-modal data analysis for Alzheimer's disease diagnosis: An ensemble model using imagery and genetic features. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 3586–3591 (IEEE, 2021).

67. Ismail, W. N., PP, F. R. & Ali, M. A. A meta-heuristic multi-objective optimization method for Alzheimer's disease detection based on multi-modal data. *Mathematics* **11**, 957 (2023).

68. Yang, L. *et al.* Deep learning based multimodal progression modeling for Alzheimer's disease. *Stat. Biopharm. Res.* 1–7 (2021).

69. Fang, X., Liu, Z. & Xu, M. Ensemble of deep convolutional neural networks based multi-modality images for Alzheimer's disease diagnosis. *IET Image Process.* **14**, 318–326 (2020).

70. Ambastha, A. K., Leong, T.-Y. & Alzheimer's Disease Neuroimaging Initiative. A deep learning approach to neuroanatomical characterisation of Alzheimer's disease. In *MEDINFO 2017: Precision Healthcare through Informatics*, 1249–1249 (IOS Press, 2017).

71. Suk, H.-I., Lee, S.-W., Shen, D. & Alzheimer's Disease Neuroimaging Initiative. Deep ensemble learning of sparse regression models for brain disease diagnosis. *Med. image analysis* **37**, 101–113 (2017).

72. Wang, H., Shen, Y., Wang, S., Xiao, T., Deng, L., Wang, X. & Zhao, X. Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease. *Neurocomputing* **333**, 145–156 (2019).

73. Ruiz, J., Mahmud, M., Modasshir, M., Kaiser, M. S. & Alzheimer's Disease Neuroimaging Initiative. 3D DenseNet ensemble in 4-way classification of Alzheimer's disease. In *International Conference on Brain Informatics*, 85–96 (Springer, 2020).

74. Chen, Y. & Xia, Y. Iterative sparse and deep learning for accurate diagnosis of Alzheimer's disease. *Pattern Recognit.* **116**, 107944 (2021).

75. Malik, A. & Tanveer, M. Graph embedded ensemble deep randomized network for diagnosis of Alzheimer's disease. *IEEE/ACM Transactions on Comput. Biol. Bioinforma.* 1–13 (2022).

76. Ganaie, M. A. & Tanveer, M. Ensemble deep random vector functional link network using privileged information for Alzheimer's disease diagnosis. *IEEE/ACM Transactions on Comput. Biol. Bioinforma.* 1–1 (2022).

77. Colbaugh, R., Glass, K. & Gallegos, G. Ensemble transfer learning for Alzheimer's disease diagnosis. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3102–3105 (IEEE, 2017).

78. Lu, D., Popuri, K., Ding, G. W., Balachandar, R. & Beg, M. F. Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. *Sci. Reports* **8**, 1–13 (2018).

79. Zhang, J. *et al.* Multi-modal cross-attention network for Alzheimer's disease diagnosis with multi-modality data. *Comput. Biol. Medicine* 107050 (2023).

80. El-Sappagh, S., Abuhmed, T., Islam, S. R. & Kwak, K. S. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. *Neurocomputing* **412**, 197–215 (2020).

81. An, N., Ding, H., Yang, J., Au, R. & Ang, T. F. Deep ensemble learning for Alzheimer's disease classification. *J. biomedical informatics* **105**, 103411 (2020).

82. Ortiz, A., Munilla, J., Gorriz, J. M. & Ramirez, J. Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease. *Int. J. Neural Syst.* **26**, 1650025 (2016).

83. Razzak, I., Naz, S., Alinejad-Rokny, H., Nguyen, T. N. & Khalifa, F. A cascaded mutliresolution ensemble deep learning framework for large scale Alzheimer's disease detection using brain MRIs. *IEEE/ACM Transactions on Comput. Biol. Bioinforma.* 1–9 (2022).

84. Kang, W. *et al.* Multi-model and multi-slice ensemble learning architecture based on 2d convolutional neural networks for Alzheimer's disease diagnosis. *Comput. Biol. Medicine* **136**, 104678 (2021).

85. Donini, M. *et al.* Combining heterogeneous data sources for neuroimaging based diagnosis: re-weighting and selecting what is important. *Neuroimage* **195**, 215–231 (2019).

86. Khanna, S. *et al.* Using multi-scale genetic, neuroimaging and clinical data for predicting alzheimer's disease and reconstruction of relevant biological mechanisms. *Sci. reports* **8**, 11173 (2018).

87. Qiu, S. *et al.* Multimodal deep learning for alzheimer's disease dementia assessment. *Nat. communications* **13**, 3404 (2022).

88. Venugopalan, J., Tong, L., Hassanzadeh, H. R. & Wang, M. D. Multimodal deep learning models for early detection of alzheimer's disease stage. *Sci. reports* **11**, 3254 (2021).

89. Buciluǎ, C., Caruana, R. & Niculescu-Mizil, A. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 535–541 (2006).

90. Cheng, W. X., Suganthan, P. N. & Katuwal, R. Time series classification using diversified ensemble deep random vector functional link and resnet features. *Appl. Soft Comput.* **112**, 107826 (2021).

91. Bengio, Y., Lecun, Y. & Hinton, G. Deep learning for ai. *Commun. ACM* **64**, 58–65 (2021).

92. Li, X. *et al.* Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowl. Inf. Syst.* **64**, 3197–3234 (2022).

93. Ali, S. *et al.* Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion* **99**, 101805 (2023).

94. Díaz-Rodríguez, N. *et al.* Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. *Inf. Fusion* 101896 (2023).

95. Kadmon, J. & Sompolinsky, H. Optimal architectures in a solvable model of deep networks. *Adv. neural information processing systems* **29** (2016).

96. Liu, W. *et al.* A survey of deep neural network architectures and their applications. *Neurocomputing* **234**, 11–26 (2017).

97. Rieke, N. *et al.* The future of digital health with federated learning. *NPJ digital medicine* **3**, 119 (2020).

98. Zhou, Z.-H. & Feng, J. Deep forest. *arXiv preprint arXiv:1702.08835* (2017).

99. Dadar, M. *et al.* Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging. *NeuroImage* **157**, 233–249 (2017).

Competing interests

Authors declare no competing interest.