Artificial Intelligence In Medicine 144 (2023) 102654

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Real-world prediction of preclinical Alzheimer's disease with a deep generative model

Uiwon Hwang [a,1], Sung-Woo Kim [b,1], Dahuin Jung [c], SeungWook Kim [b], Hyejoo Lee [d,e], Sang Won Seo [d,e], Joon-Kyung Seong [f,g,h,*], Sungroh Yoon [c,i,*], for the Alzheimer's Disease Neuroimaging Initiative[2]

[a] Division of Digital Healthcare, Yonsei University, Wonju, 26493, Republic of Korea
[b] Department of Bio-convergence Engineering, Korea University, Seoul, 02841, Republic of Korea
[c] Department of Electrical and Computer Engineering, Seoul National University, Seoul, 08826, Republic of Korea
[d] Department of Neurology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, 06351, Republic of Korea
[e] Neuroscience Center, Samsung Medical Center, Seoul, 06351, Republic of Korea
[f] Department of Artificial Intelligence, Korea University, Seoul, 02841, Republic of Korea
[g] School of Biomedical Engineering, Korea University, Seoul, 02841, Republic of Korea
[h] Interdisciplinary Program in Precision Public Health, College of Health Science, Korea University, Seoul, 02841, Republic of Korea
[i] Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, 08826, Republic of Korea

ARTICLE INFO

Keywords:
Deep generative models
Preclinical Alzheimer's disease
Real-world classification
Explainable AI

ABSTRACT

Amyloid positivity is an early indicator of Alzheimer's disease and is necessary to determine the disease. In this study, a deep generative model is utilized to predict the amyloid positivity of cognitively normal individuals using proxy measures, such as structural MRI scans, demographic variables, and cognitive scores, instead of invasive direct measurements. Through its remarkable efficacy in handling imperfect datasets caused by missing data or labels, and imbalanced classes, the model outperforms previous studies and widely used machine learning approaches with an AUROC of 0.8609. Furthermore, this study illuminates the model's adaptability to diverse clinical scenarios, even when feature sets or diagnostic criteria differ from the training data. We identify the brain regions and variables that contribute most to classification, including the lateral occipital lobes, posterior temporal lobe, and APOE ϵ4 allele. Taking advantage of deep generative models, our approach can not only provide inexpensive, non-invasive, and accurate diagnostics for preclinical Alzheimer's disease, but also meet real-world requirements for clinical translation of a deep learning model, including transferability and interpretability.

1. Introduction

The possibility of preventing the development of Alzheimer's disease (AD) is receiving increasing attention, particularly due to the limited availability of effective disease-modifying treatments (DMTs) [1,2]. Amyloid beta (Aβ) deposition is used as a biomarker in prevention since it reflects the pathophysiological changes occurring in the preclinical stage of the disease [3,4]. While direct measurements of Aβ deposition can be obtained through positron emission tomography (PET) or a lumbar puncture, these methods are inappropriate for cognitively normal (CN) individuals because it is costly, time-consuming, and involves exposure to radiation or considerable pain. Proxy measures can be obtained relatively inexpensively and safely using structural magnetic resonance imaging (MRI) [5,6] or tests of cognitive function [7,8]. However, the relationships between these measures and the extent of Aβ deposition are complicated.

Deep learning (DL) is capable of capturing complicated relationships between features and outcomes, enabling the prediction of outcomes based on given features [9,10]. If the outcome is a dichotomous variable, DL constructs a non-linear decision boundary that separates data with one outcome from those with the other. In clinical applications, however, the dataset used for training is often imperfect: some features or outcome values are missing, or the number of examples with each outcome is very different. It can be caused by the cost of clinical tests, the cost of physician diagnosis, or the prevalence of diseases/the rarity

* Corresponding author.
E-mail addresses: jkseong@korea.ac.kr (J.-K. Seong), sryoon@snu.ac.kr (S. Yoon).
[1] These authors contributed equally to this work.
[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. (adni.loni.usc.edu).

https://doi.org/10.1016/j.artmed.2023.102654
Received 22 January 2023; Received in revised form 29 August 2023; Accepted 29 August 2023
Available online 4 September 2023

of hospital visits by healthy individuals. These problems can hinder the generalization of the model by reducing the available training data or introducing biases towards specific classes. In addition, different clinicians may diagnose diseases based on different sets of features or labeling criteria, which restricts the applicability of observations to new datasets and diminishes the efficiency and utility of the trained model. Furthermore, the non-linear nature of deep neural networks makes it difficult for clinicians to interpret the model's predictions, hindering their ability to explain these predictions to patients. Addressing these issues is critical for improving the effectiveness and generalizability of the model.

In this study, we used a deep neural network to predict whether CN individuals are in the preclinical stage (or amyloid positive CN individuals) using data from structural MRI scans, demographic information, and clinical scores. Our model demonstrated accurate predictions, and its development also embodies three significant steps towards real-world clinical application. Firstly, we trained our model on a dataset with missing features and labels, and with imbalanced classes: some cognitive scores have missing values, amyloid positivities are missing for some participants, and the size of the Aβ+ group is different from that of the Aβ− group. To address these problems, we adopted the HexaGAN framework. Originally designed as a deep generative model (DGM) for synthesizing real-world data through adversarial learning [11,12], HexaGAN showed promising prediction performance when training datasets are imperfect [12]. Additionally, we implemented enhancements to HexaGAN, enabling it to proficiently handle high-dimensional multimodal data. Secondly, we dealt with the scenario of deploying a trained model on data collected from different hospitals, where differences in feature sets and diagnostic criteria may exist. We developed an appropriate prediction scheme to ensure the adaptability of HexaGAN in such situations. Our findings demonstrate that HexaGAN can predict amyloid positivity even when the testing features do not perfectly align with those used in model training. Thirdly, using explainable artificial intelligence (XAI) techniques, we identified discriminative regions and variables that represent the features that are important in the prediction of early amyloid pathology. These are determined at the population level and thus provide physicians and patients with the reliability of the model's predictive ability. Therefore, our considerations for clinical implementation, ranging from model architecture to application, showed that our model can be successfully used in real-world situations. The contribution of this work can be summarized as follows:

- We develop a deep generative model that predicts the preclinical stage of AD in cognitively normal individuals using structural MRI scans, demographic information, and clinical scores.
- The proposed model effectively addresses the imperfect dataset problem, including missing features or labels, and imbalanced classes, by modifying the HexaGAN framework.
- Leveraging the generation capability of DGMs, the learned model demonstrates robust generalization to data collected from different hospitals, which have different feature sets and diagnostic criteria.
- XAI techniques are used to identify discriminative regions and variables that significantly contribute to the prediction of early amyloid pathology. This provides physicians and patients with reliable insights into the model's predictive capability.

The rest of this paper is organized as follows. Section 2 provides materials and the proposed methodology, including a description of the datasets, preprocessing techniques, our deep generative model, and the XAI method. Section 3 presents the experimental results obtained from our model and comparisons with existing approaches. Section 4 provides in-depth discussions on the results, highlighting the strengths and limitations of the proposed method.

2. Material and methods

2.1. Participants

The discovery dataset was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.adni.loni.usc.edu). The discovery dataset was used to construct the model to address the imperfect dataset problem and to verify that our model can robustly predict amyloid positivity. It includes 539 CN participants enrolled in the ADNI-1, ADNI-GO, ADNI-2, and ADNI-3 cohorts, who had T1-weighted magnetic resonance imaging (MRI) scans from a screening visit. The details of the ADNI design, participant recruitment, and diagnostic criteria are published on the ADNI website (http://www.adni.loni.usc.edu/methods/).

The practice dataset was obtained from the Samsung Medical Center (SMC). The practice dataset consists of features and diagnostic criteria different from the discovery dataset and was used to verify that our model trained with data from a different cohort (in our study, the discovery dataset) can make good predictions on data from one hospital. It includes 343 CN participants who had T1-weighted MRI scans. All participants underwent a detailed clinical interview and a neurologic examination [13]. The study protocol was approved by the Institutional Review Board of the SMC.

2.2. MRI acquisition

To construct the discovery dataset, we downloaded T1-weighted MRI scans for 539 participants from ADNI's database. Specifically, we used 1.5 T non-accelerated magnetization-prepared rapid acquisition gradient echo (MP-RAGE) scans for the ADNI-1 cohort, 3 T non-accelerated MP-RAGE or inversion recovery spoiled gradient echo (IR-SPGR) scans for the ADNI-2 and ADNI-GO cohorts, and 3 T non-accelerated MP-RAGE scans for the ADNI-3 cohort. Details of the MRI protocols employed are available on the ADNI website.

For T1-weighted MRI scans of the practice dataset, we used 3 T turbo field echo images acquired at SMC. Detailed information about acquisition protocols is described in the previous study [13].

2.3. Image preprocessing

T1-weighted MRI scans acquired from 539 participants were preprocessed according to the standard procedures of image preprocessing with FreeSurfer v.6.0.0 (http://surfer.nmr.mgh.harvard.edu). These include non-linear registration to Talairach space, intensity correction, skull stripping, and anatomical segmentation of entire brain regions (including the cerebral gray matter, white matter, and cerebellum) according to the Hammersmith (HM) atlas [14]. The intracranial volume (ICV) of each subject was computed for further analysis.

Originally, all the MRI scans that were non-linearly transformed to Talairach space had a resolution of $256 \times 256 \times 256$. Due to the limitation of our graphical processing unit (GPU) memory, however, we performed the two following steps. First, we reduced the field-of-view (FOV) of our MRI scan volumes to $168 \times 190 \times 210$, excluding the non-brain background. We then extracted every 10th slice, from the 20th to the 190th slice. Thus, the intensities of MRI scans had a resolution of $168 \times 190 \times 18$. MRI scans were normalized to the range of 0–1.

After preprocessing, one participant was excluded from the ADNI dataset due to an image preprocessing failure. Therefore, the final number of participants used in this study is 538.

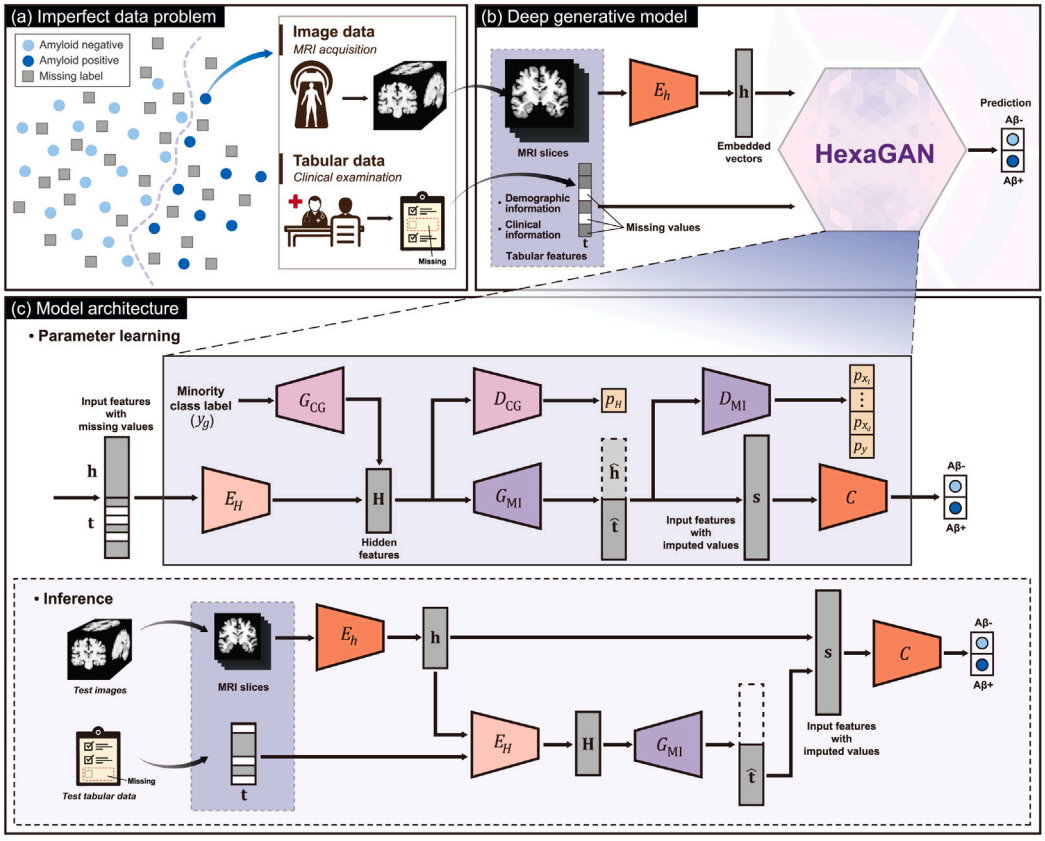

Fig. 1. Methodological overview. (a) We aim to predict amyloid positivity from structural magnetic resonance imaging (MRI) scans, demographic information, and cognitive scores from cognitively normal (CN) individuals. In its raw state, this dataset is imperfect to be fed directly into a classifier. (b) We address these real-world problems simultaneously by enhancing the HexaGAN framework. (c) The generators (G_{CG}, G_{MI}), discriminators (D_{CG}, D_{MI}), encoders (E_h, E_H), and classifier (C) are designed to learn the model from dataset having real-world problems.

2.4. Amyloid positivity determination

For the discovery dataset, global amyloid positivity was determined from PET scans, which consisted of four 5 min frames, acquired 50–70 min after an injection of 370 ±37 MBq of [^{18}F]-AV45. Amyloid positivity is defined as a cortical AV45 standardized uptake value ratio (SUVR) > 1.11 [15]. AV45 SUVRs were average values of frontal, anterior cingulate, precuneus, and parietal cortex relative to the cerebellum, which were extracted from the ADNIMERGE file. Further details of PET imaging protocols are published on the ADNI website and elsewhere [15].

For the practice dataset, global amyloid positivity was determined by the scoring system as described in the previous study [16], where three medical experts visually assessed [^{18}F]-Florbetaben PET scans [17] acquired using PET/CT scanner at SMC.

2.5. Tabular data

The tabular data consists of a total of 27 demographic, genetic, and clinical variables relating to individuals, which are available in the ADNI database. The demographic variables are age, sex, years of education, intracranial volume, and handness. The genetic variables are the number of APOE $\varepsilon 4$ alleles. The 21 clinical variables measure cognitive function and instrumental activities of daily living (IADL).

Three of the clinical variables measure AD-related global cognitive impairment: a mini-mental state examination (MMSE) score; a clinical dementia rating sum of boxes (CDR-SB); and a score for the Alzheimer's disease assessment scale, 13-item version (ADAS13). In addition, 18 variables were extracted from the ADNI neuropsychological battery considering redundancy in the cognitive domain [18], which are listed in Table 2. From the results of a Rey auditory verbal learning test (AVLT), we created learning and memory variables by using the 'Immediate recall', 'Delayed recall', and 'Recognition' scores, and by computing 'Learning', 'Intrusion error (IntErr)', 'Proactive interference (ProINTFC)', and 'Retroactive interference (RetroINTFC)' scores [19]. Finally, two variables were created from measurements of instrumental activities of daily living (IADL): scores from a functional activity questionnaire (FAQ) and for everyday cognition assessed by the patient (ECogPt). Where more than one set of scores was available for the period within 90 days of a participant's MRI scan, we chose the set with the lowest proportion of missing entries.

All 27 variables in the tabular data were normalized to the range 0–1.

2.6. Formulation of the imperfect dataset problem

The dataset we used includes observations with missing labels or missing values in clinical information, and amyloid negative (A$\beta-$) individuals outnumber amyloid positive (A$\beta+$) individuals (Fig. 1(a)). Therefore, we define the imperfect dataset problem as sub-problems, including missing data, class imbalance, and missing label problems. To formulate these sub-problems, the dataset D is defined. D consists of MRI scans I, tabular features $\mathbf{t} \in \mathbb{R}^d$ ($d = 27$), and amyloid positivities $y \in \{0, 1\}$. The numbers of individuals in A$\beta+$ ($y = 1$) and A$\beta-$ ($y = 0$) groups are denoted as n_1 and n_0 respectively. $N_l = n_1 + n_0$ is the number of labeled participants, and N_u is the number of unlabeled participants. We now introduce two Boolean objects $\mathbf{m} \in \{0, 1\}^d$ and $o \in \{0, 1\}$ that represent the missingness of tabular features and labels. We also introduce y_g to represent the minority class, which is the A$\beta+$ group in this study. We can now divide D into labeled data (denoted

as $D_l = \{(\mathbf{t}^n, \mathbf{m}^n), I^n, (y^n, o^n = 1)\}_{n=1}^{N_l}$), and unlabeled data (denoted as $D_u = \{(\mathbf{t}^n, \mathbf{m}^n), I^n, o^n = 0\}_{n=1}^{N_u}$).

Variables created from labeled data are marked with the subscript l, unlabeled data with the subscript u, and synthesized data with the subscript g. Two or more subscripts are combined to simplify the notation (e.g., $\mathbf{t}_l$ and $\mathbf{t}_u$ can be combined into $\mathbf{t}_{lu}$).

2.7. The deep generative model

We modified HexaGAN [12] to predict early amyloid pathology from the imperfect dataset described above. Our model consists of seven components (Fig. 1(b) and (c)). The auxiliary encoder E_h is a new addition to HexaGAN, which maps high-dimensional image data to a low-dimensional embedded vector $\mathbf{h}$. The auxiliary encoder and the remaining six components, employed in the HexaGAN originally, play different roles in classifying Aβ+ and Aβ− groups, as explained below:

- E_h is an encoder that receives an image and synthesizes an embedded vector $\mathbf{h}$.
- E_H is an encoder that receives table data and $\mathbf{h}$ to synthesize hidden vector $\mathbf{H}$.
- G_{MI} is a generator that receives $\mathbf{H}$, fills missing data, and synthesizes $\hat{\mathbf{h}}$ (for conditional generation).
- D_{MI} is a discriminator that distinguishes between real (non-missing) and fake (missing) elements of imputed data and embedding vector elements.
- G_{CG} is a generator that receives (minority) class labels and generates a hidden vector $\mathbf{H}_g$.
- D_{CG} is a discriminator that receives class labels and hidden vectors $\mathbf{H}$ to distinguish between real and fake hidden vectors.
- C is a classifier that predicts amyloid positivity by receiving imputed data $\hat{\mathbf{t}}$ and an embedded vector $\mathbf{h}$.

Our model receives 18 neuropsychological variables and 18 slices from the total 210 coronal slices via decimation as input. Among them, MRI slices I_l and I_u go through E_h to create embedded vectors $\mathbf{h}_l$ and $\mathbf{h}_u$, respectively (Fig. 1(b)). This allows us to deal with high-dimensional MRI scans because E_h extracts vectors of reduced dimensionality ($\mathbf{h}$). In the case of tabular data $\mathbf{t}_l$ and $\mathbf{t}_u$, we replace missing elements with noise using a Hadamard product as follows:

$$\mathbf{h}_{lu} = E_h(I_{lu}) \tag{1}$$

$$\tilde{\mathbf{t}}_{lu} = \mathbf{m}_{lu} \circ \mathbf{t}_{lu} + (1 - \mathbf{m}_{lu}) \circ \mathbf{z}_{lu}, \tag{2}$$

where the operator $\circ$ indicates the Hadamard product of two vectors, and $\mathbf{z}_l$ and $\mathbf{z}_u \in [0,1]^d$ are random noise sampled from $U(0,1)$.

2.7.1. Missing data imputation

The components used for missing data imputation are E_h, E_H, G_{MI}, and D_{MI} (MI stands for 'missing imputation'). In this section, the subscript of all variables without a subscript is lu. To impute missing elements, E_H receives $\tilde{\mathbf{t}}$, $\mathbf{h}$, and $\mathbf{m}$ and creates the hidden vector $\mathbf{H} \in [-1,1]^{d_H}$ (Fig. 1(c)). Then G_{MI} receives $\mathbf{H}$ and creates $\bar{\mathbf{t}}$ and $\hat{\mathbf{h}} \in [0,1]^{d_h}$. For tabular data, only the missing elements of $\mathbf{t}$ are replaced with $\bar{\mathbf{t}}$, and imputed tabular data is called $\hat{\mathbf{t}}$:

$$\mathbf{H} = E_H(\tilde{\mathbf{t}}, \mathbf{h}, \mathbf{m}) \tag{3}$$

$$(\bar{\mathbf{t}}, \hat{\mathbf{h}}) = G_{\text{MI}}(\mathbf{H}) \tag{4}$$

$$\hat{\mathbf{t}} = \mathbf{m} \circ \mathbf{t} + (\mathbf{1} - \mathbf{m}) \circ \bar{\mathbf{t}}, \tag{5}$$

$\mathbf{h}$ and the synthesized $\hat{\mathbf{t}}$ are fed into D_{MI} with the class label. For labeled data, y is used as the class label, and for unlabeled data, the prediction from the classifier C (e.g., for binary class, $\lfloor C(\hat{\mathbf{t}}_u, \mathbf{h}_u) + 0.5 \rfloor$) is used and called y_u. Then, D_{MI} computes $(d + d_h + 1)$ outputs, each of which is a value that measures whether elements of $\hat{\mathbf{t}}$, elements of $\mathbf{h}$, and labels are missing (fake) or non-missing (real). For missing data imputation, $(d + d_h)$ elements of the output of D_{MI} (except the last element for a label) are used for loss functions. E_H and G_{MI} are learned via element-wise WGAN hinge loss as follows:

$$\mathcal{L}^{lu}_{G_{\text{MI}}} = -\mathbb{E}_{\hat{\mathbf{t}}|\mathbf{h},y,\mathbf{m}} \left[\frac{1}{\sum_{i=1}^{d+d_h} 1 - m_i^{th}} \sum_{i=1}^{d+d_h} (1 - m_i^{th}) \cdot D_{\text{MI}}(\hat{\mathbf{t}}, \mathbf{h}, y)_i \right] \tag{6}$$

$$= -\mathbb{E}_{\hat{\mathbf{t}}|\mathbf{h},y,\mathbf{m}} \left[\frac{1}{\sum_{i=1}^{d} 1 - m_i} \sum_{i=1}^{d} (1 - m_i) \cdot D_{\text{MI}}(\hat{\mathbf{t}}, \mathbf{h}, y)_i \right], \tag{7}$$

where $D_{\text{MI}}(\cdot)_i$ is the ith element of the output of D_{MI} and $\mathbf{m}^{th}$ denotes the missingness labels for the elements of $\hat{\mathbf{t}}$ and $\mathbf{h}$. If the element is missing, the label is marked as 0, and if it is non-missing, the label is 1. For labeled and unlabeled data, the missingness label for $\hat{\mathbf{t}}$ ($\mathbf{m}^t$) is $\mathbf{m}$, and the missingness label for $\mathbf{h}$ ($\mathbf{m}^h$) is $\mathbf{1}$ and $(\mathbf{1} - \mathbf{m}^h) = \mathbf{0}$, because there is no missing element in the image data. Therefore, $\mathcal{L}^{lu}_{G_{\text{MI}}}$ uses only the first d elements for loss. D_{MI}, which plays a critical role in missing data imputation, uses the following element-wise adversarial loss:

$$\mathcal{L}^{lu}_{D_{\text{MI}}} = -\mathbb{E}_{\hat{\mathbf{t}}|\mathbf{h},y,\mathbf{m}} \left[\frac{1}{\sum_{i=1}^{d+d_h} m_i^{th}} \sum_{i=1}^{d+d_h} m_i^{th} \cdot \min(0, -1 + D_{\text{MI}}(\hat{\mathbf{t}}, \mathbf{h}, y)_i) \right] \tag{8}$$

$$- \mathbb{E}_{\hat{\mathbf{t}}|\mathbf{h},y,\mathbf{m}} \left[\frac{1}{\sum_{i=1}^{d+d_h} 1 - m_i^{th}} \sum_{i=1}^{d+d_h} (1 - m_i^{th}) \cdot \min(0, -1 - D_{\text{MI}}(\hat{\mathbf{t}}, \mathbf{h}, y)_i) \right]. \tag{9}$$

In terms of a distance metric, Wasserstein distance is weaker than the Jensen–Shannon divergence used in a vanilla GAN [11], so it converges better with complex data distributions [20], and facilitates the imputation of missing values in the data. We expanded a projection layer [21] in an element-wise manner to inject the conditional information of y into D_{MI}.

We also used a reconstruction loss that improves the imputation process by learning from non-missing data as follows:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{\hat{\mathbf{t}},\hat{\mathbf{h}}|\mathbf{m},\mathbf{t},\mathbf{h}} \left[\frac{1}{d} \sum_{i=1}^{d} m_i \cdot (t_i - \hat{t}_i)^2 + \frac{1}{d_h} \sum_{i=1}^{d_h} m_i^h \cdot (h_i - \hat{h}_i)^2 \right] \tag{10}$$

$$= \mathbb{E}_{\hat{\mathbf{t}},\hat{\mathbf{h}}|\mathbf{m},\mathbf{t},\mathbf{h}} \left[\frac{1}{d} \sum_{i=1}^{d} m_i \cdot (t_i - \hat{t}_i)^2 + \frac{1}{d_h} \sum_{i=1}^{d_h} (h_i - \hat{h}_i)^2 \right], \tag{11}$$

where $\mathbf{m}^h = \mathbf{1}$, so it can be omitted. In Eq. (11), E_h and E_H serve as the encoders of the autoencoder [22], while G_{MI} plays a role as a decoder, learning the distribution of non-missing elements. Therefore, E_H is trained to optimize the following objective:

$$\min_{E_H} \mathcal{L}^{lu}_{G_{\text{MI}}} + \lambda_1 \mathcal{L}_{\text{recon}}, \tag{12}$$

where λ_1 is a hyperparameter that adjusts the ratio between losses: a value of 10 was used for experiments. Loss functions for G_{MI} are also related to class conditional generation; therefore, they will be described later.

2.7.2. Class conditional generation

The class imbalance problem is addressed by G_{CG} and D_{CG}, which synthesize hidden vectors conditioned on the minority class label y_g, and by G_{MI}, which oversamples the minority class by imputing all features. G_{CG} (CG stands for 'conditional generation') receives the noise vector $\mathbf{z}_g \in [0,1]^d$ sampled from $U(0,1)$ and the minority class label y_g, and creates the hidden vector $\mathbf{H}_g$. Then G_{MI} receives $\mathbf{H}_g$ and synthesizes the oversampled data $\hat{\mathbf{t}}_g$ and $\hat{\mathbf{h}}_g$:

$$\mathbf{H}_g = G_{\text{CG}}(\mathbf{z}_g, y_g) \tag{13}$$

$$(\hat{\mathbf{t}}_g, \hat{\mathbf{h}}_g) = G_{\text{MI}}(\mathbf{H}_g). \tag{14}$$

Here, $\hat{\mathbf{t}}_g$ is computed on the basis that all elements are missing, and $\hat{\mathbf{h}}_g$ is not obtained from a real image. Therefore, the missingness labels $\mathbf{m}^t_g$ and $\mathbf{m}^h_g$ are zero vectors $\mathbf{0}$ in the d and d_h dimensions, respectively.

D_{CG} is trained to minimize WGAN hinge loss, considering $\mathbf{H}_l$ as real data and $\mathbf{H}_g$ as fake data, as follows:

$$\mathcal{L}_{D_{\mathrm{CG}}} = \mathbb{E}_{\mathbf{H}_l|y}\left[\max(0, 1 - D_{\mathrm{CG}}(\mathbf{H}_l, y))\right] + \mathbb{E}_{\mathbf{H}_g|y_g}\left[\max(0, 1 + D_{\mathrm{CG}}(\mathbf{H}_g, y_g))\right] \tag{15}$$

$$\min_{D_{\mathrm{CG}}} \mathcal{L}_{D_{\mathrm{CG}}}. \tag{16}$$

To inject the conditional information of y into D_{CG}, we used a projection layer [21].

The goal of G_{CG} can be subdivided into three. The first is to synthesize a realistic $\mathbf{H}_g$, which can be learned through an adversarial loss so that D_{CG} can be fooled by G_{CG} (Eq. (17)). The second is to synthesize realistic $\hat{\mathbf{t}}_g$ and $\hat{\mathbf{h}}_g$, which can be learned through element-wise adversarial loss to deceive D_{MI} (Eq. (18)). Third, since oversampled data should be well-conditioned on the minority class, a prediction is obtained by feeding $\hat{\mathbf{t}}_g$ and $\hat{\mathbf{h}}$ to C, and then trying to minimize the cross-entropy between the prediction and y_g (Eq. (19)).

G_{CG} is trained to minimize the following three loss functions:

$$\mathcal{L}_{G_{\mathrm{CG}}} = -\mathbb{E}_{\mathbf{H}_g|y_g}\left[D_{\mathrm{CG}}(\mathbf{H}_g, y_g)\right] \tag{17}$$

$$\mathcal{L}^g_{G_{\mathrm{MI}}} = -\mathbb{E}_{\hat{\mathbf{t}}_g,\hat{\mathbf{h}}_g|y_g}\left[\frac{1}{d+d_h}\sum_{i=1}^{d+d_h} D_{\mathrm{MI}}(\hat{\mathbf{t}}_g, \hat{\mathbf{h}}_g, y_g)_i\right] \tag{18}$$

$$\mathcal{L}^g_C = \mathbb{E}_{\hat{\mathbf{t}}_g,\hat{\mathbf{h}}_g|y_g}\left[-y_g \cdot \log C(\hat{\mathbf{t}}_g, \hat{\mathbf{h}}_g)\right] \tag{19}$$

$$\min_{G_{\mathrm{CG}}} \mathcal{L}_{G_{\mathrm{CG}}} + \lambda_2 \mathcal{L}^g_{G_{\mathrm{MI}}} + \lambda_3 \mathcal{L}^g_C, \tag{20}$$

where λ_2 and λ_3 are hyperparameters used as coefficients for $\mathcal{L}^g_{G_{\mathrm{MI}}}$ and $\mathcal{L}^g_C$ respectively. In experiments, we used $\lambda_2 = 1$ and $\lambda_3 = 0.01$. Since missingness labels for $\hat{\mathbf{t}}_g$ and $\hat{\mathbf{h}}_g$ are 0, $\mathcal{L}^g_{G_{\mathrm{MI}}}$ can be derived. In addition, as a critic for the synthesized data, D_{MI} is also trained through adversarial loss, as follows:

$$\mathcal{L}^g_{D_{\mathrm{MI}}} = -\mathbb{E}_{\hat{\mathbf{t}}_g|\hat{\mathbf{h}}_g,y_g}\left[\frac{1}{d+d_h}\sum_{i=1}^{d+d_h} \min(0, -1 - D_{\mathrm{MI}}(\hat{\mathbf{t}}_g, \hat{\mathbf{h}}_g, y_g)_i)\right]. \tag{21}$$

D_{MI} is trained to accurately predict all input features ($\hat{\mathbf{t}}_g$ and $\hat{\mathbf{h}}_g$) as missing.

We now have a missingness label for all labeled, unlabeled, and synthesized data. The adversarial loss for labeled and unlabeled data is $\mathcal{L}_{G^{lu}_{\mathrm{MI}}}$ and that for synthesized data is $\mathcal{L}^g_{G_{\mathrm{MI}}}$. In addition, we can use a reconstruction loss $\mathcal{L}_{\mathrm{recon}}$ for non-missing data to learn G_{MI} as follows:

$$\min_{G_{\mathrm{MI}}} \mathcal{L}^{lu}_{G_{\mathrm{MI}}} + \mathcal{L}^g_{G_{\mathrm{MI}}} + \lambda_1 \mathcal{L}_{\mathrm{recon}}, \tag{22}$$

where λ_1, the coefficient for the reconstruction loss, was set to 10, which is the same value used to update E.

2.7.3. Classification and semi-supervised learning

The missing label problem is addressed by semi-supervised learning of two components; the classifier C and the last neuron of the output layer of D_{MI}. When a mini-batch of data enters the model, we set the number of y_g as the difference between the amount of majority class data and that of minority class data in the mini-batch. Thus, the class labels of input features $\mathbf{s}_l = (\mathbf{h}_l, \hat{\mathbf{t}}_l)$ and $\mathbf{s}_g = (\hat{\mathbf{h}}_g, \hat{\mathbf{t}}_g)$ together are now balanced and used to train the classifier C. Then, the mini-batch and oversampled data are used for calculating the cross-entropy loss as follows:

$$\mathcal{L}^{lg}_C = \mathbb{E}_{\hat{\mathbf{t}}_l|\mathbf{h}_l,y_l}\left[-y_l \cdot \log C(\mathbf{s}_l)\right] + \mathbb{E}_{\mathbf{s}_g|y_g}\left[-y_g \cdot \log C(\mathbf{s}_g)\right]. \tag{23}$$

In addition, C can synthesize a pseudo-label y_u for $\hat{\mathbf{t}}_u$ and $\mathbf{h}_u$. The synthesized pseudo-label can be assessed by D_{MI}, as follows:

$$\mathcal{L}^y_C = -\mathbb{E}_{y_u|\hat{\mathbf{t}}_u,\mathbf{h}_u\sim p_C}\left[D_{\mathrm{MI}}(\hat{\mathbf{t}}_u, \mathbf{h}_u, \mathbf{y}_u)_{d+d_h+1}\right] \tag{24}$$

$$\mathcal{L}^y_{D_{\mathrm{MI}}} = \mathbb{E}_{y_u|\hat{\mathbf{t}}_u,\mathbf{h}_u\sim p_C}\left[D_{\mathrm{MI}}(\hat{\mathbf{t}}_u, \mathbf{h}_u, y_u)_{d+d_h+1}\right] - \mathbb{E}_{y_l|\hat{\mathbf{t}}_l,\mathbf{h}_l}\left[D_{\mathrm{MI}}(\hat{\mathbf{t}}_l, \mathbf{h}_l, y)_{d+d_h+1}\right], \tag{25}$$

where p_C is the distribution of the classifier. Eqs. (24) and (25) can be interpreted as representing the Wasserstein adversarial loss between C and $D_{\mathrm{MI}}(\cdot)_{d+d_h+1}$. In this context, C functions as the label generator, while $D_{\mathrm{MI}}(\cdot)_{d+d_h+1}$ acts as the label discriminator. It has been shown that such adversarial learning helps supervised learning by minimizing the ODM cost [12,23]. Finally, C and D_{MI} are trained with the following objectives:

$$\min_C \mathcal{L}^{lg}_C + \lambda_4 \mathcal{L}^y_C \tag{26}$$

$$\min_{D_{\mathrm{MI}}} \mathcal{L}^{lu}_{D_{\mathrm{MI}}} + \mathcal{L}^g_{D_{\mathrm{MI}}} + \lambda_5 \mathcal{L}^y_{D_{\mathrm{MI}}}. \tag{27}$$

We set λ_4 and λ_5 to 0.005 for our experiments.

For inference, our model was designed to respond to scenarios where test features also have missing values or even when all tabular features are missing, using MRI scans alone to predict amyloid positivity (Fig. 1(c)).

The encoder E_h should synthesize the embedded vector $\mathbf{h}$ that serves to make the classifier C achieve a high classification performance. In addition, E_h should also help E_H and G_{MI} synthesize the over-sampled embedded vector $\hat{\mathbf{h}}$ whose distribution resembles that of $\mathbf{h}$. Thus, we trained E_h using the following loss function $\mathcal{L}_{E_h}$ which is a linear combination of a reconstruction loss function $\mathcal{L}^h_{\mathrm{recon}}$ and a cross-entropy loss function $\mathcal{L}^l_C$:

$$\mathcal{L}^h_{\mathrm{recon}} = \|\mathbf{h} - \hat{\mathbf{h}}\|_2^2 \tag{28}$$

$$\mathcal{L}^l_C = \left[-y \cdot \log C(\mathbf{s}_l)\right]. \tag{29}$$

Therefore, E_h is trained to optimize the following objective:

$$\min_{E_h} \mathcal{L}^h_{\mathrm{recon}} + \mathcal{L}^l_C. \tag{30}$$

Our method has additional methodological contributions to HexaGAN. The original HexaGAN uses a zero-centered gradient penalty [24] to stabilize adversarial learning. However, keeping the gradient near zero makes training slow, and computing gradient penalties in an element-wise manner for high-dimensional data requires impractical computation time. Therefore, we applied spectral normalization [25] to the weights of all the layers in our network, which promotes stable convergence much more quickly. To improve the generalization of the encoder E_h, we employed DenseNet pretrained on the ImageNet dataset [26,27]. Representations learned from a large-scale dataset help to improve generalization performance as a rule of thumb even for a specific medical application with data having different input dimensions and the number of channels [28]. The network architecture of each component and training details are described in Table A.4 in Appendix and Appendix A.2.

2.7.4. Model comparison

When training our model with MRI scans alone, E_H and G_{MI} are not required, and G_{CG} synthesizes embedded vectors $\mathbf{h}$ instead of hidden vectors $\mathbf{H}$. When training our model with only tabular data, E_h is not used. For comparison, we used various machine learning methods for imperfect dataset problems. We constructed several benchmark models consisting of cascade combinations of the machine learning methods and the same classifier as C of our model. To solve the imperfect dataset problem for the benchmark models, we used column-wise deletion, mean imputation, k-nearest neighbor (kNN) imputation [29] and multiple imputation by chained equation (MICE) [30] to deal with the missing data problem; kNN and label propagation [31] to deal with the missing label problem; and synthetic minority oversampling technique (SMOTE) [32], adaptive synthetic (ADASYN) method [33], cost sensitive loss (CS) [34] and class rectification loss (CRL) [35] to deal with the class imbalance problem. Since all labeled data have one or more missing values, it is impossible to deal with missing values via a list-wise deletion when both labeled and unlabeled data are used. We computed the area under the receiver operating characteristic

(AUROC) curve estimated by 5-fold cross validation implemented in Scikit-learn [36] to evaluate generalized classification performance. We also performed feature ablation trials on a partition of the feature set consisting of MRI scan and tabular data to check our assumption that combining these types of data improves performance.

We evaluated the effect of imputation on the classification performance of our model as part of our exploratory process of determining where high accuracy came from. We randomly removed between 10% and 90% of observed tabular feature values and imputed these missing values using our model and the aforementioned imputation techniques. This process was repeated 100 times, each starting at a different random state for each missing rate. We then computed the 5-fold average root mean square errors (RMSE) between the imputed and original values.

2.8. Analysis by t-stochastic neighbor embedding

We inspected the changes that occurred to the spaces of the hidden features (**H**) and input features (**s**) as the model was trained. For this analysis, we trained a randomly initialized model with all the data. At each epoch, we visualized the observed hidden features $\mathbf{H}_l$ synthesized by E_h and E_H; and the synthetic hidden features $\mathbf{H}_g$ synthesized by G_{CG} by reducing their dimensionality using t-stochastic neighbor embedding (t-SNE) [37]. We visualized the labeled input features $\mathbf{s}_l = (\hat{\mathbf{t}}_l, \mathbf{h}_l)$, and the synthesized input features $\mathbf{s}_g = (\hat{\mathbf{t}}_g, \hat{\mathbf{h}}_g)$ in a similar way.

2.9. Discriminative regions and variables

Feature attributions allow us to quantify how discriminative regions and variables are. We first chose the model that produced the best predictions out of the five sets of models learned during stratified 5-fold cross validation. We then trained the model from these models using the whole of each dataset until training AUROC reached 1 from the chosen model. We then computed *feature attributions* A^n for nth participant using the integrated gradient (IG) algorithm [38]:

$$A^n = \frac{\mathrm{IG}_f\left(X^n, X_{\min}\right) + \mathrm{IG}_f\left(X^n, X_{\max}\right)}{2}, \tag{31}$$

where IG_f is an integrated gradient functional depending on a deep learning model f, $X^n = (\mathbf{t}^n, I^n)$ is observed features which consist of tabular features and an MRI scan, and $X_{\min}$ and $X_{\max}$ are features that have the same size as X^n and have the values all 0 and 1, respectively. $\mathrm{IG}_f(X^n, X_{\min})$ was calculated by summing the gradients of the model output with respect to a sequence of $k+1$ progressively interpolated features between $X_{\min}$ and X, as follows:

$$\mathrm{IG}_f(X^n, X_{\min}) = (X^n, X_{\min}) \times \int_0^1 \nabla f\left(\alpha \times X^n + (1-\alpha) \times X_{\min}\right) d\alpha, \tag{32}$$

$$\approx (X^n, X_{\min}) \times \sum_{i=0}^{k} \nabla f\left(\frac{i}{k} \times X^n + \frac{k-i}{k} \times X_{\min}\right). \tag{33}$$

$\mathrm{IG}_f(X^n, X_{\max})$ was calculated in similar manner.

For tabular data **t**, we define the *importance* of the ith tabular feature t_i as $\mathcal{I}^n_{t_i} = (A^n_{t_i})^2$, where $A^n_{t_i}$ is the attribution value of tabular feature t_i in the attribution map A^n. The importance value represents the amount of influence for predicting early amyloid pathology regardless of its sign.

For MRI scans I, attributions are computed at the region-of-interest (ROI) level, not at the voxel level, according to XRAI framework [39]. Instead of Felzenszwalb's graph-based method used in the original XRAI paper for determining parcels, we used an atlas-tased segmentation technique with the HM atlas that structurally divides a brain image, including cerebral gray/white matters, ventricles, cerebellum, and brainstem into pre-defined regions. Similar with tabular data, we defined the *importance* of ith brain region r_i, as follows:

$$\mathcal{I}^n_{r_i} = \left(\frac{1}{n_{r_i}} \sum_{v \in r_i} A_{I_v}\right)^2, \tag{34}$$

where A_{I_v} is the attribution value of voxel v in MRI scan I, and n_{r_i} is the number of voxels of a brain region r_i. Since we are interested in brain regions that affect the early amyloid pathology as tabular features, we focused on the amount of influence for each region regardless of its direction that represents whether the region should be bright or dark to be classified into the correct class. We consider population-level attribution as the square root of the average importance value over labeled participants.

Finally, we defined discriminative regions on variables as ROIs or tabular features with statistically significant attributions. Statistical significance is determined based on Bonferroni-corrected 95% confidence intervals for the population-level attribution computed by the studentized bootstrap procedure with 10,000 bootstrap resamples.

3. Results

3.1. Dataset information

Table 1 shows the demographic, genetic, and clinical characteristics of the participants in our discovery and practice datasets. In the discovery datasets, APOE ϵ4, age, and sex differed significantly between the amyloid positive (Aβ+) and negative (Aβ−) groups. In the practice dataset, APOE ϵ4 and age were significantly different between the two groups, as in the discovery dataset.

Three aspects of the imperfect dataset problem are relevant to this study. Firstly, all the records in the discovery dataset lack at least one of the 18 neuropsychological variables, which include measures of cognitive functioning and IADL. These absences are concentrated in records without a diagnosis (Table 2). In the practice dataset, not only does the tabular data have missing values, but no neuropsychological variables were observed. Secondly, 260 records in the discovery dataset and 187 in the practice dataset correspond to participants who did not have amyloid PET scans. Therefore, amyloid positivity is unknown in almost half of the records. These records were used as unlabeled data in the later process. Thirdly, the ratios between the sizes of the Aβ+ and Aβ− groups are about 1:2 and 1:5 in the discovery and practice datasets, respectively.

3.2. Improved classification performance with imperfect datasets

Fig. 2(a), (b), and (c) contains the ROC curves of our method and the top three combinations with the highest AUROC values from Tables B.6, B.7, and B.8 in Appendix, respectively. The combinations are arranged in the order of addressing the missing data, class imbalance, and missing label problems. In cases where a particular method is not employed for a problem, it is omitted from the name of the combination. For example, MLP/CNN + Mean + ADASYN means that the classifier consists of multilayer perceptrons (MLP) and convolutional neural networks (CNN) for both image and tabular data, and data are preprocessed by mean imputation for the missing data problem, ADASYN oversampling for the class imbalance problem, and no method is used for the missing label problem.

Our model achieved an average AUROC for 5-fold cross validation of 0.8609, which was 17.2% higher than the best of the other DL models such as MLP and CNN with combinations of machine learning methods (Fig. 2(a)). At optimal cut-offs determined by the Youden index [40], our model showed an average accuracy of 0.8244, an average sensitivity of 0.8415, and an average specificity of 0.8178. Performance on each fold of our model and comparative results can be found in Tables B.5 and B.6 in Appendix, respectively.

3.3. Synergistic effects of image and tabular data

When trained with just one of these types of data, our model produced more accurate classifications than other DL models (Fig. 2(b) and (c)); and these classifications improved when our model was

Table 1
Demographic and clinical characteristics of participants.

Characteristics	Discovery dataset				Practice dataset			
	Total	Aβ+	Aβ−	p values	Total	Aβ+	Aβ−	p values
N	538	93	186		343	28	128	
Age, years[a]	74.21 (5.84)	74.68 (5.87)	72.32 (5.98)	0.0020*	66.76 (12.04)	71.00 (6.54)	64.77 (15.85)	0.0012*
Sex (female)[b], no.	128	30	98	0.0019*	217	18	78	0.9081
Education, years[a]	16.38 (2.68)	16.15 (2.70)	16.91 (2.43)	0.0178	11.52 (4.69)	11.46 (4.39)	11.43 (4.83)	0.9728
ICV, ℓ[a]	1.50 (0.16)	1.47 (0.17)	1.49 (0.16)	0.6111	1.52 (0.17)	1.58 (0.18)	1.53 (0.16)	0.2201
Handness (LH), no.[b]	28	10	18	0.9438	N/A[d]	N/A	N/A	N/A
APOE ε4 (0/1/2), no.[c]	369/137/12	50/39/4	147/36/3	< 0.001*	253/73/9	8/16/4	103/20/1	< 0.001*
MMSE[a]	29.06 (1.14)	29.03 (0.96)	29.00 (1.32)	0.8168	28.07 (2.04)	27.68 (1.56)	28.22 (1.53)	0.0960
CDR-SB[a]	0.040 (0.139)	0.065 (0.184)	0.040 (0.137)	0.2632	0.677 (0.694)	0.893 (0.600)	0.659 (0.470)	0.0276
ADAS13[a]	9.23 (4.33)	9.65 (4.42)	8.68 (4.37)	0.0836	N/A	N/A	N/A	N/A

N, Sample size; no. Number; ICV, Intracranial volume; MMSE, Mini-mental state examination; CDR-SB, Clinical dementia rating, sum of boxes; ADAS13, Alzheimer's disease assessment scale, 13-item version.

[a] Data are given as the mean and standard deviation; p values were calculated by two-sided Student's or Welch's two-sample t-tests.

[b] For dichotomous data, p values were calculated by Yates-corrected Chi-square tests.

[c] Data are given as frequencies of the numbers of alleles; p value was calculated by a two-sided Fisher's exact test.

[d] Statistics are not available because features are not observed in the dataset.

* $p < 0.05$. Significance adjusted after Bonferroni correction.

Table 2
Neuropsychological variables fed into our deep generative model, with rates of missing data.

Domain	Measure	Missing rates in discovery dataset (%)			
		Aβ+	Aβ−	Label missing	Total
Language	BNTTOTAL	0	0	10.81	5.20
	CATANIMSC	0	0	10.04	4.83
Learning & memory	AVLT-immediate	0	0.54	11.58	5.76
	AVLT-delayed	0	0	10.42	5.02
	AVLT-recog	0	0	10.42	5.02
	Learning	0	0	10.81	5.20
	IntErr	0	0.54	11.58	5.75
	ProINTFC	0	0	10.81	5.20
	RetroINTFC	0	0.54	11.58	5.76
Attention	TRAASCOR	0	0	10.04	4.83
	DIGITSCOR	100	100	14.29	58.74
Executive	TRABSCOR	0	0.54	10.04	5.02
	DSPANFOR	100	100	14.29	58.74
	DSPANBAC	100	100	14.29	58.74
Perceptual-motor	CLOCKSCOR	0	0	10.04	4.83
	COPYSCOR	0	0	10.42	5.02
IADL	FAQ	0	0	10.81	5.20
	ECogPT	1.08	0	95.75	46.28

BNTTOTAL, Boston naming test, total score; CATANIMSC, Category fluency, animals; AVLT, Rey auditory verbal learning test; IntErr, intrusion error; ProINTFC, proactive interference; ReteroINTFC, retero interference [19]; TRAASCOR/TRABSCOR, trail making test, Part A/Part B; DIGITSCOR/DSPANFOR/DSPANBAC, digit span, digit symbol/forward#:total correct/backward#:total correct; CLOCKSCOR, clock drawing, total; COPYSCOR, clock copy, total; FAQ, functional activity questionnaire; ECogPT, everyday cognition, participants.

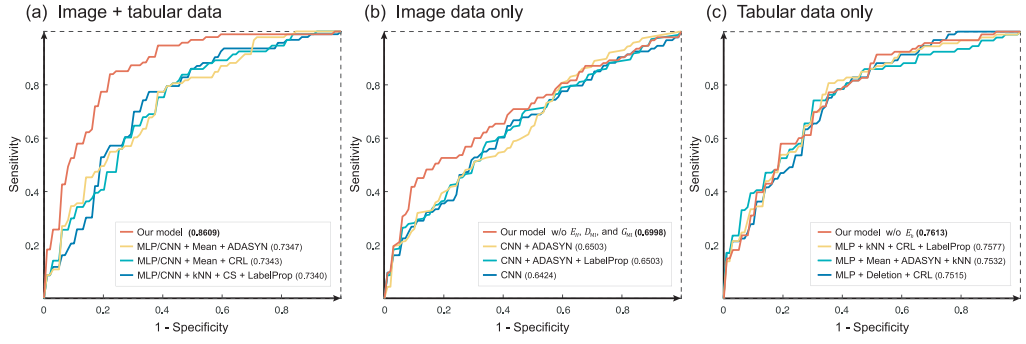

Fig. 2. Classification Performance of our model. ROC curves over 5-folds for our model and comparative classifiers, trained with both MRI scans and tabular data (a), with only MRI scans (b), and with only tabular data (c).

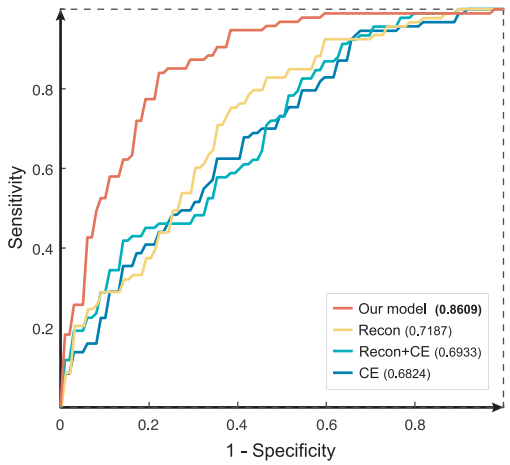

Fig. 3. ROC curves over 5-folds for our method and pretrained encoders, trained with both MRI scans and tabular data.

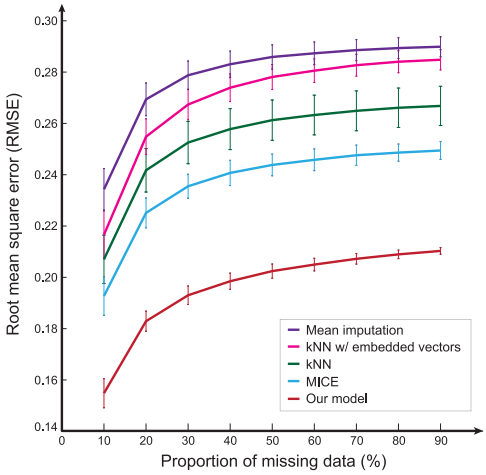

Fig. 4. Performance comparison on the imputation of missing data.

Table 3
Performance comparison on the feature extraction. Recon and CE denote the reconstruction loss and cross-entropy loss for pretraining, respectively.

Method	AUROC	F1-score	PRAUC
Ours	**0.8609**	**0.7596**	**0.6421**
Recon	0.7187	0.6095	0.5869
Recon+CE	0.6933	0.5751	0.5713
CE	0.6824	0.5949	0.4593

trained with both types of data. This contrasts with our observations of discriminative DL models, which do not perform much differently when MRI scans are added to tabular data. Further details of these comparisons are presented in Tables B.6, B.7, and B.8 in Appendix.

3.4. Effectiveness of feature extraction

To evaluate the effectiveness of the encoder's feature extraction, we conducted comparative experiments between pretrained embeddings and our method on the discovery dataset. Specifically, we pretrained networks with the same architecture as our encoders (E_H and E_h) using either the reconstruction loss, the cross-entropy loss, or both. Mean imputation and ADASYN oversampling are applied to mitigate the potential negative impact of the imperfect dataset problem during pretraining, which were identified as the most effective combination in Table B.6 in Appendix. Subsequently, we integrated these pretrained networks, with their weights frozen, in place of the encoders, and trained the remaining components of HexaGAN.

Fig. 3 and Table 3 illustrate the results of the performance comparison. Our method outperformed pretrained embeddings across all performance metrics. We attribute this to the interactions between components of our model during training, particularly through the adversarial loss with the discriminator. Additionally, the use of the cross-entropy loss calculated on better imputed data contributed to the enhanced performance.

3.5. Imputation of missing values

We compared the performances of our model and comparative methods using root mean square error (RMSE) values between the imputed values and the original values computed from 100 sets of test data. We removed 10% to 90% of features at random from test data. The error bar represents the standard deviation of RMSE values.

For our model, the average RMSE value increased from 0.16 to 0.21 as the proportion of missing data increased from 10% to 90%, suggesting our model was better at imputation even at 90% missing. For all proportions of missing data, over 100 trials, our model showed the lowest average RMSE values, followed by MICE, kNN, and mean imputation, in that order (Fig. 4). In addition, our model showed the lowest standard deviation of RMSE values over repeated trials at all missing rates compared to the other models. Taken together, we verify that our model achieved the best classification performance by imputing missing values with more accurate and robust values.

We also found that, with all the methods, the RMSE increases rapidly as the proportion of missing data increases, and then becomes saturated (Fig. 4). This seems to be because the amount of meaningful information converges downward when the missing rate reaches a certain percentage in the data we used. Notwithstanding, our method imputes missing values more robustly than other methods when there is less information available (even at a 90% missing rate), which is consistent with the results in the original HexaGAN paper [12].

3.6. Addressing the class imbalance problem

The class imbalance problem not only leads to poor generalization [41], but is also related to poor accuracy. To address this issue, we employed class-conditional generation to oversample the minority class. This is considered an imputation of all features, conditioned on the specific class. We verified that class conditional generation, adopted by our framework to deal with this issue, was performed successfully. We reduced the dimensionality of the hidden features ($\mathbf{H}$) and the input features ($\mathbf{s} = (\hat{\mathbf{t}}, \mathbf{h})$) of both the observed and synthesized data to two, using the t-stochastic neighbor embedding (t-SNE) algorithm [37], and visualize their distributions. The distributions of the labeled hidden features $\mathbf{H}_l$ and the labeled input features $\mathbf{s}_l$ of observed data were changed as the model was trained because the encoders E_h and E_H and the generator G_{MI} were updated. We observed the distributions of the synthesized hidden features $\mathbf{H}_g$ and the synthesized input features $\mathbf{s}_g$ as the model was trained (Fig. 5, columns 1–3). The distances between clusters of synthesized and observed data seemed to increase during some epochs, but the distributions of $\mathbf{H}_g$ and $\mathbf{s}_g$ eventually chased those of $\mathbf{H}_l$ and $\mathbf{s}_l$ (Fig. 5, columns 4–5). The convergence of the distributions of the synthesized and observed values was seen to be accompanied by improved separation of the Aβ+ and Aβ− groups. This suggests that our model accurately learned the data manifold occupied by the hidden features, and that it synthesized data that is realistic for each group. This solution to the class imbalance problem can be expected to improve the classification accuracy of the model.

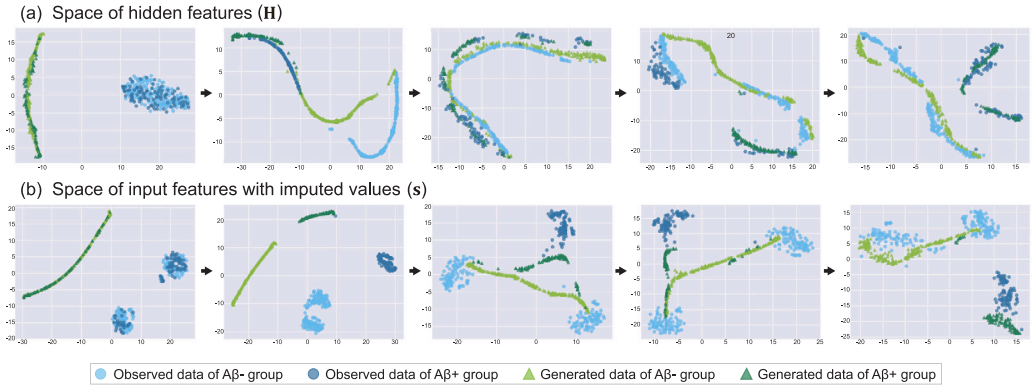

Fig. 5. t-SNE analysis of the class conditional generation performance of our model. Successive charts, from left to right, show the displacement of hidden features (a) and input features with imputed values (b) as the model is trained.

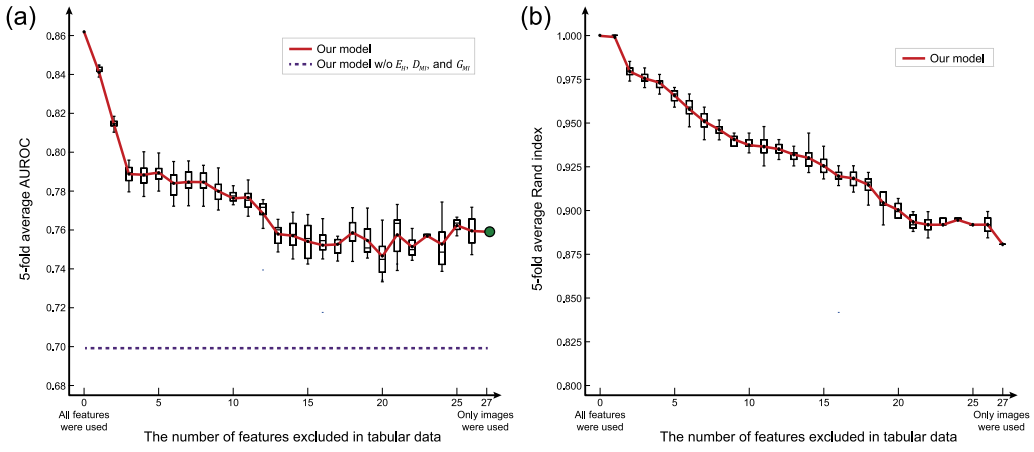

Fig. 6. Inference using partial features. (a) The classification performance. The average AUROCs over all possible combinations of features are shown as black dots, each within a box-and-whisker plot. The green solid circle shows the performance when a model trained on both tabular features and MRI scans was tested with only MRI scans. The purple dashed line shows the performance of a model trained on MRI scans alone. (b) The ability of our model to make the same prediction when using partial features in inference. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.7. Beyond missing values: imputing missing features

We tested the ability of our model to cope with features that are missing entirely. We measured the 5-fold average of test AUROCs using the trained model as features are removed one by one, measured by 5-fold cross validation. We greedily removed the features which produced the largest drop in AUROC. Fig. 6(a) shows that the AUROC drops sharply, as expected, but the AUROC remains above 0.74. Even when the fully trained model was tested with only MRI scans (Fig. 6(a), green solid circle), it showed better performance than the model that was trained with only MRI scans (Fig. 6(a), purple dashed line). This model is similar to TripleGAN [42] but it performs oversampling in an embedded vector space as well as semi-supervised learning.

We also tested the ability of our model trained with both MRI scans and all tabular features to make the same predictions when using different subsets of the features in inference. We used Rand indices [43] to compare predictions of unlabeled data with all tabular features (reference) and those of subsets of features. We greedily removed the features that produced the largest drop in the Rand index. Predictions become less consistent as features are excluded, as we would expect. Nevertheless, only 12.3% of predictions changed when all the tabular features were excluded (Fig. 6(b)).

This was evaluated using the Rand index [43] to 'reference' labels (predictions) of unlabeled data when all tabular features were used as features were removed one by one.

3.8. Translation from discovery to practice datasets

It is desirable to use the rich information available in the discovery dataset from a large cohort study when we train the model with a practice dataset collected in another clinical setting. However, both the feature sets and the diagnostic criteria found in the practice dataset from a specific clinical context can differ from those in the discovery dataset (Fig. 7(b), left plot). The discrepancy between the data distributions of the large cohort data and the practice dataset is addressed by transfer learning and fine-tuning (Fig. 7(a)). Specifically, we initially trained the model using the weights obtained from fold 3 of the discovery dataset, which exhibited the best AUROC (as shown in Table B.5). Subsequently, we continued training this model on the entire discovery dataset until achieving an AUROC of 1. We then trained the model on the practice dataset. This allows a model trained on the discovery dataset to synthesize features missing in the practice data due to not being collected by the relevant hospital, and then to use them to make predictions from the practice data.

We can visualize the labeled input features $\mathbf{s}_l$ by reducing their dimensionality using t-SNE. Looking at $\mathbf{s}_l$ obtained from the model trained with the discovery dataset, the A$\beta-$ and A$\beta+$ groups in the practice datasets are not clearly differentiated. After fine-tuning the model with the practice dataset, the separation of the A$\beta-$ and A$\beta+$ groups in the practice dataset is much clearer (Fig. 7(b), right plot), indicating that the model has been adjusted to the practice dataset.

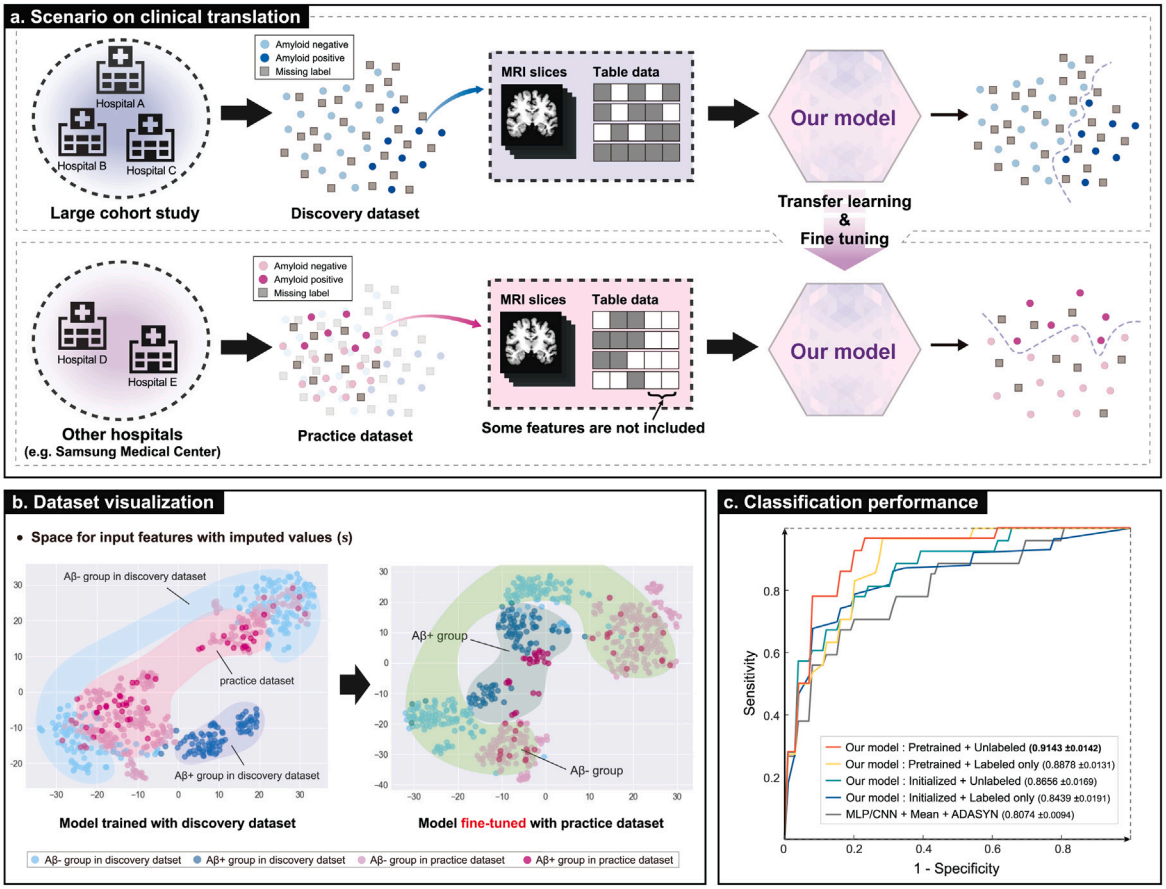

Fig. 7. Transferring information obtained from a large cohort study to specific clinical practice settings. (a) Scenario on clinical translation from the discovery dataset (obtained from the large cohort study) to the practice dataset (collected in another clinical setting). (b) Visualization of input features with imputed values (s) before and after fine-tuning. (c) ROC curves for our model fine-tuned with the practice dataset and comparative models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

After fine-tuning, our model achieved an average AUROC of 0.9143, an average accuracy of 0.8528, an average sensitivity of 0.9667, and an average specificity of 0.8286 at optimal thresholds determined by the Youden index (Fig. 7(c), red line). The effectiveness of fine-tuning is demonstrated by the significantly worse performance of a model trained from randomly initialized weights (Fig. 7(c), cyan line). Our method also outperforms other deep learning models with machine learning techniques designed for imperfect datasets (Fig. 7(c), gray line). We can also see that semi-supervised learning improves the prediction performance (Fig. 7(c), red versus yellow and cyan versus blue).

3.9. Discriminative regions and variables related to amyloid positivity

The three most discriminative regions of amyloid positivity exist in the right posterior temporal lobe, and in the right and left lateral remainders of the occipital lobes (Fig. 8(a) and (b)). The most discriminative tabular features are sex and the number of APOE $\epsilon 4$ alleles (Fig. 8(c)).

When our model is trained using fine-tuning with the practice dataset, the most discriminative regions or variables were consistent with those with the discovery dataset. However, the significance patterns were somewhat different from each other. The top three discriminative regions are the left and right lateral remainders of the occipital lobes, and the brainstem (Fig. 8(d) and (e)). Again, the most discriminative tabular features are sex, and the number of APOE $\epsilon 4$ alleles are highly discriminative of amyloid positivity. The most discriminative tabular feature is the number of APOE $\epsilon 4$ alleles (Fig. 8(f)).

4. Discussion

We have proposed a deep learning technique for predicting the preclinical stage of Alzheimer's disease in cognitively normal individuals from structural magnetic resonance imaging scans, demographic variables and various cognitive scores. The proposed model can cope with real-world situations in which (a) the training dataset is imperfect, (b) the feature set of the test data is a subset of the features used in training, and (c) different labels are used in the test and training data. Our deep generative model overcomes these real-world problems in the clinical implementation of deep learning for the determination of early Aβ pathology by implicitly estimating a joint distribution of features and outcomes [12]. We also show that non-linear discriminative regions can be used to explain how our model allocates data to Aβ+ or Aβ−.

The HexaGAN framework successfully addressed the imperfect dataset problem in terms of 'imputation' by creating additional realistic data: substitutes for missing values were created by optimizing element-wise adversarial loss between a generator G_{MI} and a discriminator D_{MI}. This is shown to be more accurate than previous competitive techniques; pseudo-labels to replace missing labels were generated using a classifier C, which is trained with the discriminator D_{MI} in adversarial fashion; and instances of the minority class were oversampled by the class-conditional generation. Our model based on the HexaGAN framework efficiently utilizes imperfect data and, therefore, effectively predicts diagnoses in a real-world situation while interplaying between components to solve these sub-problems simultaneously, not separately in a different order.

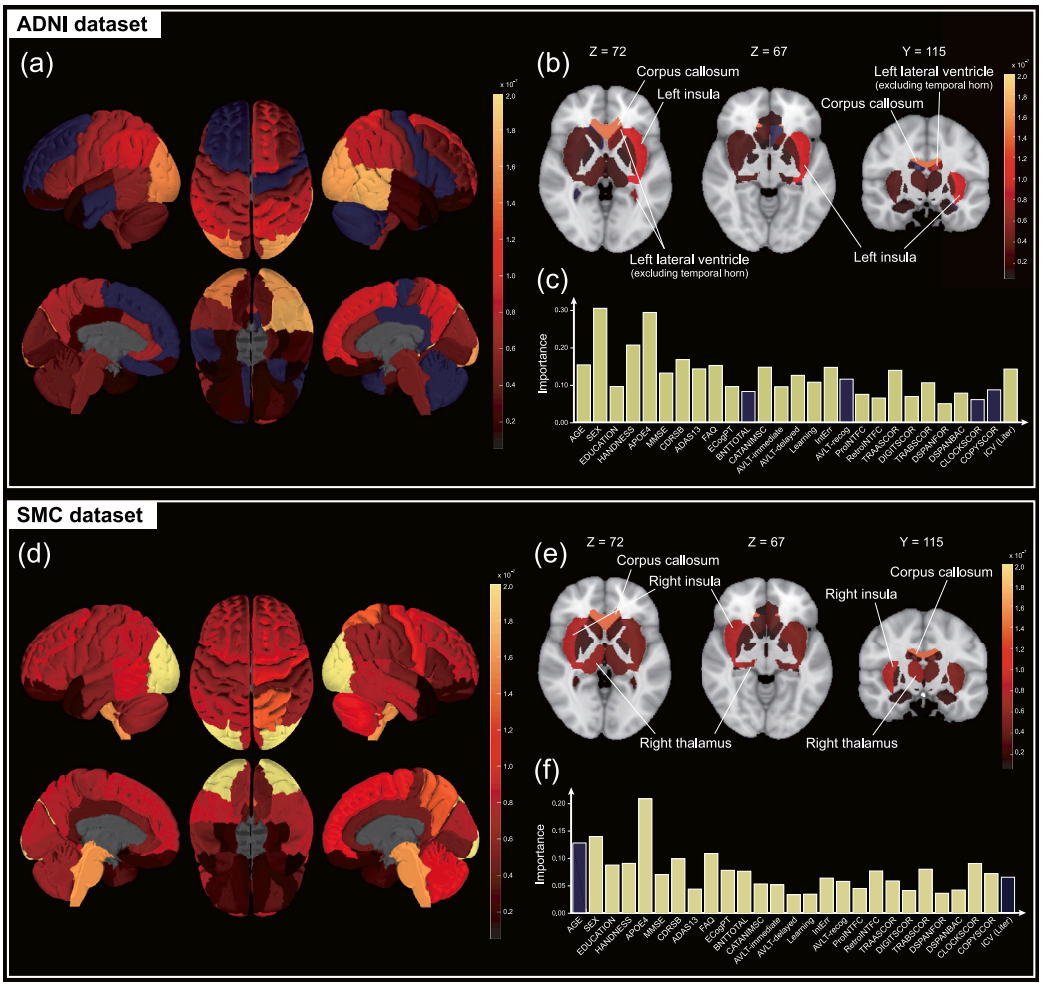

Fig. 8. Discriminative regions and variables of amyloid positivity determined by our model. (a) Brain maps for interpreting how our model trained with the discovery dataset predicted amyloid positivity. Color scale represents importance values, and dark blue regions make no significant contribution to classifying amyloid positivity. (b) Discriminative regions that cannot be shown on the cortical surface. (c) Discriminative features: those colored dark blue make no significant contribution to classifying amyloid positivity. (d)–(f) Discriminative regions and variables, fine-tuned with the practice dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The imputation of missing data is based on the data distribution. However, it is difficult to estimate the distribution of high-dimensional features of MRI scans together with tabular data. Thus, we reduce the dimensionality of this distribution. Firstly, we selected 18 slices from the total of 210 coronal slices of MRI scans. The prediction performance of our model with only a small number of slices rather than the whole image indicates that deep learning can be applied to predict the preclinical stage of AD using MRI scans acquired in clinical practice. Secondly, we introduced the encoder E_h into HexaGAN to extract low-dimensional embedded vectors $\mathbf{h}$ from the MRI scans. Our model can oversample embedded vectors $\mathbf{h}$ with hidden features from the image data, which makes imputation much more straightforward, rather than a high-dimensional MRI scan per se, which could depress the classification performance by the curse of dimensionality [44].

The prediction performance of our model trained with the discovery dataset was the AUROC of 0.8609±0.0266 (Fig. 2), which was higher than previous models trained with cross-sectional structural MRI scans [45–48], and comparable with a model trained on longitudinal structural MRI scans [49]. We attribute this performance improvement to the capability of DL models to find non-linear relationships between features and predictions. In our model, this capability is increased by using all observed data efficiently via imputation, which increases the number of data available for training. This theoretically brings down the upper bound on the difference between the generalization error and the empirical error [50]. In addition, reducing the dimensionality of the MRI scan is a pragmatic way to reduce the size of the problem and to make the predictions of our model more comprehensible. These features of our model are combined with deep generative models and appear to give better results than previous studies [51] with the currently feasible capability of DL models.

For the class imbalance problem, our approach leverages the adversarial game between G_{CG} and D_{CG} to learn the distribution of hidden features corresponding to the minority class ($p(\mathbf{H}_g|y_g)$). Additionally, G_{MI} models the distribution of input features given the hidden features ($p(\mathbf{s}|\mathbf{H})$). By combining G_{CG} and G_{MI}, we can model the distribution of input features given the minority class $\left(p(\mathbf{s}_g|y_g) = p(\mathbf{H}_g|y_g)\ p(\mathbf{s}_g|\mathbf{H}_g)\right)$. This allows us to effectively generate high-quality synthetic data for the minority class through the excellent distribution estimation capability of HexaGAN (Fig. 5).

Fig. 2(a) suggests that the predictions made by our model were most accurate when both image and tabular data were used as features. The accuracy dropped when only tabular data was used and dropped further when only image data was used. This suggests that the encoder E_h is effective in extracting features from reduced-dimensionality images obtained from MRI scans, although the tabular features are more important. In addition, E_H of our model combined the essential information of image data with tabular data and exerted a synergistic effect on the performance of our model. On the other hand, the performance of

discriminative DL models such as MLP or CNN was not much affected by the omission of the image data. This is because CNNs are poor at extracting information from image data [44].

We would expect our network to be applicable to many clinical situations, in which both the available features and their labeling differ from those in the dataset used to train the model (Fig. 7(a)). Our model can impute values for features absent from the practice dataset by estimating the class-conditional density $p(\mathbf{H}|y)$. The effectiveness of this is shown in Fig. 6(a). The problem of inconsistent labeling can be addressed by fine-tuning. Although fine-tuning confines the search area to learn the model for new data, models can employ information from the large data inherent in pretrained models [27]. Several studies corroborated that fine-tuning is also effective in constructing prediction models for biomedical data [52,53]. In addition, fine-tuning can also be beneficial for generative models when the size of a training dataset is limited to generate synthetic data [54]. Fig. 7(b) shows that fine-tuning is successful in adjusting our model to the practice dataset: the difference in the class-conditional density $p(\mathbf{s}|y)$ between the discovery dataset and the practice dataset is reduced, and the model's predictions are more accurate (Fig. 7(c)). Fine-tuning with the practice dataset had little effect on the discriminative profile of either the image or the tabular data (Fig. 8).

Explainable AI is a method that allows humans to understand the outputs generated by machine learning models. It helps to characterize model accuracy and fairness and describes the expected effects and potential biases of the model [55]. Individual-level explainability describes the prediction of an instance created by the model. This may provide an understanding of each individual (e.g., false positives or false negatives). Population-level explainability is measured by pooling information on individual-level explainability. This provides abstraction and summary of the model's decision boundary through discriminative variables and provides the model's reliability within the comprehension capability of humans (e.g., physicians).

DL models are known to be difficult to make or verify hypotheses. Nevertheless, it is often valuable to know how the features influence a prediction. Discriminative regions and variables can help physicians understand model predictions and diagnose diseases, while also providing reliability to patients. The top three discriminative regions are more associated with the regions exhibiting pathological changes, including Aβ and tau deposition. It is expected because individuals in the preclinical AD stage experience these changes before the structural change or cognitive decline [56]. Cerebral regions with the highest discriminative power are mostly in the posterior region of the brain, linked to the late stages of Aβ pathology [57,58]. High discriminative power was also observed in the midbrain in which Aβ is deposited at an earlier stage of symptomatic AD [59]. Although this deposition occurs after the preclinical AD stage, our PET-based criteria for determining amyloid positivity leads to its high discriminative power, as it tends to favor the much later stages of amyloid pathology compared to cerebrospinal fluid (CSF) measures [60]. In addition, the midbrain (and cerebellum) is related to the early stage of tau pathology [61–63], consistent with tau deposition occurring after amyloid deposition [56]. These discriminative regions also overlap with regions from previous classifications [64,65]. We note that this discriminative power may not be a direct result of amyloid or tau deposition, but a concomitant alteration of texture or shape derived from pathological changes. For the tabular features, the association between the presence of APOE $\varepsilon 4$, the most discriminative variable in our model, and Aβ deposition has been previously reported [47,66,67]. Sex, an important discriminator in our network trained on the discovery dataset alone, is known to be a factor in AD [68–70], but the discriminative power of this feature may be, in part, influenced by a characteristic of the dataset, which is skewed toward female participants (Table 1). Our ablation studies (Fig. 6(a)) also suggested that these two features were the most discriminative. We believe that this is the first study in which XAI techniques have been applied to multi-modal (image and table) data and provided discriminative regions and features for the entire dataset as well as for each participant.

Our model has some limitations. Firstly, an MRI scan is always required for inference. This is mainly because all records in the datasets we used have MRI scans. We plan to extend our method to enable inference from tabular data alone by incorporating the missingness of MRI scans. Secondly, it is not clear whether the importance values that are used to interpret model predictions arise from the morphological characteristics of the brain, such as cortical thickness or cortical widening, rather than solely from the MRI intensity. We leave this avenue of research for future work. Furthermore, there have been studies on predicting the progression of AD from longitudinal data [71–73]. We believe that extending our research to predict preclinical AD from longitudinal data is also an intriguing future work.

In conclusion, we developed a method for detecting the amyloid positivity of CN individuals using proxy measures, including structural MRI scans, demographic information, and clinical scores with a deep generative model. In tandem with the growth of artificial intelligence (AI) systems for electronic health records (EHRs), deep generative models will effectively address the imperfect dataset problem in EHRs and allow us to successfully perform clinical translation. Moreover, the fusion of XAI techniques and statistical tests will help locate important regions and features for detecting diseases and provide the reliability of the model in the real world.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A3B1077720); the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2023; Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)); and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A5A708390811). This research was also supported by the NRF grant funded by the Ministry of Science and ICT (MSIT) of the Korea government (No. 2023R1A2C2006201).

Appendix A. Details on learning deep neural networks

A.1. Network architectures

The detailed architecture of seven components in our model can be found in Table A.4. For each column in Table A.4, the rows represent layers of a given component in sequence, beginning with the input layer and ending with the output layer.

A.2. Training details

Each component of the whole system is updated in order. Since the distribution of H is altered by updating E_H and E_h, we update G_{CG} and D_{CG} 30 times for each update of the other components, in order to estimate the distribution of H accurately. That is, $30\times(G_{\mathrm{CG}} \rightarrow D_{\mathrm{CG}}) \rightarrow E_H \rightarrow G_{\mathrm{MI}} \rightarrow D_{\mathrm{MI}} \rightarrow E_h \rightarrow C$.

We adapted HexaGAN to deal with problems derived from the size of MRI data in two ways. Firstly, we added an additional encoder

Table A.4
Network architectures.

E_h	E_H	G_{MI}	D_{MI}
$I \in \mathbb{R}^{168\times190\times18}$	$(\tilde{\mathbf{t}}, \mathbf{m}) \in \mathbb{R}^{27+27}$	$\mathbf{H} \in \mathbb{R}^{256}$	$\mathbf{t} \in \mathbb{R}^{27}$
DenseNet121	FC(512) + SN + ReLU	FC(512) + SN + ReLU	FC(2048) + SN + LReLU
FC(2048) + Sigmoid	FC(2048) + SN + ReLU	FC(512) + SN + ReLU	Concat($\mathbf{h}$) $\in \mathbb{R}^{2048+2048}$
	Concat($\mathbf{h}$) $\in \mathbb{R}^{2048+2048}$	FC(512) + SN + ReLU	FC(512) + SN + LReLU
	FC(512) + SN + ReLU	FC(512) + SN + ReLU	FC(512) + SN + LReLU
	FC(512) + SN + ReLU	FC(512) + SN + ReLU	FC(512) + SN + LReLU
	FC(256) + Tanh	FC(512) + SN + ReLU	FC(512) + SN + LReLU
		FC(27+2048) + Sigmoid	Projection(y, 27+2048+2) + Sigmoid
G_{CG}	D_{CG}	C	
$(\mathbf{z}, y) \in \mathbb{R}^{27+2}$	$(\mathbf{H}, y) \in \mathbb{R}^{256+2}$	$\mathbf{t} \in \mathbb{R}^{27}$	
FC(512) + SN	FC(512) + SN + LReLU	FC(2048) + ReLU + Dropout(0.5)	
FC(512) + SN	FC(512) + SN + LReLU	Concat($\mathbf{h}$) $\in \mathbb{R}^{2048+2048}$	
FC(512) + SN	FC(512) + SN + LReLU	FC(512) + ReLU + Dropout(0.5)	
FC(256) + Tanh	FC(512) + SN + LReLU	FC(512) + ReLU + Dropout(0.5)	
	Projection(y, 1) + Sigmoid	FC(2) + Softmax	

FC(m): Fully-connected layer with hidden dimension m.
SN: Spectral normalization [25].
DenseNet121: DenseNet model [26] pretrained on the ImageNet dataset.
ReLU, LReLU, Tanh, Sigmoid, Softmax: Activation functions.
Concat(a): The output of the previous layer is concatenated with a.
Dropout(p): Output neurons are randomly turned off with probability p [74].
Projection(y, d): Projection layer [21] to inject the conditional information of y with output dimension d.

Table B.5
Classification performance of our model over five folds.

Dataset	Fold	AUROC	Cut-off[a]	Sensitivity[b]	Specificity[c]	Accuracy[d]
Discovery dataset	Fold 1	0.8757	0.0101	1.0000	0.7105	0.8036
	Fold 2	0.8165	0.1168	0.6842	0.8919	0.8214
	Fold 3	**0.8834**	**0.3193**	**0.7895**	**0.8649**	**0.8393**
	Fold 4	0.8720	0.0039	0.7895	0.8378	0.8214
	Fold 5	0.8574	0.0001	0.9444	0.7838	0.8364
Practice dataset	Fold 1	0.8333	0.7339	0.8333	0.8077	0.8125
	Fold 2	0.9000	0.0048	1.0000	0.7692	0.8065
	Fold 3	0.9385	0.0035	1.0000	0.8462	0.8710
	Fold 4	**0.9667**	**0.6872**	**1.0000**	**0.9200**	**0.9355**
	Fold 5	0.9333	0.0002	1.0000	0.8000	0.8387

AUROC, area under the receiver operating characteristic curve.
BOLD: Folds where the highest AUROC was achieved.
[a] Optimal cut-offs were determined by the Youden index.
[b] Sensitivity = (number of individuals correctly identified as Aβ+)/(total number of Aβ+ individuals).
[c] Specificity = (number of individuals correctly identified as Aβ−)/(total number of Aβ− individuals).
[d] Accuracy = (number of correctly identified individuals)/(total number of individuals).

Table B.6
Classification performance with MRI scans and tabular data.

Architecture	Missing data	Class imbalance	Missing label	AUROC	F1-score	PRAUC
MLP + CNN	Column-wise deletion	–	–	0.7043	0.2128	0.5523
		SMOTE	–	0.7111	0.5255	0.5837
		ADASYN	–	0.7231	0.5104	0.5769
		Cost-sensitive	–	0.7032	0.2057	0.5589
		Class rectification loss	–	0.7066	0.3061	0.5560
	Mean imputation	–	–	0.7099	0.2334	0.5641
		SMOTE	–	0.7104	0.4999	0.5767
		ADASYN	–	**0.7347**	0.4040	0.5990
		Cost-sensitive	–	0.7143	0.0986	0.5801
		Class rectification loss	–	**0.7343**	0.5067	0.5820
		ADASYN	k-nearest neighbors	0.6889	0.5093	0.5547
		ADASYN	Label propagation	0.6761	0.4999	0.5156
	k-nearest neighbors	–	–	0.7141	0.3182	0.5570
		SMOTE	–	0.7110	0.3986	0.5553
		ADASYN	–	0.7121	0.4217	0.5778
		Cost-sensitive	–	0.7280	0.2124	0.6042
		Class rectification loss	–	0.7221	0.2228	0.5593
		Cost-sensitive	k-nearest neighbors	0.7044	0.1013	0.5809
		Cost-sensitive	Label propagation	**0.7340**	0.2532	0.5858

(*continued on next page*)

Table B.6 (*continued*).

Architecture	Missing data	Class imbalance	Missing label	AUROC	F1-score	PRAUC
		–	–	0.7180	0.1000	0.5689
		SMOTE	–	0.7150	0.5013	0.5757
		ADASYN	–	0.7100	0.5198	0.5789
	MICE	Cost-sensitive	–	0.7004	0.0000	0.5600
		Class rectification loss	–	0.7075	0.3174	0.5665
		SMOTE	*k*-nearest neighbors	0.6818	0.4999	0.5452
		SMOTE	Label propagation	0.6840	0.5200	0.5568
Our model	HexaGAN	HexaGAN	HexaGAN	**0.8609**	**0.7596**	**0.6421**

AUROC, area under the receiver operating characteristic curve.
PRAUC, the area under the precision–recall curve.

Table B.7
Classification performance with MRI scans.

Architecture	Class imbalance	Missing label	AUROC	F1-score	PRAUC
	–	–	**0.6424**	0.4054	0.5439
	SMOTE	–	0.6158	0.1013	0.5166
	ADASYN	–	**0.6503**	0.3986	0.5439
CNN	Cost-sensitive	–	0.6021	0.0000	0.5128
	Class rectification loss	–	0.6217	0.0000	0.5364
	ADASYN	*k*-nearest neighbors	0.6229	0.1013	0.5175
	ADASYN	Label propagation	**0.6503**	0.4041	0.5526
Our model w/o E_H, D_{MI}, and G_{MI}	HexaGAN	HexaGAN	**0.6998**	0.5715	0.6086

AUROC, area under the receiver operating characteristic curve.
PRAUC, the area under the precision–recall curve.

Table B.8
Classification performance with tabular data.

Architecture	Class imbalance	Class imbalance	Missing label	AUROC	F1-score	PRAUC
		–	–	0.7481	0.6224	0.5884
		SMOTE	–	0.7511	0.6624	0.6277
	Column-wise deletion	ADASYN	–	0.7496	0.6170	0.6263
		Cost-sensitive	–	0.7492	0.6126	0.6191
		Class rectification loss	–	**0.7515**	0.6084	0.6170
		–	–	0.7373	0.6003	0.6216
		SMOTE	–	0.7335	0.6279	0.6184
		ADASYN	–	0.7409	0.6238	0.6218
	Mean imputation	Cost-sensitive	–	0.7311	0.5768	0.6219
		Class rectification loss	–	0.7339	0.5899	0.6158
		ADASYN	*k*-nearest neighbors	**0.7532**	0.6093	0.5836
		ADASYN	Label propagation	0.7445	0.5999	0.5779
MLP		–	–	0.7409	0.5872	0.6271
		SMOTE	–	0.7402	0.6212	0.6202
		ADASYN	–	0.7419	0.6171	0.6265
	k-nearest neighbors	Cost-sensitive	–	0.7411	0.5933	0.6305
		Class rectification loss	–	0.7423	0.5933	0.6271
		Class rectification loss	*k*-nearest neighbors	0.7458	0.6061	0.6199
		Class rectification loss	Label propagation	**0.7577**	0.5983	0.6172
		–	–	0.7424	0.5955	0.6268
		SMOTE	–	0.7377	0.6287	0.6073
		ADASYN	–	0.7380	0.6311	0.6127
	MICE	Cost-sensitive	–	0.7345	0.5832	0.6197
		Class rectification loss	–	0.7392	0.5958	0.6198
		–	*k*-nearest neighbors	0.7441	0.6002	0.6079
		–	Label propagation	0.7503	0.6222	0.6005
Our model w/o E_h	HexaGAN	HexaGAN	HexaGAN	**0.7613**	0.7514	0.6313

AUROC, area under the receiver operating characteristic curve.
PRAUC, the area under the precision–recall curve.

network E_h in front of HexaGAN to reduce the dimensionality of the data. To train the encoder E_h more efficiently, its training is augmented by transfer learning using DenseNet with parameters pre-trained on ImageNet [26,27] except the last two layers. We applied spectral normalization to the weights of every layer of E_H, D_{CG}, G_{CG}, D_{MI}, and G_{MI} except output layers of E_H, G_{CG}, and G_{MI}, to stabilize the learning process [25]. Further, we added projection discriminators to D_{CG} and D_{MI}, in order to reinforce class conditioning of the GAN [21].

We implemented deep neural networks using the Tensorflow library [75]. We used the Adam optimizer to update the networks with the aim of minimizing loss functions through back-propagation. We used the Adam optimizer with $\beta_1 = 0$ and $\beta_2 = 0.9$. The learning rate of C was set to 0.0001, and those of D_{CG} and D_{MI} were set to 0.00016 by applying the two time-scale update rule (TTUR) [76], and that of the other components were set to 0.00004. We used one NVIDIA TITAN V for training the model.

We calculated test AUROCs for all epochs during model training and stored the model parameters at the epochs that show the highest test AUROCs at each fold. We averaged these five test AUROCs to produce a final classification performance. The CNNs in comparison

models are trained in a similar way to ours, by transfer learning from DenseNet parameters pre-trained on ImageNet. We also set their hyperparameters, such as the learning rate, to the same values that we used in our model.

Appendix B. Comparative experiment results

See Tables B.5–B.8.

References

[1] Crous-Bou M, Minguillón C, Gramunt N, Molinuevo JL. Alzheimer's disease prevention: from risk factors to early intervention. Alzheimer's Res Ther 2017;9(1):71.

[2] Cummings J, Lee G, Ritter A, Sabbagh M, Zhong K. Alzheimer's disease drug development pipeline: 2019. Alzheimer's Dementia Transl Res Clin Intervent 2019;5:272–93.

[3] Iaccarino L, Tammewar G, Ayakta N, Baker SL, Bejanin A, Boxer AL, et al. Local and distant relationships between amyloid, tau and neurodegeneration in Alzheimer's Disease. NeuroImage Clin 2018;17:452–64.

[4] Mormino EC, Betensky RA, Hedden T, Schultz AP, Amariglio RE, Rentz DM, et al. Synergistic effect of β-amyloid and neurodegeneration on cognitive decline in clinically normal individuals. JAMA Neurol 2014;71(11):1379–85.

[5] Andrews KA, Frost C, Modat M, Cardoso MJ, Rowe CC, Villemagne V, et al. Acceleration of hippocampal atrophy rates in asymptomatic amyloidosis. Neurobiol Aging 2016;39:99–107.

[6] Becker JA, Hedden T, Carmasin J, Maye J, Rentz DM, Putcha D, et al. Amyloid-β associated cortical thinning in clinically normal elderly. Ann Neurol 2011;69(6):1032–42.

[7] Gurevich P, Stuke H, Kastrup A, Stuke H, Hildebrandt H. Neuropsychological testing and machine learning distinguish Alzheimer's disease from other causes for cognitive impairment. Front Aging Neurosci 2017;9:114.

[8] Ko H, Ihm J-J, Kim H-G, Initiative ADN, et al. Cognitive profiling related to cerebral amyloid beta burden using machine learning approaches. Front Aging Neurosci 2019;11.

[9] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.

[10] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform 2017;18(5):851–69.

[11] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in Neural Information Processing Systems. Vol. 27. 2014, p. 2672–80.

[12] Hwang U, Jung D, Yoon S. HexaGAN: Generative adversarial nets for real world classification. In: Proceedings of the 36th International Conference on Machine Learning. 2019, p. 2921–30.

[13] Cho SH, Choe YS, Kim YJ, Kim HJ, Jang H, Kim Y, et al. Head-to-head comparison of 18F-florbetaben and 18F-flutemetamol in the cortical and striatal regions. J Alzheimer's Dis 2020;(Preprint):1–10.

[14] Gousias IS, Rueckert D, Heckemann RA, Dyet LE, Boardman JP, Edwards AD, et al. Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. Neuroimage 2008;40(2):672–84.

[15] Landau SM, Breault C, Joshi AD, Pontecorvo M, Mathis CA, Jagust WJ, et al. Amyloid-β imaging with Pittsburgh compound B and florbetapir: comparing radiotracers and quantification methods. J Nucl Med 2013;54(1):70–7.

[16] Farrar G. Regional visual read inspection of [18F] flutemetamol brain images from end-of-life and amnestic MCI subjects. J Nucl Med 2017;58(supplement 1):1250.

[17] Hahn A, Kim YJ, Kim HJ, Jang H, Cho H, Choi SH, et al. The preclinical amyloid sensitive composite to determine subtle cognitive differences in preclinical Alzheimer's disease. Sci Rep 2020;10(1):1–11.

[18] Battista P, Salvatore C, Castiglioni I. Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: a machine learning study. Behav Neurol 2017;2017.

[19] Thomas KR, Eppig J, Edmonds EC, Jacobs DM, Libon DJ, Au R, et al. Word-list intrusion errors predict progression to mild cognitive impairment. Neuropsychology 2018;32(2):235.

[20] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning. 2017, p. 214–23.

[21] Miyato T, Koyama M. cGANs with projection discriminator. In: International Conference on Learning Representations. 2018.

[22] Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw 2015;61:85–117.

[23] Sutskever I, Jozefowicz R, Gregor K, Rezende D, Lillicrap T, Vinyals O. Towards principled unsupervised learning. 2015, arXiv preprint arXiv:1511.06440.

[24] Mescheder L, Geiger A, Nowozin S. Which training methods for GANs do actually converge? In: International Conference on Machine Learning. 2018, p. 3481–90.

[25] Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations. 2018, URL https://openreview.net/forum?id=B1QRgziT-.

[26] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, p. 4700–8.

[27] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2009;22(10):1345–59.

[28] Stawiaski J. A pretrained densenet encoder for brain tumor segmentation. In: International MICCAI Brainlesion Workshop. Springer; 2018, p. 105–15.

[29] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17(6):520–5.

[30] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatric Res 2011;20(1):40–9.

[31] Zhu X, Ghahramani Z. Learning from Labeled and Unlabeled Data with Label Propagation. Citeseer; 2002.

[32] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.

[33] He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks. IEEE; 2008, p. 1322–8.

[34] Sun Y, Kamel MS, Wong AK, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit 2007;40(12):3358–78.

[35] Dong Q, Gong S, Zhu X. Imbalanced deep learning by minority class incremental rectification. IEEE Trans Pattern Anal Mach Intell 2018;41(6):1367–81.

[36] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[37] Maaten Lvd, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(Nov):2579–605.

[38] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning. Vol. 70. JMLR.org; 2017, p. 3319–28.

[39] Kapishnikov A, Bolukbasi T, Viégas F, Terry M. XRAI: Better attributions through regions. In: Proceedings of the IEEE International Conference on Computer Vision. 2019, p. 4948–57.

[40] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32–5.

[41] Ali A, Shamsuddin SM, Ralescu AL. Classification with class imbalance problem. Int J Adv Soft Comput Appl 2013;5(3).

[42] LI C, Xu T, Zhu J, Zhang B. Triple generative adversarial nets. In: Advances in Neural Information Processing Systems. Vol. 30. 2017, p. 4088–98.

[43] Rand WM. Objective criteria for the evaluation of clustering methods. J Amer Statist Assoc 1971;66(336):846–50.

[44] Friedman JH. On bias, variance, 0/1—loss, and the curse-of-dimensionality. Data Min Knowl Discov 1997;1(1):55–77.

[45] Langford O, Raman R, Sperling R, Cummings J, Sun C-K, Jimenez-Maggiora G, et al. Predicting amyloid burden to accelerate recruitment of secondary prevention clinical trials. J Prevent Alzheimer's Dis 2020;7(4):213–8.

[46] Pekkala T, Hall A, Ngandu T, van Gils M, Helisalmi S, Hänninen T, et al. Detecting amyloid positivity in elderly with increased risk of cognitive decline. Front Aging Neurosci 2020;12.

[47] Ten Kate M, Redolfi A, Peira E, Bos I, Vos SJ, Vandenberghe R, et al. MRI predictors of amyloid pathology: results from the EMIF-AD Multimodal Biomarker Discovery study. Alzheimer's Res Ther 2018;10(1):1–12.

[48] Tosun D, Veitch D, Aisen P, Jack Jr CR, Jagust WJ, Petersen RC, et al. Detection of β-amyloid positivity in Alzheimer's Disease Neuroimaging Initiative participants with demographics, cognition, MRI and plasma biomarkers. Brain Commun 2021;3(2):fcab008.

[49] Petrone PM, Casamitjana A, Falcon C, Artigues M, Operto G, Cacciaglia R, et al. Prediction of amyloid pathology in cognitively unimpaired individuals using voxel-wise analysis of longitudinal structural brain MRI. Alzheimer's Res Ther 2019;11(1):72.

[50] Xu H, Mannor S. Robustness and generalization. Mach Learn 2012;86(3):391–423.

[51] Abrol A, Fu Z, Salman M, Silva R, Du Y, Plis S, et al. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. Nat Commun 2021;12(1):1–17.

[52] Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021;4(1):1–13.

[53] Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. J Am Med Inf Assoc 2014;21(4):699–706.

[54] Wang Y, Wu C, Herranz L, van de Weijer J, Gonzalez-Garcia A, Raducanu B. Transferring gans: Generating images from limited data. In: Proceedings of the European Conference on Computer Vision. 2018, p. 218–34.

[55] El-Sappagh S, Alonso-Moral JM, Abuhmed T, Ali F, Bugarín-Diz A. Trustworthy artificial intelligence in Alzheimer's disease: state of the art, opportunities, and challenges. Artif Intell Rev 2023;1–148.

[56] Jack CR, Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol 2013;12(2):207–16.
[57] Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathologica 1991;82(4):239–59.
[58] Grothe MJ, Barthel H, Sepulcre J, Dyrba M, Sabri O, Teipel SJ, et al. In vivo staging of regional amyloid deposition. Neurology 2017;89(20):2031–8.
[59] Koychev I, Hofer M, Friedman N. Correlation of Alzheimer disease neuropathologic staging with amyloid and tau scintigraphic imaging biomarkers. J Nucl Med 2020;61(10):1413–8.
[60] Palmqvist S, Schöll M, Strandberg O, Mattsson N, Stomrud E, Zetterberg H, et al. Earliest accumulation of β-amyloid occurs within the default-mode network and concurrently affects brain connectivity. Nat Commun 2017;8(1):1–13.
[61] Uematsu M, Nakamura A, Ebashi M, Hirokawa K, Takahashi R, Uchihara T. Brainstem tau pathology in Alzheimer's disease is characterized by increase of three repeat tau and independent of amyloid β. Acta Neuropathol Commun 2018;6(1):1–18.
[62] Uchihara T. Neurofibrillary changes undergoing morphological and biochemical changes–How does tau with the profile shift of from four repeat to three repeat spread in Alzheimer brain? Neuropathology 2020;40(5):450–9.
[63] Jové M, Mota-Martorell N, Torres P, Ayala V, Portero-Otin M, Ferrer I, et al. The causal role of lipoxidative damage in mitochondrial bioenergetic dysfunction linked to Alzheimer's disease pathology. Life 2021;11(5):388.
[64] Casamitjana A, Petrone P, Tucholka A, Falcon C, Skouras S, Molinuevo JL, et al. MRI-based screening of preclinical Alzheimer's disease for prevention clinical trials. J Alzheimer's Dis 2018;64(4):1099–112.
[65] de Vries BM, Golla SS, Ebenau J, Verfaillie SC, Timmers T, Heeman F, et al. Classification of negative and positive 18 F-florbetapir brain PET studies in subjective cognitive decline patients using a convolutional neural network. Eur J Nucl Med Mol Imag 2021;48(3):721–8.
[66] Landau SM, Horng A, Fero A, Jagust WJ, Alzheimer's Disease Neuroimaging Initiative, et al. Amyloid negativity in patients with clinically diagnosed Alzheimer disease and MCI. Neurology 2016;86(15):1377–85.
[67] Risacher SL, Kim S, Shen L, Nho K, Foroud T, Green RC, et al. The role of apolipoprotein E (APOE) genotype in early mild cognitive impairment (E-MCI). Front Aging Neurosci 2013;5:11.
[68] Bai F, Zhang Z, Watson DR, Yu H, Shi Y, Zhu W, et al. Absent gender differences of hippocampal atrophy in amnestic type mild cognitive impairment. Neurosci Lett 2009;450(2):85–9.
[69] Li R, Singh M. Sex differences in cognitive impairment and Alzheimer's disease. Front Neuroendocrinol 2014;35(3):385–403.
[70] Mielke MM, Vemuri P, Rocca WA. Clinical epidemiology of Alzheimer's disease: assessing sex and gender differences. Clin Epidemiol 2014;6:37.
[71] El-Sappagh S, Saleh H, Sahal R, Abuhmed T, Islam SR, Ali F, et al. Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data. Future Gener Comput Syst 2021;115:680–99.
[72] El-Sappagh S, Saleh H, Ali F, Amer E, Abuhmed T. Two-stage deep learning model for Alzheimer's disease detection and prediction of the mild cognitive impairment time. Neural Comput Appl 2022;34(17):14487–509.
[73] El-Sappagh S, Ali F, Abuhmed T, Singh J, Alonso JM. Automatic detection of Alzheimer's disease progression: An efficient information fusion approach with heterogeneous ensemble classifiers. Neurocomputing 2022;512:203–24.
[74] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929–58.
[75] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation. 2016, p. 265–83.
[76] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems. Vol. 30. 2017, p. 6626–37.