RESEARCH ARTICLE

Adversarial Learning for MRI Reconstruction and Classification of Cognitively Impaired Individuals

XIAO ZHOU[1], AKSHARA R. BALACHANDRA[2], MICHAEL F. ROMANO[3,4], SANG PETER CHIN[1,5,6], RHODA AU[7,8,9,10,11], AND VIJAYA B. KOLACHALAMA[3,10,12], (Senior Member, IEEE)

[1]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA
[2]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA
[3]Department of Medicine, Boston University Chobanian and Avedisian School of Medicine, Boston, MA 02118, USA
[4]Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA 94143, USA
[5]Thayer School of Engineering, Dartmouth College, Hanover, NH 03755, USA
[6]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, USA
[7]Department of Anatomy and Neurobiology, Boston University Chobanian and Avedisian School of Medicine, Boston, MA 02118, USA
[8]Department of Neurology, Boston University Chobanian and Avedisian School of Medicine, Boston, MA 02118, USA
[9]Framingham Heart Study, Boston University Chobanian and Avedisian School of Medicine, Boston, MA 02118, USA
[10]Boston University Alzheimer's Disease Research Center, Boston, MA 02118, USA
[11]Department of Epidemiology, Boston University School of Public Health, Boston, MA 02118, USA
[12]Department of Computer Science, Faculty of Computing and Data Sciences, Boston University, Boston, MA 02215, USA

Corresponding author: Vijaya B. Kolachalama (vkola@bu.edu)

ABSTRACT Game theory-inspired deep learning using a generative adversarial network provides an environment to competitively interact and accomplish a goal. In the context of medical imaging, most work has focused on achieving single tasks such as improving image resolution, segmenting images, and correcting motion artifacts. We developed a dual-objective adversarial learning framework that simultaneously 1) reconstructs higher quality brain magnetic resonance images (MRIs) that 2) retain disease-specific imaging features critical for predicting progression from mild cognitive impairment (MCI) to Alzheimer's disease (AD). We obtained 3-Tesla, T1-weighted brain MRIs of participants from the Alzheimer's Disease Neuroimaging Initiative (ADNI, N=342) and the National Alzheimer's Coordinating Center (NACC, N = 190) datasets. We simulated MRIs with missing data by removing 50% of sagittal slices from the original scans (i.e., diced scans). The generator was trained to reconstruct brain MRIs using the diced scans as input. We introduced a classifier into the GAN architecture to discriminate between stable (i.e., sMCI) and progressive MCI (i.e., pMCI) based on the generated images to facilitate encoding of disease-related information during reconstruction. The framework was trained using ADNI data and externally validated on NACC data. In the NACC cohort, generated images had better image quality than the diced scans (Structural similarity (SSIM) index: 0.553 ± 0.116 versus 0.348 ± 0.108). Furthermore, a classifier utilizing the generated images distinguished pMCI from sMCI more accurately than with the diced scans (F1-score: 0.634 ± 0.019 versus 0.573 ± 0.028). Competitive deep learning has potential to facilitate disease-oriented image reconstruction in those at risk of developing Alzheimer's disease.

INDEX TERMS Image reconstruction, generative adversarial network, magnetic resonance imaging, Alzheimer's disease.

The associate editor coordinating the review of this manuscript and approving it for publication was Jinhua Sheng.

I. INTRODUCTION

Detection of preclinical Alzheimer's disease (AD) is advancing greatly through novel developments in structural (e.g.,

MRI) and functional (e.g., PET) neuroimaging [1]. The quality of the MRI scan, determined largely by the scanner itself, can play an important role in accurately representing regions affected by disease [2]. Since raw MRI scans are acquired in k-space, image reconstruction algorithms are needed to transform the acquired raw data into interpretable image representations. MRI reconstruction may involve several signal processing steps such as noise pre-whitening for phased array data acquisition, interpolation, filtering, and k-space to image space transformation [3]. Since all these steps are executed on the raw MRI data, the reconstruction process is not necessarily aimed towards accurate disease detection. Modifying the objective of reconstruction to ensure the generated images are of high quality with attention to accurate diagnosis of a disease can have profound practical implications.

We contend that a generative adversarial network (GAN) is ideally suited to accomplish such a task by leveraging the principles of the two-player zero-sum game from game theory. GANs utilize the minimax principle from game theory, where the generator and the discriminator compete with each other [4]. In the context of medical imaging, a classical GAN contains two players: 1) the generator, which takes images from one domain (e.g., low-quality brain MRIs) as input and attempts to create images similar to original training data (e.g., high-quality brain MRIs) and 2) the discriminator, which tries to distinguish the generator's output from the original training data. The discriminator mainly serves to ensure that the generated MRIs adhere to a distribution of original brain MRIs.

A. RELATED WORK

GAN-based approaches have been successful in MRI reconstruction. Yang et al. developed a conditional GAN-based model for fast compressed sensing MRI reconstruction [5], resulting in output images with preserved perceptual image details. In a related study, Quan et al. developed a GAN framework with deeper generator and discriminator networks and cyclic data consistency loss for improved interpolation in the given undersampled k-space data, followed by the use of a chained network to improve the quality of image reconstruction [6]. Additionally, Shaul et al. proposed a two-stage GAN framework for MRI reconstruction from undersampled k-space data, which estimated the missing k-space samples and fixed aliasing artifacts in the image space [7]. Chen and collaborators developed a multi-level densely connected super-resolution framework that incorporated GAN loss to enhance low resolution T1-weighted MRI volumes [8]. Others have developed GAN-based deep learning frameworks to enable reconstruction of high-quality MRI scans from low quality scans with fast acquisition time [9], [10], [11].

Most of these efforts focused on accurate reconstruction as the sole objective, where the frameworks were optimized for producing higher quality images and/or for time-efficiency. Further, there is limited literature on the use of MRI reconstruction techniques to clarify specific clinical concerns, for example distinguishing between progressive (pMCI) and stable mild cognitive impairment (sMCI) within the Alzheimer's disease spectrum. Iglesias et al. recently published a public AI tool, SynthSR, that creates high quality isotropic T1-weighted brain MRIs using any clinical brain input scan [12]. Though they validated this tool with brains from persons with Alzheimer's disease, the main objective of their tool was not for enhancing diagnostic ability, but rather for use in 3D morphometry software. While the study established a proof-of-principle that competitive deep learning can be leveraged to achieve multiple objectives, the cases that were utilized to achieve the goal (i.e., scans of multiple magnetic field strengths obtained on the same individuals taken at the same time) are not routinely available in most real-world clinical scenarios. Therefore, GAN approaches that can simultaneously address multiple objectives in a more practical setting can be tremendously helpful in the clinical setting, aiding physicians in prognostication of MCI earlier in the disease course without the use of invasive testing (i.e., lumbar puncture), since not all persons with MCI progress to AD [13].

Broadly, MCI is clinically diagnosed when individuals show deficits on cognitive testing while still being able to carry out their instrumental activities of daily living (IADLs; e.g., cooking, managing finances, etc.). Once individuals with MCI are unable to do any one IADL, they have clinically progressed to dementia (i.e., they had pMCI). Although medial temporal atrophy affecting the hippocampus is usually seen in MCI due to AD pathology and greater atrophy portends a worse prognosis, it is often difficult to determine hippocampal atrophy by visual inspection or with measures of cortical thickness early in the disease course (i.e., when individuals are still in the MCI stage), as the imaging changes can be very subtle [14], [15]. In related AD research, classifying between pMCI and sMCI (sMCI; i.e., individuals with MCI who do not progress to dementia) is generally considered to be challenging as the structural differences in the brain can be subtle [16], [17], [18]. This challenge may be compounded when MRIs are of lower quality, whether due to artifacts such as motion or low magnetic field strength, for example.

In this work, we trained a dual-objective GAN to (1) construct higher quality brain MRIs with missing data, while simultaneously (2) retaining disease-specific features critical for classification of AD progression (pMCI versus sMCI). Given that classification of AD progression among persons with MCI is a clinically difficult task, this model serves as a proof-of-concept which may be applied to other diagnostic challenges. To achieve these objectives, we extend the classical GAN architecture by introducing a classifier that distinguishes the reconstructed brain MRIs from the generator as belonging to a subject with pMCI versus sMCI, thus encouraging the generator to create realistic brain images while retaining disease-related information. The classifier can also help to mitigate the feature hallucination problem,

TABLE 1. Study population.

(a) ADNI[a] cohort		
	pMCI[b]	sMCI[b]
N (% Male)	116 (57.8)	226 (57.1)
Age, years	73.0 ± 7.2 (116)	71.2 ± 7.3 (226)
MMSE[b] (N)	27.1 ± 1.8 (116)	28.2 ± 1.6 (226)
Years of educ. (N)	15.9 ± 2.7 (116)	16.4 ± 2.6 (226)
APOE4 (#0/1/2 - N)	38/58/20	138/75/13
(b) NACC[a] cohort		
N (% Male)	95 (48.4)	95 (62.1)
Age, years	75.9 ± 9.4	73.5 ± 10.2
MMSE[b] (N)	25.9 ± 2.5 (55)	27.3 ± 2.0 (75)
Years of educ. (N)	15.4 ± 2.9 (95)	15.7 ± 2.4 (95)
APOE4 (#0/1/2 - N)	38/33/12	58/31/4

Demographics summary of the participants from the ADNI and NACC datasets. Age, MMSE, and years of education are reported as means with the respective standard deviations.
[a] ADNI = Alzheimer's Disease Neuroimaging Initiative, NACC = National Alzheimer's Coordinating Center
[b] pMCI = progressive mild cognitive impairment, sMCI = stable mild cognitive impairment, MMSE = Mini Mental State Exam

where generative networks can sometimes add or remove key elements from images [19]. We implemented this framework with MRIs in image-space. This allows generalizability to existing national and international working groups with access to large imaging studies which typically lack k-space data.

B. CONTRIBUTIONS
The main contributions of this paper are summarized below:
- We developed a dual-objective adversarial learning framework that is able to reconstruct higher quality brain MRIs that retain disease-related information in the setting of MCI due to AD pathology. We show that including a classifier in the GAN architecture increases the predictive value of the generated images, likely by encoding more disease-related information in the reconstructed scans.
- Our framework is able to address stability problems commonly encountered in training GANs by balancing the learning speeds of the generator, discriminator, and classifier with fine-tuned learning rates and having additional training iterations for the generator and classifier.
- We validate the predictive value of the generated scans in distinguishing pMCI from sMCI to address potential concerns of feature hallucination in the reconstructed images.
- We use T1-weighted 3T brain MRIs and corresponding clinical data from two independent datasets and demonstrated that our dual-objective GAN is generalizable.

II. MATERIALS AND METHODS
A. STUDY POPULATION
We obtained access to clinical and neuroimaging data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the National Alzheimer's Coordinating Center (NACC) to perform this study.

ADNI is a multi-center project focused on curating clinical, imaging, genetic, and biochemical biomarkers for studying AD [21]. Our study utilized 3T T1-weighted MRIs that were obtained at a visit where the respective participant received a diagnosis of MCI and had cerebrospinal fluid data available. Where multiple such events were present for a single person, MRIs were refined by selecting the scan collected at the most recent date, and by utilizing fully sampled scans where possible. Participants were assigned diagnosis labels of pMCI if they progressed to AD within (0-36] months (rounded down to the nearest month) and sMCI if they did not progress to AD within 36 months (Table 1).

NACC is a large database containing research data from Alzheimer's Disease Research Centers across the United States. We utilized clinical and imaging data from a data freeze on December 12, 2020. For each subject, we first identified the visits at which they were diagnosed with amnestic or non-amnestic MCI. MRIs were chosen to minimize the time between a clinical diagnosis of MCI and the date of MRI acquisition. If the shortest time between an MCI clinical visit and MRI was longer than 6 months, the subject was excluded. Occasionally, participants had multiple MRIs that satisfied these criteria; in these cases, only one of these MRIs was utilized (Table 1).

B. IMAGE PROCESSING
We pre-processed MRIs using SPM12 (Statistical Parametric Mapping, https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) and MATLAB version 2020a. T1-weighted MRIs were bias-corrected, normalized into MNI space and skull-stripped. We adopted an approach to skull stripping similar to the one described by Mitchell et al. [22] in which SPM12-derived probabilistic maps of GM, WM, and CSF were summed and thresholded at 0.2 to produce binary whole-brain masks. We performed quality checks of each individual binary mask to ensure skull stripping was performed correctly and 1 subject from the ADNI cohort was excluded due to poor pre-processing results. We opted to use the SPM12 pipeline for skull stripping as it has been shown to accurately estimate total intracranial volume when compared with expert manual estimation [23]. The binary masks were then applied to the normalized, bias-corrected MRIs. Additional details can be found in our previously published work [24]. Henceforth, these normalized, bias-corrected, and masked MRIs will be referred to as 'original scans'. 'Diced' scans were created by randomly selecting half of the sagittal slices of the original scans and replacing their values with zeroes. These diced scans were used as input to the generator. A schematic of our image processing workflow is provided in Fig. 1, and an overview of the classifier is presented in Fig. 2.

Tissue probability maps obtained from SPM12 corresponding to CSF were thresholded at 0.2, and plotted as blue over the reconstructed images (Fig. 3). A background mask was obtained by identifying relevant brain parenchyma as the sum of white matter, gray matter, and CSF tissue probability maps, thresholding at 0.2, and setting the background as the
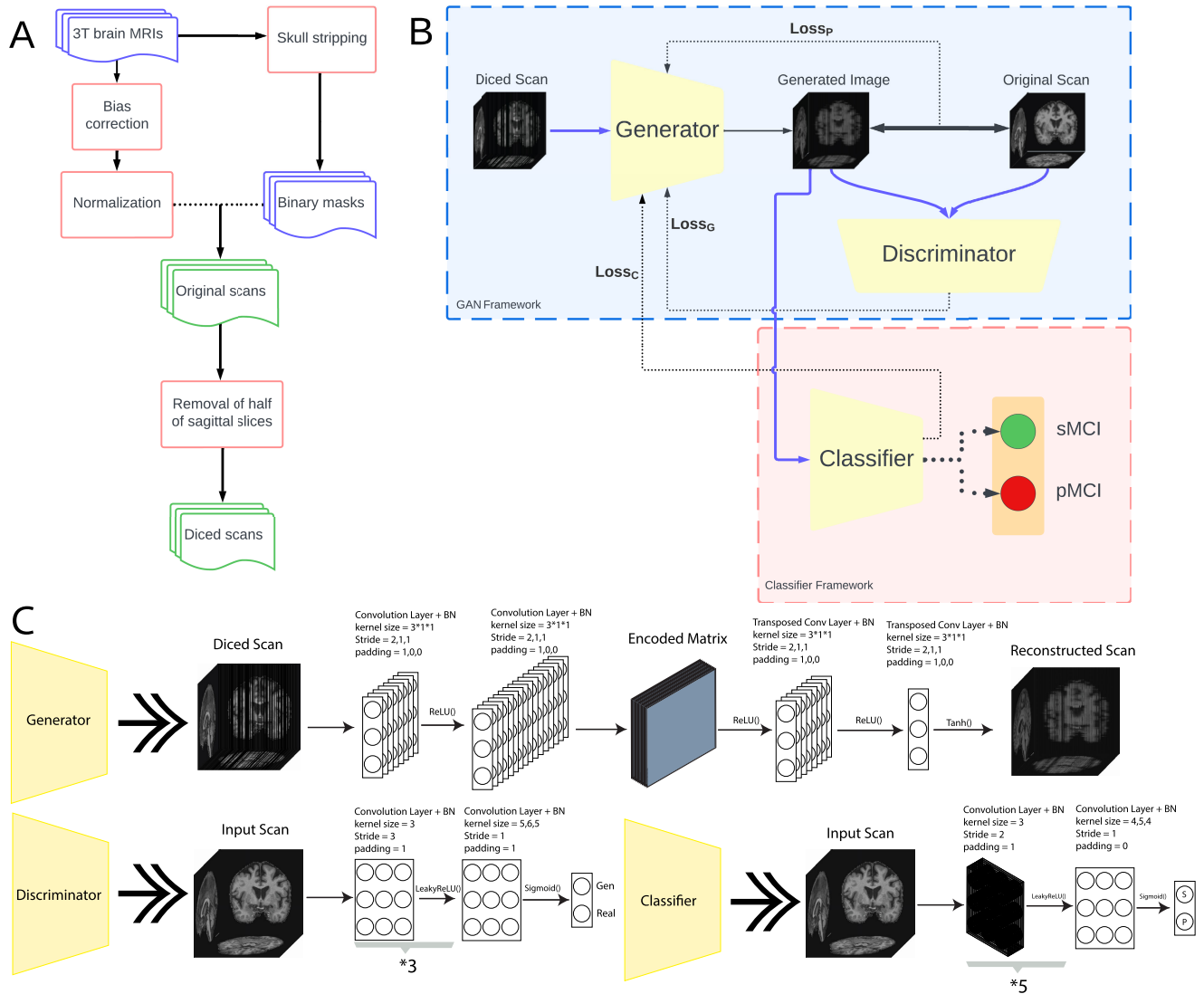
FIGURE 1. Image processing workflow and adversarial learning framework. (A) Raw T1 brain MRIs were pre-processed using SPM12 with bias-correction, normalization, and skull stripping. Binary whole-brain masks were applied to the normalized MRIs yielding the "original" images. Diced scans were created by randomly selecting half of the sagittal slices from the original scans and replacing them with zeroes. (B) Diced scans serve as input to the generator. The generator incorporates losses from the discriminator (Loss$_G$), classifier (Loss$_C$), and a new 'perception loss' (Loss$_P$) derived from the comparison of the generated images with the corresponding reference images (i.e., Original scans). While the discriminator attempts to differentiate between the generated images and the original scans, the classifier uses the generated images to classify persons who have sMCI from those with pMCI. (C) The internal compositions of the generator, discriminator, and classifier are shown. Detailed descriptions can be found in the text (See Section II-C).

voxels not part of the brain parenchyma. These voxels were set to black in Fig. 3.

C. COMPUTATIONAL FRAMEWORK

We constructed an adversarial network for three-dimensional MR image reconstruction while simultaneously attempting to classify persons as having sMCI or pMCI. In our framework, we distinguish between two types of GANs: the standard GAN (GAN-VAN) and a variant with a modified loss function (GAN-NOV). GAN-VAN follows the conventional GAN architecture, while GAN-NOV incorporates novel loss functions, enhancing the model's ability to generate more diverse and realistic 3D images. This framework (GAN-NOV) consisted of one generator, one discriminator,

and one classifier. The generator takes diced scans as input and attempts to generate a more continuous version of the image. At the same time, the generated image is compared with the original scan via the discriminator. Concomitantly, the generated image is used as input to the classifier to predict if the original scan is from someone with sMCI or pMCI. Thus, our framework is performing two tasks, image reconstruction and classification. We incorporated three separate loss functions during model training, each corresponding to one part of our framework: 1) discriminator loss, 2) classification loss, and 3) perception loss (Fig. 1). After the model was trained, we compared the generated images with the original scans using various image quality metrics. We also evaluated the predictive ability of the
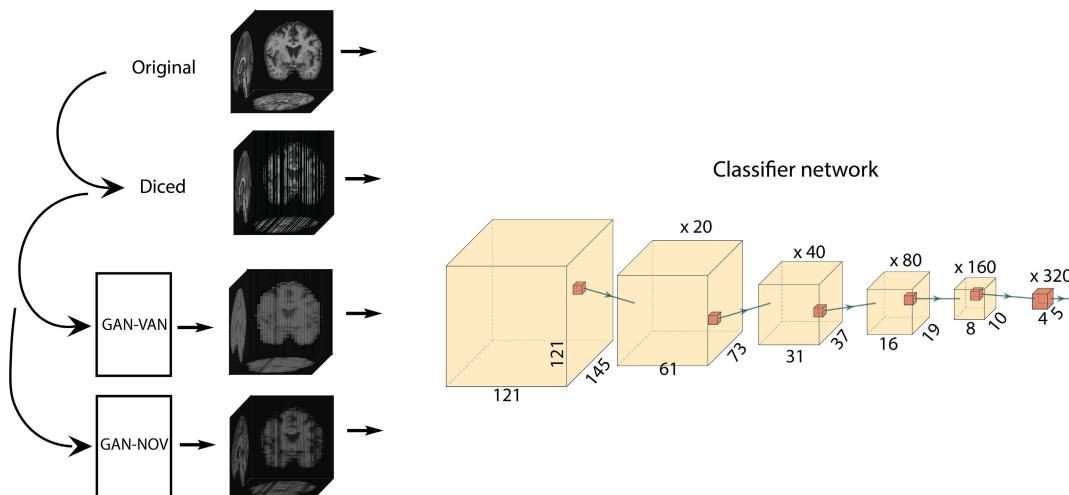
FIGURE 2. Schematics of the classifier. An exemplar MRI before ('Original') and after ('Diced') random removal of sagittal slices. Also shown are examples of brain MRIs following two different GAN-based methods of imputing the missing slices, GAN-VAN and GAN-NOV. On the right is a visual representation of the three-dimensional classifier network used to differentiate between stable and progressive MCI cases. The classifier network representation was generated using PlotNeuralNet v1.0.0 [20].
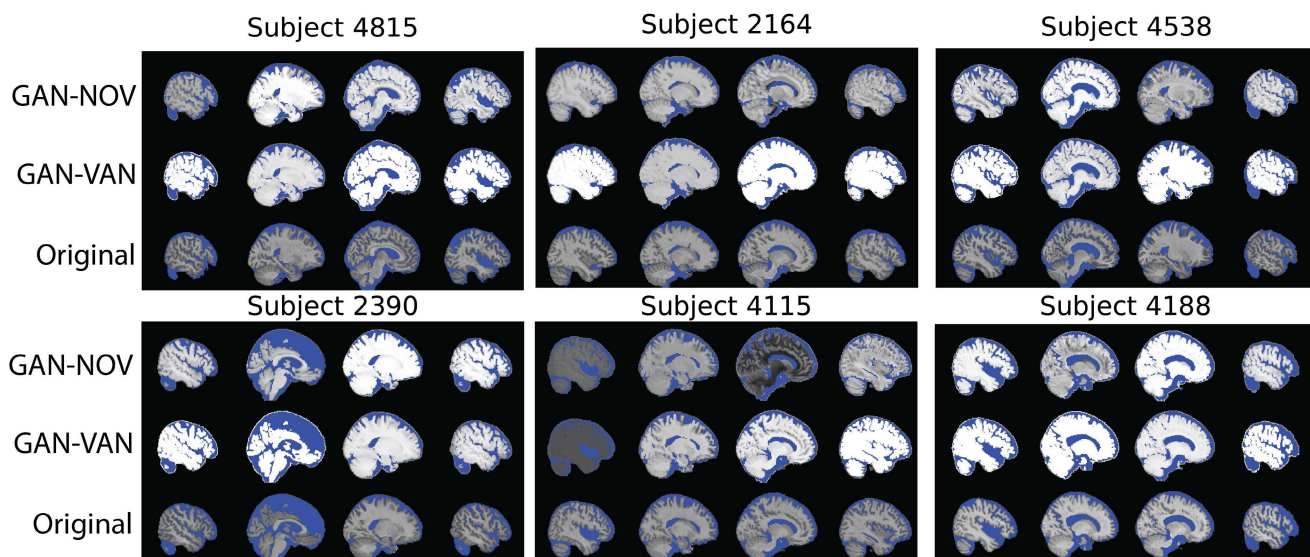

FIGURE 3. Sample reconstructed brain slices. Sample sagittal slices from six T1-weighted MRIs, labeled "Original" (bottom), and reconstructions, with signal intensities rescaled between the range [0, 1]. Reconstructed images are from either a generator utilizing perception loss, GAN loss, and classification loss together (GAN-NOV), or a generator utilizing perception loss and GAN loss only (GAN-VAN). Blue indicates CSF, while backgrounds were set to black.

generated images by comparing the performance of a classifier trained on the generated images with a classifier trained on the original scans.

Let Z denote the set of input slices to be reconstructed, z as an instance of the set; $g \in G$ as the reconstructed image, and $t \in T$ as the original scan. To reconstruct a complete 3D volume given a few 2D slices, a model capable of generating missing slices based on spatial context is necessary. In an adversarial setting, the two networks are competing against each other and the objective can be formulated as

$$\min_{G} \max_{D} V(D, G) = E_t[\log(D(t))] + E_z[\log(1 - D(G(z)))],$$

(1)

where V is the function value, $D(t)$ is the discriminator's output given t as the input, and $G(z)$ is the generator's output given z as input.

The objective of the discriminator is to distinguish between the generated and reference data. Since the output contains binary outcomes, binary cross entropy was used to measure the loss defined as

$$L_D = -\text{mean}[\log(D(t)) + \log(1 - D(G(z)))].$$

(2)

The objective of the generator is to create outputs that are indistinguishable from the real data. The loss for generator is

defined as

$$L_G = -\log(D(G(z))). \quad (3)$$

To guide the generator to reconstruct an image that captured disease-related information, we included binary cross entropy loss from the classifier, defined as

$$L_C = -\text{mean}[\log(C(t_{pMCI})) + \log(1 - C(t_{sMCI}))], \quad (4)$$

where $sMCI$ signifies that the scan originates from a person with sMCI (stable Mild Cognitive Impairment), while $pMCI$ indicates that the scan is from an individual with pMCI (progressive Mild Cognitive Impairment). This specific loss can penalize the network if the generated images are incorrectly classified based on AD (Alzheimer's Disease) status. Such mechanism helps the generator in focus on disease-related information, thereby enhancing the model's ability to distinguish between these critical clinical states. We also introduced a perception loss that estimated the distance between the reconstructed image and the target image to restrict the generator from deviating from the target image, defined as

$$L_P = -\text{abs}(|G(z) - t|). \quad (5)$$

This loss function evaluates the output image by calculating the absolute difference between the generated image $G(z)$ and the target image t. Specifically, L_P quantifies the magnitude of deviation, without considering the direction of this difference. This loss measures how far off each pixel value of the generated image is from its corresponding value in the target image. By minimizing this loss, the generator learns to produce image that match the target better, and ensuring a higher degree of image quality in the output image.

The complete objective function for the generator is defined as

$$L_{Gs} = \lambda_G \cdot L_G + \lambda_C \cdot L_C + \lambda_P \cdot L_P, \quad (6)$$

where λ_G, λ_C, and λ_P are the weights assigned for the corresponding losses. In our experiments, we observed that assigning higher weights to a specific loss component can enhance the model's performance in certain areas, such as significantly boosting classification accuracy. However, this improvement often comes with a trade-off, leading to a deterioration in performance in other aspects, such as image quality. Based on these findings, we propose that an equivalent weight assignment across different loss components yields the best overall results, striking an optimal balance between classification accuracy and image quality.

Overall, the generator was composed of 2 convolutional and 2 transposed convolutional layers, where each of them have a kernel of size [3,1,1] with a padding of [1,0,0]. The convolutional layers had a stride of [2,1,1] to perform in lieu of discrete pooling layers. Batch normalization was applied for each layer's output to rescale the parameters. The activation for the last layer was $tanh$, while the other layers used ReLU as activation function. The tanh activation

function is used as the final layer of the generator because it has several benefits that helps the network have better results and greater training speed: (1) The tanh function outputs values in the range $[-1,1]$, which is compatible with our pre-normalized MRI scan data. (2) It increases training stability and minimizes the risk of unbounded outputs by mapping extreme values into this bounded range (3) Due to its nature, the tanh function maps both positive and negative inputs away from zero, thereby promoting faster, yet smoother, learning.

The discriminator was composed of 4 convolutional layers, where the first three had a kernel of size [3,3,3] with a padding of [1,1,1] and a stride of 3, and the last layer had a kernel of size [5,6,5] without padding and a stride of 1. Similar to the generator, we did not apply pooling layers for the discriminator, and batch normalization was applied to each layer's output. We used LeakyReLU as the activation function for all layers except the last one. The last layer used the sigmoid activation function.

The classifier was composed of 6 convolutional layers, each with a doubled number of filters from the previous layer, starting from 20. The first five layers had a kernel of size [3,3,3] with a padding of [1,1,1] and a stride of 2. The last layer had a kernel of size [4,5,4] without padding and a stride of 1. We did not use any pooling layers. Batch normalization with a dropout rate of 0.25 was applied for each layer's output. Similar to the discriminator, LeakyReLU was used as the activation function for all layers except the last one. The last layer used the sigmoid activation function.

The activation functions used by the discriminator and classifier were LeakyReLU in the hidden layers and sigmoid in the final layer to optimize for performance and stability. Sigmoid was chosen in the last layer because (1) it maps the output to the range [0, 1], which is suitable for our prediction tasks and (2) our early optimization efforts indicated that it performs best for our tasks. We chose not to use sigmoid in the hidden layers since (1) it has a small gradient during backpropagation, which may cause the gradients to vanish when overlapped multiple times, (2) it is more expensive computationally, which would make the training slower, and (3) its output range can make the learning complex, as it prevents negative correlations between neurons. Finally, LeakyReLU was selected in hidden layers because it is more efficient than sigmoid and tanh, rendering faster training, and it allows for a small gradient for negative outputs, which helps prevent the vanishing gradient problem. These choices are supported by previous work that used similar settings [25], [26], [27] and empirical results from our early optimization experiments.

We tackled multiple technical challenges that often arise when training GAN frameworks. For example, during earlier epochs of training, it is possible for the discriminator to dominate, as it is straightforward to distinguish the generated image from the original (i.e., target) image. To mitigate this issue, we tuned the learning rate and paused the discriminator periodically to prevent it from learning faster than the

generator. Additionally, as discussed earlier, to ensure that the generator learns the appropriate way to reconstruct images with reference to the original scans, we introduced perception and classification losses as additional weighted loss terms.

We chose to construct our framework with a GAN because we observed in preliminary experiments that GAN variants like the Wasserstein GAN (WGAN) implementation did not perform optimally within our framework, potentially due to its loss function not aligning well with the added complexity of a classifier and our specific data characteristics. Additionally, considering our limited dataset size, which consists of only a few hundred samples, more complex architectures like those required for BiGAN, BEGAN, etc., could have led to significant overfitting. Furthermore, the computational resources required for 3D reconstruction with complex GAN variants would have risen significantly for our architecture.

Our network's structure and parameters are informed by insights from previous research, further refined through optimization in the current study for enhanced performance. We conducted hyperparameter tuning using a Bayesian approach, utilizing Gaussian Process (GP) within the WandB framework [28]. This method systematically evaluated the interplay between various parameters and their impact on model performance. The underlying formula for hyperparameter optimization allowed us to fine-tune our model with greater precision and efficiency:

$$
\begin{aligned}
a_{EI}(x; \{x_n, L_n\}, \theta) \\
= \sqrt{\Sigma(x, x; \{x_n, L_n\}, \theta)} \\
* [\gamma(x)\Phi(\gamma(x)) + N(\gamma(x); 0, 1)]
\end{aligned} \tag{7}
$$

where

$$
\gamma(x) = \frac{L_{best} - \mu(x; \{x_n, L_n\}, \theta)}{\sqrt{\Sigma(x, x; \{x_n, L_n\}, \theta)}} \tag{8}
$$

and

$$
\begin{aligned}
\Sigma(x, x'; \{x_n, L_n\}, \theta) \\
= K(x, x') - K(X, x)^\mathsf{T} K(X, X)^{-1} K(X, x').
\end{aligned} \tag{9}
$$

Note $\{x_n, L_n\}_1^N$ are the pairs of observations, and Φ is the cumulative distribution function of the standard Gaussian distribution ($N(0, 1)$), $\gamma(x)$ is Z-score, $K(X, x)$ is the N-dim vector of cross-co-variances between x and set X (which contains all xs). $K(X, X)$ is the Gram matrix for set X (K is a kernel function), $m : \chi \rightarrow R$ is the mean function, θ is the kernel parameters, and Matern 5/2 kernel was used (Table 2). We applied standard black-box optimization algorithms to optimize the function as described by Snoek et al. [29].

D. STATISTICS

We used Fisher exact test to statistically compare categorical data and Mann-Whitney U tests to compare continuous variables. To compare proportions of persons with different numbers of APOE4 alleles between the ADNI and NACC

TABLE 2. Variable definitions.

Variable	Description
$\{x_n, L_n\}_1^N$	observation pairs
Φ	CDF of normal distribution
$\gamma(x)$	Z-score
$K(X, x)$	N-dim vector of cross-covariances of x and X
$K(X, X)$	Gram matrix for set X
$m : \chi \rightarrow R$	mean function
θ	parameters of the kernel

Algorithm 1 GAN Model for 3D MR Image Reconstruction and Classification

1: **Initialize** Generator G, Discriminator D, and Classifier C's parameters with Kaiming uniform initialization method.
2: **for** each training iteration **do**
3:     Sample a batch of 3D MR images z from Z
4:     Generate reconstructed images $g = G(z)$
5:     Compute Discriminator Loss:
    $L_D = -\text{mean}[\log(D(t)) + \log(1 - D(G(z)))]$
6:     Update Discriminator weights using gradient descent on L_D
7:     Compute Generator Loss:
    $L_G = -\log(D(G(z)))$
    $L_C = -\text{mean}[\log(C(t_{pMCI})) + \log(1 - C(t_{sMCI}))]$
    $L_P = -\text{abs}(G(z) - t)$
    $L_{Gs} = \lambda_G \cdot L_G + \lambda_C \cdot L_C + \lambda_P \cdot L_P$
8:     Update Generator weights using gradient descent on L_{Gs}
9:     (Optionally) Update Classifier weights
10: **end for**
11: **Evaluate** the model:
12:     A. Compare generated images with original scans
13:     B. Evaluate classification performance on generated vs. original scans

cohorts, counts were pooled between bins. For example, to compare proportions of persons with 1 APOE4 allele between the two cohorts, the number of persons with 0 or 2 APOE4 alleles were combined.

E. DATA AND CODE AVAILABILITY

Both ADNI and NACC datasets are publicly available and can be obtained directly from the respective sources. Python scripts and manuals are made available on GitHub (https://github.com/vkola-lab/access2024)

F. COMPLEXITY ANALYSIS

1) TIME COMPLEXITY

In general, for convolutional operations, the time complexity is given by: $O(K^3 FWHD)$, where K is the kernel size, F is the number of filters, and W, H, D are the dimensions of the inputs. In our case, for the discriminator and classifier, the approximate time complexity can be represented as $O(3^3 \times 20 \times 121 \times 145 \times 121)$. For the generator, the time complexity is approximately: $O(3^3 \times 3 \times 121 \times 145 \times 121)$.

2) SPACE COMPLEXITY

In our model, the space complexity of the network is mostly determined by the number of parameters in the convolutional layers, which in general is $O(K^3 FC)$, where C is the number of input channels. For the discriminator and classifier, the approximate space complexity can be represented as $O(3^3 \times 20 \times 1)$. For the generator, the space complexity is approximately: $O(3^3 \times 3 \times 1)$.

3) NUMBER OF PARAMETERS

The total number of parameters in our generative adversarial framework is the sum of the parameters for the generator, discriminator, and classifier, which is 1,864,189.

4) FLOPS (FLOATING POINT OPERATIONS)

Our framework's FLOPs are determined by the number of floating-point operations required during a forward pass. The FLOPs for the generator is 1,673,021,020, for the discriminator is 2,661,531, and for the classifier is 1,696,252,460.

III. EXPERIMENTS

The ADNI dataset was randomly split 80% for training, 10% for validation, and 10% for internal testing. The optimal model was saved based on performance on the validation partition. The entire NACC dataset was used for external testing. We trained the framework (GAN-NOV) for 20 epochs with the Adam optimizer. The classifier was trained for 60 epochs with the stochastic gradient descent optimizer. Learning rates ranged from 0.001 to 0.005 for different components of the learning framework. We constructed the GAN-VAN framework to evaluate the added benefit of the classification loss (see Fig. 1). Thus, this generator only incorporates the GAN loss and perception loss without the classification loss. GAN-VAN was trained identically as GAN-NOV and used for comparison against GAN-NOV.

To assess if our reconstruction frameworks generated images that retained disease-relevant information, we trained and tested four separate CNN classifiers to discriminate between pMCI and sMCI utilizing unique sets of MRIs as inputs: original scans only, diced scans only, generated images from GAN-VAN only, and generated images from GAN-NOV only. The classifiers were initialized identically to the one in GAN-NOV and trained for 60 epochs with the stochastic gradient descent optimizer. The classifiers were trained, validated, and tested 5 times, each time with a different seed for the random splits of the ADNI cohort. We assessed the classification performance on the ADNI test partitions and on NACC.

A. COMPUTING INFRASTRUCTURE

Model training and evaluation was performed on a GPU workstation with an NVIDIA 2080Ti graphics card containing 11 GB memory. We used PyTorch (v1.13.0) to implement the model. The training speed was about 10-15 min for each epoch, or about 0.12-0.19 iterations/s, and training took

3.5-4 hours to reach convergence. The inference speed was 1.17s per MRI scan with a batch size of 1.

B. PERFORMANCE METRICS

The image quality metrics contrast-to-noise ratio (CNR), signal-to-noise ratio (SNR), and structural similarity index measure (SSIM) [30] were used to compare the differences in quality between original scans and generated images. Two no-reference algorithms were also used for comparisons: Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [31] and Perception-based Image Quality Evaluator (PIQE) [32]. Image quality metrics were calculated for each image in the ADNI test partition and the entire NACC cohort. We report the means and standard deviations of each metric for scans of the corresponding datasets in Table 3.

Assessment of the classifiers was done by first generating receiver operating characteristic (ROC) and precision-recall (PR) curves based on model predictions on both the ADNI test partition and the NACC cohort. Then, for each ROC and PR curve, we computed the area under the curve (AUC) statistic. We computed the mean AUCs and standard deviations for each model across the 5 folds. We report mean values, along with the respective standard deviations, of model accuracy, precision, F1-score and Matthews correlation coefficient (MCC) in Table 4.

IV. RESULTS

The primary objective of our study was to evaluate the efficacy of our multi-objective adversarial deep learning framework in MRI reconstruction and classification. We aimed to determine if our approach could produce high-quality images while retaining disease-relevant information when given low-quality images from patients with MCI. Our study populations did not differ in terms of sex or APOE scores in either the pMCI or sMCI groups (APOE: sMCI, all $p = 1$; pMCI, $p = 0.23$ (0 APOE4 alleles), $p = 0.58$ (1 APOE4 allele), and $p = 1$ (2 APOE4 alleles) (Bonferroni-corrected); sex: sMCI, $p = 0.46$; pMCI, $p = 0.21$). For sMCI, persons in the NACC cohort were older ($U = 12852$, $p = 0.0053$), had lower MMSE scores ($U = 5862$, $p = 4.4e - 05$, with 20 missing samples from NACC), and fewer years of education ($U = 9112.5$, $p = 0.03$). For pMCI, education did not differ significantly between the two cohorts ($U = 4858$, $p = 0.13$), though persons in the NACC cohort were older ($U = 6699$, $p = 0.007$) and had lower MMSE scores ($U = 2356$, $p = 0.0054$, 40 missing samples from NACC). There was a larger proportion of persons with at least 1 APOE4 allele in the pMCI groups in both NACC and ADNI ($p = 0.034$ (NACC) and $p < 0.0001$ (ADNI)). Details of these performance are provided in Fig. 4, Table 3 and 4, as discussed in following sections.

A. QUALITATIVE ASSESSMENT

To visually assess the quality of reconstructed images, we compared them with the original scans. This qualitative

TABLE 3. Image quality results.

(a). Results on ADNI test partition, batch normalization					
Input Images	CNR	SNR	SSIM	BRISQUE	PIQE
Original	2.111±0.274	0.686±0.052	-	43.513±1.594	41.936±2.062
Diced	0.726±0.131	0.537±0.046	0.319±0.065	43.567±0.333	78.307±5.322
GAN-VAN	**2.235±0.299** ↑	**1.857±0.214** ↑	0.544±0.051	42.485±0.741	72.028±3.068
GAN-NOV	1.934±0.335	1.437±0.395	**0.580±0.059** ↑	**42.246±0.563** ↑	**68.120±1.351** ↑
(b). Results on NACC cohort, batch normalization					
Original	2.065±0.294	0.676±0.074	-	44.292±2.940	43.182±6.298
Diced	0.696±0.156	0.526±0.074	0.348±0.108	43.742±1.256	78.442±4.695
GAN-VAN	**2.193±0.359** ↑	**1.821±0.230** ↑	0.523±0.105	42.563±1.128	72.133±4.024
GAN-NOV	1.790±0.363	1.432±0.534	**0.553±0.116** ↑	**42.505±1.066** ↑	**68.176±2.079** ↑
(c). Results on ADNI test partition, group normalization					
Input Images	CNR	SNR	SSIM	BRISQUE	PIQE
Original	2.111±0.274	0.686±0.052	1.000±0.000	43.513±1.594	41.936±2.062
Diced	0.719±0.172	0.538±0.044	0.318±0.064	43.583±0.333	78.427±4.124
GAN-VAN	**2.351±0.621** ↑	**1.278±0.167** ↑	**0.584±0.061** ↑	**40.541±2.579** ↑	**63.735±2.530** ↑
GAN-NOV	2.203±0.478	0.791±0.027	0.550±0.055	42.389±0.504	76.495±1.513
(d). Results on NACC cohort, group normalization					
Original	2.065±0.294	0.676±0.074	1.000±0.000	44.292±2.940	43.182±6.298
Diced	0.701±0.162	0.525±0.076	0.346±0.109	43.773±1.310	78.844±4.502
GAN-VAN	**2.247±0.538** ↑	**1.299±0.143** ↑	**0.560±0.126** ↑	**40.973±2.666** ↑	**65.059±3.403** ↑
GAN-NOV	2.101±0.403	0.795±0.056	0.513±0.116	42.649±1.354	77.005±1.811

Means and standard deviation of contrast to noise ratio (CNR), signal to noise ratio (SNR), structural similarity (SSIM), blind/referenceless image spatial quality evaluator (BRISQUE), and perception-based image quality evaluator (PIQE) of all the corresponding images in the ADNI test partition and in NACC are shown. The set of best performing images in each metric is highlighted in bold text. Higher scores of CNR, SNR, and SSIM indicate better quality, while lower scores of BRISQUE and PIQE indicate better quality. It can be observed that group normalization enhances GAN-VAN's performance across all image quality metrics, particularly in achieving superior contrast and signal quality as evidenced by the CNR and SNR values. ↑ indicate best performance across the 4 types of inputs.

TABLE 4. Classifier performance.

(a). Results on ADNI cohort test partition, batch normalization				
Input Images	Accuracy	Precision	F1-score	MCC
Original	0.794±0.033	0.795±0.032	0.785±0.036	0.516±0.081
Diced	0.703±0.064	0.681±0.082	0.665±0.080	0.254±0.145
GAN-VAN	**0.709±0.069** ↑	**0.711±0.084** ↑	0.653±0.079	0.270±0.130
GAN-NOV	**0.709±0.042** ↑	0.697±0.062	**0.694±0.060** ↑	**0.303±0.133** ↑
(b). Results on NACC cohort, batch normalization				
Original	0.675±0.006	0.683±0.004	0.671±0.007	0.358±0.010
Diced	0.597±0.021	0.624±0.021	0.573±0.028	0.219±0.042
GAN-VAN	0.589±0.025	0.625±0.021	0.556±0.039	0.211±0.047
GAN-NOV	**0.640±0.017** ↑	**0.650±0.014** ↑	**0.634±0.019** ↑	**0.289±0.030** ↑
(a). Results on ADNI cohort test partition, group normalization				
Input Images	Accuracy	Precision	F1-score	MCC
Original	0.743±0.057	0.751±0.067	0.732±0.060	0.4117±0.1333
Diced	0.651±0.042	0.660±0.036	0.636±0.049	0.2061±0.1117
GAN-VAN	0.674±0.084	0.641±0.131	0.641±0.098	0.1844±0.2395
GAN-NOV	**0.714±0.097**	**0.719±0.118**	**0.706±0.100**	**0.3402±0.2577**
(b). Results on NACC cohort, group normalization				
Original	0.647±0.010	0.656±0.006	0.642±0.013	0.3032±0.0154
Diced	0.566±0.039	0.559±0.057	0.540±0.089	0.1284±0.0895
GAN-VAN	0.567±0.039	0.578±0.043	0.538±0.066	0.1448±0.0810
GAN-NOV	**0.608±0.051**	**0.612±0.047**	**0.599±0.063**	**0.2207±0.0982**

Performance of CNN classifiers on ADNI (a) and NACC (b). Performance values of classifiers utilizing either the original scans (**Original**), diced scans (**Diced**), or reconstructed images from **GAN-VAN** and **GAN-NOV** are shown. All CNNs were trained with the same parameters on 5 random splits of the ADNI dataset. Weighted averages for accuracy, precision, F1-score, and Matthews correlation coefficient (MCC) were calculated. The input images leading to the best performance when compared to the original scans in each metric within each cohort is highlighted in bold text. The results indicate that GAN-VAN's performance, especially with group normalization, is not consistent across the two cohorts, suggesting potential overfitting or other challenges. GAN-NOV, on the other hand, shows promise, especially in the ADNI cohort with group normalization, but its performance drops in the NACC cohort, indicating potential generalization issues. Again, ↑ indicate best performance across the 4 types of inputs.

assessment provides an intuitive understanding of the reconstruction quality, which is relevant relevant for clinical applications. In Fig. 3, we show reconstructed missing sagittal brain slices of six sample MRIs using each reconstruction method and compare them with the corresponding original scan slice. We see that GAN-NOV captures the sulci and ventricles better than GAN-VAN. Additionally, cortical thinning and boundaries between the cortex and white matter appear more clearly on the images generated using the GAN-NOV framework. Different patterns of gray matter morphometry, particularly within the temporal and parietal lobes, can be seen in persons with MCI or AD,
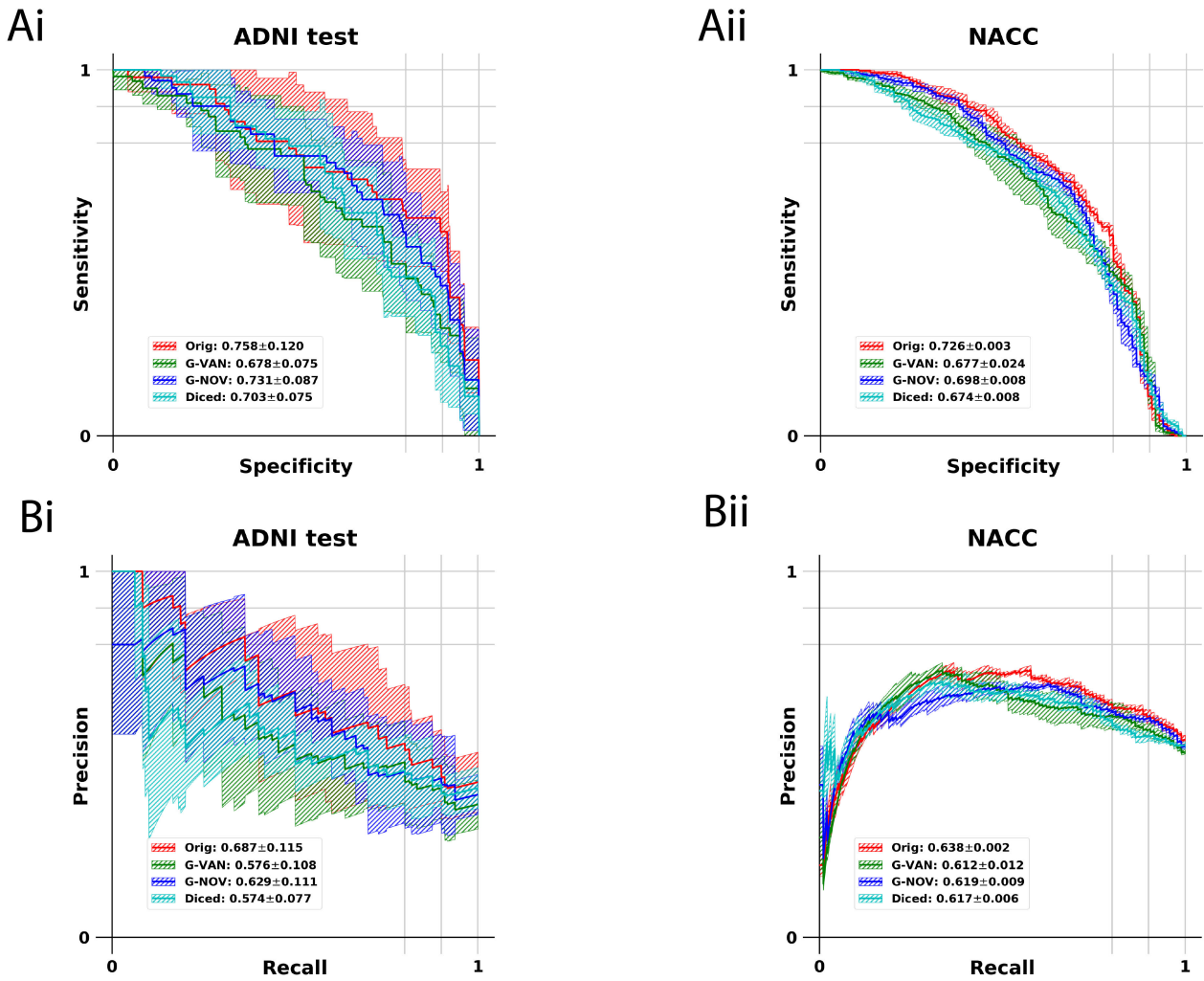
FIGURE 4. Performance of the CNN classifiers, Batch Normalization. Sensitivity-specificity (Ai and Aii) and precision-recall (Bi and Bii) curves comparing the four CNN classifiers on the ADNI test partition and NACC cohorts are shown. The classifier denoted as "Orig" was trained using original scans, "G-VAN" using generated images from GAN-VAN, "G-NOV" using generated images from GAN-NOV, and "Diced" using diced scans.

and it is therefore important to preserve these details for MCI and AD assessment [33]. Moreover, the GAN-VAN approach produces images that appear to have higher signal intensity and saturation of extreme values, whereas the images generated via the GAN-NOV framework have a more dynamic range of voxels.

B. QUANTITATIVE ASSESSMENT

Various image quality metrics, which included some no-reference algorithms, were used to objectively compare the quality of the images produced by the GAN-based reconstruction methods with the original and diced scans. Overall, the GAN-based reconstruction techniques were able to generate images with improved quality (Table 3). As expected, the diced scans demonstrated the worst image quality across all metrics tested. Moreover, the images generated via the GAN-NOV framework outperformed those generated via GAN-VAN as evaluated using SSIM, BRISQUE, and PIQE. However, generated images from the GAN-VAN framework

had improved CNR and SNR than those generated from the GAN-NOV framework. Interestingly, both GAN-VAN and GAN-NOV generated images with better BRISQUE scores than the original images (e.g., GAN-NOV, 42.505 versus Original, 44.292 on NACC).

To assess whether our framework was able to reconstruct the MR images while retaining disease-relevant information, we constructed CNN classifiers utilizing either the original scans, diced scans, or reconstructed images from GAN-VAN and GAN-NOV as input to discriminate between sMCI and pMCI cases. We trained the CNNs on 5 randomly selected folds of the ADNI dataset and performed internal testing on the corresponding ADNI test partitions and external validation on the NACC cohort. Performance of each classifier using various metrics is summarized in Table 4. ROC and PR curves for each classifier are presented in Fig. 4. As expected, the CNN classifier utilizing the original scans performed the best on both the ADNI test partition and NACC. Generated images from the GAN-NOV framework

seemed to afford better discriminative ability than the diced scans (e.g., GAN-NOV, 64% accuracy versus Diced, 59.7% accuracy on NACC). While images generated by GAN-VAN generally demonstrated the highest CNR and SNR on both ADNI and NACC (Table 3), the classifier utilizing these images tended to perform similarly to, or in some metrics worse than, the classifier utilizing the diced scans on both datasets.

Images generated by GAN-NOV performed better than those from GAN-VAN across all reported metrics on the external testing dataset (i.e., NACC). A similar trend can be seen in Fig. 4, with the classifier using GAN-VAN-generated images performing similarly to the one using diced scans (AUC 0.677, GAN-VAN versus 0.674, Diced on NACC) and the classifier using images from GAN-NOV performing better with an AUC of 0.698. Overall, these results suggest that a generative adversarial framework utilizing an additional classification loss is able to reconstruct images of better quality while maintaining disease-relevant information and also appears to be generalizable to more heterogeneous datasets.

1) CLASSIFIER PERFORMANCE

Discriminating between pMCI and sMCI is a well-known clinical challenge, as the pathological and imaging differences between these two conditions are more subtle than those between AD and normal cognition [16]. Further, predicting which individuals with MCI will progress to Alzheimer's disease becomes more difficult the longer the lead time to clinical progression, especially when trying to predict 3 or more years prior to progression [34]. In light of these challenges, others have been able to predict which persons with MCI progress to AD with varying accuracy, between 0.5-0.8, using the widely available ADNI dataset [35], [36], [37], [38], [39]. Lian et al. were able to achieve one of the highest accuracies of 81% when classifying between pMCI and sMCI on the ADNI dataset [39], similar to our baseline classifier performance using the original scans. We achieved classification accuracy comparable to published studies of 64% with the generated images from GAN-NOV when tested on a completely independent dataset (i.e., NACC). This highlights some of our framework's biggest strengths; we show GAN-NOV generates images that retain AD-relevant information essential for classification (i.e., diagnosis), and also that GAN-NOV is generalizable to unseen, heterogeneous datasets.

V. DISCUSSION

In this work, we developed and validated a multi-objective adversarial deep learning framework for MRI reconstruction and classification of persons with sMCI and pMCI using images with missing information. We extended the classical GAN architecture by incorporating a classifier to reconstruct higher quality MR images with attention to disease-related information such that the reconstructed images can be used for prognostic purposes. We systematically trained and tested

our MRI reconstruction framework using two independent datasets (ADNI & NACC), and presented the performance of our framework using well-known image quality and classification metrics.

One of the biggest challenges in identifying patients with MCI who are at high risk of progressing to AD lies in the heterogeneity of MCI presentation. For example, in the ADNI dataset, prior studies have found a significant proportion of subjects with MCI to have normal cognitive test results and no significant cortical atrophy, even though many of these subjects progressed to AD [13], [15]. Even though cortical atrophy can be subtle in a sizable proportion of pMCI patients, our competitive deep learning framework generated images with greater disease-related information than the input Diced scans, leading to improved performance in identifying persons with pMCI. That our model is able to achieve such a challenging image-based objective suggests that it may serve as a framework for development of future models targeted toward subtle but clinically important imaging tasks.

Previous work has focused on MRI reconstruction using downsampled brain scans of healthy individuals [40], [41], [42] or of patients with other neurological illnesses like multiple sclerosis [7]. Although Eo et al. proposed a deep-learning reconstruction framework in the AD population using scans from ADNI, they did not assess if the reconstructed images maintained AD-related information or validate their framework on an independent patient cohort [43]. Further, most of the published frameworks have addressed reconstruction and super-resolution in the raw MRI k-space, which limits their applicability to large, established imaging databases, such as those available for Alzheimer's disease. A final strength of our work compared to similar projects is that we mitigated the feature hallucination problem [19] inherent to generative networks to ensure accurate encoding of disease-related information. This is a necessary prerequisite to eventual use of such a framework in the clinical setting.

Our proposed framework utilizing the GAN loss, perception loss, and classification loss generated images that boosted classification accuracy from 59.7% (utilizing diced scans) to 64%. It appears that inclusion of perception loss without classification loss, like in GAN-VAN, did not yield a generalizable framework, as the classifier utilizing GAN-VAN-generated images performed worse in most metrics than the one using Diced scans (Table 4). Images generated by GAN-NOV may encode more disease-related information due to inclusion of the classification loss, thus allowing the generator network to potentially learn latent imaging features that are weakly associated with disease across MRI scans in different datasets (i.e., ADNI and NACC). These latent features may be predictive of progression from MCI to AD, but may not directly contribute to image quality. Since GAN-VAN does not incorporate classification loss, it likely cannot learn these latent features when generating images and instead focuses more on features relating to image quality.

GAN-VAN generated images with the highest SNR and CNR, likely due to the greater average pixel intensities causing oversaturation and clipping, leading to simultaneously higher mean pixel values and decreased variance. Since SNR and CNR are directly proportional to mean pixel values and inversely proportional to the variance, these measures are likely artificially inflated. In contrast, GAN-NOV generated images that demonstrated better SSIM, suggesting these images have less structural distortion when compared with the original scans than those from GAN-VAN. It is important to note image quality assessment should be done using a variety of metrics, not just SNR and CNR, especially since these latter two measures are known to be inferior metrics when comparing distorted images against corresponding references, like we do in this paper [30], [44]. For such image quality analysis, SSIM appears to be superior [30].

A. LIMITATIONS

Our study has a few limitations. We restricted the network depth in various settings due to hardware limits (GPU memory), including the number of filters (both generator and discriminator), batch size, and number of convolutional layers. We hypothesize that a larger network architecture may significantly enhance the performance of the model. However, the high-quality scans for sMCI and pMCI brains are insufficient for training a high-quality generator, which generally requires thousands of samples. Though the generator is able to obtain reasonable predictions on the NACC dataset, its broader applicability needs further verification on other datasets. It remains a challenge to maintain the balance between the different components of competitive networks even with our modified training method. For example, while the discriminator could maintain stability during the early phases of training, it may prove to be ineffective in the late phase. In other words, as the generator improves, it tends to be difficult for the discriminator to identify the differences and similarities between the generated image and the reference image, especially when the generated image belongs to the reference distribution. To alleviate this issue, potential resolutions include (1) introducing additional/alternative loss to adjust the difference during training, (2) providing additional input as extra information (i.e., spatial information) to the discriminator in later phase, (3) or using additional discriminators to create an ensemble for higher performance in later training epochs.

B. FUTURE DIRECTIONS

Sufficient spatial resolution is essential for identifying potentially minute structural brain changes that can occur in the early stages of MCI and AD. However, MRIs require long acquisition times, which can lead to patient discomfort and motion artifacts. Thus, it is essential to achieve an ideal balance between image quality and scan time. Parallel sequences can speed acquisition time, but compared to fully sampled sequences, have lower signal-to-noise-ratios [45].

Our framework could potentially be used in 2D MRI to acquire fewer images while maintaining adequate resolution of the generated slices in the setting of a specific diagnostic query. Alternatively, it could be targeted to address issues such as zipper artifacts, which appear linearly in image space [46]. It could also be extended to perform slice imputation recursively, which could enable sparser sampling or improved resolution [47]. Lastly, since we designed our framework to work with brain MRIs curated from existing imaging databases, our method could be easily extended to other imaging data banks for conditions other than AD such as vascular dementia, epilepsy, or multiple sclerosis in future research to probe other clinically relevant, subtle MRI findings.

VI. CONCLUSION

In conclusion, we propose an adversarial learning framework that takes T1-weighted MRI scans with missing slices as input and generates images with improved quality while maintaining disease-related information, assessed by discriminating between progressive MCI from stable MCI using the generated images. We trained our framework on cases from ADNI and tested it on an independent, heterogeneous dataset, NACC. Additionally, our ablation study comparing Group Normalization and Batch Normalization within this framework provided valuable insights into optimizing image quality and classifier performance, further enhancing the utility and robustness of our GAN-based approach. We note that this framework is able to meet dual objectives even when its learning ability is largely limited due to hardware and sample size. We found that the classification loss is crucial to guide the generator in producing images that retain disease-related information and to improve generalizability of the framework. Further validation of our framework, preferably using gold-standard evidence such as post-mortem neuropathology, is needed to confirm the utility of GAN-based frameworks for image reconstruction and cognitive assessment.

REFERENCES
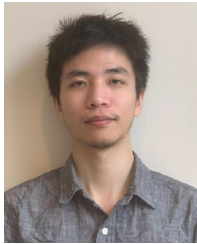
[1] X. Wang, W. Huang, L. Su, Y. Xing, F. Jessen, Y. Sun, N. Shu, and Y. Han, "Neuroimaging advances regarding subjective cognitive decline in preclinical Alzheimer's disease," *Mol. Neurodegeneration*, vol. 15, no. 1, p. 55, Dec. 2020.

[2] N. Chow, K. Hwang, S. Babakchanian, P. Thompson, C. Jack, M. Weiner, and L. Apostolova, "Comparing 3T and 1.5T MRI for mapping hippocampal atrophy in the Alzheimer's disease neuroimaging initiative (ADNI) (P06.029)," *Neurology*, vol. 80, pp. 653–660, Feb. 2013.

[3] M. S. Hansen and P. Kellman, "Image reconstruction: An overview for clinicians," *J. Magn. Reson. Imag.*, vol. 41, no. 3, pp. 573–585, Mar. 2015.

[4] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, 2nd ed. Princeton, NJ, USA: Princeton University Press, 1947.

[5] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, and D. Firmin, "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310–1321, Jun. 2018.

[6] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong, "Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1488–1497, Jun. 2018.

[7] R. Shaul, I. David, O. Shitrit, and T. Riklin Raviv, "Subsampled brain MRI reconstruction by generative adversarial neural networks," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101747.

[8] Y. Chen et al., "Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network," in *Medical Image Computing and Computer Assisted Intervention—MICCAI* (Lecture Notes of Computer Sciences). New York, NY, USA: Springer, 2018, pp. 91–99, doi: 10.1007/978-3-030-00928-1_11.

[9] K. H. Kim, W. Do, and S. Park, "Improving resolution of MR images with an adversarial network incorporating images with different contrast," *Med. Phys.*, vol. 45, no. 7, pp. 3120–3131, Jul. 2018.

[10] Z. Lai, X. Zhang, D. Guo, X. Du, Y. Yang, G. Guo, Z. Chen, and X. Qu, "Joint sparse reconstruction of multi-contrast MRI images with graph based redundant wavelet transform," *BMC Med. Imag.*, vol. 18, no. 1, pp. 1–16, Dec. 2018.

[11] L. Sun, Z. Fan, X. Fu, Y. Huang, X. Ding, and J. Paisley, "A deep information sharing network for multi-contrast compressed sensing MRI reconstruction," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6141–6153, Dec. 2019.

[12] J. E. Iglesias, B. Billot, Y. Balbastre, C. Magdamo, S. E. Arnold, S. Das, B. L. Edlow, D. C. Alexander, P. Golland, and B. Fischl, "SynthSR: A public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry," *Sci. Adv.*, vol. 9, no. 5, Feb. 2023, Art. no. eadd3607.

[13] E. C. Edmonds, L. Delano-Wood, L. R. Clark, A. J. Jak, D. A. Nation, C. R. McDonald, D. J. Libon, R. Au, D. Galasko, and D. P. Salmon, "Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors," *Alzheimer's Dementia*, vol. 11, no. 4, pp. 415–424, 2015.

[14] J. M. Ringman, W. Pope, and N. Salamon, "Insensitivity of visual assessment of hippocampal atrophy in familial Alzheimer's disease," *J. Neurol.*, vol. 257, no. 5, pp. 839–842, May 2010.

[15] E. C. Edmonds, J. Eppig, M. W. Bondi, K. M. Leyden, B. Goodwin, L. Delano-Wood, and C. R. McDonald, "Heterogeneous cortical atrophy patterns in MCI not captured by conventional diagnostic criteria," *Neurology*, vol. 87, no. 20, pp. 2108–2116, Nov. 2016.

[16] C. Wee, P. Yap, and D. Shen, "Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns," *Human Brain Mapping*, vol. 34, no. 12, pp. 3411–3425, Dec. 2013.

[17] W. Lin, Q. Gao, J. Yuan, Z. Chen, C. Feng, W. Chen, M. Du, and T. Tong, "Predicting Alzheimer's disease conversion from mild cognitive impairment using an extreme learning machine-based grading method with multimodal data," *Frontiers Aging Neurosci.*, vol. 12, p. 77, Apr. 2020.

[18] P. Coupé, S. F. Eskildsen, J. V. Manjón, V. S. Fonov, J. C. Pruessner, M. Allard, and D. L. Collins, "Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease," *NeuroImage, Clin.*, vol. 1, no. 1, pp. 141–152, 2012.

[19] J. P. Cohen, M. Luck, and S. Honari, "Distribution matching losses can hallucinate features in medical image translation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, Granada, Spain. New York, NY, USA: Springer, 2018, pp. 529–536.

[20] H. Iqbal, "HarisIqbal88/PlotNeuralNet v1.0.0," Zenodo, Switzerland, Dec. 2018, doi: 10.5281/zenodo.2526396.

[21] R. C. Petersen, P. Aisen, L. A. Beckett, M. Donohue, A. Gamst, D. J. Harvey, C. Jack, W. Jagust, L. Shaw, A. Toga, and J. Q. Trojanowski, "Alzheimer's disease neuroimaging initiative (ADNI): Clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.

[22] J. R. Mitchell, K. Kamnitsas, K. W. Singleton, S. A. Whitmire, K. R. Clark-Swanson, S. Ranjbar, C. R. Rickertsen, S. K. Johnston, K. M. Egan, and D. E. Rollison, "Deep neural network to locate and segment brain tumors outperformed the expert technicians who created the training data," *J. Med. Imag.*, vol. 7, no. 5, Oct. 2020, Art. no. 055501.

[23] I. B. Malone, K. K. Leung, S. Clegg, J. Barnes, J. L. Whitwell, J. Ashburner, N. C. Fox, and G. R. Ridgway, "Accurate automatic estimation of total intracranial volume: A nuisance variable with less nuisance," *NeuroImage*, vol. 104, pp. 366–372, Jan. 2015.

[24] M. F. Romano, X. Zhou, A. R. Balachandra, M. F. Jadick, S. Qiu, D. A. Nijhawan, P. S. Joshi, S. Mohammad, P. H. Lee, M. J. Smith, A. B. Paul, A. Z. Mian, J. E. Small, S. P. Chin, R. Au, and V. B. Kolachalama, "Deep learning for risk-based stratification of cognitively impaired individuals," *iScience*, vol. 26, no. 9, Sep. 2023, Art. no. 107522.

[25] V. Anand, S. Gupta, A. Altameem, S. R. Nayak, R. C. Poonia, and A. K. J. Saudagar, "An enhanced transfer learning based classification for diagnosis of skin cancer," *Diagnostics*, vol. 12, no. 7, p. 1628, Jul. 2022.

[26] S. Qiu, P. S. Joshi, M. I. Miller, C. Xue, X. Zhou, C. Karjadi, G. H. Chang, A. S. Joshi, B. Dwyer, and S. Zhu, "Development and validation of an interpretable deep learning framework for Alzheimer's disease classification," *Brain*, vol. 143, no. 6, pp. 1920–1933, 2020.

[27] X. Zhou, S. Qiu, P. S. Joshi, C. Xue, R. J. Killiany, A. Z. Mian, S. P. Chin, R. Au, and V. B. Kolachalama, "Enhancing magnetic resonance imaging-driven Alzheimer's disease classification performance using generative adversarial learning," *Alzheimer's Res. Therapy*, vol. 13, no. 1, pp. 1–11, Dec. 2021.

[28] L. Biewald, "Experiment tracking with weights and biases," Software, 2020, vol. 2, no. 5.

[29] J. Snoek, H. Larochelle, and R.-P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proc. NIPS*, 2012, pp. 2951–2959.

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[31] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[32] N. Venkatanath, D. Praneeth, B. H. Chandrasekhar, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Proc. 21st Nat. Conf. Commun. (NCC)*, Feb. 2015, pp. 1–6.

[33] Z. Wu, Y. Peng, M. Hong, and Y. Zhang, "Gray matter deterioration pattern during Alzheimer's disease progression: A regions-of-interest based surface morphometry study," *Frontiers Aging Neurosci.*, vol. 13, Feb. 2021, Art. no. 593898.

[34] S. F. Eskildsen, P. Coupé, D. García-Lorenzo, V. Fonov, J. C. Pruessner, and D. L. Collins, "Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning," *NeuroImage*, vol. 65, pp. 511–521, Jan. 2013.

[35] C. Salvatore, A. Cerasa, P. Battista, M. C. Gilardi, A. Quattrone, and I. Castiglioni, "Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: A machine learning approach," *Frontiers Neurosci.*, vol. 9, p. 307, Sep. 2015.

[36] J. Koikkalainen, J. Lötjönen, L. Thurfjell, D. Rueckert, G. Waldemar, and H. Soininen, "Multi-template tensor-based morphometry: Application to analysis of Alzheimer's disease," *NeuroImage*, vol. 56, no. 3, pp. 1134–1144, Jun. 2011.

[37] M. Liu, D. Zhang, and D. Shen, "Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment," *IEEE Trans. Med. Imag.*, vol. 35, no. 6, pp. 1463–1474, Jun. 2016.

[38] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova, "Residual and plain convolutional neural networks for 3D brain MRI classification," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 835–838.

[39] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 880–893, Apr. 2020.

[40] Y. Huang, F. Zheng, D. Wang, J. Jiang, X. Wang, and L. Shao, "Super-resolution and inpainting with degraded and upgraded generative adversarial networks," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 645–651.

[41] X. Zhao, T. Yang, B. Li, and X. Zhang, "SwinGAN: A dual-domain Swin transformer-based generative adversarial network for MRI reconstruction," *Comput. Biol. Med.*, vol. 153, Feb. 2023, Art. no. 106513.

[42] H. Wei, Z. Li, S. Wang, and R. Li, "Undersampled multi-contrast MRI reconstruction based on double-domain generative adversarial network," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 9, pp. 4371–4377, Sep. 2022.

[43] T. Eo, Y. Jun, T. Kim, J. Jang, H. Lee, and D. Hwang, "KIKI-net: Cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images," *Magn. Reson. Med.*, vol. 80, no. 5, pp. 2188–2201, Nov. 2018.

[44] L. S. Chow and R. Paramesran, "Review of medical image quality assessment," *Biomed. Signal Process. Control*, vol. 27, pp. 145–154, May 2016.

[45] J. Hamilton, D. Franson, and N. Seiberlich, "Recent advances in parallel imaging for MRI," *Prog. Nucl. Magn. Reson. Spectrosc.*, vol. 101, pp. 71–95, Aug. 2017.

[46] M. Hakky, S. Pandey, E. Kwak, H. Jara, and S. H. Erbay, "Application of basic physics principles to clinical neuroradiology: Differentiating artifacts from true pathology on MRI," *Amer. J. Roentgenology*, vol. 201, no. 2, pp. 369–377, Aug. 2013.

[47] K. Zhang, H. Hu, K. Philbrick, G. M. Conte, J. D. Sobek, P. Rouzrokh, and B. J. Erickson, "SOUP-GAN: Super-resolution MRI using generative adversarial networks," *Tomography*, vol. 8, no. 2, pp. 905–919, Mar. 2022.

SANG PETER CHIN is currently a Professor of engineering with Dartmouth College and the Director of the LISP—Learning, Intelligence + Signal Processing—Laboratory. He and his students are investigating fundamental questions such as, "Can Intelligence be Learned?" at the intersection of signal processing, machine learning, game theory, differential geometry, extremal graph theory, and computational neuroscience. He has served as the Co-Chair for the annual SPIE/DSS Conference on Cyber Sensing and the Symposium Chair for the GlobalSIP Conference. He is currently an Associate Editor of IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS.

XIAO ZHOU received the Ph.D. degree in computer science from Boston University, advised by Prof. Vijaya B. Kolachalama and Prof. Peter Chin. He is currently a Postdoctoral Associate with the Department of Molecular Biophysics and Biochemistry, Yale University, who specializes in integrating deep learning techniques for genomic and neurological data interpretation. His research intersects adversarial learning with Alzheimier's disease data. He has contributed to the field through publications in different journals and conferences, focusing on applications ranging from Alzheimer's disease classification to the enhancement of MRI images. His current research interests include deep learning, bioinformatics, and computational genomics.

RHODA AU is currently a Professor of anatomy and neurobiology, neurology, medicine, and epidemiology with Boston University Chobanian and Avedisian School of Medicine and the School of Public Health. She is one of PIs of the Framingham Heart Study Brain Aging Program and the Director of neuropsychology. Her work includes using technologies to promote equal opportunity science and to develop and validate multi-sensor digital biomarkers. Her long-term research objective is to enable global solutions that move the primary focus of health technologies from precision medicine to a broader emphasis on precision brain health.

AKSHARA R. BALACHANDRA received the M.S. degree from the University of California at San Diego, and the M.D. degree from Boston University Chobanian and Avedesian School of Medicine. He is currently a Resident Physician with the Department of Neurology and Neurological Sciences, Stanford University. His research interest includes the application of machine learning to neuroimaging to improve prognostication of neurologic diseases.

VIJAYA B. KOLACHALAMA (Senior Member, IEEE) is currently an Associate Professor with the Department of Medicine and the Department of Computer Science, Boston University (BU), where he is also one of the Founding Members of the Faculty of Computing and Data Sciences. His laboratory with BU is broadly interested in biomedical machine vision, representation learning, and domain generalization, with applications in neurology and cancer. His research is supported by grants from the American Heart Association, the National Institutes of Health (NIA, NHLBI, NCI, and NIDDK), the Karen Toffler Charitable Trust, Gates Ventures, Artificial Intelligence and Technology Collaboratories with NIA, and Johnson & Johnson Enterprise Innovation. He is a fellow of the American Heart Association.

MICHAEL F. ROMANO received the M.D. and Ph.D. degrees from Boston University. He is currently a Radiology Resident with the Department of Radiology and Biomedical Imaging, University of California, San Francisco, where he is also a member of the Center for Intelligent Imaging. His research interests include the clinical application of novel machine learning techniques to improve the diagnostic capabilities of MRI and CT, with a particular focus on neuroimaging.

• • •