Accelerated MRI using intelligent protocolling and subject-specific denoising applied to Alzheimer’s disease imaging

Keerthi Sravan Ravi\textsuperscript{1,2}, Gautham Nandakumar\textsuperscript{3}, Nikita Thomas\textsuperscript{3}, Mason Lim\textsuperscript{3}, Enlin Qian\textsuperscript{1,2}, Marina Manso Jimeno\textsuperscript{1,2}, Pavan Poojar\textsuperscript{4}, Zhezhen Jin\textsuperscript{5}, Patrick Quarterman\textsuperscript{6}, Girish Srinivasan\textsuperscript{3}, Maggie Fung\textsuperscript{6}, John Thomas Vaughan Jr.\textsuperscript{1,2}, Sairam Geethanath\textsuperscript{4,2} for the Alzheimer’s Disease Neuroimaging Initiative*.

\textsuperscript{1} Department of Biomedical Engineering, Columbia University in the City of New York, NY, USA
\textsuperscript{2} Columbia University Magnetic Resonance Research Center, Columbia University in the City of New York, New York, NY, USA
\textsuperscript{3} PhenoMX, Chicago, IL, USA
\textsuperscript{4} Accessible MRI Laboratory, Biomedical Engineering and Imaging Institute, Dept. of Diagnostic, Molecular and Interventional Radiology, Icahn School of Medicine at Mt. Sinai, New York, NY
\textsuperscript{5} Mailman School of Public Health, Columbia University in the City of New York, New York, NY, USA
\textsuperscript{6} MR Clinical Solutions, GE Healthcare, New York, NY, USA

**Keywords:** Autonomous MRI, deep learning, explainable AI, multi-contrast denoising, MR Value

**Abstract:** Magnetic Resonance Imaging (MRI) is expensive and time-consuming. Protocol optimization to accelerate MRI requires local expertise since each MR sequence involves multiple configurable parameters that need optimization for contrast, acquisition time, and signal-to-noise ratio (SNR). The availability and access to technical training are limited in underserved regions, resulting in a scarcity of local expertise required to operate the hardware and perform MR examinations. Along with other cultural and temporal constraints, these factors contribute to the highly inefficient utilization of MRI services diminishing their clinical value. In this work, we extend our previous effort and demonstrate accelerated MRI via intelligent protocolling of the modified brain screen protocol, referred to as the Gold Standard (GS) protocol. We leverage deep learning-based contrast-specific image-denoising to improve the image quality of data acquired using the accelerated protocol. Since the SNR of MR acquisitions depends on the volume of the object being imaged, we demonstrate subject-specific (SS) image-denoising. Utilizing the accelerated protocol resulted in a 1.94x gain in imaging throughput over

\*NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
the GS protocol. The minimum /maximum PSNR gains (measured in dB) were 1.18/11.68 and 1.04/13.15, from the baseline and SS image-denoising models, respectively.

Alzheimer’s Disease (AD) accounts for up to 60-80% of dementia cases and a global trend of longer lifespans has resulted in an increase in the prevalence of dementia/AD. Therefore, an accurate differential diagnosis of AD is crucial to determine the right course of treatment. The GS protocol constitutes 44.44% of the comprehensive AD imaging protocol defined by the European Prevention of Alzheimer’s Disease project. Therefore, we also demonstrate the potential for AD-imaging via automated volumetry of relevant brain anatomies whose atrophies have been shown to be reliable indicators of the onset of the disease. The volumetric measurements of the hippocampus and amygdala from the GS and accelerated protocols were in excellent agreement, as measured by the intra-class correlation coefficient.

In conclusion, accelerated brain imaging with the potential for AD imaging was demonstrated, and image quality was recovered post-acquisition using DL-based image denoising models.

I. **Introduction**

Although Magnetic Resonance Imaging (MRI) is a powerful imaging modality to obtain valuable information about the brain structure anatomy via the acquisition of high-resolution images, multiple challenges are involved with MR imaging. It is expensive and time-consuming, and subjects with MR-unsafe materials (such as medical device implants, prostheses, etc.) are not eligible for MR imaging. Considerable research efforts have been directed toward accelerating acquisition times by exploiting the temporal or spatiotemporal redundancies in the images (Tsao and Kozerke, 2012). However, protocol optimization to accelerate MRI requires local MR expertise. Each sequence involves multiple configurable parameters that need optimization for contrast, acquisition time, and signal-to-noise ratio (SNR). A large number of these combinations exist—for example, 29 million for 12 sequences in a protocol (Block, 2018)—and choosing an optimal combination in real-time is difficult. Since the availability and access to technical training are limited in under-served regions (Geethanath and Vaughan, 2019), this results in a scarcity of local expertise required to operate MR hardware and perform MR examinations. These factors, along with other cultural and temporal constraints contribute to the highly
inefficient utilization of MRI services diminishing their clinical value (Geethanath and Vaughan, 2019).

This combination of a very high-dimensional optimization space and inadequate local expertise necessitates a data-driven approach to augment the available manpower. Previous works involve machine learning approaches for automated RF pulse design (Shin et al., 2020), sequence design (Zhu et al., 2018), or even a joint framework for sequence generation and data reconstruction (Walker-Samuel, 2019; Loktyushin et al., 2021). We believe that augmenting human expertise by leveraging deep learning (DL) techniques across the MRI pipeline can consistently yield improved MR Value irrespective of where the service is offered or the expertise involved. MR Value is an initiative by the International Society of Magnetic Resonance in Medicine to measure the utility of MRI in the context of constantly evolving healthcare economics (https://www.ismrm.org/the-mr-value-initiative-phase-1/). We based our prior work on this premise and demonstrated preliminary results from MR Value-driven Autonomous MRI, dubbed AMRI (Ravi et al.; Ravi and Geethanath 2020).

Dementia affected 50 million people worldwide in 2018, with an estimated economic impact of US$ 1 trillion a year (Patterson, 2018; Banerjee et al., 2020). Alzheimer’s Disease (AD) accounts for up to 60-80% of dementia cases and potentially begins up to 20 years before the first symptoms emerge (Bateman et al., 2012). A global trend of longer lifespans has resulted in an increase in the prevalence of dementia/AD (Silva-Spínola et al., 2022). An accurate differential diagnosis of AD is crucial to determine the right course of treatment (Vernooij and van Buchem, 2020). Traditionally, structural neuroimaging is used to exclude treatable and reversible causes of dementia such as brain tumors, subdural hematomas, cerebral infarcts, or hemorrhages (Falahati et al., 2015). However, studies have demonstrated that atrophy of the hippocampus and amygdala volume are reliable indicators of progression from pre-dementia to AD (Simmons et al., 2011). These imaging biomarkers, or image-derived phenotypes (IDP), can be obtained from MRI. The European Prevention of Alzheimer’s Disease (EPAD, https://ep-ad.org/open-access-data/overview) prescribes four core and five advanced sequences for AD imaging. The core sequences are 3D T1-weighted (T1w), 3D fluid-attenuated inversion recovery (FLAIR), 2D T2-weighted (T2w), and 2D T2*-weighted (T2*w). The advanced sequences are 3D T2*w, 3D susceptibility weighted imaging (SWI), diffusion-weighted imaging (DWI) or dMRI, resting-state functional MRI, and arterial spin labeling. Furthermore, (Meahan et al., 2014)
reported on the adequacy of a four-sequence protocol consisting of an axial T₁w, axial T₂w FLAIR, axial DWI, and axial SWI images to evaluate new patients with neurological complaints.

In this work, we extend our previous effort and demonstrate accelerated MRI via intelligent protocolling of the modified brain screen protocol (dubbed Gold Standard, GS) employed at our institution. We leverage deep learning-based image denoising to improve the image quality of data acquired using the accelerated protocol. The GS protocol consisted of the following sequences: sagittal 3D T₁w, axial 2D T₂w, axial 2D T₂w FLAIR, axial 2D SWI, axial DWI, and axial 2D T₁w. Overall, the GS protocol constitutes 44.44% of the comprehensive EPAD imaging protocol and includes all sequences deemed adequate for neuroimaging by Mehan et al. Therefore, we also investigate the potential of the accelerated protocol for AD-screening by benchmarking volumetry of the hippocampus and amygdala against measurements from the GS protocol. This volumetry can be used for early detection of atrophy.

**A. Study design**

Figure 1 is a brief illustration of the processes involved in this work: intelligent protocolling, MR acquisition, DL-based image denoising, and quantitative evaluation. We first demonstrate accelerating the GS protocol employed at our institution using sequence-specific look-up tables (LUTs, Figure 1A). We use LUTs to optimize for accelerated acquisition durations. We perform four experiments designed to measure improvements in metrics such as (i) median object-masked local SNR (Golshan, Hasanzadeh and Yousefzadeh, 2013); (ii) peak SNR (measured in dB); (iii) variance of the Laplacian, referred to as var-Lap (Pech-Pacheco et al., 2000); (iv) multi-scale structural similarity index (MS-SSIM, Wang, Simoncelli and Bovik, 2003); and (v) MR Value. We also acquire data from five healthy subjects (Figure 1B) to test for repeatability. Subsequently, we denoise the acquired data using contrast-specific image-denoising models trained on publicly available brain MRI datasets (Figure 1C). Since the SNR of MR acquisitions depends on the volume of the object being imaged, the SNR values of acquisitions of larger subjects will be higher. We obtain preliminary results that illustrate the variability of SNR with varying whole brain volumes and base our rationale for subject-specific (SS) image denoising on this premise. Subsequently, we demonstrate SS-denoising (Figure 1D). Finally, we perform quantitative evaluation using a combination of automated and manual methods.
To the best of our knowledge, this work, dubbed Intelligent Protocolling for AMRI (AMRI-IP), is the first framework to integrate the entire pipeline of AD imaging with a particular focus on accelerating MR acquisitions. AMRI-IP provides end-to-end control of acquisition, reconstruction, image processing, and analysis. The flexible and extensible LUT approach can be utilized to accelerate a single sequence, or an entire protocol, with complete control over the optimization criterion. The contrast-specific image-denoising models are inherently generalizable by training on corresponding publicly available datasets. The SS image-denoising models enable tailored image quality recovery for each subject. Finally, the automated quantitative evaluation method enables the measurement of AD-specific IDPs such as hippocampal and amygdala volumes.

In brief, we claim that employing the accelerated protocol yields improvements in MR Value, and achieves comparable image quality to GS as per quantitative evaluation. Section II.D. maps the four experiments to the claims of this work.

II. Methods and materials

This section is organized as follows. Section II.A. describes intelligent protocolling – accelerating the routine brain screen protocol employed at our institution by consulting a LUT. Section II.B. presents the development of deep learning models to achieve SS denoising and the explainability of the models. The four experiments that were performed to investigate the hypotheses are detailed in Section II.C., and the image-derived phenotypes (IDPs) relevant to AD that were used in performance evaluation are described in Section II.D. Section II.E. discusses the quantitative evaluation methods performed to validate the hypotheses, and finally, Section II.F. describes visualizing intermediate filter outputs for explainable AI.

A. Intelligent protocolling using look-up tables

Table 1 lists the seven GS sequences and their corresponding acquisition parameters and durations. The cumulative acquisition time was 17:23 (minutes:seconds), as per the vendor console’s user interface (UI). An experienced clinical application specialist was consulted to collate a list of acquisition parameters that could be varied without compromising image contrast for each sequence in the GS protocol. These acquisition parameters were referred to as degrees of freedom (DOF), also listed in Table 1. Exhaustive combinations of these DOF or a hundred
randomly chosen combinations, whichever was smaller, were entered into the vendor console’s UI. For each combination \((P_{acq})\), the acquisition time \((t_{acq})\) and relative signal-to-noise ratio \((r\text{SNR})\) value were recorded. The \(P_{acq}\) and corresponding \(t_{acq}\) and \(r\text{SNR}\) values were stored in a LUT. These were searched to obtain the optimal \(P_{acq}\) yielding the lowest \(t_{acq}\). This procedure was repeated for each sequence in the GS protocol. Once the sequence-specific LUTs were constructed, they were consulted to derive sequence-specific optimal \(P_{acq}\) to derive the Fastest protocol. The search procedure is described as follows, applicable to each sequence individually:

1. **Compute percentage time allocated:** First, the minimum time percentage value \((y_1)\) was computed as the ratio of the shortest sequence acquisition time to the shortest protocol acquisition time \((x_1)\). Similarly, the maximum time percentage value \((y_2)\) was computed from the longest protocol acquisition time \((x_2)\). Now, for an imposed protocol acquisition time constraint \((T_{acq})\), the percentage time allocated \((%\text{TA})\) to a sequence was derived from the straight line fitting the minimum and maximum time percentage values, as described by equation 1:

\[
\text{Percentage time allocated (}\%\text{TA}) = \frac{(y_2 - y_1)}{(x_2 - x_1)} \cdot (x - x_1) + y_1 \quad \ldots (1)
\]

2. **Compute weighted rank:** The time allocated in seconds for this sequence \((t_{acq})\) was derived from the \(%\text{TA}\) value. The LUT was filtered by discarding DOF combinations whose acquisition times exceeded \(t_{acq}\). Of the remaining combinations, weighted differences of \(r\text{SNR}\) and DOF values with the corresponding default values from the GS protocol were computed. Higher weights were assigned to DOF values contributing more significantly to the image contrast (Supplementary Table 1). Finally, these weighted differences were summed to obtain a rank for each DOF combination, and the resulting LUT was sorted in ascending order of this rank value. Thus, the combination achieving the lowest rank value had the smallest difference in those DOF values which most significantly contributed to the image contrast.

3. **Obtain optimal combination:** For each time constraint, the combination with the lowest rank was chosen as the optimal set of acquisition parameters.

This process was repeated with lower imposed \(T_{acq}\) in each iteration until an optimal \(P_{acq}\) could not be obtained for every sequence in the GS protocol. In this way, the Fastest protocol was derived by consulting the sequence-specific LUTs. Data acquired utilizing the GS protocol were referred to as the GS dataset. Data acquired from the Fastest protocol for Experiments 1, 2 and 4 (see Section II.D.) were referred to as the Fastest dataset, collectively. The SWI sequence in the GS protocol was replaced by a T\(_2^*\)w sequence in the Fastest protocol. The experienced
clinical application specialist’s express protocol was dubbed the Expert Express (EE) protocol. Data was also acquired from this protocol for comparison (refer Experiment 2 in Section II.D.), referred to as the EE dataset.

**B. Image denoising using deep learning**

Two popular image denoising approaches are to directly predict the denoised image or to obtain the denoised image as the residual of the input noisy image and the predicted noise. We adopt a ResNet-inspired network architecture (demonstrated to improve training performance and stability, (He *et al.*, 2016)) to directly predict the denoised output. Individual contrast-specific denoising models were trained on pairs of noisy-denoised images from publicly available brain MRI datasets (see below). Finally, SS denoising was performed by fine-tuning the models on pairs of noisy-denoised images from the prospectively acquired Fastest dataset.

**Datasets, forward simulation, and data splits**: Publicly available datasets were utilized to train the contrast-specific image denoising models. T₁ and T₂ contrasts: IXI dataset ([https://brain-development.org/ixi-dataset/](https://brain-development.org/ixi-dataset/)); T₂*: ADNI 3 ([https://adni.loni.usc.edu](https://adni.loni.usc.edu)); T₂ FLAIR: MSSEG-2 (Commowick, Cervenansky and Cotton, 2021) and DWI: AOMICID-1000 (Snoek *et al.*, 2021). Wherever applicable, datasets were filtered to retain only the 3T data. Only the central 50% slices were utilized, and the remaining slices were discarded to avoid either unwanted anatomy or pure background noise. Supplementary Figure 1 presents the search criteria that were utilized to filter the ADNI 3 dataset for relevant results. For DWI, only the b0 images were utilized from the AOMICID-1000 dataset. All these datasets were assumed to be free of MR image artifacts, referred to as ‘clean images’. Figure 2 presents the forward modeling process to generate noise-corrupted data (‘noisy images’), described here as follows. First, the object-masked local SNR maps were computed on all acquisitions of an arbitrarily chosen subject from the Fastest dataset (see Experiment 4, Section II. D.). Object masking was based on the technique in (Jenkinson, 2003) and resulted in all background values being set to zero. All remaining non-zero values were considered to belong to the samples, which were non-skull-stripped brain images. The local SNR maps were computed based on the method reported in Golshan, Hasanzadeh and Yousefzadeh, 2013. The volume yielding the lowest median SNR was the ‘noisiest acquisition’ (Figure 2A). Next, motivated by work in (Geethanath, Poojar and Ravi, 2021; Qian *et al.*, 2022), native noise values were extracted from this noisiest acquisition. These
noise values were collaged to form a native noise block (Figure 2B). This was randomly sampled to obtain noise values, which were scaled and added to the clean images to obtain noisy images (Figure 2C). The scaling factor was determined using an iterative approach. A volume was chosen at random from the public dataset. Starting with an initial value of 1.0 (corresponding to no scaling), the scaling factor was increased by 0.5 in each iteration until the median object-masked local SNR of the corrupted volume was lesser than that of the noisiest acquisition. We chose median over mean as the guiding measure since it was less affected by skewed distributions. An 85%-10%-5% subject-wise split was performed to form the train, validation, and test sets.

**Network architectures:** Figure 3A is an illustration of the network architecture common to all contrast-specific image-denoising models. To predispose the network to learn denoising filters whilst being anatomy agnostic, we adopted a patch-wise approach in this work. Overlapping patches of size $64 \times 64$ were input to the network. Thirteen ResBlocks leveraged skip connections to improve training performance (He et al., 2016). Each ResBlock consisted of two ReLU-activated (Nair and Hinton, 2010) $3 \times 3$ 2D convolution layers. In case of a mismatch in the number of filters between the previous and current ResBlocks ($N_1, N_2$), the skip connection included a $1 \times 1$ 2D convolution with $N_2$ filters. Otherwise, the skip connection was an identity operation. Additionally, an identity skip connection was used to add the input data to the pre-final layer in the overall network. The final layer was a $3 \times 3$ 2D convolution layer with 1 filter.

All 2D convolution layers were ReLU-activated, and all development, training, and testing were performed using Keras 2.6/TensorFlow 2.6.2 (Abadi et al., 2016; M. Abadi et al., 2016) libraries.

**Loss functions:** Zhang et al. report the superiority of their mixed loss (Mix-L) function for image denoising, among other image quality restoration applications (Zhao et al., 2017). This loss is a weighted sum of $l_1$ and MS-SSIM losses:

$$Mix-L = \alpha (1 - MS-SSIM) + (1 - \alpha) l_1 \quad \ldots(2)$$

where $\alpha$ was set to 0.84. We modified Mix-L to incorporate a data-consistency term with the measured data in the Fourier domain, referred to as Mix-L+FTD:

$$f = M \odot |F(y_{pred}) - F(y_{true})| \quad \ldots(3)$$

$$Mix-L + FTD = Mix-L + \beta \left\| \frac{f}{||f||_2} \right\|_2 \quad \ldots(4)$$
... where $F$ was the 2D Fourier Transform and $M$ was a mask to only retain the central crop of the k-space of size $16 \times 16$. The $\odot$ operator represented the Hadamard product and $\beta$ determined the trade-off between the denoising and data-consistency errors. We investigated $\beta = [0, 0.01, 0.1, 1]$ in our experiments. To determine the best $\beta$, the RMSEs of the volumetric measures ($\text{RMSE}_{\text{vol}}$) were computed using an automated tool on T1 denoised outputs. The $\beta$ yielding the lowest mean $\text{RMSE}_{\text{vol}}$ was chosen to train the denoising models for the remaining contrasts. Section II.C. describes the automated T1 volumetry tool and computing $\text{RMSE}_{\text{vol}}$ in detail.

**Training:** All contrast-specific denoising models were trained for 100 epochs with a batch size of 256. The Adam optimizer (Kingma and Ba, 2014) was utilized to minimize the Mix-L+FTD loss with the optimal $\beta$, determined as stated above. During training, every input slice was cropped to a $64 \times 64$ patch. The bounds for the random crop were manually determined by examining the corresponding public dataset such that the random crops would mostly include brain anatomy. A callback was utilized to save the model achieving the lowest validation loss (corresponding to ‘best performance’). At the end of the training process, this model was chosen as the best model for evaluation, including to determine the optimal $\beta$.

**Subject-specific (SS) denoising:** SS median local SNRs were computed on masked brain volumes from the Fastest dataset to verify the premise of SS denoising. The values were computed only on the central 50% of the slices. Next, the same noise scaling factors were utilized to corrupt each subject’s noisiest acquisition from the Fastest dataset with native noise. This data was used to fine-tune the baseline denoising models to achieve SS denoising. This approach posed SS denoising as a self-supervised learning problem, mimicking the noisy-as-clean method demonstrated by (Xu et al., 2020). The initial learning rate of the Adam optimizer was reduced to avoid large modifications to the weights which would otherwise harm the learned representations (Table 2).

**C. Experiments**

We performed four experiments to investigate hypotheses regarding the throughput of the Fastest protocol, and the image quality of the Fastest dataset. Supplementary Table 2 lists the experiments performed, the protocols executed, numbers of healthy volunteers imaged, the corresponding claims and hypotheses investigated, and their respective evaluation criteria. In
total, 31 brain volumes were acquired from five volunteers across the four experiments. The data acquired from the GS and Fastest protocols are referred to as the GS and Fastest datasets, respectively.

**Experiment 1 (E1) – Throughput:** The goal of experiment one was to investigate if the Fastest protocol obtained from the LUT would yield an improvement in throughput. One volunteer was imaged using the GS and Fastest protocols, and a video recording of the entire imaging session was captured. Throughput was computed as the ratio of the table time measurement of the Fastest protocol to that of the GS protocol. In this work, table time is defined as the duration between the scanner bed reaching the center of the bore at the start of the imaging and the scanner bed returning to the home position. We also determined the MR Values of the GS and Fastest protocols, calculated as the ratio of the cumulative median object-masked local SNR values across all contrasts to the protocol’s acquisition duration, $T_{acq}$:

$$
MR Value = \frac{\sum_{c=1}^{Contrasts} median \_ object\_ masked \_ local \_ SNR_c}{T_{acq}}
$$

The object-masked local SNR maps were computed on the central 50% slices across all sequences in each protocol.

**E2 – Image quality:** Experiment two quantitatively compared the image quality of the GS, EE, and Fastest datasets using the following metrics: median object-masked local SNR and var-Lap. PSNR and SSIM were not used since they are not reference-less metrics.

**E3 – SNR recovery:** We investigated the feasibility of employing the Fastest protocol. It involved utilizing the image-denoising deep learning models described in Section II.B. to improve image quality. The metrics described in Section II.D. were utilized to determine if denoising the Fastest dataset achieved comparable quality to the GS dataset.

**E4 – Repeatability:** A repeatability test to demonstrate the consistency of the quantitative image quality metrics was performed. The GS and Fastest protocols were employed to acquire data from five subjects over five repeats. Automated and manual volumetry were performed on the acquired data as described in Section II.D.

**D. Quantitative evaluation**

Thomas et al., 2020 demonstrated an end-to-end pipeline for fully automated mental health screening. The authors leveraged a DL model to segment the various subgroups. Further
development on the previous work includes a second DL model to segment the brain tissues (white matter, gray matter, cerebrospinal fluid). This second DL model was based on the nnUnet \cite{Isensee2021}, and an evaluation of its performance is presented in Supplementary File 1.

We leveraged this tool to perform automated volumetry to measure the performance of the denoising models. HTML reports were generated containing volumetric measures of 27 brain subregions and 3 brain tissues. These were programmaticallly extracted and tabulated. RMSE$_{\text{vol}}$ was calculated as the mean of RMSEs of each of the volumetric measures. A benign White Matter Hyperintensity (WMH) was identified in data acquired from one subject. The free, open-source, and multi-platform 3D Slicer software \cite{Fedorov2012} was used to perform manual volumetry of this WMH on the T$_2$, T$_2^*$, T$_2$ FLAIR and DWI data by four different raters with 3-8 years of MR experience. All volumetries were performed on data acquired for E4 of the Fastest dataset.

Additionally, a set of image quality metrics were also computed to evaluate the denoising models. These were: median object-masked local SNR, PSNR, MS-SSIM, var-Lap and MR Value. While local SNR, PSNR, and MS-SSIM metrics are commonly used to measure image quality, we obtain var-Lap values to measure the amount of blurring. We included this metric in our evaluations since blurring negatively affected the automated volumetry on T$_1$ (preliminary experiments not reported in this work).

**E. Visualizing learned filters for explainable AI**

A $256 \times 256$ collage of four panels was assembled to investigate the features learned by the filters of the 2D convolution layers. Each of the four $64 \times 64$ panels was made up of a single intensity from the following values: $[1.25, 3.75, 6.25, 8.75] \times 10^{-1}$. This collage was corrupted with native noise from an arbitrarily chosen subject’s T$_1$w acquisition. Subsequently, it was denoised using the baseline denoising model for the T$_1$ contrast. The filter outputs of each 2D convolutional layer were obtained, and maximum intensity projection was performed to achieve dimensionality reduction. Therefore, this collapsed $N$ filter outputs into a single map. This was normalized to lie in the range $[0, 1.0]$. Finally, this collapsed feature map was hard thresholded to only retain values $> 0.75$. Figure 3B briefly illustrates this procedure.
F. Statistical analysis

The intra-class correlation coefficient (ICC) was calculated based on the analysis of variance (ANOVA) with repeated measures to assess the agreement of the volumetric measures amongst the GS, denoised baseline, and denoised SS methods. The ICC values greater than 0.9 indicate excellent agreement, values between 0.75 and 0.9 indicate good agreement, and values between 0.5 and 0.75 indicate moderate agreement.

III. Results

A. Intelligent protocolling and LUTs

rSNR was assigned the smallest weight when constructing the LUT since the aim of this work was to achieve acceleration by trading-off SNR which could be recovered post-acquisition via deep learning methods. The total duration of the Fastest protocol that was obtained by querying the LUTs was 8:34 (minutes:seconds). This was a 50.71% reduction in acquisition time from the GS protocol, which required 17:23. The EE protocol only required 7:12. It primarily achieved acceleration by employing the 3D T1 FLAIR sequence. This is not a true 3D acquisition and only required 0:37 when compared to 2:44 and 1:41 for the 3D T1w sequences in the GS and Fastest protocols, respectively. Figure 4 is a collage of one representative slice of an arbitrarily chosen subject, across contrasts. The rows represent the different datasets–GS, EE, Fastest, baseline denoised, and SS denoised. For the Fastest protocol, the sagittal T1-MPRAGE sequence was accelerated by increasing the slice-thickness from 1.0 mm in the GS protocol to 1.6 mm. Overall, this resulted in increased signal intensities and decreased variance of noise in the Fastest dataset. Therefore, the median local SNR of GS data was lower than that of the Fastest data, as is expected.

B. Image denoising using deep learning

Datasets and forward-simulation: The T1w and T2w datasets consisted of 185/13,690 and 185/11,760 volumes/slices, respectively. For T2* and T2 FLAIR, this resulted in 89/2,188 and 40/6,792 volumes/slices, respectively. Finally, the DWI dataset contained 81/2,430 volumes/slices. The final noise scaling factors determined using an iterative local SNR-guided approach were as follows. T1: 1.5; T2: 2.0; T2*: 1.0, T2 FLAIR: 1.0, DWI: 1.5. Supplementary
Figure 2 plots the maximum, minimum and mean (dashed line) local SNR values within a region of interest across the GS and forward modeled datasets, for T₁ contrast. It can be observed that the means of the GS and forward-modeled data are comparable, which validates the iterative local SNR-guided approach to determining the noise scaling factor.

**Loss functions:** Supplementary File 2 presents a tabulation of volumetric measures of denoised data obtained from the automated volumetry tool. It compares the values of data denoised using the models trained on \( \beta = [0, 0.01, 0.1, 1] \). The model trained on \( \beta = 1 \) achieved the lowest mean RMSE value, and all subsequent models were trained with the same loss formulation.

**Training:** Supplementary Figure 3 shows a plot of training and validation losses as a function of epochs for the baseline denoising models, across contrasts. The corresponding approximate training durations are also listed. Figure 5 presents the corresponding mean changes in the image quality metrics computed on the test sets. In each instance, the model with the lowest validation loss was used. The largest gains in PSNR, MS-SSIM, and var-Lap values are observed on the T₂ contrast. Since the network architecture was common to all contrast-specific denoising models, this could be attributed to the larger noise scaling factor during the forward modeling process. Consequently, this might have forced the model to learn to denoise much noisier images during training in comparison with the training data from other contrasts. The lowest gains are observed on the T₂* contrast. This could potentially be because the publicly available T₂* dataset was corrupted by native noise sourced from an SWI acquisition (GS protocol).

**SS denoising:** Figure 6 plots the subject-specific median local SNR values. The subject-dependent variability of SNR validated our rationale for subject-specific denoising. Figure 7 plots the mean changes in PSNR, MS-SSIM, and var-Lap values across contrasts and subjects. Similar to the baseline denoising models, the largest gains are observed for the T₂ contrast, and the least improvement is observed for the T₂* contrast. In addition, T₂* SS model performed significantly poorer than the baseline model, and therefore the results have not been reported. We suspect this is due to the mismatch in the training and fine-tuning datasets. The ADNI-3 training dataset consisted of T₂* GRE acquisitions with an echo train length (ETL) of 3. On the other hand, the Fastest protocol utilized an ETL of 1 in the T₂* GRE acquisitions. We attribute the
inherent mismatch in signal between the datasets to the poor fine-tuned performance. Additionally, this potentially indicates that the pre-processing steps in this work are inadequate.

C. Experiments

E1 – Throughput: The cumulative acquisition times for the GS and Fastest protocols as per the vendor UI were 17:23 and 8:34. The practical acquisition times (obtained from the video recording) were 19:12 and 9:52. This discrepancy can be attributed to the time lost during pre-scan calibration and shimming functions. Overall, imaging one healthy volunteer using the Fastest protocol yielded a 1.94x gain in throughput over the GS protocol. Supplementary File 3 presents the timestamps and calculations of durations derived from the video recordings to obtain the final acquisition durations for this experiment. The cumulative median object-masked local SNR values for the GS and Fastest data were 243.354 and 215.767, respectively. Finally, this translates to MR Values of 0.211 and 0.364, respectively. Overall, employing the Fastest protocol resulted in a 72.51% increase in MR Value. In comparison, the corresponding SNR value for EE data was 264.136. Considering a practical acquisition duration of 07:58, this resulted in an MR Value of 0.552.

E2 – Image quality: Figure 8 is a bar graph plotting the median object-masked local SNR and var-Lap values across contrasts, for the GS, EE, and Fastest datasets. The mean values are presented at the bottom of the individual bars. For local SNR, similar performance is observed from the axial DWI and axial T2 FLAIR sequences, while not for the other sequences. The higher median local SNR values of the T1 contrast from the EE protocol can be attributed to the Turbo Spin Echo-based FLAIR sequence. The GS and EE protocols also perform better than the Fastest protocol in the T2 sequence, attributed to the longer repetition times. Overall, the EE protocol yields higher local SNR values due to higher slice thickness: 5 mm across all sequences, as opposed to ranges of 1.0–3.6 mm and 1.6–5.0 mm for GS and Fastest protocols, respectively. For var-Lap, comparable performance is observed only in the T2 FLAIR contrast. The Fastest protocol performs worse than both GS and EE in T1, DWI and T2 contrasts.

E3 – SNR recovery: Figure 6 presents the changes in median object-masked local SNR, PSNR, SSIM, and var-Lap values for the baseline and SS denoising models tested on the Fastest datasets. The solid and checkerboarded bars correspond to the baseline and SS denoising models, respectively. The T1 SS denoising model does not improve PSNR over the baseline denoising
model, and only modestly improves SSIM. However, it results in a smaller increase in blurriness. For T2, the SS denoising model yields larger improvements across all metrics. For T2 FLAIR, similar improvements are observed for PSNR and MS-SSIM, along with an undesirable increase in blurriness—indicating the model potentially denoised by primarily high-pass filtering. The T2* SS denoising models deteriorated image quality in every instance, and hence their results are not presented.

**E4 – Repeatability:** Figure 9 presents the plots of volumetric measures obtained from the automated tool for T1 contrast. The top row plots values of White Matter (WM) and Gray Matter (GM), and the bottom row corresponds to measures of two IDPs for AD–hippocampal and amygdala volumes. For each of these anatomies, a representative slice with the corresponding masks overlaid is illustrated in the figure inset. Figure 10 is a plot of manual volumetric measures for T2, T2*, T2 FLAIR and DWI contrasts.

**D. Visualizing learned filters for explainable AI**

Figure 11 is a collage of intermediate layer outputs obtained from denoising a DC-biased input using the baseline image denoising model for the T1 contrast. The model appears to perform denoising (akin to low-pass filtering) in the earlier layers. Each MaxPool2D layer halves the spatial dimension, leading to reduced resolution in the later layers (refer network architecture in Figure 3A). In these layers, the model appears to be performing low-frequency denoising (high-pass filtering). Overall, no brain-specific anatomy is identifiable across any of the intermediate layer outputs (as desired), potentially attributed to the patch-wise approach adopted in this work.

**E. Statistical analysis**

Among the four methods for all 30 locations, 27 locations had excellent ICC (>=0.93); 2 had a good ICC (>0.8), 1 had moderate ICC (=0.651). Table 3 lists the individual ICC values for each of the 27 brain subregions and the 3 brain tissues.

**IV. Discussion and conclusion**

The LUT search to accelerate the GS protocol was automated. In comparison, designing the EE protocol required human expertise and manual hours. The LUT approach is also scalable–
automated recording of acquisition times and rSNR values from the vendor UI for different \( P_{acq} \) can potentially enable the construction of high-dimensional LUTs. Subsequently, high-dimensional constrained search techniques can be explored to arrive at different \( P_{acq} \). Our LUT search formulation also allows optimizing for different criteria. We optimized for shorter acquisition durations whilst trading-off SNR. However, this can easily be modified to any other criteria by suitably modifying the weights described in Supplementary Table 1. Or, the LUT search can involve finding optimal \( P_{acq} \) that satisfies an imposed acquisition time constraint, as demonstrated in our previous work (Ravi et al.). Furthermore, since domain expertise is involved in setting the weights for the DOF, the LUT search is inherently explainable.

Initially, we trained our image-denoising models for T1 and T2 contrasts on the Human Connectome Project dataset (HCP, http://www.humanconnectomeproject.org/). Preliminary results (not reported in this work) indicated poor accuracy on the automated volumetry (high RMSE\(_{vol}\)), although the denoising performance was good. We attributed this to HCP data’s superior image quality–HCP data were acquired on Siemens Prisma 3 T scanners with 80 mT/m gradient strength and 200 T/m/s slew rate. A 3D T1 MPRAGE sequence was utilized with isotropic resolution and repetition/echo times = 2530/1.15 ms. Therefore, the iterative local SNR-guided approach resulted in a higher noise scaling factor to degrade the HCP data to match the median local SNR with that of the Fastest dataset. We suspect that the denoising models trained on this data caused excessive blurring, which subsequently affected the automated T1 volumetry. Therefore, we chose to proceed with the IXI dataset for T1 contrast, and also for T2 contrast to potentially mitigate a similar issue.

For T2*, the SS denoising models failed to demonstrate better performance than the baseline denoising models. The baseline model was trained on T2* data of the ADNI 3 dataset corrupted by native noise extracted from SWI data. However, fine-tuning the baseline model involved training on T2* data corrupted by native noise extracted from T2* data itself. We suspect this sequence-specific noise distribution could have impacted the training process of the SS models.

A. Limitations and future work

Intelligent protocolling: The T1 MPRAGE sequence in the Fastest protocol achieved shorter scan durations due to higher slice thickness (1.6 mm vs 1 mm). Future work could
involve exploring the impact of interpolating anisotropic data to achieve isotropic voxel resolutions on the accuracy of automated volumetry (Deoni et al., 2022). Although our LUT search formulation was designed to avoid modifying image contrast, the Fastest dataset marginally deviates from the GS dataset’s contrast. For example, this can be observed in the T2 FLAIR contrast in Figure 4. Potentially, Virtual Scanner (Tong et al., 2019) and its digital twinning capability (Tong, Vaughan and Geethanath, 2021) can be leveraged to design a physics-informed LUT optimization approach.

**Data distribution:** For detecting AD, volumetry from T1-MPRAGE sequence is crucial. The denoising models were evaluated on a small and healthy cohort of five volunteers. Their performance on denoising pathological data has not been investigated. While we have demonstrated the value of denoising in improving the accuracy of volumetry, the robustness of the denoising models on out-of-distribution data has not been considered. A thorough evaluation will be required to assess the quality of data acquired from pathological subjects and denoising using our models. Alternatively, the training dataset could include pathological data to improve the models’ generalization capabilities. Datasets which have not undergone extensive preprocessing and/or stringent quality control are valuable during the native noise extraction process of our workflow. Future iterations could involve training on a multi-site, multi-vendor, real-world dataset such as RadImageNet (Mei et al., 2022).

**Evaluation metrics:** This work utilizes a combination of referenceless (local SNR, var-Lap) and reference-based (PSNR, SSIM) image quality metrics. The referenceless metrics were borrowed from the broader computer vision community, and might not be ideal for evaluating methods in medical imaging. In particular, since the var-Lap metric is on an arbitrary scale, it does not allow performance comparisons without controlling for the testing dataset. The reference-based metrics inherently require a gold standard (GS), and hence do not lend themselves to evaluation on real-world data which, by nature, do not have reference data.

**Inference on denoising models:** While the patch-wise implementation enables flexibility of input data sizes, this approach significantly increases the inference durations – 4x on 256 × 256 input and 8x on 512 × 512 input, when compared with full input size approaches. Furthermore, the preprocessing step of converting a full-size image into 64 × 64 patches adds an overhead that is directly proportional to the dimensions of the input image. Currently, our denoising models require approximately 3.679 seconds per slice if the input image dimensions
are $512 \times 512$, and 2.575 seconds per slice if the input image dimensions are $256 \times 256$. Potentially, this could approximately be reduced 0.459 seconds and 0.321 seconds per slice, respectively, if a full input size were instead adopted. The denoising models are also not implemented in an end-to-end workflow – currently, the data needs to be transferred to a designated system via physical storage media. Future work will potentially involve streamlining file I/O to further accelerate DL denoising durations and packaging the pipeline to be tested for deployment at beta site.

**Explainable AI:** While there exist multiple methods that aid in interpretability classification models (Zhou, Khosla and Lapedriza, 2016; Shrikumar, Greenside and Kundaje, 06–11 Aug 2017; Selvaraju, Cogswell and Das, 2017; Smilkov et al., 2017), image-to-image model outputs are difficult to explain. Explainable AI techniques such Concept Activation Vectors (CAV) (Clough et al., 2019) allow probing the latent spaces of convolutional models to determine which human-friendly concepts the models are most sensitive to. However, this technique can only be applied to network architectures that include a bottleneck layer, such as U-Nets (Ronneberger, Fischer and Brox, 2015), autoencoders (or variants thereof, such as variational-quantized autoencoders (Van Den Oord and Vinyals, 2017)). To the best of our knowledge, there is no prior work on applying CAVs to investigate the performance of image-denoising models. Future work could involve exploring these network architectures to leverage CAVs for explainability.

**B. Conclusion**

This work demonstrates an end-to-end framework tailored for AD imaging. The framework involved implementing a LUT to shorten the acquisition duration of an existing brain imaging protocol that was employed at our institution, by sacrificing image quality. Accelerated brain imaging using this faster protocol was demonstrated, and image quality was recovered post-acquisition using DL-based image denoising models. Furthermore, MR physics dictates that the amount of signal captured relates to the volume of the subject being imaged, as this directly affects the size of the proton population. This variability of SNR depending on subject size motivated the authors to implement and demonstrate subject-specific image denoising.
V. Acknowledgements

The authors received support from the Biomedical Engineering and Imaging Institute, Dept. of Diagnostic, Molecular and Interventional Radiology, Icahn School of Medicine, Mt. Sinai, New York, NY, USA. Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD). For up-to-date information, see www.adni-info.org. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.
VI. References


https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.


A. Determining native noise source

1. Random subject
2. Compute local SNR maps
3. Sort by median local SNR
4. Pick 'noisiest' acquisition

B. Extracting native noise

- Brain images from noisiest acquisition
- Mask
- Collage
- Native noise block

C. Forward modelling

- Public dataset
- Scaling factor
- Random sampling
- Native noise block
- Noisy data
Performance of baseline image denoising models on test sets of public datasets

A

B

C

- Mean Δ PSNR (dB)
- Mean Δ MS-SSIM
- Mean Δ Var-Lap

T₁, T₂, T₂*, T₂ FLAIR, DWI
Subject-dependent variability of median local SNR

![Graph showing subject-dependent variability of median local SNR](image-url)
Performance of baseline and subject-specific (SS) image denoising models on test sets of the Fastest dataset

A

B

C

Δ PSNR (dB)

Δ MS-SSIM

Δ var-Lap
Comparison of median local SNR and variance of the Laplacian values of data acquired using Gold Standard, Expert Express and Fastest protocols, from five subjects.

Median local SNR:
- Sagittal $T_1$: Gold Standard 35.9774, Expert Express 48.4003, Fastest 37.4752
- Axial DWI: Gold Standard 33.9792, Expert Express 34.4803, Fastest 30.6984
- Axial $T_2$: Gold Standard 36.9987, Expert Express 32.8964, Fastest 47.9948
- Axial $T_2$ FLAIR: Gold Standard 32.5644, Expert Express 34.5925, Fastest 32.1118

Variance of the Laplacian:
- Sagittal $T_1$: Gold Standard 0.1272, Expert Express 0.1248, Fastest 0.1488
- Axial DWI: Gold Standard 0.1902, Expert Express 0.1411, Fastest 0.1141
- Axial $T_2$: Gold Standard 0.1021, Expert Express 0.0753, Fastest 0.0638
- Axial $T_2$ FLAIR: Gold Standard 0.1249, Expert Express 0.1021, Fastest 0.0843
- Axial SWI: Gold Standard 0.1313, Expert Express 0.0697, Fastest 0.0692
Comparison of Alzheimer’s Disease imaging specific volumetric measurements across the four methods

- WM volume
- GM volume
- Amygdala volume
- Hippocampal volume

Subject 1, 2, 3, 4, 5, 6

Volume (cm³)
Comparison of volumetric measures of benign incidental White Matter Hyperintensity in $T_2$, $T_2$ FLAIR, $T_2^*$ and DWI contrasts, across the four methods.
Visualizing intermediate layer outputs for explainable AI