Randomized Denoising Autoencoders for Smaller and Efficient Imaging Based AD Clinical Trials

Vamsi K. Ithapu[1], Vikas Singh[1], Ozioma Okonkwo[1], and Sterling C. Johnson[1,2]

[1] University of Wisconsin–Madison, USA
[2] William S. Middleton Memorial VA Hospital, USA
http://pages.cs.wisc.edu/~vamsi/rda

Abstract. There is growing body of research devoted to designing imaging-based biomarkers that identify Alzheimer's disease (AD) in its prodromal stage using statistical machine learning methods. Recently several authors investigated how clinical trials for AD can be made more efficient (i.e., smaller sample size) using predictive measures from such classification methods. In this paper, we explain why predictive measures given by such SVM type objectives may be less than ideal for use in the setting described above. We give a solution based on a novel deep learning model, randomized denoising autoencoders (rDA), which regresses on training labels y while also accounting for the variance, a property which is very useful for clinical trial design. Our results give strong improvements in sample size estimates over strategies based on multi-kernel learning. Also, rDA predictions appear to more accurately correlate to stages of disease. Separately, our formulation empirically shows how deep architectures can be applied in the large d, small n regime — the default situation in medical imaging. This result is of independent interest.

1 Introduction

Alzheimer's disease (AD) affects over 20 million people worldwide, and in the last decade, efforts to identify biomarkers for AD have intensified. There is now broad consensus that the disease pathology manifests in the brain images years before the onset of AD. Various groups have adapted sophisticated machine learning methods, to *learn* patterns of pathology by classifying healthy controls from AD subjects. The success of these methods (which obtain over 90% accuracy [16]) has led to attempts at more fine grained classification tasks, such as separating controls from Mild Cognitively impaired (MCI) subjects and even identifying which MCI subjects will go on to develop AD [14,7]. Even in this difficult setting, multiple current methods have reported over 75% accuracy. While accurate classifiers are certainly desirable, one may ask if they address a real practical need — if no treatments for AD are currently available, is AD diagnosis meaningful? To this end, [9,6] showed the utility of statistical learning methods beyond diagnosis/prognosis; they can in fact be leveraged for designing *efficient* clinical trials for AD. The basic strategy here uses imaging data from two time points (i.e., TBM data or hippocampus volume change), and derives a machine learning based biomarker. Based on this measure, the top one–third quantile subjects

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 470–478, 2014.

may be selected to be included in the trial. Using this "enriched" cohort, the drug effect can then be detected with higher statistical power with far fewer subjects, making the trial more cost effective and far easier to setup/conduct.

In this work, we ask if machine learning models can play a more fundamental role. Consider a trial where participants are randomly assigned to treatment (intervened) and placebo (non-intervened) groups, and the goal is to quantify any drug effect. Traditionally, this effect is quantified based on a "primary" outcome, like cognitive measure or brain atrophy. If the distributions of this outcome for the two groups are statistically different, we conclude that the drug is effective. When the effects are subtle, the number of subjects required to see statistically meaningful differences can be huge, making the trial infeasible. Instead, one may derive a "customized outcome" (i.e., a continuous predictor) from a statistical machine learning model. Here, the system assigns predictions based on probabilities of class membership (no enrichment is used). If these customized predictions are statistically separated (classification is a special case), it directly implies that potential improvements in power and the efficiency of the trial are possible. This paper is focused on designing specialized learning architectures towards this final objective. In principle, *any* machine learning method should be appropriate for the above task. But it turns out that high statistical power in these experiments is not merely a function of the classification accuracy of the model, rather the conditional entropy of the outputs (prediction variables) from the classifier at test time. An increase in classifier accuracy does not directly reduce the variance in the predictor (from the learnt estimator). Therefore, SVM type methods *are* applicable, but significant improvements are possible by deriving a learning model with the *concurrent* goals of classifying the stages of dementia *as well as* ensuring small conditional entropy of the outcomes.

Our contributions. We achieve these goals by proposing a novel learning model based on deep learning. Deep architectures are non-parametric learning models [1,3] that have received much interest in machine learning and computer vision recently. Although powerful, it is well known that they require very large amounts of unsupervised data, which is infeasible in neuroimaging, where the dimensionality d of the data is always much larger than the number of instances (n). A naïve use of off-the-shelf deep learning models on neuroimaging data expectedly yields poor performance. In the last few months, however, independent of our work, deep learning methods have been used successfully in structural and functional neuroimaging [12,5,10]. To get around the difficulty highlighted above, [12] uses a region of interest approach whereas [5,10] sub-samples each data instance to increase n. Our work provides a mechanism where no such adjustments are necessary. The key contributions of this paper include **(a)** Scalable deep architecture(s) for learning problems in neuroimaging where number of data instances is much smaller than the data dimensionality (i.e., our models permit whole–brain analysis) and **(b)** An imaging derived continuous measure with smaller variance that leads to efficient AD clinical trials with moderate sample sizes (and based only on one time–point data).

2 Model

2.1 Stacked Denoising Autoencoders (SDA)

We motivate our formulation by highlighting the difficulty in using stacked denoising autoencoders (SDA) [1,3] directly in the $n \ll d$ regime. An autoencoder is a single layer neural network that learns robust distributed *representations* of the input data. A denoising autoencoder (DA) constructs these representations by stochastically corrupting the inputs. Denoting the d dimensional inputs by $\{\mathbf{x}_i\}_1^n$, a DA outputs $\mathbf{h}_i = \sigma(\mathbf{W}\mathbf{x}_i + p)$ (σ is a point–wise sigmoid) by minimizing the loss $\mathcal{L}(.)$ between the input and its reconstruction $\hat{\mathbf{x}}_i = \sigma(\mathbf{W}^T\mathbf{h}_i + q)$ as,

$$\mathcal{Z}_{da}(\{\mathbf{x}_i\}_1^n, \theta) \coloneqq \arg\min_{\mathbf{W},p,q} \sum_{i=1}^{n} \mathbb{E}_{\tilde{x}\sim\gamma(\tilde{x}|x)} \mathcal{L}(\mathbf{x}_i, \sigma(\mathbf{W}^T\sigma(\mathbf{W}\mathbf{x}_i + p) + q)) \quad (1)$$

where $\gamma(.)$ is the point–wise stochastic corruption [1]. A stacked denoising autoencoder (SDA) greedily concatenates $L(> 1)$ DAs, i.e., l^{th} layer outputs are the un–corrupted inputs for $(l+1)^{\text{th}}$ layer,

$$\mathcal{Z}_{sda}(\{\mathbf{x}_i\}_1^n, L, \theta) := \sum_{l=0}^{L-1} \mathcal{Z}_{da}(\{\mathbf{h}_i^l\}_1^n, \theta) \;\; ; \;\; \mathbf{h}_i^l = \sigma(\mathbf{W}^l\mathbf{h}_i^{l-1} + p^l) \; ; \; \mathbf{h}_i^0 = \mathbf{x}_i \quad (2)$$

where θ denotes the full set of stochastic gradient (SG) learning parameters (corruption rate, learning rate, hidden layer length). The transformations $\{\mathbf{W}^l, p^l, q^l\}_1^L$ serve as a *warm–start* for supervised tuning where one compares the output of the L^{th} layer to $\{\mathbf{y}_i\}_1^n$. This greedy layer–wise unsupervised training followed by supervised fitting is central to most deep architectures [3,1].

Recall that SG learning is expected to converge to a local minimum only in the asymptotic setting (of large n). Hence, the *warm–start* described above is only reliable when large amounts of unsupervised data are available, which *is* the case in computer vision but not in neuroimaging. Otherwise, the network overfits whenever d is much larger than n. In neuroimaging, d is generally on the order of millions (number of voxels) and $n < 1000$. Hence, traditional SDAs cannot be directly used (they will generalize poorly). Recent work uses deep architectures in neuroimaging either by reducing d (using anatomical ROIs or feature selective) or increased n (splitting a data instance using sets of 2D slices) [12,5,10]. Nonetheless, frameworks to perform whole–brain analysis (the de-facto input when SVMs are used for brain image classification) will yield improvements by exploiting 3D local neighborhood dependencies directly.

2.2 Randomized Denoising Autoencoders (rDA)

In Section 1, we motivated the task of concurrently optimizing two goals. Our system should be able to capture differences across different dementia stages (i.e., controls, MCI, AD) while at the same time keeping intra-stage prediction variance as small as possible. In other words, we seek to decrease the prediction variance at no cost of approximation bias. Although, these seem like competing requirements, it turns out that this behavior is exactly what is offered by ensemble learning [2]. Recall that Ensembles are bootstrap randomizations around

sets of weak learners which reduce the prediction variance in expectation. So, properly incorporating an ensemble approach within a deep architecture should yield the behavior we expect. We can generate the ensembles for a given learner in multiple ways [2] — a randomization over the number of features and/or data instances. Here, we already have $n \ll d$, so randomization over n is infeasible. Instead, we distribute/randomize over the dimensions d where each weak learner will correspond to a SDA. This randomization allows a single SDA weak learner to process pathologically correlated voxels across 3D local neighborhoods while still operating on the whole-brain image. Unlike the SVM objective which has a global optimum, SDAs can only converge to a local optimum via SG [1]. We compensate for this by including a second level of randomization that samples sets of hyperparameters from a given hyper–parameter space. This basic structure drives the performance of our randomized denoising autoencoder (rDA).

Let $\mathcal{V} = \{1, \cdots, d\}$ denote the indices of dimensions/voxels, and $\tau(v)$, a distribution over $v \in \mathcal{V}$. In the simplest case, this can be a uniform distribution. We generate a bootstrap sample of B "blocks" where each block corresponds to input data along s_b dimensions/voxels (length of the block, fixed a priori). The mapping between voxels and blocks is given by $\tau(v)$. Note that blocks may not be mutually exclusive (a voxel may belong to multiple blocks). Each block will be presented to T weak learners. Each of these weak learners correspond to a unique $\theta_t \in \Theta$ for $t = 1, \ldots, T$ where Θ is the given hyper–parameter space. This means that each sample from the hyper-parameter space yields a weak learner. Our weak learner module is a L-layered stacked denoising autoencoder (SDA). The overall rDA architecture is an ensemble of $B \times T$ SDAs. Alg. 2.2 summarizes the block–wise training of rDA. Given training data as $\{\mathbf{x}_i, \mathbf{y}_i\}_1^n$, we first learn the transformations $(\mathbf{W}_{b,t}^l, p_{b,t}^l, q_{b,t}^l)\ \forall\ b, t, l$. Denoting the L^{th} layer outputs by $\mathbf{H}_i = [[\mathbf{z}_{b,t}\mathbf{h}_{b,t}^L]]_{1,1}^{B,T}$, the weighted regression pooling gives

$$\mathbf{U} \leftarrow (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}^T\mathbf{Y} \quad ; \quad \mathbf{H} = [[\mathbf{H}_i]]_1^n \quad ; \quad \mathbf{Y} = [[\mathbf{y}_i]]_1^n \tag{3}$$

where $\mathbf{U}$ are the regression coefficients and λ is the regularization constant. $\mathbf{z}$ is the known weight vector on $B \times T$ estimated L^{th} layer outputs. $[[\cdot]]$ denotes column–wise concatenation. The prediction for a new test input x is

$$\hat{\mathbf{y}} = \mathbf{h}\mathbf{U} \quad ; \quad \mathbf{h} = [[\mathbf{z}_{b,t}\mathbf{h}_{b,t}^L]]_{1,1}^{B,T}\ ;\ \mathbf{h}_{b,t}^l = \sigma(\mathbf{W}_{b,t}^l\mathbf{h}_{b,t}^{l-1} + p_{b,t}^l)\ ;\ \mathbf{h}_{b,t}^0 = \mathbf{x} \tag{4}$$

The simplest choice for the block–wise sampler $\tau(v)$ (i.e., at $b = 1$) assigns uniform probability over all dimensions/voxels. However, we can assign large weights on local neighborhoods which are more sensitive to dementia progression, if desired. Since d is large, we modify $\tau(v)$ after each iteration (*Reweigh* step in Alg. 2.2) to prevent starvation of the previously unsampled dimensions. We can also setup $\tau(\cdot)$ based on entropy or the result of a hypothesis test. Each weak learner output $\mathbf{h}_{b,t}^L$ is an estimate of $\mathbf{y}$. Recall that SG learning is sensitive to the choice of hyper–parameters θ_t, particularly, the number of epochs and gradient learning rate influences the range and variance of these estimates [1,3]. Hence, randomization over θ_t mitigates this dependency by averaging over T such estimates for each block (i.e., set of dimensions/voxels). We can pool the block estimates via various means – average, using a ridge regression or other sophisticated schemes.

But since SDAs are already capable of learning complex concepts [1], we use a simple linear combination with ℓ_2-loss providing minimum mean squared error. This addresses our goal of reducing the stochastic error of final predictions. Observe that rDA extends easily to multi–modal inputs by first constructing individual blocks for each modality and pooling across all modalities.

Algorithm. *rDA Blocks training*
Input: $\theta_t \sim \Theta, \mathcal{V}, B, s_B, L, T, \mathcal{D} \sim \{\mathbf{x}_i, \mathbf{y}_i\}_1^n, \lambda$
Output: $(\mathbf{W}_{b,t}^l, p_{b,t}^l, q_{b,t}^l)$
  for $b = 1, \ldots, B$ **do**
    $I_b \sim \tau(\mathcal{V})$
    for $t = 1, \ldots, T$ **do**
      $(\mathbf{W}_{b,l}^l, p_{b,l}^l, q_{b,l}^l) \leftarrow \mathcal{Z}_{sda}(\mathcal{D}, L, I_b, \theta_t)$
    end for
    $\tau(\mathcal{V}) \leftarrow Reweigh(\tau(\mathcal{V}), I_b)$
  end for

The sigmoid non–linearity ensures that rDA outputs are in $\in [0, 1]$. By labeling healthy controls as 1 and AD subjects as 0, we can then project the dementia scale to $[0, 1]$. The pooled outputs, referred to as rDA measure (rDAm), are then imaging–derived continuous predictors. We can then compute the sample sizes using rDAm as a customized outcome [11]. Denoting the mean change of rDA for placebo and treatment groups by μ_p and μ_t respectively, the number of subjects per arm is given by $2(Z_{\alpha/2} + Z_{1-\beta})^2\sigma^2/(\mu_p - \mu_t)^2$ where $1 - \beta$ is the desired power at significance level α. Using a conversion rate of $\rho \in [0, 1]$ from MCI to AD, and inducing a drug effect of η (i.e., the treatment decreases the mean change by a fraction η), the sample size expression then simplifies to $2c^2(Z_{\alpha/2} + Z_{1-\beta})^2/(\eta d)^2$ where $c = \sigma/\mu$ is the coefficient of variation. Since we only use one time–point data, the proportion ρ is set based on information from previously reported studies [13] Since rDA is an ensemble designed to reduce the prediction variance (and hence c), we hope to see much smaller sample sizes compared to others.

3 Evaluations

3.1 Data and Setup

We used Amyloid, FDG–PET and MRI data at baseline for 447 subjects (210 male, 237 female) from ADNI2 (Alzheimer's Disease Neuroimaging Initiative). 131 were healthy (CN), 92 were demented (AD), 120 and 104 had early and late MCI (EMCI and LMCI) respectively. The labeling of EMCI and LMCI (done by ADNI) is based on the cognitive status of each subject. Of the 224 MCIs, 100 had maternal family history (FH) of AD, 52 had paternal and 23 had both. Preprocessing included extracting grey matter in normalized space, and correcting PET for average intensities in ponsvermis (FDG) and cerebellum (Amyloid). We train rDA on ADs (labeled 0) and CNs (labeled 1) *alone*, and test on MCIs. We use a multi-modal (MKL) ϵ–support vector regression (ϵMKm) as the baseline learning model [7]. Firstly, we evaluate if rDAm differentiates EMCI from LMCI. Additionally, we evaluated parental family history as a contributing risk factor. Since rDAm is a continuous marker, its correlations with CSF levels – τ, $p\tau$, $A\beta$,

$\tau/A\beta$ and $p\tau/A\beta$ (τ: τ-protein, $p\tau$: phospho τ-protein, $A\beta$: Amyloid-β, are the cerebrospinal fluid protein levels, and very sensitive biomarkers of AD) are also assessed to verify if it is meaningful. We then estimate sample sizes using MCI to AD conversion rate of 37.7% [13]. rDA hyper-parameters in our experiments are $L = 2$, $B = 1000$ and $T = 9$, with uniform weighting (i.e., $z_{b,t} = \frac{1}{BT} \forall b, t$).

3.2 Results and Discussion

Table 1(a) and (b) show that rDAm is highly sensitive to EMCI vs. LMCI and the influence of FH. Although the baseline ϵMKm picks up these group differences, rDAm has much higher delineation power. In particular, the p–values for rDAm for FH positive vs. negative case are an order of magnitude smaller than that of ϵMKm. These show that rDAm is at least as good as a current state-of-the-art machine learning derived measures. Table 2 shows that rDAm has significant correlation (higher than ϵMKm in all but two cases) to CSF levels, which are proven biomarkers for MCI to AD progression [15]. Note that a negative correlation with say τ implies that rDAm decreases and the subject gets demented as τ increases. Specifically, higher correlations (with $p \leq 0.01$) with $p\tau$ and $p\tau/A\beta$ suggest that rDAm is a useful continuous predictor. In most cases, these significance levels increase as more modalities are combined.

Table 3 shows that the coefficient of variation (CV) of rDAm for three different populations of interest – all MCIs, LMCIs and MCIs with positive FH. Observe that rDAm's CV is smaller than that of ϵMKm for all three populations, and all possible combinations of modalities – making it a better candidate to be used as a prediction measure. Also, the CVs for MCIs with

Table 1. Performance of rDAm vs. ϵMKm in delineating MCI sub–groups. A : Amyloid, F : FDG and T : T1GM. Each cell shows the ANOVA p–value and corresponding F–statistic.

Model	Amyloid	FDG	T1GM	A+F	A+T	F+T	A+F+T
(a) Early versus Late MCI							
MKL	$<$ **.001**, 20.5	$<$ **.001**, 16.8	$<$ **.001**, 16.5	$<$ **.001**, 16.4	$<$ **.001**, 20.4	$\ll$ **.001**, 23.6	$\ll$ **.001**, 27.9
rDA	$\ll$ **.001**, 22.1	**.001**, 9.7	$<$ **.001**, 20.0	$<$ **.001**, 19.5	$\ll$ **.001**, 24.1	$\ll$ **.001**, 21.2	$\ll$ **.001**, 27.6
(b) Family History Positive versus Negative							
MKL	*0.04*, 4.3	**0.007**, 7.5	*0.02*, 5.3	**0.007**, 7.3	**0.009**, 6.8	*0.01*, 6.6	**0.004**, 8.3
rDAm	*0.03*, 4.7	$<$ **.001**, 11.8	$<$ **.001**, 11.2	**0.009**, 6.8	$<$ **.001**, 12.4	$<$ **.001**, 13.2	$<$ **.001**, 13.3

Table 2. Correlation of CSF levels to rDAm vs. ϵSm. Note that Amyloid is used as reference modality here. Each cell represents the Spearman correlation p–value and the coefficient for the corresponding marker.

CSF	Marker	Amyloid		A+F		A+T		A+F+T	
τ	ϵMKm	*0.01*	-0.24	*0.04*	-0.20	0.14	-0.15	0.15	-0.15
	rDAm	**0.002**	-0.31	*0.01*	-0.25	0.09	-0.17	0.11	-0.16
pτ	ϵMKm	$<$ **0.001**	-0.40	$<$ **0.001**	-0.37	*0.01*	-0.27	**0.008**	-0.28
	rDAm	$<$ **0.001**	-0.39	$<$ **0.001**	-0.36	**0.008**	-0.27	**0.009**	-0.28
Aβ	ϵMKm	*0.01*	0.24	0.09	0.17	0.29	0.11	0.38	0.09
	rDAm	*0.03*	0.22	*0.03*	0.23	0.44	0.08	0.37	0.09
τ/Aβ	ϵMKm	**0.004**	-0.29	*0.01*	-0.27	0.15	-0.15	0.14	-0.15
	rDAm	$<$ **0.001**	-0.35	**0.007**	-0.28	0.08	-0.18	0.11	-0.16
pτ/Aβ	ϵMKm	**0.001**	-0.41	**0.001**	-0.37	*0.01*	-0.25	*0.01*	-0.25
	rDAm	$<$ **0.001**	-0.42	$<$ **0.001**	-0.37	*0.01*	-0.26	*0.01*	-0.26

positive FH are smaller than that of late MCIs. This also suggests that a significant number of late MCIs currently have only a mild dementia in terms of rDAm. Fig. 1 shows the estimates on the same three populations of interest as above at 80% power (Refer to the supplement for more plots). Following Table 3, it should be straight forward to expect smaller sample estimates for rDAm compared to ϵMKm, which is exactly the case as shown in Fig. 1. Particularly, FDG and MRI gave smaller estimates than that of Amyloid following their smaller CV, reflecting that $\sim 30\%$ of healthy elderly have positive Amyloid scans. FH positive MCIs (last plot in Fig. 1) lead to much smaller sizes compared to using all MCIs and late MCIs. To get a sense of the improvement with respect to non–imaging based markers, we compared the best estimate (over all modalities) of ϵMKm and rDAm with that of MMSE and CSF levels in Table 4. At 80% power, the best estimates across CSF markers was 973 and 975 (for τ and $\tau/A\beta$ respectively) compared to that of 193 using rDAm – more than 5–fold decrease. It should be noted that all these estimates use only "single time–point" data combined with known conversion rates, in contrast to direct longitudinal measurement [8,4]. Hence, the sizes using MMSE and CSF are as high as 1500, indicating that estimates on the order of two hundred (that of rDAm) are highly significant. Overall, the results show that imaging–derived markers lead to much smaller trials than cognitive scores and/or CSF levels.

Table 3. CV of rDAm vs. ϵMKm

Modality	Marker	MCIs	LMCIs	FHMCIs
Amyloid	ϵMKm	0.56	0.70	0.42
	rDAm	**0.49**	**0.57**	**0.41**
FDG	ϵMKm	0.49	0.53	0.39
	rDAm	**0.33**	**0.36**	**0.26**
T1MRI	ϵMKm	0.55	0.60	0.48
	rDAm	**0.36**	**0.42**	**0.26**
A+F	ϵMKm	0.52	0.63	0.39
	rDAm	**0.42**	**0.49**	**0.33**
A+T	ϵMKm	0.56	0.67	0.42
	rDAm	**0.41**	**0.49**	**0.29**
F+T	ϵMKm	0.51	0.58	0.41
	rDAm	**0.34**	**0.38**	**0.25**
A+F+T	ϵMKm	0.54	0.65	0.39
	rDAm	**0.41**	**0.50**	**0.28**

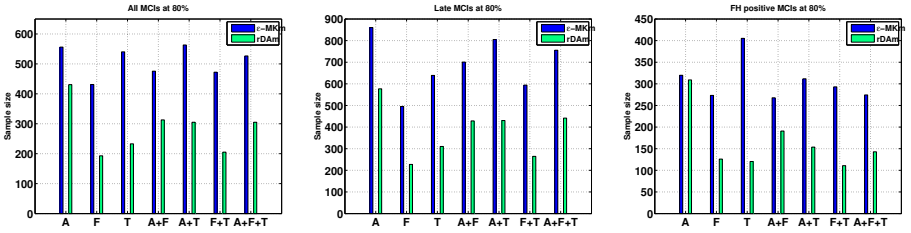

Fig. 1. Sample estimates per arm for rDAm vs. ϵMKm using all MCIs, LMCIs and FH positive MCIs respectively, at 80% power and 0.05 significance level. Conversion rate is 37.7%, and the induced drug effect is 0.25. Refer to supplement for 85% and 90% plots. ϵMKm is blue and rDAm is green.

Table 4. Best rDAm and ϵMKm sample estimates perm arm (from Fig. 1) vs. MMSE and CSF levels.

Power	MMSE	τ	$p\tau$	$A\beta$	$\tau/A\beta$	$p\tau/A\beta$	ϵMKm	rDAm
80%	> 2500	973	1447	> 2500	975	> 2000	431	**193**
85%	> 2500	1117	> 1500	> 2500	1120	> 2500	495	**221**
90%	> 2500	1303	> 1500	> 2500	1306	> 2500	577	**258**

4 Conclusions

We propose a novel deep learning architecture, randomized denoising autoencoders, that scales to very large dimensions and learns from a small number of instances. We construct a continuous predictor based on rDA and show that not only does it have high correspondence with other markers of AD, but also leads to efficient clinical trials with much smaller sample estimates.

Acknowledgments. This work was supported in part by NIH R01 AG040396; NSF CAREER award 1252725; Wisconsin Partnership proposal; UW ADRC P50 AG033514; UW ICTR 1UL1RR025011 and a VA Merit review grant I01CX000165. The contents do not represent views of the Dept. of Veterans Affairs or the US Government.

References

1. Bengio, Y.: Learning deep architectures for AI. Foundations and Trends in Machine Learning 2, 1–127 (2009)
2. Dietterich, T.G.: Machine-learning research. AI Magazine 18(4), 97–136 (1997)
3. Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P., Bengio, S.: Why does unsupervised pre-training help deep learning? JMLR 11, 625–660 (2010)
4. Grill, J.D., Di, L., Lu, P.H., Lee, C., Ringman, J., Apostolova, L.G., et al.: Estimating sample sizes for predementia Alzheimer's trials based on the Alzheimer's Disease Neuroimaging Initiative. Neurobiology of Aging 34, 62–72 (2013)
5. Gupta, A., Ayhan, M., Maida, A.: Natural image bases to represent neuroimaging data. In: Proceedings of the 30th ICML, pp. 987–994 (2013)
6. Hinrichs, C., Dowling, N.M., Johnson, S.C., Singh, V.: MKL-based sample enrichment and customized outcomes enable smaller AD clinical trials. In: Langs, G. (ed.) MLINI 2011. LNCS (LNAI), vol. 7263, pp. 124–131. Springer, Heidelberg (2012)
7. Hinrichs, C., Singh, V., Xu, G., Johnson, S.C.: Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. Neuroimage 55, 574–589 (2011)
8. Holland, D., McEvoy, L.K., Dale, A.M.: Unbiased comparison of sample size estimates from longitudinal structural measures in ADNI. Human Brain Mapping 33(11), 2586–2602 (2012)
9. Kohannim, O., Hua, X., Hibar, D.P., Lee, S., Chou, Y.Y., Toga, A.W., Jack Jr., C.R., Weiner, M.W., Thompson, P.M.: Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiology of Aging 31, 1429–1442 (2010)
10. Plis, S.M., Hjelm, D.R., Salakhutdinov, R., Calhoun, V.D.: Deep learning for neuroimaging: a validation study. arXiv preprint arXiv:1312.5847 (2013)

11. Sakpal, T.V.: Sample size estimation in clinical trial. Perspectives in Clinical Research 1(2), 67–69 (2010)
12. Suk, H.-I., Shen, D.: Deep learning-based feature representation for AD/MCI classification. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 583–590. Springer, Heidelberg (2013)
13. Tatsuoka, C., Tseng, H., Jaeger, J., Varadi, F., Smith, M.A., Yamada, T., et al.: Modeling the heterogeneity in risk of progression to Alzheimer's disease across cognitive profiles in mild cognitive impairment. Alzheimers Res. Ther. 5, 14 (2013)
14. Teipel, S.J., Born, C., Ewers, M., Bokde, A.L., Reiser, M.F., Möller, H.J., Hampel, H.: Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. Neuroimage 38, 13–24 (2007)
15. Vemuri, P., Wiste, H., et al.: MRI and CSF biomarkers in normal, MCI, and AD subjects predicting future clinical change. Neurology 73(4), 294–301 (2009)
16. Zhang, D., Wang, Y., Zhou, L., et al.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55(3), 856–867 (2011)