Machine Learning Approaches for Identification of Alzheimer's Disease using Social Determinants & Imagery

Lawrence V. Fulton, Ph.D. MSStat MHA MSS MMAS MS, CAP CQE CSSBB CStat CSci FACHE

Department of Health Administration College of Health Professions, Texas State University

Abstract

Purpose
The purpose of this study is to predict the presence of Alzheimer's Disease (AD) using socio-demographic, clinical, and Magnetic Resonance Imaging (MRI) 4D data.

Significance
Early detection of AD enables family planning and may reduce costs by delaying long-term care (Alzheimer's Association, 2018). Accurate, non-imagery methods also reduce patient costs.

Methods
Extreme Gradient Boosted random forests (XGBoost) predict Clinical Dementia Rating (CDR) presence and severity as a function of gender, age, education, socioeconomic status (SES), and Mini-Mental Status Exam (MMSE). Convolutional Neural Networks (CNN) predict CDR from MRI's transformed to Eigenbrain imagery. XGBoost also predicts CDR with additional clinical variables.

Results
XGBoost provides 93% prediction accuracy for CDR using socio-demographic and clinical non-imagery variables-92% accuracy when clinical measures are excluded. CNN using the transformed Eigenbrain imagery results in 93% prediction accuracy.

Conclusion
ML methods predict AD with high accuracy. Non-imagery analysis may be nearly as efficacious as imagery prediction at a fraction of the cost.

Background, Literature, Data, Variables

Background
Alzheimer's is a disease for which there is no cure (Mayeux & Sano, 1999). Diagnosing it early facilitates family planning and cost control (Alzheimer's Association, 2018). The cost for a brain MRI may range from approximately $600 to $1,300 (Vanvuren, 2017), and for the uninsured or underinsured, this might be infeasible.

Literature
Zhang et al. (2015, 2016) used a subset of the Open Access Series of Imaging Studies (OASIS) data (n=126) and proposed methods to identify binary-coded AD using Eigenbrain imagery. They eliminated individuals under 60 / incomplete observations, modeled AD as binary, and limited their analyses to a subset of coronal slice images. This work extends their efforts.

Data & Variables
The OASIS data provides researchers access to cross-sectional and longitudinal MRI data (OASIS, 2018) and is based on the efforts of Marcus et al. (2007). N=416 patients

Response variable: Clinical Dementia Rating. {0= nondemented; 0.5 – very mild dementia; 1 = mild dementia; 2 = moderate dementia} (Morris, 1993)

Demographic predictor variables: Gender: {0=Female, 1=Male}, Age: [18, 96], Education: {1<HS, 2=HS Grad, 3=Some College, 4=College Grad, 5=Beyond College}, Socioeconomic Status: {1=lower, 2=lower middle, 3=middle, 4=upper middle, 5=upper}

Clinical predictor variables: Mini-Mental Status Exam: [0,30] (Rubin et al., 1998), Estimated Total Intracranial Volume: mm3 (Buckner et al., 2004), Atlas Scaling Factor: [.88, 1.56] (observed) (Buckner et al., 2004), Normalized Whole Brain Volume: [.64, .90] (observed) (Fotenos et al., 2004)

Imagery predictor variables: Brain Imagery. Masked version of the gain-field corrected, atlas-registered image to the 1988 atlas space of Talairach (Buckner et al., 2004)

Objectives

1. Predict presence & severity of AD using ML methods with / without imagery
2. Provide alternatives to imagery for identification of AD for the poor & underserved

References (1 of 3)

Allaire, J.J. & Chollet, F. (n.d.). Keras: R interface to Keras. Retrieved from https://keras.rstudio.com

Alzheimer's Association. (2015). Early detection & diagnosis of Alzheimer's disease. Retrieved from https://www.alz.org/publichealth/downloads/policy-brief.pdf

Bordier, C., Dojat, M., & Lafaye de Micheaux, P. (2011, October 29). Temporal and spatial independent component analysis for fMRI data sets embedded in the AnalyzeFMRI R package. Journal of Statistical Software, 44(9), 1-24.

Buckner, R.L, Head, D., Parker, J., Fotenos, A.F., Marcus, D., Morris, J.C., Snyder, A.Z. (2004, October). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: Reliability and alidations against manual measurement of total intracranial volume. Neuroimage, 23(2), 724-38.

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2018). xgboost: Extreme gradient boosting. Retrieved from https://CRAN.R-project.org/package=xgboost

Chollet, F. (2015, June 19). How to cite keras in research paper? Retrieved from https://github.com/keras-team/keras/issues/247

References (2 of 3)

Fotenos, A.F., Snyder, A.Z., Girton, L.E., Morris, J.C., & Buckner, R.L. (2005, March 22). Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. Neurology, 64(6), 1032-39.

Marcus, D.S., Wang, T.H., Parker, J.M., Csernansky, J.G., Morris, J.C., & Buckner, R.L. 2007, September). Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented and demented older adults. Journal of Cognitive Neuroscience, 19(9), 1498-1507.

Mayeaux, R. & Sano, M. (1999, November 25). Treatment of Alzheimer's disease. New England Journal of Medicine, 341(22), 1670-9. doi: 10.1056/NEJM199911253412207

Morris, J.C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology, 43(11), 2412-4.

Open Access Series of Imagine Studies. (2018, January 7). What is OASIS? Retrieved from http://www.oasis-brains.org/app/template/index.vm

R Development Core Team. (2015, February 10). R: A language and environment for statistical computing. Retrieved from https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing

References (3 of 3)

Rubin, E.H., Storandt, M., Miller, J.P., Kinscherf, D.A., Grant, E.A., Morris, J.C., & Berg, L. (1998, March). A prospective study of cognitive function and onset of dementia in cognitively health elders. Arch Neurol, 55(3), 395-401.

Van Buuren, S. & Groothuis-Oudshoorn, K. (2011, December). Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3). Retrieved from https://www.jstatsoft.org/v45/i03/paper

Vanvuren, C. (2017, April 20). How much does a brain MRI cost? Retrieved from https://amino.com/blog/brain-mri-cost

Venables, W.N. & Ripley, B.D. (2002). Modern applied statistics with S (fourth ed.). New York, NY: Springer.

Zhang, Y., Dong, Z., Phillips, P., Wang, S., Ji, G., Yang, & Yuan, T. (2015, June 2). Detection of subjects and brain regions related to Alzheimer's disease using 3 MRI scans based on eigenbrain and machine learning. Frontiers in Computational Neuroscience, 9, 66. doi: 10.3389/fncom.2015.00066

Zhang, Y., Wang, S., Phillips, P., Yang, J., & Yuan, T.F. (2016). Three-dimensional eigenbrain for the detection of subjects and brain regions related with Alzheimer's disease. Journal of Alzheimer's Disease, 50(4), 1163-79. doi: 10.3233/JAD-150988

Methods

Exploratory Data Analysis
EDA techniques included imputation of missing sociodemographic data through Multiple Imputation using Chained Equations [MICE] (van Buuren & Groothuis-Oudshoorn, 2011) and Box-Cox transformation investigation (Venables & Ripley, 2002).

Image Manipulation
The AnalyzeFMRI package (Bordier, Dojat, & Lafaye de Micheaux, 2011) in R Statistical Software (R Development Core Team, 2016) provided image import utility. Principal Component Analysis (PCA) generated the Eigenbrain structures.

Gradient Boosted Tree Ensembles (Gradient Boosting)
The R Statistical Software Package, XGBoost (Chen et al., 2018), generated gradient boosted random forests. Hyperparameter grid search provided parameters for use in building the forests. Ten-fold cross-validation provided accuracy metrics. Individual trees grew no more than three branches to avoid overfitting.

Deep Learning with Convolutional Neural Networks (CNN).
The Keras (TensorFlow backend) package in R served as the means for evaluating images (Chollet, 2015) provided a reasonable method for imagery prediction. CNN's take advantage of pooling and are ideal for image recognition.

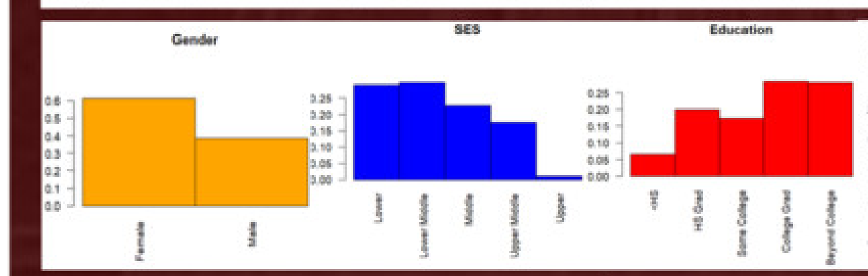
Descriptive Statistics

		Non-Demented				Demented				
Age	N	n	mean	male	female	n	mean	male	female	CDR 0.5/1/2
<20	19	19	18.53	10	9	0		0	0	0/ 0/ 0
[20,30)	119	119	22.82	51	68	0		0	0	0/ 0/ 0
[30, 40)	16	16	33.38	11	5	0		0	0	0/ 0/ 0
[40, 50)	31	31	45.58	10	21	0		0	0	0/ 0/ 0
[50, 60)	33	33	54.36	11	22	0		0	0	0/ 0/ 0
[60, 70)	40	25	64.88	7	18	15	66.13	6	9	12/ 3/ 0
[70, 80)	83	35	73.37	10	25	48	74.42	20	28	32/15/ 1
[80, 90)	62	30	84.07	8	22	32	82.88	13	19	22/ 9/ 1
[90, 100)	13	8	91.00	1	7	5	92.00	2	3	4/ 1/ 0
Total	416	316		119	197	100		41	59	70/28/ 2

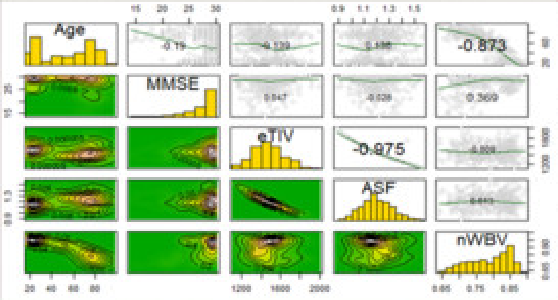
Descriptive Statistics for Dementia by Age and Gender (Adopted from OASIS , 2018)
The data include 416 patient diagnostic files with 100 of those files confirming dementia. No patients under the age of 60 were diagnosed with dementia, as this is a rare event.

Variable	n	mean	sd	median	min	max
Age	416	52.70	25.08	56	18	96
Mini-Mental Status Exam	416	27.50	3.13	29	14	30
eTIV (Intracranial Volume)	416	1480.53	158.34	1475	1123	1992
nWBV (Brain Volume)	416	0.79	0.06	0.8	0.64	0.89
ASF (Atlas Scaling Factor)	416	1.2	0.13	1.19	0.88	1.56

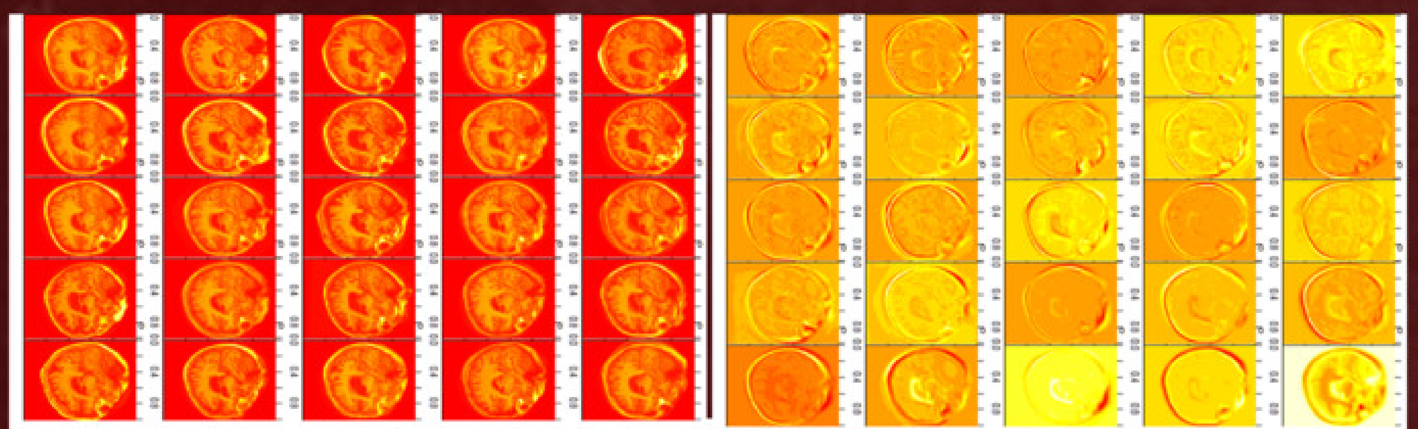
Descriptive Statistics, Quantitative Variables
The average patient in the dataset was 52.7 years old with a slightly less than perfect mental state evaluation (27.5 out of 30), an estimated brain volume of 1480.53 mm3, and 79% of the intracranial cavity occupied by the brain (nWBV).



Descriptive Statistics, Qualitative Variables
Modal responses follow: 62% women, 28% college graduates, 30% middle class. Spearman's rank correlation between SES and education indicated a negative correlation (ρ = -.715, p<.01). For the 121 individuals without a high school education, 119 were in the lower middle and lower classes.



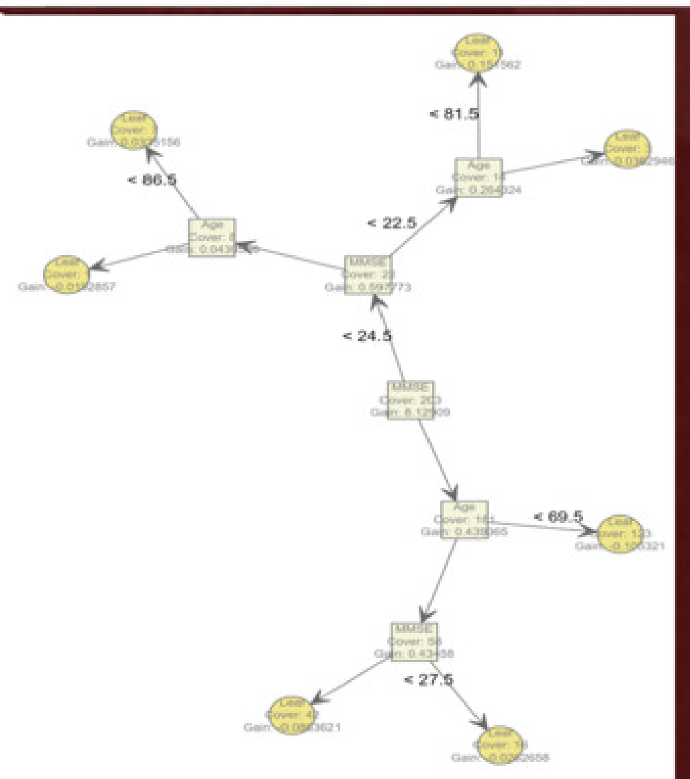
Quantitative Pairs
Correlation between ASF and eTIV is negative and significant (t414 =-88.7, p<.001) as is the correlation between age and nWBV (t414 =-36.4, p<.001). The likelihood ratio test (LRT) of all I = 1 is rejected (p<.001), as the optimal power transformation for multivariate normality is {.91, 7.0, .13, -.29, 5.29} for the variables in order of display. MMSE is particularly resistant to transformation. Tree methods and deep learning do not require normality, but rather benefit from scaling.



Transformation of MRIs and Eigenbrain Development
On the left, are axial MRI slices (orange). On the right, are Eigenbrain images derived by Principal Components Analysis where each component is maximized via $Max_{a(i)}\ L(i) = a^T_i \boldsymbol{R} a_i - \lambda_i(a^T_i a_i - 1)$. Vector *a* and scalar λ are the Eigenspace.

Results

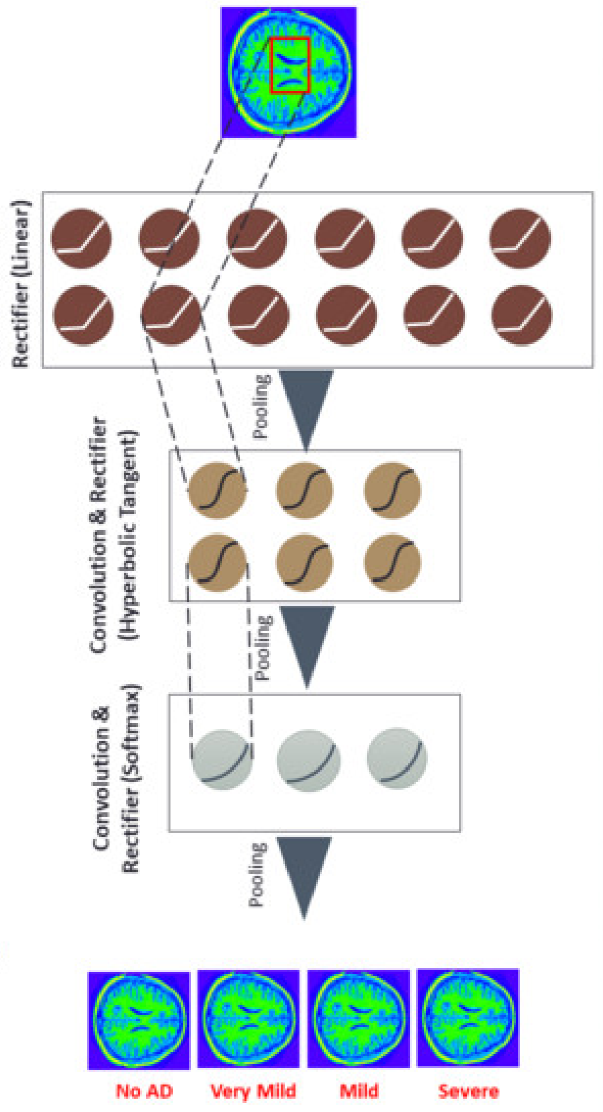
Gradient Boosting with XGBoost (Chen et al., 2018), No Imagery

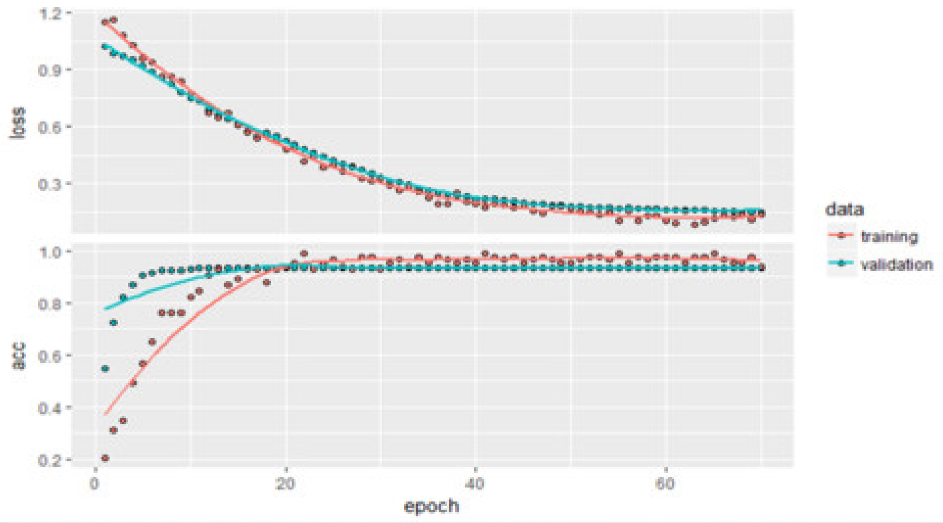


Hyperparameter grid search tuned the random forest. Ten-fold cross-validation provided accuracy metrics. To avoid overfitting, the tree depth was restrained to three. XGBoost classified 93% of the cases correctly with socio-demographic & clinical data. Gradient boosting also classified 92% accurately with only socio-demographic variables (PPV=90%, NPV=97%) and the following importance (% occurring in trees) for the variables as follows: MMSE (46%), Age (35%), Education (8%), SES (8%), Gender (3%).

Convolutional Neural Networks (Deep Learning) with Keras and Tensorflow (Allaire & Chollet, n.d.)

A Convolutional Neural Network (CNN) is a deep, layered, neural network ideal for working with image data. A picture of a convolutional neural network is to the right. A relatively simple three-layer model with {12, 6, 3} nodes and three separate activation functions {linear, hyperbolic tangent, and softmax} applied to all available image data (scaled) predicted a 30% validation set with 93% accuracy after 50 epochs (batch size of 8, categorical hinge loss function, Rmsprop optimizer). The convergence of the model is shown below. This result is similar to that of Zhang et al. (2015), although this analysis includes all available data and a multi-level response variable.


Convolution & Rectifier (Linear)
Pooling
Convolution & Rectifier (Hyperbolic Tangent)
Pooling
Convolution & Rectifier (Softmax)
Pooling
No AD
Very Mild
Mild
Severe



Conclusions

- Extreme gradient-boosted tree models provided 93% accuracy in predicting AD and associated severity with clinical data and 92% accuracy with only socio-demographics data.
- The most important variables in predicting AD with socio-demographic data only included MMSE, Age, Education, SES, and MEF, respectively.
- With imagery data, CNN accurately identified 93% of the AD by severity with a three layer model.
- Machine learning techniques might be used to identify AD in an automated fashion, and it might be possible to develop models with high predictive accuracy that do not require imagery.

TEXAS STATE UNIVERSITY
The rising STAR of Texas
MEMBER THE TEXAS STATE UNIVERSITY SYSTEM