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-invasive 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-imager variables and 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-imaging analyses may be nearly as efficacious as imaging prediction at a fraction of the cost.

Background, Literature, Data, Variables
Background
Alzheimer’s is a disease for which there is no cure (Mayoux & 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 (Vanmuren, 2017), and for the uninsured or underinsured, this might be insurmountable.

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 or incomplete observations, modeled AD as binary, and limited their analyses to a subset of coronal slices. 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 variables: Clinical Dementia Rating. (0=non-demented, 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<=15, 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: [0,0.3] (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, 30] (observed) (Potemen 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 imaging for identification of AD for the poor & underserved

Methods
Exploratory Data Analysis
EDA techniques included imputation of missing socio-demographic data through Multiple imputation using Chained Equations (MICE) (van Buuren & Groothuis-Oudshoorn, 2011) and Box-Cox transformation investigation (Venable & Ripley, 2002).

Image Manipulation
The AnalyzeFMRI package (Bordier, Doja, & 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 neural networks (Chelliot, 2015) provided a reasonable method for imagery prediction. CNNs take advantage of pooling and are ideal for image recognition.

Descriptive Statistics

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 (33%), Education (8%), SES (8%), Gender (3%).

Convolutional Neural Networks (Deep Learning) with Keras and Tensorflow (Albarel & Chelliot, 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 39% 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 in Figure 1. The result is similar to that of Zhang et al. (2015), although this analysis includes all available data and a multi-level response variable.

Conclusions

- Extreme gradient-boosted tree models provided 93% accuracy in predicting AD and associated severity with clinical data and 92% accuracy with only socio-demographic 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.

References (1 of 3)

References (2 of 3)

References (3 of 3)