Classifying Adolescent Excessive Alcohol Drinkers from fMRI Data

Yong-hun Kim, Cindy Liu, Joseph Noh

{ykim9, cliu15, jnoh2}@stanford.edu

CS229: Machine Learning, Stanford University

Abstract

Excessive alcohol drinking impacts the structural development of brain in adolescents[1], but its impact on the functional activity or connectivity of the brain has not yet been explored.

Our goal is to design a classification model to predict if a subject is a heavy drinker based on their resting-state fMRI data (stored as blood oxygen-level dependent (BOLD) signals). We used logistic regression of pre-processed data as a baseline for CNN/RNN-based models and SVMs.

Surprisingly, we found that using derived features with logistic regression yielded far better results than applying the simple, processed data to complex models.

Data and Features

Dataset

- Source: National Consortium on Alcohol and Neurodevelopment in Adolescence[2] (NCANDA) database
- fMRI scans of m = 715 adolescents and young adults (16-19 y/o), measured as BOLD signals from each voxel every between each T = 269 timesteps (2.2 seconds / timestep)
- Dataset was imbalanced (122 (17%) heavy drinkers out of 715)

Pre-processing

- Parcellate brain into regions (N) to reduce noise
- Brain activity was normalized to z-score
- Downscaling of majority class (size(1) == size(0))

Raw Features (m x T x N)

- m = 715 subjects / 244 after downscaling
- T = 269 timesteps
- N = Variable (brain regions)

Parcellation Method	Num regions (N)
Independent Component Analysis (ICA)	25
Craddock Parcellation[3]	100

Derived Features (m x N)

- Dynamic range per brain region in ICA
  - $x_{derived}(N) = \max(x_{,N}) - \min(x_{,N})$
- Demographics
  - sex, age, scanner type

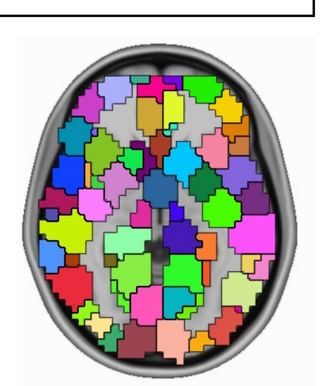

Figure 1. Craddock Parcellation example[3]

Models

Performance measured using 10-fold cross-validation. Logistic Regression implemented Newton's Method. All deep-learning models used batch binary cross-entropy as the loss and were implemented through Keras/Theano.

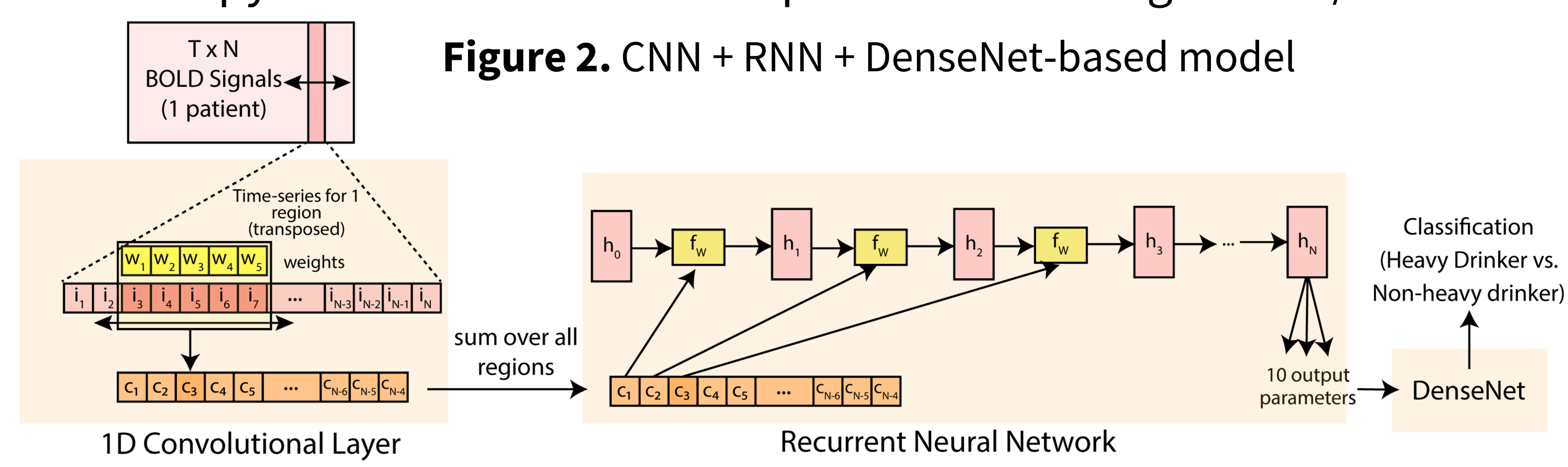

Figure 2. CNN + RNN + DenseNet-based model

- Logistic Regression with the derived features & demographics (baseline)
- Neural Networks using ICA and/or Craddock
  - Recurrent Neural Network (RNN) only
  - RNN or Convolutional Neural Network (CNN) + DenseNet
  - CNN + RNN + DenseNet (Figure 2)
- Support Vector Machines (Linear, Polynomial, Sigmoid, RBF kernels)

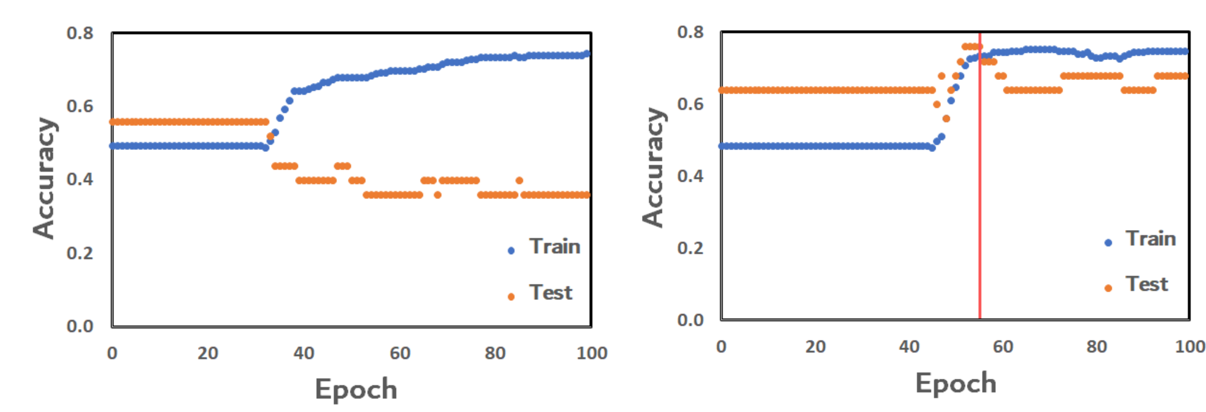

Figure 3. Train and test set accuracies over epochs. Number of epochs (55) was selected based on performance over epochs.

Results

Parcellation	Model	Train Accuracy (N = 220)	Train F1 Score	Test Accuracy (N = 24)	Test F1 Score
ICA	LR	0.799	0.800	0.713	0.709
ICA	LR - Age	0.649	0.619	0.538	0.487
ICA	C + R + N	0.535	nan	0.462	nan
ICA	R	0.507	0.142	0.495	nan
ICA	R + N	0.505	0.165	0.500	nan
ICA	L	1.000	1.000	0.496	0.541
ICA	P(2)	0.843	0.844	0.447	nan
ICA	S	0.843	0.826	0.444	0.374
ICA	RB	0.856	0.874	0.496	0.590
Craddock	C + R + N	0.583	nan	0.472	nan
Craddock	R	0.535	0.472	0.509	0.449
Craddock	R + N	0.539	0.453	0.545	0.481
Craddock	L	1.000	1.000	0.451	0.502
Craddock	P(2)	0.865	0.881	0.500	0.663
Craddock	S	0.583	0.603	0.450	0.454
Craddock	RB	0.927	0.930	0.434	nan

Table 1. Logistic Regression (LR); RNN (R); CNN (C); NN (N); SVM (L)inear, (P)oly 2, (S)igmoid, (RB)F; Yellow = Best model; Blue = Best model - age; Red = Overfitting; Green = Weak performance

Results, cont.

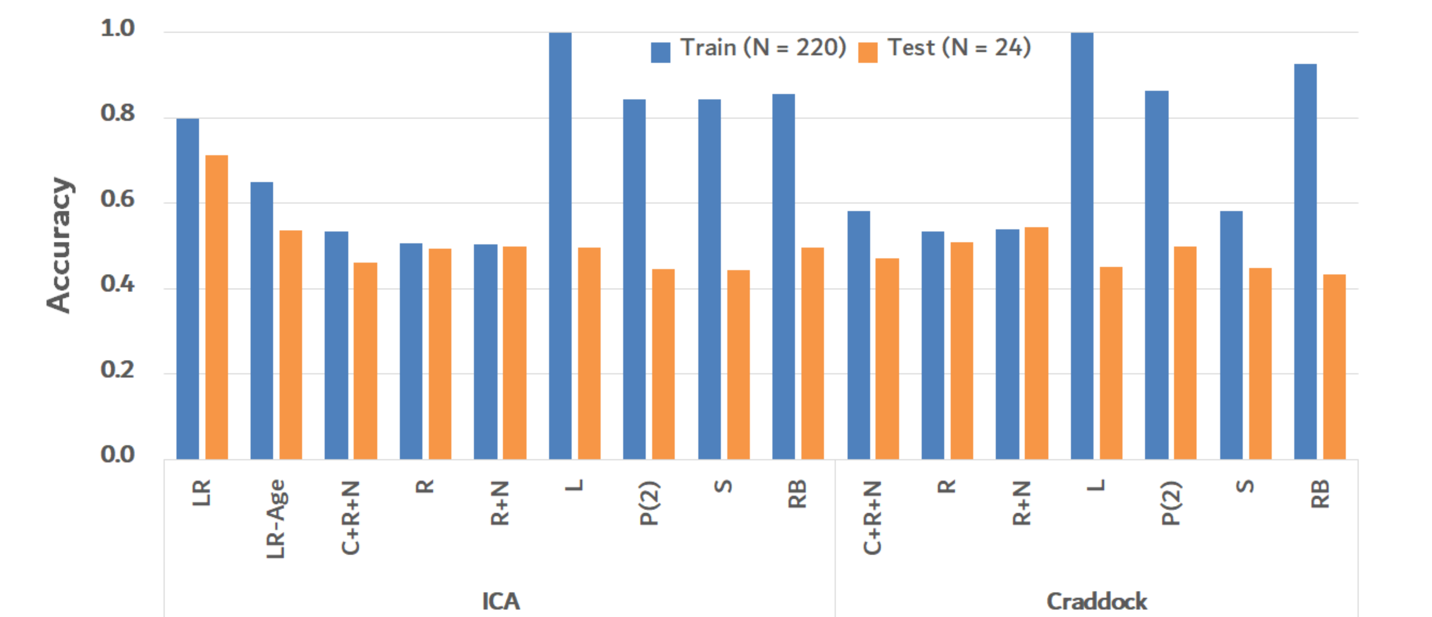

Figure 4. Performance plotted against various models

Discussion

- High variability of the deep-learning model makes adjustments of hyperparameters difficult
- Risk of overfitting deep-learning models and SVMs is high
- Small dynamic range of prediction values in deep-learning models suggests low sensitivity
- Many instances of 'nan' or bias only toward one class
- Overall suggests that our current amount of data may be insufficient to train deep-learning models
- Fairly good results from logistic regression alone when using derived features including demographics
- Removing of age as a feature decreases performance of logistic regression. Highlights the influence of demographic information toward making correct predictions

Future Steps

- Use transfer learning to circumvent small sample size
- Incorporate demographic data into deep-learning models
- Use different parcellation methods for pre-processing data
- Apply different models to condensed time-series data
- Consider different modes of preventing overfitting (regularization)

References

[1] Squeglia et al. (2014). The effect of alcohol use on human adolescent brain structures and systems. Handbook of Clinical Neurology, 125, 501–510.

[2] "NCANDA - National Consortium on Alcohol & Neurodevelopment in Adolescence." [Online]. Available: http://ncanda.org/.

[3] Craddock et al. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, *33*(8), 1914–1928.