Predicting Fluid Intelligence of Children using T1-weighted MR Images and a StackNet

Po-Yu Kao[1], Angela Zhang[1], Michael Goebel[1],
Jefferson W. Chen[2], and B.S. Manjunath[1]

[1] University of California, Santa Barbara, California, United States
{poyu_kao,manj}@ucsb.edu
[2] University of California, Irvine, California, United States

Abstract. In this work, we utilize T1-weighted MR images and StackNet to predict fluid intelligence in adolescents. Our framework includes feature extraction, feature normalization, feature denoising, feature selection, training a StackNet, and predicting fluid intelligence. The extracted feature is the distribution of different brain tissues in different brain parcellation regions. The proposed StackNet consists of three layers and 11 models. Each layer uses the predictions from all previous layers including the input layer. The proposed StackNet is tested on a public benchmark Adolescent Brain Cognitive Development Neurocognitive Prediction Challenge 2019 and achieves a mean absolute error of 82.42 on the combined training and validation set with 10-fold cross-validation.

Keywords: T1-weighted MRI · Fluid intelligence (Gf) · Machine learning · StackNet

1 Introduction

Fluid intelligence (Gf) refers to the ability to reason and to solve new problems independently of previously acquired knowledge. Gf is critical for a wide variety of cognitive tasks, and it is considered one of the most important factors in learning. Moreover, Gf is closely related to professional and educational success, especially in complex and demanding environments [7]. The ABCD Neurocognitive Prediction Challenge (ABCD-NP-Challenge 2019) provides 8556 subjects, age 9-10 years, with T1-weighted MR images and fluid intelligence which is withheld for testing subjects. The motivation of the ABCD-NP-Challenge 2019 is to discover the relationship between the brain and behavioral measures by leveraging the modern machine learning methods.

A few recent studies use structural MR images to predict fluid intelligence. Paul et al. [13] demonstrated that brain volume is correlated with quantitative reasoning and working memory. Wang et al. [19] proposed a novel framework for the estimation of a subject's intelligence quotient score with sparse learning based on the neuroimaging features. In this work, we utilize the T1-weighted MR images of adolescents to predict their fluid intelligence with a StackNet.

While whole brain volumes have been examined in relation to aspects of intelligence, to our knowledge there has been no previous work which examines the predictive ability of whole brain parcellation distributions for fluid intelligence. The main contributions of our work are two-fold: (1) to predict pre-residualized fluid intelligence based on parcellation volume distributions, and (2) to show the significance of the volume of each region on the overall prediction.

2 Materials and Methods

2.1 Dataset

The Adolescent Brain Cognitive Development Neurocognitive Prediction Challenge (ABCD-NP-Challenge 2019) [4,6,8,15,18] provides data for 3739 training subjects, 415 validation subjects and 4402 testing subjects (age 9-10 years). MR-T1 image is given for each subject, but the fluid intelligence scores are only provided for the training and validation subjects. MR-T1 images are distributed after skull-stripped and registered to the SRI 24 atlas [16] of voxel dimension $240 \times 240 \times 240$. In addition to the MR-T1 images, the distributions of gray matter, white matter, and cerebrospinal fluid in different regions of interest according to the SRI 24 atlas are also provided for all subjects. The fluid intelligence scores are pre-residualized on data collection site, sociodemographic variables and brain volume. The provided scores should, therefore, represent differences in Gf not due to these known factors.

2.2 StackNet Design

StackNet [10] is a computational, scalable and analytical framework that resembles a feed-forward neural network. It uses Wolpert's stacked generalization [20] in multiple levels to improve the accuracy of classifier or reduce the error of regressor. In contrast to the backward propagation used by feed-forward neural networks during the training phase, StackNet is built iteratively one layer at a time (using stacked generalization), with each layer using the final target as its target.

There are two different modes of StackNet: (i) each layer directly uses the predictions from only one previous layer, and (ii) each layer uses the predictions from all previous layers including the input layer that is called restacking mode. StackNet is usually better than the best single model contained in each first layer. However, its ability to perform well still relies on a mix of strong and diverse single models in order to get the best out of this meta-modeling methodology.

We adapt the StackNet architecture for our problem based on the following ideas: (i) including more models which have similar prediction performance, (ii) having a linear model in each layer (iii) placing models with better performance on a higher layer, and (iv) increasing the diversity in each layer. The resulting StackNet, shown in Fig. 1, consists of three layers and 11 models. These models include one Bayesian ridge regressor [9], four random forest regressors [1], three extra-trees regressors [5], one gradient boosting regressor [3], one kernel ridge regressor [12], and one ridge regressor. The first layer has one linear regressor and five ensemble-based regressors, the second layer contains one linear regressor and two ensemble-based regressors, and the third layer only has one linear regressor. Each layer uses the predictions from all previous layers including the input layer.

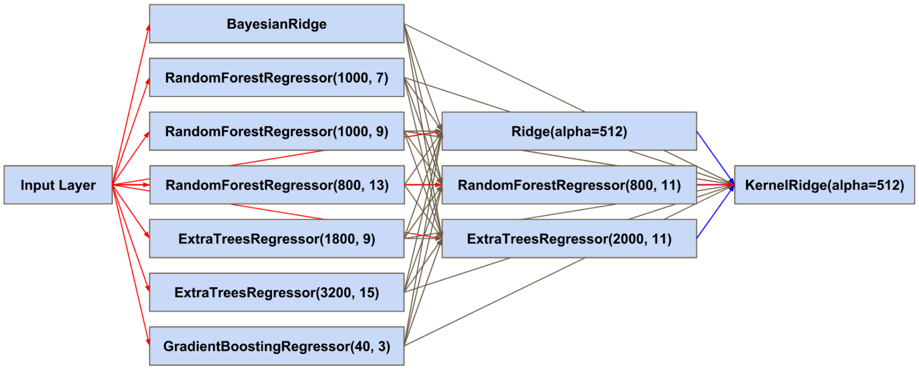
Input Layer
BayesianRidge
RandomForestRegressor(1000, 7)
RandomForestRegressor(1000, 9)
RandomForestRegressor(800, 13)
ExtraTreesRegressor(1800, 9)
ExtraTreesRegressor(3200, 15)
GradientBoostingRegressor(40, 3)
Ridge(alpha=512)
RandomForestRegressor(800, 11)
ExtraTreesRegressor(2000, 11)
KernelRidge(alpha=512)

Fig. 1. The architecture of proposed StackNet. For the ensemble-based regressor, the number of trees and the maximum depth of each tree are indicated in the first and second number, respectively.

2.3 Predicting Gf using Structural MR Images and StackNet

Fig. 2 shows the framework of predicting the fluid intelligence scores using MR-T1 images and a StackNet. The framework is implemented with the scikit-learn [2,14] Python library. In the training phase, features are extracted from the MR-T1 images of training and validation subjects. We then apply normalization and feature selection on the extracted features. In the end, these pre-processed features are used to train the StackNet in Fig. 1. In the testing phase, features are extracted from the MR-T1 images of testing subjects, and the same feature pre-processing factors are applied to these extracted features. Thereafter, the pre-processed features are used with the trained StackNet to predict the fluid intelligence of the testing subjects. Details of each step are described below.

The ABCD-NP-Challenge 2019 data includes pre-computed 122-dimension feature that characterizes the volume of brain tissues, i.e., gray matter, white matter, and cerebrospinal fluid, parcellated into SRI-24 [16] regions. The feature extracted for each subject is defined as $f_i(j)$ where i is the index of subject and $j \in \{1, 2, ..., 122\}$ is the index of feature dimension.

Normalization: We apply a standard score normalization on each feature dimension, $\overline{f}_i(j) = (f_i(j) - \mu(j))/\sigma(j)$ where i is the index of subject, j is the index

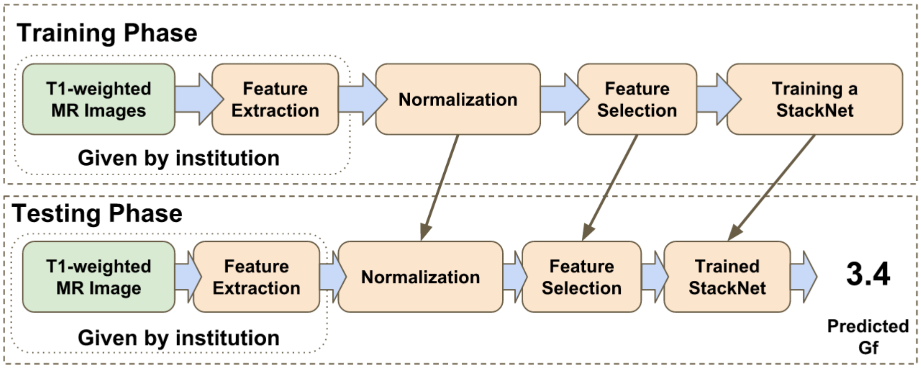

Fig. 2. The framework of predicting fluid intelligence using MR-T1 images and a StackNet.

of feature dimension, and $\overline{f}_i(j)$ and $f_i(j)$ are the normalized and raw feature dimension j of subject i, respectively. $\mu(j)$ and $\sigma(j)$ are the mean and standard deviation of feature dimension j, respectively.

Feature Selection: Feature selection consists of three steps: (i) reducing the noise of data and generating an accurate representation of data through principal component analysis (PCA) [17] with maximum-likelihood estimator [11] (ii) removing the feature dimensions with the low variance between subjects, and (iii) selecting 24 feature dimensions with the highest correlations to the ground-truth Gf scores through univariate linear regression tests. Thereafter, the feature dimensions shrink from 120 to 24.

Training a StackNet: Because the mean of the pre-residualized fluid intelligence for the training dataset ($\mu = 0.05, \sigma = 9.26$) and validation dataset ($\mu = -0.5, \sigma = 8.46$) are quite different, we combine these two datasets ($\mu = 0, \sigma = 9.19$) for hyperparemeter optimization and training a StackNet.

Predicting Fluid Intelligence: In the testing phase, we first apply the same pre-processing factors used in the training phase to the extracted features of testing subjects. We then use the trained StackNet with these pre-processed features to predict the fluid intelligence scores of testing subjects.

Evaluation Metric: The mean squared error (MSE) is used to calculate the error between the predicted Gf scores and the corresponding ground-truth Gf scores.

2.4 Computing Feature Importance

We would like to discover the correlation between the Gf score and the brain tissue volume in a region. Thus, we compute the importance of each feature dimension, and higher importance represents a higher correlation. However, after feature selection, the original data space of dimension 122 is projected and reduced to a new space of dimension 24. In this new space, we first compute the importance of each feature dimension and then backward propagate it to the original data space of dimension 122. The details are explained as follows.

After dimensionality reduction, we obtain the individual correlations between the remaining 24 feature dimensions and the ground-truth Gf scores. These correlations are first converted to F values and then normalized w.r.t. the total F values of feature dimensions, i.e., $\overline{F}_1+\overline{F}_2+...+\overline{F}_{24}=1$, where $\overline{F}_k$ is the normalized F value of feature dimension $k \in \{1,2,...,24\}$. These normalized F values are used to build a normalized F vector as $\overline{\boldsymbol{F}}_{1\times 24}=[\overline{F}_1,\overline{F}_2,...,\overline{F}_{24}]$. We then use the corresponding eigenvectors and eigenvalues from the PCA transformation to build two matrices, $\boldsymbol{U}_{122\times 24}=[\vec{u}_1,\vec{u}_2,...,\vec{u}_{24}]$ and $\boldsymbol{\Lambda}_{1\times 24}=[\lambda_1,\lambda_2,...,\lambda_{24}]$, where $\vec{u}_k$ and λ_k are the corresponding eigenvector and eigenvalue for $\overline{F}_k$, respectively. The dimension of $\vec{u}_k$ is 122. We also normalize the eigenvalue vector w.r.t. the total value of eigenvalues, i.e., $\overline{\boldsymbol{\Lambda}}_{1\times 24}=\boldsymbol{\Lambda}_{1\times 24}/\lambda_t$, where $\lambda_t=\lambda_1+\lambda_2+...+\lambda_{24}$. The normalization for eigenvalues is required to ensure that they have the same scale as the F values. Thereafter, we use $\overline{\boldsymbol{F}}$, $\overline{\boldsymbol{\Lambda}}$ and $\boldsymbol{U}$ to build the feature importance matrix $\boldsymbol{I}_{122\times 24}=[\overline{F}_1\overline{\lambda}_1\vec{u}_1,\overline{F}_2\overline{\lambda}_2\vec{u}_2,...,\overline{F}_{24}\overline{\lambda}_{24}\vec{u}_{24}]$. In the end, we sum up the absolute value of each element in every row of the matrix $\boldsymbol{I}_{122\times 24}$,

$$\vec{I}_{122\times 1}=\sum_{n=1}^{24}|\overline{F}_n\overline{\lambda}_n\vec{u}_n|$$

$\vec{I}_{122\times 1}$ is the feature importance vector in the original data space, and we also normalize it w.r.t. its total importance and rescale it,

$$\vec{I}_{nrm}=100\cdot\vec{I}_{122\times 1}/\sum_{m=1}^{122}\vec{I}_m$$

Now, $\vec{I}_{nrm}$ is the normalized feature importance vector in the original data space of dimension 122, and each value of this vector represents the importance of a brain tissue volume in a region for the task of predicting the Gf scores. Higher importance represents higher correlation to the Gf score.

3 Results and Discussion

We examine the Gf prediction performance of individual models and StackNet on the combined dataset with 10-fold cross-validation, with the quantitative results shown in Table 1. The baseline is calculated by assigning the mean fluid intelligence ($\mu=0$) to every subject in the combined dataset. From Table 1, the performance of each model is better than the baseline of guessing the mean every subject, and the performance of the StackNet is better than every individual model within itself because it takes advantage of stacked generalization.

Table 1. The quantitative results of 11 models and StackNet with 10-fold cross-validation on the combined dataset. The bold number highlights the best performance.

Model	MSE
Baseline	84.50
BayesianRidge	82.62
Ridge(alpha=512)	82.61
KernelRidge(alpha=512)	82.61
GradientBoostingRegressor(n_estimators=40, max_depth=3)	83.60
RandomForestRegressor(n_estimators=1000, max_depth=7)	83.07
RandomForestRegressor(n_estimators=1000, max_depth=9)	83.09
RandomForestRegressor(n_estimators=800, max_depth=11)	83.07
RandomForestRegressor(n_estimators=800, max_depth=13)	83.11
ExtraTreesRegressor(n_estimators=1800, max_depth=9)	83.16
ExtraTreesRegressor(n_estimators=2200, max_depth=11)	83.10
ExtraTreesRegressor(n_estimators=3200, max_depth=15)	83.16
StackNet	**82.42**

The proposed StackNet in Fig. 1 is different from the StackNet which is used to report the MSE on the validation leader board. The StackNet used to report the MSE on the validation leader board has two layers and 8 models, and it achieves an MSE of 84.04 and 70.56 (rank 7 out of 17 teams) on the training and validation set, respectively. However, we noticed that statistics between the training set and validation are quite different, so we decided to combine these two datasets and work on this combined dataset ($\mu = 0$ and $\sigma = 9.19$) using 10-fold cross-validation. In addition, we also ensured that the mean and standard deviation of each fold is similar to the mean and standard deviation of combined dataset. The source code is available on GitHub[1].

Second, we compute the importance of each dimension of the extracted feature by leveraging the F score from feature selection and eigenvectors and eigenvalues from PCA as described in Section 2.4. Each dimension of the extracted feature corresponds to the volume of a certain type of brain tissue in a certain region. Table 2 and Table 3 show the top 10 most and least important feature dimensions for the task of predicting Gf, respectively, and higher importance represents higher correlation to the Gf scores.

In conclusion, we demonstrate that the proposed StackNet with the distribution of different brain tissues in different brain parcellation regions has the potential to predict fluid intelligence in adolescents.

[1] https://github.com/pykao/ABCD-MICCAI2019

Table 2. The top 10 most important variables for the task of predicting Gf

Variable description	Importance
Pons white matter volume	1.18
Right insula gray matter volume	1.13
Right inferior temporal gyrus gray matter volume	1.11
Corpus callosum white matter volume	1.08
Cerebellum hemisphere white matter right volume	1.07
Cerebellum hemisphere white matter left volume	1.06
Left inferior temporal gyrus gray matter volume	1.06
Left insula gray matter volume	1.06
Left superior frontal gyrus, orbital part gray matter volume	1.05
Left opercular part of inferior frontal gyrus gray matter volume	1.05

Table 3. The top 10 least important variables for the task of predicting Gf

Variable description	Importance
Right hippocampus gray matter volume	0.53
Right amygdala gray matter volume	0.54
Left hippocampus gray matter volume	0.56
Right caudate nucleus gray matter volume	0.58
Right lobule IX of cerebellar hemisphere volume	0.60
Right lobule X of cerebellar hemisphere (flocculus) volume	0.60
Left lobule X of cerebellar hemisphere (flocculus) volume	0.61
Right superior parietal lobule gray matter volume	0.61
Left middle temporal pole gray matter volume	0.63
Left lobule IX of cerebellar hemisphere volume	0.63

Acknowledgement

This research was partially supported by a National Institutes of Health (NIH) award # 5R01NS103774-02.

References

1. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
2. Buitinck, L., et al.: API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning. pp. 108–122 (2013)
3. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
4. Garavan, H., et al.: Recruiting the abcd sample: Design considerations and procedures. Developmental cognitive neuroscience **32**, 16–22 (2018)
5. Geurts, P., et al.: Extremely randomized trees. Machine learning **63**(1), 3–42 (2006)
6. Hagler, D.J., et al.: Image processing and analysis methods for the adolescent brain cognitive development study. bioRxiv (2018). https://doi.org/10.1101/457739
7. Jaeggi, S.M., et al.: Improving fluid intelligence with training on working memory. Proceedings of the National Academy of Sciences **105**(19), 6829–6833 (2008)

8. Luciana, M., et al.: Adolescent neurocognitive development and impacts of substance use: Overview of the adolescent brain cognitive development (abcd) baseline neurocognition battery. Developmental cognitive neuroscience **32**, 67–79 (2018)
9. MacKay, D.J.: Bayesian interpolation. Neural computation **4**(3), 415–447 (1992)
10. Michailidis, M.: Stacknet, meta modelling framework. https://github.com/kaz-Anova/StackNet (2017)
11. Minka, T.P.: Automatic choice of dimensionality for pca. In: Advances in neural information processing systems. pp. 598–604 (2001)
12. Murphy, K.P.: Machine learning: a probabilistic perspective (2012)
13. Paul, E.J., et al.: Dissociable brain biomarkers of fluid intelligence. NeuroImage **137**, 201–211 (2016)
14. Pedregosa, F., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
15. Pfefferbaum, A., et al.: Altered brain developmental trajectories in adolescents after initiating drinking. American journal of psychiatry **175**(4), 370–380 (2017)
16. Rohlfing, T., et al.: The sri24 multichannel atlas of normal adult human brain structure. Human brain mapping **31**(5), 798–819 (2010)
17. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61**(3), 611–622 (1999)
18. Volkow, N.D., et al.: The conception of the abcd study: From substance use to a broad nih collaboration. Developmental cognitive neuroscience **32**, 4–7 (2018)
19. Wang, L., et al.: Mri-based intelligence quotient (iq) estimation with sparse learning. PloS one **10**(3), e0117295 (2015)
20. Wolpert, D.H.: Stacked generalization. Neural networks **5**(2), 241–259 (1992)