An AutoML Approach for the Prediction of Fluid Intelligence From MRI-Derived Features

Sebastian Pölsterl, Benjamín Gutiérrez-Becker, Ignacio Sarasua, Abhijit Guha Roy, and Christian Wachinger

Artificial Intelligence in Medical Imaging (AI-Med),
Department of Child and Adolescent Psychiatry,
Ludwig Maximilian Universität, Munich, Germany
{firstname}@ai-med.de

Abstract. We propose an AutoML approach for the prediction of fluid intelligence from T1-weighted magnetic resonance images. We extracted 122 features from MRI scans and employed Sequential Model-based Algorithm Configuration to search for the best prediction pipeline, including the best data pre-processing and regression model. In total, we evaluated over 2600 prediction pipelines. We studied our final model by employing results from game theory in the form of Shapley values. Results indicate that predicting fluid intelligence from volume measurements is a challenging task with many challenges. We found that our final ensemble of 50 prediction pipelines associated larger parahippocampal gyrus volumes with lower fluid intelligence, and higher pons white matter volume with higher fluid intelligence.

1 Introduction

This paper describes our method submitted to the ABCD Neurocognitive Prediction Challenge 2019. The task of the challenge is to predict fluid intelligence solely from structural T1-weighted magnetic resonance images (MRI). The challenge uses data from the Adolescent Brain Cognitive Development(ABCD) Study.

In this approach, we first extract features from MRI scans and then use an automated machine learning approach for the prediction. For the feature extraction, we use volume measurements as provided by the challenge's organizers. For the prediction, we use an automated machine learning (AutoML) approach, as determining a good machine learning pipeline is a tedious and error-prone task for humans. A typical ML pipeline includes various types of preprocessing that can be applied to input features. Afterwards, an appropriate classifier needs to be selected and the optimal hyperparameters selected to achieve high predictive performance. The goal of AutoML is to automate the whole machine learning pipeline. A recent overview of AutoML approaches together with an analysis of the results of ChaLearn AutoML Challenges over the last four years is given in [5]. AutoML has not yet been widely explored in the medical field, with PubMed listing only four articles [7,14,10,1]; none of which study MRI or neuroscience.

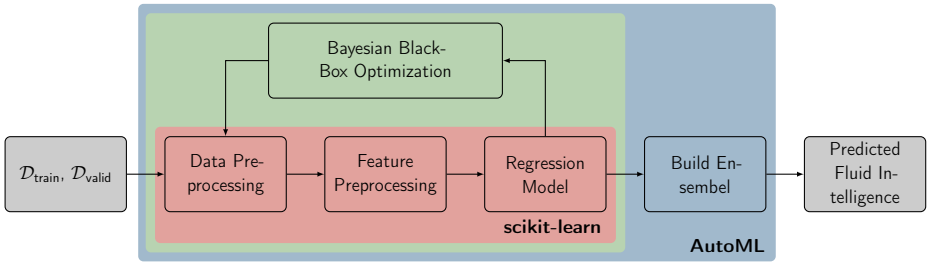

Fig. 1. Overview of our proposed AutoML pipeline for the prediction of fluid intelligence from T1-weighted MRI scans.

2 Data

Data was provided by The Adolescent Brain Cognitive Development(ABCD) Study [13], which recruited children aged 9-10. Participants were given access to T1-weighted MRI scans from 3,736 children for training, 415 children for validation, and 4,402 children for testing. Fluid intelligence scores were residualized to account for confounding due to sex at birth, ethnicity, highest parental education, parental income, parental marital status, and image acquisition site. Residualized fluid intelligence scores were provided for the training and validation data, but not for the test data. All data was obtained from the National Institute of Mental Health Data Archive.[1]

3 Methods

Our proposed pipeline for the prediction of fluid intelligence from T1-weighted MRI scans builds on the Automated Machine Learning (AutoML) framework summarized in fig. 1. Scans were acquired according to the acquisition protocol of the Adolescent Brain Cognitive Development (ABCD) study protocol.[2] For parcellation of the brain and the estimation of volume of each region of interest, we relied on the work of the challenge's organizers.

3.1 Feature-Preprocessing

We used volume measurements of 122 regions of interest extracted by the challenge's organizers from each T1-weighted MRI scan based on the SRI24 atlas [15].[3] We normalized all volume measurements while accounting for outliers by subtracting the median and dividing by the range between the 5% and 95% percentile. Thus, we reduce the impact of outliers and still obtain approximately centered

[1] https://nda.nih.gov/edit_collection.html?id=3104
[2] https://abcdstudy.org/images/Protocol_Imaging_Sequences.pdf
[3] See https://nda.nih.gov/data_structure.html?short_name=btsv01 for a full list of volumes.

features with equal scale. Finally, the provided residualized fluid intelligence scores in the training data where standardized to zero mean and unit variance; the same transformation as derived from the training data was applied to features and scores in the validation and test data. Additional pre-processing steps were selected without human interaction as described in the next section.

3.2 Automated Machine Learning

For the prediction of residualized fluid intelligence score, we used automated machine learning that leverages recent advances in Bayesian optimization, meta-learning, and ensemble construction. For every machine learning task, the fundamental problem is to decide which machine learning algorithm to use and whether and how to pre-process features. This task is extremely challenging, because there is no single algorithm that performs best on all datasets and the performance of machine learning methods depends to a large extent on their hyper-parameter settings, which can vary from one task to the next. Here, we use AutoML for the prediction of the residualized fluid intelligence score by producing test set predictions without human input within a given computational budget. Specifically, we employ Combined Algorithm Selection and Hyperparameter (CASH) optimization [3].

Let $\mathcal{A} = \{A^{(1)}, \ldots, A^{(R)}\}$ be a set of machine learning algorithms, and $\Lambda^{(j)}$ be the domain of the hyper-parameters of each algorithm. Further, we define $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ to be the training set, which we split into K cross-validation folds to obtain $\{\mathcal{D}^{(1)}_{\text{train}}, \ldots, \mathcal{D}^{(K)}_{\text{train}}\}$ and $\{\mathcal{D}^{(1)}_{\text{valid}}, \ldots, \mathcal{D}^{(K)}_{\text{valid}}\}$ with $\mathcal{D}^{(k)}_{\text{train}} = \mathcal{D}_{\text{train}} \backslash \mathcal{D}^{(k)}_{\text{valid}}$. For a particular hyper-parameter configuration $\boldsymbol{\Theta}$, we solve the CASH optimization problem

$$\underset{A^{(j)} \in \mathcal{A}, \boldsymbol{\Theta} \in \Lambda^{(j)}}{\operatorname{argmin}} \quad \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{D}^{(k)}_{\text{valid}}|} \sum_{i=1}^{|\mathcal{D}^{(k)}_{\text{valid}}|} \left(y_i - \hat{f}_{A^{(j)}_{\boldsymbol{\Theta}}}(\mathbf{x}_i \,|\, \mathcal{D}^{(k)}_{\text{train}}), \right)^2 \tag{1}$$

where $\hat{f}_{A^{(j)}_{\boldsymbol{\Theta}}}(\mathbf{x}_i \,|\, \mathcal{D}^{(k)}_{\text{train}})$ denotes the prediction on the validation set of model $A^{(j)}$ with hyper-parameters $\boldsymbol{\Theta}$ and trained on $\mathcal{D}^{(k)}_{\text{train}}$. This optimization problem can be solved via Sequential Model-based Algorithm Configuration (SMAC), a technique for Bayesian black-box optimization that uses a random-forest-based surrogate model [6]. The main idea of SMAC is to use the surrogate model to predict an algorithm's performance for a given hyper-parameter optimization. It is able to interpolate the performance of algorithms between observed hyper-parameter configurations and of extrapolating to previously unseen regions of hyper-parameter space. Thus, it enables us to focus on promising hyper-parameter configurations.

We employed the auto-sklearn toolkit (version 0.5.0), which for a given user-provided computational budget in terms of run time and memory, auto-sklearn searches for the best machine learning pipeline to predict the residualized fluid intelligence score by combining components of the scikit-learn machine learning

framework (version 0.18.2) [12]. Figure 1 depicts an overview of the AutoML framework. For data preprocessing, AutoML can choose from 11 algorithms for data transformations, such as principal component analysis. For feature preprocessing 6 feature-wise transformations are available, such as transforming each feature to have zero mean and unit variance. Finally, AutoML can choose from 13 regression models. After evaluating various machine learning pipelines, comprising data transformations, feature transformations, and regression model, the best M pipelines are combined via ensemble selection [2] to form the final prediction model. We used a budget that consisted of a total run time of 40 hours, where each pipeline was limited to 6 minutes and 4 GB of memory. The final ensemble size was $M = 50$.

3.3 Feature Importance

While complex prediction pipelines are potentially powerful, their black-box nature is often a barrier for employing such a model in clinical research. We use Shapley values to explain the predictions of our final ensemble of prediction pipelines. Shapley values are a classic solution in game theory to determine the distribution of credits to players participating in a cooperative game [16,17]. They have first been proposed for linear models in the presence of multicollinearity [8]. A Shapley value assigns an importance value ϕ_j to each feature j that reflects its effect on the model's prediction. To compute this effect, retraining the model $f(\cdot)$ on all possible feature subsets $\mathcal{S} \subseteq \mathcal{F} \backslash \{j\}$ of all features $\mathcal{F}$ is necessary. Given a feature vector $\mathbf{x} \in \mathbb{R}^{|\mathcal{F}|}$, the j-th Shapley value can then be computed as the weighted average of all prediction differences:

$$\phi_j(\mathbf{x}) = \sum_{\mathcal{S} \subseteq \mathcal{F} \backslash \{j\}} \frac{|\mathcal{S}|!(|\mathcal{F}| - |\mathcal{S}| - 1)!}{|\mathcal{F}|!} \left(\hat{f}_{\mathcal{S} \cup \{j\}}(\mathbf{x}^{\mathcal{S} \cup \{j\}}) - \hat{f}_{\mathcal{S}}(\mathbf{x}^{\mathcal{S}}) \right), \qquad (2)$$

where $\hat{f}_S(\mathbf{x}^{\mathcal{S}})$ denotes the prediction of a model trained and evaluated on the feature subset $\mathcal{S}$. The exact computation of Shapley values requires evaluating all $2^{|\mathcal{F}|}$ possible feature subsets, which is only reasonable when data consists of not more than a few dozen features. To address this problem, we employ the recently proposed SHAP (SHapley Additive exPlanations) values, which belong to the class of additive feature importance measures [9]. The exact computation of SHAP values is prohibitive, therefore we approximate SHAP values using the model-agnostic KernelSHAP approach proposed in [9]. To obtain a global measure of feature importance, we compute the average magnitude of SHAP values across all N subjects in the data:

$$\bar{\phi}_j = \frac{1}{N} \sum_{i=1}^{N} |\phi_j(\mathbf{x}_i)|. \qquad (3)$$

4 Results

The performance of the final ensemble is summarized in table 1. It reveals that predicting residualized fluid intelligence from MRI-derived volume measurements

Table 1. Performance on training, validation and test set. MSE: mean squared error. MAE: mean absolute error.

	Subjects	MSE	MAE
Training	3,736	17.027	3.206
Validation	415	69.586	6.498
Test	4,402	94.010	—

Table 2. Summary of evaluated machine learning pipelines.

Description	Number
Algorithm runs	2608
Successful algorithm runs	2179
Crashed algorithm runs	6
Algorithms that exceeded the time limit	198
Algorithms that exceeded the memory limit	225

is a challenging task. In particular, the proposed model struggles to reliably predict residualized fluid intelligence at the extremes of the distribution, i.e., very low or very high values. Consequently, we observe a relatively high mean squared error, which is an order of magnitude larger than the mean absolute error. Moreover, the large difference between the performance on the training data and the validation data indicates that overfitting seems to be an issue.

In total, we evaluated 2,608 machine learning algorithms (see table 2). The components of our final ensemble of 50 machine learning pipelines is summarized in table 3. Principal component analysis [11] was selected most often (15 times) for data pre-processing. The final ensemble was comprised of linear and non-linear regression models with ensembles of randomized regression trees [4] being selected most frequently (14 times). Looking at the top-performing pipelines in the ensemble, we noticed that combining principal component analysis with a tree-based ensemble was a frequently selected combination (5 out of the top 10 performing pipelines).

Next, we inspected which MRI-derived feature the model deems most important by computing SHAP values for each feature and subject in the training data. Figure 2 lists the top 20 features by mean absolute SHAP value ϕ. The top ranked feature is pons white matter volume ($\phi = 0.0183$), followed by left parahippocampal gyrus volume ($\phi = 0.0155$), and left lateral ventricle cerebral spinal fluid volume ($\phi = 0.0148$). However, we note that individual SHAP values are rather small, which is evidence that fluid intelligence is not strongly influenced by a single brain region, but a complex inter-relationship between different regions. Individual, subject-specific SHAP values depicted in fig. 2b indicate that larger left and right parahippocampal gyrus volume are associated with a

Table 3. Overview of selected components in the final ensemble of $M = 50$ pipelines selected by AutoML. Each pipeline consists of one data preprocessing step, one feature preprocessing step, and one regressor.

	Algorithm	Count
Data preprocessing	PCA	15
	Feature agglomeration	8
	Kernel PCA	8
	No preprocessing	6
	ICA	3
	Polynomial features	3
	Feature selection (Extra trees)	3
	Feature selection (percentile)	2
	Random trees embedding	1
	Nystroem sampler	1
Feature prepr.	Standardize	14
	None	13
	Normalize	7
	Min-max	6
	Quantile transformer	6
	Robust scaler	4
Regressor	Extra trees	14
	SGD	10
	Random forest	9
	Adaboost	5
	Decision tree	4
	Ridge regression	3
	Linear SVR	2
	ARD regression	1
	Gradient boosting	1
	k nearest neighbors	1

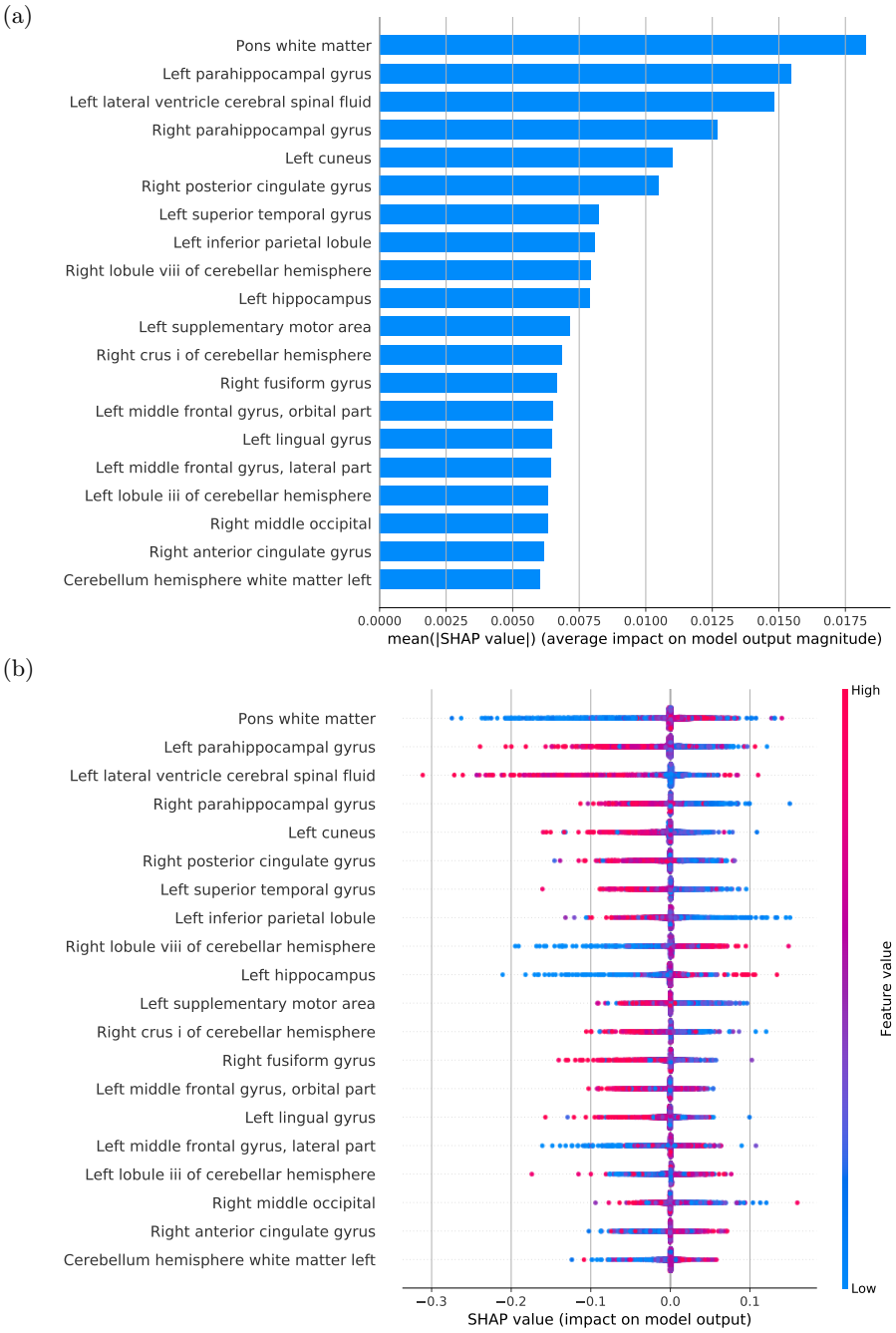

Fig. 2. (a) Top 20 features sorted by mean absolute SHAP value $\bar{\phi}_j$. (b) SHAP values of top 20 features for each subject in the training data. In each row SHAP values ϕ_j for each subject are plotted horizontally, stacking vertically to avoid overlap. Each dot is colored by the value of that feature, from low (blue) to high (red). If the impact of the feature on the model's prediction varies smoothly as its value changes then this coloring will also appear smooth.

decrease in fluid intelligence, while larger pons white matter volume is associated with an increase.

5 Conclusion

We proposed an AutoML model for the prediction of fluid intelligence from T1-weighted magnetic resonance images based on more than 2,600 evaluated machine learning pipelines. Our experiments demonstrate that it is challenging for our ensemble to reliably predict fluid intelligence from MRI scans. In particular, errors on the validation and test data were more than four times higher than on the training data, which is evidence for overfitting. We analyzed the final model's predictions using SHAP values. Results revealed that top ranked features still explain only a small fraction of the fluid intelligence score. Therefore, we concluded that current features derived from MRI are insufficient to robustly measure fluid intelligence. While current features are generic descriptors of the brain anatomy, we believe future research should focus on deriving tailor-made features from MRI, specific to the prediction of fluid intelligence, which could then be used to improve our understanding of the neurobiology underlying fluid intelligence.

Acknowledgements This research was partially supported by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

References

1. Barreiro, E., Munteanu, C.R., Cruz-Monteagudo, M., Pazos, A., González-Díaz, H.: Net-net auto machine learning (automl) prediction of complex ecosystems. Scientific reports **8**(1), 12340 (2018)
2. Caruana, R., Niculescu-Mizil, A.: Ensemble selection from libraries of models. Proceedings of the 21st International Conference on Machine Learning (ICML '04) p. 18 (2004). https://doi.org/10.1145/1015330.1015432
3. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Efficient and robust automated machine learning. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 2962–2970. Curran Associates, Inc. (2015)
4. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning **63**(1), 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1
5. Guyon, I., Sun-Hosoya, L., Boullé, M., Escalante, H.J., Escalera, S., Liu, Z., Jajetic, D., Ray, B., Saeed, M., Sebag, M., et al.: Analysis of the automl challenge series 2015–2018. In: Automated Machine Learning, pp. 177–219. Springer (2019)
6. Hutter, F., Hoos, H.H., Leyton-Brown, K.: Sequential Model-Based Optimization for General Algorithm Configuration. In: Coello, C.A.C. (ed.) Learning and Intelligent Optimization. pp. 507–523. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
7. Le, T.T., Fu, W., Moore, J.H.: Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics (2019)

8. Lipovetsky, S., Conklin, M.: Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry **17**(4), 319–330 (2001). https://doi.org/10.1002/asmb.446
9. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30. pp. 4765–4774. Curran Associates, Inc. (2017)
10. Orlenko, A., Moore, J.H., Orzechowski, P., Olson, R.S., Cairns, J., Caraballo, P.J., Weinshilboum, R.M., Wang, L., Breitenstein, M.K.: Considerations for automated machine learning in clinical metabolic profiling: altered homocysteine plasma concentration associated wtih metformin exposure. In: Pac Symp Biocomput. vol. 23. World Scientific (2017)
11. Pearson, K.: On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11), 559–572 (1901). https://doi.org/10.1080/14786440109462720
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
13. Pfefferbaum, A., Kwon, D., Brumback, T., Thompson, W.K., Cummins, K., Tapert, S.F., Brown, S.A., Colrain, I.M., Baker, F.C., Prouty, D., De Bellis, M.D., Clark, D.B., Nagel, B.J., Chu, W., Park, S.H., Pohl, K.M., Sullivan, E.V.: Altered brain developmental trajectories in adolescents after initiating drinking. American Journal of Psychiatry **175**(4), 370–380 (2018). https://doi.org/10.1176/appi.ajp.2017.17040469
14. Puri, M.: Automated machine learning diagnostic support system as a computational biomarker for detecting drug-induced liver injury patterns in whole slide liver pathology images. Assay and drug development technologies (2019)
15. Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A.: The SRI24 multichannel atlas of normal adult human brain structure. Human Brain Mapping **31**(5), 798–819 (2010). https://doi.org/10.1002/hbm.20906
16. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games **2**(28), 307–317 (1953)
17. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems **41**(3), 647–665 (2014). https://doi.org/10.1007/s10115-013-0679-x