Schizophrenia classification from brain anatomical data using machine learning

An Xuelong[1]

[1] University College London, UK
ucabxan@ucl.ac.uk

Abstract. We benchmark several machine learning architectures across different cross-validation strategies to find the optimal hyperparameter configuration to train and validate them to classify schizophrenia from brain anatomical data. We obtain a regularized logistic regression with hyperparameters found in a cross-validation paradigm guided by Bayesian optimization that can achieve a validation F1-score of 0.77 ± 0.07 and a balanced accuracy score of 0.79 ± 0.06, the highest amongst surveyed models. Our simple linear model can thus aid clinicians in the fast and accurate diagnosis of schizophrenia from brain anatomical data.

Keywords: Logistic regression, Bayesian optimization, Schizophrenia.

1 Introduction

Schizophrenia is a complex psychiatric disorder characterized by debilitating symptoms, such as cognitive deficits, hallucinations, and delusions. It's estimated that 3% of the worldwide population suffers from it, which is concerning given that the mortality rate of schizophrenic patients is higher compared to the general populace, with a prominent cause being suicide [4]. This underscores the importance of its timely and accurate diagnosis for early management. Prior research suggests that neurobiological factors, particularly changes in grey matter (GM) volume, shed insight into the pathophysiology of schizophrenia. Voxel-based morphometry (VBM) studies have also revealed complex patterns of brain atrophy particular to schizophrenic individuals [4]. Together, these findings underscore the potential of neuroimaging biomarkers in enhancing our diagnostic accuracy of schizophrenia. Further motivation behind classifying schizophrenia using neuroimaging data stems from the need for objective biomarkers underlying its pathophysiology. Diagnostic practices which primarily rely on clinical assessments of symptomatology are often subjective and unreliable due to the high inter-individual variability of schizophrenic patients, as well as the great phenotypical overlap with other conditions like bipolar disorder (BD) [4] and multiple sclerosis (MS). Thus, there is the need to "deconstruct the psychosis spectrum" by integrating multidimensional data for deep phenotyping to refine diagnostic boundaries [4]. Machine learning (ML) offers a testbed of powerful statistical models for analyzing complex and high-dimensional data, such as neuroimaging datasets. Adopting ML for classifying schizophrenia is driven by their flexibility to integrate multimodal data, achieve high diagnostic accuracy by learning intricate patterns in the data, and scalability for analyzing the sheer-scale of available neuroimaging data . These qualities can collectively alleviate the healthcare burden surrounding schizophrenic patients, namely, potential for early, accurate detection, and reduced diagnosis time [4].

In this paper, we benchmark different ML architectures and assess their generalizability and scalability in terms of computational time for training to classify schizophrenia by subjecting them to various cross-validation pipelines for hyperparameter tuning. By doing so, we aspire to train a highly accurate and scalable model to contribute to the early and accurate detection of schizophrenia, paving the way for timely intervention for individuals affected by this disorder.

2 Methods

We obtain a schizophrenia dataset from Neurospin - Université Paris-Saclay [5]. It comprises of brain anatomical data of 410 patients in the training set and 103 in the test set. For each patient, flattened 331695 3D voxel-based morphometry (VBM) measurements are collected (which we now refer to as "high" dimensional features), along with 284 features measuring detailed information on regions of interest (ROIs) of grey matter (GM) scaled for the total intracranial volume (TIV) (which we refer to as "low" dimensional features). The combination of both types of data is referred to as "all" dimensional features. This dataset can potentially capture the diffuse and complex pattern of brain atrophy associated with schizophrenia, thereby facilitating the development of a robust predictor of schizophrenia.

We benchmark a precomputed, radial basis function (RBF) kernel-based support vector machine (SVM) and random forest (RF) motivated by the extensive prior work which uses these models for classifying schizophrenic patients [5]. This is because SVMs are simple to implement, and powerful predictors that can capture non-linear relationships between label and input features, where we can flexibly customize a kernel function to improve the boundary's expressivity, as well as speed-up training time. RF are ensemble methods which are also non-linear by design, and which final prediction is the average of each trained decision trees, thus improving the expressivity of the prediction. We also benchmark logistic regression model as it's the traditional baseline for binary classification, and flexible to implement.

We assess the 3 models' generalizability, i.e., their ability to achieve a high classification accuracy on unseen data, through cross-validation (CV) accompanied with hyperparameter tuning, with their grids of values for each model shown in **Table 1**. We subject the models under 3 CV strategies:

1. Common K-Fold, which consists of randomly splitting the training set into training and validation sets for $k = 5$ different times and randomly searching over a grid of hyperparameters, outputting the hyperparameters yielding the highest average CV score. These hyperparameters are interpreted as yielding a model with high generalizability.
2. Group stratified K-Fold, which is like the above, where the main difference lies in that each k^{th} fold, input features are stratified based on patients' sex, which is to capture the known sex differences in schizophrenia [5].
3. Bayesian optimization K-Fold, which is like 1.), where the main difference lies in that hyperparameters are searched using a Gaussian process's acquisition function rather than randomly.

We also measure scalability by training each model under each CV for 3 dimensionalities of the data: low (ROI), high (VBM), and all (ROI + VBM) and measure total CV time, in seconds, to compare amongst them. An extreme spike in total CV time as dimensionality increases indicates a model's poor scalability.

Our main evaluation metrics are F1 score (F1) and balanced accuracy (BACC). The former is the harmonic mean between sensitivity and specificity, which is important for a ML with clinical implications as a false negative or false positive can lead to delayed management or patient's anxiety. Thus, a model's wrong predictions must be accounted for and penalized, unlike in alternative metrics like accuracy. A high F1 score thus indicates a model not only identifies robustly a schizophrenic patient from healthy, but also who isn't. We also record BACC to adjust accuracy by the class imbalance brought by known differences in schizophrenia incidence such as caused by sex [5], thus BACC is especially relevant in the group stratified CV pipeline.

In total, we run $3\ mdls \times 3\ CV\ \times 3\ dims = 27$ experiments.

Table 1. Hyperparameter grid per model supplied to each CV strategy. For logistic regression, we initialize it with the 'liblinear' solver and maximum iteration of 10000. The SVM is initialized with a precomputed RBF kernel. The RF is initialized with 100 estimators. All models are set with a random seed of 42 for reproducibility.

Model	Hyperparameter and grid

LogReg	Regularizer C: log-uniform space {1e-3, 1}
	Penalty: {L1, L2}

SVM	Reg. C: log-uniform space {1e-6, 1e6} for Bayesian search

RF	Max. depth: Integer range {1, 10}
	Min. samples to split: Integer range {2, 10}

3 Results

Our highest performing model is a logistic regressor with optimal hyperparameters [C: 0.037, l1 penalty] found through Bayesian search optimization which can achieve the highest CV BACC of 0.79 ± 0.06 trained on all dimensions of the data (see **Fig. 1**). This is the model we submit to the RAMP competition and achieve results of an ROC-AUC of 0.86 and (test) BACC of 0.77. Bayesian search optimization on all dimensions took ~13 mins to complete, and inference time can be done within ~5 seconds. Despite helping us find the best hyperparameter configuration for improving generalizability, we note that Bayesian search CV is also the longest to run compared to the remaining 2 CV strategies, where we can't see much difference between both (see **Figure 2**). **Table 2** summarizes our quality assessments of each CV strategy, where we recommend machine learners to opt for Bayesian CV for guided search over hyperparameters at the expense of consuming a lot of computational sources. Otherwise, either K-fold or group stratified K-fold can yield good hyperparameters.

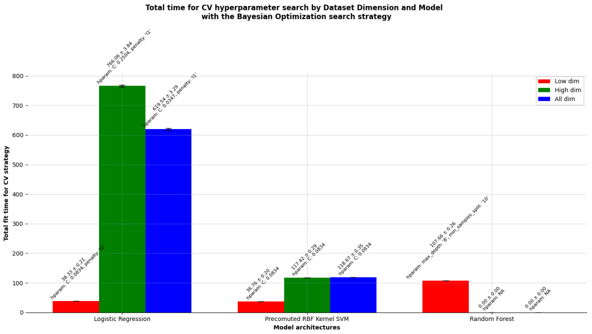


Table 2. Quality assessments of total CV time and achievable performance from hyperparameters found in each CV strategy. A "long" total CV time indicates the CV pipeline can be ran within ~5 mins. A 'high achievable performance indicates the CV strategy can find hyperparameters that yield a 'good' performance as per **Table 3** standards. We note that the total CV time varies

greatly depending on model architecture, e.g. SVM trains the fastest regardless of strategy.

CV strategy	Achievable performance	Total CV time
Common K-fold	High	Short
Group-stratified K-fold	Low	Very short
Bayesian search	Very high	Very long

We also find our logistic regression to be scalable, i.e., it can be trained under constrained resources on high dimensions of data, in around ~10 mins for doing Bayesian search CV over higher dimensions of data. Although, the most scalable is our SVM with precomputed RBF kernel (finalizing all CV in under 30 seconds), which is explained by the speed-up in computation owed to the kernel trick. Furthermore, in high-dimensional spaces, data points are more likely to be further apart, leading to sparser sets of support vectors that the SVM model can exploit. The SVM, however, generally has lower CV performance across dimensions of data, across CV strategies. Our RF achieves high performance (on par with LogReg) on low-dim data, however it can't be scaled, and thus we can't collect metrics associated to higher dimensional training, and this occurs for each CV strategy. Such behavior can be explained by the exponential increase in complexity of decision paths per tree as dimensions grow (i.e. curse of dimensionality). This is further exacerbated as we tune tree depth and min split, which requires the CV strategy to explore a wide range of configurations, **Table 3** summarizes our quality assessments.

Table 3. Quality assessments of scalability and performance achieved by the benchmarked models. A "good" scalability indicates the model can be trained on low, high and all dimensions of the data within reasonable computational budget (~10 mins). A 'good' performance indicates the model can achieve a BACC and F1-scores of ~70% and ~0.70 respectively, which is considerably better than random guessing the schizophrenia label. Of note is the 'very poor' scalability of RF when cross-validate on the high and all dimensions of data, where we time-out after $>$ 5 hours of cross-validation.

Model	Performance	Scalability
LogReg	Very good (BACC ~0.75)	Good (<10 mins)
SVM with precomputed RBF Kernel	Good (BACC ~0.6)	Very good (<1 min)
RF	Very good (BACC ~0.75)	Very poor (times out)

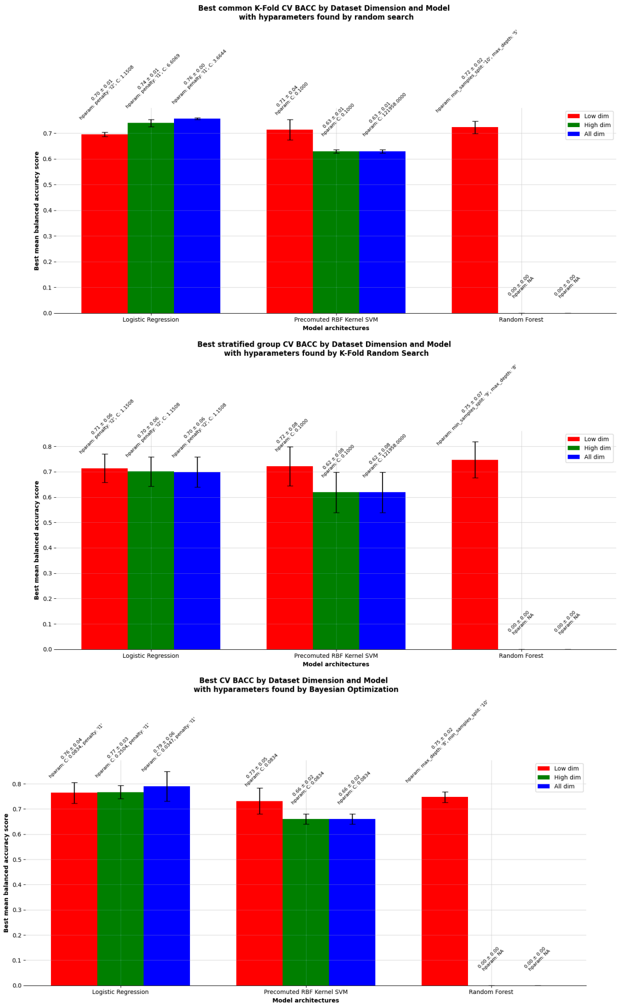
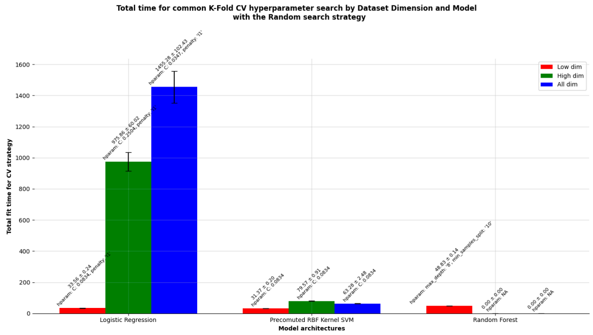


Fig. 1: Mean BACC scores with std. deviation per dataset dimension and model for the Bayesian search optimization. From top row to bottom, we show BACC scores from the common, group-stratified and Bayesian search CV. We observe the highest performance for the logistic regression consistent across data dimensions w.r.t to alternatives. Both LogReg and the RBF-SVM achieve good performance, except the RF where we can't obtain the BACC metrics when trained on the high or all dimensions of the data since we timed-out. Bars are color coded as follows: red is CV on low dim data, green is high dim, and blue is all dimensions. N/A indicates we timed-out from CV.



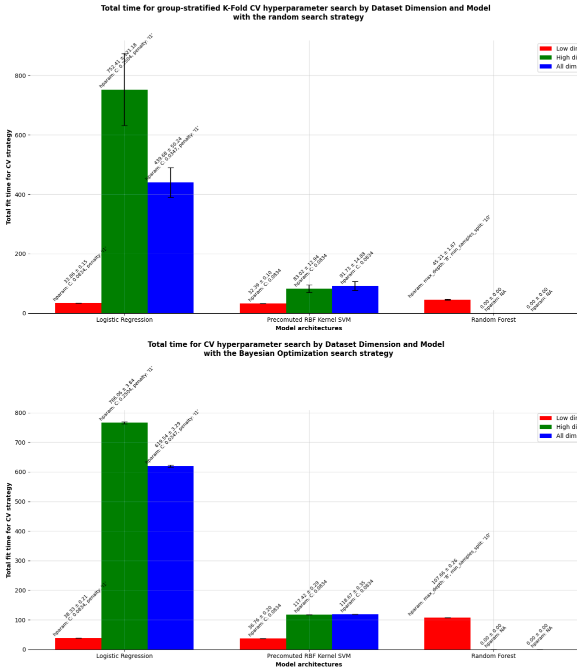


Figure 2: Total CV time for hyperparameter tuning per dimension, per model per CV strategy. From top to bottom we have C using Bayesian search CV, computed as the sum of average time for each k^{th} fold, where the total standard deviation is appropriately calculated through propagation of errors in a sum. N/A indicates we timed-out of CV, thus we have no metrics to report. Color code is the same as in **Fig. 1**

Owed to space and time constraints, we only report BACC. However, we note that BACC correlates with F1 metric, thus F1 metrics show similar patterns. For more details see appended code.

4 Discussion

We achieve interesting results given that simple linear model of logistic regression can achieve very high classification accuracy as per CV BACC. It also scales and strikes a proper balance between high accuracy and low running time. From an applied perspective, thus, a LogReg is a suitable baseline to start experimenting with the dataset for quick results and their analysis. Good results on LogReg also lends itself for interpretability thanks to its linear structure. Examination of its coefficients thus can yield insight into which ROIs are correlated with schizophrenia.

Our work can also serve as reference for using Bayesian search CV for hyperparameter tuning at the expense of computational resources. Some of the advantages that Bayesian search CV has over the alternative CV strategies is that it automates the process of selecting hyperparameters by leveraging past evaluations to inform future choices. Hyperparameters are not searched randomly, but rather adaptively sampled based on previous evaluations, focusing more on promising regions of the hyperparameter space. This adaptability can guide the search for better hyperparameters unlike K-Fold or group stratified K-Fold. However, under constrained resources, both CV strategies can serve as adequate pipelines for hyperparameter tuning.

A core limitation of our work is owed to time constraints. We have run preliminary experiments on other model variants such as Gaussian process classifiers, Naïve Bayes and neural networks (see code). We didn't perform more intricate data preprocessing to achieve better classification performance for each model, but we leave as future work.

5 Conclusion

Advances in neuroimaging techniques help bridge the world of neuropathology to psychiatric diagnoses. Machine learning is that bridge which helps connect the objective data collection of the former to the rigorous clinical application in mental disorder diagnosis.

In this work, we explore several ML models subjected to 3 different CV pipelines to find a high performing logistic regression model which can aid clinicians in classifying schizophrenia from brain anatomical data. Its hyperparameters are found using Bayesian search CV, which is a very powerful hyperparameter tuning strategy at the cost of some computational resources.

Acknowledgments. We thank the organizers of the RAMP competition for providing the dataset for this project, and colleagues for their fruitful insights when building the ML pipeline.

Disclosure of Interests. We declare no competing interests.

References

1. Montazeri, M., Montazeri, M., Bahaadinbeigy, K., Montazeri, M., & Afraz, A. (2022). Application of machine learning methods in predicting schizophrenia and bipolar disorders: A systematic review. *Health Science Reports*, *6*(1). https://doi.org/10.1002/hsr2.962
2. Brown, G. G., Lee, J.-S., Strigo, I. A., Caligiuri, M. P., Meloy, M. J., & Lohr, J. (2011). Voxel-based morphometry of patients with schizophrenia or bipolar I disorder: A matched control study. *Psychiatry Research: Neuroimaging*, *194*(2), 149–156. https://doi.org/10.1016/j.pscychresns.2011.05.005
3. Misiak, B., Samochowiec, J., Kowalski, K., Gaebel, W., Bassetti, C. L. A., Chan, A., Gorwood, P., Papiol, S., Dom, G., Volpe, U., Szulc, A., Kurimay, T., Kärkkäinen, H., Decraene, A., Wisse, J., Fiorillo, A., & Falkai, P.

(2023). The future of diagnosis in clinical neurosciences: Comparing multiple sclerosis and schizophrenia. *European Psychiatry*, *66*(1), e58. https://doi.org/10.1192/j.eurpsy.2023.2432

4. Duchesnay, E., Grigis, A., Dufumier, B., Caud, F., & Gramfort, A. (2024). *Predict schizophrenia using brain anatomy*. Ramp.studio. https://ramp.studio/events/brain_anatomy_schizophrenia_UCL_2024
5. Chilla, G. S., Yeow, L. Y., Chew, Q. H., Sim, K., & Prakash, K. N. B. (2022). Machine learning classification of schizophrenia patients and healthy controls using diverse neuroanatomical markers and Ensemble methods. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-06651-4
6. Kim, G.-W., Kim, Y.-H., & Jeong, G.-W. (2017). Whole brain volume changes and its correlation with clinical symptom severity in patients with schizophrenia: A DARTEL-based VBM study. *PLOS ONE*, *12*(5), e0177251. https://doi.org/10.1371/journal.pone.0177251