# Machine Learning Prediction of Patients with Schizophrenia from Anatomical Brain Imaging

Beiyi Pan

University College London, London WC1E 6BT, UK

**Abstract.** This study explores the use of machine learning models-logistic regression with L1 regularization, random forest, and support vector classification (SVC)—to predict schizophrenia using low and high-dimensional anatomical brain imaging data, and evaluates both simple k-fold and group stratified cross-validation methods.

**Keywords:** Schizophrenia prediction · Machine learning

## 1 Introduction

Brain atrophy in schizophrenia is associated with varied and complex patterns of atrophy [1,2]. In this project, we aim to use grey matter (GM) measurements from participants' brains to identify predictors of clinical status (schizophrenia vs. healthy control).

The pre-processed structural MRI data[1] was divided into two datasets. The first is a low-dimensional dataset with 284 regions of interest (ROIs) of GM scaled for the Total Intracranial Volume (TIV). The second is a high-dimensional dataset containing 331,695 GM 3D voxels (VBM) in Montreal Neurological Institute (MNI) space. In total, there are 513 samples, with 410 in the training set and 103 in the test set.

To achieve the project's aim, we assess the performance of different feature sets, models, and cross-validation strategies. Section 2 describes the models, pipelines, and the choice of comparison metrics used. Section 3 compares the results of different models and cross-validation strategies using both versions of the data. Section 4 discusses the findings from Section 3. Section 5 presents the main findings of the project.

## 2 Methods

### 2.1 Models

The models used are logistic regression with L1 regularization, random forest, and SVC. These three models were chosen based on the characteristics of the data. The feature-to-sample ratio of the low-dimensional data is 0.55, indicating

[1] https://ramp.studio/problems/brain$_{a}natomy_{s}chizophrenia$

that there are more samples than features, suggesting that simpler models might be sufficient and more interpretable. In contrast, the feature-to-sample ratio of the high-dimensional data is approximately 646.58, which signifies that the number of features vastly exceeds the number of samples. Therefore, models capable of handling high-dimensional data without excessive computational costs are preferred.

Logistic regression models prediction using the logistic function, which outputs values between 0 and 1. Specifically, L1 logistic regression adds a penalty equal to the absolute value of the magnitude of the coefficients. It is selected for its simplicity and interpretability, and its ability to prevent overfitting through regularization. Previous work of using l1 logistic regression to predict schizophrenia include references such as Chen et al. (2018) [3] and Ramsay et al. (2018) [4]. Random forest can handle high-dimensional data by constructing multiple decision trees and aggregating the predictions. Previous work include Marchi et al. (2022) [5], Rustam et al. (2020) [6], and Boucekine et al. (2015) [7]. SVC determines the optimal hyperplane that divides classes in the input space with the largest margin and minimal classification errors. It is known for its effectiveness in handling high-dimensional data. Previous work include Rampisela et al. (2018) [8] and Pina et al. (2015) [9].

### 2.2 Piplines

The pipline firstly extract ROIs or VBM data, then applies standard scaler to improve efficiency and performance, and finally implement the specific model. The project compared the performance of a simple k-fold cross-validation with a group stratified cross-validation by sex. K-fold cross-validation divides data into k equal-sized folds, using k-1 folds for training and the remaining fold for validation, iteratively rotating through all folds to assess model performance. Group stratify cross-validation by sex ensures that each fold preserves the same sex distribution as the original dataset, aiding in fair evaluation of model performance across gender categories. To ensure a fair comparison, the number of folds used for both methods is 2 (i.e. 2-fold simple cross-validation vs. 2-fold group stratified cross-validation).

Random search is used optimizes hyperparameters by randomly selecting combinations within predefined ranges, evaluating each to find the optimal solution efficiently (see detailed hyperparameters used for each model in jupyter notebook).

The pipline I submitted to the challenge sequentially applies the ROIs feature extraction, standard scaling, and logistic regression (with hyperparameters penalty='l1', solver='liblinear', random state=1, max iter=2000, C=0.1) steps to the data.

### 2.3 Comparison Metrics

The comparison metrics measure both model performance and computational cost. For model performance, Balanced Accuracy (BACC) and ROC-AUC (AUC) are used. The formula of BACC is:

$$\text{BACC} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{1}$$

where TP, TN, FP, FN represents True Positives, True Negatives, False Positives and False Negatives respectively.

While ROC curve is a graphical representation of the true positive rate (TPR) against the false positive rate (FPR) at various threshold levels. The AUC quantifies the entire area under ROC curve from (0,0) to (1,1).

The dataset contains about 46% patients with schizophrenia. BACC ensures balanced performance across both 'control' and 'schizophrenia' classes, preventing bias towards the majority class. AUC evaluates a model's ability to differentiate between patients with schizophrenia and healthy controls across all thresholds. Computational costs are assessed by measuring training and testing times.

## 3 Results

The results shown in Table 1 compares cross-validation performance, test performance and computational cost across different models, data version and cross-validation straties. Overall, L1 logistic regression with low-dimensional data and group stratified cross-validation achieves the best test performance and the lowest computational costs. Random forest with high-dimensional data has the lowest test performance and the largest computational costs. SVC with low-dimensional data and k-fold cross-validation achieves the highest training performance, while SVC with low-dimensional data and group stratified cross-validation achieves the lowest training performance.

## 4 Discussion

For low-dimensional data, L1 logistic regression and SVC have high performance and low computational cost compared to random forest. For high-dimensional data, random forest is the only model that manages the data without excessive computational costs. L1 logistic regression encounters memory errors due to high memory usage, while SVC's training time is prohibitive (about 12.5 minutes per training session), posing challenges in the tuning process due to limited computational resources. Additionally, random forest exhibits a significant decrease in performance and an increase in running time from low to high-dimensional data, suggesting it cannot fully capture the feature space complexity.

The impact of different cross-validation strategies on results shows that k-fold cross-validation often has more consistent training and testing scores. Scores tend to decrease when using group stratified cross-validation, likely due to the dataset's unbalanced sex distribution, which affects the models' learning capabilities.

**Table 1.** Model Performance Comparison
(The ”-” symbol indicates that the computation could not be performed due to high computational costs.)

| **Model** | **Data** | **Cross-** | **Train** | **Train** | **Test** | **Test** | **Running** |
|---|---|---|---|---|---|---|---|
| | **Version** | **Validation** | **BACC** | **AUC** | **BACC** | **AUC** | **Time (s)** |
| L1 Logistic Regression | Low | k-fold | 0.71 | 0.79 | 0.74 | 0.84 | 0.0985 |
| L1 Logistic Regression | Low | Group Stratified | 0.63 | 0.67 | 0.79 | 0.86 | 0.0470 |
| L1 Logistic Regression | High | k-fold | - | - | - | - | - |
| L1 Logistic Regression | High | Group Stratified | - | - | - | - | - |
| Random Forest | Low | k-fold | 0.68 | 0.78 | 0.74 | 0.80 | 2.0580 |
| Random Forest | Low | Group Stratified | 0.68 | 0.77 | 0.71 | 0.78 | 0.7300 |
| Random Forest | High | k-fold | 0.64 | 0.73 | 0.60 | 0.69 | 24.1831 |
| Random Forest | High | Group Stratified | 0.62 | 0.76 | 0.60 | 0.69 | 26.7055 |
| SVC | Low | k-fold | 0.72 | 0.79 | 0.74 | 0.85 | 0.1829 |
| SVC | Low | Group Stratified | 0.55 | 0.58 | 0.74 | 0.85 | 0.1319 |
| SVC | High | k-fold | - | - | - | - | - |
| SVC | High | Group Stratified | - | - | - | - | - |

From an applied perspective, predicting schizophrenia is more effective with low-dimensional data due to better model performance and computational efficiency. High-dimensional data may be more useful with dimensionality reduction or feature selection techniques, but this is beyond the scope of this work. Group stratified cross-validation is beneficial for constructing a fair and unbiased model, ensuring equal gender representation and reducing bias.

## 5 Conclusion

This work provides a comprehensive analysis of L1 logistic regression, random forest and SVC applied to both low and high-dimensional anatomical brain imaging data for schizophrenia prediction. Random search is used for hyperparameter tuning. The model are assessed by both performance (BACC and AUC) and computational cost (training and testing time). For low-dimensional data, L1 logistic regression and SVC outperforms random forest in both performance and computational cost, while random forest is the only feasible model despite its reduced performance and increased computational cost. The different cross-validation stratigies affects the model performance and choosing the suitable stratigies is essential to build a fair model. This study highlights the potential of machine learning to enhance the diagnosis of schizophrenia using brain imaging data, and also emphasizes the need for careful model and approach selection to improve performance and efficiency.

Limitations of this study include limited computational resources restricting comparisons of models with high-dimensional data, and the use of a 2-fold cross-validation strategy, which may not adequately ensure model generalizability to unseen data.

## References

1. A. Fornito, M. Yücel, J. Patti, S. J. Wood, and C. Pantelis, "Mapping grey matter reductions in schizophrenia: an anatomical likelihood estimation analysis of voxel-based morphometry studies," *Schizophrenia Research*, vol. 108, no. 1-3, pp. 104–113, 2009.
2. D. E. Job, H. C. Whalley, E. C. Johnstone, and S. M. Lawrie, "Grey matter changes over time in high risk subjects developing schizophrenia," *Neuroimage*, vol. 25, no. 4, pp. 1023–1030, 2005.
3. J. Chen, J.-s. Wu, T. Mize, D. Shui, and X. Chen, "Prediction of schizophrenia diagnosis by integration of genetically correlated conditions and traits," *Journal of Neuroimmune Pharmacology*, vol. 13, pp. 532–540, 2018.
4. I. S. Ramsay, S. Ma, M. Fisher, R. L. Loewy, J. D. Ragland, T. Niendam, C. S. Carter, and S. Vinogradov, "Model selection and prediction of outcomes in recent onset schizophrenia patients who undergo cognitive training," *Schizophrenia Research: Cognition*, vol. 11, pp. 1–5, 2018, Elsevier.
5. Mattia Marchi, Giacomo Galli, Gianluca Fiore, Andrew Mackinnon, Giorgio Mattei, Fabrizio Starace, and Gian M Galeazzi. Machine-learning for prescription patterns: random forest in the prediction of dose and number of antipsychotics prescribed to people with schizophrenia. *Clinical Psychopharmacology and Neuroscience*, 20(3):450, 2022. Publisher: Korean College of Neuropsychopharmacology.
6. Zuherman Rustam and Glori Stephani Saragih. Prediction schizophrenia using random forest. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 18(3):1433–1438, 2020.
7. Mohamed Boucekine, Laurent Boyer, Karine Baumstarck, Aurelie Millier, Badih Ghattas, Pascal Auquier, and Mondher Toumi. Exploring the response shift effect on the quality of life of patients with schizophrenia: An application of the random forest method. *Medical Decision Making*, 35(3):388–397, 2015. Publisher: SAGE Publications Sage CA: Los Angeles, CA.
8. T. V. Rampisela and Z. Rustam, "Classification of schizophrenia data using support vector machine (SVM)," in *Journal of Physics: Conference Series*, vol. 1108, 012044, 2018, IOP Publishing.
9. L. Pina-Camacho, J. Garcia-Prieto, M. Parellada, J. Castro-Fornieles, A. M. Gonzalez-Pinto, I. Bombin, M. Graell, B. Paya, M. Rapado-Castro, J. Janssen, et al., "Predictors of schizophrenia spectrum disorders in early-onset first episodes of psychosis: a support vector machine model," *European child & adolescent psychiatry*, vol. 24, pp. 427–440, 2015.