Classifying Dementia Using the Open Access Series of Imaging Studies (OASIS) Longitudinal Dataset

Sarah Bao*
https://github.com/s-bao/data1030-project.git

December 16, 2024

Contents

*Brown University Department of Mathematics & Computer Science

1 Introduction

1.1 Motivation

In this project, we aim to classify dementia using demographic, cognitive, and MRI data. This involves building an ML classification task that can predict whether a subject has dementia based on features such as age, cognitive test scores, and brain volume measurements. This classification task is crucial because early detection of dementia is vital for timely intervention and management, potentially improving patients' quality of life. A definitive diagnosis of Alzheimer's disease can only be made after an autopsy of the brain, so, typically, doctors make clinical diagnoses based off a comprehensive assessment of the patient's condition, considering many of the factors described in this dataset. Therefore, by leveraging predictive models, we can help clinicians identify high-risk individuals and support decision-making in the diagnosis of Alzheimer's disease.

1.2 Dataset Description

The dataset we use is an MRI and Alzheimer's dataset from Kaggle [1]. Specifically, we use the OASIS longitudinal dataset that contains information on 150 subjects ages 60 to 96. This dataset is longitudinal because it has information on at least 2 visits per subject, where visits are separated by at least 1 year. For each visit, all cognitive and MRI measurements are updated. In total, the dataset is 373 by 15 with non-iid data and limited missing data. It has 15 features: Subject ID, MRI ID, Group, Visit, MR Delay, M/F, Hand, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF. The target variable is "Group" with three classes: "Demented", "Nondemented", or "Converted". Converted subjects are those who, across multiple visits, converted from nondemented to demented — the dataset does not specify at which point the subject transitions. Other features to note are MMSE and CDR (minimental state examination and clinical dementia rating), which are scores that come from two different cognitive tests. And eTIV, nWBV, and ASF (estimated total volume in skull, normalized volume of whole brain, and Atlas scaling factor) are metrics that come from open access MRI scans for each person during each visit.

1.3 Previous Work

Many of the other notebooks on Kaggle relating to this dataset had similar results with accuracy levels around 90%. Other notebooks interestingly either removed CDR as one of the predictors or made CDR the target variable, in which case their accuracy was lower.

2 EDA

We begin our exploratory data analysis by exploring the target variable "Group". As seen in Figure 1, our target variable is imbalanced with much fewer data points in the "Converted" class than the other two.

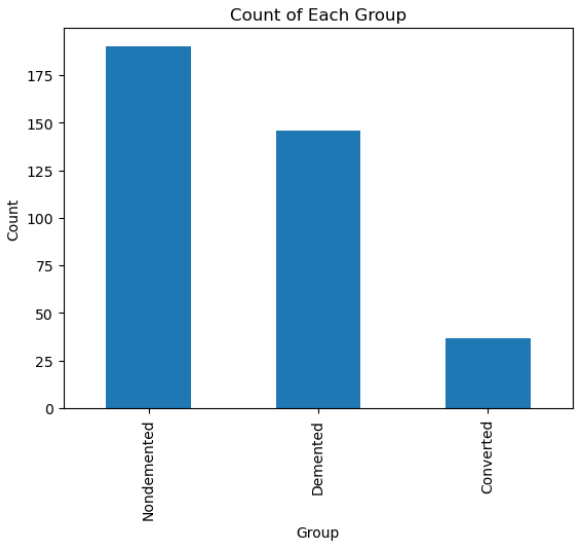
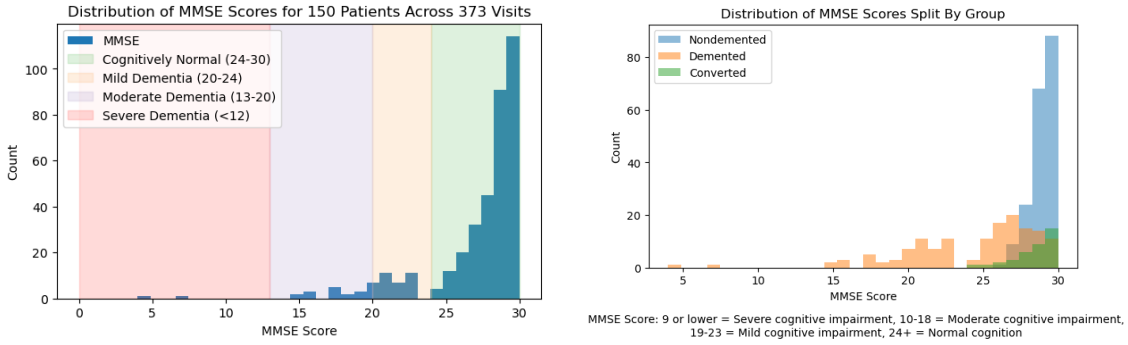


Figure 1: Distribution of Target Variable "Group"

In addition, MMSE is a cognitive test, so it has its own dementia rating system where patients who score less than 12 have severe dementia, 13-20 have moderate dementia, 20-24 have mild dementia, and 24-30 are cognitively normal. Hence, we wanted to see how the MMSE scale compared to the target variable labels.



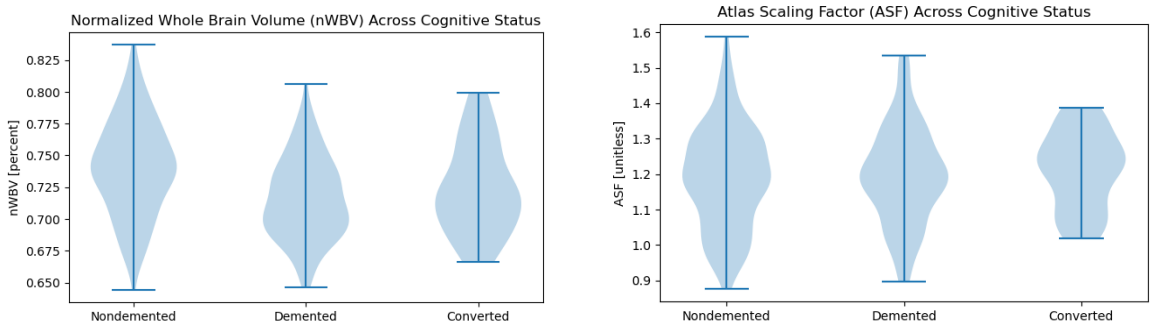
(a) Distribution of MMSE Scores

(b) Distribution of MMSE Scores by Group

Figure 2: Visualization of MMSE Score Distributions

Here, we can see that although most of the MMSE thresholds align with the classes, there is no clear-cut boundary. For example, within the 25–30 range, all three classes are represented. This implies that the MMSE score alone cannot reliably determine whether a patient is demented and further supports the notion that predicting dementia is a multifaceted task.

Next, we explored the MRI metrics, and their relationships with the target variable.



(a) Normalized Whole Brain Volume Across Group (b) Atlas Scaling Factor Across Group

Figure 3: MRI Metric vs. Cognitive Status Violin Plots

Note that nWBV is a measurement of the percentage of intracranial cavity that is occupied by brain tissue, and ASF is a numerical value calculated when comparing a brain image to a standard brain atlas. The figure tells us that, in general, demented individuals have lower normalized brain volumes than non-demented individuals, and there does not appear to be a big difference in the scaling factor measure between the two groups.

We also explored correlation across all features:

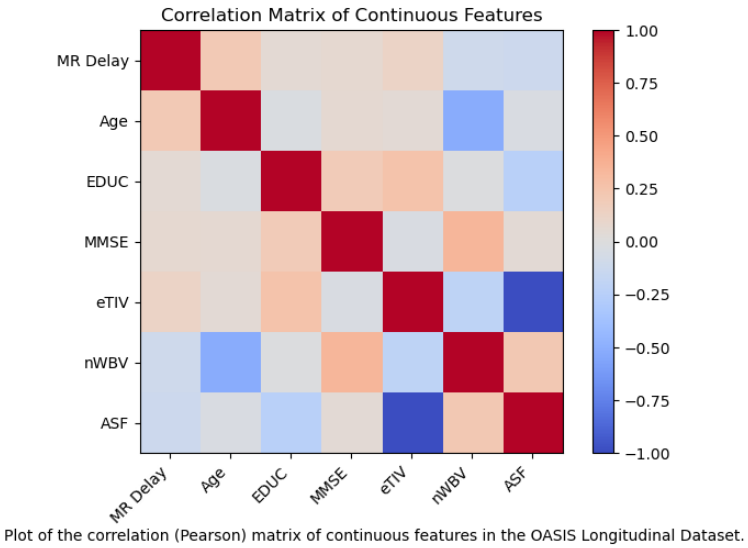


Figure 4: Correlation Matrix of Features

Here, we see that most of the features in the dataset are very slightly correlated except for eTIV and ASF. When we plot those two variables against each other in a scatter plot, we see that they are strongly negatively correlated ($r = -0.989$). Hence, in the dataset, we removed eTIV as a predictor to protect against multicollinearity issues.

3 Methods

3.1 Splitting

As mentioned above, the dataset has group structure, is imbalanced, and is quite small with only 373 data points. Thus, we grouped data by the "Subject ID" variable, stratified by the "Group" target variable, and used a StratifiedGroupKFold cross-validation procedure with an approximate 60-20-20 split.

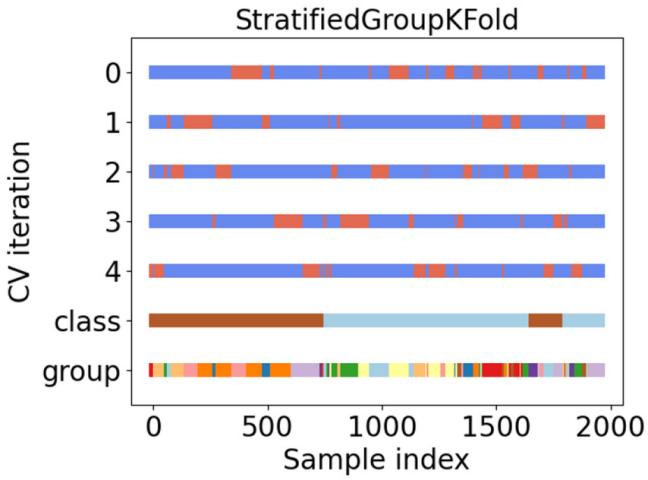


Figure 5: Stratified Group K-Fold Cross-Validation Procedure [2]

3.2 Preprocessing

To begin preprocessing, we first dropped "MRI ID", "Subject ID", "eTIV" (and kept "ASF"), and "Hand" since all patients were right-handed. Next, we observed the missing data distribution in the dataset and found that only two variables had missing values: SES, which is ordinal, and MMSE, which is continuous. The missing values in SES were handled with the ordinal encoder by adding a -1 category, and the missing values in MMSE were handled with multivariate imputation using a random forest regressor. For the remaining features, we simply categorized them by type (ordinal, categorical, or continuous) and used the corresponding preprocessor (OrdinalEncoder, OneHotEncoder, or StandardScaler, respectively). We then ran a final standard scaler through all transformed features to ensure that they were on the same scale, which helped with interpretability analysis later on.

3.3 Evaluation Metric

We evaluated several metrics, including accuracy, precision, F1 macro, and F1 weighted. While the dataset was imbalanced overall, the two key classes — demented and nondemented — were relatively balanced, so no significant adjustments were necessary for class imbalance. Ultimately, we selected F1 macro as our primary metric because it provided the greatest differentiation between model performances.

3.4 Pipeline Summary

Our overall pipeline iterates through 10 random states, performing one outer split to separate the test set from the training + validation set. The data is then preprocessed as described above. For each random state and model, we perform a grid search over the model's corresponding parameter grid to identify the parameter set that achieves the highest average validation score across 4 inner stratified group folds. Finally, the best model is evaluated on the test set to obtain the test score, and the test scores are averaged across the random states.

3.5 Model Selection

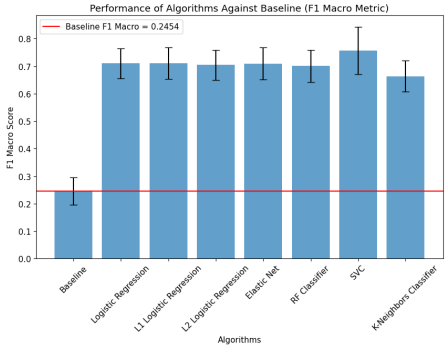
Since we did a classification task, the 7 models we fitted were all classifiers. Below are the parameters tuned for each model:

Model Name	Parameters Fitted
SimpleLogisticRegression	{}
L1LogisticRegression	{'C': [0.001, 0.01, 0.1, 1, 10, 100]}
L2LogisticRegression	{'C': [0.001, 0.01, 0.1, 1, 10, 100]}
ElasticNet	{'C': [0.001, 0.01, 0.1, 1, 10, 100], 'l1_ratio': [0.001, 0.01, 0.1, 1]}
RandomForestClassifier	{'n_estimators': [100], 'max_depth': [1, 3, 5, 10, 20, 100], 'max_features': [0.25, 0.5, 0.75, 1.0, None]}
SupportVectorClassifier	{'C': [1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3], 'gamma': [1e-5, 1e-3, 1e-1, 1e1, 1e3, 1e5]}
KNeighborsClassifier	{'n_neighbors': [3, 5, 10, 20], 'weights': ['uniform', 'distance']}

4 Results

4.1 Model Results

Here were the results from the models fitted:



Algorithm	Mean F1 Macro Score	Std Dev
Baseline	0.2454	0.0503
Logistic Regression	0.7101	0.054
L1 Logistic Regression	0.7104	0.0574
L2 Logistic Regression	0.7043	0.0552
Elastic Net	0.7094	0.0581
RF Classifier	0.7004	0.0592
SVC	0.7568	0.0853
K-Neighbors Classifier	0.6638	0.0559

Figure 6: Mean and Standard Deviations of F1 Macro Scores For Each ML Model + Baseline

As we can see, outside of k-nearest neighbors, the other models performed very well and quite similarly with mean scores over 9 standard deviations above the baseline. While the SVC model achieved the highest mean score, we also noted that it had the highest standard deviation, indicating greater variability in its performance. Despite this, we chose to focus on the SVC model for interpretability analysis.

4.2 Interpretability Analysis

When extracting the best-performing SVC model (test score: 0.9861) and its corresponding random state, we observed that the test set contained no "Converted" data points due to the randomness of splitting and the limited number of "Converted" patients in the dataset. Consequently, we opted to use the second-best SVC model (test score: 0.7795) for further analysis. To start, we plotted its confusion matrix to evaluate classification performance across all classes:

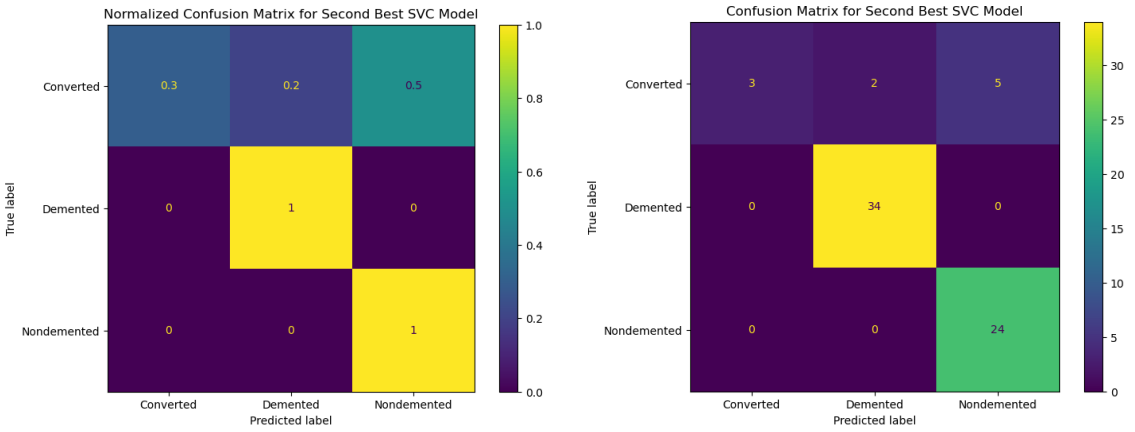


Figure 7: Confusion Matrices for Second Best SVC Model

Observe that the model perfectly predicted demented and nondemented subjects, but struggled to correctly predict converted subjects. This makes sense because converted subjects are technically a mix of nondemented and demented, so our model did pretty well in mixing its predictions for the converted data points.

4.2.1 Global Feature Importance

We then analyzed the global feature importance of the model to understand which features contributed most to its predictions across the entire dataset. Below are three plots corresponding to three different global feature importance computation methods:

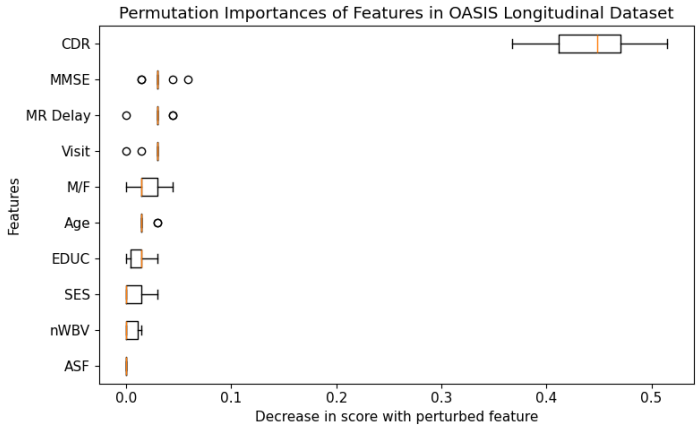


Figure 8: Global Feature Importance Using Permutation Method on SVC Model

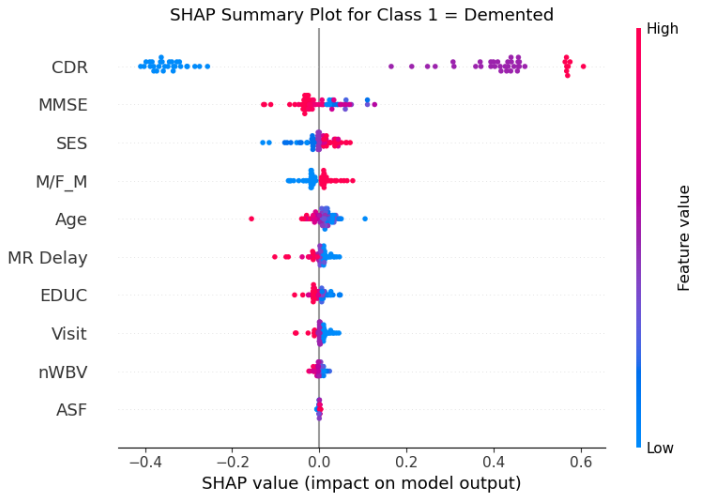


Figure 9: Global Feature Importance Using SHAP on SVC Model

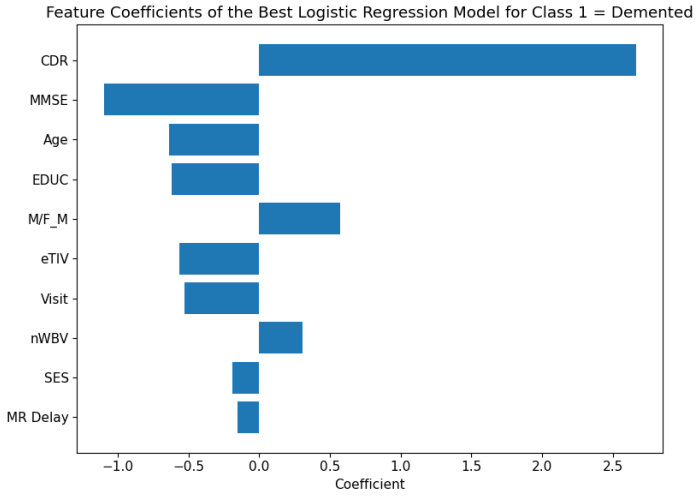


Figure 10: Global Feature Importance Using Scaled Logistic Regression Coefficients

Across all three plots, CDR and MMSE, the two cognitive assessment metrics, emerged as the most influential features, with CDR having a particularly dominant contribution. Interestingly, the MRI metrics (nWBV and eTIV) ranked as the least important features in the first two plots, suggesting that these specific MRI measurements may not be as predictive of dementia compared to cognitive assessments. Additionally, being male was associated with an increased probability of being classified as demented, hinting at a potential sex-related dynamic worthy of further exploration.

4.2.2 Local Feature Importance

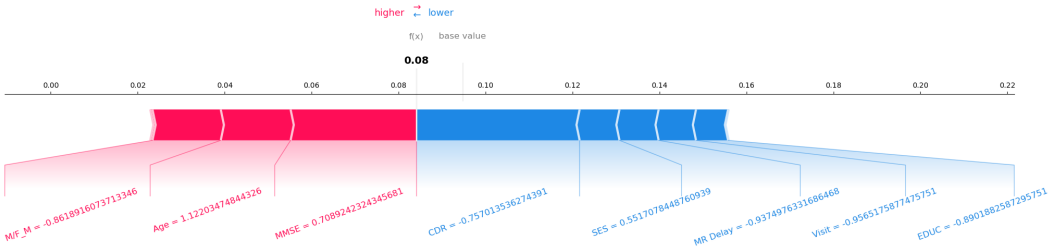


Figure 11: SHAP Force Plot For a Class 0 = Converted Patient

Here, we can see that the CDR and MMSE values are like opposing forces. The high MMSE and low CDR were conflicting for this converted patient, so the model had a difficult time correctly predicting this patient as converted.

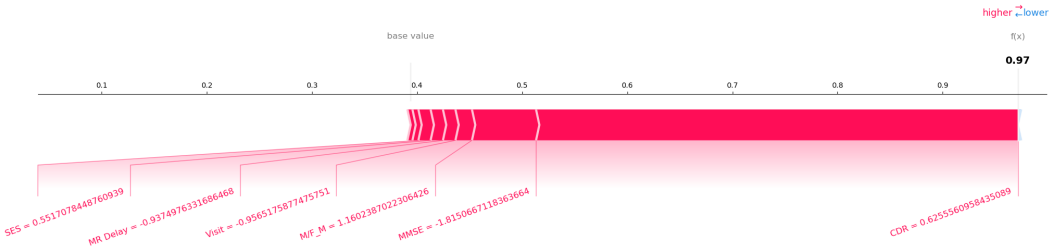


Figure 12: SHAP Force Plot For a Class 1 = Demented Patient

This force plot exemplifies the strong predictive power of CDR. With a high CDR score and a low MMSE score, the patient was very highly predicted to be demented, and they were.

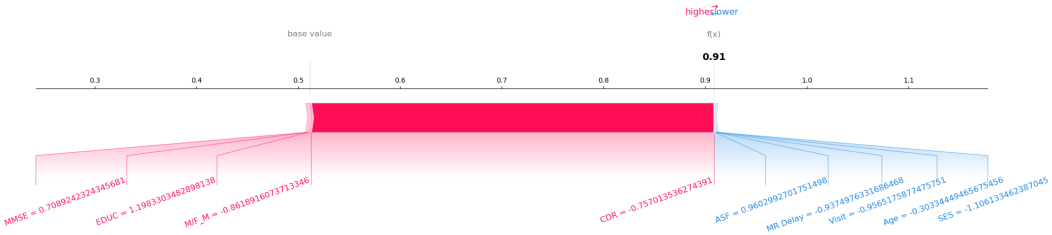


Figure 13: SHAP Force Plot For a Class 2 = Nondemented Patient

This force plot is similar to the one above, but exemplifies the predictive power of CDR in the opposing direction. A low CDR score drastically increased the probability that this patient was predicted to be nondemented, and it seems like CDR was the only factor that had even a moderate influence on the prediction for this data point.

5 Outlook

To improve predictive power, the dataset could be enhanced by adding features such as environmental factors, medical history, and more MRI metrics — this can only be done well with more domain knowledge. Additionally, other methods such as XGBoost or deep learning models could improve accuracy, but at the expense of interpretability. Also, it could be useful to, instead of imputing values for missing data, try to run the model on a subset of the available features using a reduced features model. Another way we could improve this project is by more closely considering the evaluation metric. Since false negatives are more critical in this classification task than false positives, we would want to place more emphasis on recall by using, for example, an f2 score metric.

6 References

[1] Jason Boysen. *MRI and Alzheimer's Dataset*. `https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers`. Accessed on December 15, 2024. 2023.

[2] Kaixu Chen et al. "Two-stage video-based convolutional neural networks for adult spinal deformity classification". In: *Frontiers in Neuroscience* 17 (2023). Accessed on December 15, 2024. DOI: `10.3389/fnins.2023.1278584`. URL: `https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2023.1278584`.