Yuli Jin, Violet Chen, James He
Prof. Masanao
April 17, 2022

Report for Mice Project

-Opposite Sex Experiment

Abstract

This project aims to find relationship between cells' z-score and mouse's behavior. We did exploratory data analysis to calculate the proportion of each behavior. In the model building part, we utilized Gatted Recurrent Unit (GRU) and Encoder model to predict mouse's reaction. In addition, we also combine training dataset and validation dataset together to fit training model and draw a conclusion that both two are good.

Introduction

This project intends to investigate how brain circuits control behavior using statistical analysis and deep-learning models. Department of Anatomy & Neurobiology from School of Medicine initiated the field experiments and come with data that needs to be analyzed. This report focuses on the data analysis, and I will explain the experiments briefly to give an overview of the project.

The biology team did three experiments investigating the behavior of mice with recordings of the activity of their brain cells. Our group choose to narrow our research to the opposite sex experiment to get a deeper and more comprehensive investigation. The experiment put mice into a social chamber for 10 minutes and mice are allowed to explore between two cups containing a male and a female of the same strain. (Figure 1) Their interactions with male and female cups are recorded, together with their cell activities.

The goal is to find out the relationship between mice behaviors and their cell activities. Our team will first conduct exploratory data analysis (EDA) and see if we can find any patterns. Then we will do preliminary analysis using basic models to find out possible associations between cell activities and behaviors. At last, our team will build deep-learning models to uncover the deeper relationship between brain and behavior, that is, make predictions for mice behavior based on their cell activities.

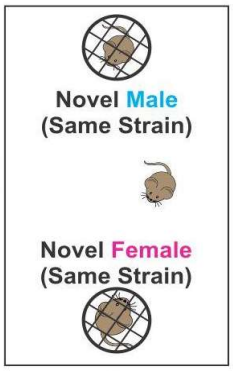


Figure 1: Opposite Sex Experiment

Exploratory Data Analysis

Data Explanation and processing

The dataset contains two data spreadsheets, which are 'binned z-score' and 'binned-behavior', record mice's cellular activity changes and their interactions with the male and female cups for 6300 seconds. The number of cells in each mouse is not the same, for example, mouse 416 has 44 recorded cells but mouse 414 has 128 recorded cells. Three outcomes of the behavior are: 'interact with male', 'interact with female' and 'no touch'. The detailed look of the data is shown below:

Table 1: Explanation of column

Data Table	Column name	Explanation
Binned z-score	$V_1, V_2, \ldots, V_n$	Each V column represents a different cell of mouse, and each row is a different temporal sample in the time series.
Binned behavior	V_1	This column contains the result for whether the mouse interact with male or not. The outcome shows in the type of 1 (interact with male) or 0.
	V_2	This column contains the result for whether the mouse interact with female or not. The outcome shows in the type of 1 (interact with female) or 0.

* In 'Binned behavior' table, the interaction results as 'no touch' when V_1 and V_2 are all equal to 0.

Data Visualization

Before performing data visualization, it is necessary to process the data by combining two data tables and checked for the NA values in the new data set. For simplicity better interpretability, we calculated the average cell z-scores at every time point and obtained 6300 average z-scores for each mouse. Then we added a column to classify the behavior into three types: 'interact with male', 'interact with female' and 'no touch'.

The dataset is a time-series data, and for the purpose of getting some figures that are easier to understand and making the model fit better, we make some line plots to explore the relationship between cell z-score and mice's behavior. Considering that there are six mice in the experiment and each mouse has a different type and number of cells, we drew six line graphs for each of them.

In the line plot below, which is for Mouse 414, the x-axis is time, and the y-axis shows average z-score of mouse's cells. Different colors indicate different behaviors: blue means 'interact with male', red means 'interact with female' and green represents 'no touch'.

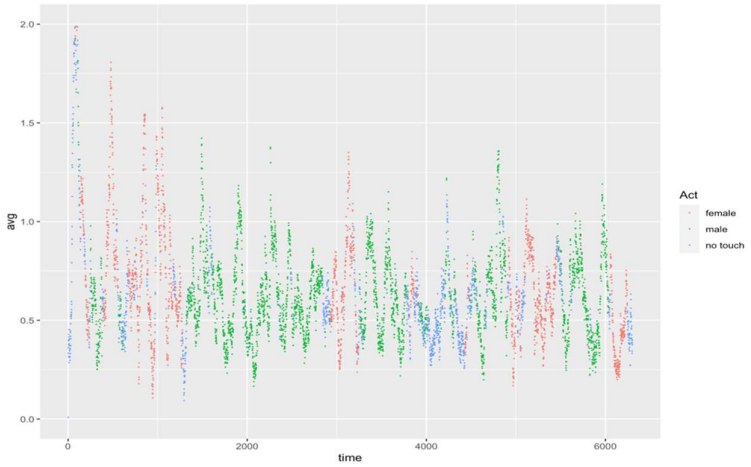


Figure 2 Mouse414 Average z_score of cells VS Time in behavior

From figure 2, it is obvious that the average z_score of cells is around 2.0, and the peak is at the beginning. As time progressed, the experimental mouse interacts with female or male mice cup. But in the time range of 1500th and 3000th time points, Mouse 414 exhibits a no contact behavior. It is only after the 3000th time point that the mouse shows contacting behaviors with other mice, which are shown as peaks in the average cells' z_scores but are below the highest peak score at the beginning of the experiment.

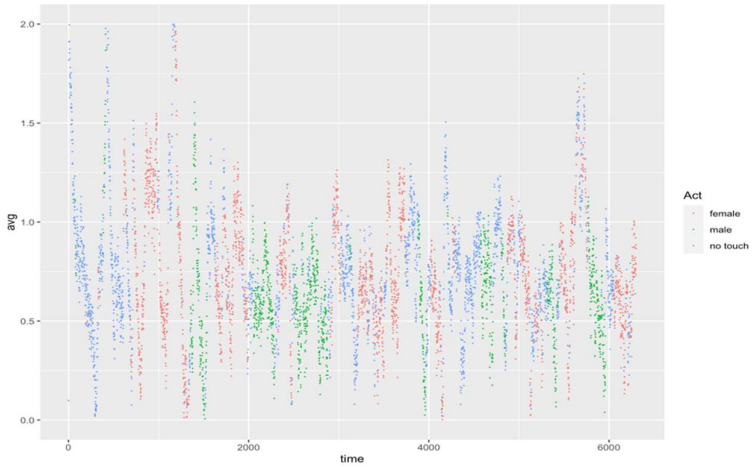


Figure 3: Mouse416 Average z-score of cells VS Time in behavior

Figure 3 illustrated the cells' average z-score of Mouse 416 and its behavior with the progress of the tine. As can be seen in this figure, three peaks of cells' average z-score occur when Mouse 416 interacts with male mouse. In addition, from the whole perspective, it looks like that Mouse 416 spent more time in contact with its own kind than stayed alone. The line plots for other mice are in the appendix.

We cannot draw any conclusions between cell's average z-score and mice's behaviors, no obvious patterns are discovered from the graphs. Besides, it is difficult to know the propensity of each mouse's interaction pattern from these graphs directly. Therefore, in order to make further analysis in mice's behaviors, it is important to show the proportion of different behavioral outcomes across the experiment. Below are the results of six mice:

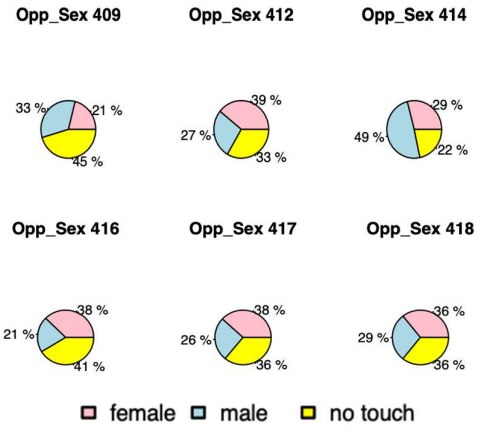


Figure 4: Proportion of different behavioral outcomes

From the pie chart above, Mouse 412, Mouse 416, Mouse 417 and Mouse 418 have relatively similar interaction outcome distribution. The proportion of 'interact with male' are around 25% and the proportion of 'interact with female' are nearly 38%. They only have no contact with other two cups about 36% of the time. However, for Mouse 414, the proportion of 'interact with male' has up to nearly 50%, which can be considered that this mouse might spend more time to interact with male mouse. On the contrary, it looked that Mouse 409 did not have contact with two cups for almost half of the time (45%).

Preliminary Analysis

Now we will choose appropriate methods to conduct preliminary analysis to the dataset. As states in the previous section, the outcome Y is a categorical variable with three types of outcomes: 'interact with male', 'interact with female' or 'no touch'. The predictors are cells, name as V_1, V_2, …, V_n, each represents a single cell with z-score indicates the active level at the time. For the categorical outcome, we decide that the binomial logistic regression model is the best fit. Multinomial logistic regression is another good model for the categorical outcome, but it is difficult to visualize, so we decide not to use this model.

The binomial logistic model requires the outcome to be binary, which means either A or B, so our group decide to fit two models to each mouse with two outcomes: V_1, interaction with male or

none, and V_2, interaction with female or none. The predictor is the average z-score, which is same as the previous section. Below is the visualization of the logistic regression model for Mouse 416:

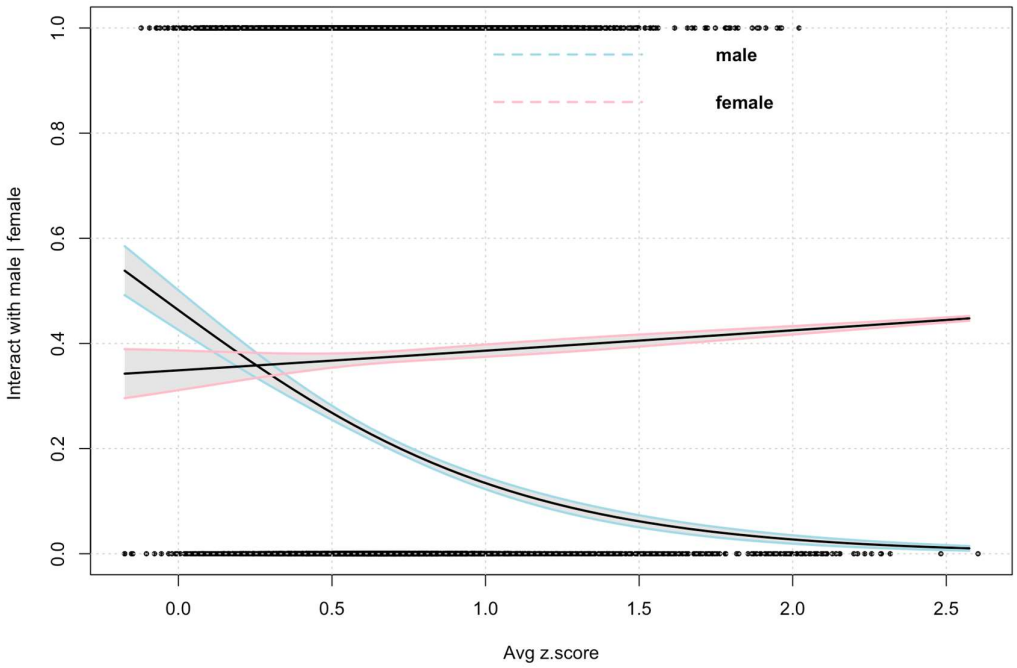


Figure 5: Logistic regression

From the graph you can see the probability of the Mouse 416's interaction result. The x-axis is the average z-score of cells, and the y-axis is the probability of interaction with male or female. The blue line represents the interaction with male and the pink line represents interaction with female. The blue interaction line, for male cup, shows a downward trend that the probability of interaction with male decreases with the increase of average z-score. This trend means that with higher cell activity level, the probability of Mouse 416 interact with male gets lower. On the contrary, the pink interaction line, for female cup, shows a smooth and upward trend. The trend demonstrates that with more cell activities, the probability of interaction with female cup is slightly higher.

However, such regression model does not have much practical value for we find the relationship between cell activities and the behaviors at the exact same time point. To make the model more accurate and are more generalized to real world, we decide to apply rolling sum to the average z-scores prior to the time point of the behavior. For example, we use the sum of the average z-score from t_{30} to t_{40}, instead of only t_{40}, to find the association with the behaviors at t_{40}. The graph below shows the roll sum logistic regression for Mouse 414:

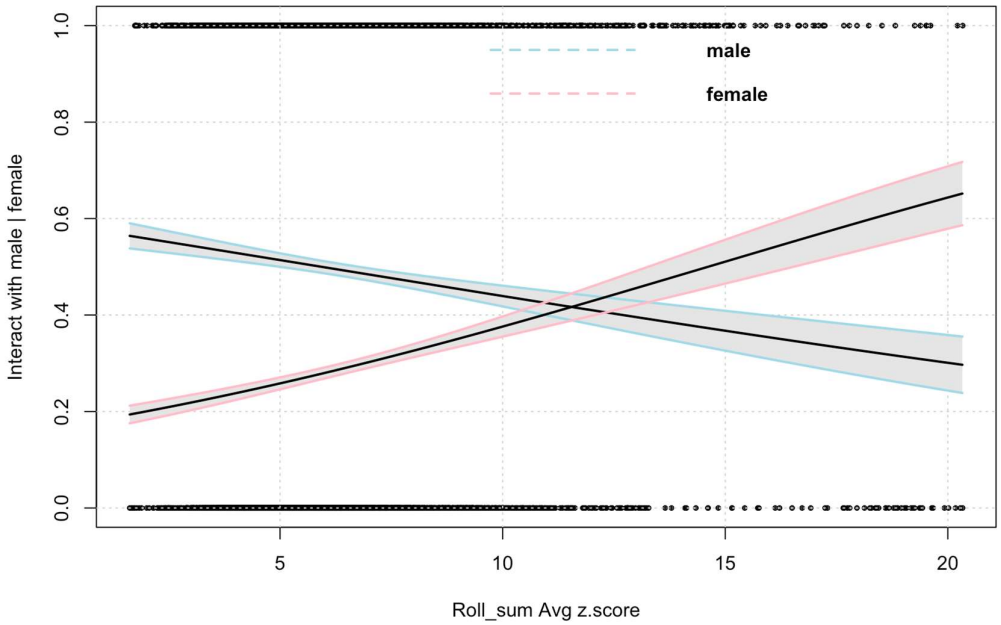


Figure 6: Roll sum logistic regression

The pink line shows the probability of interacting with female as the roll sum of the average z-scores of ten time points before the behavior t, the blue line is the probability of male interaction. From the graph, the probability of interaction with female increases as the cell activity gets higher; the probability of interaction with male, however, decreases as the cell activity gets higher.

As comparison, we also compute the regression model for Mouse 414 at the same time point and the figure is shown below:

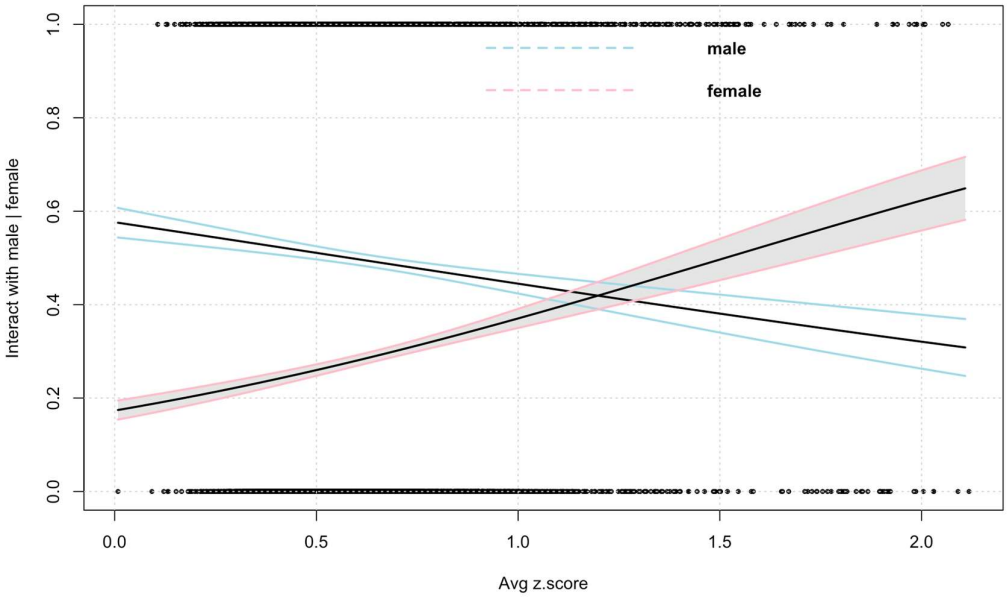


Figure 7: Logistic regression

The trends from the logistic regression without roll sum are very similar to the roll sum regression model. Therefore, for Mouse 414 we can conclude that higher cell activity level means higher probability to interact with female and low probability to male.

Model Building

Data processing

In the opposite sex dataset, we decided to use many-to-one prediction. It is a three-classification problem, and this dataset is sequential data, so the best way to split the dataset is one of the methods in natural language process. The time length is set to 90 and we assumed that the information in the latest 90-time spots had predictive power on mice's behavior. The details of cutting one observation were listed as following figure (figure 8). For each observation, we collected the most recent 90 times' cell score. As a result, the X of each observation is a matrix

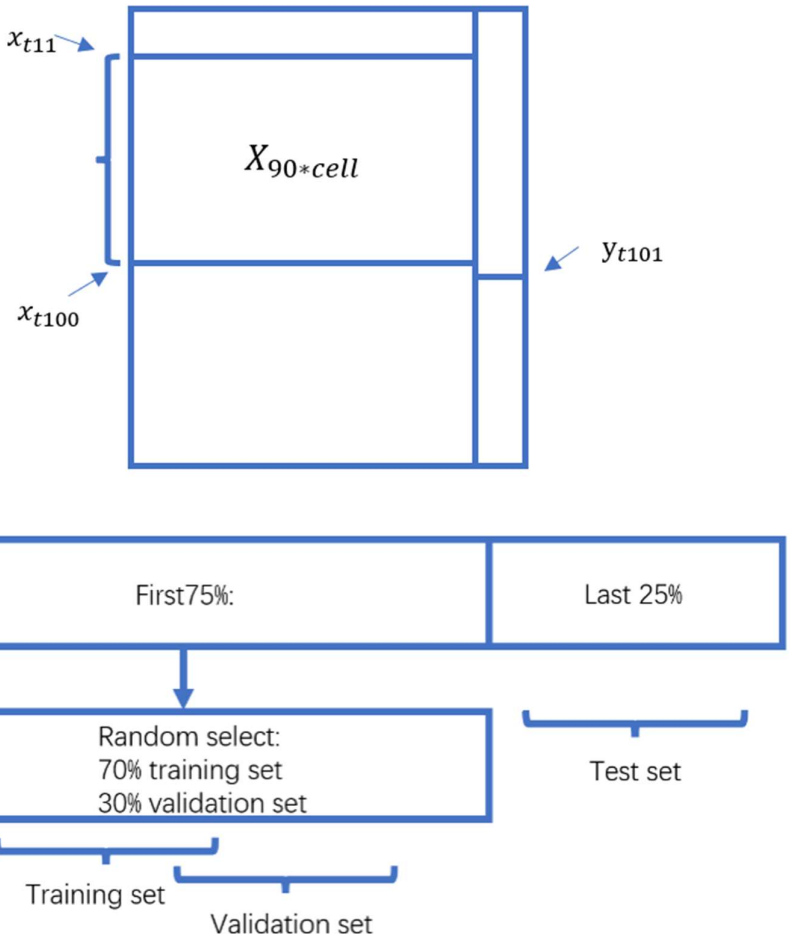


Figure 8: Observation cutting

Each mouse dataset is split into training set, validation set and test set. For each mouse, the last 25% of observations is the test set; for the remaining observations, 70% of the remaining observations were chosen to be the training set and 30% to be validation set based on random selection. In the process of model training, our group used the training set to tune the parameter and the validation set to detect the early stopping time. The test set did not involve in the training process. Also, the batch size is 64 to wrap the whole training set into an iteration.

Model Explanation

Two deep learning models were applied to make the prediction. The first model is Gated Recurrent Unit (GRU), which is suitable for the sequential dataset. LSTM is also an appropriate model, but GRU is less complex and contain less parameters. The second model is the encoder of transformer, which have two essential parts of encoder and positional encoding, and it can explore whether self-attention outperform the traditional RNN model or not. The encoder, which is one of important part of the transformer, contains the self-attention module which is also favorable to sequential data.

GRU

The framework of GRU is listed in figure 9. The GRU model has two layers, and the hidden size is 64 elements. First, we expanded the output vector into 128 elements and conducted dropout and Relu activation function to reduce 128 elements into 32 vectors for the output layer. After that, same procedure was used to reduce 32 elements into 3 elements. Finally, we applied Softmax to those 3 elements to get the probability of behavior label.

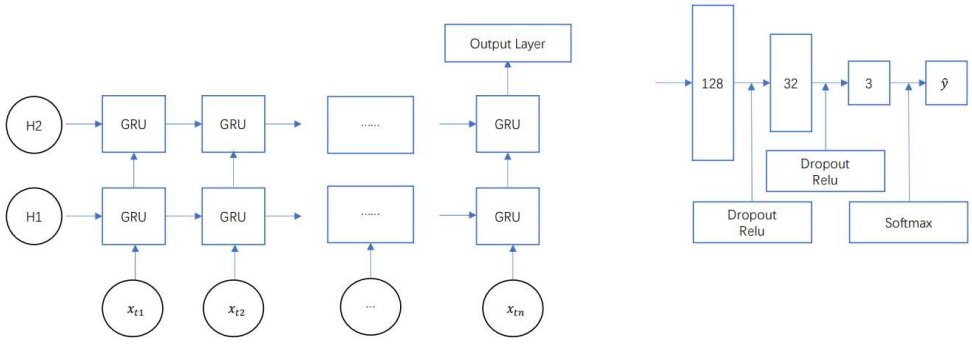


Figure 9: Framework of GRU

Encoder

The framework of Encoder is listed in the figure below (figure 10). Except the word embedding part, the whole framework in this part is identical to the encoder part in transformer. The Encoder layer is sed to 6 and the output layer is the same as the one in GRU model.

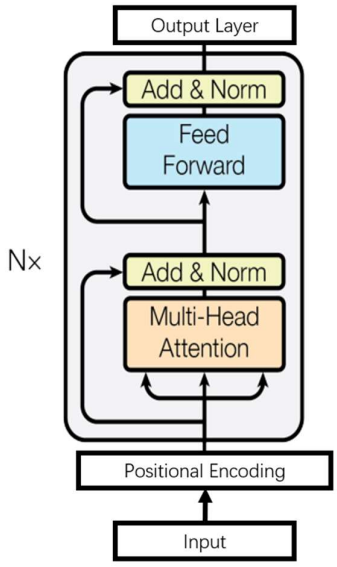


Image: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Figure 10: Framework of Encoder

Result

Original Results

The first table lists the rest of the training parameters.

Table 1: training parameter

learning rate	0.001
max epoch	30
optimizer	Adam
loss	Cross entropy loss

Table 2 and 3 give the training and test result on GRU and encoder. As states before, the dataset is split into three parts: training, validation and test. The training loss means the cross-entropy loss for the train dataset and the smaller the value is, the better fit the model to the dataset. However, the train loss should not be too small for it may lead to overfitting. To prevent overfitting, validation loss, which is the cross-entropy loss for the validation dataset, is introduced and we pick the minimum number as the standard for the train loss. Test loss is the cross-entropy loss for the test dataset and the smaller the value the better the model for the test dataset.

Test accuracy, which is the most important indicator in the table, demonstrates the accuracy of the model, and the larger the value the more accurate the model is. Stop epoch is the total number of iterations that the model run, and the stop batch is the number of iteration that the subsets run.

From the tables, the encoder model's test accuracy generally outperforms the GRU model except for Mouse 416. However, the differences between two models are small.

Table 2: training and test result on GRU

ID	train loss	validation loss	test loss	test accuracy	stop epoch	stop batch
409	0.100	0.174	0.118	0.951	15	35
412	0.254	0.240	0.294	0.878	22	15
414	0.050	0.126	0.103	0.964	26	40
416	0.251	0.197	0.253	0.900	30	60
417	0.131	0.155	0.164	0.939	25	65
418	0.106	0.128	0.112	0.957	29	30

Table 3: training and test result on Encoder

ID	train loss	validation loss	test loss	test accuracy	stop epoch	stop batch
409	0.099	0.183	0.095	0.966	23	35
412	0.149	0.280	0.234	0.916	19	60
414	0.049	0.141	0.078	0.973	29	65
416	0.286	0.262	0.303	0.882	28	65
417	0.071	0.182	0.156	0.943	14	20
418	0.053	0.182	0.115	0.959	27	55

Combining training set and validation set

The original results in the section above returns the model test accuracy for the training set with exact stopping time, which is the combination of epoch and batch. However, the results above only contain the data from training set and ignore the validation set, which may contain useful information. Therefore, leaving the two sets apart will lead to the inability for the model to learn on the important features from validation set, especially when the observation of the dataset is limited. However, if we directly combine training set and validation set with the same batch size, the number

of batches in each epoch will be higher than the training set. Fortunately, there is an easy way to deal with the delay of early stopping time. The main idea is that the number of iterations is immutable no matter how batch changes. The alternative method that we combine the training set and the validation set is described as below:

First, calculate the total iteration time based on early stopping epoch and batch. Then get the new batch number and use floor division to the total iteration. The floor integer+1 is the new stopping epoch, and the remainder is the new stopping batch. Finally, apply new stopping parameters to train the new model and use that model to make prediction on test set.

The following table illustrates the comparison of two early stopping training methods. There is an evident improvement in test loss and test accuracy on GRU after combining training set and validation set together. However, the overall performance for Encoder model shows worse results. The cross-entropy loss from only 3 out of six mice do achieve improvement but the accuracy of all mice's experiences slight decrease.

Table 4: Comparison between two training set on GRU

	before		**after**			
ID	test loss	test accuracy	test loss	test accuracy	stop epoch new	stop batch new
409	0.118	0.951	0.087	0.961	11	38
412	0.294	0.878	0.255	0.892	16	4
414	0.103	0.964	0.072	0.975	19	13
416	0.253	0.900	0.232	0.906	22	17
417	0.164	0.939	0.134	0.945	18	57
418	0.112	0.957	0.110	0.959	21	15

Table 5: Comparison between two training set on Encoder

	before		**after**			
ID	test loss	test accuracy	test loss	test accuracy	stop epoch new	stop batch new
409	0.095	0.966	0.124	0.956	16	92
412	0.234	0.916	0.203	0.910	14	37
414	0.078	0.973	0.067	0.970	21	50
416	0.303	0.882	0.289	0.882	20	78
417	0.156	0.943	0.175	0.937	10	38
418	0.115	0.959	0.160	0.931	19	96

Although the changes on GRU and Encoder models are contradictory, we believe that combining training and validation set together as the final training dataset is more scientific and rigorous. The performance of final output is always influenced by randomness. If we change the random seed or arbitrarily modify the number of iterations, the result of Encoder will outperform the origin one. It seems that identical iterations will not necessarily achieve the best performance on test set and the early stopping is to get the general optimal iteration regime but not exact number.

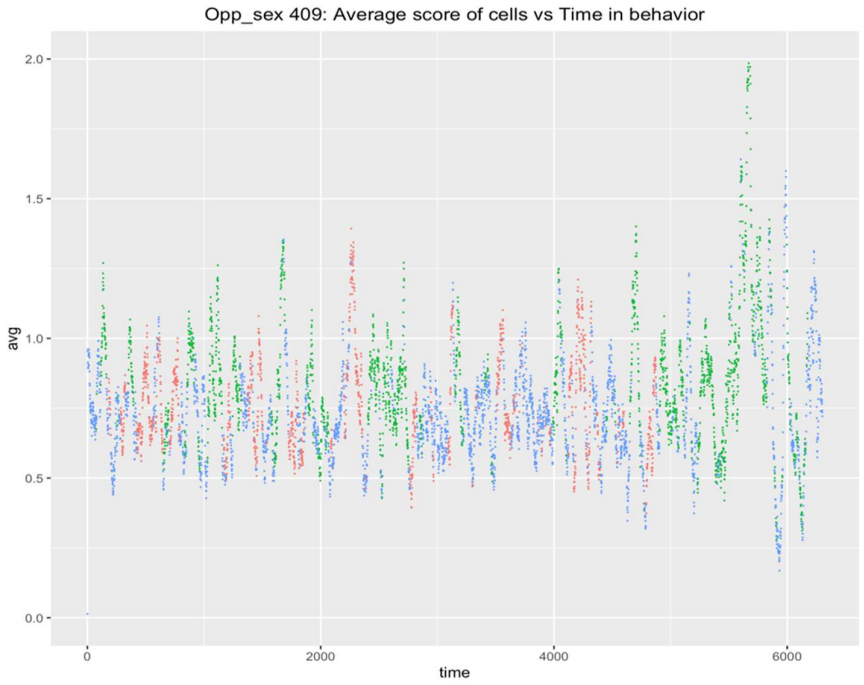
Discussion

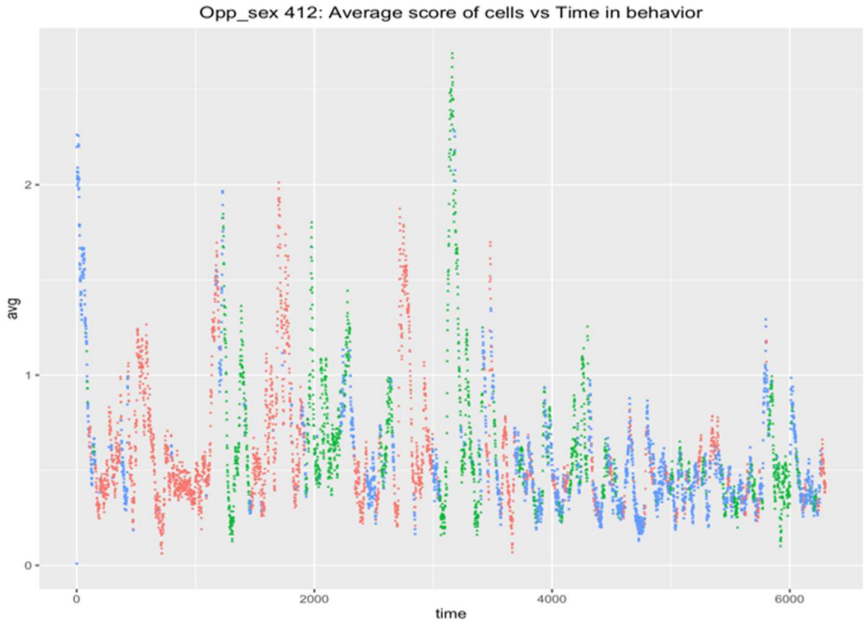
The purpose of this project is to predict mice' behavior based on statistics model. In the preliminary analysis section, we conduct a conclusion that the mouse has tendency to interact with female cup when the cell's z−score grows high. In the model construction part, two deep learning models, which are GRU and Encoder, are utilized and successfully show that cells' z−score has strongly predictive power on mice' behavior. We put more details' explanation in the appendix. Both GRU and Encoder reached high predictive power. Even though Encoder had better performance in test set, GRU had better performance in validation set. The combination modeling also shows mixed results. Therefore, there was no evidence that Encoder outperformed GRU.

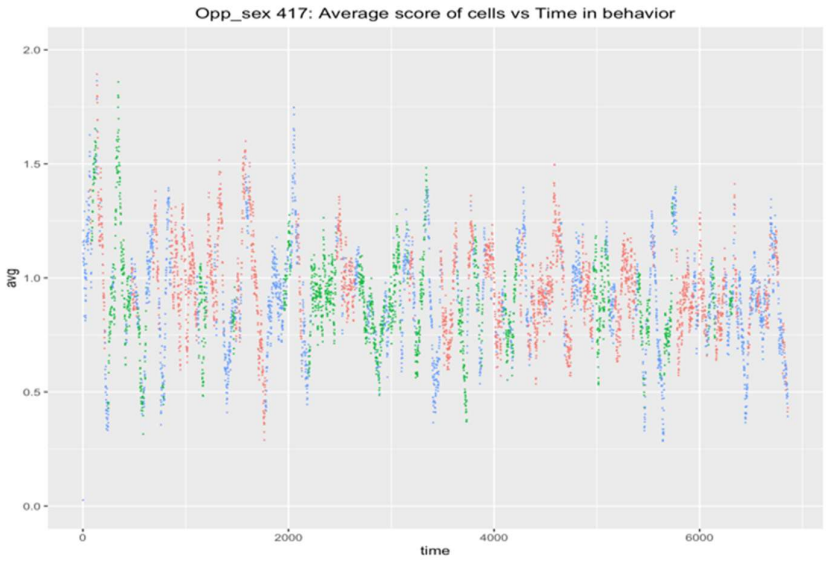
For the logistic regression model, it may be better to split the data into train and test sets so that we can test the regression fit. We can also explore more options of rolling sum like 30 and more time points before the behavior time.

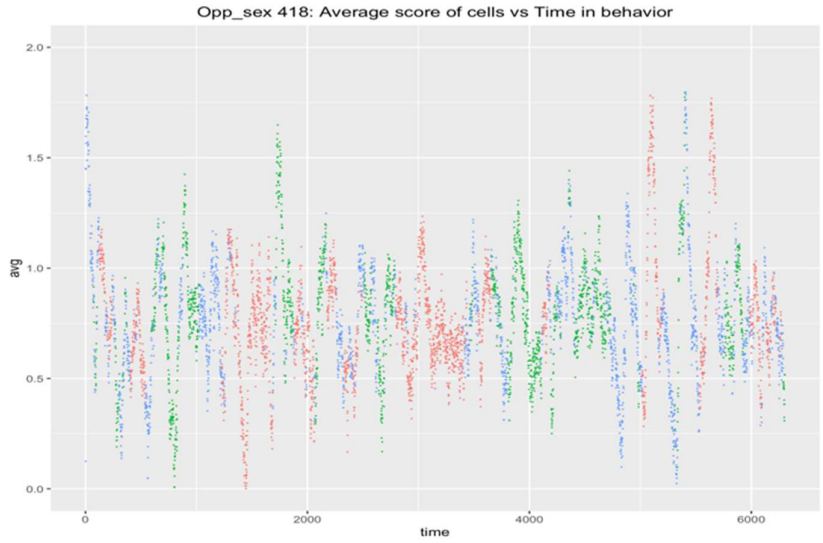
There are also some improvements for the model building. First, it is possible that 30 epochs, which is set as the maximum, is too shallow for some mice dataset and might result in underfitting. Second, we did not tune the hyper parameter which might result in comparatively lower accuracy. Thirdly, the interpretability of the model is not high. No visualizations can be generated from the model, and we did not obtain a clear relationship between cell activities and behavior. Finally, many-to-many prediction is not applied in this project, although it is feasible for this mice dataset. The relative model is similar to the one we construct in this report. For Encoder, it is necessary to use masked self-attention instead of origin self-attention.

Appendix


Opp_sex 409: Average score of cells vs Time in behavior
avg
time
0.0
0.5
1.0
1.5
2.0
0
2000
4000
6000


Opp_sex 412: Average score of cells vs Time in behavior
avg
time
0
1
2
0
2000
4000
6000





Here is our GITHUB link: https://github.com/MA679/Final-Project, you can find the code of this project on it.