Introduction to Machine Learning
Neuron Synapse Prediction Process and Results

Grace Wang
Rice University
Computational and Applied Mathematics and Operations Research
Houston, Texas 77005
gyw1@rice.edu

Didi Zhou
Rice University
Electrical and Computer Engineering
Houston, Texas 77005
dz34@rice.edu

***Abstract*—Synaptic connections allow neurons to send signals to one another, forming the foundation of understanding neural behavior and patterns.The present work outlines a machine learning approach to assess whether neurons within axonaldendritic proximity will form a synapse. The guiding principles behind our method were (1) to engineer biologically interpretable features and (2) to build a model selection process resistant to overfitting. The process began with data cleaning and preprocessing, followed by feature engineering based on an understanding of neural structures. Then, the data was split into training and query sets; with bootstrap subsamples of the training set, hyperparameters were selected for eXtreme Gradient Boosting, Random Forest, and Logistic Regression models. Finally, selected models were tested via a public and private kaggle leaderboard where our final placements were 14th and 45th respectively.**

1. Introduction

Neurons send chemical and electrical signals to each other through structures called synapses. In a synapse, the neuron whose axon sends messages is known as the presynaptic neuron and the neuron whose dendrites receive the message is known as the post-synaptic neuron [1]. In order for a synapse to form, two neurons must be within sufficiently close distance to each other, called an axonaldendritic proximity (ADP). All neurons that form a synapse arise from an ADP, but not all ADPs produce synapses [2]. Thus, it is of interest to determine what conditions apart from ADP distance determine synapse formation.

As synapses form the basis of all neural signaling, understanding their formation has become a question of immense scientific significance. As a part of the ongoing research into the formation of synapses, students in Rice University's ELEC 478/578: Introduction to Machine Learning were presented with the challenge to build a machine learning model that performs binary classification to predict whether or not an ADP will result in a synapse. The data provided originated from the MICrONS collaboration, which investigated the activity of neurons in the visual cortex of a mouse. The data consists of information on each ADP and functional and morphological data on the two neurons involved in the ADP. This report details the approach and insights of the authors to the aforementioned challenge. The approach taken by the authors was guided by two primary goals: (1) engaging in feature engineering informed by biological context and (2) building a model selection process resistant to overfitting.

2. Methods

2.1. Preprocessing

Several features were provided within the dataset, including: feature weights; morphological embeddings, axonal, nucleus, and dendritic coordinates, adp distances, skeletal to soma distances, oracle, test scores, read out location coordinates, compartment, and brain areas. When joining the morphological embeddings to presynaptic and postsynaptic neurons, several presynaptic neurons did not have corresponding morphological embeddings. Since morphological embeddings represent dendritic arbor properties as results from a graph-based self-supervised learning approach, we felt that it was necessary to have data for this feature. To do so, we employed kNN imputation. kNN imputation fills in missing data values by looking at close neighbors, so it works well when there are similarities between available morphological embeddings of observations that point to underlying distinctions between observations. As morphological embeddings correspond to physical properties, we believed that neurons that are similar in available morphological embeddings would be similar in missing morphological embeddings - hence the employment of kNN imputation.

Additionally, all categorical variables were processed using one-hot encoding. The "compartment" column in the raw data, which tells the neuronal compartment where the ADP resides on the postsynaptic neuron, was processed by grouping the compartments into three categories: cell body, axonal, and terminal areas, and one-hot encoding these three categories. Columns identifying the area of the brain

hosting the pre-synaptic and post-synaptic neurons were also processed using one-hot encoding.

2.2. Feature Engineering and Learning

As previously mentioned, a primary focus of the outlined approach is to engineer features motivated by biological principles. Through independent research and consultation with Dr. Fabricio Do Monte, a neurobiologist at McGovern Medical School, it became clear that the functional and chemical behaviors of neurons may be of great interest when determining whether or not neurons will form a synapse [3][4]. This presented itself as a slight issue, as the provided challenge data contained limited functional and chemical information, with only the morph embeddings and feature weights providing possible insight into these properties. Instead, the provided data primarily contained fields related to location data on various structures within the neurons. To reconcile this difference between desired information and provided data, the creation of features that extended location data to larger biological or functional interpretations were prioritized. The following features were engineered:

me_similarity - The cosine similarity between the morph embeddings corresponding to the pre-synaptic and post-synaptic neurons; it was hypothesized that ADPs with more similar morph embeddings would be more structurally similar and thus more likely to form a connection.

fw_similarity - The cosine similarity between the feature weights corresponding to the pre-synaptic and post-synaptic neurons; it was hypothesized that ADPs with more similar feature weights would be more functionally similar and thus more likely to form a connection.

rf_distance - The euclidean distance between the readout locations of the deep learning predictive model for the pre-synaptic and post-synaptic neurons, where the readout locations are correlated to the center location of the receptive field in visual space. It was hypothesized that connected neurons would be closer together in the receptive field. (During submission, this feature was incorrectly calculated using cosine similarity, but this has been corrected during revision.)

nuclei_adp_dist - The euclidean distance between the nucleus of the presynaptic neuron and the axonal ADP coordinates. Electrical signals originate from the nucleus of the presynaptic neuron and dissipate along the length of the unmyelinated axon [5]. It was hypothesized that neurons with a greater distance between the nucleus and the axonal ADP would have weaker electrical signals and thus be less likely to form synapses.

minicol_dist - The euclidean distance between the x and y coordinates of the nuclei of the pre-synaptic and post-synaptic neurons. Minicolumns are vertical clusters of neurons that are hypothesized to be fundamental functional units for cortical processing. In rodents, minicolumns within the primary visual cortex are about in 5 μm radius and 15 μm apart from each other [6]. As such, minicol_dist was designed to evaluate whether two neurons are close together in two-dimensional space, as neurons that are closer together are more likely to be in the same minicolumn. Additionally, studies have found that the number of minicolumns in the primary visual cortex is strongly correlated with visual acuity [7]. Given this information, it was hypothesized that two neurons in the same minicolumn may be more likely to form a connection, as a greater number of minicolumns indicates greater visual abilities which may be enabled by greater synaptic activity taking place within the minicolumns.

Following the creation of these features, all numerical features were standardized to account for differences in units.

2.3. Classification Models

To determine classification models for our machine learning approach, we reviewed several popular models that perform particularly well with classifying tabular data. The first of which was Linear Discriminant Analysis (LDA). We found that LDA performed well at the beginning of our process, but soon plateaued in performance since it is a simple linear model with fewer hyperparameters that play a large role relative to other models. Hence, we decided to move forward without LDA. Since LDA plateaued for us, we decided to try to employ a Multi-Layer Perceptron (MLP) for its expressiveness. At this point in our process, we had not fully developed a robust validation process for tuning so we found that our model performed poorly on the public leaderboard. To strike a middle ground between very expressive models and more simple linear models, we turned towards ensemble techniques.

Next, we explored Random Forest. Random Forest is an ensemble technique that uses many weak decision trees to perform classification or regression. In our case, we found that Random Forest actually did not classify particularly well without any tuning; however, after tuning how many features were considered at each split for the decision trees as well as the depth of each decision tree, Random Forest began to perform better and better without overfitting. Random Forest was included in our trio of final models.

At the same time, we investigated Logistic Regression since it places less assumptions on the input data and also typically performs better with binary classification. Logistic Regression performed well on the public leaderboard and in our internal validation process, so we included it in our trio of models, tuning it via the 'C' parameter.

We decided to continue trying ensemble techniques by shifting from a parallel ensemble of decision trees with Random Forest to employing a sequential ensemble with eXtreme Gradient Boosting (XGBoost). Sticking with decision trees, we tuned XGBoost on tree depths, learning rates, and lambda. We found that XGBoost performed better than Logistic Regression in our internal validation process and nearly performed as well as Random Forest. Hence, we considered XGBoost as one of our three models.

Generally, we tried to have a diverse set of models in our final set such that expressiveness, simplicity, and computational complexity were represented within our set. Thus, the final three models we considered for our entire machine learning process were Random Forest, XGBoost, and Logistic Regression.

2.4. Machine Learning Process

Another primary goal of this project was to build a model selection process resistant to overfitting. As such, data splitting was considered with utmost importance in the process of validating and selecting models. In our process, data was first split into a training and query set, with an 80/20 split on pre-synaptic neuron ID, which yielded an approximate 80/20 split on all neurons. Splitting was performed on pre-neuron IDs to ensure that models generalize well to unseen pre-synaptic neurons.

The validation process iterated over a grid of hyperparameters for each of the base learners. Certain hyperparameter grids were explored for submission to the competition, but following the revealing of the competition results, these hyperparameter grids were revised and validation was performed again. The before and after of our hyperparameter choices can be seen in Figure 1.

Model	Hyperparameters (Submission)	Hyperparameters (Revision)
XGBoost	Maximum tree depth = 3, 4, …, 14 Learning rates = 0.1, 0.2, …, 1.0 Lambda = 0, 0.1, …, 2.0	Maximum tree depth = 1, 2, …, 10 Learning rates = 0.1, 0.2, …, 1.0 Lambda = 0, 0.1, …, 2.0
Random Forest	Maximum tree depth = 1, 2, …., 10 Number of features = 1, 2, …, 20	Maximum tree depth = 1, 2, …., 10 Number of features = 1, 3, …, 19 Number of Estimators = 100, 1000
Logistic Regression	C value = 0.1, 0.6, 1.1, …, 9.6	C value = 0.1, 0.6, 1.1, …, 9.6

Figure 1. Models and Corresponding Hyperparameters for Validation

For each set of hyperparameters, five models were trained and validated using five bootstrap subsamples of the training data. In this process, the training data was split five times into training and validation sets with an 80/20 split on pre-synaptic neuron ID. Each of the five models was trained on its corresponding training set and evaluated on its validation set. For the submitted models, the average balanced accuracy of the five models was recorded. For the revised models, the average balanced accuracy, standard deviation of the balanced accuracies, and average training time of the five models were recorded. To select the best hyperparameters for each base learner, the average balanced accuracies and standard deviations were compared. This process can be thought of as a cross-validation strategy that has been adjusted to account for concerns surrounding isolating pre-neuron IDs to either the training or validation set for greater generalizability.

Finally, the best models from each base learner were fit to the training data with their optimal hyperparameters and evaluated on the query data. The models that performed best on the query data were chosen for submission to the private leaderboard.

From the outlined process, it is apparent that our team opted for a methodology that aims to select models and hyperparameters that generalize well to new data by undergoing a rigorous splitting and validation process.

3. Results

As a part of the competition that this research was conducted under, there were public and private leaderboards that indicated the balanced accuracy of models on a new dataset.

3.1. Public Leaderboard Performance

Before the final machine learning process outlined in the previous section was developed, we explored different base learners, feature selection strategies, and validation processes. A history of our most notable public leaderboard performances and their corresponding models and methodologies is given below.

1) 0.73853 - Random Forest (11/30/2023)
   - Maximum tree depth: 5
   - Number of features at each split: 9
   - Internal validation accuracy: 0.774
   - Internal query accuracy: 0.753
2) 0.75622 - XGBoost (11/30/2023)
   - Maximum tree depth: 3
   - Learning Rate: 1.2
   - Lambda: 0.1
   - Internal validation accuracy: 0.768
   - Internal query accuracy: 0.751
3) 0.75622 - Logistic Regression (11/30/2023)
   - C = 2.1
   - Internal validation accuracy: 0.748
   - Internal query accuracy: 0.747
4) 0.55694 - MLP (11/26/2023)
5) 0.75409 - LDA (11/19/2023)
   - XGBoost was used to select the most important feature weights, considered as a feature in the LDA model
6) 0.49963 - LDA (11/10/2023)
   - Used tSNE as a dimensionalty reduction method

3.2. Private Leaderboard Performance

For submission to the private leaderboard, we decided to choose from the three models generated from our final machine learning pipeline. We submitted the Random Forest and XGBoost models for submission, since they performed best on our internal query set and we did not want to overfit to the public leaderboard. The final scores of the three models are below:

1) 0.78559 - Logistic Regression
2) 0.76014 - Random Forest
3) 0.74307 - XGBoost

Our performance on the public and private leaderboard can be summarized in Figure 2.

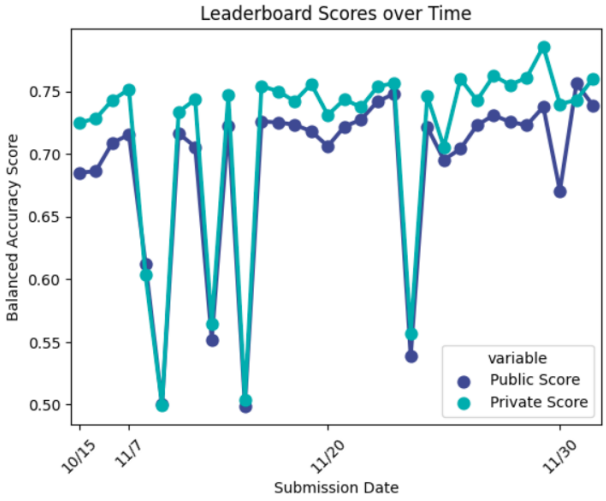

Figure 2. Leaderboard Scores over Time

3.3. Interpretability

To understand which features may be the most indicative of synapse formation, we investigated the feature importances generated by each model. As expected, adp_dist was the most important feature identified by Random Forest and XGBoost. This indicates that the distance between presynaptic and postsynaptic neurons is still a significant factor in whether or not a synapse will be formed, even when the two neurons are within an axonal dendritic proximity from each other. Furthermore, the feature RL_pre, which is a one-hot encoded column indicating that the presynaptic neuron resides in the rostrolateral visual area (RL), was also relatively important in Random Forest and XGBoost, where it was the second and third most important feature, respectively. This was interesting, as exploratory analysis of the training data revealed that ADPs with presynaptic neurons in the RL did not connect at a significantly higher rate than ADPS with presynaptic neurons elsewhere. It is possible that the importance of this feature was influenced by our data splitting methods, which divided neurons based on pre-synaptic neuron ID; it is possible that the neuron IDs chosen for the training set correspond heavily to neurons located in the RL. Lastly, for both Random Forest and XGBoost, pre_skeletal_distance_to_soma, or the path length from the axonal ADP to the pre-neuron soma, was also identified as important. This could be evidence that the distance between the originating location of a signal (the soma) and its transmission location (the axon) is significant, similar to our hypothesis surrounding the distance between the nucleus and axon.

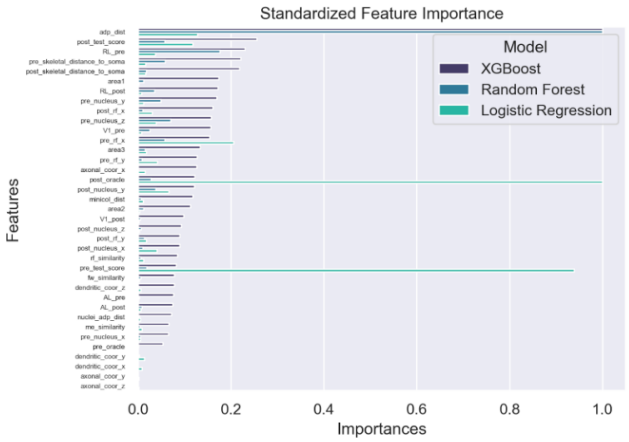

Figure 3. Feature Importances by Model

Interestingly, Logistic Regression identified different important features as opposed to the tree-based models. With a C-value of 2.1, Logistic Regression performed the most aggressive regularization on feature importances of all models. The most important feature as identified by Logistic Regression was post_oracle, which describes the neuronal response reliability of the postsynaptic neuron. Upon exploratory analysis of the training data, it was revealed that connected neurons tend to have lower post_oracle scores, which indicate less reliable responses to repeated visual stimulation (Figure 4). This is counterintuitive to our expectations, as we thought that connected neurons would more reliably respond to repeated visual stimuli, as they are able to communicate with other neurons. However, one possible explanation for this behavior is that connected postsynaptic neurons are less responsive to repeated visual stimulation because they are too overloaded with incoming information from their connections to effectively respond to additional stimulation. Lastly, in Logistic Regression, pre_test_score and post_test_score were also significant. These features describe the predictive performance of the deep learning predictive model on withheld test trials for the presynaptic and postsynaptic neurons, respectively. XGBoost also identified post_test_score as a relatively important feature. This makes sense, as neurons that are more predictable according to another deep learning predictive model will likely be more predictable using our models.

4. Discussion

Reflecting upon our results, there are many potential avenues for improvement that could be explored by future research and amendments to our current machine learning process.

4.1. Data Processing

Initially, we hoped to include dimension reduction as a form of data processing and exploration within our machine learning process. We attempted Principal Component

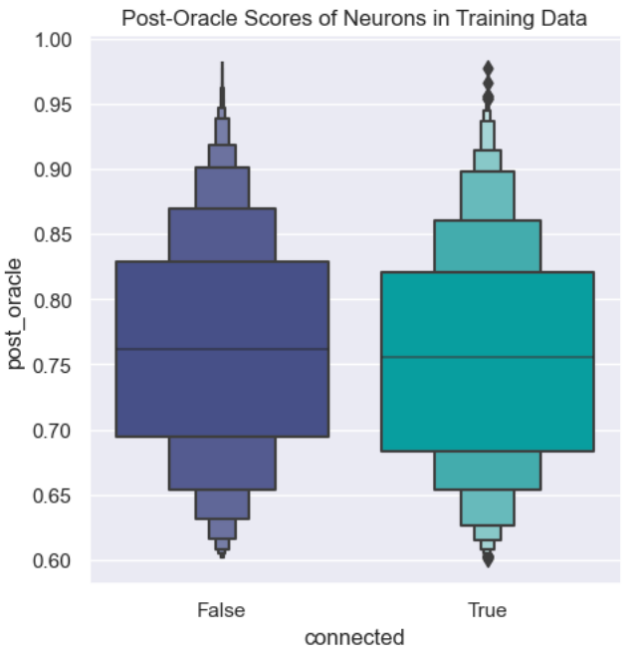

Figure 4. Post Oracle Scores of Connected and Disconnected ADPs

Analysis and t-Distributed Stochastic Neighbor Embedding to see if unsupervised methods would cause our data to form patterns that could illuminate interesting findings. In particular, we hoped that these methods might reveal clusters of neurons separated by their functionality. This was motivated by our conversation with neuroscientist Dr. Fabricio H.M. Do Monte of the McGovern Medical School. In our discussion, he emphasized the importance of the functionality of a neuron and compatible neurotransmitters between neurons in the formation of a synapse. As such, we hoped that dimension reduction would reveal clusters of neurons by function or neurotransmitter compatibility. However, we were ultimately unsuccessful and decided to move forward without using dimension reduction. Future works could further explore this avenue.

4.2. Features

It is worth noting that none of the features we engineered were identified as particularly significant by the feature importances assigned by our models. With regards to the feature minicol_dist, it could be useful to revisit the concept of minicolumns in the future when research establishes a more explicit relationship between minicolumns and neural functioning in the visual cortex. Additionally, future work could improve upon our approach to categorical features. Expanded biological knowledge could be used to create more meaningful features from categorical columns as opposed to our approach, which relied heavily upon one-hot encoding.

While our engineered features were not particularly important, feature importances revealed the significance of adp_dist, RL_pre, pre_skeletal_distance_to_soma, post_oracle, and pre_ and post_ test_score. As previously mentioned, these features could be areas of interest for future research, especially post_oracle, which claims that connected neurons respond less reliably to repeated visual stimuli. The importance of RL_pre as a feature also reveals a possible weakness in our data splitting strategy, as it is possible that a disproportionate amount of neurons with unique pre-nucleus IDs are located in the RL. Future iterations of this process could therefore investigate different data splitting methods, such as performing stratified sampling of pre-nucleus IDs across brain regions.

4.3. Model Selection

Following the release of the private leaderboard results, our team discovered that our tree-based models performed worse than our Logistic Regression model, despite the tree models performing better on our internal query set, internal validation sets, and public leaderboard. In response to this, we decided to modify our validation approach to record the standard deviation in model performance for each hyperparameter set to determine which methods produce the most stable predictions. From these results, we saw that Logistic Regression produced the greatest standard deviation in accuracy between models (0.021875) as compared to XGBoost (0.004576) and Random Forest (0.008084). It is possible that this variance worked in the model's favor in its performance on the private leaderboard, so further testing of the model's performance could be warranted. It is also possible that Logistic Regression performed the best on the private leaderboard because it is generally more resistant to overfitting than tree models. This came in handy because the model did not seem to overfit to RL_pre, which as we previously discussed was a feature that gained unintended importance due to our data splitting methods.

4.4. Adjusted Hyperparameters

To improve our process further, we decided to repeat our validation approach with an expanded number of hyperparameters following the release of the private leaderboard results. Specifically, we allowed for shallower trees in XGBoost and introduced the number of estimators for Random Forest as a hyperparameter to be tuned. This approach produced the following results:

1) Random Forest
   - Maximum tree depth: 6
   - Number of features at each split: 7
   - Number of Estimators: 100
   - Internal validation accuracy: 0.774
   - Internal query accuracy: 0.753
   - Public accuracy: 0.754
   - Private accuracy: 0.752
2) XGBoost
   - Maximum tree depth: 3

- Lambda: 0.6
- Learning rate: 0.1
- Internal validation accuracy: 0.770
- Internal query accuracy: 0.752
- Public accuracy: 0.755
- Private accuracy: 0.751

3) 0.75622 - Logistic Regression (11/30/2023)

- C-value: 0.1
- Internal validation accuracy: 0.740
- Internal query accuracy: 0.751
- Public accuracy: 0.737
- Private accuracy: 0.781
- While the hyperparameters tested for this model did not change between submission and revision, we yielded different results likely due to the change in rf_distance

These adjustments indicated an improvement in the performance of our XGBoost model on the private leaderboard. Further research could be performed by tuning a greater range of hyperparameters to identify the most optimal model.

5. Acknowledgment

The authors of this work would first like to acknowledge Dr. Genevera Allen and the Tolias Lab for the formation of this project. The contributions of author Grace Wang focused on feature engineering, data cleaning, model validation, and the final pipeline assembly. Author Didi Zhou contributed by developing the validation process and initial pipeline assembly work.

References

[1] M. J. Caire and M. Varacallo, "Physiology, Synapse," Nih.gov, Nov. 13, 2018. Available: https://www.ncbi.nlm.nih.gov/books/NBK526047/. [Accessed: Dec. 09, 2023]

[2] J. van Pelt and A. van Ooyen, "Estimating neuronal connectivity from axonal and dendritic density fields," Frontiers in Computational Neuroscience, vol. 7, no. 160, Nov. 2013, doi: https://doi.org/10.3389/fncom.2013.00160

[3] T. C. Südhof, "The cell biology of synapse formation," The Journal of Cell Biology, vol. 220, no. 7, p. e202103052, Jun. 2021, doi: https://doi.org/10.1083/jcb.202103052. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8186004/. [Accessed: Dec. 10, 2023]

[4] T. C. Südhof, "Towards an Understanding of Synapse Formation," Neuron, vol. 100, no. 2, pp. 276–293, Oct. 2018, doi: https://doi.org/10.1016/j.neuron.2018.09.040

[5] M. H. Grider, C. Q. Belcea, B. P. Covington, and Sandeep Sharma, "Neuroanatomy, Nodes of Ranvier," Nih.gov, Jun. 28, 2019. Available: https://www.ncbi.nlm.nih.gov/books/NBK537273/. [Accessed: Dec. 11, 2023]

[6] S. Kondo, T. Yoshida, and K. Ohki, "Mixed functional microarchitectures for orientation selectivity in the mouse primary visual cortex," Nature Communications, vol. 7, no. 1, Oct. 2016, doi: https://doi.org/10.1038/ncomms13210

[7] M. N. Wallace et al., "The large numbers of minicolumns in the primary visual cortex of humans, chimpanzees and gorillas are related to high visual acuity," Frontiers in Neuroanatomy, vol. 16, no. 1034264, Nov. 2022, doi: https://doi.org/10.3389/fnana.2022.1034264. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9681811/. [Accessed: Dec. 10, 2023]