Decoding Neural Connections: A Machine Learning Approach

Ahmet Kilic, Bayzhan Mukatay, Genevera Allen

Rice University, ELEC 478

abk5@rice.edu

bm61@rice.edu

Houston, Texas, USA

***Abstract*— In this study, our team explores neural connectivity using logistic regression, balanced random forest, Gaussian naive Bayes, and gradient boosting approaches. Utilizing the MICrONS dataset from the visual cortex of mice, we predict axonal-dendritic conversions to synapses in cortical pyramidal neurons. Emphasizing balanced accuracy metrics in the Kaggle competition, we tackle class imbalance and implement meticulous feature engineering. The report outlines our technical strategy, encompassing preprocessing, model selection, and creative aspects of our methodology. As we present our findings, we reflect on the competition's challenges and breakthroughs, underscoring the potential impact of our machine learning innovations on advancing neuroscience's understanding of neural connections. Our contributions include engineered features and a notable 78.8% accuracy.**

***Keywords*— Machine Learning, Neural Connectivity, Synaptic Prediction, Cortical Pyramidal Neurons, Data Analysis.**

I. Introduction

Navigating the intricacies of neural connectivity, our study unfolds against the backdrop of the MICrONS collaboration, utilizing data that seamlessly integrates 2-photon microscopy and electron microscopy to explore the visual cortex of mice. Our mission: to predict the formation of synapses, the critical junctions facilitating neural communication, and to unravel the rules dictating axonal-dendritic proximities (ADPs) that lead to synapse conversion.

Embedded in a unique competition setting, our focus lies on a binary classification task—predicting whether an ADP transforms into a synapse. However, this task comes with a distinctive challenge: an inherent imbalance in the classes, where a multitude of ADPs do not culminate in synapses. Complicating matters further, during the competition, we could gauge our performance through a public leaderboard accuracy, but the elusive private leaderboard accuracy remained concealed until the competition's close. The competition places a premium on balanced accuracy metrics, underscoring the necessity for robust predictions within the context of imbalanced classes.

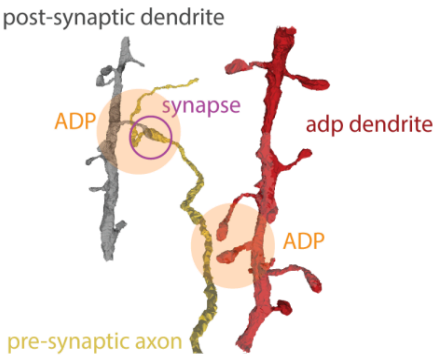

Figure 1: pre-synaptic axon forming a synapse with the post-synaptic dendrite

To comprehend our approach fully, a foundational understanding of neuronal compartments, the intricacies of ADPs, and the role of synapses in neural communication is imperative. Cortical pyramidal neurons, the focus of our dataset, exhibit distinct compartments—dendrites, soma, and axons—each playing a vital role in signal processing. Synapses materialize when axons and dendrites closely interact—form an ADP. However, not every ADP results in a synapse, making it a subject of paramount interest to uncover the determinants governing this conversion.

Our paper meticulously details the technical strategies employed in our machine learning process, starting from preprocessing the data, tp selecting appropriate models, engineering innovative features and generalizing our models. As we unveil our findings, we reflect on the challenges encountered during the competition. We underscore the potential impact of our machine learning innovations on advancing neuroscience's understanding of neural connections. Through our endeavors, we aspire to not only excel in the competition but also significantly contribute to ongoing neuroscience research, propelling our understanding of synaptic connections to new frontiers.

II. Methods

Data Exploration.

There were 185832 ADP observations in all, and some of these would lead to synaptic connections. Therefore, the first thing we did when analyzing the data was to figure out how the synapse connections in our sample were distributed. Upon initial data examination, a substantial class imbalance emerged, comprising 1366 synapse instances and 184,466 non-synapse instances. Further scrutiny of neuronal distribution unveiled that ADPs originated from 77 pre-neurons and 2663 post-neurons, with connected ADPs exclusive to 72 pre-neurons and 978 post-neurons. The pronounced class imbalance posed challenges for model learning and generalization, especially when predicting synapses for new, unseen neurons.

Feature Engineering.

Feature Weights Similarity. To capture the potential connection between neurons with similar tuning, we introduced a new feature, fw_similarity, representing the cosine similarity between the pre- and post-neuron feature weights. Feature weights encapsulate the tuning function of each neuron, and higher similarity suggests a potential synaptic connection.

Projection Regions. Brain regions, defined by projection regions, play a role in neural functions. We engineered a new feature by one-hot encoding the brain regions where pre-synaptic and post-synaptic neurons are located. This feature explores the hypothesis that similar neural tuning is associated with connections in specific projection regions.

Nucleus ID Counts. To gain insights into neuron activity and interactions, we introduced four new columns: pre_pre, pre_post, post_post, and post_pre, counting the occurrences of pre and post nuclei in the roles of axons and dendrites. This feature enhances our understanding of neuron dynamics and aids in distinguishing between more and less active neurons.

Euclidean Distances. Utilizing the spatial coordinates of pre- and post-neurons, we created columns representing the Euclidean distances between them. This feature provides a geometric perspective on neuron proximity, contributing to the model's ability to capture spatial relationships.

Morphological Distances. Initially, we attempted to capture morphological differences between neurons by calculating the distance between pre- and post-morphological embeddings. The process involved replacing missing values, applying imputation with the MissForest algorithm, and calculating distances between morphological embeddings, which is a computationally expensive task. However, due to marginal benefits and low feature importance, this feature was not used later.

Machine Learning Pipeline.

Embarking on the intricate journey into neural connectivity, our study harnesses a diverse ensemble of machine learning models to unravel the intricacies within the MICrONS collaboration dataset. Each model contributes a distinctive perspective to the predictive process:

Balanced Random Forest:

A potent solution for imbalanced class distributions, the balanced random forest constructs decision trees, aggregating predictions to withstand challenges presented by disparate synapse and non-synapse instances.

Logistic Regression:

Serving as our foundational model, logistic regression's simplicity and interpretability provide a crucial baseline, offering insights into the nuanced relationship between features and synapse likelihood.

Gaussian Naive Bayes (NB):

Grounded in probabilistic principles, Gaussian Naive Bayes introduces a probabilistic approach, navigating the dataset with efficiency and assuming conditional independence among features. Its applicability to large-scale datasets enhances its value in our ensemble.

K-Nearest Neighbors (KNN):

Adding a spatial dimension to our ensemble, K-Nearest Neighbors classifies instances based on the consensus of their nearest neighbors, demonstrating adaptability to diverse data distributions and the ability to capture local patterns.

Gradient Boosting:

Concluding our ensemble, gradient boosting employs a sequential approach to model construction. Correcting errors iteratively, it excels in capturing complex relationships within the dataset, elevating overall predictive accuracy.

The primary challenge lay in finding a model generalizable to new, unseen neurons—a task made arduous by the dataset's pronounced class imbalance. To tackle this, advanced strategies were employed, including oversampling of the minority class for Logistic Regression and Gaussian Naive Bayes, while the balanced random forest, by design, circumvented such preprocessing.

In terms of cross-validation, our approach diverged from random splits. Instead, we ensured a proportional distribution of the 72 unique pre-neurons across training, testing, and validation sets. This constraint facilitated a more representative evaluation of model performance, particularly in scenarios involving novel pre-neurons.

Our approach consisted of the following steps:

1. Forced train-test-validation split with proportional pre-neuron distribution.
2. Hyperparameter tuning on train and validation sets.
    a. For KNN, Logistic Regression and Gaussian NB, we also oversampled the minority class of ADP that formed a synapse.
3. Test Accuracy.
4. Retrain the model on all data.
5. Report on the Leaderboard data: used for model selection also.

III. Results

Predictive Performance

For nearly every one of our Kaggle submissions, we received higher private leaderboard scores than public leaderboard scores. A small number of cases resulted in slightly lower ratings, which allowed us to conclude that the models we submitted were well-generalizable and robust to overfitting. Additionally, the following table displays our predictive performance on both public and private leaderboards along with the noteworthy modification to our code:

Public Score	Private Score	Change in Code
73.759	73.284	Logistic Regression
78.092	72.441	Added Nucleus ID counts
78.447	74.667	Used Balanced Random Forest
78.254	75.083	Created validation set with new neurons that are not in the training set.
78.637	75.473	Further tuning of Balanced Random Forest.

Interpretability

Summarizing our findings, let's start by examining the feature importance plots for our final Balanced Random Forest. The model was fine-tuned to the following cross-validated hyperparameters: 1000 trees in the forest, a minimum of 2 samples required to split an internal node, a minimum of 4 samples in each leaf node, and consideration of up to 7 features at each split.

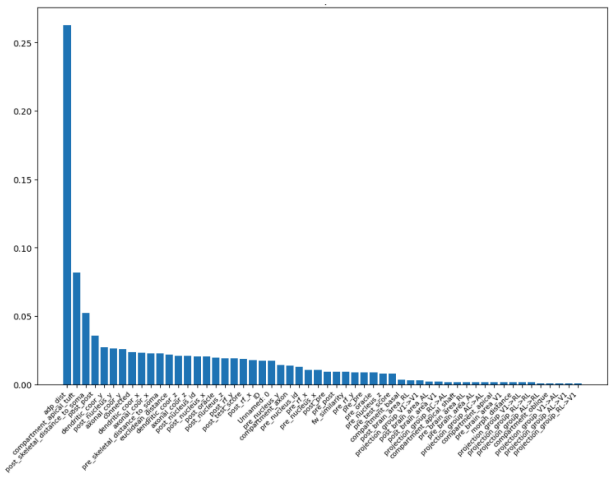

Figure 2: Balanced Random Forest feature importance

Upon examination, pivotal features in our analysis include ADP distance, compartment 'apical tuft,' and post-skeletal distance to soma. Notably, the recently engineered feature 'post_post,' reflecting occurrences of the ADP-forming post-neuron as a post neuron for synapses and other ADPs, also ranks among the most influential. Additionally, the newly crafted 'euclidean_distance' feature secured a position within the top 12, showcasing its significance comparable to other key features in our analysis.

Distilling our findings, let's now explore the significance of features through the feature importance plots for our finalized Logistic Regression model. The hyperparameters guiding this refined model are specified as follows: a maximum of 1000 iterations (max_iter), a regularization parameter C set to 0.25, a penalty term of 'l1', and the utilization of the 'liblinear' solver.

In the realm of Logistic Regression, the paramount feature is identified as the x-coordinate of the nucleus of the neuron, closely trailed by the nucleus ID and features associated with brain areas. Noteworthy is the divergence in feature importance compared to the Balanced Random Forest, with the Logistic Regression model assigning less prominence to the 'adp_distance' feature.

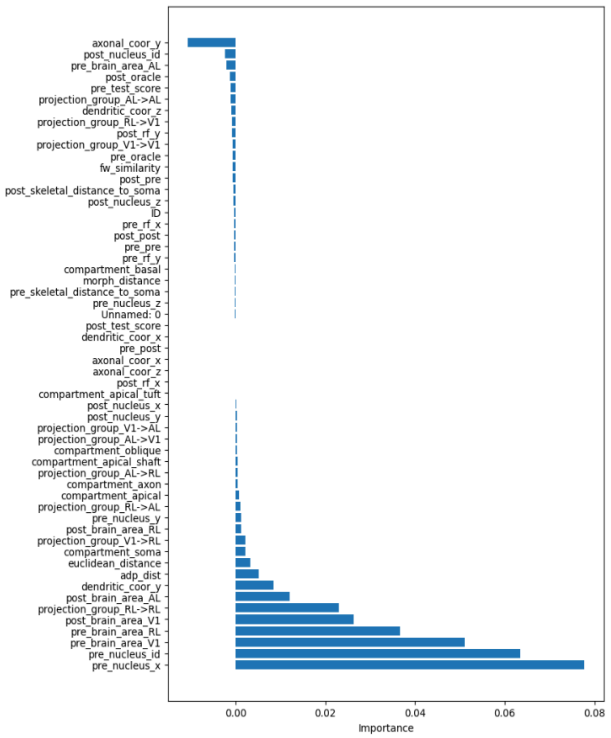

Figure 3: Logistic Regression feature importance

Predictions

Below is a table encompassing the models and their associated balanced accuracy scores on our test data, public and private leaderboards.

Model	Test Accuracy	Public Accuracy	Private Accuracy
Balanced RF	76.9%	75.8%	78%
Logistic Reg	74.9%	73.9%	78.4%
Gaussian NB	73.2%	N/A	N/A
Gradient Boost	51%	N/A	N/A

TABLE I
Model Performances

IV. Discussion

Reflecting on our journey through the competition, several key learnings and achievements emerge, paving the way for insightful discussions and considerations for future endeavors.

Balanced Random Forest (Balanced RF): Achieving a commendable 78% on the private

leaderboard demonstrates the efficacy of Balanced RF in providing generalizable and consistently strong performance. The minor overfitting observed is a testament to the inherent robustness of the random forest model, which inherently resists overfitting.

Logistic Regression. With an average performance of 76%, our logistic regression model also showcased robust capabilities. While not as high as Balanced RF, its consistent performance adds a valuable dimension to our ensemble.

Feature Engineering Contributions. Innovative Features: The introduction of six new features, including 'euclidean distance' and 'morph_distance,' proved to be instrumental. Notably, 'euclidean distance' and 'post_post' emerged with substantial feature importance, underscoring their significance in capturing relevant patterns.

Effective Data Splitting and Hyperparameter Tuning. Proportional Distribution of Minority Pre-Nucleus_ID: Our approach to train-test splitting, ensuring a proportional distribution of the minority pre-nucleus_ID, emerged as a strategic choice for the highly unbalanced data. This method not only contributed to more representative model evaluation but also played a crucial role in guiding hyperparameter tuning.

Overfitting Considerations. While our models demonstrated impressive performance, it's crucial to acknowledge the presence of some overfitting. This insight prompts us to explore ways to fine-tune models further and enhance generalization. Delving deeper into strategies to mitigate overfitting will be a key focus. This may involve refining feature selection and implementing advanced regularization techniques.

Future Directions. Ensemble Stacking: Future work will explore the synergies of ensemble stacking, combining the strengths of Balanced RF, logistic regression, and Gaussian Naive Bayes. While computational constraints limited our exploration in this direction during the competition, it presents a promising avenue for enhanced predictive capabilities.

In conclusion, our participation in the competition not only yielded commendable results but also illuminated pathways for future improvements and exploration. The interplay between model performance, feature engineering, and data preprocessing provides a rich landscape for ongoing research and innovation in the field of neural connectivity prediction.

Acknowledgment

We extend our sincere gratitude to the Tolias Lab at Baylor College of Medicine for generously providing the invaluable MICrONS collaboration dataset. Their contribution forms the bedrock of our exploration into neural connectivity.

Special appreciation goes to Genevera Allen for her insightful teachings and guidance throughout the competition. Her expertise played a pivotal role, particularly in the inception of key features such as Feature Weights Similarity and Projection Regions.

Thanks to Ahmet Kilic for his crucial role, spanning hyperparameter tuning, and extensive exploration of models and features.

We extend our gratitude to Bayzhan Mukatay for his involvement in the project, contributing to the model exploration, engineering features, and report.

References

[1] MICrONS Consortium, J Alexander Bae, Mahaly Baptiste, Caitlyn A Bishop, Agnes L Bodor, Derrick Brittain, JoAnn Buchanan, Daniel J Bumbarger, Manuel A Castro, Brendan Celii, et al. Functional connectomics spanning multiple areas of mouse visual cortex. BioRxiv, pages 2021–07, 2021.

[2] Brendan Celii, Stelios Papadopoulos, Zhuokun Ding, Paul G Fahey, Eric Wang, Christos Papadopoulos, Alexander B Kunin, Saumil Patel, J Alexander Bae, Agnes L Bodor, et al. Neurd: automated proofreading and feature extraction for connectomics. bioRxiv, 2023.

[3] Leila Elabbady, Sharmishtaa Seshamani, Shang Mu, Gayathri Mahalingam, Casey Schneider-Mizell, Agnes Bodor, J Alexander Bae, Derrick Brittain, JoAnn Buchanan, Daniel J Bumbarger, et al. Quantitative census of local somatic features in mouse visual cortex. bioRxiv, pages 2022–07, 2022.

[4] Eric Y Wang, Paul G Fahey, Kayla Ponder, Zhuokun Ding, Andersen Change, Taliah Muhammad, Saumil Patel, Zhiwei Ding, Dat T Tran, Jiakun Fu, et al. Towards a foundation model of the mouse visual cortex. bioRxiv, pages 2023–03, 2023.

[5] Marissa A Weis, Stelios Papadopoulos, Laura Hansel, Timo L¨uddecke, Brendan Celii, Paul G Fahey, J Alexander Bae, Agnes L Bodor, Derrick Brittain, JoAnn Buchanan, et al. Large-scale unsupervised discovery of excitatory morphological cell types in mouse visual cortex. bioRxiv, pages 2022–12, 2022