# <u>Predicting Synaptic Activity Based on Axonal-</u> <u>Dendritic Proximity</u>

Michael Popa (Kaggle Name: mrew01)

ELEC 478 Machine Learning

Dr. Allen Genevera

Rice University, Fall 2023

Public Leadboard - Placed 22nd

Private Leadboard - Placed 4th

## The Goal:

When the axon of one neuron comes close to the dendrite of another neuron, we call this an <u>axonal-dendritic proximity</u>, or **ADP**. Every synapse results from an ADP, but not every ADP results in a synapse. When the ADP contains a synapse, we say the ADP has been "converted" to a synapse. It is of major interest to neuroscience what rules determine whether ADP's will convert to a synapse. In other words, what drives neurons to connect to each other, given that are within proximity (i.e. physically close enough)?

The machine learning task is to predict whether an ADP results in a synapse, and is hence a binary classification problem. There are many, many more ADPs that do not result in a synapse, however, resulting in unbalanced classes. Because of this, the competition will used the balanced accuracy metric to determine the public and private leaderboard winners. Note that during the competition, submissions will be scored according to the public leaderboard and the private leaderboard will only be revealed at the close of the competition.

This competition is part of a major ongoing research project in neuroscience. Being able to predict when an ADP will convert to a synapse and explain why this happens, is an open problem in neuroscience and a new solution would represent a major advance.

## Load In and Merge Data

Concatenate the feature weights and morphological embeddings into np.array's to decrease memory usage, then merge the feature weights and morphological embeddings into the training data.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

#load in training data on each potential synapse
data = pd.read_csv("./train_data.csv")

#load in additional features for each neurona
feature_weights = pd.read_csv("./feature_weights.csv")
morph embeddings = pd.read_csv("./morph embeddings.csv")
```

```
In [2]: # join all feature_weight_i columns into a single np.array column
        feature weights["feature weights"] = (
           feature weights.filter(regex="feature weight ")
            .sort index(axis=1)
            .apply(lambda x: np.array(x), axis=1)
        # delete the feature weight i columns
        feature weights.drop(
            feature weights.filter(regex="feature weight ").columns, axis=1, inplace=True
        # join all morph embed i columns into a single np.array column
        morph embeddings["morph embeddings"] = (
           morph embeddings.filter(regex="morph emb ")
            .sort index(axis=1)
            .apply(lambda x: np.array(x), axis=1)
        # delete the morph embed i columns
        morph embeddings.drop(
           morph embeddings.filter(regex="morph emb ").columns, axis=1, inplace=True
In [3]: data = (
            data.merge(
               feature weights.rename(columns=lambda x: "pre " + x),
               how="left",
               validate="m:1",
               copy=False,
            )
            .merge(
               feature weights.rename(columns=lambda x: "post " + x),
               how="left",
               validate="m:1",
               copy=False,
            )
               morph embeddings.rename(columns=lambda x: "pre " + x),
               how="left",
               validate="m:1",
               copy=False,
            .merge(
```

# **Quantifying the Data**

how="left",
validate="m:1",
copy=False,

Characteristics of our dataset:

)

- There are a total of 185,832 potential synapses. Out of these,
  - We have 1,366 that **are** connected (synaptic connections)
  - We have 184,446 that are not connected (did not result in a synapse)

morph embeddings.rename(columns=lambda x: "post " + x),

Out of these 1,366 synapses, they result from 77 pre-synaptic neurons and 2,663 post-synaptic neurons

Clearly, the dataset is extremely unbalanced. We will have to employ resampling techniques like undersampling or oversampling to deal with this, and/or focus on ensemble methods that would be effective at dealing with misclassification.

```
In [4]: import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from tqdm import tqdm
from sklearn.metrics.pairwise import cosine_similarity

print("Number Connected (Synapses):",data['connected'].sum())

print("\nNumber Not Connected:",(~data['connected']).sum())

print(f"\nAll the adps are from {data['pre_nucleus_id'].nunique()} pre- neurons and {dat
Number Connected (Synapses): 1366

Number Not Connected: 184466

All the adps are from 77 pre- neurons and 2663 post- neurons.

**DataFrame**
```

Our dataset comes with 34 features (30 + our 4 new merged features) that we can better understand in the data documentation PDF file.

The features so far include spatial coordinates, the brain area of the neurons (AL, RL, and V1), metrics of neuronal response (oracle - response reliability to repeated presentation; test\_score - how well the model predicts response based on input), morphological embeddings, distance metrics, the ID of the neuron, and whether a connection (synapse) is present or not.

```
In [17]: print("Data Size:", data.shape)
        print("\nData:")
        data.info()
        Data Size: (185832, 34)
        Data:
        <class 'pandas.core.frame.DataFrame'>
        Int64Index: 185832 entries, 0 to 185831
        Data columns (total 34 columns):
                                           Non-Null Count Dtype
         # Column
        ---
                                            _____
                                           185832 non-null int64
         0
           ID
         1 axonal coor x
                                          185832 non-null int64
                                          185832 non-null int64
         2 axonal coor y
                                          185832 non-null int64
         3 axonal coor z
         4 dendritic_coor_x
5 dendritic_coor_y
6 dendritic_coor_z
                                          185832 non-null int64
                                          185832 non-null int64
                                          185832 non-null int64
                                           185832 non-null float64
         7
           adp dist
         8 post skeletal distance to soma 185832 non-null float64
         9 pre_skeletal_distance_to_soma 185832 non-null float64
         10 pre oracle
                                          185832 non-null float64
                                           185832 non-null float64
         11 pre test score
         12 pre rf x
                                           185832 non-null float64
         13 pre rf y
                                           185832 non-null float64
                                           185832 non-null float64
         14 post_oracle
                                          185832 non-null float64
         15 post test score
         16 post rf x
                                           185832 non-null float64
```

```
185832 non-null float64
17 post rf y
18 compartment
                                  185832 non-null object
19 pre brain area
                                 185832 non-null object
                                 185832 non-null object
20 post brain area
21 pre nucleus x
                                  185832 non-null int64
22 pre nucleus y
                                 185832 non-null int64
23 pre nucleus z
                                 185832 non-null int64
                                  185832 non-null int64
24 post nucleus x
                                 185832 non-null int64
25 post nucleus y
26 post nucleus z
                                 185832 non-null int64
27 pre nucleus_id
                                 185832 non-null int64
                                  185832 non-null int64
28 post nucleus id
29 connected
                                 185832 non-null bool
30 pre feature weights
                                 185832 non-null object
31 post feature weights
                                 185832 non-null object
32 pre morph embeddings
                                 138123 non-null object
33 post_morph_embeddings
                               185832 non-null object
dtypes: bool(1), float64(11), int64(15), object(7)
memory usage: 48.4+ MB
```

As we see, not all features have the same amount of values.

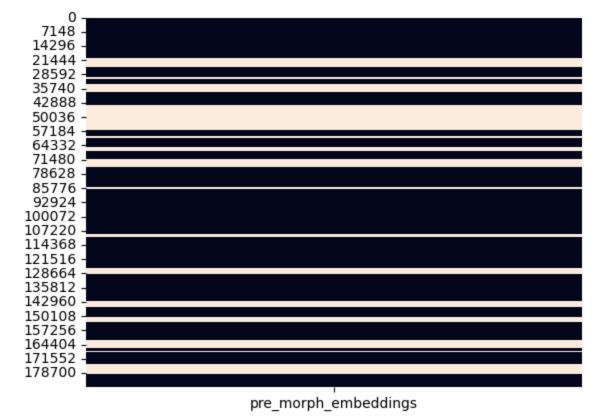
For instace, pre\_morph\_embeddings has a significant amount of missing values:

```
In [117... # Check for missing values in pre_morph_embeddings
missing_pre_morph = data['pre_morph_embeddings'].isnull()

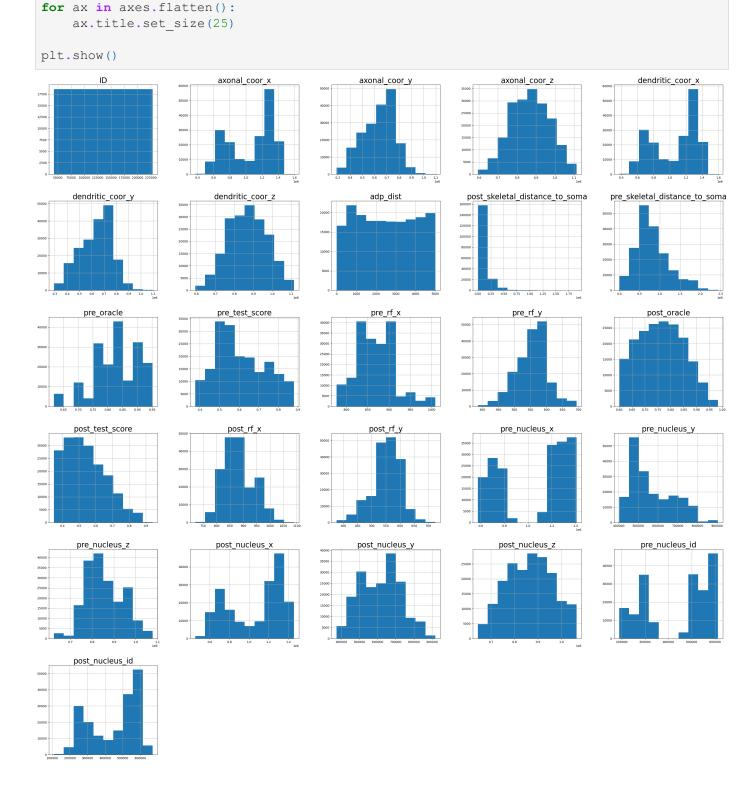
# Print the number of missing values
print(f"Number of missing values in pre_morph_embeddings: {missing_pre_morph.sum()}")

sns.heatmap(data[['pre_morph_embeddings']].isnull(), cbar=False)
plt.show()
```

Number of missing values in pre morph embeddings: 47709



In [16]: # Plot histograms of each column in the DataFrame
axes = data.hist(figsize=(40, 40))
# Loop through each plot and set the title size



# **FEATURE ENGINEERING**

</br>
In order to capture important aspects of the data that lead to improved model performance, we will have to perform feature engineering while also being aware of model complexity.

# **One-Hot Encoding**

Given our dataset, we can perform one-hot encoding to convert categorical data values into numeric values. We can use these conversions for later when we are engineering new features.

So, we will one-hot encode the following features:

- brain\_area: It refers to the brain area of which the neuron belongs to
  - We will get:
    - The V1 (Primary Visual Area)
    - The AL (Anterolateral Visual Area)
    - The RL (Rostrolateral Visual Area)
- compartment: It refers to the label of neuronal compartment where ADP resides on the post-synaptic neuron
  - We will get:
    - Axon
    - Oblique
    - Apical
    - Basal
    - o Soma
    - Apical tuft
    - Apical\_shaft

```
In [4]: from sklearn.model_selection import train_test_split

# Create a copy of the 'compartment' column
data['compartment_copy'] = data['compartment']

# Perform one-hot encoding on the 'compartment' column
data = pd.get_dummies(data, columns=['compartment'])

# One-hot encoding for 'pre_brain_area' and 'post_brain_area' features
data = pd.get_dummies(data, columns=['pre_brain_area', 'post_brain_area'])

# Split your data into training and test sets
train_data, test_data = train_test_split(data, test_size=0.2, random_state=1)
```

#### **Distance Features**

We already have some distance metrics:

- adp\_dist: Distance between axonal ADP coord. and dendritic ADP coord.
- **pre\_skeletal\_distance\_to\_soma**: Distance (on the pre-synaptic neuron) from soma to the specific point on its axon that is close to the dendrite of another neuron.
- **post\_skeletal\_distance\_to\_soma**: Distance (on the post-synaptic neuron) from soma to the specific point on its dendrite that is close to the dendrite of another neuron.

Note: These features are important since the time it takes for a signal to travel can depend on the distance. If that distance is too large, the signal may degrade, and a synapse could fail to occur.

#### We will engineer the following distance features:

- fw\_distance: Euclidean distance between feature weight vectors of pre- and post-synaptic neurons
  - How similar/dissimilar the tuning functions are of the two neurons; smaller dist. -> more similar tuning functions, bigger dist. -> less similar tuning functions

- **axon\_dendrite\_dist**: Euclidean distance of the physical gap between the axon of the pre-synaptic neuron and the dendrite of the post-synaptic neuron.
  - Note: ADP distances may not precisely match the direct Euclidean distances between axonal and dendritic coordinates, but they exhibit a high degree of correlation.
- **nucleus\_dist**: Euclidean distance between nucleus of the pre-synaptic neuron and nucleus of the post-synaptic neuron.
  - Represents how far apart the two neurons are (with respect to their nuclei)
- axon\_dendrite\_dist\_manhattan: Manhattan distance of the physical gap between the axon of the presynaptic neuron and the dendrite of the post-synaptic neuron.
  - Manhattan distance is better when dealing with sparse data and is more robust to outliers.
- nucleus\_dist\_manhattan: Manhattan distance between nucleus of the pre-synaptic neuron and nucleus
  of the post-synaptic neuron.
  - Same reason here
- rf\_dist: Euclidean distance between receptive field locations of the pre- and post-synaptic neurons.
  - The receptive field deals with the specific region in the visual field where a stimulus causes a response
  - Neurons with closer receptive fields may lead to synapses
  - Distance between these fields impacts time it takes a signal to travel between two neurons, which could also impact synaptic activity.
- pre/post\_nucleus\_rf\_dist: Euclidean distance between the nucleus and the receptive field locations of the pre- and post-synaptic neurons respectively
  - This distance deals with the pre- and post-synaptic neurons, looking at the distance between the nucleus and the specific area that causes neuronal activation
  - If the receptive field is far, the neuron may be less likely to form a synapse than for closer receptive fields

#### Did not end up using:

- morph\_distance: Euclidean distance between the morphological embeddings of the pre- and postsynaptic neurons
  - Morphological embeddings deal with the shape/structure of the neurons
  - Two neurons with similar structure or "morphology" may be more likely to form a synapse than less similar structures.

```
In [5]: def row_feature_distance(row):
    pre = row["pre_feature_weights"]
    post = row["post_feature_weights"]
    return np.sqrt(((pre - post)**2).sum())

data["fw_distance"] = data.apply(row_feature_distance, axis=1)

# Compute Euclidean distance
data['axon_dendrite_dist'] = np.sqrt(
    (data['axonal_coor_x'] - data['dendritic_coor_x'])**2 +
    (data['axonal_coor_y'] - data['dendritic_coor_y'])**2 +
    (data['axonal_coor_z'] - data['dendritic_coor_z'])**2
)
data['nucleus_dist'] = np.sqrt(
```

```
(data['pre nucleus x'] - data['post nucleus x'])**2 +
    (data['pre nucleus y'] - data['post nucleus y'])**2 +
    (data['pre nucleus z'] - data['post nucleus z'])**2
# Compute Manhattan distance
data['axon dendrite dist manhattan'] = (
    abs(data['axonal_coor_x'] - data['dendritic_coor_x']) +
    abs(data['axonal coor y'] - data['dendritic coor y']) +
    abs(data['axonal coor z'] - data['dendritic coor z'])
data['nucleus dist manhattan'] = (
    abs(data['pre nucleus x'] - data['post nucleus x']) +
    abs(data['pre nucleus y'] - data['post nucleus y']) +
    abs(data['pre nucleus z'] - data['post nucleus z'])
# Compute Euclidean distance between receptive field locations
data['rf dist'] = np.sqrt(
    (data['pre rf x'] - data['post rf x'])**2 +
    (data['pre rf y'] - data['post rf y'])**2
# Compute Euclidean distance between nucleus and receptive field locations
data['pre nucleus rf dist'] = np.sqrt(
    (data['pre_nucleus_x'] - data['pre_rf_x'])**2 +
    (data['pre nucleus y'] - data['pre rf y'])**2
data['post nucleus rf dist'] = np.sqrt(
    (data['post_nucleus_x'] - data['post_rf_x'])**2 +
    (data['post nucleus y'] - data['post rf y'])**2
```

## **Similarity Features**

We will engineer the following similarity features:

- **fw\_similarity**: Cosine similarity between the feature weight vectors of the pre- and post-synaptic neurons
  - Cosine similarity quantifies the similarity between two vectors.
  - Recall, the feature weight vectors deal with each neuron's tuning function. If two neurons have similar response patterns (similar feature weights), they might be more likely to form a synapse
- rf\_similarity: Cosine similarity between receptive field locations of the pre- and post-synaptic neurons
  - This would provide a measure of how similar the receptive field locations are for the pre- and post-synaptic neurons.

```
In [6]: #cosine similarity function
def row_feature_similarity(row):
    pre = row["pre_feature_weights"]
    post = row["post_feature_weights"]
    return (pre * post).sum() / (np.linalg.norm(pre) * np.linalg.norm(post))

# compute the cosine similarity between the pre- and post- feature weights
data["fw_similarity"] = data.apply(row_feature_similarity, axis=1)

def row_rf_similarity(row):
    pre_rf = np.array([row['pre_rf_x'], row['pre_rf_y']])
    post_rf = np.array([row['post_rf_x'], row['post_rf_y']])
    return (pre_rf * post_rf).sum() / (np.linalg.norm(pre_rf) * np.linalg.norm(post_rf))
```

```
# Compute the cosine similarity between the pre- and post- receptive fields
data['rf_similarity'] = data.apply(row_rf_similarity, axis=1)
```

## **Interaction Features**

We will engineer the following interaction features:

- pre/post\_oracle\_adp\_dist: Captures how the reliability of neural activation (of the pre- and postsynaptic neuron) changes with the distance between the axon and dendrite
  - Recall, the oracle feature deals with neuronal response reliability to repeated visual stimulus
  - So, neuronal response reliability could vary for different distances for both pre- and post-synaptic neurons
- **pre/post\_oracle\_pre\_rf\_x/y\_interaction**: Captures how the reliability of neural activation (of the post-synaptic neuron) changes with the receptive field location, measured by rf\_x & rf\_y
  - So, neuronal response reliability could vary for different receptive fields for both pre- and postsynaptic neurons
- **pre/post\_oracle\_axon\_dendrite\_dist\_interaction**: Captures how the reliability of neural activation (of the post-synaptic neuron) changes with the Euclidean distance between the pre-synaptic neuron's axon and the post-synaptic neuron's dendrite.
  - So, neuronal response reliability could vary for different Eucliean distances for both pre- and postsynaptic neurons
  - Remember, ADP distances may not precisely match the direct Euclidean distances between axonal and dendritic coordinates, but they exhibit a high degree of correlation
- **pre/post\_oracle\_test\_score\_interaction**: Represents the combined effect of a neuron's response reliability and the predictive performance of the deep learning model on withheld test trials
  - Recall, the test\_score is a measure of how well the deep learning model could predict the outcomes
  - So, this feature is important since a neuron's ability to form a synapse could be influenced both by response reliability and the performance of the model
  - Neurons with both high activation reliability and high predictive performance could form a synapse when compared to lower values
- pre\_AL/RL/V1\_post\_RL/V1/AL: Represents interactions between the specific brain areas (AL, RL, and V1) of the pre-synaptic and post-synaptic neurons
  - Interactions between different brain areas may influence whether we will get a synapse from an ADP

```
In [7]: # Compute interaction between oracle and distance features
    data['pre_oracle_adp_dist'] = data['pre_oracle'] * data['adp_dist']
    data['post_oracle_adp_dist'] = data['post_oracle'] * data['adp_dist']

# Compute interaction between oracle and receptive field information
    data['pre_oracle_pre_rf_x_interaction'] = data['pre_oracle'] * data['pre_rf_x']
    data['pre_oracle_pre_rf_y_interaction'] = data['pre_oracle'] * data['pre_rf_y']
    data['post_oracle_post_rf_x_interaction'] = data['post_oracle'] * data['post_rf_x']
    data['post_oracle_post_rf_y_interaction'] = data['post_oracle'] * data['post_rf_y']

# Compute interaction between pre_oracle and axon_dendrite_dist
    data['pre_oracle_axon_dendrite_dist_interaction'] = data['pre_oracle'] * data['axon_dendrite_dist_oracle_axon_dendrite_dist_interaction'] = data['post_oracle'] * data['axon_dendrite_dist_oracle_axon_dendrite_dist_interaction'] = data['post_oracle'] * data['axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_oracle_axon_dendrite_dist_ora
```

```
data['pre_oracle_test_score_interaction'] = data['pre_oracle'] * data['pre_test_score']
data['post_oracle_test_score_interaction'] = data['post_oracle'] * data['post_test_score

# Create interaction features
data['pre_AL_post_RL'] = data['pre_brain_area_AL'] * data['post_brain_area_RL']
data['pre_AL_post_V1'] = data['pre_brain_area_AL'] * data['post_brain_area_V1']
data['pre_RL_post_AL'] = data['pre_brain_area_RL'] * data['post_brain_area_AL']
data['pre_RL_post_V1'] = data['pre_brain_area_RL'] * data['post_brain_area_V1']
data['pre_V1_post_AL'] = data['pre_brain_area_V1'] * data['post_brain_area_AL']
data['pre_V1_post_RL'] = data['pre_brain_area_V1'] * data['post_brain_area_RL']
```

## **Difference Features**

We will engineer only one difference feature:

- **test\_score\_diff**: Represents the difference in predictive performance of the deep learning predictive model between the pre- and post-synaptic neuron
  - If the difference in test scores is significant, it might influence whether or not a synapse will form
  - A higher score may lead to a higher likelihood of a synapse

```
In [8]: data['test_score_diff'] = data['post_test_score'] - data['pre_test_score']
```

#### **Number Count Features**

We will engineer the following count features:

- **total\_synapses**: Represents the total number of synapses formed (from each pre-synaptic neuron)
  - The number of synapses a pre-synaptic neuron (the one that sends the signal) forms could potentially influnce its overall connectivity.
  - In other words, a neuron with a larger number of synapses could have a higher possibility in general of forming a synapse than those with lower number of synapses.
- **pre/post\_(compartment)\_count**: The count of each type of compartment (axon, oblique, apical, etc.) for the pre- and post-synaptic neurons
  - Recall, compartment refers to the label of neuronal compartment where ADP resides on the postsynaptic neuron.
  - Taking into account the specific number of compartments per neuron could determine its likelihood in being part of a synapse
- **compartment\_synapse\_prop**: The proportion of synapses in each neuronal compartment; quantifies the likelihood of synapse formation in each compartment.
  - Certain types of compartments may be likely to lead to a synapse than others. The propensity for synapse formation could vary between the compartments

```
In [9]: # Calculate the total number of synapses for each neuron
    total_synapses = data.groupby('pre_nucleus_id').size()

# Add the Total Synapses to your DataFrame
    data['total_synapses'] = data['pre_nucleus_id'].map(total_synapses)

# List of all compartment types
    compartments = ['apical', 'apical_shaft', 'apical_tuft', 'axon', 'basal', 'oblique', 'so

# Create new aggregated features for pre_nucleus_id
    for compartment in compartments:
```

```
# Create new feature name
    new_feature = f'pre_{compartment}_count'

# Create new feature
    data[new_feature] = data.groupby('pre_nucleus_id')[f'compartment_{compartment}'].tra

# Create new aggregated features for post_nucleus_id

for compartment in compartments:
    # Create new feature name
    new_feature = f'post_{compartment}_count'

# Create new feature
    data[new_feature] = data.groupby('post_nucleus_id')[f'compartment_{compartment}'].tr

# Calculate the proportion of synapses in each compartment
synapse_proportions = data.groupby('compartment_copy')['connected'].mean()

# Map the proportions to the original DataFrame
data['compartment_synapse_prop'] = data['compartment_copy'].map(synapse_proportions)
```

## Engineered Features That Were Not Used

There are a number of engineered features that ended up either not making a difference in model performance, made model performance worse, or were misleading in the sense that they scored incredibly well in Jupyter Notebook but performed horribly in Kaggle.

#### **Notes:**

- One group of features that were incredibly difficult to work with was the <u>morphological embeddings</u>.
   Using techniques like imputation to deal with the missing values of pre\_morph\_embeddings was harder than expected, and even when new features were engineered from the morphological embeddings (morph\_distance and morph\_similarity), they did not perform well on the model.
- The interaction features below failed to make a meaningful difference in the model, and any change in performance was usually negative. They also made the model more complex for no specific benefit.

Each commented out block of code in triple quotes represents a failed feature

```
In [28]:
        ### DISTANCE FEATURES ###
        # DID NOT USE
        # Euclidean distance function for morphological embeddings
        #def row morph distance(row):
        # pre = row["pre morph embeddings"]
            post = row["post morph embeddings"]
        # return np.linalg.norm(pre - post)
        # Compute the Euclidean distance between the pre- and post- morphological embeddings
         #data["morph distance"] = data.apply(row morph distance, axis=1)
        ### SIMILARITY FEATURES ###
        # DID NOT USE
        # Cosine similarity function for morphological embeddings
        #def row morph similarity(row):
         # pre = row["pre morph embeddings"]
           post = row["post_morph_embeddings"]
```

```
return (pre * post).sum() / (np.linalg.norm(pre) * np.linalg.norm(post))
# Compute the cosine similarity between the pre- and post- morphological embeddings
#data["morph similarity"] = data.apply(row morph similarity, axis=1)
### INTERACTION FEATURES ###
# Compute interaction between post oracle and nucleus dist
# data['pre oracle nucleus dist interaction'] = data['pre oracle'] * data['nucleus dist'
# data['post oracle nucleus dist interaction'] = data['post oracle'] * data['nucleus dis
0.00
# Compute interaction between adp dist and other features
# data['adp dist axon dendrite dist interaction'] = data['adp dist'] * data['axon dendri
# data['adp dist nucleus dist interaction'] = data['adp dist'] * data['nucleus dist'] ->
# data['adp dist rf dist interaction'] = data['adp dist'] * data['rf dist'] -> DID NOT U
0.00\,0
# Create a new feature that is True if both pre oracle and post oracle are at least 0.8
# data['high reliability'] = (data['pre oracle'] >= 0.5) & (data['post oracle'] >= 0.5)
### DIFFERENCE FEATURES ###
#data['test score ratio'] = data['post test score'] / data['pre test score'] -> DID NOT
#data['test score avg'] = (data['post test score'] + data['pre test score']) / 2 -> DID
# Compute difference between pre and post features
# data['oracle diff'] = data['post oracle'] - data['pre oracle']
# data['rf x diff'] = data['post_rf_x'] - data['pre_rf_x']
# data['rf y diff'] = data['post rf y'] - data['pre rf y']
# data['oracle int'] = data['post oracle'] * data['pre oracle']
### RATIO FEATURES ###
# List of all compartment types
compartments = ['compartment apical', 'compartment apical shaft', 'compartment apical tu
# Create a new feature that represents the total count of compartments for each neuron
data['total compartments'] = data[compartments].sum(axis=1)
\# Create new features that represent the ratio of each compartment type to the total num
for compartment in compartments:
    data[f'{compartment} ratio'] = data[compartment] / data['total compartments']
.....
11 11 11
# Synapse ratio feature
# Calculate the total number of synapses for each neuron
total synapses = data.groupby('pre nucleus id').size()
# Calculate the number of connected synapses for each neuron
```

```
connected synapses = data[data['connected'] == True].groupby('pre nucleus id').size()
# Calculate the Synapse Ratio
synapse ratio = connected synapses / total synapses
# Add the Synapse Ratio to your DataFrame
data['synapse ratio'] = data['pre nucleus id'].map(synapse ratio)
### COUNT FEATURES ###
# Compute the global mean
mean = data['connected'].mean()
# Compute the number of values and the mean of each group
agg = data.groupby('compartment copy')['connected'].agg(['count', 'mean'])
counts = agg['count']
means = agg['mean']
# Compute the smoothed means
smooth = (counts * means + 100 * mean) / (counts + 100)
# Replace each compartment with the smoothed mean
# data['compartment target enc'] = data['compartment copy'].map(smooth) -> DID NOT USE
# data['compartment count'] = data['compartment copy'].map(data['compartment copy'].valu
### IMPUTATION ###
# from sklearn.impute import SimpleImputer
# Mean imputation
# mean imputer = SimpleImputer(strategy='mean')
# data['morph distance'] = mean imputer.fit transform(data['morph distance'].values.resh
# data['morph similarity'] = mean imputer.fit transform(data['morph similarity'].values.
# Median imputation
# median imputer = SimpleImputer(strategy='median')
# data['morph distance'] = median imputer.fit transform(data['morph distance'].values.re
# data['morph similarity'] = median imputer.fit transform(data['morph similarity'].value
# Most frequent imputation
# most frequent imputer = SimpleImputer(strategy='most frequent')
# data['morph distance'] = most frequent imputer.fit transform(data['morph distance'].va
# data['morph similarity'] = most frequent imputer.fit transform(data['morph similarity'
# data['synapse ratio'] = mean imputer.fit transform(data['synapse ratio'].values.reshap
### OTHER WAYS OF DEALING WITH EMPTY VALUES ###
# Fill NaNs with the mean of the column
# mean morph similarity = train data['morph distance'].mean()
# train data['morph distance'].fillna(mean morph similarity, inplace=True)
# Fill NaNs with the mean of the column in the training data
# mean morph similarity = test data['morph distance'].mean()
# test data['morph distance'].fillna(mean morph similarity, inplace=True)
# Drop rows with NaNs in the 'morph similarity' column
# train data.dropna(subset=['morph similarity'], inplace=True)
# test data.dropna(subset=['morph similarity'], inplace=True)
# Drop rows with NaNs in the 'morph distance' column
# train data.dropna(subset=['morph distance'], inplace=True)
```

```
# test_data.dropna(subset=['morph_distance'], inplace=True)

# print(X_resampled.isnull().sum()) # Check for NaNs

# print(np.isinf(X_resampled).sum()) # Check for infinity values

# print((X_resampled > np.finfo(np.float64).max).sum()) # Check for values too large fo
```

# Check the Dataframe

We now have 88 features in total.

Plotting a histogram for each column in the Dataframe, we can get a better look at each feature's distribution of values

```
In [13]: print("Main Dataframe Size:", data.shape)
        print("\nMain Dataframe:")
         data.info()
         column names = data.columns.tolist()
        Main Dataframe Size: (185832, 88)
        Main Dataframe:
        <class 'pandas.core.frame.DataFrame'>
        Int64Index: 185832 entries, 0 to 185831
        Data columns (total 88 columns):
                                                          Non-Null Count Dtype
         # Column
        --- -----
                                                          _____
         \cap
            ID
                                                          185832 non-null int64
                                                          185832 non-null int64
         1 axonal coor x
         2 axonal coor y
                                                          185832 non-null int64
         3 axonal_coor_z
                                                          185832 non-null int64
         4 dendritic coor x
                                                          185832 non-null int64
         5 dendritic coor y
                                                         185832 non-null int64
         6 dendritic coor z
                                                         185832 non-null int64
                                                     185832 non-null float64
         7
            adp dist
         8 post skeletal distance to soma
         9 pre skeletal distance to soma
         10 pre oracle
                                                          185832 non-null float64
         11 pre test score
                                                          185832 non-null float64
         12 pre rf x
                                                         185832 non-null float64
         13 pre rf y
                                                          185832 non-null float64
         14 post oracle
                                                          185832 non-null float64
         15 post test score
         16 post rf x
                                                         185832 non-null float64
         17 post rf y
                                                          185832 non-null float64
                                                          185832 non-null int64
         18 pre_nucleus_x
                                                          185832 non-null int64
         19 pre nucleus y
         20 pre nucleus z
                                                         185832 non-null int64
         21 post nucleus_x
                                                          185832 non-null int64
                                                          185832 non-null int64
         22 post nucleus y
                                                          185832 non-null int64
         23 post nucleus z
         24 pre nucleus id
                                                          185832 non-null int64
                                                          185832 non-null int64
         25 post nucleus id
         26 connected
                                                          185832 non-null bool
                                                        185832 non-null object
138123 non-null
         27 pre feature weights
         28 post feature weights
         29 pre morph embeddings
         30 post morph embeddings
                                                         185832 non-null object
                                                         185832 non-null object
         31 compartment copy
                                                          185832 non-null uint8
         32 compartment apical
         33 compartment apical shaft
                                                          185832 non-null uint8
         34 compartment apical tuft
                                                         185832 non-null uint8
         35 compartment axon
                                                          185832 non-null uint8
```

```
36 compartment_basal
                                                                                                                                                                                                                                               185832 non-null uint8
     37 compartment oblique
                                                                                                                                                                                                                                             185832 non-null uint8

      37
      compartment_oblique
      185832 non-null uint8

      38
      compartment_soma
      185832 non-null uint8

      39
      pre_brain_area_AL
      185832 non-null uint8

      40
      pre_brain_area_RL
      185832 non-null uint8

      41
      pre_brain_area_V1
      185832 non-null uint8

      42
      post_brain_area_AL
      185832 non-null uint8

      43
      post_brain_area_RL
      185832 non-null uint8

      44
      post_brain_area_V1
      185832 non-null float64

      45
      fw_distance
      185832 non-null float64

      46
      axon_dendrite_dist
      185832 non-null float64

      47
      nucleus_dist
      185832 non-null int64

      48
      axon_dendrite_dist_manhattan
      185832 non-null int64

      49
      nucleus_dist_manhattan
      185832 non-null float64

      50
      rf_dist
      185832 non-null float64

      51
      pre_nucleus_rf_dist
      185832 non-null float64

   rr_aist

pre_nucleus_rf_dist

post_nucleus_rf_dist

float64

post_nucleus_rf_dist

fw_similarity

float64

fw_similarity

float64

fre_similarity

float64

pre_oracle_adp_dist

pre_oracle_adp_dist

float64

pre_oracle_pre_rf_x_interaction

pre_oracle_pre_rf_y_interaction

post_oracle_post_rf_x_interaction

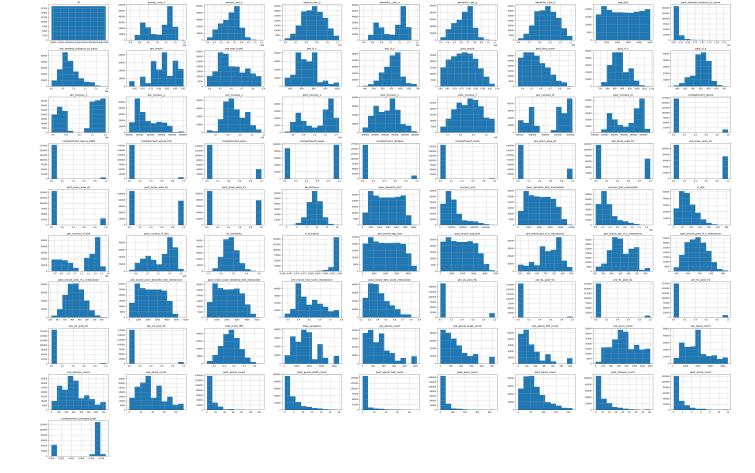
post_oracle_post_rf_x_interaction

post_oracle_post_rf_x_interaction

float64

float64
     61 pre oracle axon dendrite dist interaction 185832 non-null float64
     62 post_oracle_axon_dendrite_dist_interaction 185832 non-null float64
     63 pre_oracle_test_score_interaction 185832 non-null float64
64 post_oracle_test_score_interaction 185832 non-null float64
                                                                                                                                                                                                                                           185832 non-null uint8
     65 pre AL post RL
     66 pre AL post V1
                                                                                                                                                                                                                                             185832 non-null uint8
 67 pre_RL_post_AL 185832 non-null uint8
68 pre_RL_post_V1 185832 non-null uint8
69 pre_V1_post_AL 185832 non-null uint8
70 pre_V1_post_RL 185832 non-null uint8
71 test_score_diff 185832 non-null int64
72 total_synapses 185832 non-null int64
73 pre_apical_count 185832 non-null float64
74 pre_apical_shaft_count 185832 non-null float64
75 pre_apical_tuft_count 185832 non-null float64
76 pre_axon_count 185832 non-null float64
77 pre_basal_count 185832 non-null float64
78 pre_oblique_count 185832 non-null float64
79 pre_soma_count 185832 non-null uint8
80 post_apical_shaft_count 185832 non-null uint8
81 post_apical_shaft_count 185832 non-null uint8
82 post_apical_tuft_count 185832 non-null uint8
83 post_axon_count 185832 non-null uint8
84 post_apical_tuft_count 185832 non-null uint8
85 post_oblique_count 185832 non-null uint8
86 post_soma_count 185832 non-null uint8
87 compartment_synapse_prop 185832 non-null uint8
88 compartment_synapse_prop 185832 non-null uint8
89 compartment_synapse_prop 185832 non-null uint8
80 compartment_synapse_prop 185832 non-null uint8
81 compartment_synapse_prop 185832 non-null uint8
82 compartment_synapse_prop 185832 non-null uint8
83 compartment_synapse_prop 185832 non-null uint8
84 compartment_synapse_prop 185832 non-null uint8
     67 pre RL post AL
                                                                                                                                                                                                                                        185832 non-null uint8
dtypes: bool(1), float64(38), int64(18), object(5), uint8(26)
memory usage: 92.7+ MB
```

In [120... # Plot histograms of each column in the DataFrame data.hist(figsize=(60, 40)) plt.show()



# **MODELING**

</br>
 Now, we will work on building a model that optimizes performance but does not lead to overfitting. There are a number of models that will be tested, as well as hyperparameters that will require a lot of tuning.

To start, we will import the simple Logistic Regression classification model from the example notebook to use as reference:

## **Simple Logistic Regression**

Accuracy of approximately 99.27%, but the confusion matrix reveals that it's not performing well on predicting when a synapse will form (True Positives).

It's predicting all ADPs as not forming a synapse. This is likely due to the imbalance in your data, where the majority of ADPs do not form a synapse.

```
In [10]: # logistic regression model (connected ~ fw_similarity + adp)
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import balanced_accuracy_score, accuracy_score, confusion_matrix
    from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import StandardScaler
    from sklearn.pipeline import Pipeline

# split into a train and test set
    #(Even though we're working with the competition training set, you may want to have your
    train_data, test_data = train_test_split(data, test_size=0.2, random_state=1)

# create pipeline
```

```
pipe = Pipeline(
        [("scaler", StandardScaler()), ("model", LogisticRegression(random_state=2))]
)

# fit model
pipe.fit(train_data[["fw_similarity", "adp_dist"]], train_data["connected"])

# predict on test data
test_data["pred"] = pipe.predict_proba(test_data[["fw_similarity", "adp_dist"]])[:, 1]

# compute accuracy
print(f"accuracy: {accuracy_score(test_data['connected'], test_data['pred'] > .5)}")
```

accuracy: 0.9927085855732236

- **True Negatives (Top-Left, 36896)**: These are cases in which model correctly predicted that a synapse would not form.
- **False Positives (Top-Right, 0)**: These are cases in which model incorrectly predicted that a synapse would form.
- **False Negatives (Bottom-Left, 271)**: These are cases in which model incorrectly predicted that a synapse would not form.
- **True Positives (Bottom-Right, 0)**: These are cases in which model correctly predicted that a synapse would form.

The Balanced Accuracy is 0.5. This means that the model is performing no better than random guessing.

#### Addressing the unbalanced nature of our dataset

- Oversampling to address the imbalance in the data. It involves randomly duplicating examples in the minority class (in this case, ADPs where a synapse forms) to increase their prevalence in the training data.
- Accuracy of model is approximately 64.78%. This is lower but remember that accuracy can be misleading when dealing with imbalanced datasets.
- Balanced Accuracy of your model is approximately 72.19%. The model is doing a better job at classifying both majority and minority classes

```
# predict on test data
test_data["pred"] = pipe.predict_proba(test_data[["fw_similarity", "adp_dist"]])[:, 1]
# compute accuracy
print(f"accuracy: {accuracy_score(test_data['connected'], test_data['pred'] > .5)}")
# confusion matrix
print(confusion_matrix(test_data['connected'], test_data['pred'] > .5))
# compute balanced accuracy
print(
    f"balanced accuracy: {balanced_accuracy_score(test_data['connected'], test_data['pre
)
accuracy: 0.647752038098313
[[23859 13037]
```

```
[[23859 13037]
[ 55 216]]
balanced accuracy: 0.7218517172433216
```

# **Modeling Notes**

We will be using:

- **Resampling**: We will be using undersampling as opposed to oversampling. This involves randomly removing examples from the majoriy class to decrease instances of over-representation. Additionally, it will be incredibly beneficial in terms of computational efficiency/cost.
  - Note, oversampling (SMOTE) was used at certain points, but it failed to perform better than undersampling, and took way too long to run.
- **Standardization**: We transform the features of the dataset to have a mean of zero and a standard deviation of one. To do this, we will use StandardScaler()
  - Alternatives like MinMaxScaler and MaxAbsScaler did not yield to better results.
  - StandardScaler is also less sensitive to extreme outliers
- **Pipeline**: Allows a linear series of data transforms to be linked together. Streamlines the machine learning workflow
  - They simplify the process of fitting and transforming data, and can help prevent data leakage by ensuring preprocessing steps happen only on the training data
  - The workflow also becomes more easily reproducible
- **Confusion Matrix**: Represents the accuracy of a classification model. Displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN)
  - We can use a confusion matrix in every output to understand how many predictions were correct and the kind of errors our model made
  - Can better understand where the model became 'confused'

# **Support Vector Machines (SVMs)**

We will start with a simple SVM with default parameters (meaning our kernel will be RBF), and train this SVM only on two simple features: fw\_similarity and adp\_dist. Doing so, we will see worse results than the logistic regression from earlier on model accuracy, however a slightly better result in terms of balanced accuracy.

```
In [16]: from sklearn.svm import SVC
from sklearn.metrics import balanced_accuracy_score, accuracy_score, confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from imblearn.under_sampling import RandomUnderSampler
```

```
In [17]: # split into a train and test set
         train data, test data = train test split(data, test size=0.2, random state=1)
         # create pipeline
         pipe = Pipeline(
             [("scaler", StandardScaler()), ("model", SVC(probability=True, random state=2))]
         # undersample majority class
         rus = RandomUnderSampler(random state=0)
         X resampled, y resampled = rus.fit resample(
             train data[["fw similarity", "adp dist"]], train data["connected"]
         # fit model
         pipe.fit(X resampled, y resampled)
         # predict on test data
         test data["pred"] = pipe.predict proba(test data[["fw similarity", "adp dist"]])[:, 1]
         # compute accuracy
         print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
         # confusion matrix
         print(confusion matrix(test data['connected'], test data['pred'] > .5))
         # compute balanced accuracy
         print(
             f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
```

[[22555 14341] [ 43 228]] balanced accuracy: 0.7263206463645295

accuracy: 0.6129900180267441

#### However,

Simply going from rbf to a linear kernel, we can see a massive improvement in model accuracy, but a worse balanced accuracy.

Note: After manually trying all possible kernels (linear, poly, rbf, sigmoid), linear massively outperformed all other kernels

```
In [13]: # split into a train and test set
    train_data, test_data = train_test_split(data, test_size=0.2, random_state=1)

# create pipeline
pipe = Pipeline(
    [("scaler", StandardScaler()), ("model", SVC(kernel='linear', probability=True, rand)
```

```
# undersample majority class
rus = RandomUnderSampler(random state=0)
X resampled, y resampled = rus.fit resample(
    train data[["fw similarity", "adp dist"]], train data["connected"]
# fit model
pipe.fit(X resampled, y resampled)
# predict on test data
test data["pred"] = pipe.predict proba(test data[["fw similarity", "adp dist"]])[:, 1]
# compute accuracy
print(f"accuracy_score(test_data['connected'], test data['pred'] > .5)}")
# confusion matrix
print(confusion matrix(test data['connected'], test data['pred'] > .5))
# compute balanced accuracy
print(
    f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
accuracy: 0.6563887319396239
[[24184 12712]
```

At this point of the competition, there were two new features that were engineered: axon\_dendrite\_dist and nucleus\_dist

Adding these two to our pipeline and training them on our linear SVM, we see performance on both metrics improve drastically.

Our Kaggle public leaderboard score was around ~70.12%

balanced accuracy: 0.7188759149083251

[ 59 212]]

```
In [14]: # create pipeline
        pipe = Pipeline(
            [("scaler", StandardScaler()), ("model", SVC(kernel='linear', probability=True, rand
         # undersample majority class
         rus = RandomUnderSampler(random state=0)
        X resampled, y resampled = rus.fit resample(
             train_data[["fw_similarity", "adp dist", "axon dendrite dist", "nucleus dist"]],
             train data["connected"]
         # fit model
        pipe.fit(X resampled, y resampled)
         # predict on test data
         test data["pred"] = pipe.predict proba(test data[["fw similarity", "adp dist", "axon den
         # compute accuracy
        print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
         # confusion matrix
         print(confusion matrix(test data['connected'], test data['pred'] > .5))
```

```
# compute balanced accuracy
print(
    f"balanced accuracy: {balanced_accuracy_score(test_data['connected'], test_data['pre'])

accuracy: 0.6817068905211613
[[25121 11775]
    [ 55 216]]
balanced accuracy: 0.7389538421349088
```

Introducing three new engineered features led to increased performance on both metrics. Those three being rf\_dist, pre\_oracle\_adp\_dist, and post\_oracle\_adp\_dist

```
In [17]: | # create pipeline
         pipe = Pipeline(
            [("scaler", StandardScaler()), ("model", SVC(kernel='linear', probability=True, rand
         # undersample majority class
         rus = RandomUnderSampler(random state=0)
         X resampled, y resampled = rus.fit resample(
             train data[["fw similarity", "adp dist", "axon dendrite dist", "nucleus dist", "rf d
             train data["connected"]
         # fit model
         pipe.fit(X resampled, y resampled)
         # predict on test data
         test data["pred"] = pipe.predict proba(test data[["fw similarity", "adp dist", "axon den
         # compute accuracy
         print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
         # confusion matrix
         print(confusion matrix(test data['connected'], test data['pred'] > .5))
         # compute balanced accuracy
         print(
             f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
        accuracy: 0.6831328867005677
```

#### [[25170 11726] [ 51 220]] balanced accuracy: 0.7469979445566355

## **Hyperparameter Tuning**

We will use both <u>GridSearchCV</u> and <u>RandomizedSearchCV</u> to find the optimal parameters

Note: With RandomizedSearchCV, there will be a tradeoff between runtime and best possible solution in the  $n_i$ ter parameter

```
In [45]: from sklearn.model_selection import GridSearchCV

# define the parameter grid
param_grid = {
    'model__C': [0.01, 0.1, 1, 10],
    'model__gamma': ['scale', 'auto', 0.001, 0.01, 0.1, 1, 10, 100],
}
```

```
# create a GridSearchCV object
         grid = GridSearchCV(pipe, param grid, cv=5, scoring='balanced accuracy')
         # fit the model and tune parameters
         grid.fit(X resampled, y resampled)
         print(f"Best parameters: {grid.best params }")
         print(f"Best balanced accuracy: {grid.best score }")
        Best parameters: {'model C': 10, 'model gamma': 'scale'}
        Best balanced accuracy: 0.7351598173515981
In [47]: from scipy.stats import uniform, randint
         # define the parameter distribution
         param dist = {
            'model C': uniform(0.001, 50),
             'model gamma': uniform(0.001, 10),
         # create a RandomizedSearchCV object
         random search = RandomizedSearchCV(pipe, param distributions=param dist,
                                           n iter=10, cv=5, scoring='balanced accuracy', random
         # fit the model and tune parameters
         random search.fit(X resampled, y resampled)
         print(f"Best parameters: {random search.best params }")
         print(f"Best balanced accuracy: {random search.best score }")
        Best parameters: {'model C': 27.484123893935458, 'model gamma': 4.354223926182769}
        Best balanced accuracy: 0.7356164383561644
```

#### Manipulating one hyperparameter, C, yielded a significant difference in performance.

Specifically, making C=10 increased both model accuracy and balanced accuracy by around ~0.002

#### Notes:

- Increasing C to 10 (from a default of 1.0) means we are increasing the penalty for misclassified points (making our margin narrower). Our SVM will try to classify **all** points correctly, including the outliers. We can see the number of true positives increase and false positives decrease in our confusion matrix.
- RandomizedSearch gave us C to be 27, but to avoid any potential overfitting, C was chosen to be 10. Even if RandomizedSearchCV had C to be bigger than 10, there were not really any major advances in accuracy
- Gamma was best at scale, which is the default parameter setting.

At this point, the highest Kaggle score was around \*\*~70.55%\*\*

```
# fit model
pipe.fit(X_resampled, y_resampled)
# predict on test data
test_data["pred"] = pipe.predict_proba(test_data[["fw_similarity", "adp_dist", "axon_den
# compute accuracy
print(f"accuracy: {accuracy_score(test_data['connected'], test_data['pred'] > .5)}")
# confusion matrix
print(confusion_matrix(test_data['connected'], test_data['pred'] > .5))
# compute balanced accuracy
print(
    f"balanced accuracy: {balanced_accuracy_score(test_data['connected'], test_data['pred'] > .5)}
accuracy: 0.6853929561169855
```

```
[[25254 11642]
[ 51 220]]
balanced accuracy: 0.7481362793354733
```

#### Adding one-hot encoded brain area features

After one-hot encoding the brain area categorical features, we engineered some features looking at their interactions. Those features were pre\_AL/RL/V1\_post\_RL/V1/AL

We can see a strong increase in model performance, but a dip in balanced accuracy.

Although there was a dip in the balanced accuracy, we still managed to end up with a Kaggle score of \*\*~70.55%\*\*

```
In [56]: | # create pipeline
        pipe = Pipeline(
            [("scaler", StandardScaler()), ("model", SVC(kernel='linear', C=10.0, probability=Tr
         # undersample majority class
         rus = RandomUnderSampler(random state=0)
        X resampled, y resampled = rus.fit resample(
             train data[["fw similarity", "adp dist", "axon dendrite dist", "nucleus dist", "rf d
             train data["connected"]
         # fit model
        pipe.fit(X resampled, y resampled)
         # predict on test data
         test data["pred"] = pipe.predict proba(test data[["fw similarity", "adp dist", "axon den
         # compute accuracy
        print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
         # confusion matrix
        print(confusion matrix(test data['connected'], test data['pred'] > .5))
         # compute balanced accuracy
            f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
        accuracy: 0.6879220814163102
```

accuracy: 0.6879220814163102 [[25354 11542]

```
[ 57 214]] balanced accuracy: 0.7384213290853636
```

#### **Our Best Linear SVM**

We get these results from the various new engineered features

#### Results:

We get a model accuracy of ~69.155% and a balanced accuracy of ~73.476%

Kaggle *Public* Score -> **70.877%** 

```
In [164... # Split your data into training and test sets
         train_data, test_data = train_test_split(data, test size=0.2, random state=1)
         # create pipeline
        pipe = Pipeline(
            [("scaler", StandardScaler()), ("model", SVC(kernel='linear', C=10.0, probability=Tr
         # undersample majority class
         rus = RandomUnderSampler(random state=0)
        X resampled, y resampled = rus.fit resample(
             train data[["fw similarity", "adp dist", "axon dendrite dist", "nucleus dist", "rf d
             train data["connected"]
         # fit model
        pipe.fit(X resampled, y resampled)
         # predict on test data
         test data["pred"] = pipe.predict proba(test data[["fw similarity", "adp dist", "axon den
         # compute accuracy
         print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
         # confusion matrix
        print(confusion matrix(test data['connected'], test data['pred'] > .5))
         # compute balanced accuracy
        print(
             f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
        accuracy: 0.691554335835553
        [[25492 11404]
         [ 60
                 211]]
        balanced accuracy: 0.7347563951571866
In [ ]: from imblearn.combine import SMOTEENN
         from sklearn.svm import SVC
         # create pipeline
        pipe = Pipeline(
            [("scaler", StandardScaler()), ("model", SVC(kernel='linear', C=10.0, probability=Tr
         # apply SMOTEENN
         smote enn = SMOTEENN(random state=0)
        X resampled, y resampled = smote enn.fit resample(
             train_data[["fw_similarity", "adp_dist", "axon_dendrite_dist", "nucleus_dist", "rf_d
```

```
train_data["connected"]
)

# fit model
pipe.fit(X_resampled, y_resampled)

# predict on test data
test_data["pred"] = pipe.predict_proba(test_data[["fw_similarity", "adp_dist", "axon_den

# compute accuracy
print(f"accuracy: {accuracy_score(test_data['connected'], test_data['pred'] > .5)}")

# confusion matrix
print(confusion_matrix(test_data['connected'], test_data['pred'] > .5))

# compute balanced accuracy
print(f"balanced accuracy: {balanced_accuracy_score(test_data['connected'], test_data['pred'] > .5))
```

## **Random Forests**

As the number of engineered features would continue to increase, there slowly became a worry of overfitting due to increased model complexity. Thus, there was a transition from SVMs to ensemble methods, as they tend to be robust to overfitting.

#### So, we start with a simple Random Forest algorithm:

- The only parameter changed so far was class\_weight, and it was changed to balanced because of our imbalanced dataset.
- We also used a small subset of features to start.
- Performance on both metrics (model accuracy and balanced accuracy) plummeted.

```
from sklearn.ensemble import RandomForestClassifier
    from sklearn.preprocessing import StandardScaler
    from sklearn.pipeline import Pipeline
    from imblearn.under_sampling import RandomUnderSampler
    from sklearn.decomposition import PCA
    from sklearn.metrics import accuracy_score, confusion_matrix, balanced_accuracy_score
    import time
```

# Note: Oversampling failed miserably. It gave incredible model accuracy but awful balanced accuracy, making the following models essentially useless:

```
In [12]: from sklearn.ensemble import RandomForestClassifier
         from imblearn.over sampling import ADASYN
         # create pipeline
        pipe = Pipeline(
            [("scaler", StandardScaler()), ("model", RandomForestClassifier(class weight='balanc
         # oversample minority class using ADASYN
         adasyn = ADASYN(random state=0)
        X resampled, y resampled = adasyn.fit resample(
            train data[["fw similarity", "adp dist", "axon dendrite dist", "nucleus dist", "rf d
            train data["connected"]
         # fit model
        pipe.fit(X resampled, y resampled)
         # predict on test data
         test data["pred"] = pipe.predict proba(test data[["fw similarity", "adp dist", "axon den
         # compute accuracy
        print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
         # confusion matrix
        print(confusion matrix(test data['connected'], test data['pred'] > .5))
         # compute balanced accuracy
        print(
            f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
        accuracy: 0.9380364301665456
        [[34808 2088]
         [ 215
                  5611
        balanced accuracy: 0.5750252829935064
```

In [62]: from imblearn.over\_sampling import SMOTE
from imblearn.under\_sampling import RandomUnderSampler

```
from imblearn.pipeline import Pipeline
# define pipeline
pipe = Pipeline([
    ("scaler", StandardScaler()),
    ("over", SMOTE(sampling strategy=0.5)), # over-sample the minority class to 50%
    ("under", RandomUnderSampler(sampling strategy=0.5)), # under-sample the majority c
    ("model", RandomForestClassifier(random state=1, n estimators=100, bootstrap=False))
])
# fit model
pipe.fit(train data[["fw similarity", "fw distance", "compartment synapse prop", "axon d
train data["connected"])
# predict on test data
test data["pred"] = pipe.predict proba(test data[["fw similarity", "fw distance", "compa
# compute accuracy
print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
# confusion matrix
print(confusion matrix(test data['connected'], test data['pred'] > .5))
# compute balanced accuracy
print(
    f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
accuracy: 0.9912825893938171
         59]
[[36837
[ 265
         611
balanced accuracy: 0.5102705660350185
```

## The Second Pick

As time would pass, many new features were engineered. Those features were tested on this Random Forest.

In the following RF model, all aforementioned features, except for those we did not use (morph\_similarity, morph\_distance, etc.) and except for any features relating to compartments, were used to train and test this model.

#### This model ended up becoming my second pick in Kaggle

Out of fear that my main model would perform poorly on new data, and may have overfit to the public leadboard data, this RF model was used as a backup since Random Forests are robust to overfitting.

#### Results:

We get a model accuracy of ~73.062% and a balanced accuracy of ~77.641%

Kaggle Public Score -> 74.295%\ Kaggle Private Score -> 77.900%

Note: GridSearchCV was not helpful. The "optimal hyperparameters" it gave me performed worse than simply leaving n\_estimators to 100 and bootstrap to False. However, cv was put to 3 for computational efficiency, so it's possible a cv of 5 or 10 would've yielded different results.

```
In [38]: # create pipeline
pipe = Pipeline(
```

```
start time = time.time()
         # undersample majority class
         rus = RandomUnderSampler(random state=0)
        X resampled, y resampled = rus.fit resample(
            train data[["fw similarity", "fw distance", "compartment synapse prop", "axon dendri
             train data["connected"]
         # fit model
         pipe.fit(X resampled, y resampled)
         # predict on test data
         test data["pred"] = pipe.predict proba(test data[["fw similarity", "fw distance", "compa
         # compute accuracy
        print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
         # confusion matrix
        print(confusion matrix(test data['connected'], test data['pred'] > .5))
         # compute balanced accuracy
        print(
            f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
         # Stop timer and print time
        end time = time.time()
        print(f"Time taken: {end time - start time} seconds")
        accuracy: 0.730621250033632
        [[26932 9964]
         [ 48 223]]
        balanced accuracy: 0.7764109270537631
        Time taken: 8.174513578414917 seconds
In [74]: from sklearn.model selection import GridSearchCV
         # define the parameter grid
        param grid = {
            'model n estimators': [10, 50, 100],
             'model max depth': [10, 20, 30, 40, 50, None],
             'model min samples split': [2, 5, 10],
             'model min samples leaf': [1, 2, 4],
             'model bootstrap': [True, False]
         # create a GridSearchCV object
        grid search = GridSearchCV(estimator=pipe, param grid=param grid, cv=3, verbose=2, n job
         # fit the model
         grid search.fit(X resampled, y resampled)
         # print the best parameters
        print(grid search.best params )
         # predict on test data using the best model
         test data["pred"] = grid search.predict proba(test data[["fw similarity", "fw distance",
         # compute accuracy
         print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
```

("model", RandomForestClassifier(random state=1, n estimators=100, bootstrap=False)

[("scaler", StandardScaler()),

```
# confusion matrix
        print(confusion matrix(test data['connected'], test data['pred'] > .5))
         # compute balanced accuracy
        print(
            f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
        Fitting 3 folds for each of 324 candidates, totalling 972 fits
        {'model bootstrap': True, 'model max depth': 10, 'model min samples leaf': 2, 'model
        min samples split': 10, 'model n estimators': 100}
        accuracy: 0.6970699814351441
        [[25674 11222]
         [ 37 234]]
        balanced accuracy: 0.7796582115322455
In [39]: # Get feature importances
         importances = pipe.named steps['model'].feature importances
         # Convert importances into a pandas Series for easier handling
         importances_series = pd.Series(importances, index=X resampled.columns.tolist())
         # Sort importances
         sorted importances = importances series.sort values(ascending=False)
         # Print the sorted importances
        print(sorted importances)
        adp dist
                                                      0.079954
                                                      0.067981
        compartment synapse prop
        pre oracle axon dendrite dist interaction
                                                     0.063369
        pre oracle adp dist
                                                     0.062627
        post oracle adp dist
                                                     0.061688
        axon dendrite dist manhattan
                                                     0.059267
        post skeletal distance to soma
                                                      0.052325
        axon dendrite dist
                                                     0.050342
        post_oracle_axon_dendrite_dist_interaction 0.042476
                                                     0.040098
        nucleus dist
        nucleus dist manhattan
                                                      0.037912
        pre skeletal distance to soma
                                                     0.035993
                                                      0.033663
        fw similarity
        post nucleus rf dist
                                                      0.033283
        rf dist
                                                      0.030249
        post oracle post rf x interaction
                                                     0.028154
                                                     0.027800
        fw distance
        post oracle post rf y interaction
                                                      0.027233
                                                     0.025689
        test score diff
        post oracle test score interaction
                                                     0.024630
        total synapses
                                                      0.022520
        pre nucleus rf dist
                                                      0.020284
        pre oracle pre rf x interaction
                                                     0.018277
        pre oracle test score interaction
                                                     0.017937
        pre oracle pre rf y interaction
                                                      0.015382
        post brain area V1
                                                      0.004606
        pre brain area RL
                                                      0.002374
        post brain area RL
                                                      0.002371
        pre brain area V1
                                                      0.002123
        pre AL post RL
                                                      0.001825
        post brain area AL
                                                      0.001752
                                                      0.001578
        pre brain area AL
        pre RL post AL
                                                      0.001376
        pre RL post V1
                                                      0.001273
        pre V1 post RL
                                                      0.000589
        pre AL post V1
                                                      0.000512
                                                      0.000490
        pre V1 post AL
        dtype: float64
```

# **Model Stacking**

Model stacking was used thanks to its popularity in machine learning. We are essentially improving predictive performance by combining the predictions of multiple models. Using diverse models can also help the stacked model be more robust and may end up generalizing better to new data.

There also was an attempt to use **PCA** for dimensionality reduction. However, as more work was done on the competition, PCA proved to not be as useful as originally thought.

#### To start:

- Random classifiers were used with default parameters to see what we would get.
- Out final estimator was Logistic Regression

#### Results:

We get a model accuracy of ~66.903% and a balanced accuracy of ~73.440%

```
In [92]:
        from sklearn.ensemble import StackingClassifier
         from sklearn.linear model import LogisticRegression
         from sklearn.ensemble import RandomForestClassifier
         from imblearn.over sampling import ADASYN
         from sklearn.feature_selection import SelectFromModel
         from sklearn.ensemble import RandomForestClassifier
        from imblearn.under sampling import RandomUnderSampler
         from sklearn.ensemble import StackingClassifier, AdaBoostClassifier, GradientBoostingCla
         from sklearn.linear model import LogisticRegression
         from sklearn.neighbors import KNeighborsClassifier
        from sklearn.svm import SVC
         from imblearn.over sampling import ADASYN
         from xgboost import XGBClassifier
        from sklearn.decomposition import PCA
         from sklearn.preprocessing import MinMaxScaler
```

```
In [21]: | from sklearn.ensemble import StackingClassifier, AdaBoostClassifier, GradientBoostingCla
         from sklearn.tree import DecisionTreeClassifier
         from sklearn.linear model import LogisticRegression
         from sklearn.svm import SVC
         from imblearn.over sampling import ADASYN
         # define base models
         base models = [
            ("svc", SVC(kernel='linear', C=10.0, probability=True, random state=2)),
             ("ada", AdaBoostClassifier(random_state=2)),
             ("gbc", GradientBoostingClassifier(random state=2)),
             ("lr", LogisticRegression()),
             ("knn", KNeighborsClassifier()),
             ("rf", RandomForestClassifier(random state=1))
         # create pipeline
         pipe = Pipeline(
            [("scaler", StandardScaler()),
              ("model", StackingClassifier(estimators=base models, final estimator=LogisticRegres
```

```
# undersample majority class
rus = RandomUnderSampler(random state=0)
X resampled, y resampled = rus.fit resample(
    train data[["fw similarity", "adp dist", "axon dendrite dist", "nucleus dist", "rf d
    train data["connected"]
# fit model
pipe.fit(X resampled, y resampled)
# predict on test data
test data["pred"] = pipe.predict proba(test data[["fw similarity", "adp dist", "axon den
# compute accuracy
print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
# confusion matrix
print(confusion matrix(test data['connected'], test data['pred'] > .5))
# compute balanced accuracy
print(
    f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
accuracy: 0.6690343584362473
[[24649 12247]
```

#### Next:

54

217]]

balanced accuracy: 0.7344025032563856

- MinMaxScaler was used instead of StandardScaler just as a test, but the difference was not very noticeable, and StandardScaler often performed slightly better.
- PCA was used with n\_components = 7, as it returned the best model accuracy, however PCA proved to be problematic in our Kaggle results
- When testing various final estimators, linear SVMs proved to be best (W/ C=10.0)

#### Results:

We get a model accuracy of ~70.431% and a balanced accuracy of ~75.034%

However, our Kaggle score was significantly worse: ~67.227%

Performance jumped the second PCA was removed

```
("model", StackingClassifier(estimators=base models, final estimator=SVC(kernel='li
# undersample majority class
rus = RandomUnderSampler(random state=0)
X resampled, y resampled = rus.fit resample(
    train data[["fw similarity", "adp dist", "axon dendrite dist", "nucleus dist", "rf d
    train data["connected"]
# fit model
pipe.fit(X resampled, y resampled)
# predict on test data
test data["pred"] = pipe.predict proba(test data[["fw similarity", "adp dist", "axon den
# compute accuracy
print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
# confusion matrix
print(confusion matrix(test data['connected'], test data['pred'] > .5))
# compute balanced accuracy
print(
    f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
accuracy: 0.7043075846853392
[[25961 10935]
[ 55 216]]
```

Here, we have some significant improvement in performance.

balanced accuracy: 0.7503371899232869

All aforementioned features that we engineered (apart from the compartment features) are being used to test and train the algorithm.

- SelectFromModel was used to select features based on importance weights
  - This was done through Random Forest to estimate feature importance
  - The threshold is set to 0.8\*mean, so features whose importance is greater than this threshold will be kept
- The base models were chosen through trial and error.
  - Note: There was tuning done for the base models, but I had more success manually adding and removing various models
- After getting rid of n\_components for PCA, our performance improved by not specifying
- The final estimator was kept at linear SVM

#### Results:

We get a model accuracy of ~72.266% and a balanced accuracy of ~78.156%

Our Kaggle score improved significantly: ~73.467%

Note: This score actually ended up being my second highest privatel leaderboard score: 78.133%.

```
In [94]: # Define preprocessor
```

```
# Scales/normalizes the data, then selects the most important features through Random Fo
preprocessor = Pipeline(
    [("scaler", MinMaxScaler()),
     ("selector", SelectFromModel(RandomForestClassifier(n estimators=100), threshold='0
# define base models
# The models used for model stacking
base models = [
   ("svc", SVC(kernel='linear', C=10.0, probability=True, random state=2)),
    ("ada", AdaBoostClassifier(n estimators=50, learning rate=1.0, random state=2)),
    ("knn", KNeighborsClassifier(n neighbors=1, weights='uniform', algorithm='auto', lea
    ("rf", RandomForestClassifier(random state=1, n estimators=10, bootstrap=True)),
    ("xgb", XGBClassifier(random state=2, max depth=7, learning rate=0.2, n estimators=1
# create pipeline
pipe = Pipeline(
   [("preprocessor", preprocessor),
     ("pca", PCA()),
     ("model", StackingClassifier(estimators=base models, final estimator=SVC(kernel='li
# undersample majority class
rus = RandomUnderSampler(random state=0)
X resampled, y resampled = rus.fit resample(
    train data[["fw similarity", "fw distance", "compartment synapse prop", "axon dendri
    train data["connected"]
# fit model
pipe.fit(X resampled, y resampled)
# predict on test data
test data["pred"] = pipe.predict proba(test data[["fw similarity", "fw distance", "compa
# compute accuracy
print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
# confusion matrix
print(confusion matrix(test data['connected'], test data['pred'] > .5))
# compute balanced accuracy
print(
    f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
accuracy: 0.7226571958995883
[[26631 10265]
 [ 43 228]]
balanced accuracy: 0.7815569863471835
```

### There were a number of base models that were tested for best performance

```
In [18]: from sklearn.ensemble import RandomForestClassifier
    from sklearn.linear_model import LogisticRegression
    from sklearn.neighbors import KNeighborsClassifier
    from sklearn.tree import DecisionTreeClassifier
    from sklearn.svm import SVC
    from sklearn.model_selection import cross_val_score
    from sklearn.ensemble import StackingClassifier, AdaBoostClassifier

# define base models
base_models1 = [
```

```
("lr", LogisticRegression()),
    ("knn", KNeighborsClassifier()),
    ("cart", DecisionTreeClassifier()),
    ("rf", RandomForestClassifier(random state=1)),
    ("ada", AdaBoostClassifier(random state=2)),
    ("svc", SVC(kernel='linear', C=10.0, probability=True, random state=2))
base models2 = [
    ("svc", SVC(kernel='linear', C=10.0, probability=True, random state=2)),
    ("rf", RandomForestClassifier(random state=1)),
    ("knn", KNeighborsClassifier()),
    ("ada", AdaBoostClassifier(random state=2))
# define meta models
meta model1 = LogisticRegression()
meta model2 = LogisticRegression()
# define a list of models
models = [
    StackingClassifier(estimators=base models1, final estimator=meta model1),
    StackingClassifier(estimators=base models2, final estimator=meta model2),
]
# for each model
for model in models:
   # create pipeline
   pipe = Pipeline(
        [("scaler", StandardScaler()), ("model", model)]
    # fit model
    pipe.fit(X resampled, y resampled)
    # perform cross-validation and print the average score
    scores = cross val score(pipe, X resampled, y resampled, cv=5, scoring='balanced acc
    print(f"Average cross-validation score: {scores.mean()}")
```

# **Deep Learning**

The last modeling technique used was deep learning, more specifically, feedforward neural networks

The first attempt was a five-layer neural network with the Adam optimizer and cross entropy loss.

The accuracies ended up being subpar, and were not tested in Kaggle:

We get a model accuracy of ~63.928% and a balanced accuracy of ~68.095%

Note: If there was more time, I would've worked more on tuning the neural networks in terms of number of layers and neurons per layer.

```
In [55]: import tensorflow as tf
    from tensorflow.keras.models import Sequential
    from tensorflow.keras.layers import Dense
```

```
from tensorflow.keras.wrappers.scikit learn import KerasClassifier
# Function to create model, required for KerasClassifier
def create model():
   model = Sequential()
   model.add(Dense(20, input dim=13, activation='relu')) # Adjust input dim to match n
   model.add(Dense(40, activation='relu'))
    model.add(Dense(60, activation='relu'))
    model.add(Dense(40, activation='relu'))
    model.add(Dense(1, activation='sigmoid'))
    # Compile model
    model.compile(loss='binary crossentropy', optimizer='adam', metrics=['accuracy'])
    return model
# Create model
model = KerasClassifier(build fn=create model, epochs=100, batch size=10, verbose=0)
# Create pipeline
pipe = Pipeline([("scaler", StandardScaler()), ("model", model)])
# Fit model
pipe.fit(X resampled, y resampled)
# Predict on test data
test data["pred"] = pipe.predict proba(test data[["fw similarity", "adp dist", "axon den
# Compute accuracy
print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
# Confusion matrix
print(confusion matrix(test data['connected'], test data['pred'] > .5))
# Compute balanced accuracy
print(f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['p
C:\Users\micha\AppData\Local\Temp\ipykernel 20776\1989297797.py:20: DeprecationWarning:
KerasClassifier is deprecated, use Sci-Keras (https://github.com/adriangb/scikeras) inst
ead. See https://www.adriangb.com/scikeras/stable/migration.html for help migrating.
 model = KerasClassifier(build fn=create model, epochs=100, batch size=10, verbose=0)
accuracy: 0.6392767777867463
[[23564 13332]
[ 75 196]]
balanced accuracy: 0.6809536249091892
```

#### Deep Learning Proved to be Difficult and Misleading

Below is the best performing FFNN that was possible.

- Three layers only, the input layer having 512 neurons and the hidden layer having only 16.
- Dropout was used to prevent overfitting, although it did not help as much as expected

#### Results:

We get a model accuracy of ~75.147% and a balanced accuracy of ~75.212%

Our Kaggle score was less than ideal: ~73.832%

• The use of 300 epochs most likely led to overfitting

Nonetheless, this model was the best scoring model on Kaggle

```
from sklearn.base import BaseEstimator, ClassifierMixin
In [67]:
         from tensorflow.keras.models import Sequential
         from tensorflow.keras.layers import Dense, Dropout
         from tensorflow.keras.optimizers import Adam
         from sklearn.preprocessing import StandardScaler
         from sklearn.pipeline import Pipeline
         from imblearn.under sampling import RandomUnderSampler
         import tensorflow as tf
         import numpy as np
         np.random.seed(0)
         tf.random.set seed(0)
         class KerasClassifier(BaseEstimator, ClassifierMixin):
             def __init__(self, input_dim):
                 self.input dim = input dim
                 self.model = self. build model()
             def build model(self):
                 model = Sequential()
                 model.add(Dense(512, input dim=self.input dim, activation='relu')) # Input laye
                 model.add(Dropout(0.5))
                 model.add(Dense(16, activation='relu')) # Hidden layer
                 model.add(Dropout(0.5))
                 model.add(Dense(1, activation='sigmoid')) # Output layer
                 model.compile(loss='binary crossentropy', optimizer=Adam(), metrics=['accuracy']
                 return model
             def fit(self, X, y):
                 self.model.fit(X, y, epochs=300, batch size=5)
                 return self
             def predict proba(self, X):
                 return self.model.predict(X)
         # create pipeline
         pipe = Pipeline(
             [("scaler", StandardScaler()),
              ("model", KerasClassifier(input dim=X resampled.shape[1]))]
         # undersample majority class
         rus = RandomUnderSampler(random state=0)
         X resampled, y resampled = rus.fit resample(
             train data[["fw similarity", "fw distance", "compartment synapse prop", "axon dendri
             "compartment apical",
             "compartment apical shaft",
             "compartment apical tuft",
             "compartment axon",
             "compartment basal",
             "compartment oblique",
             "compartment soma",
             "pre apical count",
             "pre apical shaft count",
             "pre apical tuft count",
             "pre axon count",
             "pre basal count",
             "pre oblique count",
             "pre soma_count",
             "post apical count",
             "post apical shaft count",
             "post apical tuft count",
             "post axon count",
             "post basal count",
             "post oblique count",
             "post soma_count"]],
```

```
train data["connected"]
# fit model
pipe.fit(X resampled, y resampled)
# predict on test data
test data["pred"] = pipe.predict proba(test data[["fw similarity", "fw distance", "compa
  "compartment apical",
  "compartment apical shaft",
  "compartment apical tuft",
  "compartment axon",
  "compartment basal",
  "compartment oblique",
  "compartment soma",
  "pre apical count",
  "pre apical shaft count",
  "pre apical tuft count",
  "pre axon count",
  "pre basal count",
  "pre oblique count",
  "pre soma_count",
  "post apical count",
  "post apical shaft count",
  "post apical tuft count",
  "post axon count",
  "post basal count",
  "post oblique count",
  "post soma count"]])
# compute accuracy
print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
# confusion matrix
print(confusion matrix(test data['connected'], test data['pred'] > .5))
# compute balanced accuracy
print(
  f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
Epoch 1/300
Epoch 2/300
Epoch 3/300
Epoch 4/300
Epoch 5/300
Epoch 6/300
Epoch 7/300
Epoch 8/300
Epoch 9/300
Epoch 10/300
Epoch 11/300
Epoch 12/300
```

```
Epoch 13/300
Epoch 14/300
Epoch 15/300
Epoch 16/300
Epoch 17/300
Epoch 18/300
Epoch 19/300
Epoch 20/300
Epoch 21/300
Epoch 22/300
Epoch 23/300
Epoch 24/300
Epoch 25/300
Epoch 26/300
Epoch 27/300
Epoch 28/300
Epoch 29/300
Epoch 30/300
Epoch 31/300
Epoch 32/300
Epoch 33/300
Epoch 34/300
Epoch 35/300
Epoch 36/300
Epoch 37/300
Epoch 38/300
Epoch 39/300
Epoch 40/300
Epoch 41/300
Epoch 42/300
Epoch 43/300
Epoch 44/300
Epoch 45/300
```

```
Epoch 46/300
Epoch 47/300
Epoch 48/300
Epoch 49/300
Epoch 50/300
Epoch 51/300
Epoch 52/300
Epoch 53/300
Epoch 54/300
Epoch 55/300
Epoch 56/300
Epoch 57/300
Epoch 58/300
Epoch 59/300
Epoch 60/300
Epoch 61/300
Epoch 62/300
Epoch 63/300
Epoch 64/300
Epoch 65/300
Epoch 66/300
Epoch 67/300
Epoch 68/300
Epoch 69/300
Epoch 70/300
Epoch 71/300
Epoch 72/300
Epoch 73/300
Epoch 74/300
Epoch 75/300
Epoch 76/300
Epoch 77/300
Epoch 78/300
```

```
Epoch 79/300
Epoch 80/300
Epoch 81/300
Epoch 82/300
Epoch 83/300
Epoch 84/300
Epoch 85/300
Epoch 86/300
Epoch 87/300
Epoch 88/300
Epoch 89/300
Epoch 90/300
Epoch 91/300
Epoch 92/300
Epoch 93/300
Epoch 94/300
Epoch 95/300
Epoch 96/300
Epoch 97/300
Epoch 98/300
Epoch 99/300
Epoch 100/300
Epoch 101/300
Epoch 102/300
Epoch 103/300
Epoch 104/300
Epoch 105/300
Epoch 106/300
Epoch 107/300
Epoch 108/300
Epoch 109/300
Epoch 110/300
Epoch 111/300
```

```
Epoch 112/300
Epoch 113/300
Epoch 114/300
Epoch 115/300
Epoch 116/300
Epoch 117/300
Epoch 118/300
Epoch 119/300
Epoch 120/300
Epoch 121/300
Epoch 122/300
Epoch 123/300
Epoch 124/300
Epoch 125/300
Epoch 126/300
Epoch 127/300
Epoch 128/300
Epoch 129/300
Epoch 130/300
Epoch 131/300
Epoch 132/300
Epoch 133/300
Epoch 134/300
Epoch 135/300
Epoch 136/300
Epoch 137/300
Epoch 138/300
Epoch 139/300
Epoch 140/300
Epoch 141/300
Epoch 142/300
Epoch 143/300
Epoch 144/300
```

```
Epoch 145/300
Epoch 146/300
Epoch 147/300
Epoch 148/300
Epoch 149/300
Epoch 150/300
Epoch 151/300
Epoch 152/300
Epoch 153/300
Epoch 154/300
Epoch 155/300
Epoch 156/300
Epoch 157/300
Epoch 158/300
Epoch 159/300
Epoch 160/300
Epoch 161/300
Epoch 162/300
Epoch 163/300
Epoch 164/300
Epoch 165/300
Epoch 166/300
Epoch 167/300
Epoch 168/300
Epoch 169/300
Epoch 170/300
Epoch 171/300
Epoch 172/300
Epoch 173/300
Epoch 174/300
Epoch 175/300
Epoch 176/300
Epoch 177/300
```

```
Epoch 178/300
Epoch 179/300
Epoch 180/300
Epoch 181/300
Epoch 182/300
Epoch 183/300
Epoch 184/300
Epoch 185/300
Epoch 186/300
Epoch 187/300
Epoch 188/300
Epoch 189/300
Epoch 190/300
Epoch 191/300
Epoch 192/300
Epoch 193/300
Epoch 194/300
Epoch 195/300
Epoch 196/300
Epoch 197/300
Epoch 198/300
Epoch 199/300
Epoch 200/300
Epoch 201/300
Epoch 202/300
Epoch 203/300
Epoch 204/300
Epoch 205/300
Epoch 206/300
Epoch 207/300
Epoch 208/300
Epoch 209/300
Epoch 210/300
```

```
Epoch 211/300
Epoch 212/300
Epoch 213/300
Epoch 214/300
Epoch 215/300
Epoch 216/300
Epoch 217/300
Epoch 218/300
Epoch 219/300
Epoch 220/300
Epoch 221/300
Epoch 222/300
Epoch 223/300
Epoch 224/300
Epoch 225/300
Epoch 226/300
Epoch 227/300
Epoch 228/300
Epoch 229/300
Epoch 230/300
Epoch 231/300
Epoch 232/300
Epoch 233/300
Epoch 234/300
Epoch 235/300
Epoch 236/300
Epoch 237/300
Epoch 238/300
Epoch 239/300
Epoch 240/300
Epoch 241/300
Epoch 242/300
Epoch 243/300
```

```
Epoch 244/300
Epoch 245/300
Epoch 246/300
Epoch 247/300
Epoch 248/300
Epoch 249/300
Epoch 250/300
Epoch 251/300
Epoch 252/300
Epoch 253/300
Epoch 254/300
Epoch 255/300
Epoch 256/300
Epoch 257/300
Epoch 258/300
Epoch 259/300
Epoch 260/300
Epoch 261/300
Epoch 262/300
Epoch 263/300
Epoch 264/300
Epoch 265/300
Epoch 266/300
Epoch 267/300
Epoch 268/300
Epoch 269/300
Epoch 270/300
Epoch 271/300
Epoch 272/300
Epoch 273/300
Epoch 274/300
Epoch 275/300
Epoch 276/300
```

```
Epoch 277/300
Epoch 278/300
Epoch 279/300
Epoch 280/300
Epoch 281/300
Epoch 282/300
Epoch 283/300
Epoch 284/300
Epoch 285/300
Epoch 286/300
Epoch 287/300
Epoch 288/300
Epoch 289/300
Epoch 290/300
Epoch 291/300
Epoch 292/300
Epoch 293/300
Epoch 294/300
Epoch 295/300
Epoch 296/300
Epoch 297/300
Epoch 298/300
Epoch 299/300
Epoch 300/300
accuracy: 0.7514730809589152
[[27726 9170]
[ 67 204]]
balanced accuracy: 0.752115550481177
```

## **Cross-Validation**

```
In [68]: import numpy as np
    from sklearn.model_selection import KFold

class KerasClassifier(BaseEstimator, ClassifierMixin):
    def __init__(self, input_dim):
        self.input_dim = input_dim
        self.model = self._build_model()

def _build_model(self):
    model = Sequential()
```

```
model.add(Dense(512, input dim=self.input dim, activation='relu')) # Input laye
     model.add(Dropout(0.5))
     model.add(Dense(16, activation='relu')) # Hidden layer
     model.add(Dropout(0.5))
     model.add(Dense(1, activation='sigmoid')) # Output layer
     model.compile(loss='binary crossentropy', optimizer=Adam(), metrics=['accuracy']
     return model
  def fit(self, X, y):
     self.model.fit(X, y, epochs=300, batch size=5)
     return self
  def predict proba(self, X):
     return self.model.predict(X)
  def predict(self, X):
     return (self.predict proba(X) > 0.5).astype(int)
# Define the number of folds for cross-validation
n folds = 5
kfold = KFold(n splits=n folds, shuffle=True)
# Initialize list to store scores for each fold
scores = []
# Loop over each fold
for train, test in kfold.split(X resampled, y resampled):
  # Reset the pipeline
  pipe = Pipeline(
     [("scaler", StandardScaler()),
      ("model", KerasClassifier(input dim=X resampled.shape[1]))]
  # Select the training and testing data for this fold
  X train, X test = X resampled.iloc[train], X resampled.iloc[test]
  y train, y test = y resampled.iloc[train], y resampled.iloc[test]
  # Fit the model
  pipe.fit(X train, y train)
  # Evaluate the model
  score = pipe.score(X test, y test)
  # Store the score
  scores.append(score)
# Print the mean and standard deviation of the scores
print(f"Cross-validation mean accuracy: {np.mean(scores)}, std: {np.std(scores)}")
Epoch 1/300
Epoch 2/300
Epoch 3/300
Epoch 4/300
Epoch 5/300
Epoch 6/300
Epoch 7/300
Epoch 8/300
Epoch 9/300
```

```
Epoch 10/300
Epoch 11/300
Epoch 12/300
Epoch 13/300
Epoch 14/300
Epoch 15/300
Epoch 16/300
Epoch 17/300
Epoch 18/300
Epoch 19/300
Epoch 20/300
Epoch 21/300
Epoch 22/300
Epoch 23/300
Epoch 24/300
Epoch 25/300
Epoch 26/300
Epoch 27/300
Epoch 28/300
Epoch 29/300
Epoch 30/300
Epoch 31/300
Epoch 32/300
Epoch 33/300
Epoch 34/300
Epoch 35/300
Epoch 36/300
Epoch 37/300
Epoch 38/300
Epoch 39/300
Epoch 40/300
Epoch 41/300
```

Epoch 42/300

```
Epoch 43/300
Epoch 44/300
Epoch 45/300
Epoch 46/300
Epoch 47/300
Epoch 48/300
Epoch 49/300
Epoch 50/300
Epoch 51/300
Epoch 52/300
Epoch 53/300
Epoch 54/300
Epoch 55/300
Epoch 56/300
Epoch 57/300
Epoch 58/300
Epoch 59/300
Epoch 60/300
Epoch 61/300
Epoch 62/300
Epoch 63/300
Epoch 64/300
Epoch 65/300
Epoch 66/300
Epoch 67/300
Epoch 68/300
Epoch 69/300
Epoch 70/300
Epoch 71/300
Epoch 72/300
Epoch 73/300
Epoch 74/300
```

Epoch 75/300

```
Epoch 76/300
Epoch 77/300
Epoch 78/300
Epoch 79/300
Epoch 80/300
Epoch 81/300
Epoch 82/300
Epoch 83/300
Epoch 84/300
Epoch 85/300
Epoch 86/300
Epoch 87/300
Epoch 88/300
Epoch 89/300
Epoch 90/300
Epoch 91/300
Epoch 92/300
Epoch 93/300
Epoch 94/300
Epoch 95/300
Epoch 96/300
Epoch 97/300
Epoch 98/300
Epoch 99/300
Epoch 100/300
Epoch 101/300
Epoch 102/300
Epoch 103/300
Epoch 104/300
Epoch 105/300
Epoch 106/300
Epoch 107/300
```

Epoch 108/300

```
Epoch 109/300
Epoch 110/300
Epoch 111/300
Epoch 112/300
Epoch 113/300
Epoch 114/300
Epoch 115/300
Epoch 116/300
Epoch 117/300
Epoch 118/300
Epoch 119/300
Epoch 120/300
Epoch 121/300
Epoch 122/300
Epoch 123/300
Epoch 124/300
Epoch 125/300
Epoch 126/300
Epoch 127/300
Epoch 128/300
Epoch 129/300
Epoch 130/300
Epoch 131/300
Epoch 132/300
Epoch 133/300
Epoch 134/300
Epoch 135/300
Epoch 136/300
Epoch 137/300
Epoch 138/300
Epoch 139/300
Epoch 140/300
```

Epoch 141/300

```
Epoch 142/300
Epoch 143/300
Epoch 144/300
Epoch 145/300
Epoch 146/300
Epoch 147/300
Epoch 148/300
Epoch 149/300
Epoch 150/300
Epoch 151/300
Epoch 152/300
Epoch 153/300
Epoch 154/300
Epoch 155/300
Epoch 156/300
Epoch 157/300
Epoch 158/300
Epoch 159/300
Epoch 160/300
Epoch 161/300
Epoch 162/300
Epoch 163/300
Epoch 164/300
Epoch 165/300
Epoch 166/300
Epoch 167/300
Epoch 168/300
Epoch 169/300
Epoch 170/300
Epoch 171/300
Epoch 172/300
Epoch 173/300
```

Epoch 174/300

```
Epoch 175/300
Epoch 176/300
Epoch 177/300
Epoch 178/300
Epoch 179/300
Epoch 180/300
Epoch 181/300
Epoch 182/300
Epoch 183/300
Epoch 184/300
Epoch 185/300
Epoch 186/300
Epoch 187/300
Epoch 188/300
Epoch 189/300
Epoch 190/300
Epoch 191/300
Epoch 192/300
Epoch 193/300
Epoch 194/300
Epoch 195/300
Epoch 196/300
Epoch 197/300
Epoch 198/300
Epoch 199/300
Epoch 200/300
Epoch 201/300
Epoch 202/300
Epoch 203/300
Epoch 204/300
Epoch 205/300
Epoch 206/300
```

Epoch 207/300

```
Epoch 208/300
Epoch 209/300
Epoch 210/300
Epoch 211/300
Epoch 212/300
Epoch 213/300
Epoch 214/300
Epoch 215/300
Epoch 216/300
Epoch 217/300
Epoch 218/300
Epoch 219/300
Epoch 220/300
Epoch 221/300
Epoch 222/300
Epoch 223/300
Epoch 224/300
Epoch 225/300
Epoch 226/300
Epoch 227/300
Epoch 228/300
Epoch 229/300
Epoch 230/300
Epoch 231/300
Epoch 232/300
Epoch 233/300
Epoch 234/300
Epoch 235/300
Epoch 236/300
Epoch 237/300
Epoch 238/300
Epoch 239/300
```

Epoch 240/300

```
Epoch 241/300
Epoch 242/300
Epoch 243/300
Epoch 244/300
Epoch 245/300
Epoch 246/300
Epoch 247/300
Epoch 248/300
Epoch 249/300
Epoch 250/300
Epoch 251/300
Epoch 252/300
Epoch 253/300
Epoch 254/300
Epoch 255/300
Epoch 256/300
Epoch 257/300
Epoch 258/300
Epoch 259/300
Epoch 260/300
Epoch 261/300
Epoch 262/300
Epoch 263/300
Epoch 264/300
Epoch 265/300
Epoch 266/300
Epoch 267/300
Epoch 268/300
Epoch 269/300
Epoch 270/300
Epoch 271/300
Epoch 272/300
```

Epoch 273/300

```
Epoch 274/300
Epoch 275/300
Epoch 276/300
Epoch 277/300
Epoch 278/300
Epoch 279/300
Epoch 280/300
Epoch 281/300
Epoch 282/300
Epoch 283/300
Epoch 284/300
Epoch 285/300
Epoch 286/300
Epoch 287/300
Epoch 288/300
Epoch 289/300
Epoch 290/300
Epoch 291/300
Epoch 292/300
Epoch 293/300
Epoch 294/300
Epoch 295/300
Epoch 296/300
Epoch 297/300
Epoch 298/300
Epoch 299/300
Epoch 300/300
Epoch 1/300
Epoch 2/300
Epoch 3/300
Epoch 4/300
Epoch 5/300
```

Epoch 6/300

```
Epoch 7/300
Epoch 8/300
Epoch 9/300
Epoch 10/300
Epoch 11/300
Epoch 12/300
Epoch 13/300
Epoch 14/300
Epoch 15/300
Epoch 16/300
Epoch 17/300
Epoch 18/300
Epoch 19/300
Epoch 20/300
Epoch 21/300
Epoch 22/300
Epoch 23/300
Epoch 24/300
Epoch 25/300
Epoch 26/300
Epoch 27/300
Epoch 28/300
Epoch 29/300
Epoch 30/300
Epoch 31/300
Epoch 32/300
Epoch 33/300
Epoch 34/300
Epoch 35/300
Epoch 36/300
Epoch 37/300
Epoch 38/300
```

Epoch 39/300

```
Epoch 40/300
Epoch 41/300
Epoch 42/300
Epoch 43/300
Epoch 44/300
Epoch 45/300
Epoch 46/300
Epoch 47/300
Epoch 48/300
Epoch 49/300
Epoch 50/300
Epoch 51/300
Epoch 52/300
Epoch 53/300
Epoch 54/300
Epoch 55/300
Epoch 56/300
Epoch 57/300
Epoch 58/300
Epoch 59/300
Epoch 60/300
Epoch 61/300
Epoch 62/300
Epoch 63/300
Epoch 64/300
Epoch 65/300
Epoch 66/300
Epoch 67/300
Epoch 68/300
Epoch 69/300
Epoch 70/300
Epoch 71/300
```

Epoch 72/300

```
Epoch 73/300
Epoch 74/300
Epoch 75/300
Epoch 76/300
Epoch 77/300
Epoch 78/300
Epoch 79/300
Epoch 80/300
Epoch 81/300
Epoch 82/300
Epoch 83/300
Epoch 84/300
Epoch 85/300
Epoch 86/300
Epoch 87/300
Epoch 88/300
Epoch 89/300
Epoch 90/300
Epoch 91/300
Epoch 92/300
Epoch 93/300
Epoch 94/300
Epoch 95/300
Epoch 96/300
Epoch 97/300
Epoch 98/300
Epoch 99/300
Epoch 100/300
Epoch 101/300
Epoch 102/300
Epoch 103/300
Epoch 104/300
```

Epoch 105/300

```
Epoch 106/300
Epoch 107/300
Epoch 108/300
Epoch 109/300
Epoch 110/300
Epoch 111/300
Epoch 112/300
Epoch 113/300
Epoch 114/300
Epoch 115/300
Epoch 116/300
Epoch 117/300
Epoch 118/300
Epoch 119/300
Epoch 120/300
Epoch 121/300
Epoch 122/300
Epoch 123/300
Epoch 124/300
Epoch 125/300
Epoch 126/300
Epoch 127/300
Epoch 128/300
Epoch 129/300
Epoch 130/300
Epoch 131/300
Epoch 132/300
Epoch 133/300
Epoch 134/300
Epoch 135/300
Epoch 136/300
Epoch 137/300
```

Epoch 138/300

```
Epoch 139/300
Epoch 140/300
Epoch 141/300
Epoch 142/300
Epoch 143/300
Epoch 144/300
Epoch 145/300
Epoch 146/300
Epoch 147/300
Epoch 148/300
Epoch 149/300
Epoch 150/300
Epoch 151/300
Epoch 152/300
Epoch 153/300
Epoch 154/300
Epoch 155/300
Epoch 156/300
Epoch 157/300
Epoch 158/300
Epoch 159/300
Epoch 160/300
Epoch 161/300
Epoch 162/300
Epoch 163/300
Epoch 164/300
Epoch 165/300
Epoch 166/300
Epoch 167/300
Epoch 168/300
Epoch 169/300
Epoch 170/300
```

Epoch 171/300

```
Epoch 172/300
Epoch 173/300
Epoch 174/300
Epoch 175/300
Epoch 176/300
Epoch 177/300
Epoch 178/300
Epoch 179/300
Epoch 180/300
Epoch 181/300
Epoch 182/300
Epoch 183/300
Epoch 184/300
Epoch 185/300
Epoch 186/300
Epoch 187/300
Epoch 188/300
Epoch 189/300
Epoch 190/300
Epoch 191/300
Epoch 192/300
Epoch 193/300
Epoch 194/300
Epoch 195/300
Epoch 196/300
Epoch 197/300
Epoch 198/300
Epoch 199/300
Epoch 200/300
Epoch 201/300
Epoch 202/300
Epoch 203/300
```

Epoch 204/300

```
Epoch 205/300
Epoch 206/300
Epoch 207/300
Epoch 208/300
Epoch 209/300
Epoch 210/300
Epoch 211/300
Epoch 212/300
Epoch 213/300
Epoch 214/300
Epoch 215/300
Epoch 216/300
Epoch 217/300
Epoch 218/300
Epoch 219/300
Epoch 220/300
Epoch 221/300
Epoch 222/300
Epoch 223/300
Epoch 224/300
Epoch 225/300
Epoch 226/300
Epoch 227/300
Epoch 228/300
Epoch 229/300
Epoch 230/300
Epoch 231/300
Epoch 232/300
Epoch 233/300
Epoch 234/300
Epoch 235/300
Epoch 236/300
```

Epoch 237/300

```
Epoch 238/300
Epoch 239/300
Epoch 240/300
Epoch 241/300
Epoch 242/300
Epoch 243/300
Epoch 244/300
Epoch 245/300
Epoch 246/300
Epoch 247/300
Epoch 248/300
Epoch 249/300
Epoch 250/300
Epoch 251/300
Epoch 252/300
Epoch 253/300
Epoch 254/300
Epoch 255/300
Epoch 256/300
Epoch 257/300
Epoch 258/300
Epoch 259/300
Epoch 260/300
Epoch 261/300
Epoch 262/300
Epoch 263/300
Epoch 264/300
Epoch 265/300
Epoch 266/300
Epoch 267/300
Epoch 268/300
Epoch 269/300
```

Epoch 270/300

```
Epoch 271/300
Epoch 272/300
Epoch 273/300
Epoch 274/300
Epoch 275/300
Epoch 276/300
Epoch 277/300
Epoch 278/300
Epoch 279/300
Epoch 280/300
Epoch 281/300
Epoch 282/300
Epoch 283/300
Epoch 284/300
Epoch 285/300
Epoch 286/300
Epoch 287/300
Epoch 288/300
Epoch 289/300
Epoch 290/300
Epoch 291/300
Epoch 292/300
Epoch 293/300
Epoch 294/300
Epoch 295/300
Epoch 296/300
Epoch 297/300
Epoch 298/300
Epoch 299/300
Epoch 300/300
Epoch 1/300
Epoch 2/300
```

Epoch 3/300

```
Epoch 4/300
Epoch 5/300
Epoch 6/300
Epoch 7/300
Epoch 8/300
Epoch 9/300
Epoch 10/300
Epoch 11/300
Epoch 12/300
Epoch 13/300
Epoch 14/300
Epoch 15/300
Epoch 16/300
Epoch 17/300
Epoch 18/300
Epoch 19/300
Epoch 20/300
Epoch 21/300
Epoch 22/300
Epoch 23/300
Epoch 24/300
Epoch 25/300
Epoch 26/300
Epoch 27/300
Epoch 28/300
Epoch 29/300
Epoch 30/300
Epoch 31/300
Epoch 32/300
Epoch 33/300
Epoch 34/300
Epoch 35/300
```

Epoch 36/300

```
Epoch 37/300
Epoch 38/300
Epoch 39/300
Epoch 40/300
Epoch 41/300
Epoch 42/300
Epoch 43/300
Epoch 44/300
Epoch 45/300
Epoch 46/300
Epoch 47/300
Epoch 48/300
Epoch 49/300
Epoch 50/300
Epoch 51/300
Epoch 52/300
Epoch 53/300
Epoch 54/300
Epoch 55/300
Epoch 56/300
Epoch 57/300
Epoch 58/300
Epoch 59/300
Epoch 60/300
Epoch 61/300
Epoch 62/300
Epoch 63/300
Epoch 64/300
Epoch 65/300
Epoch 66/300
Epoch 67/300
Epoch 68/300
```

Epoch 69/300

```
Epoch 70/300
Epoch 71/300
Epoch 72/300
Epoch 73/300
Epoch 74/300
Epoch 75/300
Epoch 76/300
Epoch 77/300
Epoch 78/300
Epoch 79/300
Epoch 80/300
Epoch 81/300
Epoch 82/300
Epoch 83/300
Epoch 84/300
Epoch 85/300
Epoch 86/300
Epoch 87/300
Epoch 88/300
Epoch 89/300
Epoch 90/300
Epoch 91/300
Epoch 92/300
Epoch 93/300
Epoch 94/300
Epoch 95/300
Epoch 96/300
Epoch 97/300
Epoch 98/300
Epoch 99/300
Epoch 100/300
Epoch 101/300
```

Epoch 102/300

```
Epoch 103/300
Epoch 104/300
Epoch 105/300
Epoch 106/300
Epoch 107/300
Epoch 108/300
Epoch 109/300
Epoch 110/300
Epoch 111/300
Epoch 112/300
Epoch 113/300
Epoch 114/300
Epoch 115/300
Epoch 116/300
Epoch 117/300
Epoch 118/300
Epoch 119/300
Epoch 120/300
Epoch 121/300
Epoch 122/300
Epoch 123/300
Epoch 124/300
Epoch 125/300
Epoch 126/300
Epoch 127/300
Epoch 128/300
Epoch 129/300
Epoch 130/300
Epoch 131/300
Epoch 132/300
Epoch 133/300
Epoch 134/300
```

Epoch 135/300

```
Epoch 136/300
Epoch 137/300
Epoch 138/300
Epoch 139/300
Epoch 140/300
Epoch 141/300
Epoch 142/300
Epoch 143/300
Epoch 144/300
Epoch 145/300
Epoch 146/300
Epoch 147/300
Epoch 148/300
Epoch 149/300
Epoch 150/300
Epoch 151/300
Epoch 152/300
Epoch 153/300
Epoch 154/300
Epoch 155/300
Epoch 156/300
Epoch 157/300
Epoch 158/300
Epoch 159/300
Epoch 160/300
Epoch 161/300
Epoch 162/300
Epoch 163/300
Epoch 164/300
Epoch 165/300
Epoch 166/300
Epoch 167/300
```

Epoch 168/300

```
Epoch 169/300
Epoch 170/300
Epoch 171/300
Epoch 172/300
Epoch 173/300
Epoch 174/300
Epoch 175/300
Epoch 176/300
Epoch 177/300
Epoch 178/300
Epoch 179/300
Epoch 180/300
Epoch 181/300
Epoch 182/300
Epoch 183/300
Epoch 184/300
Epoch 185/300
Epoch 186/300
Epoch 187/300
Epoch 188/300
Epoch 189/300
Epoch 190/300
Epoch 191/300
Epoch 192/300
Epoch 193/300
Epoch 194/300
Epoch 195/300
Epoch 196/300
Epoch 197/300
Epoch 198/300
Epoch 199/300
Epoch 200/300
```

Epoch 201/300

```
Epoch 202/300
Epoch 203/300
Epoch 204/300
Epoch 205/300
Epoch 206/300
Epoch 207/300
Epoch 208/300
Epoch 209/300
Epoch 210/300
Epoch 211/300
Epoch 212/300
Epoch 213/300
Epoch 214/300
Epoch 215/300
Epoch 216/300
Epoch 217/300
Epoch 218/300
Epoch 219/300
Epoch 220/300
Epoch 221/300
Epoch 222/300
Epoch 223/300
Epoch 224/300
Epoch 225/300
Epoch 226/300
Epoch 227/300
Epoch 228/300
Epoch 229/300
Epoch 230/300
Epoch 231/300
Epoch 232/300
Epoch 233/300
```

Epoch 234/300

```
Epoch 235/300
Epoch 236/300
Epoch 237/300
Epoch 238/300
Epoch 239/300
Epoch 240/300
Epoch 241/300
Epoch 242/300
Epoch 243/300
Epoch 244/300
Epoch 245/300
Epoch 246/300
Epoch 247/300
Epoch 248/300
Epoch 249/300
Epoch 250/300
Epoch 251/300
Epoch 252/300
Epoch 253/300
Epoch 254/300
Epoch 255/300
Epoch 256/300
Epoch 257/300
Epoch 258/300
Epoch 259/300
Epoch 260/300
Epoch 261/300
Epoch 262/300
Epoch 263/300
Epoch 264/300
Epoch 265/300
Epoch 266/300
```

Epoch 267/300

```
Epoch 268/300
Epoch 269/300
Epoch 270/300
Epoch 271/300
Epoch 272/300
Epoch 273/300
Epoch 274/300
Epoch 275/300
Epoch 276/300
Epoch 277/300
Epoch 278/300
Epoch 279/300
Epoch 280/300
Epoch 281/300
Epoch 282/300
Epoch 283/300
Epoch 284/300
Epoch 285/300
Epoch 286/300
Epoch 287/300
Epoch 288/300
Epoch 289/300
Epoch 290/300
Epoch 291/300
Epoch 292/300
Epoch 293/300
Epoch 294/300
Epoch 295/300
Epoch 296/300
Epoch 297/300
Epoch 298/300
Epoch 299/300
```

Epoch 300/300

```
Epoch 1/300
Epoch 2/300
Epoch 3/300
Epoch 4/300
Epoch 5/300
Epoch 6/300
Epoch 7/300
Epoch 8/300
Epoch 9/300
Epoch 10/300
Epoch 11/300
Epoch 12/300
Epoch 13/300
Epoch 14/300
Epoch 15/300
Epoch 16/300
Epoch 17/300
Epoch 18/300
Epoch 19/300
Epoch 20/300
Epoch 21/300
Epoch 22/300
Epoch 23/300
Epoch 24/300
Epoch 25/300
Epoch 26/300
Epoch 27/300
Epoch 28/300
Epoch 29/300
Epoch 30/300
Epoch 31/300
Epoch 32/300
```

Epoch 33/300

```
Epoch 34/300
Epoch 35/300
Epoch 36/300
Epoch 37/300
Epoch 38/300
Epoch 39/300
Epoch 40/300
Epoch 41/300
Epoch 42/300
Epoch 43/300
Epoch 44/300
Epoch 45/300
Epoch 46/300
Epoch 47/300
Epoch 48/300
Epoch 49/300
Epoch 50/300
Epoch 51/300
Epoch 52/300
Epoch 53/300
Epoch 54/300
Epoch 55/300
Epoch 56/300
Epoch 57/300
Epoch 58/300
Epoch 59/300
Epoch 60/300
Epoch 61/300
Epoch 62/300
Epoch 63/300
Epoch 64/300
Epoch 65/300
```

Epoch 66/300

```
Epoch 67/300
Epoch 68/300
Epoch 69/300
Epoch 70/300
Epoch 71/300
Epoch 72/300
Epoch 73/300
Epoch 74/300
Epoch 75/300
Epoch 76/300
Epoch 77/300
Epoch 78/300
Epoch 79/300
Epoch 80/300
Epoch 81/300
Epoch 82/300
Epoch 83/300
Epoch 84/300
Epoch 85/300
Epoch 86/300
Epoch 87/300
Epoch 88/300
Epoch 89/300
Epoch 90/300
Epoch 91/300
Epoch 92/300
Epoch 93/300
Epoch 94/300
Epoch 95/300
Epoch 96/300
Epoch 97/300
Epoch 98/300
```

Epoch 99/300

```
Epoch 100/300
Epoch 101/300
Epoch 102/300
Epoch 103/300
Epoch 104/300
Epoch 105/300
Epoch 106/300
Epoch 107/300
Epoch 108/300
Epoch 109/300
Epoch 110/300
Epoch 111/300
Epoch 112/300
Epoch 113/300
Epoch 114/300
Epoch 115/300
Epoch 116/300
Epoch 117/300
Epoch 118/300
Epoch 119/300
Epoch 120/300
Epoch 121/300
Epoch 122/300
Epoch 123/300
Epoch 124/300
Epoch 125/300
Epoch 126/300
Epoch 127/300
Epoch 128/300
Epoch 129/300
Epoch 130/300
Epoch 131/300
```

Epoch 132/300

```
Epoch 133/300
Epoch 134/300
Epoch 135/300
Epoch 136/300
Epoch 137/300
Epoch 138/300
Epoch 139/300
Epoch 140/300
Epoch 141/300
Epoch 142/300
Epoch 143/300
Epoch 144/300
Epoch 145/300
Epoch 146/300
Epoch 147/300
Epoch 148/300
Epoch 149/300
Epoch 150/300
Epoch 151/300
Epoch 152/300
Epoch 153/300
Epoch 154/300
Epoch 155/300
Epoch 156/300
Epoch 157/300
Epoch 158/300
Epoch 159/300
Epoch 160/300
Epoch 161/300
Epoch 162/300
Epoch 163/300
Epoch 164/300
```

```
Epoch 165/300
Epoch 166/300
Epoch 167/300
Epoch 168/300
Epoch 169/300
Epoch 170/300
Epoch 171/300
Epoch 172/300
Epoch 173/300
Epoch 174/300
Epoch 175/300
Epoch 176/300
Epoch 177/300
Epoch 178/300
Epoch 179/300
Epoch 180/300
Epoch 181/300
Epoch 182/300
Epoch 183/300
Epoch 184/300
Epoch 185/300
Epoch 186/300
Epoch 187/300
Epoch 188/300
Epoch 189/300
Epoch 190/300
Epoch 191/300
Epoch 192/300
Epoch 193/300
Epoch 194/300
Epoch 195/300
Epoch 196/300
Epoch 197/300
```

```
Epoch 198/300
Epoch 199/300
Epoch 200/300
Epoch 201/300
Epoch 202/300
Epoch 203/300
Epoch 204/300
Epoch 205/300
Epoch 206/300
Epoch 207/300
Epoch 208/300
Epoch 209/300
Epoch 210/300
Epoch 211/300
Epoch 212/300
Epoch 213/300
Epoch 214/300
Epoch 215/300
Epoch 216/300
Epoch 217/300
Epoch 218/300
Epoch 219/300
Epoch 220/300
Epoch 221/300
Epoch 222/300
Epoch 223/300
Epoch 224/300
Epoch 225/300
Epoch 226/300
Epoch 227/300
Epoch 228/300
Epoch 229/300
Epoch 230/300
```

```
Epoch 231/300
Epoch 232/300
Epoch 233/300
Epoch 234/300
Epoch 235/300
Epoch 236/300
Epoch 237/300
Epoch 238/300
Epoch 239/300
Epoch 240/300
Epoch 241/300
Epoch 242/300
Epoch 243/300
Epoch 244/300
Epoch 245/300
Epoch 246/300
Epoch 247/300
Epoch 248/300
Epoch 249/300
Epoch 250/300
Epoch 251/300
Epoch 252/300
Epoch 253/300
Epoch 254/300
Epoch 255/300
Epoch 256/300
Epoch 257/300
Epoch 258/300
Epoch 259/300
Epoch 260/300
Epoch 261/300
Epoch 262/300
Epoch 263/300
```

```
Epoch 264/300
Epoch 265/300
Epoch 266/300
Epoch 267/300
Epoch 268/300
Epoch 269/300
Epoch 270/300
Epoch 271/300
Epoch 272/300
Epoch 273/300
Epoch 274/300
Epoch 275/300
Epoch 276/300
Epoch 277/300
Epoch 278/300
Epoch 279/300
Epoch 280/300
Epoch 281/300
Epoch 282/300
Epoch 283/300
Epoch 284/300
Epoch 285/300
Epoch 286/300
Epoch 287/300
Epoch 288/300
Epoch 289/300
Epoch 290/300
Epoch 291/300
Epoch 292/300
Epoch 293/300
Epoch 294/300
Epoch 295/300
Epoch 296/300
```

```
Epoch 297/300
Epoch 298/300
Epoch 299/300
Epoch 300/300
Epoch 1/300
Epoch 2/300
Epoch 3/300
Epoch 4/300
Epoch 5/300
Epoch 6/300
Epoch 7/300
Epoch 8/300
Epoch 9/300
Epoch 10/300
Epoch 11/300
Epoch 12/300
Epoch 13/300
Epoch 14/300
Epoch 15/300
Epoch 16/300
Epoch 17/300
Epoch 18/300
Epoch 19/300
Epoch 20/300
Epoch 21/300
Epoch 22/300
Epoch 23/300
Epoch 24/300
Epoch 25/300
Epoch 26/300
Epoch 27/300
Epoch 28/300
Epoch 29/300
```

```
Epoch 30/300
Epoch 31/300
Epoch 32/300
Epoch 33/300
Epoch 34/300
Epoch 35/300
Epoch 36/300
Epoch 37/300
Epoch 38/300
Epoch 39/300
Epoch 40/300
Epoch 41/300
Epoch 42/300
Epoch 43/300
Epoch 44/300
Epoch 45/300
Epoch 46/300
Epoch 47/300
Epoch 48/300
Epoch 49/300
Epoch 50/300
Epoch 51/300
Epoch 52/300
Epoch 53/300
Epoch 54/300
Epoch 55/300
Epoch 56/300
Epoch 57/300
Epoch 58/300
Epoch 59/300
Epoch 60/300
Epoch 61/300
Epoch 62/300
```

```
Epoch 63/300
Epoch 64/300
Epoch 65/300
Epoch 66/300
Epoch 67/300
Epoch 68/300
Epoch 69/300
Epoch 70/300
Epoch 71/300
Epoch 72/300
Epoch 73/300
Epoch 74/300
Epoch 75/300
Epoch 76/300
Epoch 77/300
Epoch 78/300
Epoch 79/300
Epoch 80/300
Epoch 81/300
Epoch 82/300
Epoch 83/300
Epoch 84/300
Epoch 85/300
Epoch 86/300
Epoch 87/300
Epoch 88/300
Epoch 89/300
Epoch 90/300
Epoch 91/300
Epoch 92/300
Epoch 93/300
Epoch 94/300
Epoch 95/300
```

```
Epoch 96/300
Epoch 97/300
Epoch 98/300
Epoch 99/300
Epoch 100/300
Epoch 101/300
Epoch 102/300
Epoch 103/300
Epoch 104/300
Epoch 105/300
Epoch 106/300
Epoch 107/300
Epoch 108/300
Epoch 109/300
Epoch 110/300
Epoch 111/300
Epoch 112/300
Epoch 113/300
Epoch 114/300
Epoch 115/300
Epoch 116/300
Epoch 117/300
Epoch 118/300
Epoch 119/300
Epoch 120/300
Epoch 121/300
Epoch 122/300
Epoch 123/300
Epoch 124/300
Epoch 125/300
Epoch 126/300
Epoch 127/300
Epoch 128/300
```

```
Epoch 129/300
Epoch 130/300
Epoch 131/300
Epoch 132/300
Epoch 133/300
Epoch 134/300
Epoch 135/300
Epoch 136/300
Epoch 137/300
Epoch 138/300
Epoch 139/300
Epoch 140/300
Epoch 141/300
Epoch 142/300
Epoch 143/300
Epoch 144/300
Epoch 145/300
Epoch 146/300
Epoch 147/300
Epoch 148/300
Epoch 149/300
Epoch 150/300
Epoch 151/300
Epoch 152/300
Epoch 153/300
Epoch 154/300
Epoch 155/300
Epoch 156/300
Epoch 157/300
Epoch 158/300
Epoch 159/300
Epoch 160/300
Epoch 161/300
```

```
Epoch 162/300
Epoch 163/300
Epoch 164/300
Epoch 165/300
Epoch 166/300
Epoch 167/300
Epoch 168/300
Epoch 169/300
Epoch 170/300
Epoch 171/300
Epoch 172/300
Epoch 173/300
Epoch 174/300
Epoch 175/300
Epoch 176/300
Epoch 177/300
Epoch 178/300
Epoch 179/300
Epoch 180/300
Epoch 181/300
Epoch 182/300
Epoch 183/300
Epoch 184/300
Epoch 185/300
Epoch 186/300
Epoch 187/300
Epoch 188/300
Epoch 189/300
Epoch 190/300
Epoch 191/300
Epoch 192/300
Epoch 193/300
Epoch 194/300
```

```
Epoch 195/300
Epoch 196/300
Epoch 197/300
Epoch 198/300
Epoch 199/300
Epoch 200/300
Epoch 201/300
Epoch 202/300
Epoch 203/300
Epoch 204/300
Epoch 205/300
Epoch 206/300
Epoch 207/300
Epoch 208/300
Epoch 209/300
Epoch 210/300
Epoch 211/300
Epoch 212/300
Epoch 213/300
Epoch 214/300
Epoch 215/300
Epoch 216/300
Epoch 217/300
Epoch 218/300
Epoch 219/300
Epoch 220/300
Epoch 221/300
Epoch 222/300
Epoch 223/300
Epoch 224/300
Epoch 225/300
Epoch 226/300
Epoch 227/300
```

```
Epoch 228/300
Epoch 229/300
Epoch 230/300
Epoch 231/300
Epoch 232/300
Epoch 233/300
Epoch 234/300
Epoch 235/300
Epoch 236/300
Epoch 237/300
Epoch 238/300
Epoch 239/300
Epoch 240/300
Epoch 241/300
Epoch 242/300
Epoch 243/300
Epoch 244/300
Epoch 245/300
Epoch 246/300
Epoch 247/300
Epoch 248/300
Epoch 249/300
Epoch 250/300
Epoch 251/300
Epoch 252/300
Epoch 253/300
Epoch 254/300
Epoch 255/300
Epoch 256/300
Epoch 257/300
Epoch 258/300
Epoch 259/300
Epoch 260/300
```

```
Epoch 261/300
Epoch 262/300
Epoch 263/300
Epoch 264/300
Epoch 265/300
Epoch 266/300
Epoch 267/300
Epoch 268/300
Epoch 269/300
Epoch 270/300
Epoch 271/300
Epoch 272/300
Epoch 273/300
Epoch 274/300
Epoch 275/300
Epoch 276/300
Epoch 277/300
Epoch 278/300
Epoch 279/300
Epoch 280/300
Epoch 281/300
Epoch 282/300
Epoch 283/300
Epoch 284/300
Epoch 285/300
Epoch 286/300
Epoch 287/300
Epoch 288/300
Epoch 289/300
Epoch 290/300
Epoch 291/300
Epoch 292/300
Epoch 293/300
```

```
Epoch 294/300
   Epoch 295/300
   Epoch 296/300
   Epoch 297/300
   Epoch 298/300
   Epoch 299/300
   Epoch 300/300
   Cross-validation mean accuracy: 0.7264840182648402, std: 0.014611872146118704
In [64]: from sklearn.model_selection import cross_val_score
   # Perform 5-fold cross validation
   scores = cross val score(pipe, X resampled, y resampled, cv=5)
   # Print the mean cross-validation score
   print("Cross-Validation Mean Score:", scores.mean())
   Cross-Validation Mean Score: 0.7680365296803653
```

## Model That Performs Best (as of December 1st, 2023)

Model stacking with the following base models perform best:

- Linear SVM W/ C=10
- Adaptive Boosting (AdaBoost) W/ default parameter values
- **K-Nearest Neighbors** W/ K=1 and Manhattan Distance (L1 Norm)
- Random Forest W/ 10 Trees
- (Extreme) Gradient Boosting W/ A maximum tree depth of 7, learning rate of 0.2, and subsample of 0.5

and the final estimator was a **Linear SVM** with the same parameters as before.

For computational efficiency, RandomizedSearchCV was used as opposed to GridSearchCV, however a lot of hyperparameter tuning came from trial and error as well.

#### **Accuracies**

We get a model accuracy of ~73.164% and a balanced accuracy of ~78.059%

After performing **5-fold** cross-validation, we get a mean CV score of ~77.2603%

Results:

Kaggle Public Score -> **75.228%**\ Kaggle Private Score -> **79.611%** -> **4th Highest Score!** 

```
In [12]: from imblearn.under_sampling import RandomUnderSampler
from sklearn.ensemble import StackingClassifier, AdaBoostClassifier, GradientBoostingCla
```

```
from sklearn.linear model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from imblearn.over sampling import ADASYN
from imblearn.over sampling import SMOTE
from xgboost import XGBClassifier
from sklearn.decomposition import PCA
from sklearn.preprocessing import MinMaxScaler
import time
# define base models
base models = [
   ("svc", SVC(kernel='linear', C=10.0, probability=True, random state=2)),
    ("ada", AdaBoostClassifier(n estimators=50, learning rate=1.0, random state=2)),
    ("knn", KNeighborsClassifier(n neighbors=1, weights='uniform', algorithm='auto', lea
    ("rf", RandomForestClassifier(random state=1, n estimators=10, bootstrap=True)),
    ("xgb", XGBClassifier(random state=2, max depth=7, learning rate=0.2, n estimators=1
# Start timer
start time = time.time()
# create pipeline
pipe = Pipeline(
    [("scaler", StandardScaler()),
     ("model", StackingClassifier(estimators=base models, final estimator=SVC(kernel='li
# undersample majority class
rus = RandomUnderSampler(random state=0)
X resampled, y resampled = rus.fit resample(
    train data[["fw similarity", "fw distance", "compartment synapse prop", "axon dendri
    "compartment apical",
    "compartment apical shaft",
    "compartment apical tuft",
    "compartment axon",
    "compartment basal"
    "compartment oblique",
    "compartment soma",
    "pre apical count",
    "pre apical shaft count",
    "pre apical tuft count",
    "pre axon count",
    "pre basal count",
    "pre oblique count",
    "pre soma count",
    "post apical count",
    "post apical shaft count",
    "post apical tuft count",
    "post axon count",
    "post basal count",
    "post oblique count",
    "post soma count"]],
    train data["connected"]
# fit model
pipe.fit(X resampled, y resampled)
# predict on test data
test data["pred"] = pipe.predict proba(test data[["fw similarity", "fw distance", "compa
    "compartment apical",
    "compartment apical shaft",
    "compartment apical tuft",
    "compartment axon",
    "compartment basal"
```

```
"compartment soma",
             "pre apical count",
             "pre apical shaft count",
             "pre apical tuft count",
            "pre axon count",
            "pre basal count",
             "pre oblique count",
            "pre soma count",
            "post apical count",
             "post apical shaft count",
            "post apical tuft count",
            "post axon count",
             "post basal count",
             "post oblique count",
             "post soma count"]])[:, 1]
         # compute accuracy
        print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
         # confusion matrix
        print(confusion matrix(test data['connected'], test data['pred'] > .5))
         # compute balanced accuracy
        print(
            f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
         # Stop timer and print time
         end time = time.time()
        print(f"Time taken: {end time - start time} seconds")
        accuracy: 0.7316436623886782
        [[26968 9928]
         [ 46 225]]
        balanced accuracy: 0.7805888217164912
        Time taken: 119.38371515274048 seconds
In [23]: from sklearn.model selection import RandomizedSearchCV
         from scipy.stats import uniform, randint
         # define the parameter distribution
         param dist = {
            'model svc C': uniform(0.1, 10),
             'model ada n estimators': randint(30, 70),
             'model ada learning rate': uniform(0.5, 1.5),
             'model knn n neighbors': randint(1, 5),
             'model rf n estimators': randint(10, 100),
             'model rf max depth': [None, 5, 10],
             'model xgb max depth': randint(5, 9),
             'model xgb learning rate': uniform(0.1, 0.3),
             'model xgb n estimators': randint(50, 150),
             'model xgb subsample': uniform(0.3, 0.7)
         # create a RandomizedSearchCV object
         random search = RandomizedSearchCV(estimator=pipe, param distributions=param dist, n ite
         # fit the model
         random search.fit(X resampled, y resampled)
         # print the best parameters
        best params = random search.best params
         for param, value in best params.items():
            print(f"{param}: {value}")
```

"compartment oblique",

```
Fitting 5 folds for each of 10 candidates, totalling 50 fits
[CV] END model ada learning rate=1.1539923532130056, model ada n estimators=38, mode
1 knn n neighbors=3, model rf max depth=10, model rf n estimators=85, model svc
C=4.30367802087489, model__xgb__learning_rate=0.19910044630116225, model__xgb__max_depth
=8, model xgb n estimators=125, model xgb subsample=0.7334896764454646; total time=
52.4s
[CV] END model ada learning rate=1.1539923532130056, model ada n estimators=38, mode
  knn n neighbors=3, model rf max depth=10, model rf n estimators=85, model svc
C=4.30367802087489, model xgb learning rate=0.19910044630116225, model xgb max depth
=8, model xgb n estimators=125, model xgb subsample=0.7334896764454646; total time=
54.3s
[CV] END model ada learning rate=1.1539923532130056, model ada n estimators=38, mode
1 knn n neighbors=3, model rf max depth=10, model rf n estimators=85, model svc
C=4.30367802087489, model xgb learning rate=0.19910044630116225, model xgb max depth
=8, model__xgb__n_estimators=125, model__xgb__subsample=0.7334896764454646; total time=
49.9s
[CV] END model ada learning rate=1.1539923532130056, model ada n estimators=38, mode
1 knn n neighbors=3, model rf max depth=10, model rf n estimators=85, model svc
C=4.30367802087489, model xgb learning rate=0.19910044630116225, model xgb max depth
=8, model xgb n estimators=125, model xgb subsample=0.7334896764454646; total time=
[CV] END model ada learning rate=1.1539923532130056, model ada n estimators=38, mode
  knn n neighbors=3, model rf max depth=10, model rf n estimators=85, model svc
C=4.30367802087489, model xgb learning rate=0.19910044630116225, model xgb max depth
=8, model xgb n estimators=125, model xgb subsample=0.7334896764454646; total time=
51.6s
[CV] END model ada learning rate=0.9494820105117847, model ada n estimators=56, mode
1 knn n neighbors=1, model rf max depth=None, model rf n estimators=47, model svc
C=6.930259944094713, model xgb learning rate=0.2495683510159957, model xgb max dep
th=7, model xgb n estimators=101, model xgb subsample=0.8497346034716713; total time
= 1.2min
[CV] END model ada learning rate=0.9494820105117847, model ada n estimators=56, mode
 knn n neighbors=1, model rf max depth=None, model rf n estimators=47, model svc
 C=6.930259944094713, model xgb learning rate=0.2495683510159957, model xgb max dep
th=7, model xgb n estimators=101, model xgb subsample=0.8497346034716713; total time
= 1.2min
[CV] END model__ada__learning_rate=0.9494820105117847, model__ada__n_estimators=56, mode
  _knn__n_neighbors=1, model__rf__max_depth=None, model__rf__n_estimators=47, model__svc
 C=6.930259944094713, model xgb learning rate=0.2495683510159957, model xgb max dep
th=7, model xgb n estimators=101, model xgb subsample=0.8497346034716713; total time
= 1.1 min
[CV] END model ada learning rate=0.9494820105117847, model__ada__n_estimators=56, mode
1 knn n neighbors=1, model rf max depth=None, model rf n estimators=47, model svc
 C=6.930259944094713, model xgb learning rate=0.2495683510159957, model xgb max dep
th=7, model xgb n estimators=101, model xgb subsample=0.8497346034716713; total time
= 1.2min
[CV] END model ada learning rate=0.9494820105117847, model ada n estimators=56, mode
 _knn__n_neighbors=1, model__rf__max_depth=None, model__rf__n_estimators=47, model svc
 C=6.930259944094713, model xgb learning rate=0.2495683510159957, model xgb max dep
th=7, model xgb n estimators=101, model xgb subsample=0.8497346034716713; total time
[CV] END model__ada__learning_rate=1.7809629389592332, model__ada__n_estimators=33, mode
  C=3.070183571076878, model xgb learning rate=0.1863606447610013, model xgb max dep
th=6, model xgb n estimators=81, model xgb subsample=0.42720892644682507; total time
[CV] END model ada learning rate=1.7809629389592332, model ada n estimators=33, mode
1 knn n neighbors=2, model rf max depth=None, model rf n estimators=78, model svc
 C=3.070183571076878, model xgb learning rate=0.1863606447610013, model xgb max dep
th=6, model xgb n estimators=81, model xgb subsample=0.42720892644682507; total time
= 50.2s
[CV] END model ada learning rate=1.7809629389592332, model ada n estimators=33, mode
  _knn__n_neighbors=2, model__rf__max_depth=None, model__rf__n_estimators=78, model svc
 _C=3.070183571076878, model__xgb__learning_rate=0.1863606447610013, model xgb max dep
th=6, model xgb n estimators=81, model xgb subsample=0.42720892644682507; total time
= 43.7s
```

```
[CV] END model__ada__learning_rate=1.7809629389592332, model__ada__n_estimators=33, mode
1 knn n neighbors=2, model rf max depth=None, model rf n estimators=78, model svc
 C=3.070183571076878, model xgb learning rate=0.1863606447610013, model xgb max dep
th=6, model xgb n estimators=81, model xgb subsample=0.42720892644682507; total time
= 42.3s
[CV] END model ada learning rate=1.7809629389592332, model ada n estimators=33, mode
1 knn n neighbors=2, model rf max depth=None, model rf n estimators=78, model svc
 C=3.070183571076878, model xgb learning rate=0.1863606447610013, model xgb max dep
th=6, model__xgb__n_estimators=81, model__xgb__subsample=0.42720892644682507; total time
= 47.6s
[CV] END model ada learning rate=1.2414346515768613, model ada n estimators=42, mode
  _knn__n_neighbors=3, model__rf__max_depth=None, model__rf__n_estimators=73, model__svc
 C=2.303062070705597, model xgb learning rate=0.2049478855009898, model xgb max dep
th=7, model xgb n estimators=58, model xgb subsample=0.44122025838547574; total time
= 42.1s
[CV] END model ada learning rate=1.2414346515768613, model ada n estimators=42, mode
1 knn n neighbors=3, model rf max depth=None, model rf n estimators=73, model svc
 C=2.303062070705597, model xgb learning rate=0.2049478855009898, model xgb max dep
th=7, model xgb n estimators=58, model xgb subsample=0.44122025838547574; total time
= 39.8s
[CV] END model ada learning rate=1.2414346515768613, model ada n estimators=42, mode
 knn n neighbors=3, model rf max depth=None, model rf n estimators=73, model svc
 C=2.303062070705597, model xgb learning rate=0.2049478855009898, model xgb max dep
th=7, model xgb n estimators=58, model xgb subsample=0.44122025838547574; total time
[CV] END model ada learning rate=1.2414346515768613, model ada n estimators=42, mode
  knn n neighbors=3, model rf max depth=None, model rf n estimators=73, model svc
 C=2.303062070705597, model xgb learning rate=0.2049478855009898, model xgb max dep
th=7, model xgb n estimators=58, model xgb subsample=0.44122025838547574; total time
= 37.9s
[CV] END model ada learning rate=1.2414346515768613, model__ada__n_estimators=42, mode
1 knn n neighbors=3, model rf max depth=None, model rf n estimators=73, model svc
 C=2.303062070705597, model xgb _learning_rate=0.2049478855009898, model__xgb__max_dep
th=7, model xgb n estimators=58, model xgb subsample=0.44122025838547574; total time
= 37.3s
[CV] END model ada learning rate=1.4606100878223722, model ada n estimators=38, mode
  knn n neighbors=4, model rf max depth=5, model rf n estimators=32, model svc C
=8.036374544415771, model xgb learning rate=0.274001253663342, model xgb max depth=
6, model xgb n estimators=140, model xgb subsample=0.6566681135446268; total time=
1.2min
[CV] END model ada learning rate=1.4606100878223722, model ada n estimators=38, mode
=8.036374544415771, model xgb learning rate=0.274001253663342, model xgb max depth=
6, model xgb n estimators=140, model xgb subsample=0.6566681135446268; total time=
1.3min
[CV] END model ada learning rate=1.4606100878223722, model ada n estimators=38, mode
1 knn n neighbors=4, model rf max depth=5, model rf n estimators=32, model svc C
=8.036374544415771, model__xgb__learning_rate=0.274001253663342, model__xgb__max_depth=
6, model xgb n estimators=140, model xgb subsample=0.6566681135446268; total time=
1.2min
[CV] END model ada learning rate=1.4606100878223722, model ada n estimators=38, mode
  knn n neighbors=4, model rf max depth=5, model rf n estimators=32, model svc C
=8.036374544415771, model xgb learning rate=0.274001253663342, model xgb max depth=
6, model xgb n estimators=140, model xgb subsample=0.6566681135446268; total time=
1.2min
[CV] END model ada learning rate=1.4606100878223722, model ada n estimators=38, mode
=8.036374544415771, model xgb learning rate=0.274001253663342, model xgb max depth=
6, model_xgb_n_estimators=140, model_xgb_subsample=0.6566681135446268; total time=
1.2min
[CV] END model ada learning rate=0.8259048896601802, model ada n estimators=62, mode
  knn n neighbors=3, model rf max depth=10, model rf n estimators=18, model svc
C=2.6873317142982747, model xgb learning rate=0.34275258579123147, model xgb max dep
th=7, model xgb n estimators=110, model xgb subsample=0.627489745430074; total time=
[CV] END model ada learning rate=0.8259048896601802, model ada n estimators=62, mode
```

```
knn n neighbors=3, model rf max depth=10, model rf n estimators=18, model svc
C=2.6873317142982747, model xgb learning rate=0.34275258579123147, model xgb max dep
th=7, model xgb n estimators=110, model xgb subsample=0.627489745430074; total time=
[CV] END model ada learning rate=0.8259048896601802, model ada n estimators=62, mode
1 knn n neighbors=3, model rf max depth=10, model rf n estimators=18, model svc
C=2.6873317142982747, model xgb learning rate=0.34275258579123147, model xgb max dep
th=7, model__xgb__n_estimators=110, model__xgb__subsample=0.627489745430074; total time=
  37.1s
[CV] END model ada learning rate=0.8259048896601802, model ada n estimators=62, mode
  knn n neighbors=3, model rf max depth=10, model rf n estimators=18, model svc
C=2.6873317142982747, model xgb learning rate=0.34275258579123147, model xgb max dep
th=7, model xgb n estimators=110, model xgb subsample=0.627489745430074; total time=
[CV] END model ada learning rate=0.8259048896601802, model ada n estimators=62, mode
  _knn__n_neighbors=3, model__rf__max_depth=10, model__rf__n_estimators=18, model__svc__
C=2.6873317142982747, model xgb learning rate=0.34275258579123147, model xgb max dep
th=7, model xgb n estimators=110, model xgb subsample=0.627489745430074; total time=
  35.4s
[CV] END model ada learning rate=0.9112598890856902, model__ada__n_estimators=52, mode
1 knn n neighbors=3, model rf max depth=None, model rf n estimators=92, model svc
 C=3.763424016750204, model__xgb__learning_rate=0.35525515120135054, model__xgb__max_de
pth=6, model xgb n estimators=120, model xgb subsample=0.3190416561263947; total tim
e = 52.4s
[CV] END model ada learning rate=0.9112598890856902, model ada n estimators=52, mode
  _knn__n_neighbors=3, model__rf__max_depth=None, model rf n estimators=92, model svc
 C=3.763424016750204, model xgb learning rate=0.35525515120135054, model xgb max de
pth=6, model xgb n estimators=120, model xgb subsample=0.3190416561263947; total tim
e = 55.2s
[CV] END model__ada__learning_rate=0.9112598890856902, model__ada__n_estimators=52, mode
l_knn__n_neighbors=3, model__rf__max_depth=None, model_ rf n estimators=92, model_ svc
 C=3.763424016750204, model xgb learning rate=0.35525515120135054, model xgb max de
pth=6, model xgb n estimators=120, model xgb subsample=0.3190416561263947; total tim
e = 55.8s
[CV] END model ada learning rate=0.9112598890856902, model ada n estimators=52, mode
1 knn n neighbors=3, model rf max depth=None, model rf n estimators=92, model svc
 C=3.763424016750204, model xgb learning rate=0.35525515120135054, model xgb max de
pth=6, model__xgb__n_estimators=120, model__xgb__subsample=0.3190416561263947; total tim
e = 52.8s
[CV] END model ada learning rate=0.9112598890856902, model ada n estimators=52, mode
  _knn__n_neighbors=3, model__rf__max_depth=None, model__rf__n_estimators=92, model svc
 _C=3.763424016750204, model__xgb__learning_rate=0.35525515120135054, model xgb max de
pth=6, model xgb n estimators=120, model xgb subsample=0.3190416561263947; total tim
e = 54.0s
[CV] END model ada learning rate=0.87076585849603, model ada n estimators=48, model
knn n neighbors=1, model rf max depth=10, model rf n estimators=97, model svc C=
9.805803133771734, model xgb learning rate=0.34007750533977604, model xgb max depth=
7, model_xgb_n_estimators=122, model_xgb_subsample=0.835471902316177; total time= 1.
6min
[CV] END model ada learning rate=0.87076585849603, model ada n estimators=48, model
knn n neighbors=1, model rf max depth=10, model rf n estimators=97, model svc C=
9.805803133771734, model xgb learning rate=0.34007750533977604, model xgb max depth=
7, model xgb n estimators=122, model xgb subsample=0.835471902316177; total time= 1.
[CV] END model ada learning rate=0.87076585849603, model__ada__n_estimators=48, model_
_knn__n_neighbors=1, model__rf__max_depth=10, model__rf__n_estimators=97, model__svc__C=
9.805803133771734, model xgb learning rate=0.34007750533977604, model xgb max depth=
7, model xgb n estimators=122, model xgb subsample=0.835471902316177; total time= 1.
7min
[CV] END model ada learning rate=0.87076585849603, model ada n estimators=48, model
knn n neighbors=1, model rf max depth=10, model rf n estimators=97, model svc C=
9.805803133771734, model xgb learning rate=0.34007750533977604, model xgb max depth=
7, model xgb n estimators=122, model xgb subsample=0.835471902316177; total time= 1.
6min
[CV] END model ada learning rate=0.87076585849603, model ada n estimators=48, model
_knn__n_neighbors=1, model__rf__max_depth=10, model__rf__n_estimators=97, model__svc__C=
```

```
9.805803133771734, model xgb learning rate=0.34007750533977604, model xgb max depth=
7, model xgb n estimators=122, model xgb subsample=0.835471902316177; total time= 1.
7min
[CV] END model ada learning rate=0.7538381698762668, model ada n estimators=46, mode
1 knn n neighbors=2, model rf max depth=5, model rf n estimators=53, model svc C
=0.5567896524540005, model xgb learning rate=0.39494603360716385, model xgb max dept
h=7, model xgb n estimators=87, model xgb subsample=0.6528003075654067; total time=
19.3s
[CV] END model ada learning rate=0.7538381698762668, model ada n estimators=46, mode
1 knn n neighbors=2, model rf max depth=5, model rf n estimators=53, model svc C
=0.5567896524540005, model xgb learning_rate=0.39494603360716385, model__xgb__max_dept
h=7, model xgb n estimators=87, model xgb subsample=0.6528003075654067; total time=
20.6s
[CV] END model ada learning rate=0.7538381698762668, model ada n estimators=46, mode
  =0.5567896524540005, model xgb learning rate=0.39494603360716385, model xgb max dept
h=7, model xgb n estimators=87, model xgb subsample=0.6528003075654067; total time=
20.1s
[CV] END model ada learning rate=0.7538381698762668, model ada n estimators=46, mode
1 knn n neighbors=2, model rf max depth=5, model rf n estimators=53, model svc C
=0.5567896524540005, model xgb learning rate=0.39494603360716385, model xgb max dept
h=7, model xgb n estimators=87, model xgb subsample=0.6528003075654067; total time=
19.4s
[CV] END model ada learning rate=0.7538381698762668, model ada n estimators=46, mode
1 knn n neighbors=2, model rf max depth=5, model rf n estimators=53, model svc C
=0.5567896524540005, model__xgb__learning_rate=0.39494603360716385, model__xgb__max_dept
h=7, model xgb n estimators=87, model xgb subsample=0.6528003075654067; total time=
[CV] END model ada learning rate=0.9853119763020483, model ada n estimators=43, mode
  _knn__n_neighbors=3, model__rf__max_depth=None, model__rf__n_estimators=32, model svc
 _C=7.384984962121942, model__xgb__learning_rate=0.20672608866707592, model xgb max de
pth=8, model xgb n estimators=77, model xgb subsample=0.5534191877236918; total time
= 1.2min
[CV] END model ada learning rate=0.9853119763020483, model ada n estimators=43, mode
1 knn n neighbors=3, model rf max depth=None, model rf n estimators=32, model svc
 C=7.384984962121942, model xgb learning rate=0.20672608866707592, model xgb max de
pth=8, model xgb n estimators=77, model xgb subsample=0.5534191877236918; total time
= 1.1min
[CV] END model ada learning rate=0.9853119763020483, model ada n estimators=43, mode
l knn n neighbors=3, model rf max depth=None, model rf n estimators=32, model svc
C=7.384984962121942, model xgb learning rate=0.20672608866707592, model xgb max de
pth=8, model xgb n estimators=77, model xgb subsample=0.5534191877236918; total time
= 1.0min
[CV] END model ada learning rate=0.9853119763020483, model ada n estimators=43, mode
  _knn__n_neighbors=3, model__rf__max_depth=None, model__rf__n_estimators=32, model__svc
 C=7.384984962121942, model xgb learning rate=0.20672608866707592, model xgb max de
pth=8, model xgb n estimators=77, model xgb subsample=0.5534191877236918; total time
= 1.0 min
[CV] END model ada learning rate=0.9853119763020483, model ada n estimators=43, mode
1 knn n neighbors=3, model rf max depth=None, model rf n estimators=32, model svc
C=7.384984962121942, model xgb learning rate=0.20672608866707592, model xgb max de
pth=8, model xgb n estimators=77, model xgb subsample=0.5534191877236918; total time
= 1.1 min
model ada learning rate: 0.9853119763020483
model ada n estimators: 43
model knn n neighbors: 3
model rf max depth: None
model rf n estimators: 32
model svc C: 7.384984962121942
model xgb learning rate: 0.20672608866707592
model xgb max depth: 8
model xgb n estimators: 77
model xgb subsample: 0.5534191877236918
```

```
# Perform 5-fold cross validation
scores = cross_val_score(pipe, X_resampled, y_resampled, cv=5)

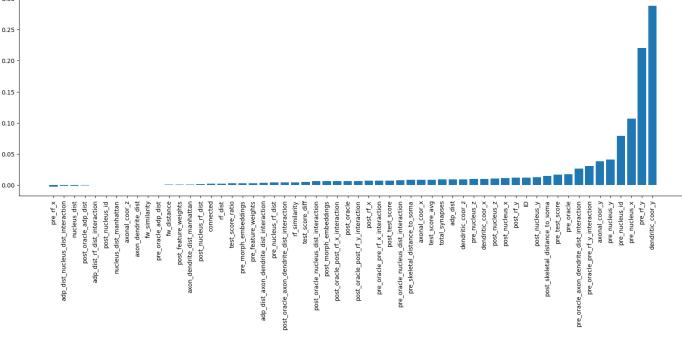
# Print the mean cross-validation score
print("Cross-Validation Mean Score:", scores.mean())
```

Cross-Validation Mean Score: 0.7726027397260273

### **Perform Permutation Importance**

Calculates mean importance of each feature by repeatedly shuffling the values of each feature and measuring how much the model's accuracy decreases

```
In [27]:
         from sklearn.inspection import permutation importance
         import matplotlib.pyplot as plt
         import numpy as np
         # perform permutation importance
         results = permutation importance(pipe, X resampled, y resampled, scoring='accuracy')
         # get importance
         importance = results.importances mean
         # get feature names
         feature names = train data.columns.tolist()
         # plot feature importance
In [28]:
         plt.figure(figsize=(20, 6)) # Adjust the width and height of the figure
         importance sorted, feature names sorted = zip(*sorted(zip(importance, feature names)))
         plt.bar(feature names sorted, importance sorted)
         plt.xticks(rotation=90) # Set rotation of the x-axis labels to 90 for vertical labels
         plt.show()
        0.30
```

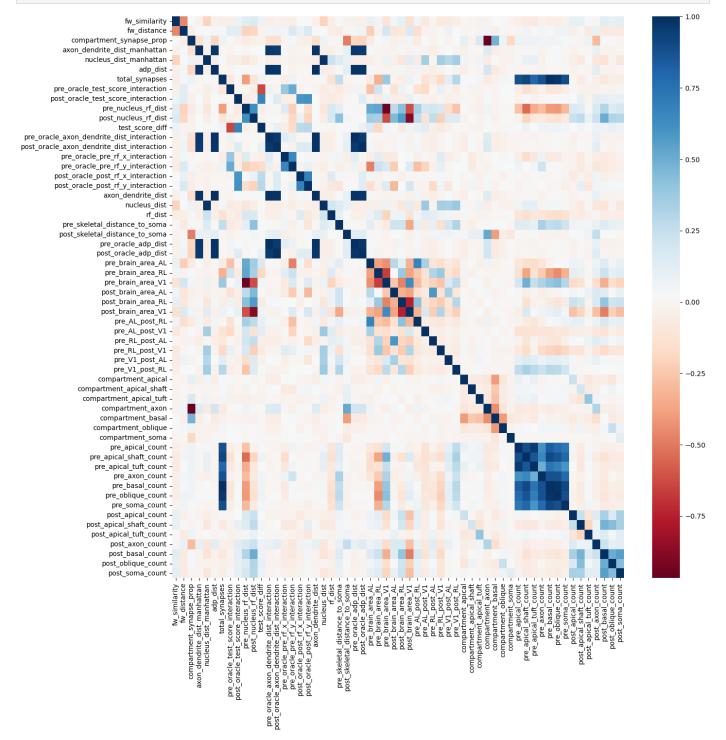


```
In [41]: import seaborn as sns
  import matplotlib.pyplot as plt

# calculate correlation matrix
  corr = X_resampled.corr()

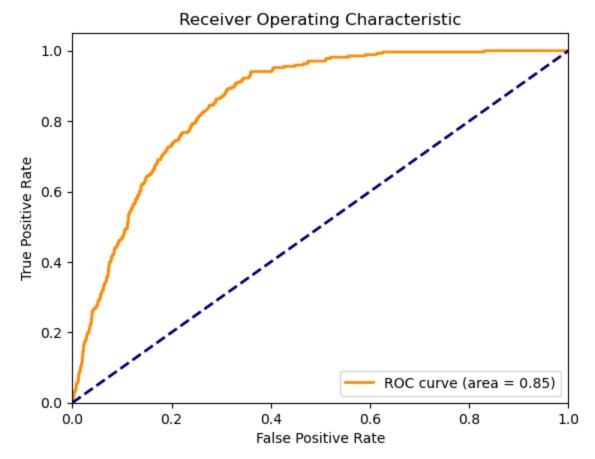
# create a larger figure
```

```
plt.figure(figsize=(15, 15))
4
# create a heatmap
sns.heatmap(corr, xticklabels=corr.columns, yticklabels=corr.columns, cmap='RdBu')
# rotate x-axis labels for better readability
plt.xticks(rotation=90)
# show the plot
plt.show()
```



```
In [30]: from sklearn.metrics import roc_curve, auc
    fpr, tpr, thresholds = roc_curve(test_data['connected'], test_data['pred'])
    roc_auc = auc(fpr, tpr)
    plt.figure()
    plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve (area = %0.2f)' % roc_auc)
    plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.05])
```





# A Potentially Better Model (After Competition Ended)

After the competition ended, using these base models with the following parameters yielded better accuracy scores than our best model (although not by a huge amount):

- Linear SVM W/ C=10
- Logistic Regression W/ default parameter values
- Random Forest W/ 10 Trees
- (Extreme) Gradient Boosting W/ A maximum tree depth of 7, learning rate of 0.2, and subsample of 0.5

and the final estimator was a **Linear SVM** with the same parameters as before.

Note: The model ran faster as well

#### **Accuracies**

We get a model accuracy of ~73.24% and a balanced accuracy of ~78.1%

After performing **5-fold** cross-validation, we get a mean CV score of ~**77.306**%

Kaggle Results Unavailable As The Competition Has Ended

```
from imblearn.under sampling import RandomUnderSampler
from sklearn.ensemble import StackingClassifier, AdaBoostClassifier, GradientBoostingCla
from sklearn.linear model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from imblearn.over sampling import ADASYN
from imblearn.over sampling import SMOTE
from xgboost import XGBClassifier
from sklearn.decomposition import PCA
from sklearn.preprocessing import MinMaxScaler
# define base models
base models = [
    ("svc", SVC(kernel='linear', C=1.0, probability=True, random state=2)), # C is the
    ("logreg", LogisticRegression(C=1.0, random_state=2)), # C is the inverse of regula
    ("rf", RandomForestClassifier(random state=1, n estimators=10, bootstrap=True)),
    ("xgb", XGBClassifier(random state=2, max depth=7, learning rate=0.2, n estimators=1
# create pipeline
pipe = Pipeline(
    [("scaler", StandardScaler()),
     ("model", StackingClassifier(estimators=base models, final estimator=SVC(kernel='li
# undersample majority class
rus = RandomUnderSampler(random state=0)
X resampled, y resampled = rus.fit resample(
    train data[["fw similarity", "fw distance", "compartment synapse prop", "axon dendri
    "compartment apical",
    "compartment apical shaft",
    "compartment apical tuft",
    "compartment axon",
    "compartment basal"
    "compartment oblique",
    "compartment soma",
    "pre apical count",
    "pre apical shaft count",
    "pre apical tuft count",
    "pre axon_count",
    "pre basal count",
    "pre oblique count",
    "pre soma count",
    "post apical count",
    "post apical shaft count",
    "post apical tuft count",
    "post axon count",
    "post basal count",
    "post oblique count",
    "post soma count"]],
    train data["connected"]
# fit model
pipe.fit(X resampled, y resampled)
# predict on test data
test data["pred"] = pipe.predict proba(test data[["fw similarity", "fw distance", "compa
    "compartment apical",
    "compartment apical shaft",
    "compartment apical tuft",
    "compartment axon",
    "compartment basal"
    "compartment oblique",
    "compartment soma",
```

```
"pre apical count",
    "pre apical shaft count",
    "pre apical tuft count",
    "pre axon count",
    "pre basal count",
    "pre oblique count",
    "pre soma count",
    "post apical count",
    "post apical shaft count",
    "post apical tuft count",
    "post axon count",
    "post basal count",
    "post oblique count",
    "post soma count"]])[:, 1]
# compute accuracy
print(f"accuracy: {accuracy score(test data['connected'], test data['pred'] > .5)}")
# confusion matrix
print(confusion matrix(test data['connected'], test data['pred'] > .5))
# compute balanced accuracy
print(
    f"balanced accuracy: {balanced accuracy score(test data['connected'], test data['pre
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

```
Increase the number of iterations (max iter) or scale the data as shown in:
            https://scikit-learn.org/stable/modules/preprocessing.html
        Please also refer to the documentation for alternative solver options:
            https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
          n iter i = check optimize result(
        C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
         genceWarning: lbfgs failed to converge (status=1):
        STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
         Increase the number of iterations (max iter) or scale the data as shown in:
            https://scikit-learn.org/stable/modules/preprocessing.html
        Please also refer to the documentation for alternative solver options:
            https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
          n iter i = check optimize result(
        accuracy: 0.7323701132725267
         [[26995 9901]
         [ 46 225]]
        balanced accuracy: 0.7809547150382605
In [64]: from sklearn.model selection import cross val score
         # Perform 5-fold cross validation
         scores = cross val score(pipe, X resampled, y resampled, cv=5)
         # Print the mean cross-validation score
         print("Cross-Validation Mean Score:", scores.mean())
        C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
        genceWarning: lbfgs failed to converge (status=1):
        STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
         Increase the number of iterations (max iter) or scale the data as shown in:
            https://scikit-learn.org/stable/modules/preprocessing.html
         Please also refer to the documentation for alternative solver options:
            https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
          n_iter_i = _check_optimize_result(
        C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
        genceWarning: lbfgs failed to converge (status=1):
        STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
         Increase the number of iterations (max iter) or scale the data as shown in:
            https://scikit-learn.org/stable/modules/preprocessing.html
        Please also refer to the documentation for alternative solver options:
            https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
          n iter i = check optimize result(
        C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
        genceWarning: lbfgs failed to converge (status=1):
        STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
         Increase the number of iterations (max iter) or scale the data as shown in:
            https://scikit-learn.org/stable/modules/preprocessing.html
        Please also refer to the documentation for alternative solver options:
            https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
          n iter i = check optimize result(
        C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
        genceWarning: lbfgs failed to converge (status=1):
        STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
         Increase the number of iterations (max iter) or scale the data as shown in:
            https://scikit-learn.org/stable/modules/preprocessing.html
        Please also refer to the documentation for alternative solver options:
            https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
          n iter i = check optimize result(
        C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
```

```
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
  n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
  n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
```

STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

```
Increase the number of iterations (max iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
  n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
 n iter i = check optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
  n_iter_i = _check_optimize result(
C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:460: Conver
genceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
Increase the number of iterations (max iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
```

```
https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
          n iter i = check optimize result(
        C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
        genceWarning: lbfgs failed to converge (status=1):
        STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
        Increase the number of iterations (max iter) or scale the data as shown in:
            https://scikit-learn.org/stable/modules/preprocessing.html
        Please also refer to the documentation for alternative solver options:
            https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
          n iter i = check optimize result(
        C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
        genceWarning: lbfgs failed to converge (status=1):
        STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
        Increase the number of iterations (max iter) or scale the data as shown in:
            https://scikit-learn.org/stable/modules/preprocessing.html
        Please also refer to the documentation for alternative solver options:
            https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
          n iter i = check optimize result(
        C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
        genceWarning: lbfgs failed to converge (status=1):
        STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
        Increase the number of iterations (max iter) or scale the data as shown in:
            https://scikit-learn.org/stable/modules/preprocessing.html
        Please also refer to the documentation for alternative solver options:
            https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
          n iter i = check optimize result(
        C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
        genceWarning: lbfgs failed to converge (status=1):
        STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
        Increase the number of iterations (max iter) or scale the data as shown in:
            https://scikit-learn.org/stable/modules/preprocessing.html
        Please also refer to the documentation for alternative solver options:
            https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
          n iter i = check optimize result(
        C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
        genceWarning: lbfgs failed to converge (status=1):
        STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
        Increase the number of iterations (max iter) or scale the data as shown in:
            https://scikit-learn.org/stable/modules/preprocessing.html
        Please also refer to the documentation for alternative solver options:
            https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
          n iter i = check optimize result(
        C:\Users\micha\anaconda3\lib\site-packages\sklearn\linear model\ logistic.py:460: Conver
        genceWarning: lbfgs failed to converge (status=1):
        STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
        Increase the number of iterations (max iter) or scale the data as shown in:
            https://scikit-learn.org/stable/modules/preprocessing.html
        Please also refer to the documentation for alternative solver options:
            https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
          n_iter_i = _check_optimize result(
        Cross-Validation Mean Score: 0.7730593607305936
In [31]: print("Main Dataframe Size:", data.shape)
        print("\nMain Dataframe:")
```

```
data.info()
Main Dataframe Size: (185832, 109)
```

Main Dataframe:

<class 'pandas.core.frame.DataFrame'>

```
Int64Index: 185832 entries, 0 to 185831
Columns: 109 entries, ID to synapse_ratio
dtypes: bool(2), float64(56), int64(20), object(5), uint8(26)
memory usage: 121.2+ MB
```

### Leaderboard Data

From here we have the code used to create the csv file used in Kaggle

```
#we need to first load and merge the leaderboard data to have the same format as the tra
In [63]:
         lb data = pd.read csv("./leaderboard data.csv")
        lb data = (
In [64]:
             lb data.merge(
                feature weights.rename(columns=lambda x: "pre " + x),
                how="left",
                validate="m:1",
                copy=False,
             )
             .merge(
                feature weights.rename(columns=lambda x: "post " + x),
                how="left",
                validate="m:1",
                copy=False,
             .merge(
                morph embeddings.rename(columns=lambda x: "pre " + x),
                how="left",
                validate="m:1",
                copy=False,
             )
             .merge(
                 morph embeddings.rename(columns=lambda x: "post " + x),
                how="left",
                validate="m:1",
                copy=False,
In [30]: print(lb data.columns)
         Index(['ID', 'axonal coor x', 'axonal coor y', 'axonal coor z',
                'dendritic coor x', 'dendritic coor y', 'dendritic coor z', 'adp dist',
                'post skeletal distance to soma', 'pre skeletal distance to soma',
                'pre oracle', 'pre test score', 'pre rf x', 'pre rf y', 'post oracle',
                'post test score', 'post rf x', 'post rf y', 'compartment',
                'pre brain area', 'post brain area', 'pre nucleus x', 'pre nucleus y',
                'pre nucleus z', 'post nucleus x', 'post nucleus y', 'post nucleus z',
                'pre nucleus id', 'post nucleus id', 'pre feature weights',
                'post_feature_weights', 'pre_morph embeddings',
                'post morph embeddings'],
              dtype='object')
In [65]: | # compute the cosine similarity between the pre- and post- feature weights
         lb data["fw similarity"] = lb data.apply(row feature similarity, axis=1)
         lb data["fw distance"] = lb data.apply(row feature distance, axis=1)
         # Create a copy of the 'compartment' column
```

```
lb data['compartment copy'] = lb data['compartment']
# Map the proportions to the original DataFrame
lb data['compartment synapse prop'] = lb data['compartment copy'].map(synapse proportion
# Compute Euclidean distance for leaderboard data
lb data['axon dendrite dist'] = np.sqrt(
    (lb_data['axonal_coor_x'] - lb_data['dendritic_coor_x'])**2 +
    (lb data['axonal coor y'] - lb data['dendritic coor y'])**2 +
    (lb data['axonal coor z'] - lb data['dendritic coor z'])**2
lb data['nucleus dist'] = np.sqrt(
    (lb data['pre nucleus x'] - lb data['post nucleus x'])**2 +
    (lb data['pre nucleus y'] - lb data['post nucleus y'])**2 +
    (lb data['pre nucleus z'] - lb data['post nucleus z'])**2
# Compute interaction between oracle and distance features for train data
lb data['pre oracle adp dist'] = lb data['pre oracle'] * lb data['adp dist']
lb data['post oracle adp dist'] = lb data['post oracle'] * lb data['adp dist']
# Compute Euclidean distance between receptive field locations for train data
lb data['rf dist'] = np.sqrt(
    (lb data['pre rf x'] - lb data['post rf x'])**2 +
    (lb_data['pre_rf_y'] - lb_data['post rf y'])**2
lb data = pd.get dummies(lb data, columns=['pre brain area', 'post brain area'])
# Perform one-hot encoding on the 'compartment' column
lb data = pd.get dummies(lb data, columns=['compartment'])
# Create interaction features
lb data['pre AL post RL'] = lb data['pre brain area AL'] * lb data['post brain area RL']
lb data['pre AL post V1'] = lb data['pre brain area AL'] * lb data['post brain area V1']
lb data['pre RL post AL'] = lb data['pre brain area RL'] * lb data['post brain area AL']
lb_data['pre_RL_post_V1'] = lb_data['pre_brain_area_RL'] * lb_data['post_brain_area_V1']
lb data['pre V1 post AL'] = lb data['pre brain area V1'] * lb data['post brain area AL']
lb data['pre V1 post RL'] = lb data['pre brain area V1'] * lb data['post brain area RL']
lb data['pre skeletal distance to soma']
lb data['pre skeletal distance to soma']
# Compute the Euclidean distance between the pre- and post- morphological embeddings
lb data["morph distance"] = lb data.apply(row morph distance, axis=1)
# Fill NaNs with the mean of the column
mean morph distance = lb data['morph distance'].mean()
lb data['morph distance'].fillna(mean morph distance, inplace=True)
# Drop rows with NaNs in the 'morph distance' column
lb data.dropna(subset=['morph distance'], inplace=True)
lb data['pre oracle pre rf x interaction'] = lb data['pre oracle'] * lb data['pre rf x']
lb data['pre oracle pre rf y interaction'] = lb data['pre oracle'] * lb data['pre rf y']
lb data['post oracle post rf x interaction'] = lb data['post oracle'] * lb data['post rf
lb data['post oracle post rf y interaction'] = lb data['post oracle'] * lb data['post rf
# Compute interaction between pre oracle and axon dendrite dist
lb data['pre oracle axon dendrite dist interaction'] = lb data['pre oracle'] * lb data['
lb data['post oracle axon dendrite dist interaction'] = lb data['post oracle'] * lb data
# Compute interaction between post oracle and nucleus dist
# 1b data['pre oracle nucleus dist interaction'] = 1b data['pre oracle'] * 1b data['nucl
# lb data['post oracle nucleus dist interaction'] = lb data['post oracle'] * lb data['nu
```

```
lb data['test score diff'] = lb data['post test score'] - lb data['pre test score']
# Compute Euclidean distance between nucleus and receptive field locations
lb data['pre nucleus rf dist'] = np.sqrt(
    (lb data['pre nucleus x'] - lb data['pre rf x'])**2 +
    (lb data['pre nucleus y'] - lb data['pre rf y'])**2
lb data['post nucleus rf dist'] = np.sqrt(
    (lb data['post nucleus x'] - lb data['post rf x'])**2 +
    (lb data['post nucleus y'] - lb data['post rf y'])**2
# Compute interaction between oracle and test score
lb data['pre oracle test score interaction'] = lb data['pre oracle'] * lb data['pre test
lb data['post oracle test score interaction'] = lb data['post oracle'] * lb data['post t
# Compute Manhattan distance
lb data['axon dendrite dist manhattan'] = (
    abs(lb_data['axonal_coor_x'] - lb_data['dendritic_coor_x']) +
    abs(lb data['axonal coor y'] - lb data['dendritic coor y']) +
    abs(lb data['axonal coor z'] - lb data['dendritic coor z'])
lb data['nucleus dist manhattan'] = (
    abs(lb data['pre nucleus x'] - lb data['post nucleus x']) +
    abs(lb data['pre nucleus y'] - lb data['post nucleus y']) +
    abs(lb data['pre nucleus z'] - lb data['post nucleus z'])
# Calculate the total number of synapses for each neuron
total synapses = lb data.groupby('pre nucleus id').size()
# Add the Total Synapses to your DataFrame
lb data['total synapses'] = lb data['pre nucleus id'].map(total synapses)
# List of all compartment types
compartments = ['compartment apical', 'compartment apical shaft', 'compartment apical tu
# Create a new feature that represents the total count of compartments for each neuron
lb data['total compartments'] = lb data[compartments].sum(axis=1)
# Create new features that represent the ratio of each compartment type to the total num
for compartment in compartments:
    lb data[f'{compartment} ratio'] = lb data[compartment] / lb data['total compartments
# List of all compartment types
compartments = ['apical', 'apical shaft', 'apical tuft', 'axon', 'basal', 'oblique', 'so
# Create new aggregated features for pre nucleus id
for compartment in compartments:
   # Create new feature name
   new feature = f'pre {compartment} count'
    # Create new feature
   lb data[new feature] = lb data.groupby('pre nucleus id')[f'compartment {compartment}
# Create new aggregated features for post nucleus id
for compartment in compartments:
   # Create new feature name
   new feature = f'post {compartment} count'
    # Create new feature
    lb data[new feature] = lb data.groupby('post nucleus id')[f'compartment {compartment
```

```
# predict on leaderboard data
lb_data["pred"] = pipe.predict_proba(lb_data[["fw_similarity", "fw_distance", "compartme
# create a boolean prediction solution
lb_data["connected"] = lb_data["pred"] > .5

In [67]: # columns should be ID, connected
submission_data = lb_data.filter(['ID','connected'])

In [68]: #writing csv files
submission_data.to_csv('submission_data_19_RF2.csv',index=False)

In []:
```