By: Muskan Gupta

SUMMER PROJECT DELIVERBALE

07/31/2025

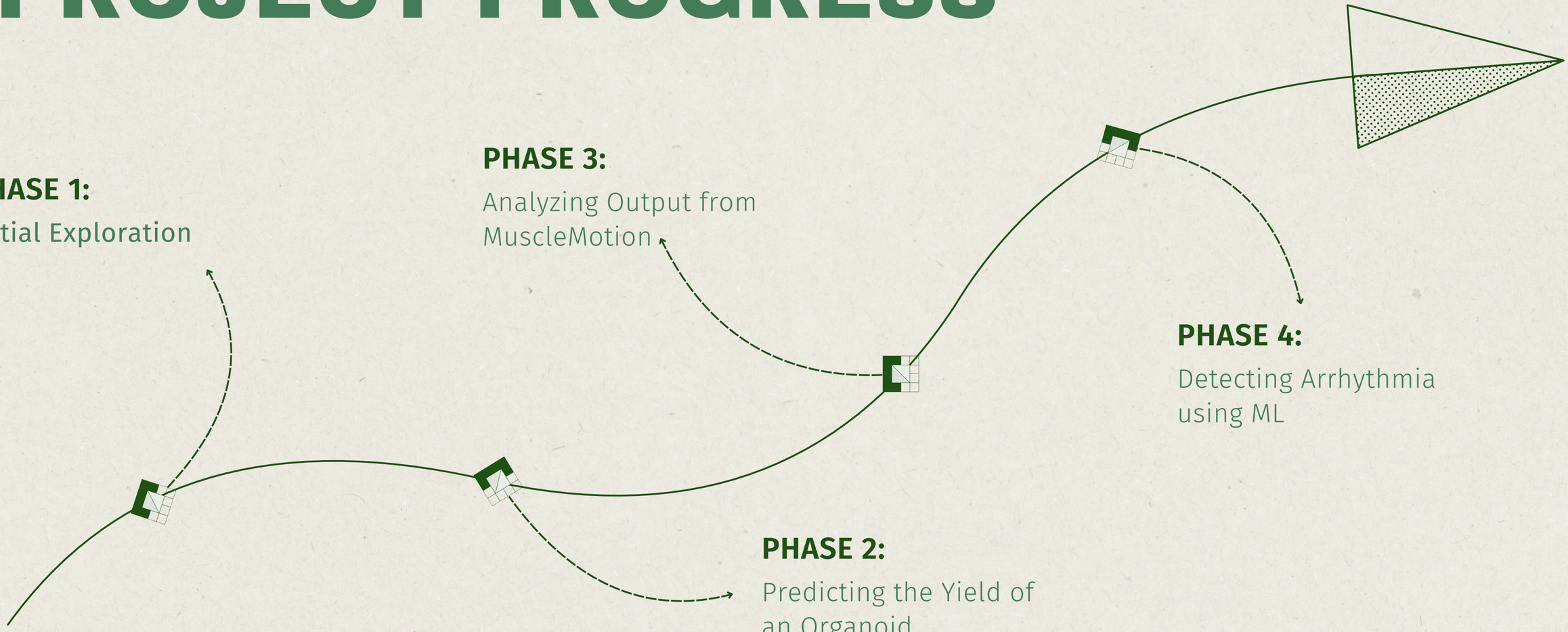
PROJECT PROGRESS

PHASE 1:
Initial Exploration

PHASE 2:
Predicting the Yield of an Organoid

PHASE 3:
Analyzing Output from MuscleMotion

PHASE 4:
Detecting Arrhythmia using ML



PHASE 1: Initial Exploration

MODELS

Supervised Models: Use labeled data in which each data instance has a known category or value to which it belongs

- Common examples include linear regression, decision trees, logistic regression, random forest, and boosting algorithms (including XGBoost)

Why Supervised Models?
Since we had labels available to train the data, we can use supervised models!
Unsupervised models can't label data, instead they primarily focus on finding patterns and clustering

MUSCLEMOTION

Algorithm

- Each reference pixel is subtracted from the corresponding pixel in the frame of interest
  - Unchanged (low) values represented with black
  - Changed (high) values represented with white

Explored

- Analyzed multiple MuscleMotion output files per well, including contraction.txt and speed-of-contraction.txt
  - Found that contraction.txt provided sufficient data to find and analyze peak-to-peak patterns

HEART

Key Concepts

- Used details of real human hearts
  - organoid can be modeled to mimic real heart behavior
- A heartbeat corresponds to a peak in contraction data
- Normal rhythms show evenly spaced, consistent peaks
- Determined what makes a heartbeat arrhythmic or not
  - Arrhythmic patterns include irregular spacing, missing peaks, or erratic fluctuations

PHASE 2: Predicting the Yield of an Organoid

PROCESS INPUT

- Loaded and processed the data (matrix.mtx, barcodes.tsv, features.tsv) using Scanpy, and converted them into an AnnData object
- Filtered out low-quality cells and genes. Performed normalization.
- Clustered cells using the Leiden algorithm and identified marker genes to characterize the cell types.

1

MERGE TOGETHER

- Generated features:
  - % of cells per cluster
  - top marker gene statistics
- Merged with metrics.csv from the cellranger_count file to get full features list

2

SPLIT TRAIN/TEST

- Split the combined dataset into training and test sets using train_test_split from sklearn.
  - model will be evaluated on unseen data

3

TRAIN MODEL

- Trained a regression model using XGBoost, a powerful gradient boosting framework.
- Features used included both biological and sequencing metrics.

4

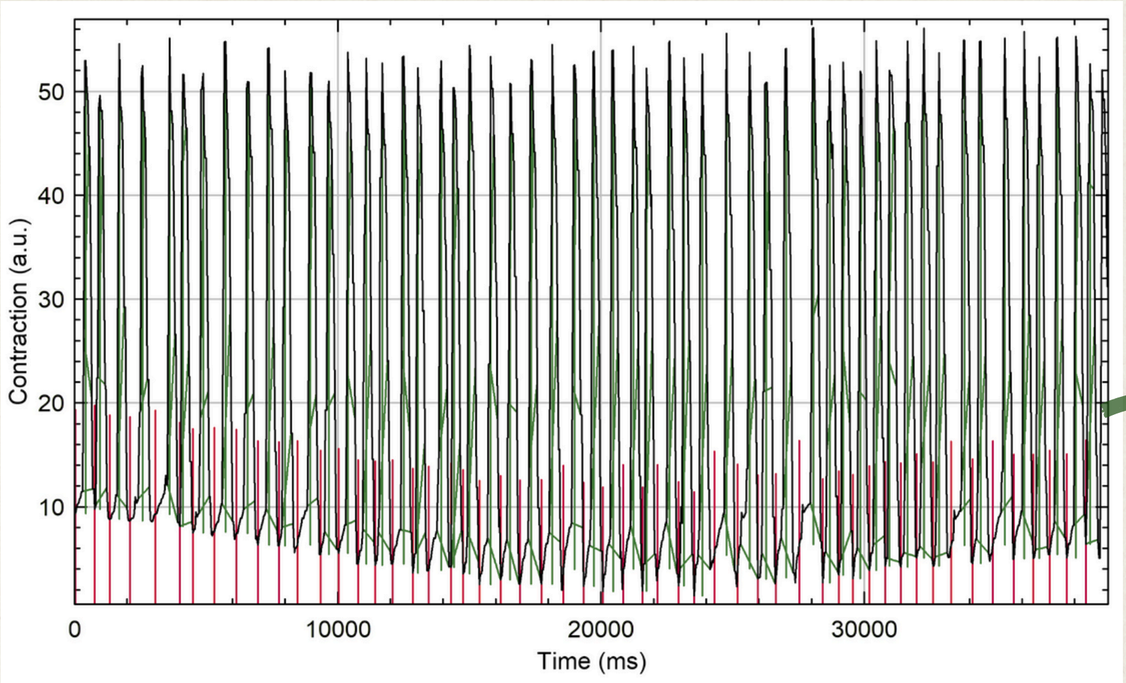
EVALUATE

- Root Mean Squared Error (RMSE) provides a measure of the average prediction error in cell count
  - Reduce RMSE by incorporating more training data

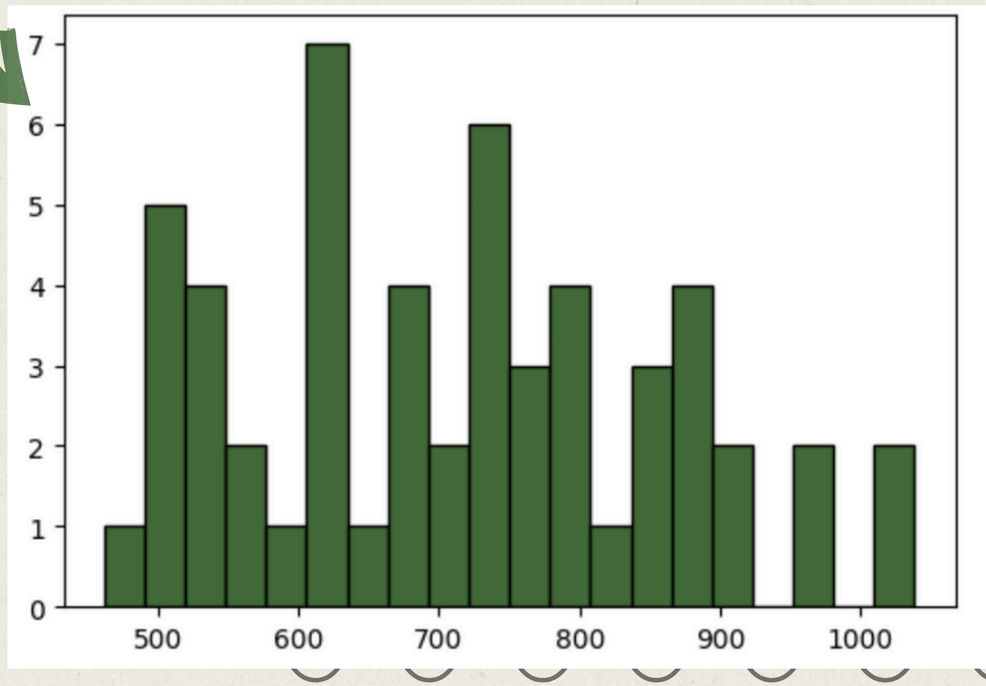
5

PHASE 3: Analyzing Output from MuscleMotion

Sub 1: Finding Peaks



Given the Contraction.txt file for a specific organoid, analyze where are the peaks and all peak-to-peak intervals



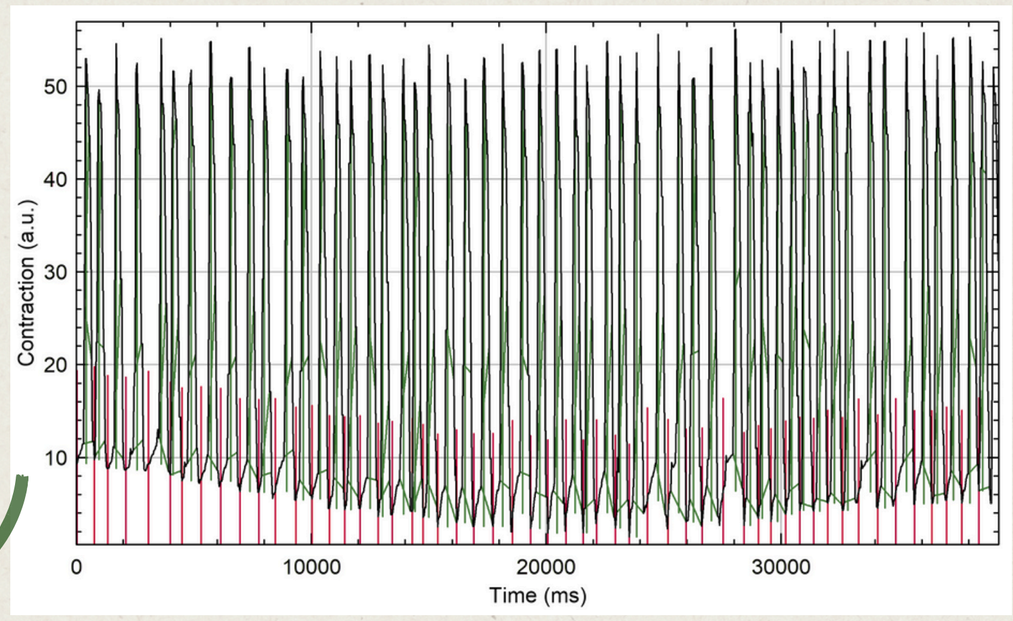
Print a graph of how many times each interval occurs

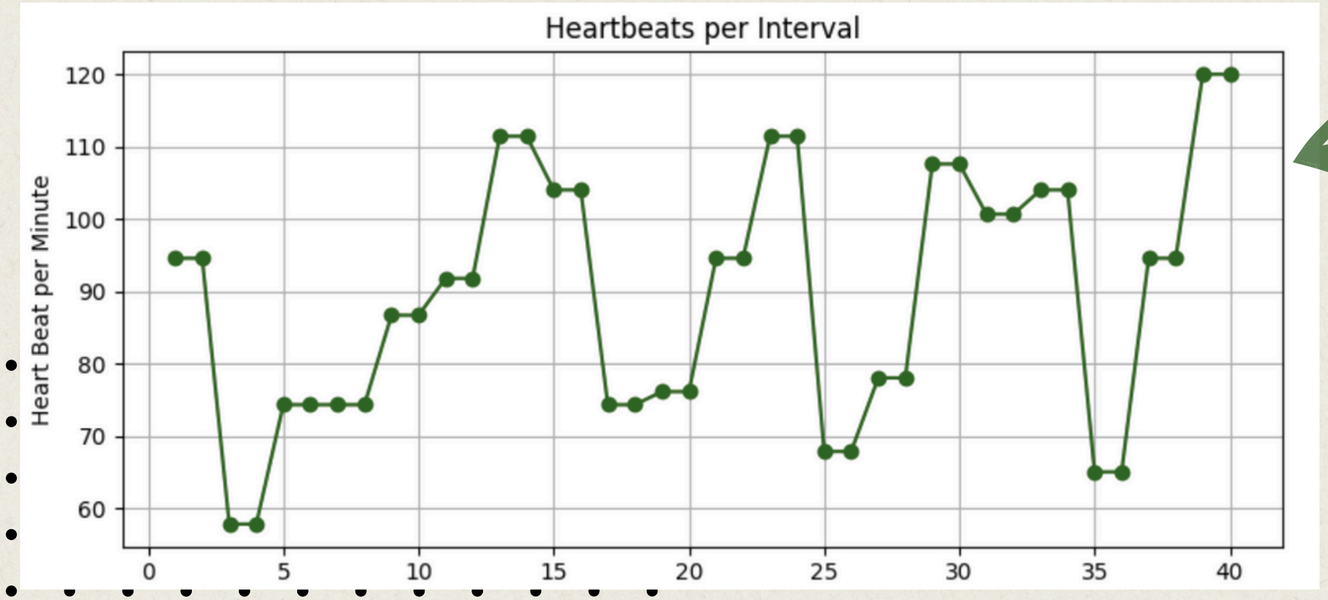
Average peak-to-peak interval: 715.099
Stand Dev: 148.503
Heart Beat per Minute: 4.615

PHASE 3: Analyzing Output from MuscleMotion

Sub 2: Finding the Change in Heart Beats per Minute

Given the Contraction.txt file for a specific organoid, divide the data into 2000 ms intervals to analyze the individual heart beat per minutes





Print a graph showing the heart beats per minute, representing how it changes over the course of the video

Time	Sample #	Type	Sub	Chan	Num
0:00.050	18	+	0	0	0
0:00.214	77	N	0	0	0
0:01.028	370	N	0	0	0
0:01.839	662	N	0	0	0
0:02.628	946	N	0	0	0
0:03.419	1231	N	0	0	0
0:04.208	1515	N	0	0	0
0:05.025	1809	N	0	0	0
0:05.678	2044	A	0	0	0
0:06.672	2402	N	0	0	0
0:07.517	2706	N	0	0	0
0:08.328	2998	N	0	0	0
0:09.117	3282	N	0	0	0
0:09.889	3560	N	0	0	0
0:10.728	3862	N	0	0	0
0:11.583	4170	N	0	0	0
0:12.406	4466	N	0	0	0
0:13.233	4764	N	0	0	0
0:14.056	5060	N	0	0	0
0:14.850	5346	N	0	0	0
0:15.647	5633	N	0	0	0
0:16.439	5918	N	0	0	0
0:17.261	6214	N	0	0	0

PHASE 4:
Detecting Arrhythmia Using ML
Sub 1: Finding a Dataset

MIT-BIH Arrhythmia Database

- ECG readings taken from 47 patients at a hospital in Boston between 1975 and 1979
  - 30 minute long readings, 48 total
- first publicly available annotated arrhythmia ECG database, accepted as a benchmark for arrhythmia detection since 1980
- sample # tells you the position of a data point in the ECG recording
  - since the ECG is recorded at a constant rate (360 samples per second), can use the sample # to figure out the exact time something happened in the signal

PHASE 4: Detecting Arrhythmia Using ML

Sub 2: Model Workflow

- Uploaded the dataset and selected key columns: time, sample #, and arrhythmic
- Performed initial cleanup by removing missing values and ensuring all data types were consistent for modeling

- Divided each patient's data into 5000 ms intervals and extracted heart rate variability (HRV) features from each segment
- Features served as the input variables (X) for the model
- Each interval's label as arrhythmic or not based on the dataset annotations served as the output (Y)

List of Features:

- Average time between consecutive peaks
- Standard deviation of intervals
- Min/Max: Extremes in intervals
- RMSSD: Root mean square of successive interval differences
- pNN50: Proportion of successive intervals that differ by more than 50 ms

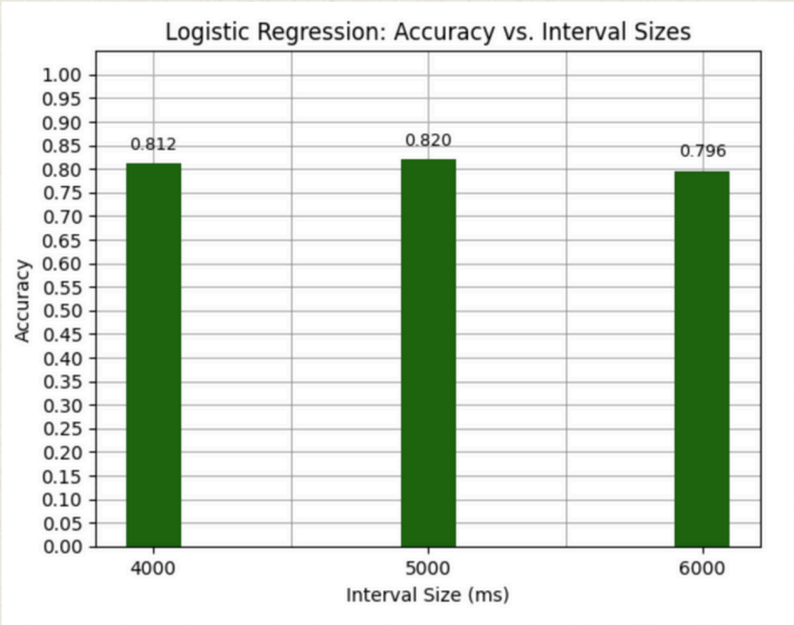
- Split the HRV features (X) and corresponding arrhythmia labels (Y) into training and testing sets (80/20 split)
- Trained an XGBoost classifier and evaluated its performance on the test set by calculating accuracy

- Used the trained model to analyze organoid contraction data MuscleMotion output was divided into 2500 ms intervals, and the same HRV features were extracted from each interval
- Each interval then classified as arrhythmic or normal using the model

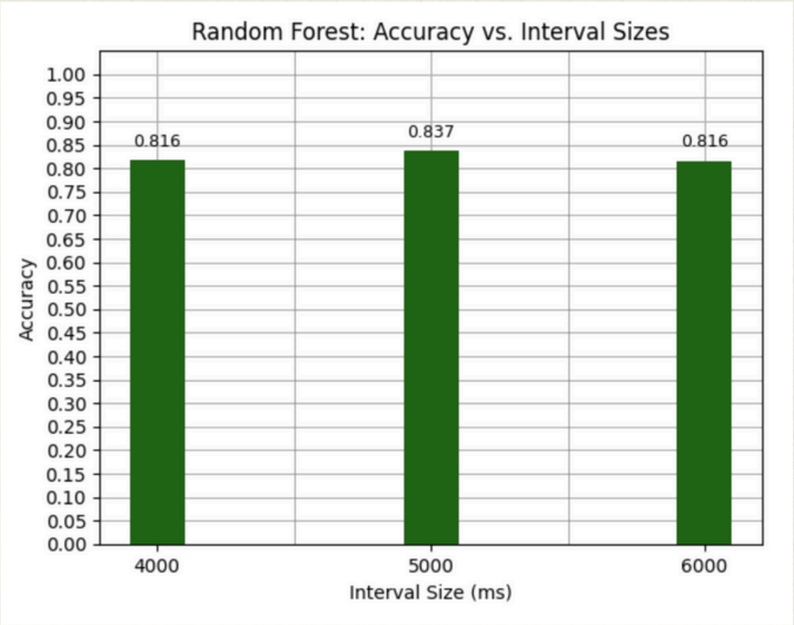
PHASE 4: Detecting Arrhythmia Using ML

Sub 3: Determining the Best Model

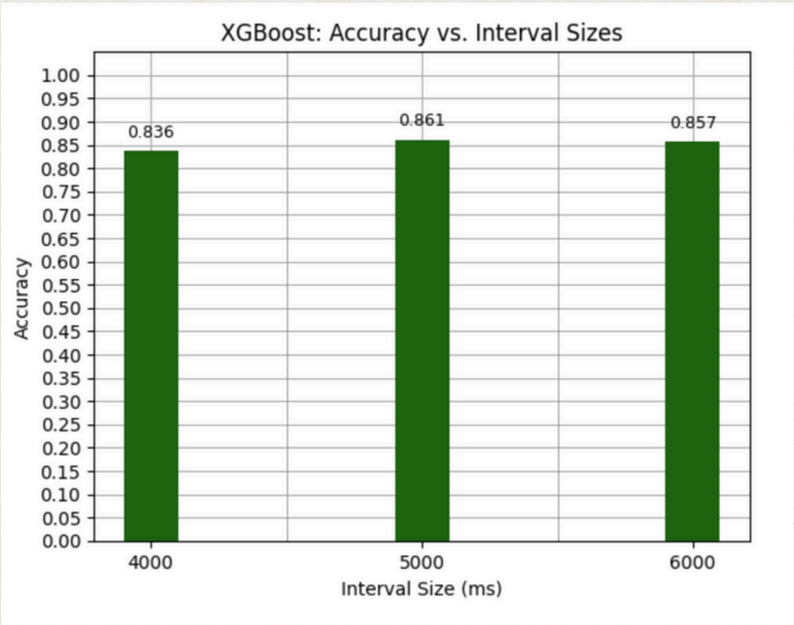
Logistic Regression



Random Forest



XGBoost

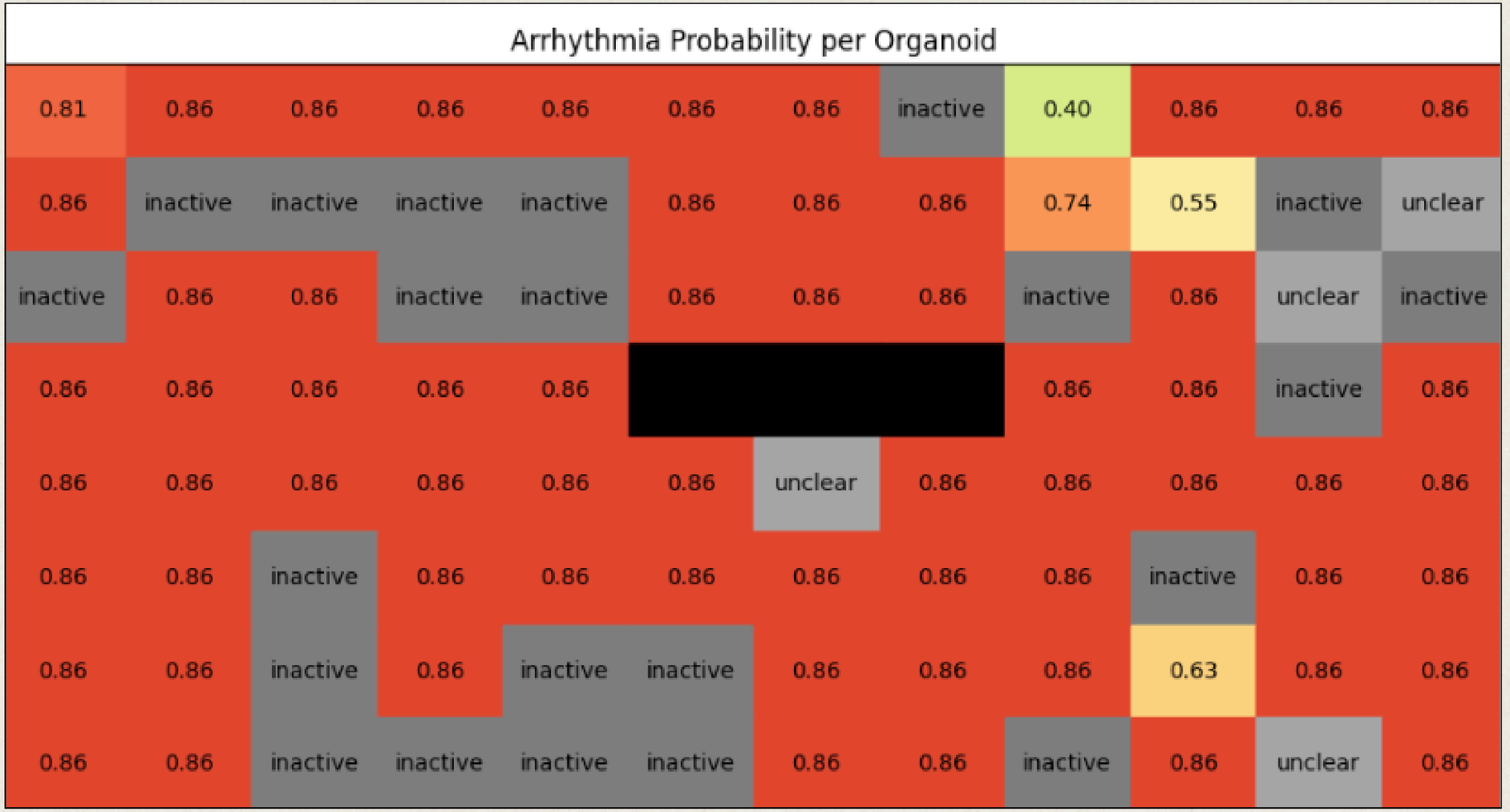


Highest Accuracy: 86.1%

XGBoost, 5000 ms intervals

PHASE 4: Detecting Arrhythmia Using ML

Sub 4: Implementing the Model



More Likely to Have Arrhythmia

50% Chance

More Likely to Not Have Arrhythmia

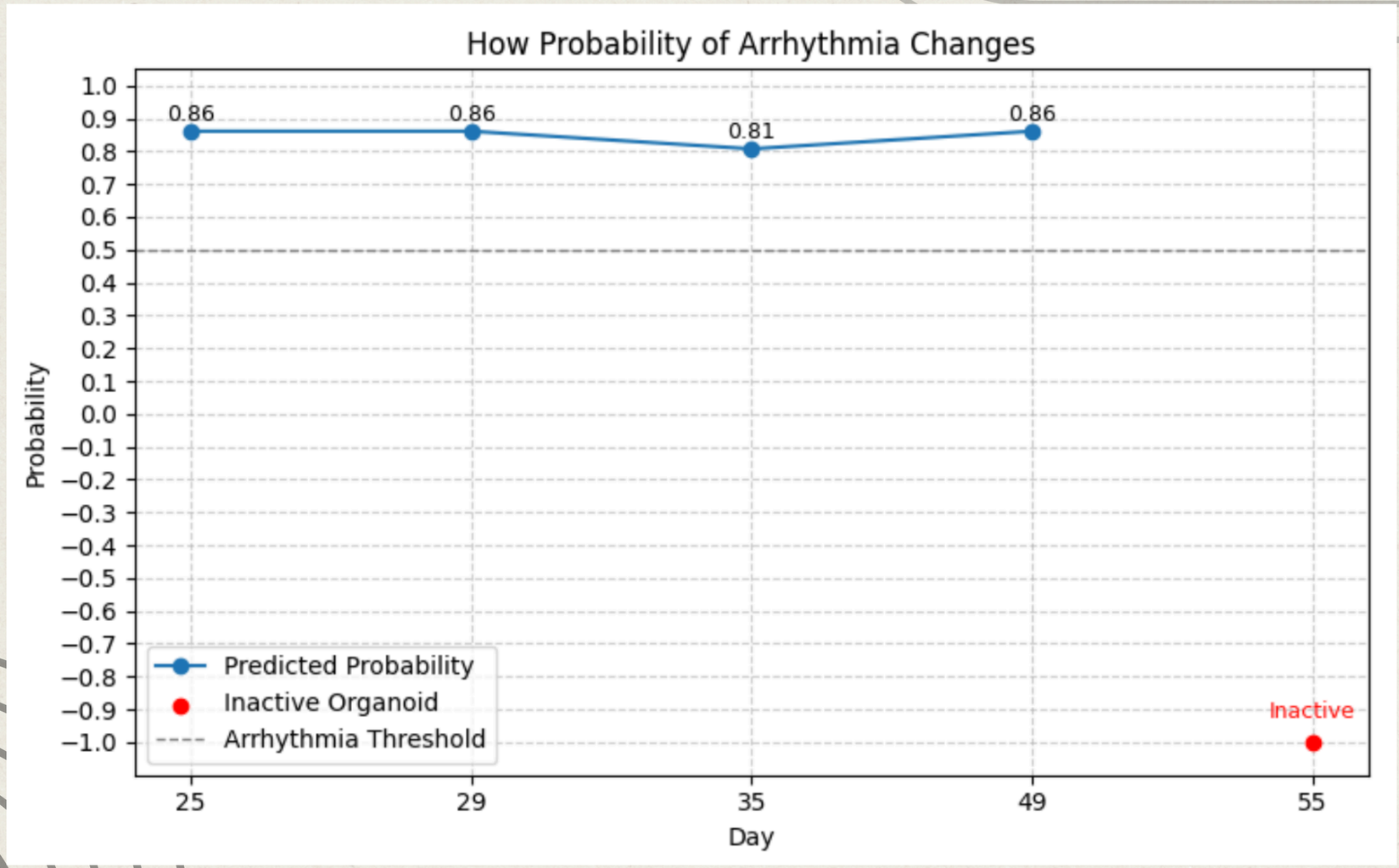
Inactive Organoid

Unclear Data

PHASE 4: Detecting Arrhythmia Using ML

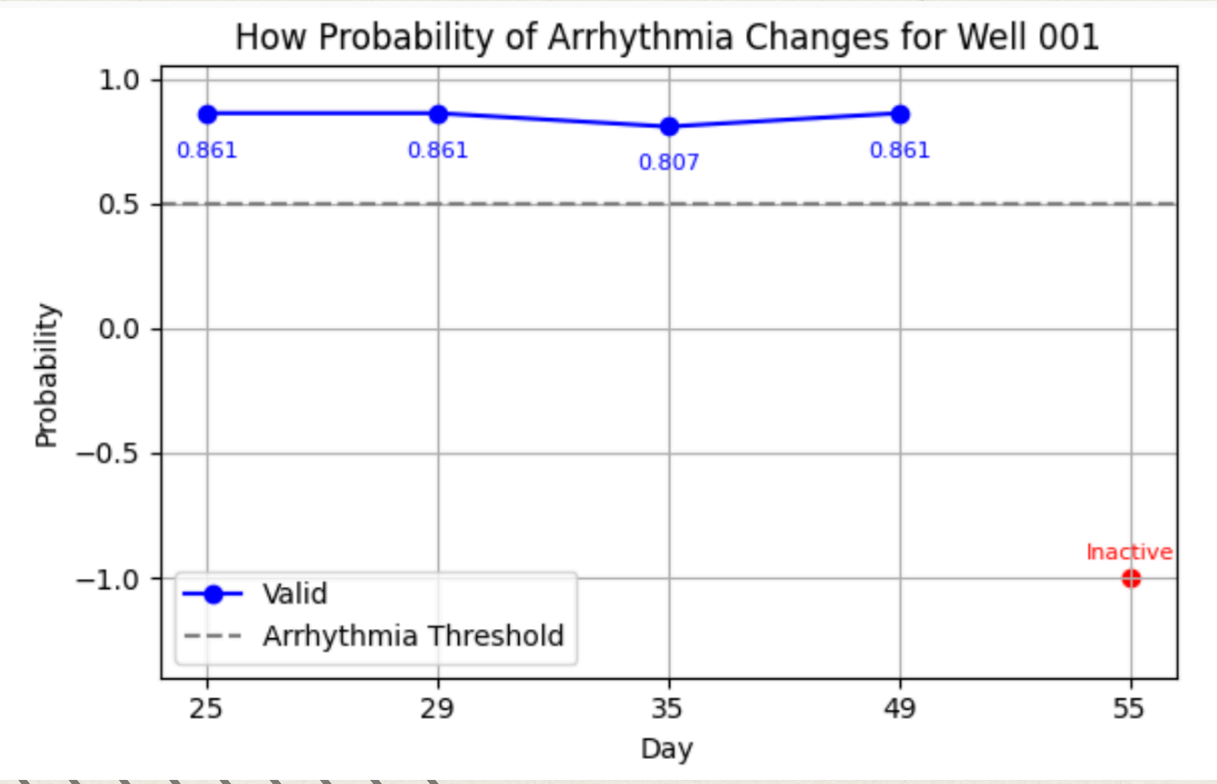
Sub 5: Probability Over Time

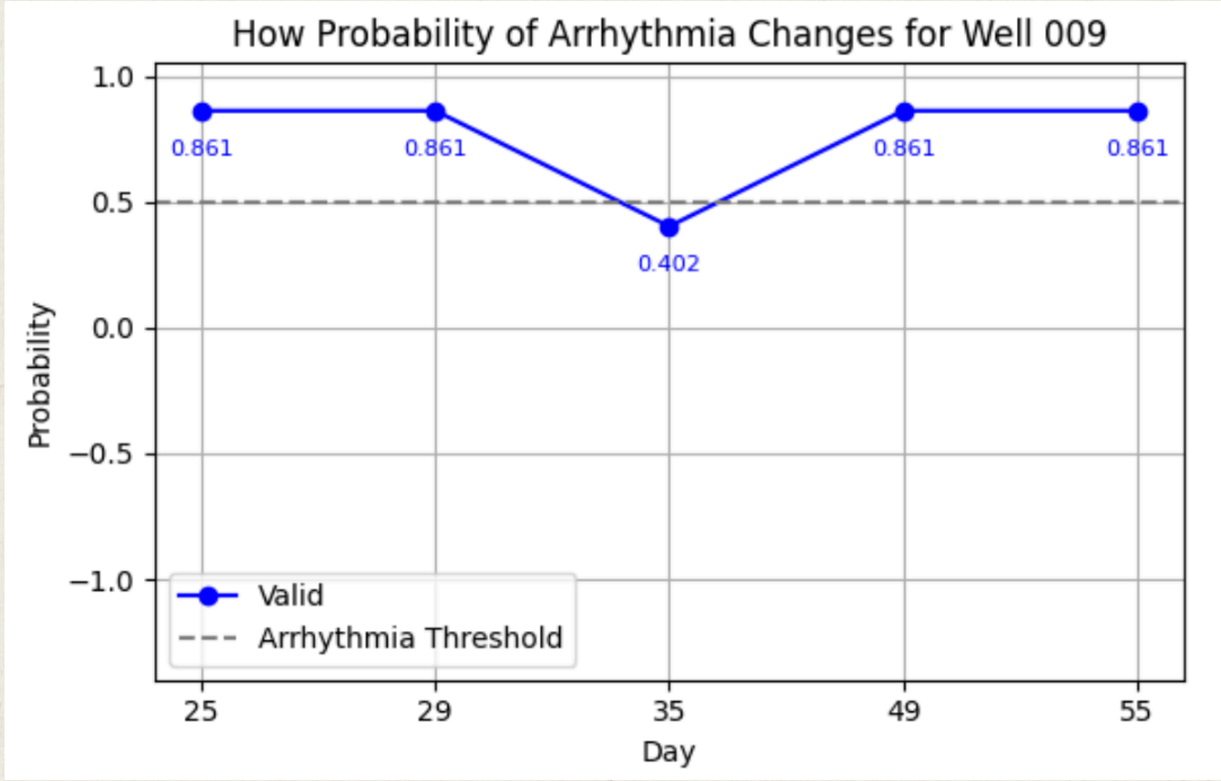
Graph for CP011, Well 001 over the course of the days data was taken



PHASE 4: Detecting Arrhythmia Using ML

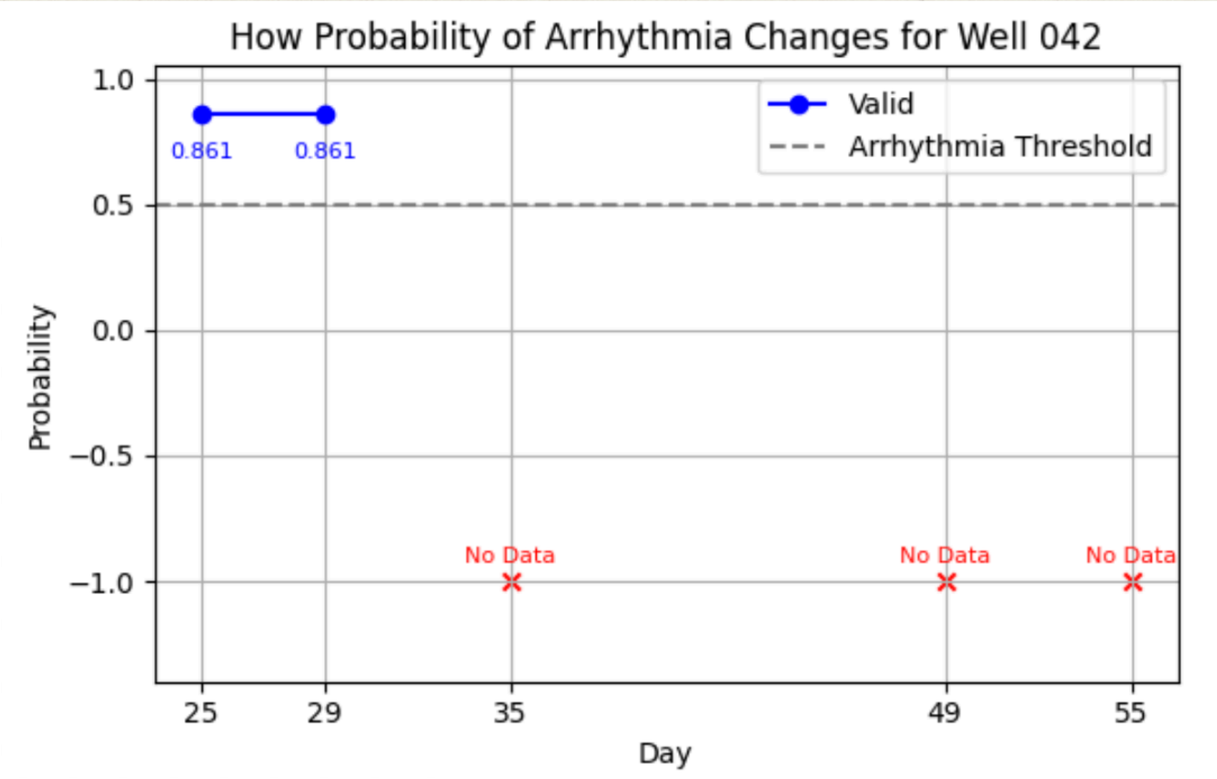
Sub 6: Probability of Plate Over Time

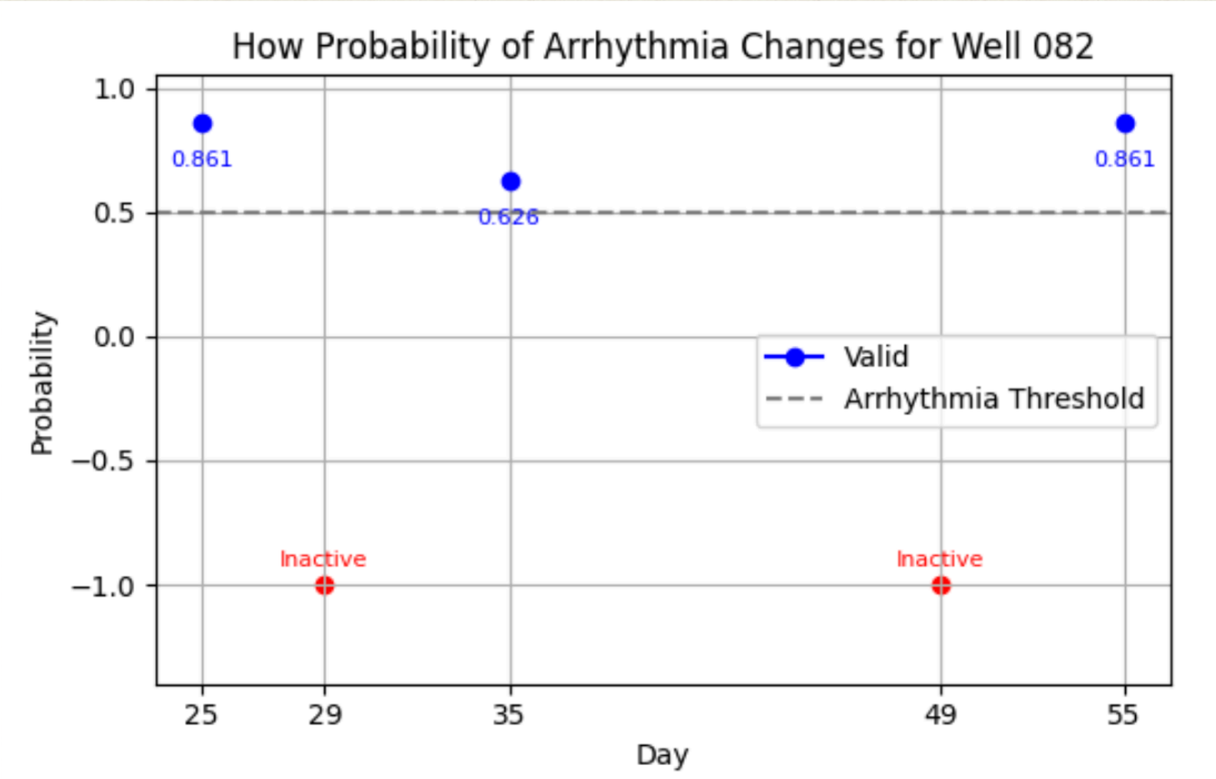




PHASE 4: Detecting Arrhythmia Using ML

Sub 6: Probability of Plate Over Time





THANK YOU

for listening to my presentation