Empowering Prenatal Diagnostics and Care with Machine Learning and Optimization

Angel Gantzia (angel_g@mit.edu)
Aditi Singh (aditis93@mit.edu)

15.095 Machine Learning under Modern Optimization Lens
December 12, 2023

1 Problem Overview

Infant mortality continues to be a significant issue in healthcare systems worldwide, even with the advancements in diagnostic tools like cardiotocography (CTG). In gynecology, CTG plays an essential role in assessing fetal well-being and identifying signs of fetal distress, both before and during labor. It measures and monitors fetal attributes such as heart rate and the mother's uterine contractions, aiding healthcare providers in determining the fetus's oxygenation and identifying any immediate medical concerns. However, despite its effectiveness, interpreting CTG data in a timely manner is challenging, especially in areas without access to expert obstetricians. This makes the method to be used on a case-by-case basis, but even then, it is not only time-consuming but also largely inefficient.

To address these issues, our project proposes the use of machine learning and optimization models for an efficient and accurate classification of fetal health. By employing a multi-class model to classify CTG features into three categories – healthy, suspect, and pathological – we aim to streamline the diagnostic process, significantly reducing the reliance on constant specialist intervention. Through this, we aim to enhance efficiency and lower fetal mortality rates, a goal underscored by the United Nations' 2022 report [1], which notes that over 4.5 million women and babies die each year during pregnancy, childbirth, or in the first weeks after birth, equating to one death every seven seconds. Most of these deaths are from preventable or treatable causes, emphasizing the need for proper care and the potential impact of improved diagnostic processes.

2 Data

The dataset used in this analysis is derived from cardiotocographic (CTG) readings, which are pivotal in monitoring fetal health during pregnancy. The dataset encompasses a comprehensive array of indicators reflecting the interplay between fetal heart rate (FHR) and uterine contractions, pivotal markers for assessing fetal distress. Integral metrics include the baseline FHR, the incidence of FHR accelerations and decelerations, fetal movement frequency, and the cadence of uterine contractions. Further depth is added through advanced statistical analysis of the FHR histogram, including measures like histogram width, peak counts, and variability indices.

The 21 total predictors enable a classification of fetal conditions into 'Normal', 'Suspect', or 'Pathological' states, providing a crucial tool for early intervention. Drawn from a substantial clinical sample size of 2126 cases, the dataset is characterized by an imbalanced class distribution, predominantly featuring 'Normal' cases (1655), followed by 'Suspect' (295), and 'Pathological' (176) cases. This imbalance reflects the real-world prevalence of these health states and presents an analytical challenge in ensuring accurate predictive modeling for fetal health assessment. The following figure succinctly visualizes the relationship between each of the predictors in the dataset by displaying scatterplots of each pair (lower half), alongside histograms along the diagonal and their correlation matrix (upper half). The axes labels in each row and column corresponds to a different variable, labeled at the extreme left for the rows and at the bottom for the columns, and the individual points plotted in the scatterplots show the observed values of the variables.

This exploratory data analysis was further broken down and analyzed especially in the (Clustering section) for patterns that indicate relationships between the variables. Furthermore, it was also used when trying to pinpoint potential non-linear transformations that could be introduced to the model (individually in each feature, as well as between by multiplying the features).

3 Challenges

Through the project, we had to deal with multiple challenges that stemmed from the dataset, which was imbalanced, had limited original predictors due correlations, and was riddled with medical jargon.

As aforementioned, there was a stark imbalance in class distribution, with most cases being labeled as 'Normal', and only few as 'Suspect', and 'Pathological'. This disproportion resulted in models having a biased understanding, favoring the majority class ('Normal'). To counteract this, we tried the naïve approach of trimming the dataset's Normal class, to even the rest of the statuses out. However, seeing that our data was too few to begin with, this worsened our metrics. Thus, we tried training the models using Synthetic Minority Over-sampling Technique (SMOTE), to generate synthetic examples for the underrepresented classes—'Suspect' and 'Pathological'—by interpolating between existing, similar cases.

[1] United Nations, *United Nations Report 2022*

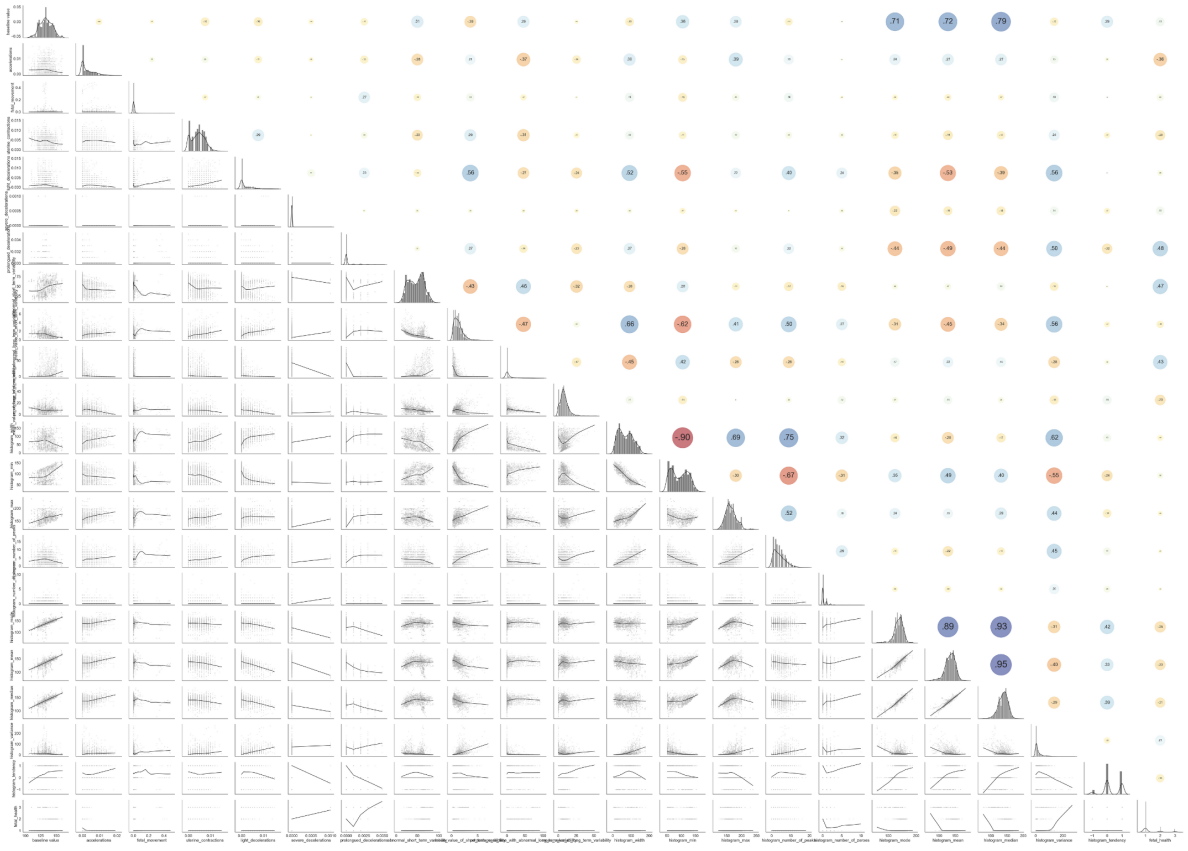

Figure 1: Plot with pairwith feature relations (lower-half) and correlation matrix (upper-half)

This approach balanced the dataset, enhancing the models' ability to generalize and accurately identify all classes.

However, this method was applied only on the training set, as it's inapplicable to our test set. This means that it still remained imbalanced, making it a bad fit for just using accuracy as an evaluation metric, as it would favour the percentage of correctly classified cases. That is why instead, we turned to metrics such as weighted F1 scores (which is the harmonic mean of precision and recall, our main metric), by calculating the F1 score for each class, weighing it, and then averaging it to obtain a singular value.

Apart from that, another challenge was that the initial dataset included only 21 predictors, many of which were correlated, counting similar metrics (for instance, histogram mode, histogram mean, histogram median). To expand our feature space and extract more nuanced patterns from the data, we designed a function to calculate additional non-linear features from the existing ones. This function augmented our dataset with new columns, each representing a transformed version of the original features. The transformations we tested and included in the pipeline were squared of the original feature, square root, and logarithmic transformations. Furthermore, getting the opinion of a resident doctor, Dr. Fournari, we also located which important columns would capture more complex relationships within the data, were they to be multiplied with each other. These interaction features included (histogram_mean*mean_value), (fetal_movement*uterine_contractions), and (baseline*accelerations).

Lastly, since this project involved intricate medical terminology for MBAn students with no medical background, we bridged our knowledge gap by consulting with Dr. Fournari. This collaboration was crucial in interpreting the findings and understanding the medical context in some of the CTG features (to see which ones would create good patterns that would raise performance, as explained above).

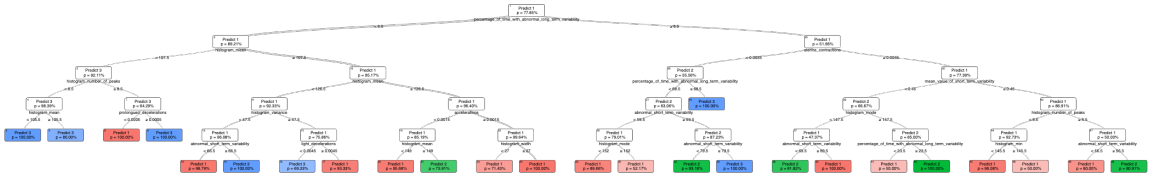

Figure 2: Best OCT model output (Scaling: False, SMOTE: False, Transformations: False)

4 Methodology

4.1 Modeling

In the beginning of our project, we implemented a baseline model using a simple logistic regression with no hyper tuning to serve as a point of comparison for more complex algorithms and help to identify the strengths and weaknesses of the relationships between features and the target variable. The baseline yielded an accuracy of 85%, a weighted F1 score of 85%, a weighted precision score of %83, and a weighted recall score of %85. Individually, with a precursory look, these results are acceptable, but upon closer examination we notice that individual F1-scores can be as low as 40% (e.g. in the case of Suspect cases).

To improve upon these scores, we built a pipeline that run 5 models including Logistic Regression, CART, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier, and Optimal Classification Trees, with OCT being the most interpretable. The selection of each model was intentional where we started by hyper-tuning logistic regression because of its simplicity. Then, we used SVC due to its applicability in high-dimension space and the use of kernels. We also chose to use tree-based models like decision trees or CART for its interpretative advantages and ability to mirror human decision-making.

Building upon the foundation of Decision Trees, we also used Random Forest for ensemble learning to improve predictive performance. By aggregating the results of many decision trees, it reduced the risk of overfitting and provided a more accurate and stable prediction. Together, these models span from simplicity to complexity, from transparency to black-box models, each with its strengths in capturing different aspects of the data. The Logistic Regression model offers a quick and understandable approach, while the Decision Tree provides a visual representation of decision paths. The Random Forest and SVC provide more sophisticated and nuanced predictions, and the OCT bridges the gap by offering both high interpretability and optimized performance. Figure 2 showcases the resulting best OCT which includes percentage_of_time_with_abnormal_long_term_variability, histogram_mean, and uterine_contractions as some of the important features when considering classification into fetal states.

During this procedure, we went back and forth over the columns we wanted to include due to the highly correlated features issue and frequently consulted with Dr. Fournari to ensure that the variables we chose made sense. With robustness and accuracy in mind and due to the challenges posed above, we adopted a comprehensive approach by applying scaling, SMOTE, and various transformations across all possible combinations, resulting in the creation of 40 distinct models. An extensive grid search and cross-validation were conducted to find the best hyperparameters. All of this was compiled into a single pipeline that ran all the models with the mentioned specifications. In the end, we also implemented two versions of soft voting, one for models with transformations and one for models without transformations that all used scaling and SMOTE. While these classifiers gave us useful insights, we wanted to present our findings in an actionable way where healthcare professionals could understand and enact them. Thus,

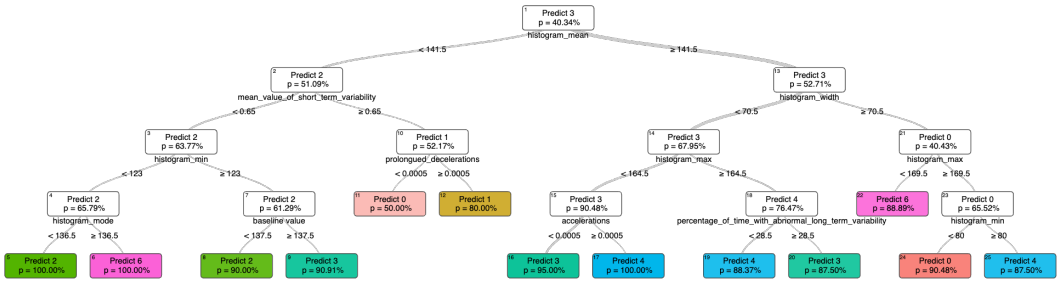

Figure 3: Interpretable clustering for suspect cases

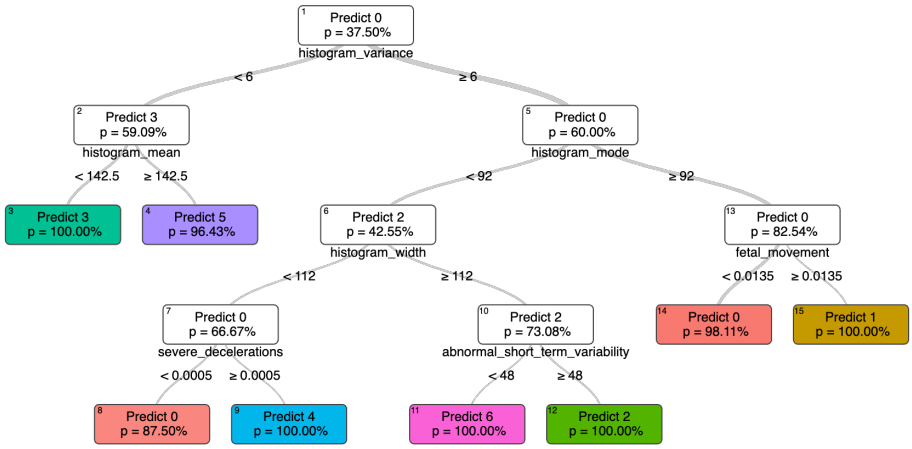

Figure 4: Interpretable clustering for pathological cases

we decided to continue with interpretable clustering.

4.2 Clustering

To emphasize interpretability, we executed interpretable clustering using the two-stage approach to identify key features of the CTG readings that result in the respective fetal classification. We divided the clustering into 2 sections: hierarchical clustering for only suspect, and hierarchical clustering for only pathological. First, we scaled our data, as clustering is sensitive to this. We utilized dendrograms(using the method ward) and scree plots, employing the elbow method, to find an optimal k ie. the number of clusters. Instead of choosing one specific k, we tried a few different values that seemed promising to find the best interpretable result. We also executed grid search to find the best parameters for clustering. We settled for k=7 for the suspect-only clustering and k=7 for the pathological-only clustering as they seemed to provide applicable insights while also being interpretable. Figures 3 and 4 showcase the respective trees.

For suspect-only clustering, the most important features seem to be histogram_mean, prolonged_decelerations, and histogram_min. For pathological-only clustering, the most important features seem to be histogram_variance, histogram_mean, and histogram_mode. The different clusters can be viewed as different cases which lead to being labeled as either suspect or pathological, which we can, then, translate into actions by executing different actions. Essentially, we're translating these cluster characteristics into targeted interventions, adopting specific actions tailored to each case. The way healthcare professionals can prescribe care based on this will be detailed in key insights.

5 Results and Implications

	SVC	Random Forest	OCT	CART	Logistic Regression
Applied Scaling	Yes	No	No	No	Yes
Applied SMOTE	No	Yes	No	Yes	Yes
Non-linear transformations	No	Yes	Yes	No	Yes
Yes					
Accuracy	92.48%	91.54%	89.10%	89.10%	88.35%
Weighted F1 Score	92.30%	91.83%	89.73%	89.39%	89.20%
Weighted Precision	92.22%	92.34%	91.36%	89.84%	91.02%
Weighted Recall	92.49%	91.55%	89.10%	89.10%	88.35%

Table 1: Scores for Best Performing Tuned Classification Model in Each Classifier Type

The Support Vector Classifier model shows the highest Accuracy and F1 Score, indicating it is quite effective for this particular classification task. It also has the highest Precision, suggesting it is less likely to label a negative sample as positive. However, its Recall is slightly less than the Random Forest Classifier, meaning it might miss a few more positive cases.

The Random Forest Classifier has the second-highest F1 Score, which suggests that it balances Precision and Recall relatively well. It is a versatile model that can handle both linear and non-linear data, reflected in the use of SMOTE (Synthetic Minority Over-sampling Technique) and non-linear transformations to enhance its performance. However, it might be more complex and less interpretable than simpler models, such as Logistic Regression or CART (Classification and Regression Trees).

OCT (Optimal Classification Trees) and CART are tree-based methods that usually offer more interpretability due to their decision-tree structure, which can be visualized and understood by humans. They both have the same Accuracy, but CART has a slightly better F1 Score and Precision. The simplicity of these models can be advantageous when the interpretability of decisions is crucial, even though they might not achieve the highest possible accuracy.

For the first version of soft-voting, all models that were specified as scaling equal to true, SMOTE equal to true, and transformations equal to true were considered. For the second version of soft-voting, all models that were specified as scaling equal to true, SMOTE equal to true, and transformations equal to false were considered. This is because we wanted to consider the cases where scaling and SMOTE helped deal with the challenges in our data and also check if transformations were improving our metrics. The resulting metrics are provided in Table 2. The metrics were pretty similar regardless of whether transformations was true or false and yielded the best F1 score compared to all the individual models.

	Tranformations=True	Tranformations=False
Accuracy	93%	92%
Weighted F1 Score	93%	92%

Table 2: Soft Voting Results

6 Key Insights

The Insights section of our report is comprised of two sections, in which we present the analysis of the interpretable clusters we created with IAI, for only the pathological and only the suspect cases. This analysis is meant to guide practitioners to further clarify upon the model's predictions, and find the next steps that need to be taken to give proper care to the baby.

6.1 Clustering Pathological Cases For Actionable Insights

For the Pathological clustering, we break down the analysis into 1) the pathology and 2) the action(s) that practitioners can take. We have 7 distinct clusters, which can be seen in Figure 3, in the above section.

The first critical split is on histogram variance. When histogram variance is low, the Dark Green cluster, for instance, epitomizes a pattern of uniformity in heart rate, with low histogram variance and mean. This potentially signifies a worrisome non-reactive fetal state — an insight that triggers a cascade of clinical responses, including the deployment of non-stress tests (NST) and biophysical profiles (BPP).

Next, in the Purple cluster, we discern an elevated baseline in fetal heart rate (due to low histogram variance with high mean), a pattern that might be reflective of underlying stress, either maternal or fetal in origin. This insight can guide clinicians to investigate maternal health parameters and fetal behavior, with an evaluation for fever or infection.

On the other hand, when on histogram variance is high we have a higher number of problematic situations.

To begin with, the Red Cluster exposes fetal compromise in two scenarios. First, when we have Low Histogram Mode & Width with Few Severe Decelerations, this may reflect a less pronounced variability in heart rate, with occasional but not frequent severe decelerations, possibly indicating intermittent fetal distress. Secondly, this cluster can also be characterised by High Histogram Mode & Low Fetal Movement, suggesting periods of increased baseline heart rate coupled with reduced fetal activity, which can be concerning for fetal well-being, potentially signaling a hypoxic event. In both cases, the cluster necessitates a rigorous assessment of uteroplacental function, often leading to the initiation of continuous monitoring protocols and potentially, interventions to enhance fetal oxygenation, since there's a potential for fetal lung immaturity.

In the subcase of Low Histogram Mode & Width with high Severe Decelerations, Blue cluster, signals the need for immediate and decisive action, often translating to emergency delivery if the clinical scenario permits. This cluster's criticality is underscored by its direct association with acute fetal distress, a state that mandates swift intervention to ensure fetal well-being.

In the case of high Histogram Width, we can be led into one of the two following clusters. Specifically, the Pick cluster with low abnormal short-term variability suggests less acute stress, calling for close monitoring and serial assessments. Its more worrisome counterpart is the Light Green cluster with high abnormal short-term variability, instead, suggests significant fluctuations in the fetal heart rate in a very short time. This can be a sign of acute fetal distress, potentially due to umbilical cord complications or rapid changes in fetal oxygenation. Clinical action should involve immediate and thorough assessment, including possible fetal scalp blood sampling if the situation allows, and continuous electronic fetal monitoring to determine the appropriate intervention, which could range from conservative management to expedited delivery based on the overall clinical picture.

The last cluster, the Yellow one, is characterized by high histogram variance and mode, with high fetal movement. While flagged as a pathological case, this may indicate an acute compensatory response to a stressor. The heightened fetal activity could be an attempt to improve oxygenation in response to transient hypoxic events. The high variance and mode suggest fluctuations in heart rate that are not stabilizing, a sign often associated with fetal distress. This necessitates immediate investigation, such as detailed ultrasonography and possibly fetal heart rate monitoring, to ascertain fetal status and guide potential interventions.

6.2 Clustering Suspect Cases For Actionable Insights

In the suspect case analysis, we discern six clusters with different pathological implications and corresponding clinical actions, as outlined in Figure 3. The initial split in our decision tree hinges on the mean value of short-term variability. With low variability, we encounter the Green Cluster, where a relatively stable heart rate suggests a vigilant wait-and-see approach with consistent monitoring. Conversely, high variability propels us into more concerning territories.

The clusters here diverge based on the presence of prolonged decelerations and baseline values. In the Blue Cluster, prolonged decelerations without high mean variability point to potential episodic fetal oxygenation issues, calling for NSTs and possible immediate intervention. When faced with high mean variability in the Red Cluster, the implications of potential acute distress necessitate a comprehensive evaluation for expedited action, possibly including delivery.

As we delve further into clusters characterized by the width of the histogram, we encounter the Pink and Light Green Clusters. The Pink Cluster, with its combination of wide histogram and few decelerations, could indicate a recovering fetus and demands close but not emergent monitoring. The Light Green Cluster, with a wide histogram yet marked by frequent decelerations, requires a more proactive approach, including fetal blood sampling and continuous monitoring, to quickly address any arising complications.

Lastly, the Yellow Cluster presents a unique paradox: high fetal movement and high variance, typically signs of a healthy, active fetus, yet in a suspect context, it may mask underlying stress. Thus, it necessitates a nuanced response, balancing between non-invasive monitoring and more definitive investigative measures as clinical indicators evolve.

7 Conclusion

In conclusion, if signs of fetal distress and their sources are overlooked before or during labor, the consequences can be dire. By utilizing this data-driven predictive and clustering approach, we can raise alarms that flag suspect and pathological fetal statuses and as such prevent irreversible organ damage or even death. With such timely warnings, medical actions can be employed, for each of various reasons for flagging distress, including administering oxygen to the mother, hydrating her, or giving medications to control uterine contractions, or in more severe cases, if the fetus is considered to be in immediate danger, an emergency delivery might be necessary (usually an emergency C-section).

That is why, in our project, it is highly important to have a reliable model that will not cause unnecessary alarm or ignore the possibility of danger. That is why we employ as a primary metric the F1-score, which gives us a balance between the two. Our results across multiple models give way to high reliability even when tested out of sample. However, in the trade off of performance and accuracy, we are in favour of more explainable models such as OCT, which are very easily interpretable, as this project is meant to guide and lessen the workload of practitioners who will ultimately be deciding over each case. This was important as it allows more trust in the diagnosis. Lastly, apart from diagnosing the health status, we presented the clustering part of the analysis, which aims to give doctors insight into where the irregularities stem from, and then allow them to act on it either without alarming the patient or intervening immediately when it is critical. The clustering can give insights to individual pathologies or suspeted pathologies that may plague the fetus, going over suggestions and next steps that can be taken. As such, this project fully equips healthcare professionals with the diagnostic ability to discern between cases needing immediate intervention and those that warrant careful monitoring or intervention.

8 Contributions

In their project, Angel and Aditi each played key roles with distinct responsibilities.

Angel focused on the initial stages of the project. She cleaned the data, conducted exploratory data analysis, and created the baseline model. Her major contribution was developing a pipeline of models, which included Logistic Regression, SVC, Decision Tree, and Random Forest. This pipeline also accounted for scaling and SMOTE, resulting in 16 different model variations. Beyond this, she fine-tuned the parameters of these models for better performance.

Aditi extended the project's capabilities. She integrated IAI (Interpretable Artificial Intelligence) with the existing Python codebase. Her work involved introducing non-linear transformations to all the existing models, adding Optimal Classification Trees into the pipeline, and implementing soft voting techniques. Aditi also developed the interpretable clustering segment, focusing on clustering for specific cases (only suspect and only pathological) and creating OCTs for various cluster numbers.

Together, Angel and Aditi communicated with Dr. Fournari mulitple times to benefit from her knowledgeable feedback. Lastly, both teammates collaborated on compiling the final report. Angel primarily prepared the presentation slides, for which Aditi supplied the necessary OCT graphics.