2018

Fetal Heart Rate Exploration

CAPSTONE REPORT BY:

AHMED, HUMZA A

PROJECT OVERVIEW

Childbirth has been deconstructed into occurring in three general stages [17].The first stage beginning with the onset of contractions as the cervix begins to dilate to its maximum size. During this stage uterine contractions are around 30-45 seconds, and irregular in nature occurring around 20 minutes apart. Progressively these contractions become more frequent and get closer to lasting between 60-90 seconds. The second stage begins after the cervix has reached its maximal diameter, and involves regular contractions between 60-90 seconds occurring around every 5 minutes. During this stage the baby will actively move through the birth canal until it is delivered. The final stage ends with the delivery of the placenta. From the middle of the gestational period to the time of delivery, the early detection of fetal complications can help improve outcomes for both child and parent (e.g. through Cesarean Section).

Monitoring of fetal state during pregnancy most commonly occurs through the use of Cardiotocography (CTG), or Fetal Electrocardiography (fECG). Both of these methods have the ability to measure fetal heart rate (FHR) which can indicate forms of fetal distress. However, the complexity in making precise judgements based on FHR data has caused division between clinicians about its value to consistently to improve patient outcomes. In order to standardize clinical practice, the American College of Obstetricians and Gynecologists (ACOG) has published a set of guidelines and definitions to facilitate interpretation of FHR data [2]. These definitions in addition to more familiar heart rate characteristics used in adults has allowed clinicians to now describe changes in FHR quantitatively. Although these guidelines have made FHR interpretation more quantitative, they are based upon a review of studies showing mixed results in patient outcomes. These studies and others have even shown an increase in the number of unneeded interventions conducted in healthy patients. [1, 8]. Thus, it seems that there is still a need to further solidify our understanding of normal FHR patterns.

The ACOG guidelines in addition to most clinical literature analyzing FHR tends to ignore the temporal differences between Stage I and II of labor. Clinical literature either tends to conduct a global analysis of FHR during labor, or focus on a single stage [4, 12, 19]. It is however known that during uterine contractions FHR shows patterns of deceleration and acceleration [6]. Thus, naturally it would be expected that FHR patterns differ between Stage I and II of labor due to the differences in contraction duration and regularity.

Recent machine learning work has found significant differences in model performance and parameter estimates when separating FHR data by labor stage. Spika's work explored the impact of separating data by labor stage on the use of FHR to classify cases of healthy fetuses, and those with acidosis [16]. Not only did creating separate classifiers improve performance, the respective models preferred different input features. In a separate paper, Granero-Belinchon shows that several entropy measure estimates are significantly different when FHR data is separated by labor stage [10]. These observations suggest that the ACOG guidelines may need

revision to specify separate recommendations for interpreting FHR data during each stage of labor.

PROBLEM STATEMENT

In summary there is a need to determine if and how normal FHR characteristics differ between Stage I and II of labor. Previous work has been done using machine learning techniques to differentiate heart rate characteristics for both fetuses and adults [3, 5, 7, 10, 11, 13, 14, 16]. Models such as these can be evaluated based on various measures of accuracy/purity and compactness. The objective of this work was to investigate the link of FHR characteristics to their origin in time within the first two stages of childbirth. This was accomplished through two main analysis after labelling sample FHR time-series as originating from within Stage I or II of labor (i.e. as binary labels):

- **Unsupervised Clustering:** K-means and hierarchical clustering models were developed, optimized, and visualized. This allowed for the determination of the number of distinct groups naturally present in FHR data. Furthermore, the link between each group to a particular labor stage was investigated by looking at cluster composition in terms of the binary sample labels. Visualization would serve as a means of clustering model verification and would give insight into the behavior of the supervised models described below.
- **Supervised Learning:** If FHR characteristics do vary between each stage of labor, these variations should be able to predict the underlying binary labor stage labels. This predictive ability was assessed by developing decision tree and random forest models. These classification models were chosen because of previous work showing information theory estimates of entropy (linked to the concept of information gain) being different within each labor stage[10]. These models were assessed for accuracy in comparison to the accuracy of a benchmark model always predicting the dominant class (after outlier removal). Furthermore, model stability was also assessed when cluster labels were used as inputs to the two classification models.

METRICS

The unsupervised and supervised models were evaluated using the metrics below:

Unsupervised Clustering Metrics: Clustering models were evaluated based on metrics of compactness, uniformity, and similarity to the "true" clustering that assigned each binary label to its own cluster. Multiple measures were used to help determine the simplest clustering model during optimization.

1. **Silhouette Coefficient:** This metric was used to evaluate if clusters could be regarded as spatially distinct. The Silhouette Coefficient is the calculated as the difference between mean inter-cluster distance and mean intra-cluster distance divided by the max of the two listed distances. The coefficient ranges from -1 to 1. Values above 0 would indicate clusters that are on average further apart than the width of an average cluster.
2. **Homogeneity**: This metric was used to measure how perfectly clusters contained samples from a single labor stage. This metric ranges from 0 to 1, where 1 would mean clusters do not contain more than one class. In sklearn this is calculated by building a contingency matrix of the predicted clustering and true clustering labels. A mutual information score is then calculated based on this matrix. This information score is then divided by the entropy of the true labels.
3. **Purity:** This would serve as an easy way to describe the distribution of samples from a particular labor stage. Purity was calculated as the average accuracy of clusters if all samples within an individual cluster were labeled to be the most frequented class. For the given problem this metric would range between 0.5 and 1. The lower bound is 0.5 for this problem because there are only two classes, and thus the most frequented class within a cluster would always have an accuracy above 0.5.
4. **Adjusted Rand Index:** This measure was used to determine if clustering results were similar to the binary labels of labor stage. The adjusted rand index measures how well the model puts pairs of similarly labeled examples within the same clusters. The measure varies from -1 to 1, where 1 would correspond to identical clustering (in sklearn the adjusted rand score acts symmetrically). This measure is adjusted so that scores >0 correspond to two clusterings more similar than expected by chance.
5. **Adjusted Normalized Mutual Information**: This metric would also be used to determine if clustering was related to the binary labor stage labels. Mutual Information can describe the reduction in entropy of class labels based on if we know the cluster labels. The measure varies between -1 and 1, where 1 would correspond to perfectly homogenous and complete clusters. This measure is adjusted so that scores >0 correspond to clusterings more homogenous and complete than that expected by chance.

Supervised Classification Metrics: Classification models were evaluated based on accuracy, precision, and recall. Percision and recall would help to evaluate the models for bias due to class imbalance after outlier removal.

DATASET CONSTRUCTION

The open source Intrapartum CTG dataset available on PhysioNet [9, 18] was used as the basis of this work. This data was all collected at the University Hospital in Brno, Czech Republic. The dataset contains FHR time-series readings for at most 90 minutes before labor. Recordings met the following timing criteria during Stages I and II of labor:

- Stage I recordings were at least 30 minutes, and at most 60 minute.
- Stage II recordings to delivery were kept to at most 30 minutes.

There were a total of 552 subject recordings. Only records of patients having an umbilical pH > 7.15 after birth were included. This criteria is commonly used to determine if a Fetus has respiratory hypoxia or acidosis. Only subjects that were delivered Vaginally were included. Caesarian Sections were excluded as they may indicate the presence of other confounding factors. Furthermore, the organizers of the dataset only included full-term infants (>37 weeks of gestation), infants free of opiate administration, and only singleton pregnancies. One example FHR time-series and uterine contraction signal recorded by the clinicians of a subject is shown in Figure 1.

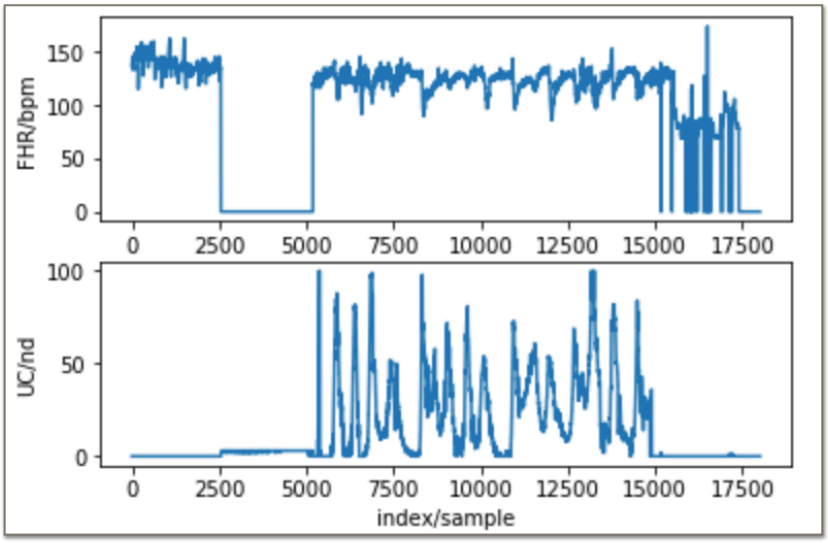


Figure 1: Example plot of a raw recorded FHR and Uterine Contraction time-series. Times with an unmeasurable heart rate can be seen where the heart rate drops to 0.

The following processing steps were conducted to create samples from each time-series:

- The first 2.5 minutes of each recording was discarded. This was done to remove any potential inaccurate readings during the CTG monitoring setup.
- Signals were compressed to remove any time where the CTG was not able to record a FHR signal. Compression was done only including nonzero signal samples.

- An equal number of 10 minute Stage I and II FHR time-series samples were collected. For participants with lengthy recordings, multiple samples were collected. Multiple samples were collected by selecting a contiguous time-series rounded down to the nearest 5 minutes of Stage II length (e.g. For a participant with a 23 minute long Stage II, I randomly select a 20 minute sample of both Stage I and II to use in my analysis). Ten minute segments were collected with an overlapping windows of 5 minutes. In order to help get reproducible results for this paper, contiguous time-series were from the first viable signal datapoint (**not random**).
- Based on the above points participants with compressed signal lengths for Stage I or II of less than 10 minutes were excluded.

This signal processing resulted in a total of **384** samples evenly split across labor stage.

INPUT FEATURES

A total of eight time domain and frequency domain features were computed.

Time Domain Features:

The time domain features were meant to closely resemble ACOG definitions describing FHR data.

- **Baseline Heart Rate (BHR)**: Mean heart rate over a period of 10 minutes. Since a single sample had a length of 10 minutes this would be equivalent to the mean of a particular signal sample.
- **Baseline Heart Rate Variability (HRV)**: Standard deviation from baseline heart rate.
- **Number of Accelerations (nAcc)**: An increase of 15 beats per minute (bpm) from baseline for over 15 seconds. The ACOG defines an increase of 15bpm for over 10 minutes to be a change in baseline. Thus, accelerations would only be counted if they returned to within 15 bpm of baseline within a given sample time-series.
- **Number of Decelerations: (nDec)** A decrease of 15 bpm from baseline for over 15 seconds (Opposite of nAcc). The ACOG defines multiple types of decelerations which also takes into account the simultaneous occurrence of uterine contractions. Since uterine contraction signals were inconsistently recorded at times when an FHR signals was able to be acquired information about contractions was ignored.

Frequency Domain Features:

Frequency domain features were included to help characterize nervous system activity controlling heart rate. These features were calculated using the magnitude of the squared Fast Fourier Transform. Energy spectrum densities were calculated as the Reiman sum of each respective frequency band. Bands are based on the frequencies used in [15].

- **Power estimated from very low frequency bands (VLFB)** (<0.03Hz). Used to describe thermal regulation mechanisms.
- **Power estimated from low frequency bands (LFB)** (0.03-0.15Hz). Used to describe sympathetic neural activety.
- **Power estimated from high frequency bands (HFB)** (0.15-Nyquist Frequency). Used to describe parasympathetic neural activity.
- **Ratio of LFB/HFB Energy Spectrum:** Used to describe the balance between sympathetic and parasympathetic activity.

DATASET EXPLORATION AND PREPROCESSING

The input space was explored by plotting feature histograms shown in Figure 2. Many of the features showed a rightward skew. In order to help these distributions better approximate gaussian distributions a natural logarithm transform was applied to all of the features. Furthermore, in order to limit feature dominance in unsupervised clustering distance metrics a minmax scaling transformation was applied. Feature distributions after these transformations can also be seen in Figure 2. After scaling the distributions show more spread, however there is still skewness present particularly in frequency domain features. These observations seem to indicate the presence of many outliers.

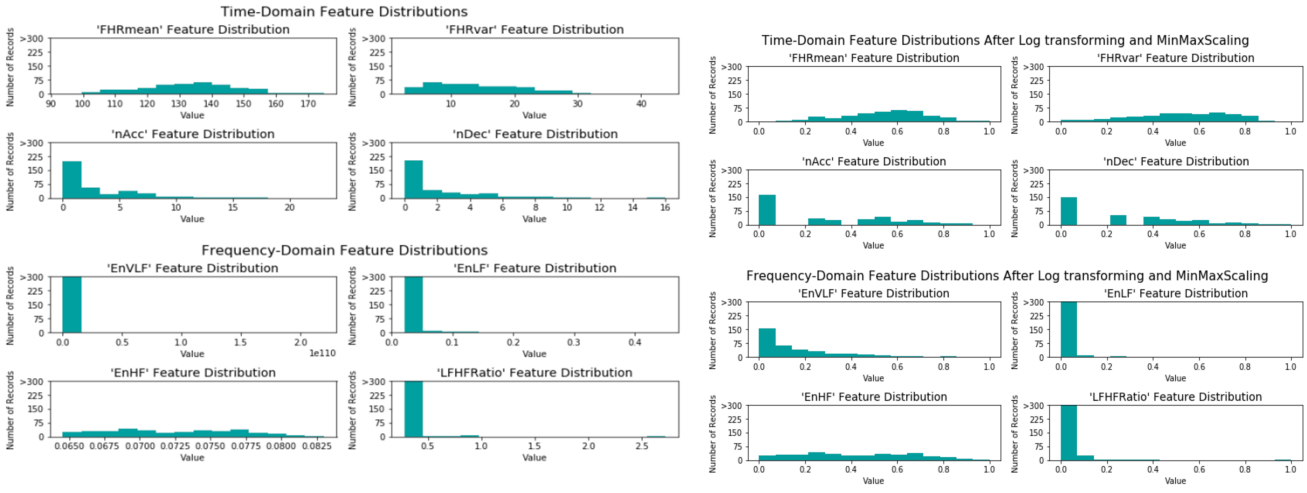


Figure 2: Raw feature distributions (left), and feature distributions after log transformation and minmax scaling (right).

Figure 3 shows a box plot of the scaled features. Outliers (denoted as 1.5*IQR of each feature) are shown as unshaded circles. Almost all of the outliers can be seen to occur within the frequency domain features, which is in agreement with the extreme skewness seen in the histograms. In particular EnLF and LFHF features without outliers show very little variance. Thus, the outliers could potentially have a large effect on any dimensionality reduction, and calculation of distance metrics for clustering. It was decided that all outliers would be removed from the dataset. **The final sample count after outlier removal was 267 (150 Stage I samples, and 117 Stage II samples.**

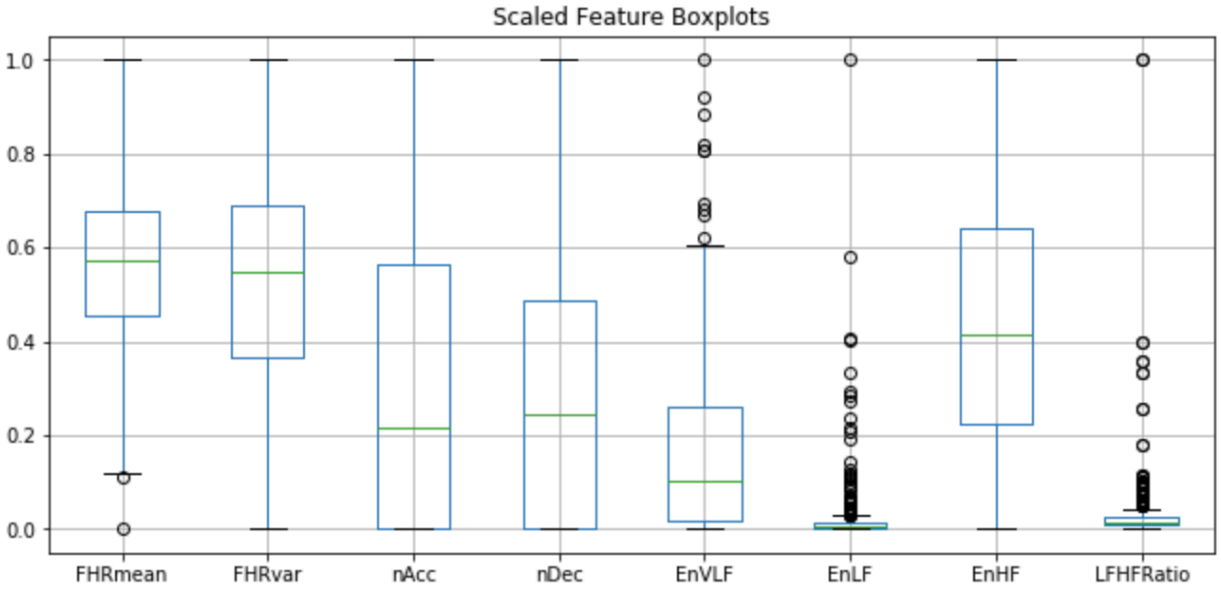


Figure 3: Boxplot of input features after scaling. Outliers are denoted as open circles if they were ±1.5*IQR of each respective feature.

Table 1 and Figure 4 show descriptive statistics and distributions of the final dataset respectively. The EnLF and LFHFRatio features show much less skewness, and all features generally show more spread than before. There is still a noticeable skew within the nDec, nAcc, and EnVLF features. This remaining skewness should at least give more representative estimations of the proposed non-parametric unsupervised and supervised models than before outlier detection. FHRmean and FHRvar show the most variance and are also the most normally distributed. This is to be expected due to heart rate being the nonlinear outcome of a large number of physiological controls. The large spread in FHRvar values is in agreement with the datasets large spread in nAcc and nDec. Lastly, it seems that parasympathetic modulation (EnHF) must be the main neurological component determining FHRmean and FHRvar. Due to the small standardized values of EnLF, the LFHFRatio feature also shows little variability irrespective of the standardized values of EnHF. These observations are in agreement with the

covariance matrix seen in Figure 4. In addition, this covariance matrix shows a relatively strong negative covariance between nAcc and nDec. This may indicate that a Fetus either tends to have large drops or increases in heart rate (but not both). The large number of potentially significant interactions indicate that dimensionality reduction to two or three features would allow for a reasonable summarization of the dataset.

Table 1: Descriptive statistics of the raw input feature space.

	FHRmean	FHRvar	nAcc	nDec	EnVLF	EnLF	EnHF	LFHFRatio
count	267.000000	267.000000	267.000000	267.000000	267.000000	267.000000	267.000000	267.000000
mean	0.571604	0.467029	0.270620	0.270678	0.149686	0.006190	0.514665	0.014746
std	0.170803	0.197285	0.301702	0.281483	0.158355	0.005531	0.212609	0.006959
min	0.117633	0.000000	0.000000	0.000000	0.000000	0.000000	0.014113	0.000000
25%	0.484289	0.327750	0.000000	0.000000	0.015662	0.002990	0.341598	0.010253
50%	0.586499	0.474372	0.218104	0.244651	0.101084	0.004168	0.539591	0.013692
75%	0.684037	0.613605	0.563791	0.489301	0.224312	0.006650	0.673057	0.017520
max	1.000000	0.927919	1.000000	1.000000	0.605886	0.028870	1.000000	0.042555

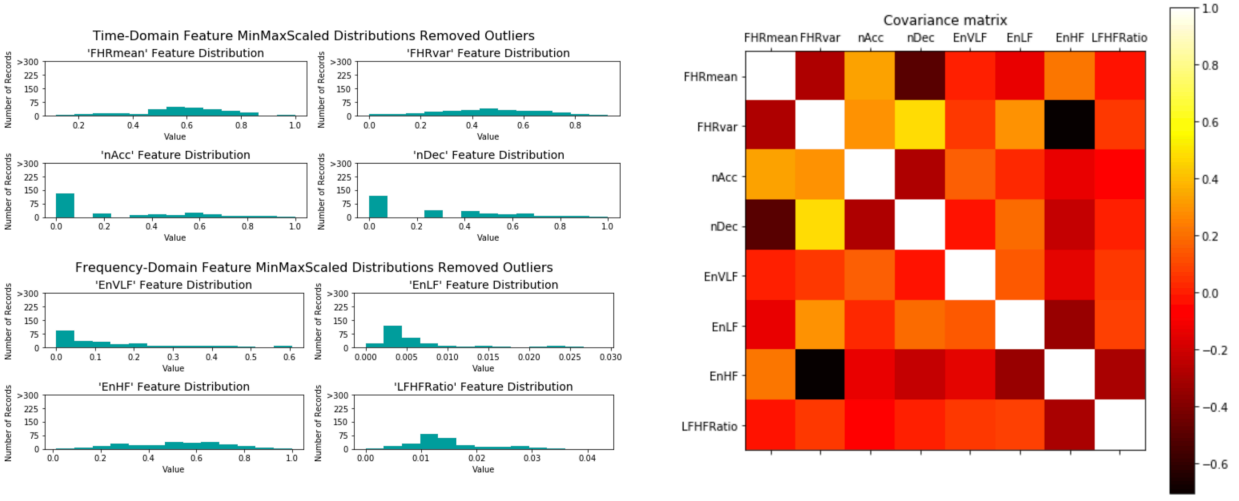


Figure 4: Feature distributions and covariance matrix of the raw input space after outlier removal.

Dimensionality reduction was conducted by applying PCA to the dataset. Figure 5 describes the first 5 components and their respective explained variance. The first two components have a combined explained variance of 0.72, and the first three components have a combined explained variance of 0.84. Looking at the explained variance plot the first component can be thought of as the ratio of nAcc to nDec for a given FHRmean. The second component seems to describe the total amount of variability in heart rate relative to a given EnHF or parasympathetic activity. The third component seems to describe stable (i.e. Acc or Dec rather

than brief spikes which could be more likely detected as a large FHRvar) changes in heart rate regulated by changes in EnHF or parasympathetic activity.

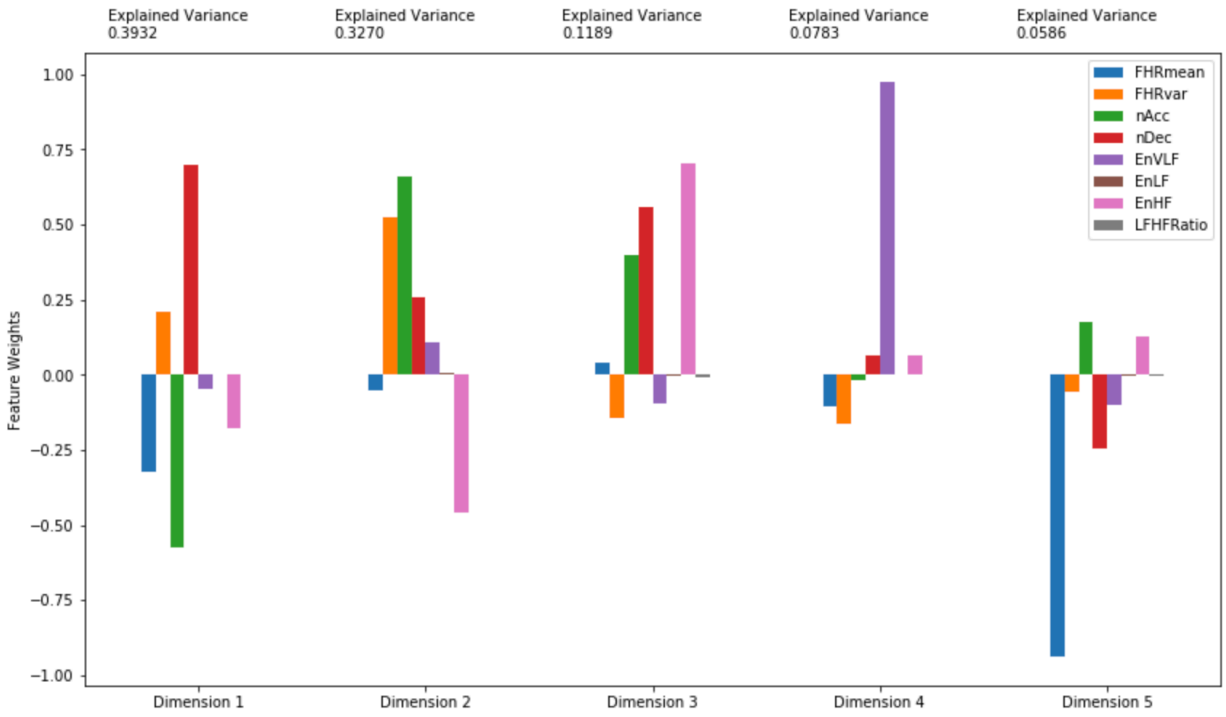


Figure 5: PCA reduction feature weights for the first 5 dimensions.

A 2D biplot, and a 3D scatter plot in Figure 6 show projections of the first two and three principal component feature space. The 2D projection shows that both stages have a similar amount of spread in each of these dimensions. Both distributions have a similar shape, but Stage II tends to show components that are slightly higher in value. Larger Dimension 1 scores in Stage 2 samples may indicate a higher nDec/nAcc ratio compared to Stage I. Looking at the 3D scatter plot, potential clusters do not seem to be very pure in terms of labor stage. The relationship between the binary labels looks to be more complex. In particular when there is a low amount of relative change in heart rate described by dimension 1 scores below 0, samples from Stage 2 show a larger amount of variability not regulated by parasympathetic activity. However, dimension 1 scores between 0 and 0.2 samples from both stages tend to have a noticeable overlap. When dimension 1 is above 0.2, Stage 2 tends to show the opposite relationship seen when dimension 1 was negative. Dimension 3 does not show any clear differentiation based on labor stage, and reinforces the observation that potential clusters may not have much purity. Based on these plots I see most meaningful splits for the proposed classifiers to occur along the dimension 1 and 2 projection.

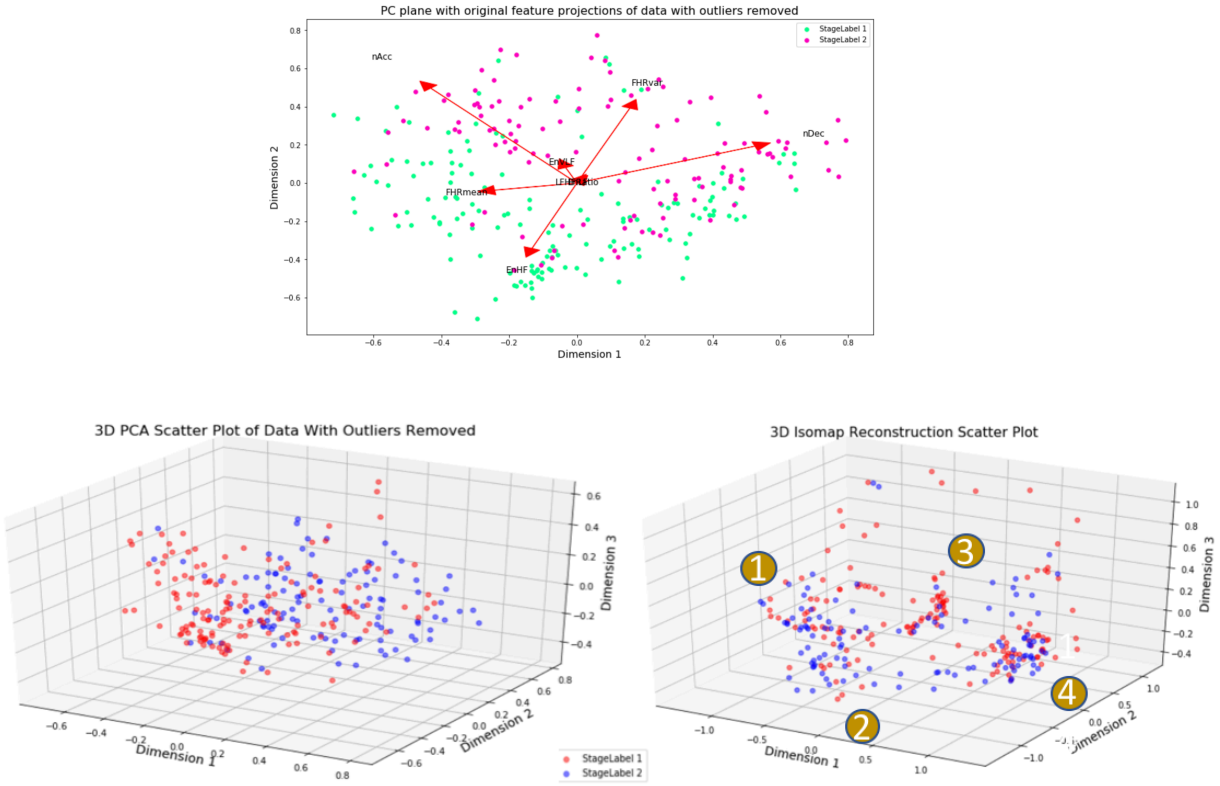


Figure 6: Projections of the reduced PCA and Isomap presentations of the dataset in 2D and 3D.

In order to attempt to visually find an optimal number of clusters for the k-means and hierarchical models, an isomap dimensionality reduction was calculated. Isomap components can help describe the dataset by computing a nonlinear geometrical representation of the data. A 3D scatterplot of the first three isomap components can be seen in Figure 6. The reconstruction error of the first three components was 0.14 (Explained variance of 0.86) which is 2% greater than the first three components of the PCA reduction. The 3D isomap scatter plot shows four major clusters. Clusters labeled 1 and 2 on the 3D scatter plot are more spread out than 3 and 4. Only cluster 3 looks to be dominated by samples from Stage 1. Whereas the others look to be fairly impure. Clusters 1 and 2 looks like they share a boundary, however most samples look like they can be easily categorized into one of the four clusters. This observation gives some confidence in the ability to find clusters with a positive silhouette coefficient.

ALGORITHMS AND TECHNIQUES

Based upon the above outlier detection, PCA visualization, and Isomap visualizations the important characteristics of the data can be summarized within two or three dimensions. In relation to model performance the above exploration can be summarized in the following points:

1. Outlier detection is necessary to gain representative weights of frequency domain features in dimensionality reduction.
2. Potential clusters have varying widths.
3. Potential clusters may have varying degrees of purity based on labor stage.
4. The dataset shows regions of significant overlap between classes. This results in the need to control for overfitting to ensure model stability.
5. Increasing the number of PCA components does not provide a significantly different potential decision boundary.

Unsupervised Model:

As a reminder the goal of the clustering is to find compact groups of samples that also group the data by labor stage. Furthermore, if clusters show a high purity, they can potentially be used as a new feature set to stabilize a classifier. Given these goals k-means and ward hierarchical clustering become the most obvious methods to try. Both would optimize variance (k-means globally, ward stepwise) helping to minimize the impact of any single sample. This would be particularly important to keep cluster shapes generally constant when training a supervised classifier on a training subset. During optimization of the number of clusters, k-means models would use 15 initializations. When evaluating a particular k-means model, 100 initializations would be conducted. Since the dataset is relatively small the default 300 iterations would be used.

Supervised Model:

Based on the previous work showing different entropy measures varying between labor stage, it seems that a decision tree model would naturally be the first model to try. In addition this model would be easy to interpret and relate back to observations seen in the literature due to its ability to be easily visualized. This model can also handle dummy coded variables which would be needed if cluster labels were to be used as inputs. Furthermore, the difference in the 2D and 3D scatterplots do not show much potential improvement in classification in a higher dimensional space. Since there is significant overlap between classes in the reduced pca space, a support vector model with a nonlinear kernel may only contribute to overfitting. Due to the small sample size a neural network model may not be able to adequately learn feature weights, or would tend to overfit to the dataset. In order to address potential instabilities inherent in decision tree models, a random forest model was also evaluated. Both models would be optimized and validated using GridSearchCV (details described in the implementation section).

BENCHMARK METRICS

Unsupervised Model:

- **Silhouette Coefficient (SI):** A score > 0 would indicate clusters that are on average further apart than the width of an average cluster.

- **Homogeneity Score (HS):** A score greater than > 150/267 (0.562) would indicate the clusters with a homogeneity greater than the class balance.

- **Purity Score (PS):** After disregarding outliers a purity > 150/267 (0.562) would guarantee average cluster purity greater than the class balance.

- **Adjusted Rand Index (ARI):** A score > 0 would correspond to clustering that is more similar to the binary labels better than chance.

- **Adjusted Normalized Mutual Information Score (ANMI):** A score > 0 would indicate a gain in knowledge based on a paritular clustering.

Supervised Model:

- The main benchmark metric would be to have an accuracy better than 0.562 which would be the accuracy if all samples were predicted to belong to Stage I.

IMPLEMENTATION

Note: Evaluation results in some tables are truncated to 3 decimal places.

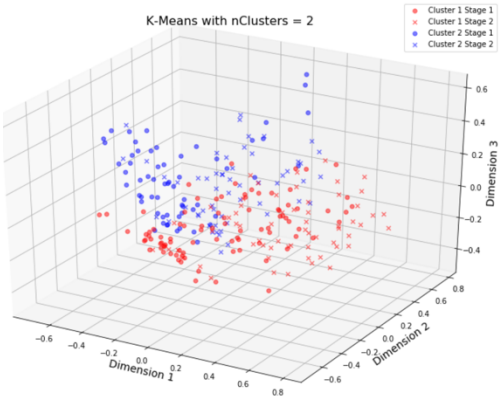
Note: All of the below algorithms were trained on the reduced dataset of the first 3 PCA components.

Unsupervised Clustering:

K-means:

Initial Model**:**

To begin with a k-means model was constructed using nClusters = 2 which was equivalent to the number of labels used for classification. The model was calculated using the entire dataset. 100 initializations were used to help avoid local optimums. 300 iterations were used for each initialization in order to help ensure a stable solution was found. Figure 7 shows a 3-D scatterplot of the clustering results, and lists results of the evaluation metrics. All evaluation metrics except purity_score were used from the sklearn library. The purity_score metric implementation was duplicated from **jhumigas'** GitHub.



Metric	Score
SI	0.384
HS	0.003
Purity	0.561
ARI	0.003
ANMI	0.001

Figure 7: 3D projection of the resulting K-means clusters (left). Table of evaluation scores (right).

Refinement:

The initial model evaluation metrics resemble those that would be expected based on a random clustering. Visually due to the large spread of these clusters there is little opportunity to localize smaller groups of samples originating from either labor stage. In order to optimize the number of clusters, a range of k-means models were constructed with nClusters varying from 2 to 25. Each model was constructed with 15 initializations, 300 iterations, and were fit to the entire dataset. Each model had the same evaluation metrics calculated. These evaluation metrics were compiled in the scatter plots shown in Figure 8. Based on the plots it seems that nClusters = 4 would result in clusters with the highest ARI and ANMI scores. This model produced a purity score which was the beginning of a fairly flat plateau in purity. Models with 2 or 3 clusters did have higher silhouette scores but lower ARI and ANMI scores. Models with > 4 clusters did have a higher homogeneity, but would see large drops in the silhouette coefficient.

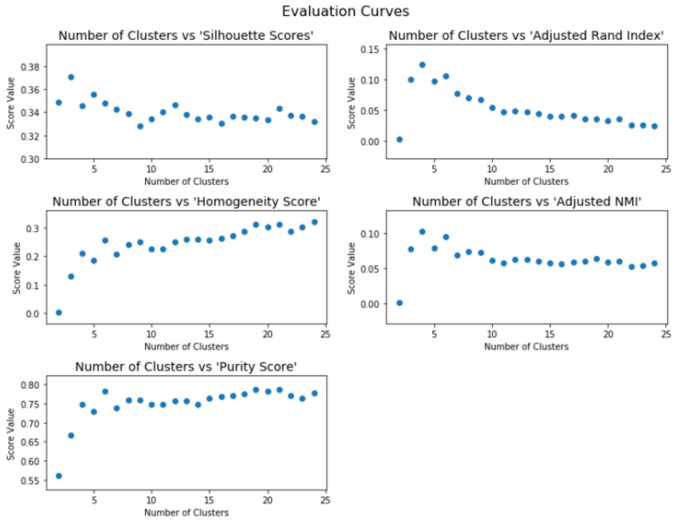


Figure 8: Evaluation curves for k-means models ranging from 2 to 25 clusters.

Thus nClusters = 4 was used in the final model. The model was fit using 100 initializations and 300 iterations. Tables 2-3 show the final models evaluation metrics, and each clusters composition by labor stage. Figure 9 shows a 3D scatterplot of the four clusters.

Tables 2-3: Evaluation Results of the final k-means model with 4 clusters.

Metric	Score
SI	0.345
HS	0.212
Purity	0.749
ARI	0.124
ANMI	0.102

	nInCluster	nStage1	nStage2	percStage1	percStage2
Cluster 1	46	40	6	0.869565	0.130435
Cluster 2	79	64	15	0.810127	0.189873
Cluster 3	67	17	50	0.253731	0.746269
Cluster 4	75	29	46	0.386667	0.613333

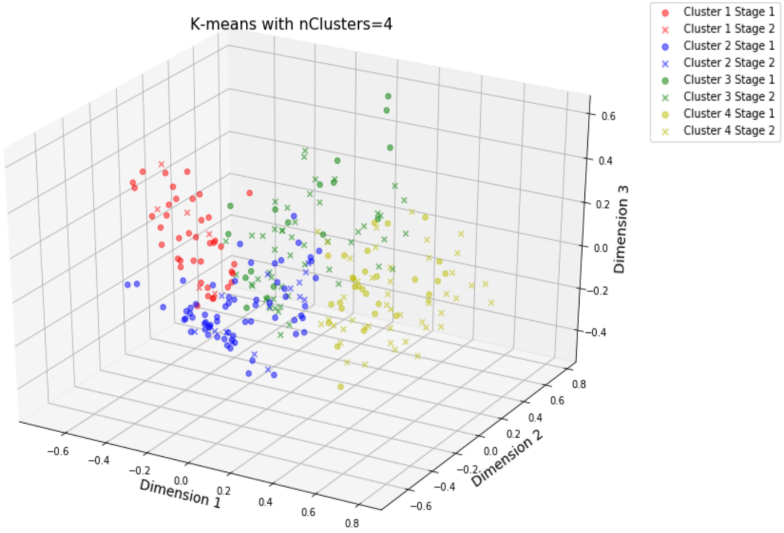


Figure 9: 3D scatterplot of the k-means model with 4 clusters.

Hierarchical Clustering:

Initial Model:

A hierarchical model using ward linkage was created using the scipy library. A full dendrogram for all of the samples can be seen in Figure 10. The dendrogram shows that the two major splits are not even.

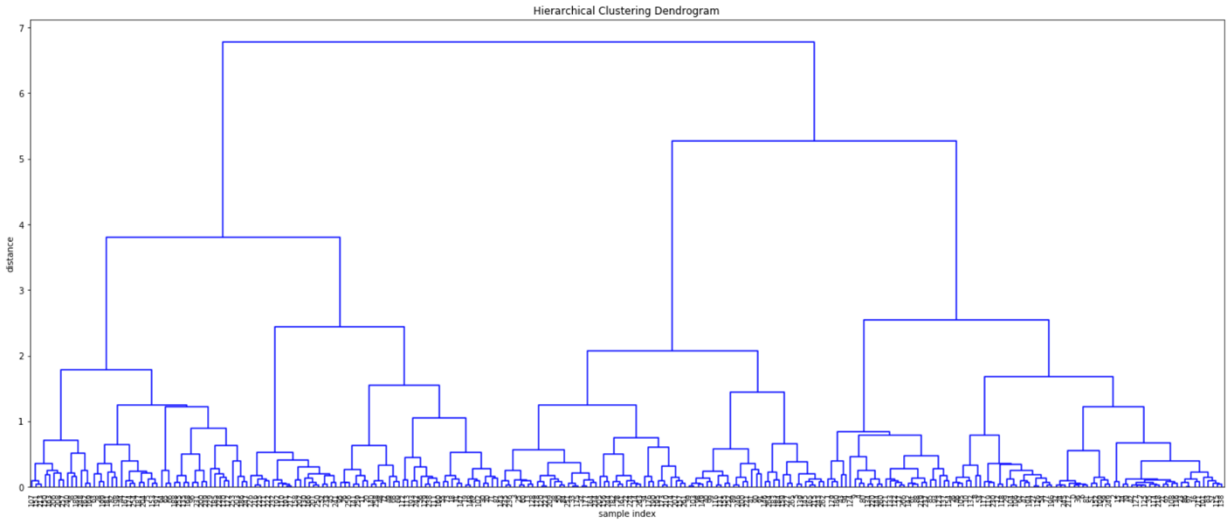
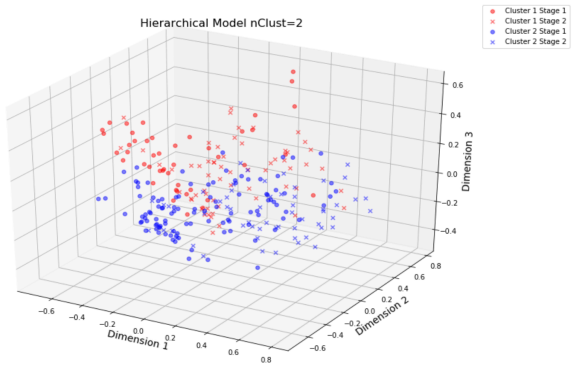


Figure 10: Ward clustering dendogram.

As with k-means an initial model was evaluated with only two clusters. The resulting evaluation metrics and 3D scatterplot can be seen in Figure 11.



Metric	Score
SI	0.334
HS	0.034
Purity	0.617
ARI	0.051
ANMI	0.031

Figure 11: 3D scatterplot of Ward Hierarchical model with 2 splits (left), and a summary of evaluation metrics (right).

Refinement:

Although better than the initial k-means model, the low homogeneity, ARI, and ANMI scores are still very close to what would be expected by a random clustering scheme. Visually the clusters closely resemble those found by the two cluster k-means model. Results from the same optimization process as conducted with the k-means model can be seen in Figure 12. In this case a model with 4 or 6 clusters seemed to be the optimum. A model with 6 clusters seemed to have higher purity, homogeneity, and slightly higher NMI. However a model with 4 clusters had a higher ARI, and a significantly larger silhouette coefficient. Since results were mixed and in order to ensure enough samples would be included in each cluster the simpler 4 cluster solution was chosen. In addition, fewer clusters would help limit overfitting due to the small dataset size.

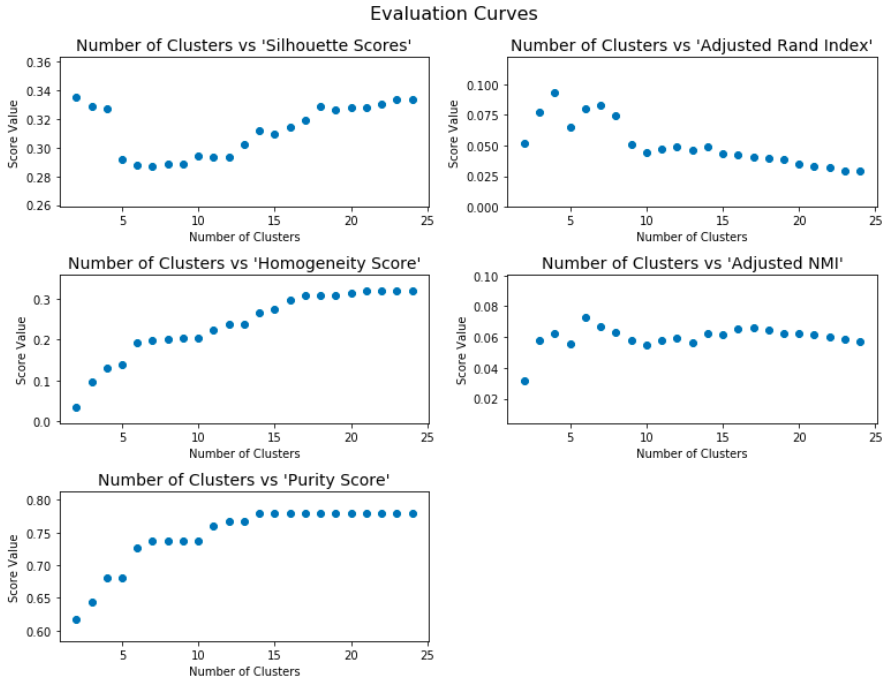


Figure 12: Evaluation curves for ward hierarchical models ranging from 2 to 25 clusters.

The 4 cluster 3D scatterplot, cluster makeup, and evaluation metrics can be seen in Figure 13 and Tables 4-5 respectively.

Tables 4-5: Evaluation metrics and composition of Ward Hierarchical model with 4 clusters.

Metric	Score
SI	0.326
HS	0.132
Purity	0.681
ARI	0.092
ANMI	0.062

	nInCluster	nStage1	nStage2	percStage1	percStage2
Cluster 1	49	12	37	0.244898	0.755102
Cluster 2	56	33	23	0.589286	0.410714
Cluster 3	73	33	40	0.452055	0.547945
Cluster 4	89	72	17	0.808989	0.191011

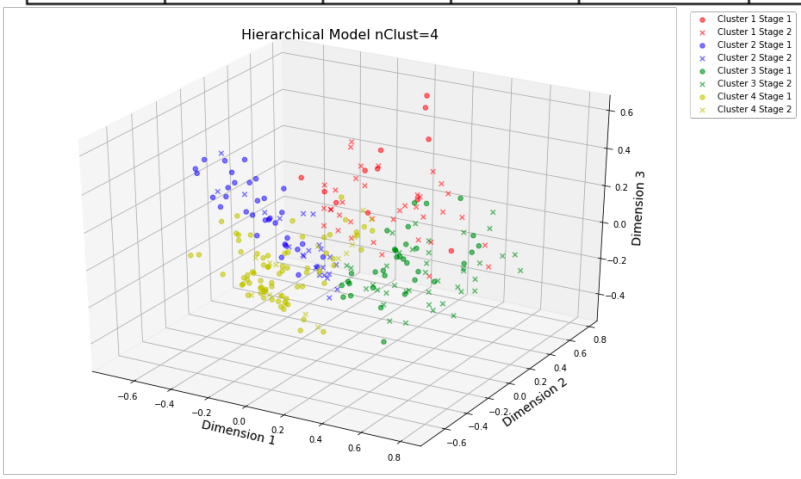


Figure 13: 3D Scatterplot of the 4 cluster Ward Hierarchical Model.

Unsupervised Model Discussion:

Both refined clustering models outperformed better than the set benchmarks on SI, ARI, ANMI, and purity. Although both models had ARI and ANMI scores >0 they were also <0.125. Thus the clustering was only slightly better than chance when compared to the "true" binary labels. This similarity would not be enough on its own to make generalizations about fetal treatment. This result could however be used to justify further research into the limiting factors of clustering FHR data. Since ANMI incorporates homogeneity, the low homogeneity scores of 0.212 for k-means and 0.132 for ward clustering were expected. This low homogeneity can be attributed to each labor class appearing in all of the four clusters. Furthermore, both models contain at least one cluster where the distribution is relatively even. These clusters would skew the homogeneity scores closer to 0.

The k-means model outperformed the hierarchical model in all evaluation criteria, most notably in PS, HS, and ANMI. Visually both models produce clusters in the same general regions. However the hierarchical clustering tends to follow the data's geometrical distribution more closely. This is most noticeable in cluster 1 in the k-means model, and cluster 2 in the ward model. Although both models produce a clustering purity greater than the class balance, the significant class overlap near the center of the coordinate system limits purity and homogeneity. This is reflected in relatively even cluster makeup of cluster 2 and 3 in the ward model, and 4 in the k-means model. Looking at the full dendogram and ward evaluation curves the clusters may contain further subgroups. However due to the small sample size, these subclusters would be made up of few data points and would make unreliable generalizations. With 4 clusters in each model the smallest k-means cluster contained 46 samples, and the smallest ward cluster contained 49 samples. This would correspond to both models containing >17% of the dataset in each cluster. If the supervised classifiers were to be fit using a training set which is 75% of the entire dataset, on average about 34 samples should be expected to come from the smallest clusters.

Descriptive Statistics of each k-means cluster can be found in the Appendix. Clusters 1 and 2 are both made up >80% of samples originating from Stage 1. While Cluster 3 is made up ~75% of Stage 2 samples. Looking at the feature statistics in the reduced PCA space the following observations can be made:

- Cluster 3 predominantly containing samples from Stage 2 differs from clusters 1 and 2 predominantly containing samples from Stage 1, by having a much higher and positive mean in dimension 2. This could be interpreted as changes in FHR within Stage 2 samples to be much greater relative to changes in parasympathetic activity represented by EnHF.
- Cluster 3 also differs from clusters 1 and 2 in the standard deviation of dimensions 1 and 3. Both of these dimensions have a standard deviation of 0.195 and 0.229 in cluster 3 respectively, as compared to the standard deviations of clusters 1 and 2 in these dimensions of around ~0.15.
- Cluster 4 which contains 61% Stage 2 samples shows a large positive dimension 1 component. This would correspond to a larger number of nDec relative to nAcc.

These observations agree with the projects background describing Stage 2 of labor to show more variance in heart rate due to more regular uterine contractions. Furthermore, Cluster 4 although only slightly favoring the Stage 2 class also agrees with the clinical observation the

nDec should tend to increase as the fetus movement increases. If uterine contractions are driving fetal responses (indirectly impacting heart rate), then changes in heart rate should be much greater than any neurological control internal to the fetus. This agrees with the positive dimension 2 component of cluster 3.

Supervised Classification:

Decision Tree Model:

An optimized Decision Tree model was constructed using the three dimensional PCA reduction. The data was split into a shuffled training and testing sets with a test size of 0.25. Optimal decision tree parameters were found using GridSearchCV. Table 6 shows the various parameter combinations explored.

Table 6: GridSearchCV parameter grid.

Parameter	Grid
criterion	gini, entropy
min_samples_split	2, 10, 20
max_depth	2, 3, 4
min_samples_leaf	1, 5, 10
max_leaf_nodes	2 ,3, 4, 5, 6

Models were cross validated using 4 shuffled folds on the training set. The resulting best estimator was used to find the accuracy on the unseen testing set. This process was conducted 10 times in order to get different samples within the training and testing sets. Descriptive statistics on the resulting accuracy from the 10 trials can be seen in Table 7. A summary of accuracy, precision, recall, F1-score, and confusion matrices of the worst and best decision tree performance on their respective test sets can be found in Tables 8-9.

Tables 7-9: Decision Tree performance using reduced PCA features.

Metric	Worst Tree	Best Tree
Accuracy	0.641	0.791
Percision	0.531	0.8
Recall	0.653	0.689
F1-Score	0.293	0.370
10 Trial Acc Mean	0.729	
10 Trial Acc Std.	0.052	

Worst Tree Confusion Matrix				
	Prediction			
Actual		True	False	**Total**
	True	17	9	26
	False	15	26	41
	Total	32	35	67

Best Tree Confusion Matrix				
	Prediction			
Actual		True	False	**Total**
	True	20	9	29
	False	5	33	38
	Total	25	42	67

Decision Tree Results:

Based on the above results the decision tree model surpassed the benchmark result of an accuracy of >0.562 even in its worst case. Comparing the best to the worst performing tree it is evident that most of the increase in accuracy is largely due to an increase in precision relative to the increase in recall. The mean accuracy of 0.729 is 29.7% better than the accuracy if all samples were predicted to originate from Stage 1.

Random Forest Model:

A similar procedure was conducted to construct an optimized Random Forest model using the three dimensional PCA reduction. Table 10 shows the various parameter combinations inputted into the GridSearchCV.

Table 10: Random Forest GridSearchCV parameter grid.

Parameter	Grid
criterion	gini, entropy
min_samples_split	2, 10, 20
max_depth	2, 3, 4
min_samples_leaf	1, 5, 10
max_leaf_nodes	2 ,3, 4, 5, 6

Descriptive statistics on the resulting accuracy from the 10 trials can be seen in Table 11. A summary of accuracy, precision, recall, F1-score, and confusion matrices of the worst and best performing random forests on their respective test sets can be found in Tables 12-13.

Tables 11-13: Random Forest performance using reduced PCA features.

Metric	**Worst Tree**	**Best Tree**
Accuracy	0.641	0.791
Percision	0.531	0.75
Recall	0.653	0.692
F1-Score	0.293	0.359
10 Trial Acc Mean	0.728	
10 Trial Acc Std.	0.042	

Worst Forest Confusion Matrix				
	Prediction			
Actual		True	False	**Total**
	True	17	9	26
	False	15	26	41
	Total	32	35	67

Best Forest Confusion Matrix				
	Prediction			
Actual		True	False	**Total**
	True	18	8	26
	False	6	35	41
	Total	24	43	67

Random Forest Results:

The random forest model surpasses the established benchmark accuracy of 0.562 by an average of 29.5%. Similar to the decision tree behavior, the best random forest model improves compared to the worst forest due largely to improved precision. As expected the random forest has a smaller standard deviation (0.042 compared to 0.052) in accuracy compared to the decision tree model. This can be attributed to the bootstrapping and ensemble nature of random forests. The mean accuracy is within a standard deviation of the mean of the decision tree acciracu, and thus will not be considered to be significantly different.

Decision Tree and Random Forest Models using cluster labels as inputs:

Based on the results of the optimized hierarchical and k-means models, the k-means model outperforms the hierarchical model in all evaluation criteria. Thus, a k-means model would be used during this refinement step. The k-means results show clusters that are on average not overlapping (i.e. >0 Silhouette Coefficient). Thus, grouping samples into a cluster should result in better model generalization by showing less variation in model performance on different training/testing sets. Geometrically this can be thought of as occurring due to an increase in margin from a decision boundary by moving samples into discrete points on a coordinate system. Since the average purity across clusters is within a standard deviation of the mean accuracies of the models fit on the reduced pca features I do not expect there to be an improvement in model accuracy.

The dataset was first split into training and test subsets with a test size of 0.25. The training set was then clustered using a k means model with 4 clusters, 100 initializations, and 300 iterations. This model was then used to predict the respective cluster labels of the test set. Clustering was not fit on the entire dataset in order to limit the introduction of model bias for the subsequent classification evaluation. The k-means training and test sets were then dummy coded. These dummy coded features were then used as the sole inputs into the respective classifiers using the same GridSearchCV parameters. The procedure was repeated 10 times as done with the previous models.

Descriptive statistics on the resulting accuracy from the 10 trials for each model can be seen in Tables 14-15. A summary of accuracy, precision, recall, F1-score, and confusion matrices of the worst and best performing decision tree and random forest on their respective test sets can be found in Tables 16-17.

Tables 14-17: Classifier performance using k-means cluster labels as inputs.

Metric	Worst Tree	Best Tree
Accuracy	0.641	0.791
Percision	0.518	0.676
Recall	0.560	0.884
F1-Score	0.269	0.383
10 Trial Acc Mean	0.726	
10 Trial Acc Std.	0.048	

Metric	Worst Forest	Best Forest
Accuracy	0.641	0.791
Percision	0.518	0.676
Recall	0.560	0.884
F1-Score	0.269	0.383
10 Trial Acc Mean	0.726	
10 Trial Acc Std.	0.048	

Worst Tree Confusion Matrix				
	Prediction			
Actual		True	False	**Total**
	True	14	11	25
	False	13	29	42
	Total	27	40	67

Worst Forest Confusion Matrix				
	Prediction			
Actual		True	False	**Total**
	True	14	11	25
	False	13	29	42
	Total	27	40	67

Best Tree Confusion Matrix				
	Prediction			
Actual		True	False	**Total**
	True	23	3	26
	False	11	30	41
	Total	34	33	67

Best Forest Confusion Matrix				
	Prediction			
Actual		True	False	**Total**
	True	23	3	26
	False	11	30	41
	Total	34	33	67

Results of Decision Tree and Random Forest models using cluster label inputs:

Using categorical cluster inputs resulted in both the decision tree and random forest models to have identical performance. This is expected since both models classify samples based on information gain resulting in similar splits for categorical inputs. The standard deviation of accuracy of these models was less than the original decision tree model, but larger than the random forest model. The accuracy of the models fit on the clustering inputs was within a standard deviation to their pca fit counterparts. The models fit on the clustering inputs tended to improve accuracy largely due to an increase in recall as opposed to percision.

Supervised Classification Discussion:

In order to explore the behavior of the models fit on the reduced pca features and too understand why they may prefer improving precision, decision boundaries for both models were plotted. The decision boundaries for each model's worst and best performing fit were projected into the three orthogonal planes. These projections can be seen in Figure 14. Decision boundaries from both models show that there is little ability to make predictions based on projections on dimension 2 and 3 plane. This is in agreement with the earlier observation that samples from Stage 2 tend to shift from negative dimension 2 scores to positive dimension 2 scores along dimension 1. Thus, when looking at the dimension 2 and 3 projection, there is little opportunity to reduce entropy. The best model fits tend to improve performance by creating decision boundaries that better localize stage 2 samples which can be seen most clearly in the dimension 1 and 3 projections. By focusing on the edge cases both models are able to more precisely detect Stage 2 samples while also lowering the number of misclassifications of Stage 1. The best random forest fit shows a more complex decision boundary in the dimension 1 and 2, and dimension 1 and 3 projections. The dimension 1 and 2 projection boundary again reinforces the idea that the models tend to localize edge cases. In the dimension 1 and 3 projection the model tries to capture more Stage 2 samples, but due to increasing class overlap around dimension 1 scores of 0 there is also an increase in misclassification. Looking back at the k-means scatterplot with 4 clusters and cluster makeup, we would expect models fit on cluster labels to improve recall mostly through the ability to localize Stage 1 samples into clusters 1 and 2.

Based on these results all of the models surpass the benchmark accuracy metric, however there does not seem to be a clearly dominant model across the optimizations. When considering the implications on fetal health, both precision and recall are equally important. A false positive or negative of either labor stage could result in clinical procedures or treatments being undertaken at inopportune times. Similar to the unsupervised results, these classification results can be used to help justify the need for further research. They however should not be used to make reliable generalizations about FHR data in general. While an accuracy of ~0.72 is better than the benchmark, it is not convincing enough to make judgements to drive clinical treatment. If a high precision model was to be chosen, I would choose the random forest model fit on the reduced pca components. This is because the model is more stable compared to the decision tree with a similar level of accuracy. If a high recall model was to be chosen, I would choose the decision tree model fit on k-means clustering labels. Although both the decision tree and random forest models preformed identically when fit on the cluster labels, the decision tree representation would be simpler. To make the most informed decisions regarding a particular

sample I would look at the output of both of these high precision and high recall models, and other factors seen physically during the labor process.

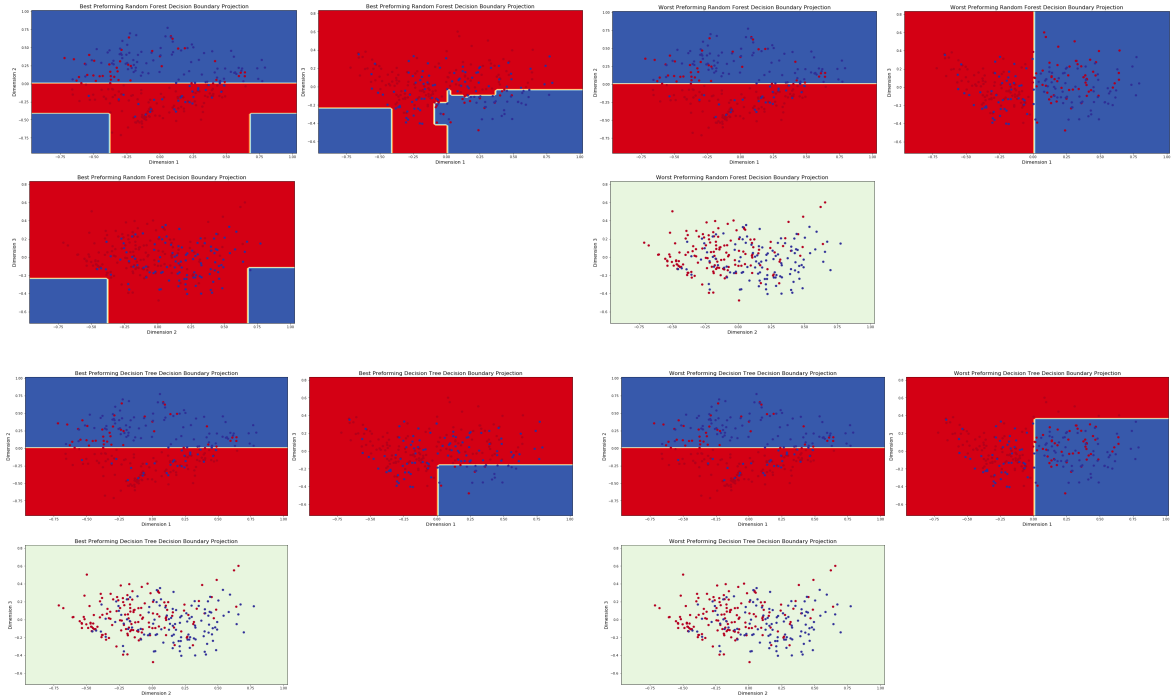


Figure 14: Best and worst random forest (above), and decision tree (below) decision boundaries. Stage 1 (red) and Stage 2 (blue). Projections showing no splits are a neutral greenish-grey color.

SUMMARY:

This work composed of 5 major stages:

1) Data import and feature construction
2) PCA Dimensionality Reduction
3) K-means and ward clustering
4) Supervised decision tree and random forest classification on reduced pca features.
5) Supervised decision tree and random forest classification on k-means cluster labels.

The most difficult part of this project was to determine how to best replicate input features used by clinicians and previous machine learning work. The ACOG clinical guidelines defined input features which required long 10 minute segments. This would result in excluding many subjects, and an unneeded over sampling frequency domain features. This was especially detrimental on sample size when considering the large number of outliers needing to be removed.

The most interesting part of this project was using visualizations of the models using the reduced pca feature space in order to help describe model behavior.
Based on the results seen during clustering and classification it is clear that there is some preference of heart rate data to group subjects by labor stage. In order to get even more insight

into the characteristics of each labor stages heart rate characteristics the following improvements could be done:

1) Gather more data for training and testing in order to apply more complex models such as neural networks, and get better generalizations of results.
2) Gain more precision and recall by classifying subjects only in clusters which show purity greater than a certain threshold.
3) Classify subjects into a labor stage if both the high precision random forest, and the high recall decision tree vote similarly.

References

1. Alfirevic, Z., Devane, D., & Gyte, G. M. (2006). Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. *Cochrane Database Syst Rev*, *3*(3).
2. American College of Obstetricians and Gynecologists. (2009). ACOG Practice Bulletin No. 106: Intrapartum fetal heart rate monitoring: nomenclature, interpretation, and general management principles. *Obstetrics and gynecology*, *114*(1), 192.
3. Baldzer, K., Dykes, F. D., Jones, S. A., Brogan, M., Carrigan, T. A., & Giddens, D. P. (1989). Heart rate variability analysis in full-term infants: spectral indices for study of neonatal cardiorespiratory control. *Pediatric research*, *26*(3), 188.
4. Beard, R. W., Filshie, G. M., Knight, C. A., & Roberts, G. M. (1971). The significance of the changes in the continuous fetal heart rate in the first stage of labour. *BJOG: An International Journal of Obstetrics & Gynaecology*, *78*(10), 865-881.
5. Chan, H. L., Fang, S. C., Ko, Y. L., Lin, M. A., Huang, H. H., & Lin, C. H. (2006). Heart rate variability characterization in daily physical activities using wavelet analysis and multilayer fuzzy activity clustering. *IEEE Transactions on Biomedical Engineering*, *53*(1), 133-139.
6. Electronic Fetal Heart Monitoring: Healthwise Medical Information on eMedicineHealth. (n.d.). Retrieved January 27, 2018, from https://www.emedicinehealth.com/script/main/art.asp?articlekey=129181
7. Ferrario, M., Signorini, M. G., & Magenes, G. (2009). Complexity analysis of the fetal heart rate variability: early identification of severe intrauterine growth-restricted fetuses. *Medical & biological engineering & computing*, *47*(9), 911-919.
8. Freeman, R. K. (2002). Problems with intrapartum fetal heart rate monitoring interpretation and patient management. *Obstetrics & Gynecology*, *100*(4), 813-826.
9. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]; 2000 (June 13).
10. Granero-Belinchon, C., Roux, S. G., Abry, P., Doret, M., & Garnier, N. B. (2017). Information Theory to Probe Intrapartum Fetal Heart Rate Dynamics. *Entropy*, *19*(12), 640.
11. Inbarani, H. H., Banu, P. N., & Azar, A. T. (2014). Feature selection using swarm-based relative reduct technique for fetal heart rate. *Neural Computing and Applications*, *25*(3-4), 793-806.
12. Langer, B., Carbonne, B., Goffinet, F., Le Gouëff, F., Berkane, N., & Laville, M. (1997). Fetal pulse oximetry and fetal heart rate monitoring during stage II of labour. 1. *European Journal of Obstetrics and Gynecology and Reproductive Biology*, *72*(1), S57-S61.
13. Phongsuphap, S., Pongsupap, Y., Chandanamattha, P., & Lursinsap, C. (2008). Changes in heart rate variability during concentration meditation. *International journal of cardiology*, *130*(3), 481-484.
14. Spilka, J. (2013). Complex approach to fetal heart rate analysis: A hierarchical classification model. *Czech Technical University, Faculty of Electrical Engineering, Prague*, 35-47.
15. Romano, M., Iuppariello, L., Ponsiglione, A. M., Improta, G., Bifulco, P., & Cesarelli, M. (2016). Frequency and time domain analysis of foetal heart rate variability with traditional indexes: a critical survey. *Computational and mathematical methods in medicine*, *2016*.
16. Spilka, J., Leonarduzzi, R., Chudácek, V., Abry, P., & Doret, M. (2016, November). Fetal Heart Rate Classification: First vs. Second Stage of Labor. In *Proceedings of the 8th International Workshop on Biosignal Interpretation, Osaka, Japan* (pp. 1-3).
17. Stages of labor and birth: Baby, it's time! (2016, June 22). Retrieved January 26, 2018, from https://www.mayoclinic.org/healthy-lifestyle/labor-and-delivery/in-depth/stages-of-labor/art-20046545
18. Václav Chudáček, Jiří Spilka, Miroslav Burša, Petr Janků, Lukáš Hruban, Michal Huptych, Lenka Lhotská. Open access intrapartum CTG database. *BMC Pregnancy and Childbirth* 2014 **14**:16.
19. Williams, K. P., & Galerneau, F. (2003). Intrapartum fetal heart rate patterns in the prediction of neonatal acidemia. *American Journal of Obstetrics & Gynecology*, *188*(3), 820-823.

APPENDIX

Descriptive Statistics of the nClusters = 4 k-means model.

```
Cluster 0 (1 of 4) Statistics
```

	Dimension 1	Dimension 2	Dimension 3	Cluster_0	Cluster_1	Cluster_2	Cluster_3
count	46.000000	46.000000	46.000000	46.0	46.0	46.0	46.0
mean	-0.453118	-0.032575	0.088399	1.0	0.0	0.0	0.0
std	0.127898	0.170085	0.157066	0.0	0.0	0.0	0.0
min	-0.718754	-0.397337	-0.173837	1.0	0.0	0.0	0.0
25%	-0.558087	-0.165757	-0.048259	1.0	0.0	0.0	0.0
50%	-0.426750	-0.047627	0.065038	1.0	0.0	0.0	0.0
75%	-0.356304	0.094326	0.203034	1.0	0.0	0.0	0.0
max	-0.221470	0.355884	0.381477	1.0	0.0	0.0	0.0

	FHRmean	FHRvar	nAcc	nDec	EnVLF	EnLF	EnHF	LFHFRatio
count	46.000000	46.000000	46.000000	46.000000	46.000000	46.000000	46.000000	46.000000
mean	0.715554	0.315661	0.546547	0.005318	0.236778	0.004598	0.668322	0.012825
std	0.108734	0.127693	0.198656	0.036072	0.143373	0.002402	0.156566	0.006092
min	0.477586	0.023089	0.218104	0.000000	0.000294	0.000683	0.309004	0.003340
25%	0.657861	0.234627	0.345687	0.000000	0.174054	0.002958	0.544642	0.009991
50%	0.705873	0.334398	0.563791	0.000000	0.215712	0.004163	0.661597	0.011229
75%	0.790123	0.405875	0.706973	0.000000	0.295016	0.006229	0.804669	0.015107
max	0.933891	0.533570	1.000000	0.244651	0.605018	0.011763	0.974135	0.032944

Cluster 1 (2 of 4) Statistics

	Dimension 1	Dimension 2	Dimension 3	Cluster_0	Cluster_1	Cluster_2	Cluster_3
count	79.000000	79.000000	79.000000	79.0	79.0	79.0	79.0
mean	-0.013350	-0.353627	-0.027017	0.0	1.0	0.0	0.0
std	0.161612	0.141565	0.164079	0.0	0.0	0.0	0.0
min	-0.360857	-0.711266	-0.392157	0.0	1.0	0.0	0.0
25%	-0.134089	-0.458573	-0.130656	0.0	1.0	0.0	0.0
50%	-0.066642	-0.354455	-0.043926	0.0	1.0	0.0	0.0
75%	0.128871	-0.239959	0.059314	0.0	1.0	0.0	0.0
max	0.324561	-0.056603	0.503303	0.0	1.0	0.0	0.0

	FHRmean	FHRvar	nAcc	nDec	EnVLF	EnLF	EnHF	LFHFRatio
count	79.000000	79.000000	79.000000	79.000000	79.000000	79.000000	79.000000	79.000000
mean	0.608041	0.300211	0.030369	0.151918	0.067120	0.004846	0.659502	0.013940
std	0.126123	0.130650	0.075990	0.197884	0.136081	0.002514	0.142757	0.005897
min	0.311480	0.000000	0.000000	0.000000	0.000000	0.000767	0.272183	0.000000
25%	0.524716	0.213732	0.000000	0.000000	0.000062	0.003137	0.563135	0.010195
50%	0.611082	0.298835	0.000000	0.000000	0.006129	0.004231	0.670171	0.013408
75%	0.682385	0.381674	0.000000	0.244651	0.059165	0.006056	0.740937	0.016069
max	1.000000	0.651299	0.218104	0.686821	0.583082	0.013793	1.000000	0.032422

Cluster 2 (3 of 4) Statistics

	Dimension 1	Dimension 2	Dimension 3	Cluster_0	Cluster_1	Cluster_2	Cluster_3
count	67.000000	67.000000	67.000000	67.0	67.0	67.0	67.0
mean	-0.158807	0.390824	-0.013257	0.0	0.0	1.0	0.0
std	0.195005	0.154948	0.229136	0.0	0.0	0.0	0.0
min	-0.532877	0.111509	-0.404591	0.0	0.0	1.0	0.0
25%	-0.291432	0.272826	-0.199775	0.0	0.0	1.0	0.0
50%	-0.197532	0.399207	0.013469	0.0	0.0	1.0	0.0
75%	0.002891	0.487121	0.117572	0.0	0.0	1.0	0.0
max	0.252329	0.774243	0.600861	0.0	0.0	1.0	0.0

	FHRmean	FHRvar	nAcc	nDec	EnVLF	EnLF	EnHF	LFHFRatio
count	67.000000	67.000000	67.000000	67.000000	67.000000	67.000000	67.000000	67.000000
mean	0.613843	0.652526	0.612851	0.250547	0.171059	0.007675	0.356259	0.015809
std	0.121152	0.104628	0.147831	0.241330	0.149928	0.006794	0.162916	0.007757
min	0.256056	0.437756	0.345687	0.000000	0.006603	0.001847	0.014113	0.000000
25%	0.549442	0.589829	0.506422	0.000000	0.054144	0.003180	0.246449	0.010353
50%	0.613962	0.660239	0.563791	0.244651	0.136601	0.004626	0.337544	0.014119
75%	0.692493	0.721468	0.691374	0.387762	0.229314	0.010199	0.488270	0.019890
max	0.843977	0.927919	0.926491	0.846352	0.592359	0.028386	0.705408	0.042555

Cluster 3 (4 of 4) Statistics

	Dimension 1	Dimension 2	Dimension 3	Cluster_0	Cluster_1	Cluster_2	Cluster_3
count	75.000000	75.000000	75.000000	75.0	75.0	75.0	75.0
mean	0.433842	0.043330	-0.013917	0.0	0.0	0.0	1.0
std	0.153275	0.168383	0.189090	0.0	0.0	0.0	0.0
min	0.122555	-0.267052	-0.475490	0.0	0.0	0.0	1.0
25%	0.332847	-0.095927	-0.161471	0.0	0.0	0.0	1.0
50%	0.416114	0.023317	-0.031264	0.0	0.0	0.0	1.0
75%	0.536104	0.156654	0.128626	0.0	0.0	0.0	1.0
max	0.792143	0.455537	0.402419	0.0	0.0	0.0	1.0

	FHRmean	FHRvar	nAcc	nDec	EnVLF	EnLF	EnHF	LFHFRatio
count	75.000000	75.000000	75.000000	75.000000	75.000000	75.000000	75.000000	75.000000
mean	0.407200	0.569874	0.048724	0.576510	0.164147	0.007257	0.409368	0.015825
std	0.156563	0.128101	0.112800	0.191849	0.159540	0.007209	0.177194	0.007483
min	0.117633	0.288101	0.000000	0.244651	0.000027	0.000000	0.032116	0.003537
25%	0.268148	0.481004	0.000000	0.387762	0.037099	0.002913	0.278845	0.010547
50%	0.395060	0.552104	0.000000	0.568061	0.110362	0.004016	0.389736	0.013763
75%	0.542240	0.638990	0.000000	0.710386	0.303943	0.007085	0.541860	0.018713
max	0.805420	0.881351	0.436209	1.000000	0.605886	0.028870	0.760695	0.032944