Interpretable Clustering for Patient Phenotyping using Advanced Machine Learning Models

Roy S. Zawadzki, Saman Parvaneh

Edwards Lifesciences, Irvine, USA

Abstract

Essential to personalized treatment in cardiology is identifying patient phenotypes and interpreting their association with post-interventional outcomes, such as in transcatheter aortic valve replacement. After obtaining the phenotypes via unsupervised clustering methods, one method is to use machine learning to predict cluster membership in order to interpret patient characteristics defining each cluster. In previous work, we used decision trees due to their ease of interpretability, which is limited in capturing complex relationships in the data. We mitigate this issue by integrating SHapley Additive exPlanations (SHAP) values with advanced machine learning algorithms. To demonstrate our approach, we identified six distinct patient clusters using single-center data from 581 TAVR patients using k-means and applied various classifiers to classify cluster memberships, with Light Gradient Boosting performing best (F1-score: 0.879). SHAP values were used to interpret cluster memberships, revealing key features influencing outcomes. For example, creatinine level and annulus area were significant predictors for high-risk phenotypes. This SHAP-based approach balances model complexity with interpretability, providing robust, detailed insights into patient phenotypes.

1. Introduction

Understanding heterogeneity in the characteristics of patients with cardiovascular disease and the impact on post-intervention outcomes is critical for developing safe and effective personalized treatments. Because patients are multi-faceted, grouping patients with only a few variables (e.g., age and sex) may paint an incomplete picture of their health status. On the other hand, considering many variables simultaneously can lead to largely uninterpretable and spurious conclusions. Patient phenotyping strikes a balance between these two sides of the continuum by combining multi-modal data sources (e.g., demographic, echocardiograph, and computed tomography) via unsupervised clustering into a set of distinct subpopulations. However, a key challenge in patient phenotyping is ensuring that the found subpopulations are clinically meaningful.

In Zawadzki, Johnson, and Parvaneh (2022) (ZJP2022), we introduced an interpretable patient phenotyping framework that utilizes decision trees after k-means to profile each cluster using a classification based on cluster membership [1]. Yet, while decision trees are one of the more interpretable machine learning (ML) algorithms, they may not adequately capture the complex, non-linear relationships between a patient's features and their cluster membership. Furthermore, they may overfit the current data, leading to spurious results. This naturally motivates using more advanced algorithms, such as gradient-boosted learning and random forests. Nevertheless, these algorithms comprise significantly more complex architecture, leading to the common judgment that they are largely "black box."

To address the trade-off between model complexity and interpretability, we propose using SHapley Additive exPlanations (SHAP) values for interpreting patient phenotypes [2]. SHAP values provide a model-agnostic measure of feature importance, offering insights into how each predictor influences the likelihood of cluster membership via the odds ratio (OR). This method allows for using a wide array of ML algorithms while maintaining interpretability, increasing the flexibility in modeling choices for data scientists who wish to use this approach for interpretable clustering. More importantly, by leveraging SHAP values, we can better understand the nuanced relationships between patient characteristics, outcomes, and discovered phenotypes, potentially leading to more effective personalized treatment strategies. In this work, we detail our proposed SHAP-based method and revisit the analysis conducted in ZJP2022 phenotyping transcatheter aortic valve replacement (TAVR) patients.

2. Methods

We use single-center data collected in Germany (n = 581) on TAVR patients and their outcomes [3]. The data contains different modalities, including demographics, medical history, echocardiography, and computed tomography. The data can be accessed online and may be used under a Creative Commons 4.0 license. More details about the dataset can be found in Pollari et al. [3]. Furthermore, details on variable selection and imputation can be found in ZJP2022. Briefly, we removed the less clinically meaningful feature in pairwise correlations of over 0.6 and then imputed missing data using random

forests. The final set of variables is as follows: age (years), body mass index (kg/m²), creatinine levels (mg/dl), creatinine clearance (mL/min), Euroscore II, ejection fraction (%), pulmonary hypertension (mmHg), aortic valve (AV) delta mean (mmHg), AV effective orifice area (cm²), annulus area (cm²), distance from annulus to right coronary artery (RCA) (mm), distance from annulus to left CA (LCA) (mm), valve oversizing, eccentricity index (mm³), total calcium in AV (mm³), total calcium in left ventricular outflow tract (LVOT) (mm³), RCC calcium in LVOT (mm³).

Using the above variables in ZJP2022, six clusters were found using k-means via three methods: locating the elbow in the sum of squares error, maximizing the Bayesian information criterion (BIC), and maximizing the silhouette score. The summary statistics for each feature are presented in Table 1. Cluster 5 had a significantly higher proportion of those who experienced myocardial injury during operation (75 percent) compared to the other clusters, which varied from 20 to 35 percent.

In this work, we first predict cluster memberships as a function of the 17 variables using a set of several classifiers using the "pycaret" Python package [4]. Specifically, we fit logistic regression (LR), gradient boosting classifier (GBC), catboost classifier (CBC), decision tree, random forest (RF), light gradient boosting (LGB), and extra trees (ET). Using 10-fold cross-validation, we selected the model with the best F1 score and re-fit it with the entire dataset. Then, we extracted the SHAP scores from this model and investigated the values to extract insights regarding the key factors behind cluster membership.

SHAP values provide a unified approach to interpreting the output of machine learning models rooted in cooperative game theory. They quantify the contribution of each feature to a specific prediction, ensuring that the interpretation is fair and comprehensive. The calculation of SHAP values involves several steps. First, the model's expected output, or baseline prediction, is computed over the entire dataset. Then, for a given instance, the SHAP value for each feature is determined by considering all possible subsets of features that exclude the feature in question. A feature's contribution to the model's prediction is evaluated by measuring the change in the model's output when the feature is added to each subset. This change is known as the marginal contribution. The SHAP value for a feature is the weighted average of its marginal contributions across all possible subsets and observations. For binary outcomes, the SHAP value can be directly interpreted as log odds. As such, we will examine the factors that identify one cluster versus the others (i.e., one vs. all). The "shap" Python package was used to compute and visualize the SHAP values [5].

3. Results

Among the models tested, LGB performed best with respect to a 10-fold cross-validated F1-Score of 0.879, though many algorithms also performed well (Table 2). Notably, the more sophisticated algorithms vastly outperformed the decision tree, with the F1-score LGB improving 0.217 (33 percent) compared to the decision tree.

Table 2: Machine Learning Classifier Performance for Identifying Cluster Membership, Ranked by F1-Score

Classifier	F1-score	AUC
Light Gradient Boosting	**0.879**	0.879
Extra Trees	0.875	0.883
Gradient Boosting	0.872	**0.888**
Catboost	0.871	0.884
Random Forest	0.847	0.875
Logistic Regression	0.816	0.866
Decision Tree	0.662	0.711

The SHAP summary plot for LGB is shown in Figure 1, which is constructed by summing the average SHAP scores for each feature to identify each cluster. The size of the bar of a particular color indicates how impactful the features were for classifying that cluster. For example, the annulus area was the most impactful feature for predicting cluster membership, particularly for Clusters 1 and 2. Meanwhile, creatinine clearance was overwhelmingly impactful for Cluster 3 and less so for Cluster 1.

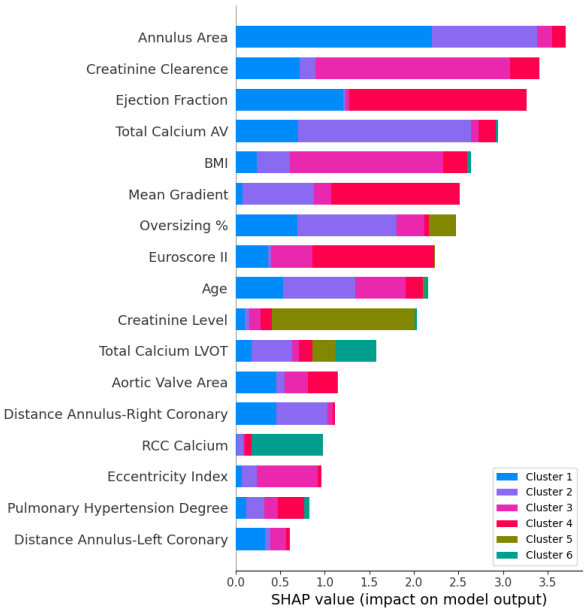

Figure 1: SHAP Summary Plot for Features Determining Cluster Membership

Interestingly, creatinine level is the primary determinant for Cluster 5 but not the others. Cluster 5 is the cluster that had a significantly higher rate of myocardial injury. We can examine the individual SHAP values for the baseline

creatinine level, which appears to be the main defining factor of this cluster, using the partial dependence plot. This plot is constructed by examining how the SHAP values for classifying into this cluster change as the feature values within the dataset change, holding all the other features constant. Figure 2 shows that when the creatinine value is roughly above 4 mg/dL, the data point will likely be classified into Cluster 5. This cluster, therefore, identifies those with kidney dysfunction.

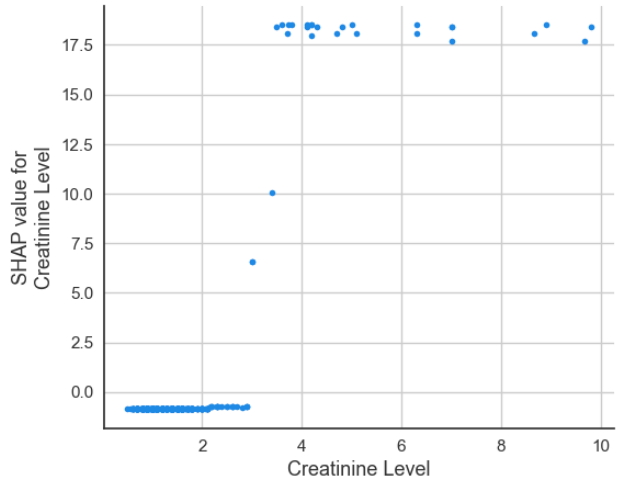

Figure 2: Partial Dependence Plot for the Impact of Creatinine for Predicting Cluster 5

The importance of each data point for each feature simultaneously can be examined using the swarm plot, which additionally colors each point by the value of the feature. This gives us more insight into what is driving cluster membership. Figure 3 gives the summary plot for Cluster 4, which had a lower rate of myocardial injury but was not statistically significant. As the top feature, we can see that subjects with lower ejection fraction are more likely to be in this cluster. Furthermore, a similar conclusion can be made when Euroscore II is high.

4. Discussion

In this work, we have expanded our developed interpretable clustering framework in ZJP2022 to accommodate a wider array of classifiers to explain the factors that drive each cluster. This is primarily driven by the SHAP score, which is model-agnostic and can give comparable conclusions across models of varying complexity. Importantly, using SHAP scores ameliorates the possible trade-off between model fit and interpretability. As such, the prediction performance of LGB, the best algorithm in terms of F1 score, was notably better than that of the decision tree. This fact naturally leads to an increase in confidence in results. In summary, our framework has three steps that are model agnostic: (i) choose an unsupervised clustering algorithm to form clusters and identify cluster membership, (ii) train a classifier to predict cluster membership using the features used for clustering, and (iii) interpret clusters with ML and SHAP values with descriptive statistics on key outcomes.

Our analysis has several advantages over ZJP2022 in making conclusions on patient phenotypes in post-TAVR outcomes. Firstly, as done in ZJP2022, a multi-class decision tree offers less interpretability on which features drive specific cluster membership rather than a one-vs.-all approach. Often, a subset of the clusters does not map to any meaningful difference in clinical outcomes. Therefore, we are not concerned about examining those specific clusters in the one-vs.-all approach.

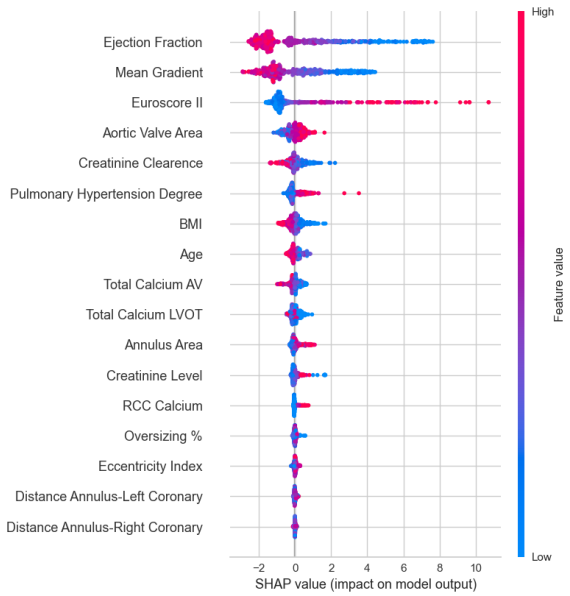

Figure 3: Swarm Plot for Features Impactful for Predicting Cluster 4

Next, though we may visualize decision trees to understand better what drives individual cluster membership as we did in ZJP2022, there are two primary limitations. First, we must restrict the depth of the tree to one that can be manageably interpreted, risking oversimplification of the interaction between features. With SHAP values, we do not have to restrict model complexity. Second, the cut-offs for splitting may further overfit the training data and, thus, may be misleading. With SHAP values, the swarm plot presents the features continuously as opposed to via binarization.

Finally, the SHAP approach enhances the replicability of results as the interpretation of cluster membership is consistent across models. Thus, one could try an array of machine learning models across multiple datasets and would be able to straightforwardly compare the conclusions made.

Despite the advantages of SHAP values for interpreting clusters, our analysis has some limitations. Firstly, the findings in this study are limited to that of a single center in Germany and, thus, should be interpreted in this context. Second, the computation of SHAP values can be resource-intensive, especially for complex models and large datasets, which may limit their scalability. We may additionally be concerned about how sampling variability may affect SHAP values. Future work may use techniques like bootstrapping or cross-fitting to construct confidence intervals on SHAP values based on recent work for inference on SHAP values [6].

Table 1: Descriptive statistics of clusters where all numbers are reported as mean (SD). All variables are significant at the 0.003 level (Bonferroni adjustment). Reproduced from ZJP2022 with permission from authors.

Variable	Overall	Cluster 1 (n = 193)	Cluster 2 (n = 123)	Cluster 3 (n = 123)	Cluster 4 (n = 104)	Cluster 5 (n = 25)	Cluster 6 (n = 13)
Age	81.7 (6.1)	84.3 (4.6)	84.3 (4.6)	77.5 (6.3)	80.6 (5.5)	74.8 (7.0)	81.2 (5.5)
BMI	27.1 (4.8)	26.2 (4.1)	25.3 (3.4)	31.6 (5.3)	25.7 (3.3)	27.6 (5.1)	24.7 (3.6)
Creatinine Levels	1.5 (1.0)	1.3 (0.4)	1.3 (0.4)	1.2 (0.3)	1.5 (0.5)	5.5 (2.0)	1.3 (0.4)
Creatinine Clearance	45.0 (19.6)	37.6 (12.4)	44.1 (13.0)	67.4 (20.2)	41.2 (14.3)	14.3 (6.3)	41.7 (14.9)
Euroscore II	0.1 (0.1)	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.2 (0.1)	0.1 (0.0)	0.1 (0.1)
Ejection Fraction	52.6 (12.9)	59.9 (8.4)	53.2 (10.8)	55.0 (10.6)	37.1 (11.5)	49.3 (8.2)	46.1 (15.4)
PHT Measurement	50.6 (13.8)	50.9 (13.6)	48.6 (12.4)	47.6 (12.8)	54.1 (14.9)	58.3 (15.9)	49.1 (14.7)
AV Delta Mean	44.6 (14.7)	47.6 (12.5)	53.2 (10.8)	42.2 (12.4)	30.0 (10.5)	45.1 (14.0)	48.1 (14.0)
AV Effective Orifice Area	0.7 (0.2)	0.6 (0.1)	0.7 (0.1)	0.8 (0.1)	0.8 (0.2)	0.7 (0.1)	0.6 (0.2)
Annulus Area	4.7 (0.9)	3.9 (0.6)	5.3 (0.8)	4.9 (0.8)	4.9 (0.9)	5.1 (0.8)	4.6 (1.1)
Distance from Annulus to RCA	15.6 (3.8)	14.1 (3.2)	17.1 (4.0)	16.2 (4.0)	15.7 (3.7)	16.4 (3.3)	14.5 (3.3)
Distance from Annulus to LCA	13.6 (3.0)	12.5 (2.6)	14.2 (3.2)	14.4 (3.1)	13.6 (2.8)	14.0 (2.4)	13.0 (2.6)
Valve Oversizing	13.8 (15.7)	20.0 (17.1)	3.9 (11.0)	14.4 (13.9)	13.5 (14.1)	11.4 (11.5)	14.9 (21.9)
Eccentricity Index	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)
Total Calcium in AV	839.2 (572.4)	596.7 (313.0)	1435 (607.5)	757.7 (396.4)	604.6 (418.9)	796.4(443.8)	1529.1 (1167.8)
Total Calcium in LVOT	73.2 (123.3)	53.6 (91.1)	116.2 (145.8)	52.5 (82.9)	39.9 (68.2)	66.0 (77.3)	431.5 (292.0)
RCC Calcium in LVOT	5.9 (21.0)	2.9 (7.8)	3.8 (8.6)	2.0 (7.0)	4.2 (9.0)	1.8 (5.7)	128.5 (40.4)

References

[1] R. S. Zawadzki, T. Johnson and S. Parvaneh, "Patient Phenotyping Using Interpretable Clustering to Study Clinical Outcomes in TAVR Patients," 2022 Computing in Cardiology (CinC), Tampere, Finland, 2022, pp. 1-4, doi: 10.22489/CinC.2022.345.

[2] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.

[3] F. Pollari, W. Hitzl, F. Vogt, et al. "Aortic Valve Calcification as A Risk Factor for Major Complications and Reduced Survival After Transcatheter Replacement," Journal of cardiovascular computed tomography, vol. 14, no. 4, pp. 307-313, 2020.

[4] Ali, M. PyCaret: An open-source, low-code machine learning library in Python. 2020; Available from: https://www.pycaret.org

[5] SHAP: SHapley Additive exPlanations. Available at: https://pypi.org/project/shap/.

[6] Li X, Zhou Y, Dvornek NC, Gu Y, Ventola P, Duncan JS. Efficient Shapley Explanation For Features Importance Estimation Under Uncertainty. Med Image Comput Comput Assist Interv. 2020;12261:792-801. doi: 10.1007/978-3-030-59710-8_77. Epub 2020 Sep 29. PMID: 34308439; PMCID: PMC8299327.

Address for correspondence:
Saman Parvaneh
1 Edwards Way, Irvine, CA, USA 92614
parvaneh@ieee.org