

A fast and interpretable prediction system for the Site Of Origin (SOO) of Outflow Tract Ventricular Arrhythmias (OTVAs)

Jana Casau, Marc Mallol, Carla Pairés, Oriol Pont

“Computational Models and Data Science for Biomedical Engineering” Course
Universitat Pompeu Fabra (UPF)
Barcelona, Spain

Abstract—Outflow tract ventricular arrhythmias (OTVAs) require precise localization for effective catheter ablation, yet current methods rely heavily on expert interpretation of 12-lead ECGs. We developed a lightweight, two-stage system of machine learning models that uses basic demographic data and directly extracted ECG features. The goal is to predict, in Part 1, whether an OTVA originates from the left ventricular outflow tract (LVOT) or the right ventricular outflow tract (RVOT), and, in Part 2, to further distinguish between the right coronary cusp (RCC) and the aortomitral commissure sub-regions. Each task employs gradient-boosted trees optimized via stratified cross-validation and leverages SHAP values for transparent, case-level explanations. For Part 1, on a dataset of over 25,000 confirmed OTVA cases, our best-performing model generalized well to unseen data, achieving a total accuracy of 86.70%; while a more lightweight version with only 17 decision trees only lagged 5 to 10% behind in most metrics while being 10 times more lightweight and faster to run. For Part 2, due to a very reduced training set of only 12 patients, the model struggled to generalize. By combining high accuracy with fully interpretable outputs, this approach holds promise for guiding preprocedural planning and enhancing clinician confidence in SOO predictions.

Index Terms—Ventricular tachycardia, Outflow tract ventricular arrhythmias, Site of origin, Machine learning, XGBoost, Interpretable machine learning

I. INTRODUCTION

Ventricular tachycardia (VT) is a potentially life-threatening cardiac arrhythmia that can occur even in structurally normal hearts. In such cases, there is an overtaking of the sino-atrial activation, that is manifested as a premature ventricular contractions (PVC), which disrupts the heart’s normal rhythm [1].

VT is a serious condition that may lead to Sudden Cardiac Death (SCD) if left untreated. Among idiopathic ventricular tachycardias, which are those that occur in ventricles and are not linked to any detectable structural heart disease, Outflow Tract Ventricular Arrhythmias (OTVAs) are the most common subtype[2].

Currently, two main treatments exist for managing OTVAs: antiarrhythmic drugs and radiofrequency ablation (RFA). This study focuses on RFA, a procedure in which targeted energy is used to burn and destroy the myocardial tissue responsible for initiating the arrhythmia [3]. A critical factor for a successful RFA intervention is the accurate localization of the arrhythmia’s Site of Origin (SOO). Identi-

fying whether the SOO is located in the left ventricular outflow tract (LVOT) or the right ventricular outflow tract (RVOT) is essential, as it determines the appropriate vascular access route for the ablation catheter. Early and accurate localization can improve procedural success rates, reduce intervention time, and minimize patient risk [4].

There are numerous potential SOOs where OTVAs may arise. However, this study focuses specifically on two anatomical locations: the right coronary cusp and commissure. Anatomically, the aortic valve consists of three cusps: the right coronary cusp (RCC), the left coronary cusp (LCC), and the non-coronary cusp (NCC), as illustrated in Figure 1. The commissure refers to the junction between two cusps, for example the left-right commissure lies between the LCC and RCC [5].

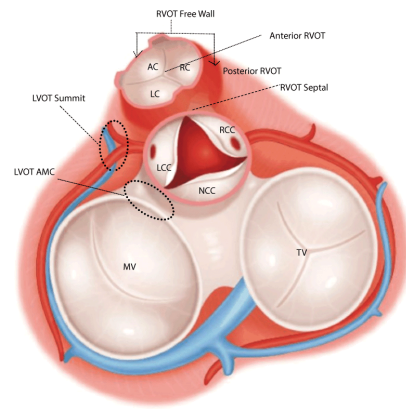


Fig. 1: Detailed view of the right coronary cusp (RCC), highlighting its structural features and anatomical significance in the aortic valve.

The goal of this study is twofold. First, we aim to classify whether the SOO of the arrhythmia is in the LVOT or RVOT using a combination of demographic information and ECG-derived features. Second, the approach is more specific since it is sought to further localize the SOO by distinguishing between two specific anatomical sites: the RCC and the commissure. Both tasks leverage machine learning models trained on clinical and electrophysiological data to enhance pre-procedural planning and support more targeted interventions.

II. METHODS

To understand how the processed data was obtained, it is important to first outline the multimodal system used in this study. The process begins with what we call Model A, which takes as input a dataframe containing ECG data from the selected patients. This model performs a dimensionality reduction using Principal Component Analysis (PCA), helping to identify and retain only the most relevant features from the raw ECG signals.

These selected ECG features are then combined with the patients' demographic data, which has been preprocessed separately. The merged dataset is passed into Model B, which is responsible for the first classification task: distinguishing between LVOT and RVOT origins. This forms the focus of Part 1 of the project.

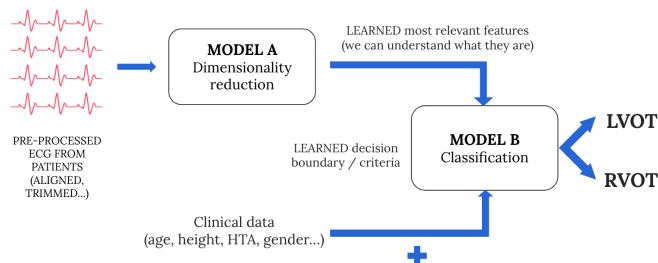


Fig. 2: System architecture for Part 1, illustrating the two-stage machine learning approach.

For Part 2, we refine the dataset further by filtering out any patients whose outcomes are not classified as either RCC or COMMISSURE. Since we are only interested in these two outcomes at this stage, all others are excluded. The resulting data is then used as input for Model C, which focuses on classifying between RCC and COMMISSURE outcomes.

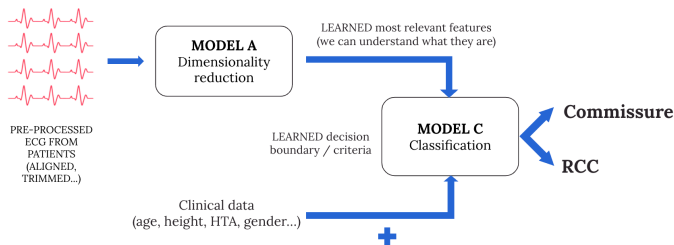


Fig. 3: System architecture for Part 2, showing the modularity of the system and how it builds upon the structure established in Part 1.

The following sections describe each data processing and modeling step in detail.

A. Data Preprocessing

a) Demographic data:

Preprocessing began with cleaning and organizing the dataset from 'all_points_may_2024.pkl', which contained comprehensive patient-level information including demographics, clinical indicators, and ECG structure data. Each row corresponded to a unique patient and included fields such as patient ID, sex, age, hypertension (HTA), diabetes

(DM), smoking status, PVC transition, BMI, and others, along with a Structures column containing the ECGs. For the initial steps, only demographic and clinical features were considered, and the Structures column was excluded to simplify preprocessing focused on patient-level attributes.

Although demographic data was initially assumed to have limited value in predicting the site of origin (SOO) or its sub-regions, evidence from [6] showed that including these features significantly improved classification accuracy (from 67% to 89%) regardless of whether full ECGs or derived features were used. Based on this, demographic preprocessing was the first step done. All categorical variables such as sex, HTA, and smoker status were binarized, with values standardized to 0 and 1. The original dataset also stored many fields in nested lists, which were flattened to allow proper manipulation. Missing data in continuous features such as 'weight' and 'height' was handled using statistical imputation based on outlier presence. For height, where no outliers were detected, the mean value was used to fill in 33 missing entries. In contrast, the weight column had four outliers, so the median was used to impute 34 missing values. For BMI, where 89 entries were missing but both height and weight were available, BMI was recalculated using the standard formula. This resulted in a complete BMI column with no remaining gaps.

Label definition followed, targeting two classification problems: Model B for binary SOO classification (LVOT vs. RVOT) and Model C for sub-region classification within these chambers (RCC vs. COMMISSURE). The SOO_chamber and SOO columns in the original dataset included highly detailed anatomical labels, which were mapped to these broader categories. Two Excel files were used in this mapping process ('label2.xlsx'). The first served as the primary reference, using the SOO_Chamber and Region_Simplified columns to map raw anatomical labels to standard categories. The second file offered additional mappings and was used to confirm and supplement the primary mapping. These resources enabled a full relabeling of all 93 anatomical labels into two clean outcome columns: Outcome_B and Outcome_C. Once created, the original SOO_chamber and SOO columns were removed.

Further inspection addressed missing values in remaining categorical fields. For the sex column, three patients with missing data were dropped due to the potential impact of sex on ECG characteristics and the fact that these patients would have been excluded from model training regardless, due to having irrelevant SOO labels. Four patients missing HTA values were also removed. Similarly, eight patients missing PVC transition data were excluded. PVC transition is a particularly informative feature in distinguishing between LVOT and RVOT origins, with early transitions (e.g., V1-V2) suggesting LVOT and later ones (e.g., V4-V6) suggesting RVOT, as emphasized in the literature. Given the importance of this feature, missing values were not imputed. One additional patient missing a DM value was also dropped. This patient was not relevant for Model C and only marginally for Model B, and excluding them avoids introducing assumptions about their diabetes status.

b) ECG signal:

Once demographic preprocessing is complete, the ECG signals must be prepared for use as inputs to the machine learning models. The ECG data is stored in a column named Structures within the dataframe as dictionaries, where each entry corresponds to a patient and contains multiple ECG recordings from different anatomical regions and within each anatomical region, different ECG positions are provided. In other words, each patient has several ECGs captured from distinct positions.

The first step is to restructure this data into a unified dataframe in which each row corresponds to a single ECG, with separate columns for each lead. Each ECG is uniquely identified using the patient ID and the ECG position. Given that each ECG is 2.5 seconds long and sampled at 1000 Hz, each recording consists of 2500 samples.

Once the ECGs are unified, segmentation, and alignment are performed. Prior to alignment, filtering and segmentation are necessary to ensure clean and interpretable signals. To facilitate segmentation, the ECGs are temporarily down-sampled to 250 Hz. This step simplifies the detection of waveforms, but the original 1000 Hz signals are retained for model training to preserve full temporal resolution and signal variability.

Segmentation is performed using ‘modelos’ provided by the instructors. These models output binary arrays, where a value of 1 indicates the likely presence of a specific waveform or complex (P-wave, QRS complex, T-wave). To make this information human-readable and practical for further processing, the start and end sample indices of each detected waveform, along with its label, are stored. This data is sufficient to segment the ECGs and prepare them for alignment. For each ECG, a new dataframe is created that contains the intervals corresponding to each identified complex.

```
{
  "0": {
    "P": [[723, 851], [1392, 1519]],
    "QRS": [[892, 1023], [1564, 1676], [1945, 2088]],
    "T": [[0, 214], [1088, 1295], [1745, 1936], [2105, 2340]]
  }
}
```

Listing 1: Example JSON format given to the output from the ECG segmentation model showing the detected P-waves, QRS complexes, and T-waves with their corresponding sample indices.

Following segmentation, the ECGs are aligned. The alignment criterion is to position the final R-peak (the last QRS complex in the recording) exactly at the 2-second timestep. This step is critical for model A, which performs dimensionality reduction. Aligning the waveforms ensures the model focuses on morphological features rather than temporal variability and shifting.

To align the ECGs, the time difference between the detected R-peak and the 2nd second is computed and used to shift the signal accordingly. This process is applied to all ECGs. An outlier analysis is then conducted to identify ECGs that require excessive shifting. Any ECGs deemed outliers are removed from the dataset. Additionally, to ensure all

signals have the same length post-alignment, cropping is applied as needed. Originally, each ECG contained 2500 samples, after alignment and cropping, this number is reduced to 2353 samples. Consequently, the aligned R-peak appears at approximately 1.93 seconds instead of exactly 2.0 seconds.

At the end of this process, the ECGs are fully preprocessed and ready to serve as inputs for model A, which performs dimensionality reduction.

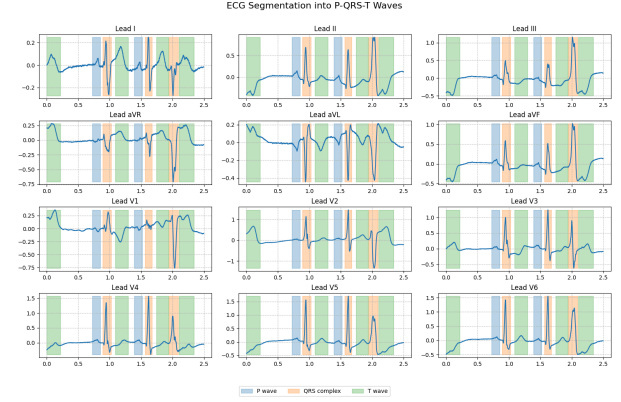


Fig. 4: Validation results of ECG segmentation before alignment, demonstrating the initial segmentation output on the validation dataset.

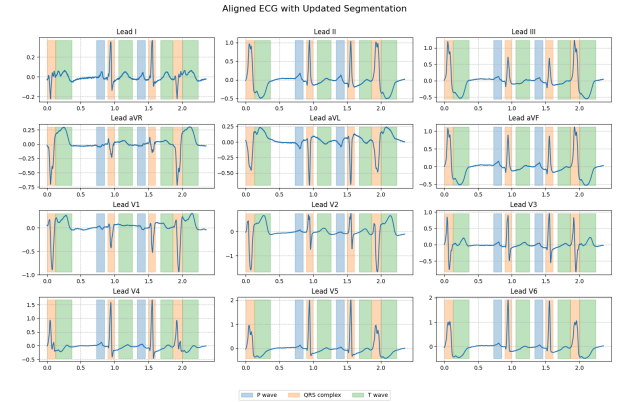


Fig. 5: Aligned ECG segmentation results, showing improved temporal alignment of cardiac cycles.

B. Dimensionality Reduction (Model A)

To ensure a more reliable and objective analysis, instead of manually selecting the most important features for distinguishing between LVOT and RVOT, we apply dimensionality reduction using Principal Component Analysis (PCA). This allows the algorithm to automatically identify the key features that contribute most to the classification task, instead of using subjective criteria, improving precision. To further improve interpretability, we apply Varimax rotation to the principal components. Varimax is an orthogonal rotation technique that redistributes the variance across components, making the loadings more distinct and sparse. This helps clarify which original features contribute most to each component, facilitating a more meaningful understanding of the underlying structure that differentiates LVOT from RVOT.

Applying dimensionality reduction is important for several reasons. First, it prevents overfitting, since working

with too many features can cause the model to fit noise rather than true patterns, reducing generalization to new data. Additionally, handling high-dimensional data is computationally expensive and complex, and the curse of dimensionality makes it difficult for models to learn meaningful relationships when the number of features is very large. Therefore, reducing dimensionality simplifies the problem, improves model performance, and optimizes computation.

C. Model training runs

Both demographic (17 features) and ECG feature data (200 features) were independently preprocessed and stored in separate dataframes. To enable model training, these dataframes were merged into a single dataset by retaining only patients present in both, identified by a common patient ID. This ensured consistency and prevented data mismatches.

To evaluate model performance in a robust and unbiased manner, the data was split into training, validation, and test sets using an 60/20/20 partition strategy based on patient-aware stratified sampling. Stratification preserved the distribution of the target variable across subsets, which was especially important given the class imbalance observed in some of the classification tasks. Each patient’s ECG recordings were assigned exclusively to one set to avoid data leakage and ensure generalizability. The resulting training set contained 15,285 samples from 101 patients; the validation set included 4,787 samples from 34 patients; and the test set comprised 4,910 samples from 34 patients. In the RVOT vs. LVOT classification task, the class distribution was 11,941 RVOT and 3,344 LVOT in the training set; 3,709 RVOT and 1,078 LVOT in the validation set; and 3,782 RVOT and 1,128 LVOT in the test set.

An initial model comparison phase evaluated a variety of classification model families, prioritizing those offering a balance between interpretability and inference efficiency. For each model family, hyperparameter tuning was conducted using grid search and performance was assessed on the validation/test sets. However, this exhaustive search approach proved computationally inefficient, and early results consistently indicated that gradient boosting models outperformed alternatives such as regular random forests. Consequently, the focus shifted exclusively to XGBoost due to its strong empirical performance, interpretability through tree-based structures and feature importance metrics, and rapid training and inference capabilities. A unified training pipeline was developed to support both Model B and Model C, each using their respective input dataframes. Feature selection was performed using SHAP values, computed from an initial XGBoost model trained on a balanced subset of the training data. The 50 most influential features were retained to reduce dimensionality and potential noise, improving training efficiency and model robustness.

Hyperparameter optimization was conducted using randomized search over a predefined parameter space per model. Each configuration was evaluated using five-fold cross-validation, ensuring stable estimates of model performance. Within each fold, multiple decision thresholds ranging from 0.01 to 0.99 were evaluated to identify the

threshold that maximized macro-F1 score on the validation set. This metric was chosen due to its ability to account for class imbalance by equally weighting performance across classes, unlike accuracy or AUC, which may be misleading in imbalanced settings. This threshold tuned on the validation set was then applied to the test set to evaluate final model performance and assess generalization capabilities.

For Model C, where class imbalance was particularly pronounced, SMOTE was applied during training to over-sample the minority class, enhancing the model’s ability to learn from underrepresented patterns. After identifying the best-performing hyperparameter configuration and decision threshold, the model was retrained on the entire training set. The optimal threshold was re-estimated using the validation set and then used for final evaluation on the test set. All models, parameters, and evaluation results were persisted on disk to facilitate reproducibility and further experimentation.

An additional comparison was conducted between a full XGBoost model and a constrained “Lite” variant for Model B. The Lite model was restricted to a reduced amount of maximum learned estimator trees of depth of 1, meaning each tree consisted of only a single (weighted) binary criteria. This is beneficial as constraining the model to use trees of depth 1 offers several practical advantages. This simplification reduces the risk of overfitting by limiting model complexity, making it more robust, especially in smaller or noisy datasets. It also improves interpretability, since each tree performs a single, easily understandable decision based on one feature threshold. Additionally, training and inference times are significantly reduced, which is valuable in real-time or resource-constrained environments. Incorporating such a “Lite” model into the analysis allows for a clearer understanding of how much predictive power is retained with minimal complexity, and whether strong performance can be achieved with a simpler, more transparent decision process.

For Task 2, a filtered version of the dataset was created by including only patients whose arrhythmia originated from the Right Coronary Cusp (RCC) or the Left-Right Commisure. This resulted in a dataset of 1,835 ECGs from (just) 22 patients. The entire training pipeline which consists of the data splitting, SMOTE application, SHAP-based feature selection, hyperparameter optimization, threshold tuning, and final evaluation was replicated for this binary classification task thanks to the modular design of the system.

III. RESULTS AND DISCUSSION

A. Model A: Dimensionality reduction

After performing PCA, the number of features is reduced from 28,236 to 200, while still preserving 95.58% of the total variance in the data. In Figure 6 it is displayed a heatmap representing the loadings (the contribution weights) of each ECG lead over time samples for the Varimax PCA performed. Red indicates positive loading, in other words higher contributions, blue negative loadings, and white means close to zero contributions. This figure clearly shows which parts

of the ECG signal (across time and across leads) contribute most strongly to this specific component.

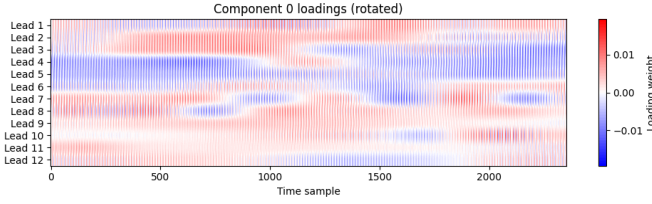


Fig. 6: First principal component analysis (PCA) visualization, highlighting the main pattern of variation in the ECG data.

In examining the principal components derived from the ECG signals, we can approximate the regions of the waveform that contribute most to variability across the dataset. Based on the alignment of the ECG with component 0, it appears that the greatest variability is associated with the QRS complex and extends into the QT interval, suggesting that differences in ventricular depolarization and repolarization dynamics are key sources of variation captured by the PCA. We can also see in Figure 7 the ECG signal with the relevant parts of the first principal component’s loadings highlighted. This visualization helps to understand the temporal location of the main sources of signal variability.

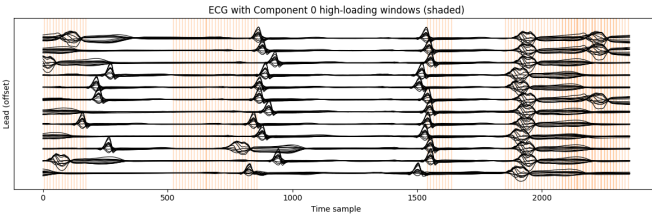


Fig. 7: Electrocardiogram (ECG) signal with highlighted components, showing the key cardiac electrical activity phases and their characteristic waveforms.

B. Model B: Classification of LVOT and RVOT

During the training process for Model B, we explored a wide range of hyperparameter configurations for the XGBoost model, including the number of estimators, tree depth, and regularization parameters, across more than 1,000 iterations. As part of this experimentation, we tested a version of the model with a restricted amount of estimators (less than 20) and a fixed depth of 1. Surprisingly, this extremely simple configuration achieved a comparable performance. After hyperparameter tuning, the best-performing setup under this constraint used only 17 such trees. While it was not the top-performing model overall, this Lite (restricted) version stood out for its simplicity and its ability to generalize well, making accurate predictions using just 17 decision rules.

TABLE I: CONFUSION MATRIX FOR MODEL B-LITE

Actual \ Predicted	LVOT	RVOT
LVOT	832	296
RVOT	540	3242

TABLE II: TEST METRICS FOR MODEL B-LITE

Metric	Value
Macro Accuracy	0.8297
Macro Precision	0.7614
Macro Recall	0.7974
Macro F1-Score	0.7757
ROC AUC	0.8833
PR AUC	0.9620

TABLE III: CLASSIFICATION REPORT FOR MODEL B-LITE

	Precision	Recall	f1-score	support
LVOT	0.6064	0.7376	0.6656	1128
RVOT	0.9163	0.8572	0.8858	3782
accuracy			0.8297	4910
macro average	0.7614	0.7974	0.7757	4910
weighted average	0.8451	0.8297	0.8352	4910

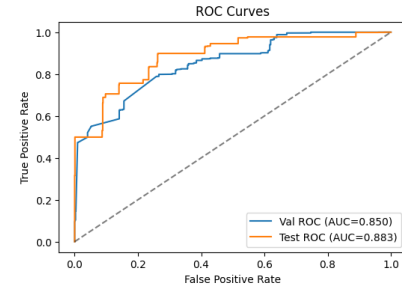


Fig. 8: Receiver Operating Characteristic (ROC) curve for the “Lite” version of Model B, showing its strong performance in distinguishing LVOT and RVOT origins.

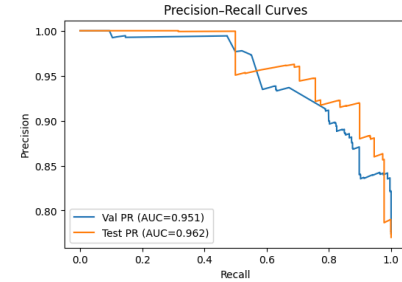


Fig. 9: Precision-Recall curve for the “Lite” version of Model B, illustrating its performance in classifying left ventricular outflow tract (LVOT) and right ventricular outflow tract (RVOT) origins.

In contrast, when training without such constraints the best-performing configuration required 163 trees with a depth of 1. This more complex model yielded slightly better predictive performance but at the cost of increased training time and reduced interpretability. The comparison between these two setups is noteworthy. While the full model delivers optimal results, the lite version achieves competitive performance. This demonstrates that the input features, selected by SHAP, carry a strong predictive signal, and that even shallow, low-complexity models can generalize effectively. Moreover, the simplicity of the lite model translates to faster inference and easier interpretation, making it especially valuable in

practical, real-world scenarios where model transparency and efficiency are critical.

The comparison between Model B and its compressed counterpart, Model B-Lite, highlights the trade-offs between model complexity and predictive performance. While Model B-Lite offers a dramatic reduction in model size (from 123.46 KB to 15.17 KB, an 88% reduction) and the number of estimators (from 163 to 17), this comes at a small measurable cost in accuracy and overall classification quality. Across all data splits, Model B consistently outperforms Model B-Lite, with a 6.87% drop in accuracy, 9.14% in F1-score, and 4.55% in ROC AUC in the compressed model. Most notably, LVOT classification performance deteriorates more sharply in Model B-Lite, as seen in a lower recall (from 87.60% to 75.30%), which may have downstream clinical relevance. However, despite the performance gap, Model B-Lite still maintains acceptable discrimination, suggesting it could be a viable option in resource-constrained or real-time deployment settings where model size and inference speed are critical. The slight shift in optimal decision thresholds (0.71 vs. 0.69) is minimal and unlikely to alter decision boundaries substantially, and indicates that both models learned similar probability distributions for the classes.

For Model B, the optimal threshold found was 0.71. The classification results demonstrate strong and balanced model performance in distinguishing between LVOT and RVOT origins. With a macro F1-score of 0.7962 on the test set, the model maintains a high level of generalization while effectively handling class imbalance. Notably, it achieves a recall of 82.4% for LVOT, indicating good sensitivity to the minority class, which is crucial in clinical settings. The overall ROC AUC of 0.89 and PR AUC of 0.96 further confirm the model’s ability to distinguish between classes with high confidence. These results suggest that the chosen threshold of 0.71 yields a well-calibrated classifier that performs reliably across both classes.

TABLE IV: CONFUSION MATRIX FOR MODEL B

Actual Predicted	LVOT	RVOT
LVOT	929	199
RVOT	589	3193

TABLE V: TEST METRICS FOR MODEL B

Metric	Value
Macro Accuracy	0.8395
Macro Precision	0.7767
Macro Recall	0.8339
Macro F1-Score	0.7962
ROC AUC	0.8919
PR AUC	0.9660

TABLE VI: CLASSIFICATION REPORT FOR MODEL B

	Precision	Recall	f1-score	support
LVOT	0.6120	0.8236	0.7022	1128
RVOT	0.9413	0.8443	0.8902	3782
accuracy			0.8395	4910
macro average	0.7767	0.8339	0.7962	4910
weighted average	0.8657	0.8395	0.8470	4910

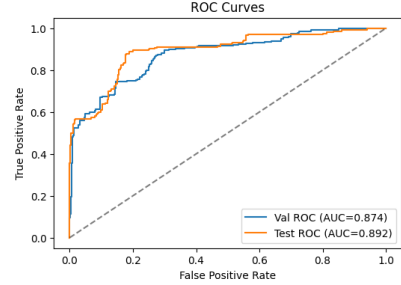


Fig. 10: Receiver Operating Characteristic (ROC) curve for Model B, illustrating the model’s stronger performance in comparison to Model B-Lite.

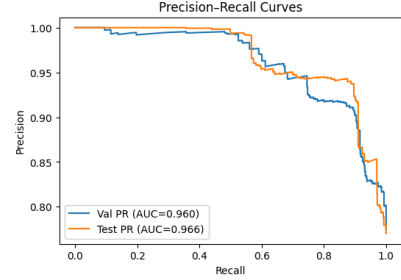


Fig. 11: Precision-Recall curve for Model B, depicting the trade-off between precision and recall at different thresholds, with an area under the curve (AUC) of 0.96 across both validation and test sets, indicating high classification quality.

C. Model C: Classification of RCC and COMMISSURE

Contrary to Part 1, for Part 2, however, the lite approach was not done since with preliminary trials, lightweight models (few estimators or shallow depth) consistently led to poor performance, especially due to the increased difficulty of the classification task. As a result, we allowed the model to explore a broader and more expressive parameter space (number of estimators from 10 to 200 and max_depth of 5 to 15) to find the optimum approach.

The final trained model for Task 2 was evaluated using macro-F1 score, per-class precision/recall/F1, as well as ROC AUC and PR AUC metrics, using a decision threshold of 0.14, which was computed to be the one that maximized the macro-F1.

TABLE VII: CONFUSION MATRIX FOR MODEL C

Actual Predicted	COMMISSURE	RCC
COMMISSURE	275	0
RCC	96	25

TABLE IX: CLASSIFICATION REPORT FOR MODEL C

	Precision	Recall	f1-score	support
COMMISURE	0.7412	1.0000	0.8514	275
RCC	1.0000	0.2066	0.3425	121
accuracy			0.7576	396
macro average	0.8706	0.6033	0.5969	396
weighted average	0.8203	0.7576	0.6959	396

TABLE VIII: TEST METRICS FOR MODEL C

Metric	Value
Macro Accuracy	0.7576
Macro Precision	0.8706
Macro Recall	0.6033
Macro F1-Score	0.5969
ROC AUC	0.2095
PR AUC	0.3611

Analyzing the results obtained, the model performed very well on the validation set, with a strong macro-F1 score (0.8568), precision and recall, and good separability (PR AUC > 0.87). However, a significant performance drop was observed on the test set, particularly in recall for the RCC class, which dropped to 0.2066, resulting in a much lower macro-F1 score (0.5969) and very low ROC AUC (0.2095). This discrepancy between validation and test results suggests that the model may have overfitted to the validation data despite the cross-validation process, or that the RCC class is particularly hard to generalize, due to general lack of enough data. In contrast, the commissure class was consistently well predicted, with perfect recall (1.0) in both validation and test sets, indicating the model’s high confidence and reliability in identifying this class for both the validation and test sets. From a clinical point of view, the model’s ability to predict commissure-originating arrhythmias with high precision and recall suggests strong and stable features from the signal in this subgroup. However, the poor generalization to RCC cases underscores the need for further data collection from RCC cases to improve generalizability.

D. SHAP features interpretation

The SHAP summary plots for Model B reveal both consistent and diverging patterns in feature importance between the two modeling strategies. In both models, `ecg_feature_3` and `PVC_transition_V1` emerge as key drivers of classification, suggesting that these features are robustly informative for distinguishing between LVOT and RVOT origins.

In fact, having `PVC_transition_V1` as one of the top features aligns well with established electrophysiological understanding since the precordial transition zone, especially when it occurs early in lead V1, is a strong indicator of a right ventricular outflow tract (RVOT) origin. This is because arrhythmias originating from the RVOT typically produce a characteristic electrical activation pattern that causes the R-wave to become dominant earlier in the precordial leads, particularly in V1. As a result, an early transition suggests a

right-sided origin, making this feature both physiologically meaningful and diagnostically valuable.

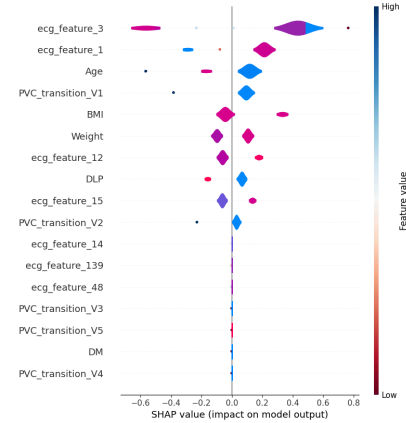
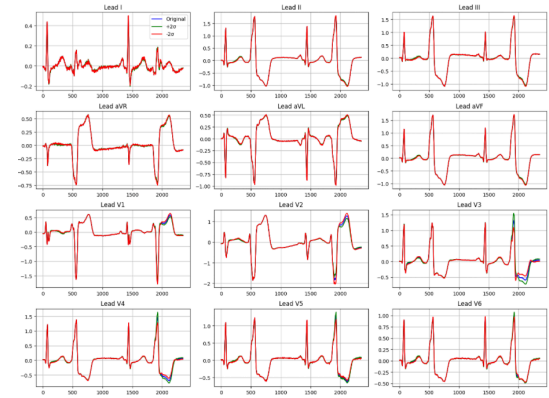


Fig. 12: Comprehensive SHAP value analysis showing feature contributions in the lite model version, providing detailed insights into the model’s decision-making process.

We also tried interpreting the ECG features selected. For example, `ecg_feature_3` and `ecg_feature_2` were analyzed in detail due to their relevance in predicting the origin of the OTVA. Where `ecg_feature_3` exhibits significant activity, specifically around samples 250, 1400, 1900, and beyond 2000. These time points correspond closely to the QRS complex and the T wave in the ECG waveform. This observation suggests that `ecg_feature_3` captures critical electrophysiological events related to ventricular depolarization and repolarization, which are essential for distinguishing the site of origin in outflow tract ventricular arrhythmias. By focusing on these key segments of the ECG signal, the model leverages physiologically relevant information that underpins its predictive capability.

Fig. 13: Detailed interpretation of `ecg_feature_3`, highlighting its critical role in the model’s decision-making process.

Moreover, while the full model incorporates additional clinical variables like BMI and DLP, the lite version focuses more narrowly on ECG-derived features. This difference suggests that the full model may be leveraging subtle correlations with comorbidities or patient-specific characteristics to refine predictions, whereas the lite model prioritizes

direct electrophysiological patterns, yet still manages to maintain competitive performance. This contrast highlights the strength of the ECG signal itself in distinguishing between the two OTVA subtypes.

In terms of ECG features, a few of them, like `ecg_feature_1`, `ecg_feature_3`, and `ecg_feature_63`, are strongly associated with the Commissure group. They might probably capture features like QRS width, transition zone, or axis, which are known to change depending on where the PVC originates. For example, PVCs from the Commissure might have a different QRS pattern or show up with a later R-wave transition. Meanwhile, `PVC_transition_V3` seems to have more influence when it's lower, pushing the model toward RCC. That fits with what we know clinically, PVCs from the RCC tend to have earlier precordial transitions because they come from a more anterior and right-sided position in the heart.

So overall, the model appears to be picking up on a mix of patient characteristics and ECG signal patterns that actually make sense when we think about the underlying anatomy and electrophysiology. It's reassuring to see that the SHAP results align with what we'd expect clinically.

Now we can discuss the SHAP values obtained from the models, which provide insights into the feature contributions and their impact on model predictions. The SHAP values for Model B and Model C are presented below, illustrating how each feature influences the classification decisions made by the models.

For the lite model, the SHAP interpretation gives a good insight on all the decisions the model is taking, as the model is very simple. The SHAP values for the model, which are shown in the figure below, explain that for the lite model, the most important features are the `ecg_feature_3` and `1`, the age and `PVC_transition_V1`, and BMI, we can see that the higher the value for the `ecg_feature_3`, the more likely is to become a LVOT, this makes sense as the `ecg_feature_3` is related to the QRS complex and T-wave. This makes sense in a biomedical context, as the QRS complex and T-wave are critical components of the ECG waveform that reflect ventricular depolarization and repolarization, respectively. The model's reliance on these features suggests that it effectively captures the underlying electrophysiological characteristics that differentiate LVOT from RVOT origins.

Now for the model C, the SHAP values reveal that the most influential features for distinguishing between the Commissure and RCC origins are `ecg_feature_1`, `ecg_feature_3`, and age. The model assigns positive SHAP values to these features when predicting the Commissure origin, indicating that higher values of these features increase the likelihood of classifying an arrhythmia as originating from the Commissure. Conversely, negative SHAP values for these features suggest a stronger association with the RCC origin.

In this case the importance of the age is merely a correlation, as the model is trained on a very small dataset. This means that the model may be picking up on this correlation rather than a true causal relationship between age and arrhythmia origin. This highlights the importance of having

a diverse and representative training dataset to ensure that the model learns meaningful patterns rather than spurious correlations.

IV. CONCLUSION

This study successfully demonstrates the feasibility and clinical utility of machine learning models for the localization of the site of origin (SOO) in outflow tract ventricular arrhythmias (OTVAs) using ECG-derived features combined with demographic data. The classification between left ventricular outflow tract (LVOT) and right ventricular outflow tract (RVOT) origins achieved strong predictive performance, with the full and lite XGBoost models both showing promising accuracy, sensitivity, and generalization capacity. Notably, the lite model, despite being simple, maintained competitive performance, making it a viable option for real-time clinical environments.

Further refinement was attempted in distinguishing between two specific anatomical SOOs within the RVOT region: the right coronary cusp (RCC) and the Commissure. While the model performed well on validation data, generalizability to unseen test data was limited, especially for RCC predictions. This highlights the intrinsic challenge in classifying closely related anatomical sites with subtle electrophysiological differences and underscores the need for larger, more diverse datasets to improve robustness.

The feature importance analysis aligns well with established electrophysiological and anatomical knowledge, confirming that key ECG markers such as precordial transition zones and demographic factors (e.g., age and height) meaningfully contribute to model predictions. These insights reinforce the clinical relevance and interpretability of the machine learning approach.

Overall, this work paves the way for enhanced pre-procedural planning in OTVA ablation by enabling non-invasive, data-driven SOO localization. Future efforts should focus on expanding training datasets, integrating multimodal data, and validating models prospectively to translate these findings into routine clinical practice.

REFERENCES

- [1] R. Doste *et al.*, "Training machine learning models with synthetic data improves the prediction of ventricular origin in outflow tract ventricular arrhythmias," *Front Physiol*, vol. 13, p. 909372, 2022, doi: 10.3389/fphys.2022.909372.
- [2] R. D. Anderson *et al.*, "Differentiating Right- and Left-Sided Outflow Tract Ventricular Arrhythmias," *Circulation: Arrhythmia and Electrophysiology*, vol. 12, no. 6, 2019, doi: 10.1161/circep.119.007392.
- [3] S. A. Strickberger, E. G. Daoud, K. C. Man, M. J. Niebauer, R. C. Semelka, and M. E. Josephson, "Focal Atrial Tachycardia Arising From the Pulmonary Veins," *Journal of Cardiovascular Electrophysiology*, vol. 12, no. 11, pp. 1215–1221, 2001, doi: 10.1046/j.1540-8167.2001.01309.x.
- [4] P. Santangeli and F. Marchlinski, "Outflow tract ventricular arrhythmias: Anatomical considerations and ablation strategies," *Heart Rhythm*, vol. 14, no. 9, pp. 1367–1377, Sep. 2017, doi: 10.1016/j.hrthm.2017.05.030.
- [5] University of Minnesota. VHLab, "Anatomy Tutorial – Cardiac Valve Nomenclature." 2025.
- [6] Á. Bocanegra-Pérez *et al.*, "Automatic and interpretable prediction of the site of origin in outflow tract ventricular arrhythmias: machine learning integrating electrocardiograms and clinical data," *Front. Cardiovasc. Med.* 11:1353096, 2024.