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

Jana Casaú, 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.

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. A critical factor for a successful RFA intervention is the accurate localization of the arrhythmia's Site of Origin (SOO). Identifying 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.

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 [2].

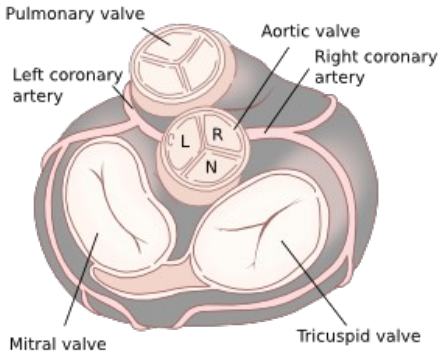

Fig. 1: Anatomical representation of the aortic valve showing the right coronary cusp (RCC), left coronary cusp (LCC), and non-coronary cusp (NCC).

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.

For Part 2, we refine the dataset further by filtering out any patients whose outcomes are not classified as either RCC or COMMISURE. 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 COMMISURE outcomes.

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

A. Data Preprocessing

a) *Demograpgic 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 Bocanegra-Pérez et al. (2024) [3] 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 (POSAR NOM EXCEL). 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, since HTA is a clinically relevant feature. 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 are 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 postion. 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 downsampled 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.

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.

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 2459 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.

c) *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.

B. MODEL TRAINING

Both demographic and ECG signal data 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 80/20 partition strategy based on patient-level 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 (e.g., RVOT vs. LVOT, RCC vs. commissure). Each patient's ECG recordings were assigned exclusively to one set to avoid data leakage and ensure generalizability. The resulting training set contained 12,023 samples from 101 patients; the validation set included 6,454 samples from 34 patients; and the test set comprised 6,505 samples from 34 patients. In the RVOT vs. LVOT classification task, the class distribution was 5,020 RVOT and 1,485 LVOT in the training set; 5,640 RVOT and 814 LVOT in the validation set; and 8,772 RVOT and 3,251 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 set. However, this exhaustive search approach proved computationally inefficient, and early results consistently indicated that gradient boosting models outperformed alternatives. 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 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. 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. 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.

For Model C, where class imbalance was particularly pronounced, SMOTE was applied during training to oversample the minority class, enhancing the model's ability to learn from underrepresented patterns. After identifying the

TABLE I: Classification Report for LVOT vs. RVOT

	Precision	Recall	f1-score	support
LVOT	0.6568	0.6976	0.6766	1078
RVOT	0.9105	0.8940	0.9022	3709
accuracy			**0.8498**	4787
macro average	0.7836	0.7958	0.7894	4787
weighted average	0.8534	0.8498	0.8514	4787

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 saved 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 maximum tree depth of 1, meaning each tree consisted of only a single binary split. 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 Commissure. This resulted in a dataset of 1,835 ECGs from 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.

III. Results and discussion


```
{
  "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 output from the ECG segmentation model showing the detected P-waves, QRS complexes, and T-waves with their corresponding sample indices.

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

TABLE II: Table 2. Confusion Matrix (RVOT vs. LVOT)

Actual / Predicted	LVOT	RVOT
LVOT	752	326
RVOT	410	3299

TABLE III: Classification Report for LVOT vs. RVOT

	Precision	Recall	f1-score	support
LVOT	0.6568	0.6976	0.6766	1078
RVOT	0.9105	0.8940	0.9022	3709
accuracy			**0.8498**	4787
macro average	0.7836	0.7958	0.7894	4787
weighted average	0.8534	0.8498	0.8514	4787

TABLE IV: Overall Test Metrics

Metric	Value
Macro Accuracy	0.8568
Macro Precision	0.7954
Macro Recall	0.8436
Macro F1-Score	0.8139
ROC AUC	0.8992
PR AUC	0.9669

variance in the data. In Figure 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 order 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.

Part 1: 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. As part of this experimentation, we tested a version of the model with a maximum tree depth limited to 1 which means that each decision tree was restricted to a single binary split. Surprisingly, this extremely simple configuration achieved very strong 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" version stood out for its simplicity and its ability to generalize well, making accurate predictions using just 17 decision rules.

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 using a much simpler structure. 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.

With the initial model train, after hyperparameter optimization, the following most optimal parameters were obtained and used (posar parameters obtinguts). The following table shows the classification report obtained from this model for the validation set:

A. POSAR RESULTATS

For this model, 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.

Using the "lite" model, with only 17 decisions and 1 for depth, the following results were obtained:

B. POSAR RESULTATS LITE

Analyzing our results we can confirm the effectiveness of this simpler XGboost model, maintaining good performance while increasing inference. Although slightly better results are obtained with the previous model, the decrease of number of estimators (163 to 17) clearly justifies the use of this model. At an optimized threshold of 0.75, the model continues to demonstrate robust performance, with a macro F1-score of 0.8139 on the test set, indicating a strong balance between precision and recall across both classes. Notably, LVOT recall improves to 81.9%, enhancing sensitivity to the minority class, which is vital for clinical reliability. Although precision for LVOT is slightly lower at 64.9%, this tradeoff results in a better overall F1-score for the class, suggesting improved detection capability without substantially increasing false positives. Moreover, high ROC AUC (0.8992) and PR AUC (0.9669) confirm excellent separability and reliability, making this threshold a well-calibrated choice that enhances performance on underrepresented cases while maintaining strong accuracy and generalization.

Part 2: Classification of RCC and COMMISSURE:

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 the one that maximized the macro-F1 On the validation set, the model achieved strong performance

C. RESULTS

On the test set, however, model performance dropped noticeably:

D. POSAR RESULTS

The model maintained high precision for RCC on the test set, indicating that when it predicted RCC, it was correct most of the time. However, recall for RCC was only 0.369, meaning that approximately 63% of actual RCC cases were missed. The ROC AUC obtained (0.2095) for the test set, confirms that the model struggles to differentiate between the two classes, especially RCC. It does slightly better than random guessing. In contrast, commissure classification remained robust across both validation and test sets, with F1 scores above 0.85.

Despite the strong validation metrics, the drop in test performance, particularly for RCC, suggests a generalization gap. This is likely driven by the small number of test patients (n=5) and significant class imbalance. Since only 149 RCC samples were present in the test set, misclassifications in just a few patients could disproportionately impact recall and F1 scores. Additionally, the two sites of origin display really similar ECG behaviour due to how close their location is.

E. FALTA INTERPRETACIO DE LES DE SHAP

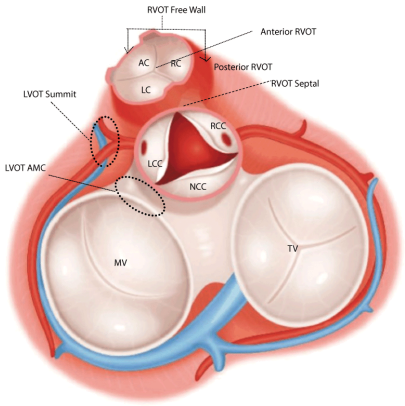

Fig. 2: Detailed view of the right coronary cusp (RCC), highlighting its structural features and anatomical significance in the aortic valve.

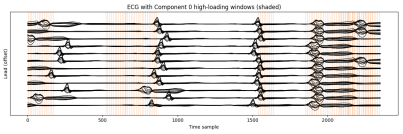

Fig. 3: Electrocardiogram (ECG) signal with highlighted components, showing the key cardiac electrical activity phases and their characteristic waveforms.

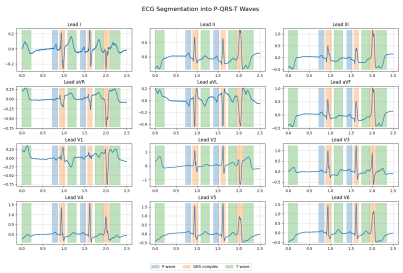

Fig. 4: Validation results of ECG segmentation before alignment, demonstrating the initial segmentation output on the validation dataset.

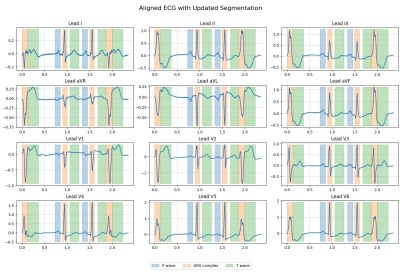

Fig. 5: Aligned ECG segmentation results, showing improved temporal alignment of cardiac cycles after applying alignment procedures.

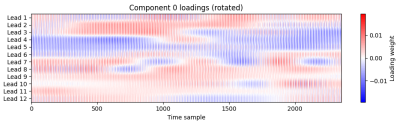

Fig. 6: First principal component analysis (PCA) visualization, highlighting the main pattern of variation in the ECG data.

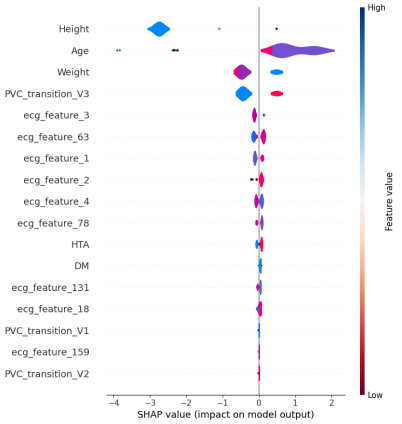

Fig. 7: SHAP (SHapley Additive exPlanations) values for Model C, illustrating feature importance and their impact on model predictions.

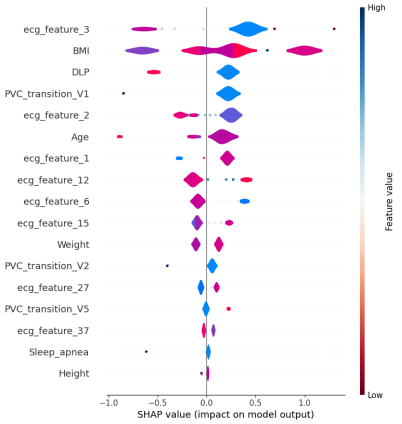

Fig. 8: Comprehensive SHAP value analysis showing feature contributions in the full model version, providing detailed insights into the model's decision-making process.

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] University of Minnesota. VHLab, "Anatomy Tutorial – Cardiac Valve Nomenclature." 2025.

[3] 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.