Machine Learning for Ventricular Origin Detection: Integrating Clinical and ECG-based features

I.Expósito, M.Fairey, X.Miret, and T.Pazos

***Abstract*— *Objective:* To develop an interpretable machine learning pipeline for classifying the origin of ventricular arrhythmias using 12-lead ECGs and patient metadata, focusing on Left vs Right outflow tract and RCC vs RVOTSEPTUM differentiation. *Methods*: ECG signals were resampled, filtered, and aligned to R-peaks, followed by morphological feature extraction using a deep learning segmentation ensemble. These features, combined with clinical data, were filtered via ANOVA F-test and Mutual Information. Multiple classifiers were trained and evaluated, including Random Forest, XGBoost, SVM, and Logistic Regression. *Results:* Tree-based models outperformed others for Left vs Right classification, while simpler models like Logistic Regression performed better for RCC vs RVOTSEPTUM. Feature selection and patient-level class balancing were key to stability and interpretability. *Conclusion:* The proposed pipeline integrates signal processing, feature selection, and interpretable modeling to classify arrhythmic origin effectively. *Significance:* This work supports ECG-based diagnostic decision-making and demonstrates a practical, explainable approach for clinical arrhythmia classification.**

***Index Terms*— Ventricular Arrhythmias, Premature Ventricular Contractions, Electrocardiography (ECG), Site of Origin, Left vs Right Classification, RVOT Septum vs Right Coronary Cusp (RCC), Machine Learning, Random Forest, XGBoost, Logistic Regression, SHAP Explainability, Class Imbalance, R-Peak Alignment.**

Code availability: The code implementing the training pipeline, feature selection, and model evaluation described in this work is publicly available at https://github.com/taniapazospuig/compbiomed-seminars.git.

I. INTRODUCTION

Idiopathic ventricular arrhythmias (VAs) arising from the ventricular outflow tracts, most notably the right ventricular outflow tract (RVOT) and the left ventricular outflow tract (LVOT), constitute a significant proportion of premature ventricular contractions (PVCs) and idiopathic ventricular tachycardias (VTs), particularly in structurally normal hearts. Among these, RVOT-origin PVCs are the most common, accounting for up to 70-80% of idiopathic VAs, while those originating from the LVOT, including the aortic sinus cusps, constitute a smaller but clinically significant subset [1].

The electrocardiographic (ECG) characterization of these arrhythmias is essential for non-invasive localization of the site of origin (SOO) [2]. Consequently, considerable research has been dedicated to refining the diagnostic accuracy of surface ECG markers to differentiate RVOT from LVOT PVCs.

Typically, both RVOT and LVOT arrhythmias exhibit a left bundle branch block (LBBB) pattern with an inferior axis on surface ECG. However, more nuanced distinctions, such as the precordial transition point, the R/S amplitude ratios in inferior and precordial leads, and specific QRS morphologies in leads V1-V3, have emerged as critical tools for localization.

Clinicians often rely on visual inspection of 12-lead electrocardiogram (ECG) signals to infer the likely origin of PVCs, based on heuristic rules involving QRS morphology, polarity, and transition zones. However, this manual approach is limited by significant inter-observer variability and reduced diagnostic accuracy in morphologically ambiguous cases. As a result, there is increasing interest in the development of computational tools that assist in PVC origin classification using quantitative and automated methods.

Several recent studies have applied machine learning techniques to classify PVC origin based on ECG data [2,3]. For instance, Bocanegra-Pérez et al. [2] proposed data-driven approaches leveraging ECG features to distinguish between origins within the outflow tract, demonstrating the potential of automated methods in reducing subjectivity.

Automating the classification of PVC origin is crucial because current manual approaches are inherently subjective, time-consuming, and dependent on the experience of the clinician. Visual interpretation of 12-lead ECGs requires expertise to identify subtle morphological cues, and even among experienced electrophysiologists, inter-observer variability can be significant, particularly in cases with ambiguous or overlapping features.

Moreover, manual analysis is not scalable in high-throughput clinical environments or when rapid decision-making is needed, such as during ablation procedures. An automated, interpretable system can provide consistent, reproducible assessments, reduce diagnostic uncertainty, and serve as a valuable decision-support tool, ultimately improving patient outcomes by guiding more accurate and efficient intervention planning.

In this context, we propose an interpretable machine learning pipeline that integrates signal processing, morphological

feature extraction, and classical classifiers to support SOO prediction from 12-lead ECGs. Our approach specifically addresses two key classification challenges: (1) distinguishing between Left and Right ventricular outflow tract origins, and (2) differentiating RVOT subregions –RVOT Septum vs. Right Coronary Cusp (RCC)– in morphologically ambiguous cases. By combining deep learning-based segmentation with handcrafted ECG features and clinically meaningful metadata, our goal is to deliver a reproducible and explainable tool to assist electrophysiologists in PVC localization. This work contributes to the growing effort to bridge clinical insight and computational intelligence in cardiology, supporting more accurate and scalable arrhythmia diagnosis.

II. METHODS

The dataset used in this study was provided anonymously by the supervisors of the course *Ciència de Dades i Models Computacionals en Biomedicina* at Universitat Pompeu Fabra, and originates from the Teknon Medical Centre, Barcelona. It consists of anonymized ECG recordings and associated clinical metadata from 190 patients, each identified by a PatientID. All data was shared in compliance with data protection and ethical requirements.

Each patient has multiple catheter-based intracardiac ECG recordings acquired from distinct anatomical regions of the heart. The data is organized by region (e.g., "2-LV" for Left Ventricle), with the prefix "Re" indicating a repeated acquisition. Within each region, several measurement points are defined where a catheter was placed during surgery, and simultaneous 12-lead ECG signals were recorded. This enabled us to have a complete and rich dataset of real-world cases.

The dataset, stored in all_points_may_2024.pkl, was originally organized as a deeply nested dictionary under the root key structures, structured by region, then point, and finally the associated 12-lead ECG signals. For analysis, the data was transformed into a flat, tabular format where each ECG sample corresponds to a single row, linked to its respective patient and clinical metadata.

We were also given metadata that included 16 variables: demographic and anthropometric data (Sex, Age, Height, Weight, BMI), cardiovascular risk factors (HTA, DM, DLP, Smoker, COPD, Sleep_apnea), and electrophysiological parameters (PVC_transition, SOO_chamber, CLINICAL_SCORE, SOO, and OTorigin).

To standardize anatomical labels for classification, two supplementary sheets from labels_FontiersUnsupervised.xlsx were used. These provided mappings from the original SOO and binary OTorigin values to simplified anatomical categories, ensuring consistency across patients and enabling reliable supervised learning.

Preprocessing

The flattened raw dataset included 29,153 individual ECG samples across 190 patients. Only samples containing all 12 standard ECG leads were retained to ensure consistency across inputs.

Anatomical labels were mapped and standardized using the file labels_FontiersUnsupervised.xlsx. Specifically, detailed SOO labels were converted to broader chamber categories (SOO_Chamber_Mapped) and then normalized into binary labels: Right for sites corresponding to the right ventricular outflow tract (RVOT, right ventricle, tricuspid annulus, coronary sinus), and Left for left ventricular regions (LVOT, left ventricle, mitral annulus). Fine-grained anatomical labels (region_label) were also extracted to support the region-specific classification task, distinguishing between RVOT Septum and RCC.

Missing clinical data was handled through imputation. Numerical fields such as age, height, and weight were imputed using the median value, while categorical variables, including sex, smoker status, and comorbidities, were imputed using the mode. Missing BMI values were computed from the available height and weight measurements (with height converted from centimeters to meters). The field CLINICAL_SCORE was excluded as it was considered leaky. Incomplete or ambiguous OTorigin values were completed using the second sheet of labels_FontiersUnsupervised.xlsx.

Signal preprocessing was then applied to all ECG recordings. The original signals were sampled at 1000 Hz over 2.5 seconds. To reduce computational and storage demands, all signals were uniformly downsampled to 250 Hz. Bandpass filtering was performed independently on each lead, using a 0.5-100 Hz filter to remove baseline drift and high-frequency noise.

For temporal alignment, the R peak was identified in Lead II within the 1-2 second interval. A fixed 1.25-second segment (312 samples) centered around the R peak was extracted, ensuring consistent beat alignment across all recordings. This windowed extraction enhances morphological comparability between signals and supports downstream modeling. The final processed signal for each sample is represented as a matrix of shape [312, 12], corresponding to 312 timepoints across 12 leads.

Train/test split

To ensure independence between training and testing samples, the dataset was split at the patient level, not the individual ECG level. Each patient was assigned a label by computing the mode of their associated ECG-derived labels, thus enabling patient-level stratification for each classification task.

A stratified train/test split was then performed based on these patient-level labels, preserving the overall class distribution in both subsets. All ECG recordings were assigned to either the training or test set according to their patient ID, preventing any overlap of individuals across splits and avoiding data leakage.

In the train set, for the Left vs Right classification task, all samples labeled as Left were retained, while a soft balancing

strategy was applied to the Right class: a random sample containing 20% more patients than the Left group was selected from the available Right-labeled patients. A similar approach was used for the RCC vs RVOT Septum task, where all RCC-labeled patients were retained, and 20% more patients were randomly selected from the RVOT Septum group.

This soft balancing strategy allowed for a more balanced training set without discarding minority class samples, while the test set remained imbalanced to reflect the clinical class distribution.

ECG Segmentation and Morphological Feature Extraction

To extract physiologically meaningful waveform features from each ECG sample, we applied a deep learning-based segmentation pipeline using an ensemble of five pretrained Swiss Army Knife (SAK) models obtained from the GitHub repository of Guillermo Jiménez Pérez [4]. Each model independently processed the 12-lead ECG signal and produced binary segmentation masks of shape [3, t], where 3 corresponds to the P wave, QRS complex, and T wave regions, and t represents the number of timepoints in the ECG signal.

The ensemble outputs were aggregated to yield final binary masks delineating the location of each waveform component per sample. These masks were then used to guide the extraction of handcrafted morphological features from each of the 12 leads.

For each lead, the following features were computed:

- R-wave amplitude
- S-wave amplitude
- R/S amplitude ratio
- QRS duration (in milliseconds)
- T-wave polarity, encoded categorically as positive (+1), negative (-1), or isoelectric (0)

This resulted in a total of 60 features per ECG sample. These features were extracted independently for each task after the train/test split.

Feature Selection

To enrich the morphological feature set with patient-level clinical information, each ECG sample was linked to its corresponding metadata using a unique triplet identifier: (PatientID, SampleID, Structure). This allowed the integration of demographic and clinical variables –including both categorical (e.g., Sex, Smoker, Sleep_apnea) and numerical features (e.g., Age, BMI)– into the ECG-level feature matrix. Categorical variables were one-hot encoded to ensure compatibility with downstream machine learning models.

Before feature selection, a cleaning step was performed to eliminate columns that could lead to data leakage or were not informative for prediction. Specifically, the following types of features were removed:

- Target variables and derived labels: label, normalized_label, region_label, SOO, SOO_chamber, SOO_Chamber_Mapped
- Identifiers: PatientID, SampleID, Structure
- Raw signal dictionaries: Leads

The resulting matrix contained a mixture of handcrafted morphological features and encoded clinical metadata for each ECG sample.

To identify the most informative predictors, two complementary statistical methods were applied:

1. ANOVA F-test, which evaluates linear associations between numerical features and class labels. It provides a straightforward ranking of features by discriminative power under the assumption of normality and equal variance.
2. Mutual Information (MI), which captures both linear and nonlinear dependencies between features and the target variable. It is especially suitable for datasets combining continuous and categorical features, as in this study.

Each method was applied separately to rank features by importance. From these rankings, the top 30 features were selected for each classification task. This dimensionality reduction aimed to maintain model interpretability and computational efficiency while minimizing overfitting risks associated with the original high-dimensional space.

Model Training and Evaluation

To classify the SOO of PVCs, we trained a diverse set of seven machine learning classifiers on the combined set of handcrafted ECG features and patient-level clinical metadata. We wanted to focus on ML as in a clinical setting it provides stronger explainability and interpretability, so we chose models that span a range of complexity and learning paradigms, including linear models, distance-based classifiers, kernel-based methods, ensemble trees, and neural networks. All models were trained and evaluated under two separate feature selection conditions: one using the top 30 features from ANOVA F-test rankings and another using the top 30 from Mutual Information rankings. This allowed a direct comparison of model behavior across linear and nonlinearly informative feature subsets.

The classifiers evaluated were:

1. Logistic Regression (LR): A linear baseline model trained with L2 regularization. It provides interpretable coefficients and serves as a reference for linear separability of the selected features.
2. k-Nearest Neighbors (k-NN): A non-parametric model that classifies samples based on the majority label among their nearest neighbors in the feature space. It was tuned for optimal k and distance metric via grid search.
3. Support Vector Machine (SVM) with RBF kernel: A kernel-based model capable of learning nonlinear

decision boundaries. Hyperparameters for the kernel coefficient (gamma) and regularization term (C) were selected using cross-validated grid search.

4. Multi-Layer Perceptron (MLP): A fully connected neural network trained with ReLU activations, dropout, and early stopping. This model is suited for learning hierarchical patterns from mixed feature types.
5. Random Forest (RF): An ensemble of decision trees trained using bootstrap aggregation. The number of trees and maximum depth were tuned to balance variance and interpretability.
6. Gradient Boosting (GB): A sequential ensemble method that builds trees to correct errors of previous ones. It was trained with early stopping and learning rate tuning.
7. XGBoost (XGB): An optimized gradient boosting implementation with built-in regularization and early stopping. It is particularly effective on structured data with complex feature interactions.

Hyperparameter tuning for all models was performed using GridSearchCV with stratified 5-fold cross-validation on the training set to ensure fair comparison and mitigate overfitting. All features were standardized before model fitting.

Finally, performance was assessed on the test set using the classification report and confusion matrices generated from the model predictions, summarizing metrics such as accuracy, precision, recall, and F1-score. In addition, SHAP (SHapley Additive exPlanations) values were computed for selected models to interpret feature contributions and provide insights into decision boundaries, enhancing the model's clinical transparency.

III. Results

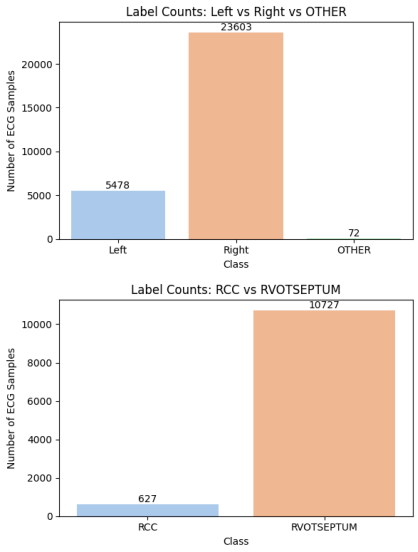

Fig. 1. Distribution of ECG samples across classification tasks. The top panel shows the number of ECG samples labeled as Left, Right, or OTHER. The bottom panel displays sample counts for the RCC vs. RVOTSEPTUM task.

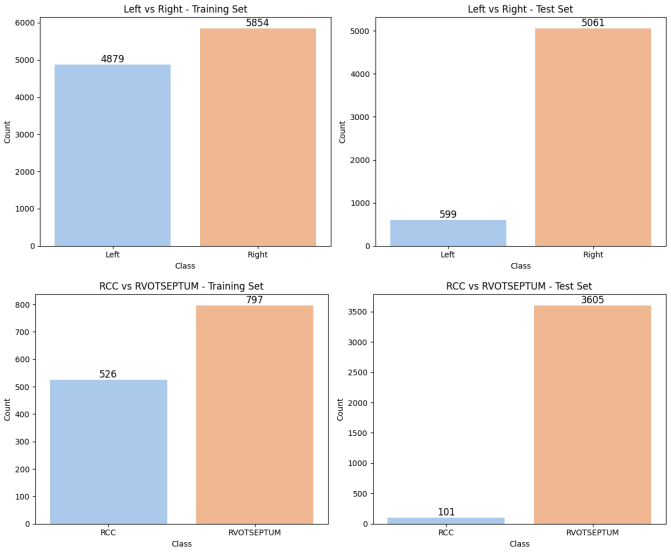

Fig. 2. Distribution of ECG samples in training and test sets after patient-level soft balancing. Top: Class counts for the Left vs Right classification task. Bottom: Class counts for the RCC vs RVOTSEPTUM task. In both cases, patient-level stratified splitting was applied, followed by soft balancing of the training set to reduce class imbalance while maintaining a realistic class ratio. The test sets remain unaltered to reflect real-world distribution.

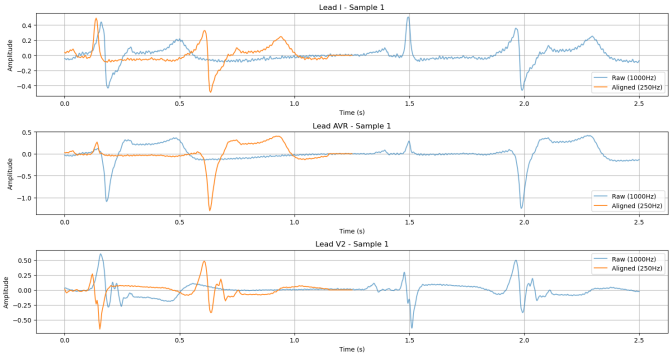

Fig. 3. Preprocessing and alignment of raw ECG signals. Comparison of raw ECG signals (sampled at 1000 Hz) and their corresponding preprocessed versions (resampled to 250 Hz and R-peak aligned). Three selected leads (I, AVR, V2) are shown for one sample from the test set.

TABLE I
HYPERPARAMETER GRIDS USED IN MODEL SELECTION VIA GRID SEARCH. EACH MODEL WAS TUNED INDEPENDENTLY USING 3-FOLD CROSS-VALIDATION.

Model	Hyperparameter	Values Tested
Logistic Regression	C	[0.01, 0.1, 1, 10]
Logistic Regression	penalty	["l2"]
Logistic Regression	solver	["lbfgs"]
SVM (RBF)	C	[0.1, 1, 10]
SVM (RBF)	gamma	["scale", "auto"]
SVM (RBF)	kernel	["rbf"]
k-NN	n_neighbors	[3, 5, 7]
k-NN	weights	["uniform", "distance"]
Gradient Boosting	n_estimators	[50, 100]
Gradient Boosting	learning_rate	[0.01, 0.1]
Gradient Boosting	max_depth	[3, 5]
MLP	hidden_layer_sizes	[(50,), (100,), (100, 50)]
MLP	activation	["relu", "tanh"]
Random Forest	n_estimators	[100, 200]
Random Forest	max_depth	[None, 10, 20]
XGBoost	n_estimators	[100, 200]

TABLE II
MODEL PERFORMANCE (ACCURACY AND MACRO RECALL) FOR LEFT VS RIGHT CLASSIFICATION USING ANOVA AND MUTUAL INFORMATION FEATURE SELECTION.

Model	Feature Selector	ACC	Macro Recall
Logistic Regression	ANOVA	0,53	**0,68**
Logistic Regression	MI	0,60	**0,70**
SVM (RBF)	ANOVA	0,72	0,62
SVM (RBF)	MI	**0,87**	0,55
k-NN	ANOVA	0,71	0,64
k-NN	MI	0,56	0,38
Gradient Boosting	ANOVA	0,76	**0,68**
Gradient Boosting	MI	0,84	**0,68**
MLP	ANOVA	0,64	0,50
MLP	MI	0,65	0,59
Random Forest	ANOVA	**0,86**	**0,68**
Random Forest	MI	0,81	0,53
XGBoost	ANOVA	0,76	0,62
XGBoost	MI	0,84	0,65

TABLE III
MODEL PERFORMANCE (ACCURACY AND MACRO RECALL) FOR RCC VS RVOTSEPTUM USING ANOVA AND MUTUAL INFORMATION FEATURE SELECTION.

Model	Feature Selector	ACC	Macro Recall
Logistic Regression	ANOVA	**0,63**	**0,81**
Logistic Regression	MI	0,39	0,66
SVM (RBF)	ANOVA	**0,61**	**0,80**
SVM (RBF)	MI	0,56	0,77
k-NN	ANOVA	0,57	0,48
k-NN	MI	0,57	0,48
Gradient Boosting	ANOVA	0,48	0,73
Gradient Boosting	MI	0,51	0,56
MLP	ANOVA	0,51	0,75
MLP	MI	0,43	0,70
Random Forest	ANOVA	0,50	0,75
Random Forest	MI	0,49	0,74
XGBoost	ANOVA	0,48	0,73
XGBoost	MI	0,35	0,67

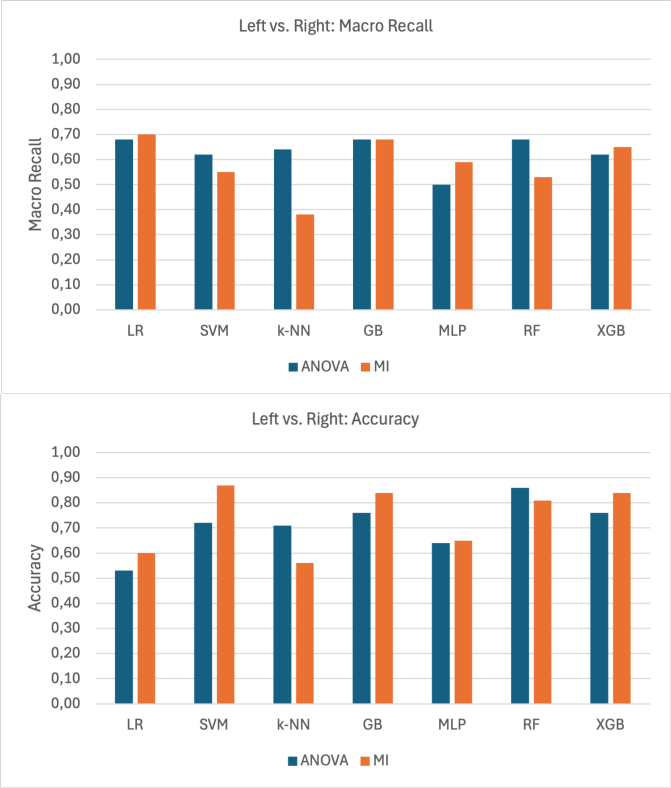

Fig. 4. Classifier performance on the Left vs. Right task using ANOVA vs. MI feature selection. Top: Accuracy. Bottom: Macro recall. Models: LR (Logistic Regression), SVM (RBF), k-NN, GB (Gradient Boosting), MLP, RF (Random Forest), XGB (XGBoost)

Fig. 5. Classifier performance on the RCC vs. RVOTSEPTUM task using ANOVA vs. MI feature selection. Top: Accuracy. Bottom: Macro recall. Models: LR (Logistic Regression), SVM (RBF), k-NN, GB (Gradient Boosting), MLP, RF (Random Forest), XGB (XGBoost).

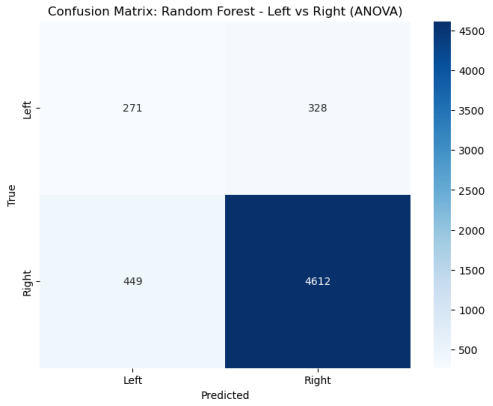

Fig. 6. Confusion matrix of the best-performing model for Left vs. Right classification: Random Forest with ANOVA-selected features.

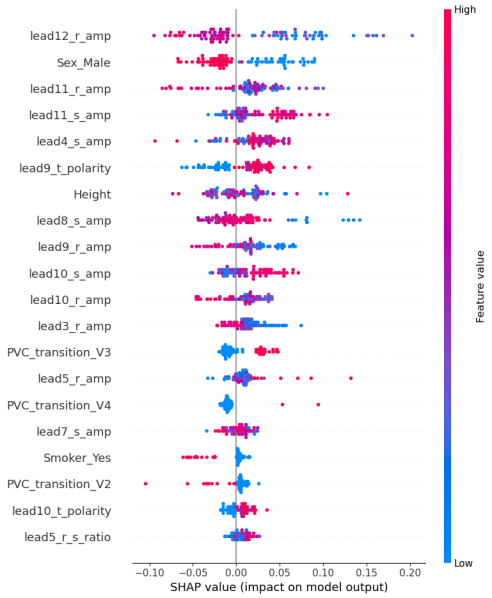

Fig. 7. SHAP analysis of the Random Forest model, showing feature contributions to predictions in the Left vs. Right task.

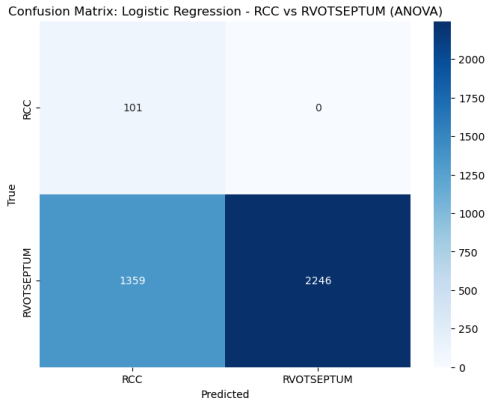

Fig. 8. Confusion matrix of the best-performing model for RCC vs. RVOTSEPTUM classification: Logistic Regression with ANOVA-selected features.

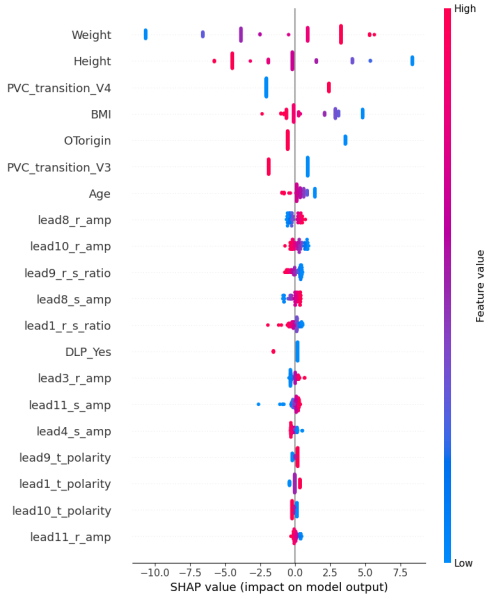

Fig. 9. SHAP analysis of the Logistic Regression model, showing feature contributions to predictions in the RCC vs. RVOTSEPTUM task.

IV. Discussion

Preprocessing and Class Imbalance

Figures 1 and 2 clearly illustrate the significant class imbalance present in the original dataset, with Right-origin PVCs outnumbering Left-origin cases by approximately 4:1. To address this, we applied patient-level soft balancing during training, which reduced this disparity (Fig. 2, bottom panels), enhancing model sensitivity and recall for the minority Left class without distorting the realistic distribution in the test set. Moreover, standardizing ECG signal input through preprocessing (Fig. 3) also proved critical. Resampling to 250 Hz and precise R-peak alignment enhanced feature consistency, particularly around the QRS complex. This uniform representation enabled accurate segmentation and morphological feature extraction, forming the foundation for model learning.

Feature Selection: ANOVA vs MI

Across both tasks, the feature selection strategy had a notable impact. Mutual Information (MI) and ANOVA F-test, the two methods evaluated, produced markedly different outcomes depending on the task and model. In the Left vs Right classification task (Fig. 4), MI generally improved accuracy across most classifiers, including Logistic Regression, SVM, and Gradient Boosting, likely due to its ability to capture non-linear dependencies between features and labels. However, ANOVA consistently led to higher macro recall, which is particularly relevant in imbalanced classification settings as it reflects performance on both classes equally. In contrast, for the RCC vs RVOTSEPTUM task (Fig. 5), ANOVA outperformed MI in both accuracy and macro recall across nearly all models. Notably, MI severely degraded the performance of models reliant on linear separability, with Logistic Regression dropping from 0.63 to 0.39 in accuracy. These results underscore the importance of choosing a feature selection strategy aligned with both the data characteristics and model assumptions, particularly when dealing with subtle inter-class distinctions or imbalanced distributions.

In the Left vs Right task, ANOVA-selected features included categorical clinical indicators like Sex_Male, Smoker_Yes, and COPD_Yes, while MI favored continuous variables such as Age, BMI, and Weight, suggesting different modes of class separation. In the RCC vs RVOTSEPTUM task, both methods agreed on core ECG amplitudes and T-wave polarity features across leads 1-12. ANOVA prioritized interpretable features like OTorigin, while MI highlighted more subtle dependencies such as PVC_transition patterns and ECG amplitude ratios.

Model Behaviour Across Tasks

Left vs Right Classification
Tree-based models such as Random Forest and XGBoost consistently delivered superior performance. Specifically, the Random Forest combined with ANOVA feature selection achieved the highest macro recall of 0.68, alongside strong accuracy metrics (0.86). Notably, XGBoost (MI) and Gradient Boosting (MI) also performed well, reflecting their strength in modeling nonlinear feature interactions.

Although SVM with MI features reached the highest accuracy (0.87), it underperformed in macro recall (0.55), indicating overfitting to the majority class and reduced sensitivity to the minority Left class. Ensemble methods, by contrast, maintained strong recall, confirming their robustness in imbalanced settings.

RCC vs RVOTSEPTUM Classification
In contrast to the previous task, simpler models relying on fewer assumptions performed better here. Logistic Regression with ANOVA-selected features achieved the best overall performance, reaching 0.63 accuracy and 0.81 macro recall (Table III). SVM with an RBF kernel, despite being a non-linear model, also performed competitively, suggesting that while some non-linear interactions may be present, the feature space is still sufficiently well-structured for models with lower complexity to perform effectively.

Tree-based and neural models underperformed in this scenario, with Gradient Boosting, Random Forest, MLP, and XGBoost showing lower accuracy (0.48-0.51) and moderate macro recall (0.56-0.75). This may reflect the task's smaller and more focused dataset, where high-capacity models are more prone to overfitting.

Feature Importance and SHAP Analysis

The outcomes suggest that the nonlinear interactions between ECG morphological features and clinical metadata are critical in distinguishing Left vs. Right origins. Tree-based models effectively capture these complex relationships better than linear models. The SHAP analysis of the Random Forest model (Fig. 7) supports this, highlighting the significant impact of ECG features from R-wave amplitudes in mid and lateral precordial leads such as lead12_r_amp and lead11_r_amp, as well as key clinical features like Sex and Height on the model's predictions. These align with known anatomical differences in ventricular depolarization.

Conversely, in the RCC vs. RVOTSEPTUM classification task, linear models (notably Logistic Regression) outperformed others, particularly when paired with ANOVA feature selection. The SHAP explanation for this logistic model (Fig. 9) reveals that core clinical features such as Weight, PVC_transition, and OTorigin, alongside specific ECG amplitude ratios, strongly drive prediction. This aligns with clinical insights, where subtle but direct morphological differences characterize RCC and RVOT septal origins.

Comparison Between Classification Tasks

When comparing classification tasks, we can see how accuracy was generally higher in the Left vs. Right task across models, likely due to the larger sample size and more distinct sight morphology. On the other hand, macro recall peaked in the

RCC vs RVOTSEPTUM task (0.81), suggesting more balanced performance between classes, possibly due to a more linearly separable feature space.

V. Conclusion

In this project, we have built a complete pipeline to classify the site of origin of PVCs, starting from raw intracardiac ECGs enriched with patient clinical data. Throughout our work, we have addressed real-world challenges like noisy and misaligned signals, class imbalance, and the complexity of combining heterogeneous data sources. Our workflow integrated precise preprocessing, handcrafted feature extraction, soft balancing, and a thorough evaluation of seven machine learning models using two complementary feature selection strategies.

A central goal of our work was to move beyond accuracy alone and focus on interpretability, a critical factor in medical decision-making. By combining ECG morphology with clinical metadata and applying SHAP-based explainability, we showed that ML models can be both effective and transparent, offering insights that align and can potentially complement clinical reasoning.

Our results revealed task-dependent differences in model performance. Tree-based models such as Random Forest and XGBoost achieved the highest accuracy and macro recall in the broader Left vs. Right classification, benefiting from their ability to capture complex, nonlinear patterns in mixed feature types. In contrast, simpler models (Logistic Regression and SVM) performed better in the more localized RCC vs. RVOTSEPTUM task, suggesting a more linearly separable feature space in that setting.

Feature selection also played a critical role. ANOVA-based selection consistently led to more stable and better-balanced models compared to MI. While MI improved accuracy in some settings, especially in the Left vs Right task, ANOVA provided more reliable results overall, particularly in terms of macro recall. This was especially evident in the RCC vs RVOTSEPTUM task, where ANOVA outperformed MI across most models. These findings highlight the importance of aligning the feature selection method with the specific demands of the classification task.

While our results are promising, there are still limitations, most notably the smaller sample sizes in certain subgroups and the lack of external validation. Future directions include incorporating multi-beat context and validating on larger, multicenter datasets.

Overall, this work demonstrates that ML pipelines can go beyond black-box predictions to provide interpretable and clinically meaningful support in cardiology, giving a starting point for relevant future advances.

References

[1] P. Santangeli and F. E. Marchlinski, "Ventricular arrhythmias originating from the outflow tract: how to map and ablate," in Heart Rhythm, vol. 15, no. 8, Philadelphia, PA, USA: Elsevier, 2018, pp. 1264-1272.

[2] Á. J. Bocanegra-Pérez, G. Piella, R. Sebastian, G. Jimenez-Perez, G. Falasconi, A. Saglietto, D. Soto-Iglesias, A. Berruezo, D. Penela, and O. Camara, "Automatic and interpretable prediction of the site of origin in outflow tract ventricular arrhythmias: machine learning integrating electrocardiograms and clinical data," *Front. Cardiovasc. Med.*, vol. 11, p. 500508, 2024.

[3] P. Ponikowski, A. A. Voors, S. D. Anker, H. Bueno, J. G. F. Cleland, A. J. S. Coats, V. Falk, J. R. González-Juanatey, V. Harjola, E. A. Jankowska, M. Jessup, C. Linde, P. Nihoyannopoulos, J. T. Parissis, B. Pieske, J. P. Riley, G. M. C. Rosano, L. M. Ruilope, F. Ruschitzka, and P. Van Der Meer, "2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure," *Eur. Heart J.*, vol. 37, no. 27, pp. 2129–2200, 2016.

[4] G. Jiménez Pérez, "Swiss Army Knife (SAK) ECG segmentation models," GitHub repository, 2024. [Online]. Available: https://github.com/guillermo-jimenez/sak.