Confidence-based Training for Clinical Data Uncertainty in Image-based Prediction of Cardiac Ablation Targets

Rocío Cabrera-Lozoya[1], Jan Margeta[1], Loïc Le Folgoc[1], Yuki Komatsu[2], Benjamin Berte[2], Jatin Relan[2], Hubert Cochet[2], Michel Haïssaguerre[2], Pierre Jaïs[2], Nicholas Ayache[1], Maxime Sermesant[1]

[1] Inria, Asclepios team, Sophia Antipolis, France
[2] Hôpital Cardiologique du Haut-Lévêque, l'Université Victor Segalen Bordeaux II, Institut LYRIC, Bordeaux, France

Abstract. Ventricular radio-frequency ablation (RFA) can have a critical impact on preventing sudden cardiac arrest but is challenging due to a highly complex arrhythmogenic substrate. This work aims to identify local image characteristics capable of predicting the presence of local abnormal ventricular activities (LAVA). This can allow, pre-operatively and non-invasively, to improve and accelerate the procedure. To achieve this, intensity and texture-based local image features are computed and random forests are used for classification. However using machine-learning approaches on such complex multimodal data can prove difficult due to the inherent errors in the training set. In this manuscript we present a detailed analysis of these error sources due in particular to catheter motion and the data fusion process. We derived a principled analysis of confidence impact on classification. Moreover, we demonstrate how formal integration of these uncertainties in the training process improves the algorithm's performance, opening up possibilities for non-invasive image-based prediction of RFA targets.

1 Introduction

Sudden cardiac arrest (SCA) is a leading cause of death in the world, with 350,000 deaths per year in the USA, and similarly in Europe. Its main cause is cardiac arrhythmia with RFA increasingly being used to treat it but with an unsatisfying success rate due to the difficulty to find the ablation targets. There is therefore a need to identify the arrhythmia substrates and the optimal ablation strategy to substantially improve its success rate.

Most arrhythmias occur on structurally diseased hearts with fibrotic scar, where bundles of surviving tissue promote electrical circuit re-entry. These can be identified using electrophysiological (EP) mapping, a lengthy and invasive method that records cardiac electrical signals through intra-cardiac catheters. LAVA, sharp fractionated bipolar potentials occurring during or after the far-field electrogram, indicate surviving fibres within the scar and have been successfully used as targets for RFA [3].

Late-enhancement magnetic resonance imaging (LE-MRI) enables a non-invasive 3D assessment of scar topology and heterogeneity with millimetric spatial resolution. It has been hypothesised that areas of intermediate signal intensity in LE-MRI, referred to as the grey zone, are likely to host both scarred and surviving myocardium related to arrhythmia in ischemic populations [3]. However, consistent EP correlations are still missing. The ability to relate imaging features to LAVA might have direct clinical applications to pre-emptively define mapping and ablation targets, to increase the success rate and to decrease the duration of such procedures (currently>6h). In this manuscript, we evaluated the predictive power of locally computed intensity and texture-based MRI features to identify RFA targets. On the methodological side, we used random forests with advanced image features and classifier parameter estimation using nested cross-validation. However, using machine-learning approaches on such complex multimodal data can prove difficult due to the inherent errors in the training set. We present a detailed analysis of these error sources, due in particular to catheter motion and the data fusion process, their formal integration in the training process, which is rarely done in machine learning approaches, and demonstrate an improved algorithm performance.

2 Clinical Data

Three patients referred for cardiac ablation for post-infarction in ventricular tachycardia were included in this study. They underwent cardiac MRI prior to high-density EP contact mapping of the endocardium (Patients 1 and 2) and epicardium (Patient 3).

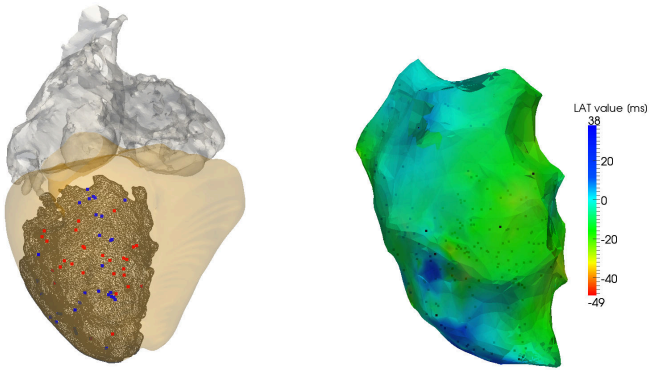

Fig. 1. [left] Anatomical model and EAPs (blue: healthy, red: LAVA) in scar regions (black). [right] CARTO reconstruction of the endocardial cavity, with activation times.

Electrophysiological Data: The CARTO mapping system (Biosense Webster) enables the 3D localization of the catheter tip and provides the distribution of

EP signals on cardiac surfaces. Contact mapping was achieved in sinus rhythm on the endocardium (trans-septal approach) and the epicardium (sub-xiphoid approach) with a multi-spline catheter (PentaRay, Biosense Webster). Signals were categorised as normal or LAVA by an experienced electrophysiologist. Table 1 summarizes the characteristics of the electrophysiological datasets for each of the patients in this study.

	No. Points	Map Source	No. Healthy	No. Lava
Patient 1	91	endocardium	54	37
Patient 2	83	endocardium	50	33
Patient 3	124	epicardium	113	11

Table 1. Patient Electrophysiological Dataset Characteristics.

Imaging Data: Scar was imaged on a 1.5 Tesla clinical device (Avanto, Siemens Medical Systems) 15 minutes after the injection of a gadolinium contrast agent. A whole heart image was acquired using an inversion-recovery prepared, ECG-gated, respiratory-navigated, 3D gradient-echo pulse sequence with fat-saturation ($1.25\times1.25\times2.5mm^3$). The myocardium was manually segmented on reformatted images of isotropic voxel size ($0.625mm^3$). Abnormal myocardium (dense scar and grey zone areas) was segmented using adaptive thresholding of the histogram, with a cut-off at 35% of maximal signal intensity. Segmentations were reviewed by an experienced radiologist, with the option of manual correction.

3 Confidence-based Learning

3.1 Sources of Uncertainty

We aim to identify differences in regional image characteristics between LAVA inducing and healthy tissue. Nevertheless, despite the integrated catheter localisation in the EP system and previous registration with anatomical data, there remains three main sources of uncertainty between the EP measurements and the imaging data.

Temporal Displacement. Due to breathing and cardiac motions, the recording catheter is displaced throughout the 2.5 seconds of recording time. Magnitudes varying significantly among EAPs as is shown in Figure 2.

Data Fusion. Meshes generated by EP mapping systems are a rough approximation of the shape of the ventricular cavity, as seen in Figures 1 and 2. Therefore, a registration is needed between the EP recording locations and the image segmentation. This is done manually by using landmarks and matching between low voltage areas and scars. Then the EAPs are projected on the image-based mesh by finding the closest cell, and an evaluation of the registration uncertainty

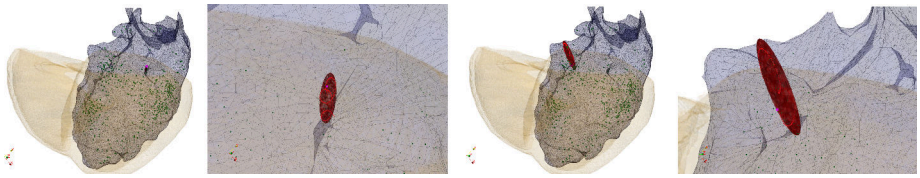

Fig. 2. Location of two EAPs and temporal position variation described by a red ellipsoid (left: low displacement, right: high displacement).

is present in the resulting projection distance (PD).

Sensing Range. The volume of tissue that influences the recording at a particular EAP (catheter's sensing range) is represented by a sphere of 10mm radius.

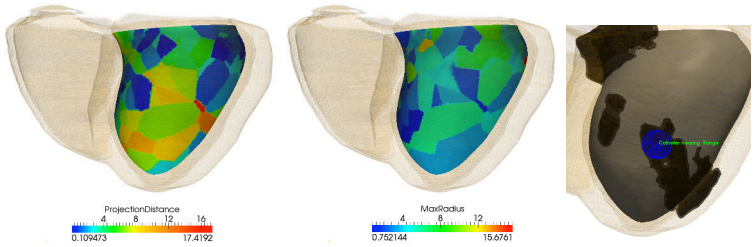

Fig. 3. Posterior endocardial maps for uncertainty attributes [mm] (left: projection distance, centre: major ellipsoid axes). Right: spherical representation of the catheter's hearing range (blue) for a given location. Scar regions are shown in black.

3.2 Image Feature Computation

These properties allow us to define the scale at which the image features will be looked at and how to quantify the errors.

Intensity-based Features: Voxels contained inside the sensing range were used to compute intensity-based features, including minimal, maximal, mean and standard deviation values. Another feature, defined as the standard deviation over the average intensity in the region was included. Myocardium thickness was calculated and the scar transmurality was defined as the extent of scar through the entire myocardium thickness.

Texture-based Features: Grey level co-occurence matrices (GLCM) are matrices of the joint probability of occurrence of a pair of grey values separated by a displacement $d = (dx, dy, dz)$. Haralick features are statistics computed on

GLCM that emphasize specific texture properties and have been extensively used in medical image analysis [7]. In our study, the GLCM were computed around the center of the myocardium were the EAP had been projected using a ROI of window size of 11×11×11 pixels (∼9.4×9.4×9.4mm). Three distances from the central pixel (1, 2 and 4 pixels), 13 directions and 12 Haralick features were considered, resulting in a 468 element texture feature vector per EAP analyzed. Concatenation of the seven intensity and the texture features yielded a final image-based feature vector of 475 dimensions which was used for classification.

3.3 Classification Framework

Random Forests Classifier: Random forests are discriminative classifiers created in an intuitive and easily understandable structure that also provide informative uncertainty measures on the classification results [1]. They have successfully found multiple applications in medical image processing [1][6]. We used the Python implementation from the scikit-learn library [8].

Nested Cross-validation: Cross-validation has been shown to be among the best ways to estimate performance [4]. We used stratified cross-validation and optimized the classifier for precision performance. The use of nested cross-validation, with a parameter-tuning inner loop and an outer loop for performance estimation, avoided an optimistic bias introduction into generalization estimate [9].

Feature Selection: Due to the high dimensionality of our feature vector, the effect of feature space reduction was next assessed. Univariate t-Test statistics were used to assess feature significance [2]. Three reduced datasets were created, including the 50%, 25% and the 10 most relevant features (MRF). Additionally, a feature subset was generated containing only the intensity-based features.

Uncertainty Assessment: We derived a principled analysis of confidence impact on classification. Inspired by cost-sensitive learning, we formulate the problem as samples (x, y, c) drawn from a distribution D on a domain $X \times Y \times C$ with X being the input feature space, Y corresponding to the binary output class and C to the confidence associated with each sample. We aim to learn a classifier $h : X \to Y$ which minimizes the new expected classification error:

$$E(x, y, c \sim D)[cI(h(x) \neq y)] \tag{1}$$

Using the Translation Theorem 2.1 in [10] we can compute and draw samples from a distribution D' such that the optimal error rate classifiers for D' are optimal cost minimizers for data drawn from D. We derive how this modifies the training using weights to simulate the expectation of finite data $E(x, y \sim D)[f(x, y)]$ as:

$$E(x, y \sim D)[f(x, y)] = \frac{1}{\sum c} \sum cf(x, y) \tag{2}$$

equivalent to importance sampling for D' using distribution D, so the modified expectation is an unbiased Monte Carlo estimate of the expectation with respect to D' [10]. In random forests, the node split criterion is information gain:

$$IG = H(S) - \sum_{i=1,2} \frac{|S^i|}{|S|} H(S^i) \tag{3}$$

with $|S|$ being the number of samples in a node before split, $|S^i|$ being the number of samples of each children node and $H(S)$ the Shannon entropy:

$$H(S) = -\sum_{c \in C} p(c) log(p(c)) \tag{4}$$

where p(c) is calculated as normalized empirical histogram of labels corresponding to the training points in S, $p(c) = \frac{|S^i|}{|S|}$. Using weighted instances, $p(c)$ is replaced by $p_w(c)$, which has the following formulation:

$$p_w^i(c) = \frac{\sum \text{Weights of samples of class c in node i}}{\sum \text{Weights of samples in node i}} = \frac{\sum_{S_c} W^i}{\sum_S W^i} \tag{5}$$

This yields a sample weighted formulation of the information gain that can be written as:

$$IG = H(W) - \sum_{i=1,2} \frac{\sum_S W^i}{\sum_S W} H(W^i) \tag{6}$$

where W are sample weights at the parent node and W^i are sample weights that have been passed to each child node. $H(W)$ is given by:

$$H(W) = -\sum_{c \in C} p_w(c) log(p_w(c)) \tag{7}$$

To our knowledge, it is the first time such formulation is derived in this context and we believe it strengthens our approach's methodological ground.

We analysed the influence of two factors affecting the certainty in EAPs and imaging data correspondences by weighting our training samples according to the confidence we have on their image features.

Projection Distance. More confidence is assigned to the imaging features computed from EAPs with low PD with respect to those with high values. It is defined as the Euclidean distance between the EAP and the center of the cell in the mesh closest to the given point:

$$PD = \|CellCenter - EAP\| \tag{8}$$

Temporal Displacement. The covariance of the position matrix is obtained and an ellipsoid with radii $2\sqrt{diag(D)}$ is generated, where D is the matrix containing the eigenvalues along the main diagonal. The major ellipsoid radius defines the temporal displacement, as following:

$$TD = max(2\sqrt{diag(D)}) \tag{9}$$

Intuitively, image features from EAPs with smaller major ellipsoid radius are more reliable as they are less affected by movement.
Each EAP is assigned a confidence value by linearly scaling either the PD or the temporal displacement to a weight parameter with range of [0.5, 1] where 0.5 corresponds to the lowest confidence and 1 to the highest. Additionally, a *combined uncertainty* weight is defined as the product of both uncertainty sources. We explore the uncertainty inherently introduced to our dataset due to these factors with three extra experiments conducted on the dataset of Patient 1 by using the previously described sample weighting schemes.

Evaluation Metrics: The results for the classification results will be assessed using precision-recall (PR) and receiver operating characteristic (ROC) curves. A PR curve illustrates the trade-off between the proportion of positively labelled examples that are truly positive (precision) as a function of the proportion of correctly classified positives (recall). A good performance line lies in the upper-right portion of the graph and has a high area under the curve (AUC) value.

A ROC curve depicts relative trade-offs between benefits (true positives) and costs (false positives). Any deviation from the diagonal (representing random guess and with 0.5 AUC) into the plot's upper triangle represents a better performance than chance. Classifiers with higher ROC AUC values are said to have a better average performance.

4 Results and Discussion

The results of classification using the full feature set for each patient are shown in Figure 4 and their mean AUC values are summarized in Table 2. The AUC PR ranges from 0.75 in Patient 3 to 0.88 in Patient 2. For the AUC ROC metric, values range from 0.80 in Patient 1 to 0.92 in Patient 3. We can therefore argue that an overall good performance was achieved throughout the patients.

By closely looking at the PR curves, we can conclude that the classifier is able to retrieve approximately more than half of the LAVA instances without having a considerable drop in the precision. It is when the totality of the LAVA instances are recovered that precision is compromised. This might correspond to areas in which the LAVA regions are spatially close to healthy ones, therefore finer descriptors of the adjacent areas should be explored. Nonetheless, for our approach some precision can be compromised as the ablation procedure might not be able to distinguish between closely spaced cardiac regions. Plots for Patient 3 show that it presents great variability between folds. Some have perfect scores, probably representing the typical LAVA image signatures, while others perform poorly, possibly due to outlier LAVA present in the testing phase. This can also be due to the patient having an epicardial study compared to the endocardial mapping of Patients 1 and 2.

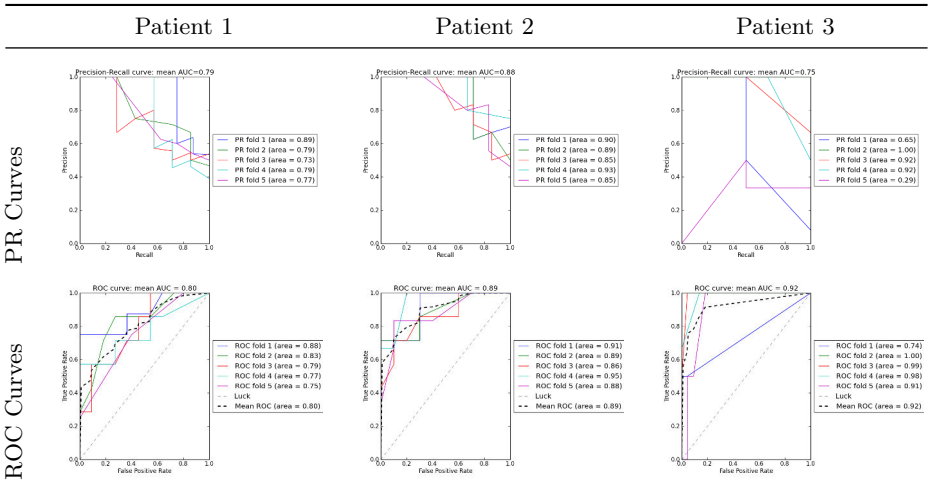

Fig. 4. Precision-recall and ROC curves for patients 1-3 after nested cross-validation with precision optimisation. Line colors represent each fold (curves with AUC = 1 are aligned with axes) and the dotted line represents the average curve for all folds.

	PR AUC			ROC AUC		
	Patient 1	Patient 2	Patient 3	Patient 1	Patient 2	Patient 3
Full Features	0.79	0.88	0.75	0.80	0.89	0.92
50% MRF	0.72	0.86	0.72	0.76	0.86	0.91
25% MRF	0.73	0.91	0.81	0.73	0.91	0.91
10 MRF	0.69	0.80	0.66	0.73	0.81	0.88
Intensity Features	0.78	0.75	0.31	0.73	0.78	0.63

Table 2. Area under the curves for Precision-Recall and ROC.

Results Using Subsets of MRF: The classification results obtained using only 50%, 25% and the top 10 MRF in each dataset are summarized in Table 2. The purpose of this task was to investigate the redundancy in the feature set. In general, the use of 50% of MRF resulted only in a small drop in AUC PR and AUC ROC scores, suggesting some feature redundancy. Nonetheless, while Patient 2 and 3 had a slight increase in performance when using its 25% MRF, all patients suffered a considerable drop in score when using 10 MRF. Due to the characteristics of our current feature selection scheme (univariate filtering method), it failed to assess groups of features that work together to better discriminate between classes.

The results for the assessment of intensity feature importance w.r.t. the full set of features are included in Table 2. Using only intensity features for classification led to a large drop in overall classification performance. This shows that advanced texture patterns are required to describe the complex intertwining of

myocardial fibres in scarred and grey zone areas responsible of LAVA generation.

Uncertainty Impact on the Prediction: Results of classification with the full feature set and weighted samples are shown in Table 3. Furthermore, the optimal random forests construction parameters found during nested cross validation with precision optimization and and temporal displacement weighting are included in Table 4.

A general increase in performance is observed when weighting samples according to their proximity to the location of image feature computation and their temporal position stability. This confirms our hypothesis that a lower confidence should be assigned during training to EAPs with large PD or temporal displacements. Weighting samples with a combination of both uncertainties results in a similar improvement in the classification performance as when the elements were used independently. Currently, this combined uncertainty is a naive product of both elements. A different fusion should be explored to better exploit both uncertainty sources and construct a more reliable estimate of the confidence of a given EAP.

Confidence-based Training

	Patient 1		Patient 2		Patient 3	
Area Under the Curves	PR	ROC	PR	ROC	PR	ROC
No Confidence Weighting	0.79	0.80	0.88	0.89	0.75	0.92
Projection Distance Weighting	0.85	0.85	0.94	0.94	0.96	0.99
Temporal Displacement Weighting	0.86	0.87	0.95	0.95	0.95	0.99
Combined Uncertainty Weighting	0.84	0.87	0.94	0.94	0.94	0.99

Table 3. Classification performance scores using sample uncertainty weights.

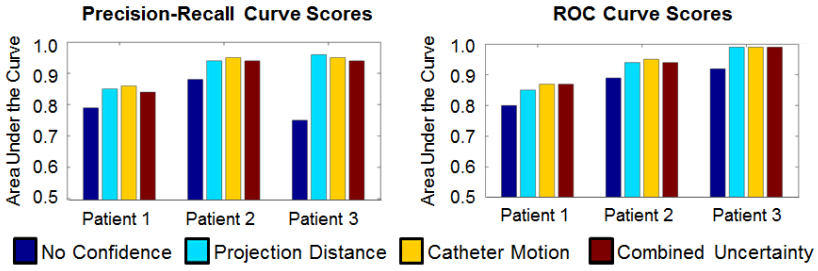

Fig. 5. Bar plots for classification performance scores using sample uncertainty weights

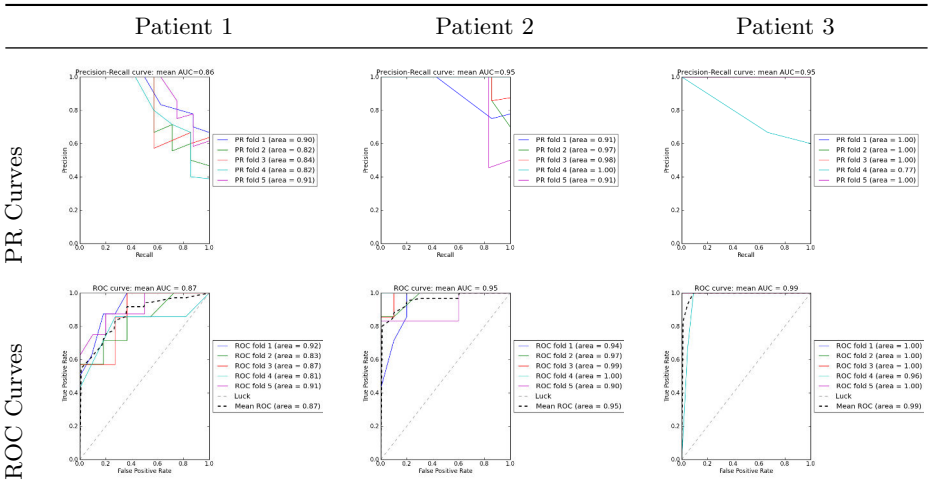

Fig. 6. Precision-recall and ROC curves for patients 1-3 after nested cross-validation with precision optimisation and temporal displacement weighting. Line colors represent each fold (curves with AUC = 1 are aligned with axes) and the dotted line represents the average curve for all folds.

Random Forest Construction Parameters

	PR Score	No. Trees	No. Features	Split Criteria	No. Inner Folds
Patient 1	0.86	25	60	entropy	5
Patient 2	0.95	25	60	entropy	5
Patient 3	0.95	10	20	gini	5

Table 4. Random forests construction parameters when performing nested cross validation with precision optimization and temporal displacement weighting.

A visual interpretation of the results obtained using temporal displacement weighting and their comparison to ground truth are shown in Figure 6. The central image is a preliminary output to be used in a clinical environment. The endocardial map shows regions classified as being in risk of presenting LAVA and that should be ablation targets. The rightmost image shows that prediction errors are primarily present in regions with low classification confidence.

5 Conclusions

We presented the use of intensity and texture-based local imaging features in the vicinity of myocardial scar and grey zones towards the prediction of RFA target localisation. Additionally, we detailed the uncertainty in the data and explored its impact on the classification results. For both PR AUC and ROC AUC, we scored above 0.75 and an extra 0.05 was gained when using uncertainty

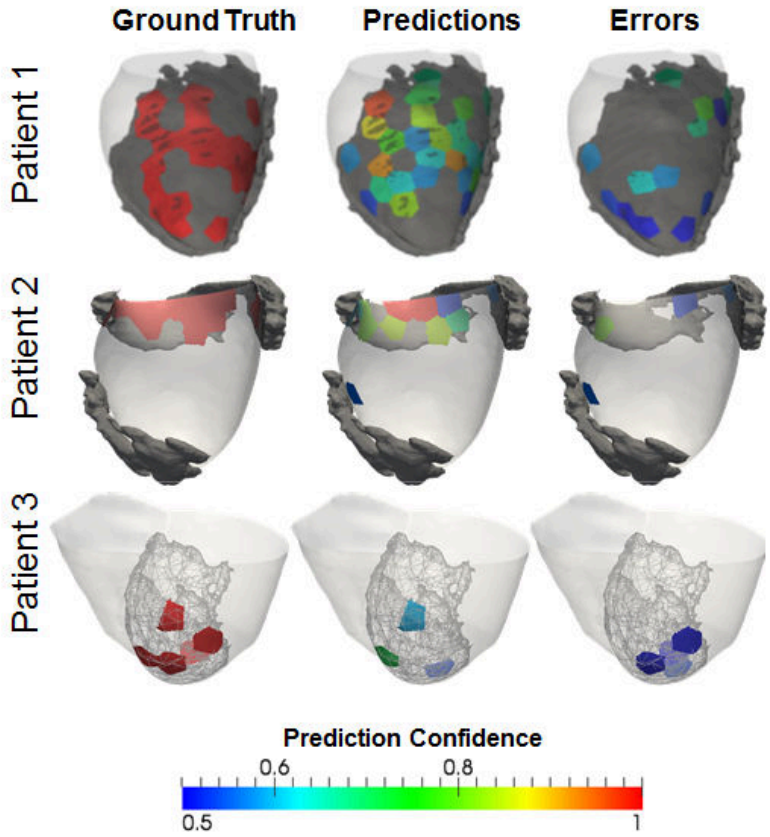

Fig. 7. [left] LAVA regions from ground truth. [center] LAVA regions from predictions. [right] Prediction errors. (Color coding by classification confidence)

evaluation to weight the training data. Finally, a preliminary output with visual interpretation and potential use in a clinical environment was presented.

The aim of the current work was to analyse the feasibility of classification using solely image-based features on complex multi-modal data. So far, the choice of features produced encouraging results but a further refinement and exploration of different texture filters is considered for the following stages of the project to better describe the underlying tissue inhomogeneity. In the feature selection stage, the use of wrapper methods should be explored to evaluate the impact in the classification performance of feature subsets rather than individual features [5]. Future work also includes the improvement of the weighting scheme to better exploit uncertainties on catheter position and motion.

We are aware that an increase in the size of the patient database is required in order to aim for inter-patient analysis, but this work serves as a proof of concept

with results good and encouraging enough to warrant further investigation and open up possibilities for non-invasive cardiac arrhythmia ablation planning.

6 Acknowledgments

Part of this work was funded by the European Research Council through the ERC Advanced Grant MedYMA 2011-291080 (on Biophysical Modeling and Analysis of Dynamic Medical Images).

References

1. Criminisi A, Shotton J, Konukoglu E. Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. no. MSR-TR-2011-114, 28 October 2011.
2. Guyon I and Elisseeff A. 2003, An Introduction to Variable and Feature Selection. In Journal of Machine Learning Research. pp. 1157-1182.
3. Jais P, Maury P, Khairy P, Sacher F, Nault I, Komatsu Y, Hocini M, Forclaz A, Jadidi AS, Weerasooryia R, Shah A, Derval N, Cochet H, Knecht S, Miyazaki S, Linton N, Rivard L, Wright M, Wilton SB, Scherr D, Pascale P, Roten L, Pederson M, Bordachar P, Laurent F, Kim SJ, Ritter P, Clementy J, Haissaguerre M. Elimination of Local Abnormal Ventricular Activities: A New End Point for Substrate Modification in Patients with Scar-Related Ventricular Tachycardia. Circulation. 2012 May 8;125(18):2184-96.
4. Kohavi, R. 1995. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'95. pp. 1137-1143.
5. Kohavi, R and John G. 1997. Wrappers for feature subset selection. In Artificial Intelligence. pp. 273-324.
6. Lempitsky V, Verhoek M, Alison Noble J and Blake A. Random Forest Classification for Automatic Delineation of Myocardium in Real-Time 3D Echocardiography. In FIMH, LNCS 5528, pages 447-456. Springer, 2009.
7. Ludvik T, Smutek D, Shimizu A and Kobatake H. 2007, 3D Extension of Haralick Texture Features for Medical Image Analysis. In Proceedings of the Fourth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications. SPPRA 07 pp. 350-355.
8. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011. Vol 12, pp. 2825-2830
9. Ruschhaupt M, Huber W, Poustka A and Mansmann U. A compendium to ensure computational reproducibility in high-dimensional classification tasks. Stat Appl Genet Mol Biol, 3 : Article37, 2004.
10. Zadrozny B et al. 2003. Cost-Sensitive Learning by Cost-Proportionate Example Weighting. Proc 3rd IEEE Conf Data Mining, p 435-442.