Exploring Interpretability in Deep Learning Prediction of Successful Ablation Therapy for Atrial Fibrillation

Shaheim Ogbomo-Harmitt[1] SHAHEIM.OGBOMO-HARMITT@KCL.AC.UK
Marica Muffoletto[1] MARICA.MUFFOLETTO@KCL.AC.UK
Aya Zeidan[1] AYA.ZEIDAN@KCL.AC.UK
Ahmed Qureshi[1] AHMED.QURESHI@KCL.AC.UK
Andrew P. King[1] ANDREW.KING@KCL.AC.UK
Oleg Aslanidi[1] OLEG.ASLANIDI@KCL.AC.UK
[1] *School of Biomedical Engineering and Imaging Sciences, King's College London, 3rd Floor Lambeth Wing, St Thomas' Hospital, London, SE1 7EH*

Editors: Under Review for MIDL 2022

Abstract

Radiofrequency catheter ablation (RFCA) therapy is the first-line treatment for atrial fibrillation (AF), the most common type of cardiac arrhythmia globally. However, the procedure currently has low success rates in dealing with persistent AF, with a reoccurrence rate of $\sim$ 50% post-ablation. Therefore, deep learning (DL) has increasingly been applied to improve and optimise RFCA treatment for AF. However, for a clinician to trust the prediction of a DL model, the model's decision process needs to be interpretable and have biomedical relevance. This study explores DL interpretability in predicting the success of RFCA strategies simulated using MRI-derived 2D left atrial (LA) tissue models. The developed DL model had an AUC (area under the receiver operating characteristic curve) of 0.78 $\pm$ 0.04 for predicting the success of the pulmonary vein isolation strategy, 0.923 $\pm$0.016 for a fibrosis-based (FIBRO) strategy and 0.77 $\pm$0.02 for a rotor-based strategy (ROTOR). Three feature attribution (FA) map methods were used to investigate interpretability: GradCAM, Occlusions and LIME. GradCAM was found to have the highest percentage of RFCA ablation lesions (known from 2D LA simulations, but unseen by DL model) within informative regions within the FA maps (62% for FIBRO and 71% for ROTOR). Most of the ablation lesions coincided with informative regions of the FA maps (for ROTOR and FIBRO), suggesting that the DL model leveraged structural features of MR images to identify pro-arrhythmogenic regions to make its prediction. In the future, such techniques can help predict the success of RFCA strategies from patient imaging data.

Keywords: Deep learning, Atrial fibrillation, Cardiac modelling, Interpretability, Medical imaging.

1. Introduction

Atrial fibrillation (AF), the rapid, uncoordinated contraction of the atria, is a heart condition that affects 33 million people worldwide - making it the most common type of cardiac arrhythmia globally (Hart and Halperin, 2001; Chugh et al., 2014). Currently, the direct cause of AF is unclear. However, there is evidence that ectopic electrical beats originating from the pulmonary veins (PVs) can trigger AF (Chen et al., 1999). The triggers can then generate re-entrant drivers (rotors) that sustain AF, and spatial fibrosis distributions in the

left atria (LA) have been demonstrated to facilitate such drivers (Morgan et al., 2016; Roy et al., 2020). One treatment for AF is radiofrequency catheter ablation (RFCA) therapy. RFCA involves using induced heat from a rapidly alternating current in a catheter to ablate (isolate or destroy) the arrhythmogenic area of atrial tissue that harbours triggers or rotors, thus restoring sinus rhythm and the mechanical function of the heart (Townsend and Sabiston, 2001). Presently, the success rate of RFCA is ∼70% for paroxysmal AF - which is relatively high (Oketani et al., 2012). However, the procedure is much less successful when dealing with persistent AF, which has a reoccurrence rate of ∼75% post-intervention. Therefore, with the high reoccurrence rate of AF, there is a need for improvements within RFCA (Wang et al., 2017; Yubing et al., 2018). Image-based computational modelling has been used to understand the structure-function relationship that determines re-entrant atrial drivers for AF with the aim of improving RFCA outcomes. As a result, computational modelling has been introduced to improve RFCA outcomes, ultimately leading to the FIRM (Focal Impulse and Rotor Modulation) trial study. The FIRM trial study investigated ablating areas of the LA fostering phase singularities – showing AF reversal in 80.3% of the patients (Narayan et al., 2014). However, the procedure is still awaiting a multi-centre study as the ablation strategy has not proven to be better than the pulmonary vein isolation (PVI) ablation strategy (current gold standard) (Brachmann et al., 2019). With the recent rise of artificial intelligence (AI), machine and deep learning (DL) have been applied to patient medical imaging data and computational cardiac modelling with the aim to develop effective RFCA treatment. These applications of DL include predicting AF reoccurrence post-RFCA and the origins of AF triggers and ablation sites (Kim et al., 2020; Liu et al., 2020). However, DL is limited by its black-box nature. This is an issue when considering European Union's General Data Protection Regulation (GDPR), as any algorithmic decision used in patient care requires an explanation for transparency. Moreover, clinicians have also argued that if AI can outperform human diagnosis, understanding the AI model's decision process will be beneficial in discovering new biological processes and furthering medical knowledge (Watson et al., 2019).

1.1. Related Work

Muffoletto et al. were the first to apply DL to this problem and developed a convolutional neural network (CNN) to predict suitable ablation strategies for a given patient, using synthetic tissue-based atrial models with randomly distributed fibrotic patches. The approach proved effective (79% accuracy) and illustrated the proof-of-concept (Muffoletto et al., 2019). Ultimately, this led to the approach being applied to MRI-derived data to predict the patient-specific optimal RFCA strategy. As a result, the developed CNN had a 100% accuracy for classifying optimal fibrosis- (FIBRO) and rotor-based (ROTOR) strategies success and 33% accuracy for the PVI ablation strategy (Muffoletto et al., 2021).

Currently, research in interpretability for DL AF management is minimal. For example, one study by Alhusseini et al. used gradient-weighted class activation mapping (GradCAM) to show that their feature attribution (FA) map closely replicated rules used by clinicians. However, only one method was validated within this study, and a comparison between other methods was not investigated. Furthermore, the study used spatial maps of the activation phase derived from electrocardiogram data from a basket catheter. Hence, there has been

no investigation into DL interpretability for models which use medical imaging data to make explainable predictions for cardiac arrhythmias and anti-arrhythmic treatments (Alhusseini et al., 2020).

1.2. Contributions

In this study, we will make the following novel contributions to the field of DL-based RFCA outcome prediction:

1. We present a novel qualitative and quantitative comparison of established DL interpretability methods for medical imaging and image-based cardiac modelling.
2. We propose a new quantitative metric to assess interpretability of FA maps for image-based cardiac modelling.
3. We present the first investigation of DL interpretability methods for cardiac modelling with medical imaging data.

2. Methods and Materials

2.1. Overview

In this study, we propose a DL model to 1) accurately predict the outcomes of RFCA therapy based on biophysical modelling and simulations and 2) interpret the decision process of the DL model. To achieve this, standardised 2D LA models were produced with patient-specific distributions of fibrosis derived from late gadolinium-enhanced (LGE) MR imaging data. Simulations of AF and its termination with RFCA were performed to predict the success of multiple strategies, and the simulation results were compared with interpretability maps to identify proarrhythmogenic locations. Three established interpretability approaches were also compared qualitatively and quantitatively to interpret the CNN's predictions.

Thus, this study simulates three RFCA strategies on patient-specific 2D LA tissue models. The latter is derived from the LGE MR images from which the LA and fibrosis have been segmented.

2.2. Data Acquisition and Pre-processing

The datasets used in this study were derived from 122 LGE MRI patient scans: 86 datasets with spatial resolution of 0.625x0.625x0.625 mm^3 were acquired from the Atrial Segmentation Challenge at the STACOM 2018 workshop; additionally, 36 LGE MRI images were collected at St. Thomas' Hospital London with resolution of 1.3x1.3x4 mm^3 (specifically, 18 AF patients were scanned both pre-and post-intervention) (Xiong et al., 2021; Chubb et al., 2018).

Generating 2D LA models with fibrosis first required manual segmentation of patient LGE MRI data to produce 3D patient-specific endocardial LA surface meshes. The LGE MR image intensities were then mapped to these models and the image intensity ratio thresholding technique was applied to quantify and visualise LA fibrosis (Roy et al., 2020). Finally, the 3D LA fibrosis maps were unwrapped using the LA standardised unfold mapping technique to produce models in the 2D LA disk format for the DL network, as shown in

Appendix 2 (Williams et al., 2017; Qureshi et al., 2020). Furthermore, to increase the size of the dataset, synthetic 2D LA disks were generated by weighted-averaging the patient-specific datasets to vary the fibrosis distribution and PVs. The creation of synthetic disks consisted of three steps. First, 65 MRI images were extracted from the STACOM 2018 dataset and were each weighted by assigning a random weight (between zero to one) to all voxels of a given image. Then the extracted fibrosis distribution was further augmented by applying one or multiple affine transformations (translation, rotation and flipping). The fibrosis threshold value and the types of transformation were randomly selected. Lastly, the PVs were varied by assigning one of 6 different variants (which included changing PV size and position) (Muffoletto et al., 2021). This resulted in a total of 199 synthetic 2D LA tissue models in addition to the 122 patient-specific models, totalling 321 2D LA tissue models.

2.3. Atrial Tissue Modelling and AF simulation

Equation (1) represents the Fenton-Karma semi-physiological model, which consists of three ionic currents representing the overall ion current in the electrical dynamics of atria cells; I_{fi} represents the fast inward current (Na^+), I_{so} is the slow outward current (K^+) and I_{si} is the slow inward current (Ca^+) (Fenton and Karma, 1998):

$$I_{ion} = I_{fi} + I_{so} + I_{si} \tag{1}$$

Equation (2) is the standard monodomain equation to describe electrical wave propagation.

$$\frac{\partial V_m}{\partial t} = \nabla . D \nabla V_m - \frac{I_{ion}}{C_m} \tag{2}$$

Here V_m is the membrane potential, C_m is the membrane capacitance, D is a tensor that represents the diffusion of the electrical coupling within the tissue. Equation (2) with ion current determined in Equation (1) was solved using the forward Euler method with a finite-difference approximation of the Laplacian. Therefore, Equation (1) and Equation (2) were solved using each 2D tissue disk as a spatial domain to simulate electrical waves sustaining AF. Such waves in the form of rotors were generated using the standard cross-field protocol at 28ms into the simulation (Tobón et al., 2014). The numerical integration steps were 0.01 ms time step and 0.3 mm spatial step. Additionally, healthy tissue had a D value of 0.1 $mm^2 s^{-1}$ to match the physiological value of healthy myocardium tissue. Meanwhile, fibrotic tissue had D value of 0.015 $mm^2 s^{-1}$. The three most common ablation strategies were simulated to terminate persistent AF: PVI, FIBRO and rotor-based ROTOR RFCA strategies. The FIBRO strategy involved ablating the perimeter of the fibrotic tissue, while PVI consisted of ablating the circumference of the PVs and Rotor ablated the phase singularities of the electrical wave. The ablation strategy was deemed successful for a tissue if AF was terminated within 2000 ms and less than 40% of the tissue was ablated (Muffoletto et al., 2021). Therefore, using the stated simulation pipeline, the success of the three RFCA strategies was determined for AF simulations in the 2D LA tissues (real patient data and synthetic). Furthermore, since multiple strategies can be successful/unsuccessful for a given 2D LA tissue, the classification task was multi-label.

2.4. Deep Learning

We employed the convolutional neural network with hyperparameters based on the study by Muffoletto et al. as the basis of our interpretability framework (Muffoletto et al., 2021). Four convolutional layers of 32x32 filters followed by Rectified Linear Unit (ReLU) activation. The convolution block was followed by a Max pooling of pool size two and three linear layers (2048, 128 and 3 units, respectively) and had ReLU activation. A Dropout layer followed this at a rate of 0.8 and a sigmoid function (Paszke et al., 2019). Since we address a multi-label classification problem (i.e., multiple ablation strategies), we modified the loss function to be a mean-squared error tailored to perform multi-label classification for the three ablation strategies.

$$MSE(y_{score}, y) = \frac{\sum_{i=1}^{N} (y_{score}^{i} - y_i)^2}{N} \tag{3}$$

Equation (3) is the mean-squared error function formulation, where y_{score} is the predicted class score array and y is the RFCA strategy success ground truth (where 1 = success and 0 = unsuccessful). Meanwhile, N represents the number of classes/strategies (three in this study) and i is the index of a class in the class score array. To train and effectively test the CNN, a leave-one-out cross-validation was used where four folds were used to train the CNN explicitly, and the last fold was used as a validation set to select the optimal CNN model state (model state with the lowest loss during training) (Raschka, 2018). This was the approach employed in our previous study (Muffoletto et al., 2021). In total, there were 271 2D LA tissues in the leave-one-out cross-validation dataset (96 real and 175 synthetic). Within each fold the DL model was trained for 100 epochs using an ADAM optimiser with a learning rate of 1e-4 (Kingma and Ba, 2014). Finally, the optimal state was tested on a randomly chosen hold-out set of 50 2D LA tissues (26 real and 24 synthetic) from the total dataset to evaluate the DL model's performance. The process was repeated five times and tested on the same hold-out test set to evaluate the error for the model's prediction.

2.5. Interpretability

Three popular local post-hoc interpretability methods were used to interpret the CNN's predictions - GradCAM, occlusions and local interpretable model-agnostic explanations (LIME) (Selvaraju et al., 2017; Zeiler and Fergus, 2013; Ribeiro et al., 2016; Kokhlikyan et al., 2020). The DL model state from the most accurate fold of leave-one-out cross-validation was used to produce the FA maps for the three methods on the hold-out test set. Moreover, the GradCAM method was applied to the last convolutional layer of the CNN. Each FA map was thresholded above the respective map's average FA to highlight the most informative features. Furthermore, these informative features were compared to the ablation lesions from the RFCA simulations to evaluate if the DL model focuses on relevant biomedical features by measuring the percentage of simulation RFCA ablation lesions within thresholded informative regions (lesion percentage) and the Jacquard index (IoU) of the informative regions and ablation lesions.

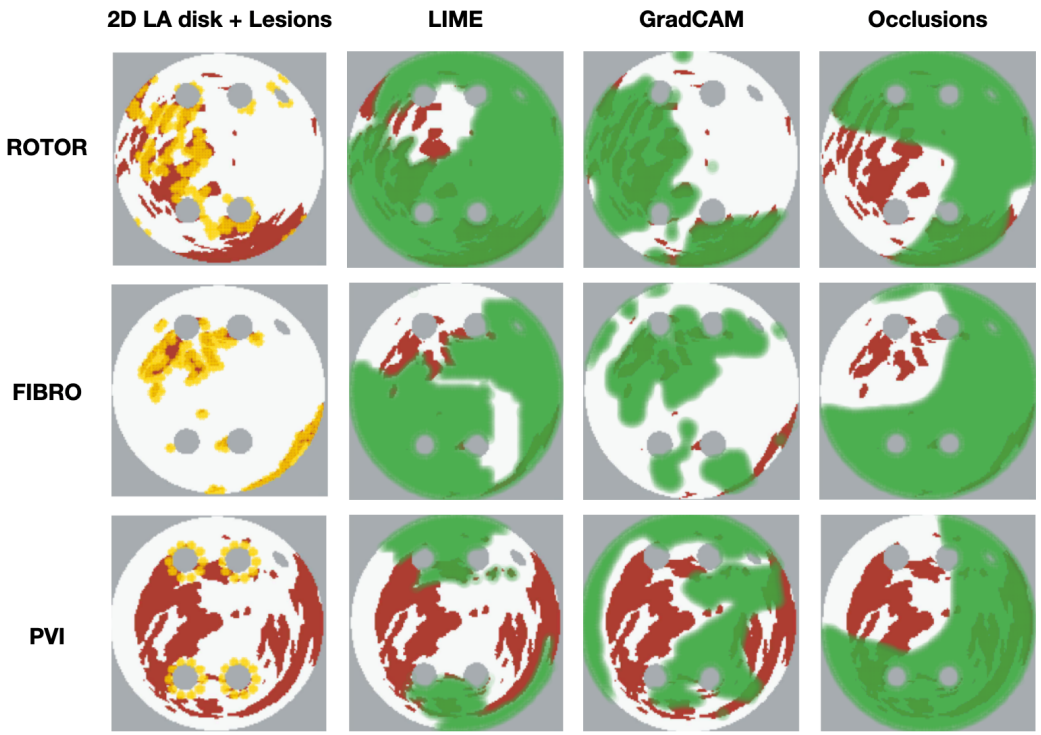

Figure 1: Diagram of 2D LA tissues with highlighted feature attribution maps. White areas in the 2D tissues are healthy tissue and red areas are fibrosis. Ablation lesion locations known from simulations are shown (yellow) for all three RFCA strategies, along with respective FA maps for LIME, GradCAM and occlusions and highlighted thresholded informative regions (green).

Table 1: Mean area under the receiver operating characteristic curve (AUC) score, recall and precision on independent hold-out test set (with standard deviation) for each RFCA strategy.

Strategy	**AUC**	**Recall**	**Precision**
PVI	0.78 ± 0.04	0.35 ± 0.07	0.68 ± 0.28
FIBRO	0.92 ± 0.02	0.89 ± 0.03	0.82 ± 0.02
ROTOR	0.77 ± 0.02	0.93 ± 0.04	0.76 ± 0.02

Table 2: Mean lesion percentage and IoU of the informative region and ablation lesions with errors (standard deviation for each FA map method and RFCA strategy.

Ablation Strategy	**Method**	**Lesion Percentage**	**IoU**
PVI	LIME	0.44 ± 0.24	**0.077 ± 0.023**
	Occlusions	**0.55 ± 0.15**	0.065 ± 0.17
	GradCAM	0.47 ± 0.17	0.063 ± 0.029
FIBRO	LIME	0.57 ± 0.19	0.18 ± 0.09
	Occlusions	0.45 ± 0.14	0.19 ± 0.11
	GradCAM	**0.62 ± 0.25**	**0.256 ± 0.11**
ROTOR	LIME	0.62 ± 0.16	0.12 ± 0.07
	Occlusions	0.53 ± 0.16	0.14 ± 0.06
	GradCAM	**0.71 ± 0.13**	**0.20 ± 0.08**

3. Results

As shown in Table 2, GradCAM is characterised with the highest lesion percentage and IoU for the FIBRO and ROTOR strategies. Additionally, Figure 1 shows that in FA maps obtained with GradCAM for ROTOR and FIBRO, the informative regions coincide with most ablation lesions. Using the Wilcoxon signed-rank test, ROTOR strategy lesion percentage for GradCAM was significantly greater ($p < 0.017$ using Bonferroni correction) than that for occlusions, but not for LIME ($p = 3.1e\text{-}8$ and $p = 0.0253$, respectively). Moreover, for FIBRO strategy, the lesion percentage for GradCAM was significantly higher than that for the occlusions method, but again not for LIME ($p = 4.0$ e-6, $p = 0.06$, respectively). However, the IoU scores for GradCAM were significantly greater ($p < 0.017$) than those for occlusions and LIME for ROTOR ($p = 3.3e\text{-}6$ and $p = 2.1e\text{-}9$, respectively) and FIBRO ($p = 4.2e\text{-}6$ and $p = 1.6e\text{-}9$, respectively). Therefore, GradCAM produces more interpretable FA maps than LIME (for FIBRO and ROTOR) as the informative regions are more focused on areas with a high number of ablation lesions – reflected in GradCAM having a significantly greater IoU score than LIME. These findings show little dependence on the threshold between informative and uninformative regions. As shown in Appendix F, when the threshold value is set to 25% above and below the average feature attribution, Grad-CAM still has the highest lesion percentage and IoU compared to LIME and Occlusions for the ROTOR and FIBRO strategies. For the PVI strategy, the occlusions method provided FA maps with the greatest lesion percentage and LIME FA maps had the highest IoU score. The difference in best FA map methods in terms of lesion percentage and IoU score can be seen in Figure 1, as informative regions in the occlusions' FA maps cover vast area highlighting the ablation lesions, but are not local to the PVs. Meanwhile, the LIME FA map highlights areas around the PVs, but does not cover many ablation lesions.

4. Discussion and Conclusion

Predicting RFCA outcomes from imaging data is a challenging task, as shown by Kim et al., who predicted AF recurrence post-RFCA with a 0.61 accuracy from a CNN which used a combination of MRI data and patient demographics (Kim et al., 2020). Therefore, developing a successful DL model to predict RFCA outcomes in AF simulations is the natural first step to predict real RFCA outcomes in AF patients. Hence, this study (i) demonstrates a multi-label classification CNN for the success of ablation strategies in patient-specific simulations of AF, with AUC scores of 0.92 ± 0.02 for FIBRO, 0.78 ± 0.04 for PVI and 0.77 ± 0.02 for ROTOR, and (ii) explores different methods of DL interpretability in the classification, with GradCAM shown to provide the most interpretable FA maps for the ROTOR and FIBRO strategy, suggesting that the DL model utilises pro-arrhythmogenic regions to make its prediction. A possible explanation for why GradCAM performed better than the other methods is that LIME is susceptible to unstable generated interpretations due to random perturbations and feature selection. Moreover, LIME and occlusions are not class discriminative – meaning that they cannot localise the class (RFCA strategy) within the feature space. Grad-CAM is gradient-based (does not randomise parameters to obtain FA maps) and is class discriminative, allowing it to localise pro-arrhythmogenic regions more faithfully than LIME and occlusions (Zafar and Khan, 2021; Selvaraju et al., 2017).

Classification of the PVI strategy was difficult to interpret. A possible reason for this difficulty is that the PVI strategy in the clinic is based on ablating PV triggers that typically initiate AF. However, these initial PV triggers were not present in the 2D LA tissue models. Therefore, the three FA methods could not produce interpretable maps. Lastly, limitations of the study include using a large amount of synthetic LA data and 2D data. The RFCA strategy that has the highest magnitude of lesion percentage (ROTOR) also had the lowest AUC score in testing (Table 1). Showing that the interpretability of a FA map does not increase with the accuracy of the strategies prediction. This observation demonstrates that the need for interpretability in RFCA strategy prediction likely goes beyond FA, and in future work, we will investigate the incorporation of confidence in prediction outputs to enable our method to be used as a decision support tool to help clinicians select the appropriate therapy. Future work should also focus on; exclusively real patient LA data, extension to 3D simulations and investigating intrinsically interpretable DL models such as ICAM developed by Bass et al. (Bass et al., 2020).

The purpose of FA maps is not to be directly applied in the clinic to predict ablation lesions in a patient – but to explain why the DL approach is making a certain prediction, and to increase clinical confidence in this approach. The EU's GDPR requires an explanation for any algorithmic decision used in patient care. Unclear what form this explanation should take; we believe our work represents a significant step to meet this requirement. Most of the ablation lesions in our study coincided with informative regions of the GradCAM FA maps (ROTOR and FIBRO, Figure 1), suggesting that the DL model learns from structural features of MR images even without knowledge of the LA function – since CNN training is blinded to the locations of RFCA lesions. The explanation is that the structural features constitute pro-arrhythmogenic LA regions (e.g., fibrotic regions are well-known for their ability to harbour rotors sustaining AF) that need to be targeted by ablation. Such mechanistic explanations should increase clinician's confidence in using the DL predictions.

Acknowledgments

This research was supported by the UK Medical Research Council.

References

Mahmood I Alhusseini, Firas Abuzaid, Albert J Rogers, Junaid A B Zaman, Tina Baykaner, Paul Clopton, Peter Bailis, Matei Zaharia, Paul J Wang, Wouter-Jan Rappel, and Sanjiv M Narayan. Machine learning to classify intracardiac electrical patterns during atrial fibrillation. *Circ. Arrhythm. Electrophysiol.*, 13(8):e008160, August 2020.

Cher Bass, Mariana da Silva, Carole Sudre, Petru-Daniel Tudosiu, Stephen M Smith, and Emma C Robinson. ICAM: Interpretable classification via disentangled representations and feature attribution mapping. June 2020.

Johannes Brachmann, John D Hummel, David J Wilber, Anne E Sarver, Joshua Rapkin, S Shpun, and T Szili-Torok. Prospective randomized comparison of rotor ablation vs conventional ablation for treatment of persistent atrial fibrillation—the REAFFIRM trial. *Heart Rhythm*, 16(6):963–965, 2019.

S A Chen, M H Hsieh, C T Tai, C F Tsai, V S Prakash, W C Yu, T L Hsu, Y A Ding, and M S Chang. Initiation of atrial fibrillation by ectopic beats originating from the pulmonary veins: electrophysiological characteristics, pharmacological responses, and effects of radiofrequency ablation. *Circulation*, 100(18):1879–1886, November 1999.

Henry Chubb, Rashed Karim, Sébastien Roujol, Marta Nuñez-Garcia, Steven E Williams, John Whitaker, James Harrison, Constantine Butakoff, Oscar Camara, Amedeo Chiribiri, Tobias Schaeffter, Matthew Wright, Mark O'Neill, and Reza Razavi. The reproducibility of late gadolinium enhancement cardiovascular magnetic resonance imaging of post-ablation atrial scar: a cross-over study, 2018.

Sumeet S Chugh, Rasmus Havmoeller, Kumar Narayanan, David Singh, Michiel Rienstra, Emelia J Benjamin, Richard F Gillum, Young-Hoon Kim, John H McAnulty, Jr, Zhi-Jie Zheng, Mohammad H Forouzanfar, Mohsen Naghavi, George A Mensah, Majid Ezzati, and Christopher J L Murray. Worldwide epidemiology of atrial fibrillation: a global burden of disease 2010 study. *Circulation*, 129(8):837–847, February 2014.

Flavio Fenton and Alain Karma. Vortex dynamics in three-dimensional continuous myocardium with fiber rotation: Filament instability and fibrillation. *Chaos*, 8(1):20–47, March 1998.

R G Hart and J L Halperin. Atrial fibrillation and stroke : concepts and controversies. *Stroke*, 32(3):803–808, March 2001.

Ju Youn Kim, Younghoon Kim, Gil-Hwan Oh, Sun Hwa Kim, Young Choi, Youmi Hwang, Tae-Seok Kim, Sung-Hwan Kim, Ji-Hoon Kim, Sung-Won Jang, Yong-Seog Oh, and Man Young Lee. A deep learning model to predict recurrence of atrial fibrillation after pulmonary vein isolation. *J. Interv. Card. Electrophysiol.*, 21(1):1–7, November 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch. September 2020.

Chih-Min Liu, Shih-Lin Chang, Hung-Hsun Chen, Wei-Shiang Chen, Yenn-Jiang Lin, Li-Wei Lo, Yu-Feng Hu, Fa-Po Chung, Tze-Fan Chao, Ta-Chuan Tuan, Jo-Nan Liao, Chin-Yu Lin, Ting-Yung Chang, Cheng-I Wu, Ling Kuo, Mei-Han Wu, Chun-Ku Chen, Ying-Yueh Chang, Yang-Che Shiu, Henry Horng-Shing Lu, and Shih-Ann Chen. The clinical application of the deep learning technique for predicting trigger origins in patients with paroxysmal atrial fibrillation with catheter ablation. *Circ. Arrhythm. Electrophysiol.*, 13 (11):e008518, November 2020.

Ross Morgan, Michael A Colman, Henry Chubb, Gunnar Seemann, and Oleg V Aslanidi. Slow conduction in the border zones of patchy fibrosis stabilizes the drivers for atrial fibrillation: Insights from Multi-Scale human atrial modeling. *Front. Physiol.*, 7:474, October 2016.

Marica Muffoletto, Xiao Fu, Aditi Roy, Marta Varela, Paul A Bates, and Oleg V Aslanidi. Development of a deep learning method to predict optimal ablation patterns for atrial fibrillation. In *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–4, July 2019.

Marica Muffoletto, Ahmed Qureshi, Aya Zeidan, Laila Muizniece, Xiao Fu, Jichao Zhao, Aditi Roy, Paul A Bates, and Oleg Aslanidi. Toward Patient-Specific prediction of ablation strategies for atrial fibrillation using deep learning, 2021.

Sanjiv M Narayan, Tina Baykaner, Paul Clopton, Amir Schricker, Gautam G Lalani, David E Krummen, Kalyanam Shivkumar, and John M Miller. Ablation of rotor and focal sources reduces late recurrence of atrial fibrillation compared with trigger ablation alone: extended follow-up of the CONFIRM trial (conventional ablation for atrial fibrillation with or without focal impulse and rotor modulation). *J. Am. Coll. Cardiol.*, 63 (17):1761–1768, May 2014.

Naoya Oketani, Hitoshi Ichiki, Yasuhisa Iriki, Hideki Okui, Maenosono Ryuichi, Namino Fuminori, Yuichi Ninomiya, Sanemasa Ishida, Shuichi Hamasaki, and Chuwa Tei. Catheter ablation of atrial fibrillation guided by complex fractionated atrial electrogram mapping with or without pulmonary vein isolation, 2012.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, High-Performance deep learning library. In H Wallach, H Larochelle, A Beygelzimer, F dAlché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Ahmed Qureshi, Aditi Roy, Henry Chubb, Adelaide de Vecchi, and Oleg Aslanidi. Investigating strain as a biomarker for atrial fibrosis quantified by patient cine MRI data. In *2020 Computing in Cardiology*, pages 1–4, September 2020.

Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. November 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you? In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, August 2016. ACM.

Aditi Roy, Marta Varela, and Oleg Aslanidi. Image-Based computational evaluation of the effects of atrial wall thickness and fibrosis on re-entrant drivers for atrial fibrillation. *Front. Physiol.*, 9:1352, October 2018.

Aditi Roy, Marta Varela, Henry Chubb, Robert MacLeod, Jules C Hancox, Tobias Schaeffter, and Oleg Aslanidi. Identifying locations of re-entrant drivers from patient-specific distribution of fibrosis in the left atrium. *PLoS Comput. Biol.*, 16(9):e1008086, September 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Catalina Tobón, Laura C Palacio, Juan E Duque, Esteban A Cardona, Juan P Ugarte, Andrés Orozco-Duque, Miguel A Becerra, Javier Saiz, and John Bustamante. Simple ablation guided by ApEn mapping in a 2D model during permanent atrial fibrillation. In *Computing in Cardiology 2014*, pages 1029–1032, September 2014.

Courtney M Townsend and David C Sabiston. *Sabiston Review of Surgery*. Saunders, 2001.

Marta Varela, Ross Morgan, Adeline Theron, Desmond Dillon-Murphy, Henry Chubb, John Whitaker, Markus Henningsson, Paul Aljabar, Tobias Schaeffter, Christoph Kolbitsch, and Oleg V Aslanidi. Novel MRI technique enables Non-Invasive measurement of atrial wall thickness. *IEEE Trans. Med. Imaging*, 36(8):1607–1614, August 2017.

Yubing Wang, Yanping Xu, Zhiyu Ling, Weijie Chen, Li Su, Huaan Du, Peilin Xiao, Zengzhang Liu, and Yuehui Yin. GW28-e1219 radiofrequency catheter ablation for paroxysmal atrial fibrillation: over 3-year follow-up outcome, 2017.

David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher Em Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. Clinical applications of machine learning algorithms: beyond the black box. *BMJ*, 364:l886, March 2019.

Steven E Williams, Catalina Tobon-Gomez, Maria A Zuluaga, Henry Chubb, Constantine Butakoff, Rashed Karim, Elena Ahmed, Oscar Camara, and Kawal S Rhode. Standardized unfold mapping: a technique to permit left atrial regional data display and analysis. *J. Interv. Card. Electrophysiol.*, 50(1):125–131, October 2017.

Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, Pheng-Ann Heng, Dong Ni, Caizi Li, Qianqian Tong, Weixin Si, Elodie Puybareau, Younes Khoudli, Thierry Géraud, Chen Chen, Wenjia Bai, Daniel Rueckert, Lingchao Xu, Xiahai Zhuang, Xinzhe Luo, Shuman Jia, Maxime Sermesant, Yashu Liu, Kuanquan Wang, Davide Borra, Alessandro Masci, Cristiana Corsi, Coen de Vente, Mitko Veta, Rashed Karim, Chandrakanth Jayachandran Preetha, Sandy Engelhardt, Menyun Qiao, Yuanyuan Wang, Qian Tao, Marta Nuñez-Garcia, Oscar Camara, Nicolo Savioli, Pablo Lamata, and Jichao Zhao. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Med. Image Anal.*, 67:101832, January 2021.

Wang Yubing, Xu Yanping, Ling Zhiyu, Chen Weijie, Su Li, Du Huaan, Xiao Peilin, Liu Zengzhang, and Yin Yuehui. Long-term outcome of radiofrequency catheter ablation for persistent atrial fibrillation. *Medicine*, 97(29):e11520, July 2018.

Muhammad Rehman Zafar and Naimul Khan. Deterministic local interpretable Model-Agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, 3(3):525–541, June 2021.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. November 2013.

Appendix A. Data Preprocessing

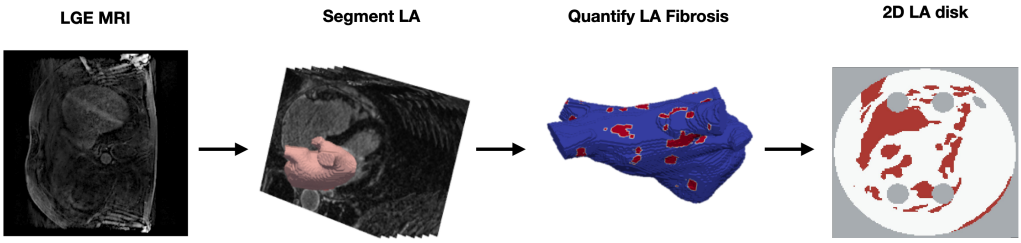

Figure 2: Workflow of 2D LA tissue disk processing pipeline. The figure illustrates the process of how the 2D LA tissue models are obtained from LGE MRI by LA segmentation, thresholding fibrosis from healthy tissue and mapping onto 2D tissues.

Appendix B. 2D LA Simulation Justification

The reason for using 2D LA simulations in our study was the efficiency in providing the needed proof of concept: (i) running 3D atrial simulations for several hundred cases would take years of simulations on a supercomputer, which would be a misuse of computational power; (ii) standardised 2D unfolded LA images allowed us to easily generate a large number of additional synthetic images, which is crucial for training CNNs. Hence, image-based 2D LA models provided a sensible balance between realistic details (such as fibrosis distributions) and computational efficiency (i.e., the ability to run a large number of simulations and train the CNN). Our previous work has shown that atrial wall thickness is distributed more or less evenly in the LA outside of PVs (Varela et al., 2017) and that slow conduction in fibrotic areas is the main determinant of the rotor dynamics (Roy et al., 2018; **?**).

Appendix C. Future Direction and Clinical Translation

The main benefit of the DL approach is the ability of CNNs to make fast predictions from a combination of structural (imaging) and functional (simulation) data. While image-based simulations can provide useful information about structure-function relationships during AF, its downsides include (i) huge computational power needed to simulate multiple AF scenarios in the detailed 3D atrial models, and (ii) the need to rerun the models each time novel data is integrated into them, which makes the application of models in a clinical setting impractical. DL can overcome these limitations and (after careful validation/integration of clinical data) provide a fast and flexible tool to predict ablation strategies for a large patient population. The simulations in our study only provide labels (RFCA strategy success) for a patient image, and the labelled image is used to train the CNN. The trained CNN can

then make predictions about suitable RFCA strategies from patient images only. This builds confidence in the DL approach, which can then be used for images labelled using real data from patients (should such data be available/reliable) and help make clinically valid predictions. Moreover, the approach is not restricted to three specific RFCA strategies considered in our study – it can include any other promising strategies, and a trained CNN can help pick the most suitable one for each patient.

Appendix D. Deep Learning Model

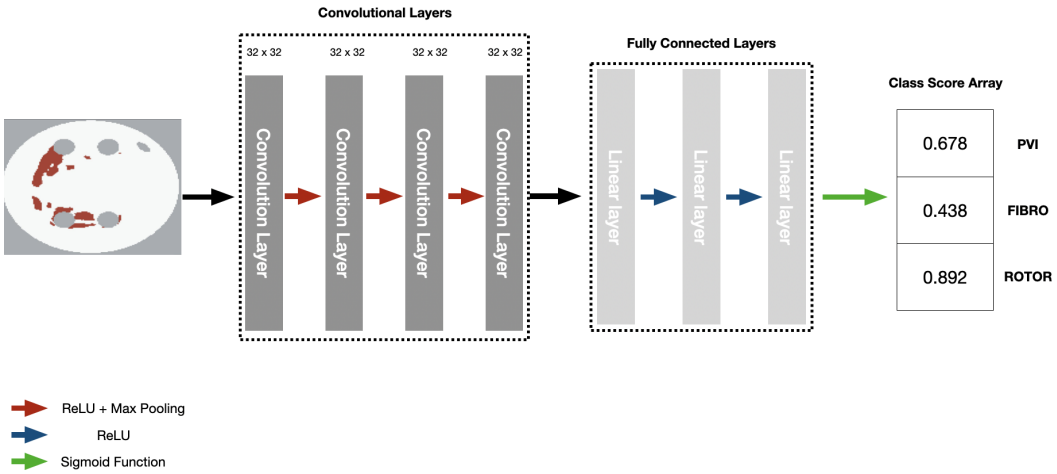

Figure 3: Diagram of CNN with parameters to predict RFCA simulation strategy success from 2D left atrium tissue disk.

Appendix E. Synthetic and Real Data Accuracy Metrics

Table 3: Mean AUC score (with standard deviation) of DL model trained with real data only and with synthetic and real data from a leave-one-out cross-validation on a hold-out test (of $\sim$ 20 % of the total of the respective dataset).

Strategy	**Real Data**	**Real + Synthetic Data**
PVI	0.73 $\pm$ 0.03	**0.78 $\pm$ 0.04**
FIBRO	0.80 $\pm$ 0.03	**0.92 $\pm$ 0.02**
ROTOR	0.49 $\pm$ 0.06	**0.77 $\pm$ 0.02**

To train a DL model effectively, a considerable amount of data is needed. There are only 122 real data subjects in this study, while if we combine this with synthetic data subjects, we get a total dataset of 321 subjects. Furthermore, this is reflected in (Table 3), where the DL model has higher prediction accuracy (AUC) with a real and synthetic dataset.

Table 4: Mean AUC score on independent hold-out test set (with standard deviation) for each RFCA strategy and type of data.

Strategy	**Real Data**	**Real + Synthetic Data**
PVI	0.67 $\pm$ 0.03	0.78 $\pm$ 0.04
FIBRO	0.85 $\pm$ 0.02	0.92 $\pm$ 0.02
ROTOR	0.62 $\pm$ 0.05	0.77 $\pm$ 0.02

Appendix F. Sensitivity to FA Maps Thresholding

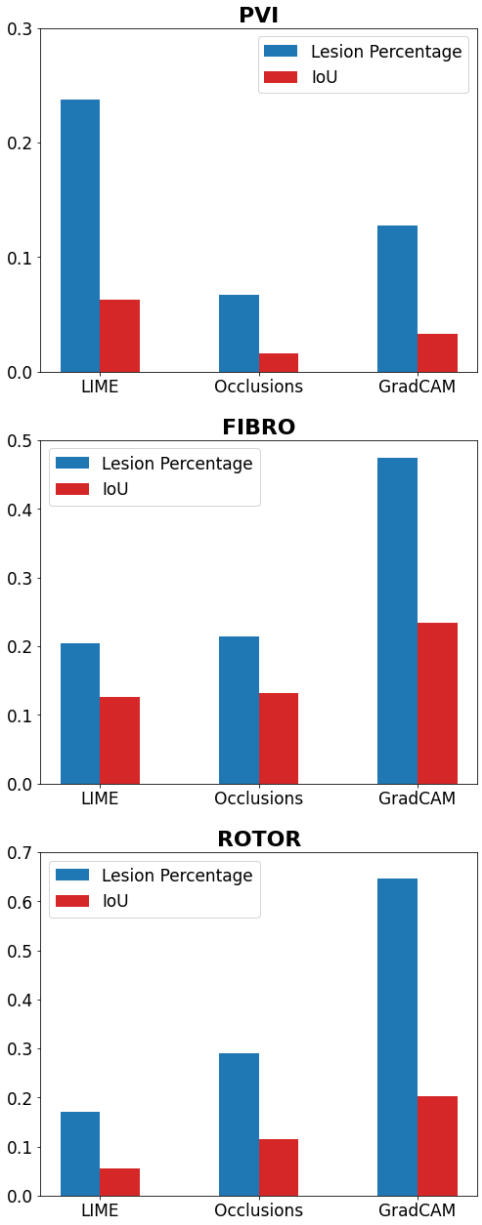

Figure 4: Lesion percentage and IoU values for each interpretability method and ablation strategy with informative threshold value 25 % above the average.

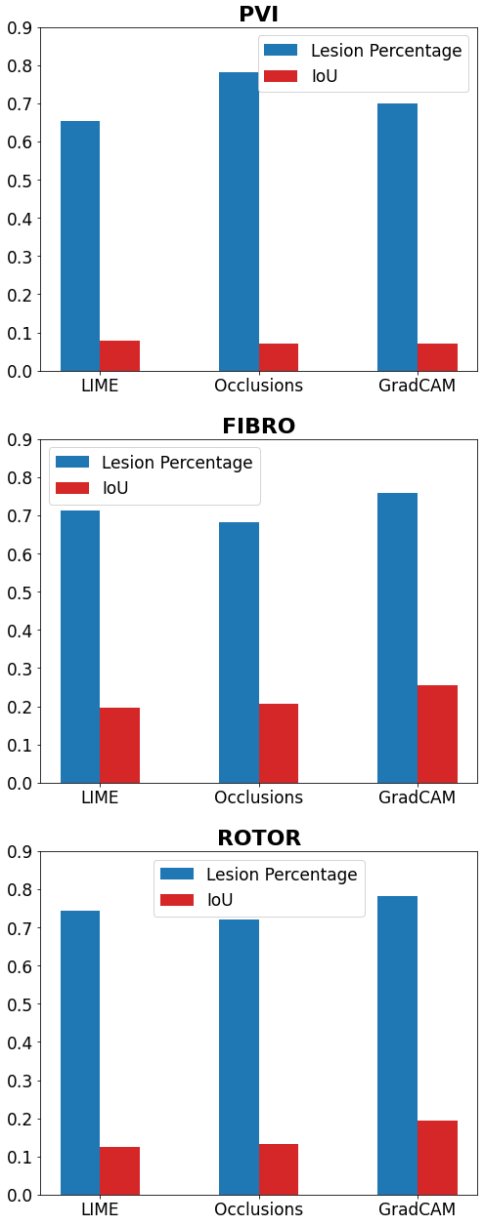

Figure 5: Lesion percentage and IoU values for each interpretability method and ablation strategy with informative threshold value 25 % below the average.

Appendix G. Examples of FA maps

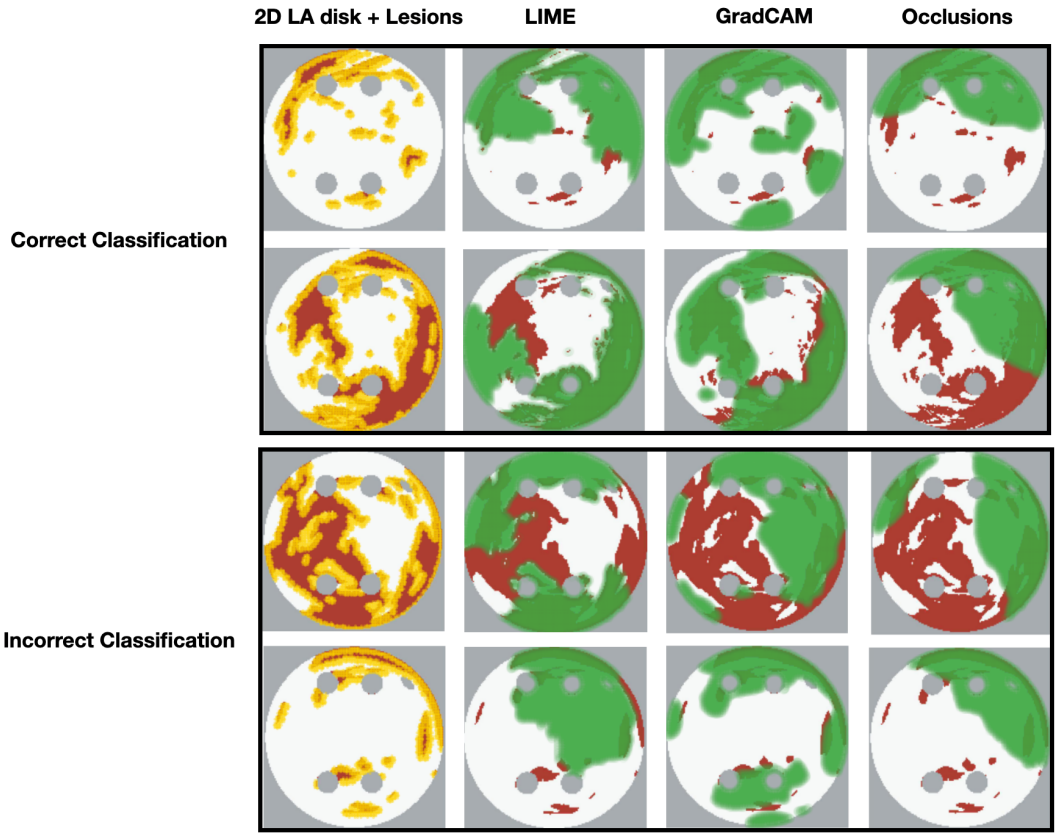

Figure 6: Correct and incorrect classification examples of FA maps (LIME, GradCAM and occlusions) for FIBRO.

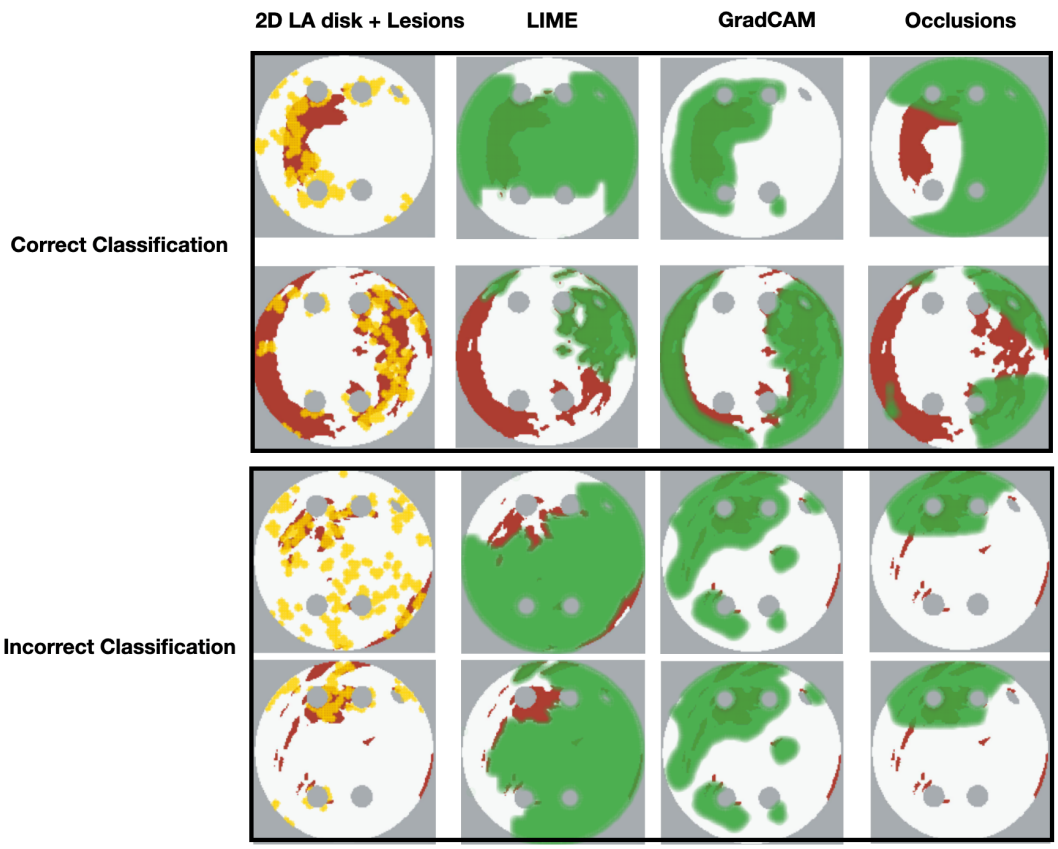

Figure 7: Correct and incorrect classification examples of FA maps (LIME, GradCAM and occlusions) for ROTOR.

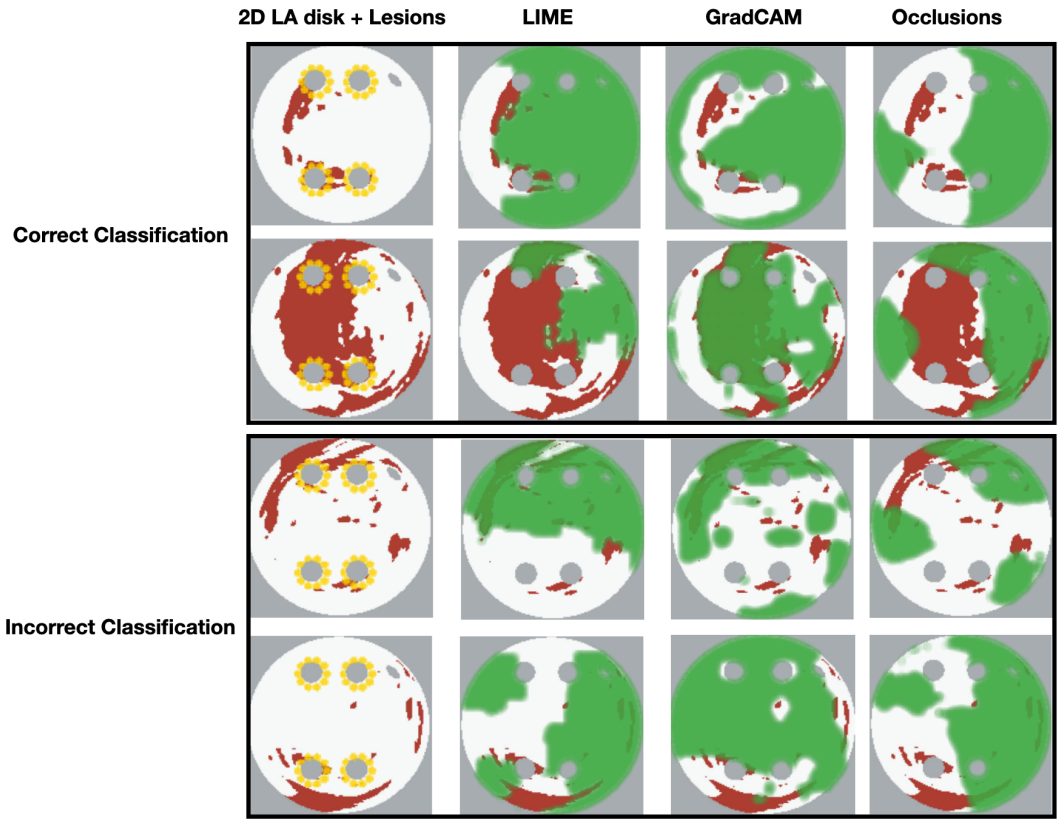

Figure 8: Correct and incorrect classification examples of FA maps (LIME, GradCAM and occlusions) for PVI.