Localizing cardiac dyssynchrony in M-mode echocardiography with attention maps

Marta Saiz-Vivó, Isaac Capallera, Nicolas Duchateau, Gabriel Bernardino, Gemma Piella, Oscar Camara

Localizing cardiac dyssynchrony in M-mode echocardiography with attention maps

Marta Saiz-Vivó[1][0000−0003−0448−9045], Isaac Capallera[1], Nicolas Duchateau[2,3][0000−0001−8803−2004], Gabriel Bernardino[1][0000−0001−8741−2566], Gemma Piella[1][0000−0001−5236−5819], and Oscar Camara[1][0000−0002−5125−6132]

[1] Physense, BCN Medtech, Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain
{marta.saiz, gabriel.bernardino, gemma.piella,oscar.camara}@upf.edu, isaac.capallera01@estudiant.upf.edu
[2] Univ Lyon, Université Claude Bernard Lyon 1, INSA-Lyon, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, Lyon, France
[3] Institut Universitaire de France (IUF)
nicolas.duchateau@creatis.insa-lyon.fr

Abstract. Cardiac Resynchronization Therapy (CRT) is a treatment aimed at restoring the electrical synchronization in patients with heart failure and intraventricular conduction delay. However, over 30% of patients do not respond to CRT. Septal Flash (SF), an abnormality characterized by a rapid inward-outward abnormal motion at early systole, has been linked to an improved response to CRT in patients with Left Bundle Branch Block (LBBB). In clinical practice, the detection of SF is usually performed manually through echocardiographic acquisitions, which is subjective and dependent on the operator's experience. To address this issue, a deep classification model for automatic SF detection from 2D anatomical M-mode images has been proposed. Additionally, this work focuses on SF localization from gradient-based attention maps, which provide a visual explanation of the output prediction of the model. Two models based on convolutional neural networks (CNNs) were trained with original and cropped M-modes from 143 patients, and achieved an accuracy of 0.83 and 1.0 respectively, on 29 testing patients. The attention map visualization showed that in SF cases, the models effectively identified the discriminant regions, while in non-SF cases, the maps appeared more dispersed. Further research is necessary to quantitatively evaluate the attention map results.

Keywords: Localization · M-mode echocardiography · heart failure · cardiac resynchronization therapy · attention maps

1 Introduction

Cardiac resynchronization therapy (CRT) is a medical treatment developed to recover electrical synchrony in heart failure patients with intraventricular conduction delay, characterized by disordered ventricular contractions and adverse

clinical prognosis [3]. However, a significant amount of patients do not respond well to the treatment, with CRT non-response rate still over 30% [15]. Among the identified mechanisms amenable to CRT response, septal flash (SF) is indicative of electrically mediated dyssynchrony such as left bundle branch block (LBBB) [17]. It is characterized by a fast inward-outward motion of the septum that occurs during early systole or isovolumetric contraction period and mostly ends before aortic valve opening. Due to the early contraction of the right ventricle (RV), a transseptal pressure difference is created that tethers the septum towards the left ventricle (LV) [3]. Several studies [1, 6] have shown SF to be a strong predictor of CRT response in patients with LBBB, thus accurate identification of SF is of interest for CRT patient selection and response prediction.

In clinical practice, SF is commonly assesed through simple 'eye-balling' from 2D transthoracic echocardiography [3]. However, visual assessment methods rely heavily on operator's experience, which could lead to subjective diagnosis. Consequently, several works have focused on more automatic detection of SF. For example, statistical atlases and representation learning were proposed to automatically detect and quantify abnormal ventricular motion, including SF [8, 11, 16, 7]. With recent computational advancements, deep learning methods have emerged as powerful tools for image analysis tasks such as object classification and detection, segmentation and registration, offering new perspectives for the automatic characterization of mechanisms relevant to CRT response [13]. In the field of medical image classification, several authors employ convolutional neural network (CNN)- based methods. However, these models are often seen as 'black boxes' offering little to no explanation on why the output prediction was chosen. In the context of our application, a CRT response prediction model was proposed by Puyol-Antón et al. [12] from cardiac magnetic resonance images with an interpretable variational autoencoder. More generally, many works build upon gradient-based class activation mapping (CAM) approaches, such as Grad-CAM++ [5, 18], to offer a visual explanation of the output prediction via localization maps.

Regarding SF detection, Qu et al. [13] recently proposed a CNN-based approach to classify 2D+t B-mode sequences with SF. To capture both spatial and temporal context (required to detect the abnormal septal motion) the authors proposed a linear attention cascaded net (LACnet) with CNN-based encoders and a LSTM- based decoder for temporal feature extraction. However, the design of a complex deep learning architecture with temporal units capable of handling long-term dependencies was required, and handling sequential image data may lead to increased computational complexity compared to simple 2D image processing. Furthermore, the localization of the SF region, employed in the model's output prediction, was not provided.

Motion-mode (M-mode) is an echocardiography technique that provides a one-dimensional view of the tissue along a specific ultrasound beam in different temporal instances, enabling the analysis of tissue motion from 2D images, where the x-axis represents time. Due to its high temporal resolution, it is often used for septal motion analysis and SF detection [6]. A variant of M-mode, known

as anatomical M-mode [4], allows the extraction of M-mode images along freely specified lines as a post-processing.

In this work, we obtain localization maps of SF prediction from 2D echocardiographic sequences. We take advantage of virtual M-mode images to use gradient-based class activation maps that consider both spatial and temporal aspects of the sequence. We thoroughly evaluate their ability to localize SF by their provision of discriminant image regions employed in the model's output prediction, which can help to better interpret echocardiographic images, in a longer term objective of assisting less experienced clinicians to find SF and develop an interactive process.

2 Methods

2.1 Dataset

In this study, the dataset was provided by the Hospital Clínic de Barcelona. It consisted of 2D ultrasound sequences (GE Vingmed Ultrasound A.S., Horten, Norway) with the corresponding electrocardiogram (ECG) from 89 CRT patients acquired both at baseline (n = 89) and 12 months follow-up (n = 89) after the CRT implant. The CRT inclusion criteria corresponded to the international guidelines: symptomatic heart failure with QRS duration >120 ms, and NYHA classification III-IV or NYHA II who covered less than 500 meters in the 6 minutes walking test. The transthoracic echocardiography sequences were acquired in an apical four-chamber view, useful for the assessment of abnormal septal motion. After the deletion of corrupted sequences, we analyzed a total of 143 sequences of two subpopulations, with (n = 55) and without SF (n = 88).

The SF/non-SF labels were annotated by one experienced observer and controlled by another experienced observer. Septal segmentations at the end-diastolic frame were provided as a sequence of 62 points with spatial and temporal normalization, also manually marked.

2.2 Virtual M-mode generation

Ultrasound data was provided in echoline format and a B-mode scan reconstruction was performed with isotropic pixel size of 0.03 mm. The average frame rate was 60 fps. Using the R peaks of the ECG, the frames corresponding to one complete cardiac cycle were extracted (30-150 frames). The reconstructed image sequence was resized to 500×500 pixels, reducing computational complexity for posterior M-mode generation.

The generation of a virtual M-mode image post-hoc from a B-mode sequence implies reading 2D pixel samples along a specified line for each frame. To achieve this whilst capturing the septal motion, a cut plane in the direction perpendicular to the septum was applied to slice the B-scan considering the concatenated temporal frames as third dimension. Figure 1 illustrates the steps involved in the M-mode generation. Two points above and below the septal point of interest (mid-basal region), automatically extracted from the septum segmentation

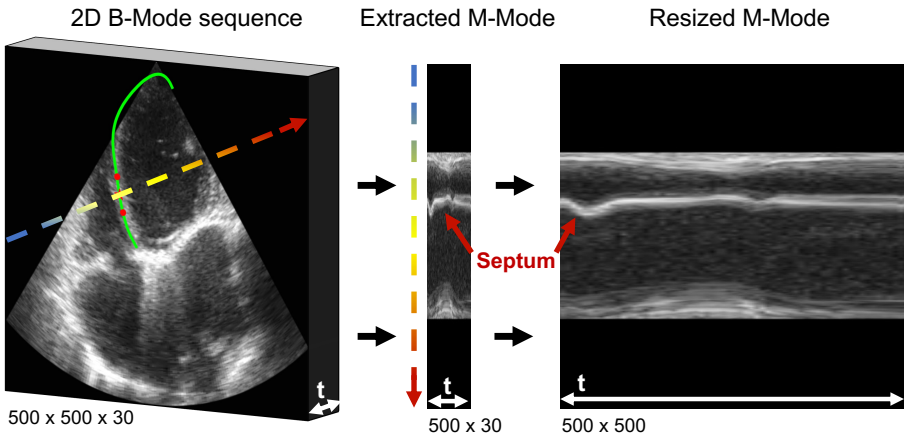

Fig. 1. Steps for virtual M-mode generation. Left: B-mode sequence with septum perpendicular vector (arrow) defined from two points (in red) sampled from septum segmentation (in green). Middle: synthetic M-mode. Right: resized M-mode

as the 10th and 20th point, were used to define the tangent direction of the mid-myocardium, from which the normal vector, in the direction towards the LV blood pool, was computed (Fig. 1-left). To slice the B-scan sequence as a 2D+t volumetric image, the *vtkImageReslice()* function from the VTK library in Python was employed. Fig. 1-middle illustrates the cut image, where the x-axis corresponds to the number of frames and the y-axis to the grayscale pixel intensities along the vector perpendicular to the septum, from the RV region (above) to the LV lateral wall (below). Finally, a temporal resampling was applied to generate the final M-mode image with 500×500 pixels (Fig. 1-right). The proposed approach enabled the semi-automatic extraction of anatomical M-modes, in contrast to conventional methods [4].

2.3 Preprocessing and data augmentation

To evaluate the impact of extracting the region of interest on SF recognition rate, an additional dataset was created by cropping the synthetic M-modes. Specifically, the M-mode images were cropped along the spatial axis, considering only the upper half region of the image scan that contains the septum. Moreover, given that SF occurs very early in the cardiac cycle, the M-mode was cropped along the temporal axis to retain the initial half of the cardiac cycle. Both the original and cropped M-mode datasets were used to develop SF classification models described in Sec. 2.4.

Data augmentation allows increasing the performance of image classification tasks. In this work, data augmentation was applied both offline and online for each dataset. The offline data augmentation increased the number of M-mode image training samples by a factor of 11 through sampling 5 additional septum

segmentation points above and below the original septal point and extracting the M-mode at different septal locations, without varying the normal vector. Random perturbations of the image contrast were applied online in the range $[0.5, 2]$, adding further image variations after every epoch. For image normalization purposes, mean subtraction was applied to every input image from the mean image of the training dataset. All input images were resized to 256×256 pixel size.

2.4 CNN model architecture and experimental settings

The CNN-based model implemented in this work adopted the DenseNet121 architecture, relevant for classification problems, as a backbone, and it was initialized with weights from RadImageNet [10] pre-trained models, designed for transfer learning in medical imaging applications to reduce computational and data expenses [2]. A fully connected layer with 2 output units and a softmax activation function were added to the DenseNet121 backbone to obtain the final binary classification output. After initializing with the pre-trained weights, the full architecture was retrained. The model was trained for 5 epochs (which was enough for convergence, probably thanks to pre-training) with an initial learning rate of 0.001, batch size of 4 and Adam optimization.

Two experiments were performed, on the cropped and original datasets with similar settings. The stratified training/test dataset split was performed with a ratio of 80:20 based on patient identifier to avoid placing highly correlated samples of the same patient in different groups and reduce overfitting. The same procedure was followed to generate the validation set from the training set. All experiments were implemented on NVIDIA Tesla T4 GPU with Python and Keras library. Once trained, the testing samples were employed to evaluate the diagnostic performance of the model with the area under the curve (AUC) and the accuracy as evaluation metrics.

2.5 Implementation of Grad-CAM++ attention method

To localize SF, the discriminative image regions used by the CNN to predict the output were identified through Grad-CAM++ attention maps [5] on the test samples. This algorithm is a visualization tool to better understand CNN model predictions through class-specific activation maps. The activation maps represent regions (with higher intensity) of the input image where the CNN model has "looked" to output the predicted class.

To estimate these regions, Grad-CAM++ computes the gradients of the predicted class with respect to the last convolutional layer. In this work the last convolutional layer of DenseNet121 was selected with feature map pixel dimensions of 8×8. The obtained heatmap was normalized in the range [0,1] and upsampled to input image resolution. The attention algorithm was implemented through the 'tf-keras-vis' visualization toolkit [9].

Table 1. Test accuracy and area under the curve (AUC) of convolutional neural network (CNN) models trained with virtual original M-mode and cropped M-mode images

CNN Model	Accuracy	AUC
w. original M-mode	0.83	0.95
w. cropped M-mode	1.00	1.00

3 Results

The models were trained and evaluated with 91 training patients and 23 validation patients (respectively 5005 and 253 samples due to off and on-line data augmentation), and 29 testing patients.

3.1 Results on the test dataset

To evaluate the trained models, the test dataset was preprocessed as done for the training set but without the data augmentation. Table 1 shows the test results for the models trained with the original M-mode and the cropped M-mode. The network trained with cropped images obtained the highest accuracy (100%), although accuracy for the model trained with original M-mode was reasonably high (83%). The AUCs for the cropped and original M-mode models were 1.0 and 0.95, respectively.

3.2 Grad-CAM++ visualization on the test dataset

Figure 2 shows Grad-CAM++ visualization results on 4 test patients with correct output predictions from original M-mode (Fig. 2, 1st and 3rd column) and cropped M-mode (Fig. 2, 2nd and 4th column) CNN models. On patients with SF (top rows) we observe that both CNN models accurately localize the region of SF occurrence, both along the spatial (vertical) and temporal (horizontal) axis, as the regions with high activation occur between mitral valve closure (MVC) and aortic valve opening (AVO). In the non-SF cases (bottom rows) the heatmaps appear with less localized attention, specially in the cropped M-mode model where other areas towards the RV region have been activated.

Figure 3 shows Grad-CAM++ visualization results for patients that were incorrectly predicted by the original M-mode model. As observed in the corresponding attention map, the network failed to detect the discriminant image features related to SF, thus leading to activated regions far from the septal region. In the first patient (Fig. 3, 1st column) the M-mode exhibits low contrast between septum and LV blood pool, which might explain the failure in the localization. The last two columns correspond to incorrect predictions of SF for non-SF cases. In the first one, the attention map appears focused on the septum, likely due to the detection of a perturbation similar to SF. On the second one, regions corresponding to the LV lateral wall appear mistakenly activated.

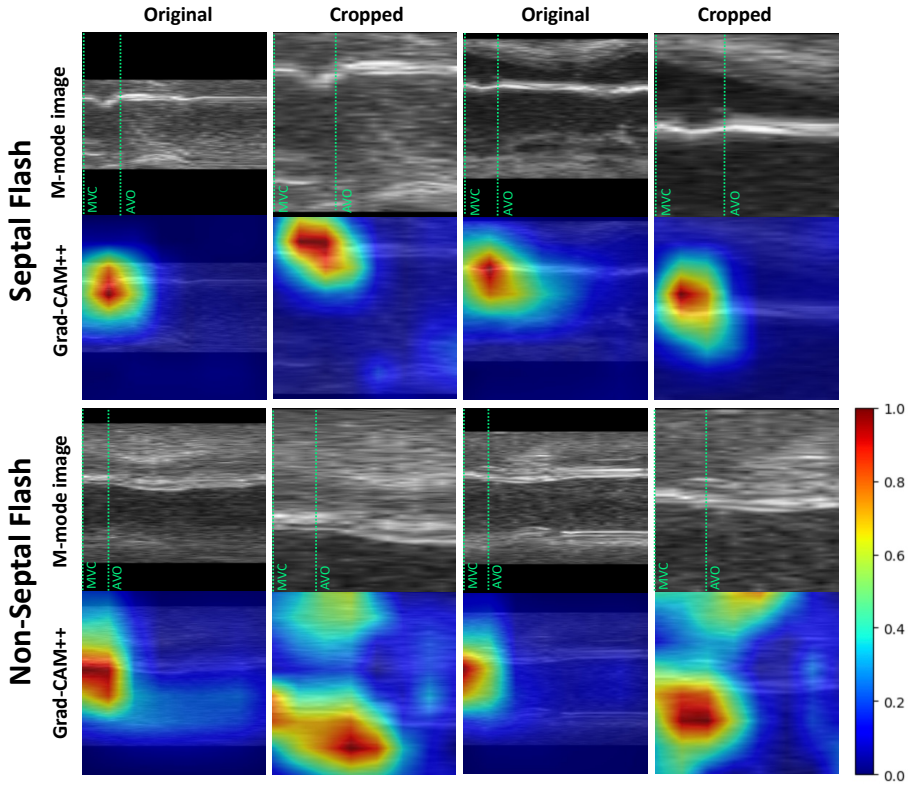

Fig. 2. Grad-CAM++ heatmap results for septal flash (SF) and non-SF correct predictions. 1st and 3rd column: extracted from original M-mode trained model. 2nd and 4th column: extracted from cropped M-mode trained model. MVC; mitral valve closure. AVO; aortic valve opening

4 Discussion

In this work we proposed the development of 2D SF detection models from generated anatomical M-mode images and the localization of SF through gradient-based attention maps.

In the test dataset, the model trained with original virtual M-modes obtained a reasonably high accuracy, with a higher precision for non-SF cases. The network's performance further improved to a perfect classification accuracy of 100% when trained with cropped images around the temporal and spatial region of interest. However, this increased (and perfect) accuracy may be due to the over-simplification of the SF classification problem by reducing the search area. The experiment demonstrated the feasibility of SF detection from the generated M-modes. Nonetheless, it is crucial to acknowledge the potential limitations of

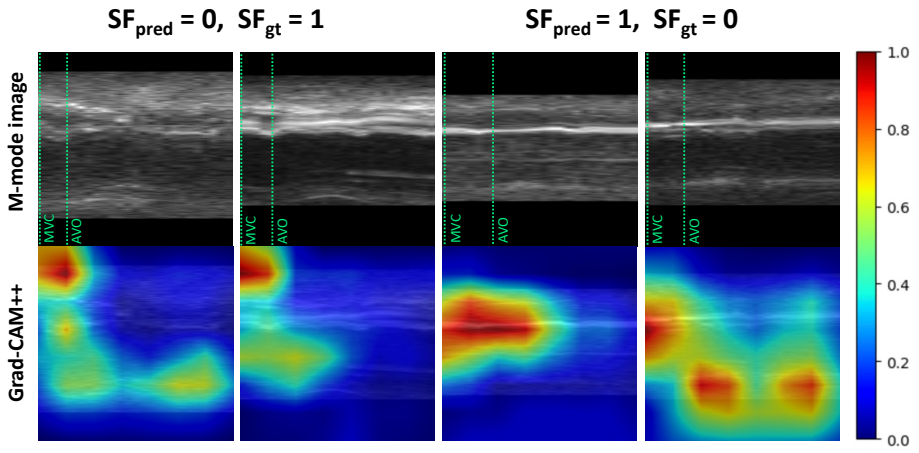

Fig. 3. Grad-CAM++ heatmap results for septal flash (SF) and non-SF incorrect predictions from original M-mode model. pred: predicted class, gt: ground-truth class. MVC; mitral valve closure. AVO; aortic valve opening

the experiment and model, including the generalizability of the results to different datasets.

The accuracy of the cropped M-mode classification model surpasses the performance reported in previous studies, including state-of-the-art CNN-based detection from US sequences (LACNet, 91% [13]) or machine learning-based classification from myocardial strain (linear discriminant analysis (LDA), 94% [16]). These findings indicate that the cropped M-mode model has the potential to be an effective approach for SF detection in medical imaging. However, each test dataset is unique and further evaluation with large and external testing databases is necessary to comprehensively understand the effectiveness and limitations of the model for SF classification.

Gradient-based class activation maps were employed to visualize the differentiating image regions for the network's output prediction. As observed in Fig. 2, for SF prediction (Fig. 2, 2nd row) the attention maps show a localized activation in the septum, specifically in the motion abnormality temporal region (between mitral valve closure and aortic valve opening). This suggests that the network has successfully learnt the discriminant image features characterizing SF in images with different septum localization and different septum/LV blood pool contrast, a necessary requirement for generalization purposes. On the other hand, for the case of non-SF patients (Fig. 2 bottom row), the attention maps show higher activations in regions other than the septum and in general, a higher dispersion. This is especially observed on the cropped M-mode activation maps (Fig. 2, 2nd and 4th column). Our hypothesis is that upon not finding the characteristic motion perturbation of SF, the networks focuses on random locations; however, further testing is needed to confirm this. Regarding the wrongly clas-

sified SF cases (Fig. 3, 1st and 2nd column), the attention maps have difficulty focusing on the SF region, as expected. Whilst for the non-SF cases, (Fig. 3, 3rd and 4th column), it appears that the network wrongly focused on specific temporal (3rd column) and spatial regions (4th column) not corresponding to the SF region, which could have led to the incorrect classification as SF.

The development of 2D SF detection models from generated anatomical M-modes has shown promising results. Nonetheless, concerning the proposed methodology of anatomical M-mode generation, the processing algorithms required to downsize the original reconstructed B-mode sequence for computational complexity reasons, with the corresponding loss of spatial resolution. Also, although superimposing gradient-based activation maps allows qualitative assessment of where the network focused its attention, to draw deeper conclusions, we propose as future work to extract quantitative metrics from the attention maps, similar to *Schöttl et al.* [14], and to analyse the differences in the attention distribution for SF and non-SF populations. Finally, the proposed SF detection model should be further evaluated with external databases. Further work will be conducted to evaluate SF localization from gradient-based approaches in models trained for CRT response prediction.

Acknowledgements This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016496 (SimCardioTest), from the French ANR (LABEX PRIMES of Univ. Lyon [ANR-11-LABX-0063] and the JCJC project "MIC-MAC" [ANR-19-CE45-0005]). G. Piella is supported by ICREA under the ICREA Academia programme. They are also grateful to M. Sitges and A. Doltra (Hospital Clínic de Barcelona, Spain) for providing the imaging data related to the studied population, and to B. Bijnens (ICREA Barcelona, Spain) for initial discussions on this topic.

References

1. Bennett, S., Tafuro, J., Duckett, S., Heatlie, G., Patwala, A., Barker, D., Cubukcu, A., Ahmed, F.Z., Kwok, C.S.: Septal flash as a predictor of cardiac resynchronization therapy response: A systematic review and meta-analysis. Journal of Cardiovascular Echography **31**(4), 198 (2021)
2. Cadrin-Chênevert, A.: Moving from imagenet to radimagenet for improved transfer learning and generalizability. Radiology: Artificial Intelligence **4**(5), e220126 (2022)
3. Calle, S., Delens, C., Kamoen, V., De Pooter, J., Timmermans, F.: Septal flash: At the heart of cardiac dyssynchrony. Trends in Cardiovascular Medicine **30**(2), 115–122 (2020)
4. Carerj, S., Micari, A., Trono, A., Giordano, G., Cerrito, M., Zito, C., Luzza, F., Coglitore, S., Arrigo, F., Oreto, G.: Anatomical m-mode: An old–new technique. Echocardiography **20**(4), 357–361 (2003)
5. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional

networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 839–847. IEEE (2018)
6. Doltra, A., Bijnens, B., Tolosana, J.M., Borràs, R., Khatib, M., Penela, D., De Caralt, T.M., Castel, M.Á., Berruezo, A., Brugada, J., et al.: Mechanical abnormalities detected with conventional echocardiography are associated with response and midterm survival in CRT. JACC: Cardiovascular Imaging **7**(10), 969–979 (2014)
7. Duchateau, N., De Craene, M., Piella, G., Frangi, A.F.: Constrained manifold learning for the characterization of pathological deviations from normality. Medical image analysis **16**(8), 1532–1549 (2012)
8. Duchateau, N., De Craene, M., Piella, G., Silva, E., Doltra, A., Sitges, M., Bijnens, B.H., Frangi, A.F.: A spatiotemporal statistical atlas of motion for the quantification of abnormal myocardial tissue velocities. Medical image analysis **15**(3), 316–328 (2011)
9. Kubota, Y.: tf-keras-vis (11 2022), `https://keisen.github.io/tf-keras-vis-docs/`
10. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., et al.: Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), e210315 (2022)
11. Peressutti, D., Bai, W., Jackson, T., Sohal, M., Rinaldi, A., Rueckert, D., King, A.: Prospective identification of CRT super responders using a motion atlas and random projection ensemble learning. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 493–500. Springer (2015)
12. Puyol-Antón, E., Chen, C., Clough, J.R., Ruijsink, B., Sidhu, B.S., Gould, J., Porter, B., Elliott, M., Mehta, V., Rueckert, D., et al.: Interpretable deep models for cardiac resynchronisation therapy response prediction. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 284–293. Springer (2020)
13. Qu, M., Wang, Y., Li, H., Yang, J., Ma, C.: Automatic identification of septal flash phenomenon in patients with complete left bundle branch block. Medical Image Analysis **82**, 102619 (2022)
14. Schöttl, A.: Improving the interpretability of gradcams in deep classification networks. Procedia Computer Science **200**, 620–628 (2022)
15. Sieniewicz, B.J., Gould, J., Porter, B., Sidhu, B.S., Teall, T., Webb, J., Carr-White, G., Rinaldi, C.A.: Understanding non-response to cardiac resynchronisation therapy: common problems and potential solutions. Heart failure reviews **24**, 41–54 (2019)
16. Sinclair, M., Peressutti, D., Puyol-Antón, E., Bai, W., Rivolo, S., Webb, J., Claridge, S., Jackson, T., Nordsletten, D., Hadjicharalambous, M., et al.: Myocardial strain computed at multiple spatial scales from tagged magnetic resonance imaging: Estimating cardiac biomarkers for crt patients. Medical image analysis **43**, 169–185 (2018)
17. Smiseth, O.A., Russell, K., Skulstad, H.: The role of echocardiography in quantification of left ventricular dyssynchrony: state of the art and future directions. European Heart Journal–Cardiovascular Imaging **13**(1), 61–68 (2012)
18. Zhang, Y., Hong, D., McClement, D., Oladosu, O., Pridham, G., Slaney, G.: Grad-cam helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. Journal of Neuroscience Methods **353**, 109098 (2021)