frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 11 September 2023
DOI 10.3389/fcvm.2023.1213290

OPEN ACCESS

EDITED BY
Grigorios Korosoglou,
GRN Klinik Weinheim, Germany

REVIEWED BY
Thomas Hadler,
Charite University Medicine Berlin, Germany
Elisa Scalco,
National Research Council (CNR), Italy

*CORRESPONDENCE
Ricardo A. Gonzales
✉ ricardo.gonzales@cardiov.ox.ac.uk

†A list of HCMR and OxAMI Investigators is provided in Supplementary Material

RECEIVED 27 April 2023
ACCEPTED 16 August 2023
PUBLISHED 11 September 2023

CITATION
Gonzales RA, Ibáñez DH, Hann E, Popescu IA, Burrage MK, Lee YP, Altun İ, Weintraub WS, Kwong RY, Kramer CM, Neubauer S, Ferreira VM, Zhang Q and Piechnik SK (2023) Quality control-driven deep ensemble for accountable automated segmentation of cardiac magnetic resonance LGE and VNE images.
Front. Cardiovasc. Med. 10:1213290.
doi: 10.3389/fcvm.2023.1213290

Quality control-driven deep ensemble for accountable automated segmentation of cardiac magnetic resonance LGE and VNE images

Ricardo A. Gonzales[1]*, Daniel H. Ibáñez[1,2], Evan Hann[1], Iulia A. Popescu[1], Matthew K. Burrage[1,3], Yung P. Lee[1], İbrahim Altun[1], William S. Weintraub[4], Raymond Y. Kwong[5], Christopher M. Kramer[6], Stefan Neubauer[1], Hypertrophic Cardiomyopathy Registry (HCMR) Investigators, Oxford Acute Myocardial Infarction (OxAMI) Study†, Vanessa M. Ferreira[1], Qiang Zhang[1] and Stefan K. Piechnik[1]

[1]Oxford Centre for Clinical Magnetic Resonance Research (OCMR), Division of Cardiovascular Medicine, Radcliffe Department of Medicine, John Radcliffe Hospital, University of Oxford, Oxford, United Kingdom, [2]Artificio, Cambridge, MA, United States, [3]Faculty of Medicine, University of Queensland, Brisbane, QLD, Australia, [4]MedStar Health Research Institute, Georgetown University, Washington, DC, United States, [5]Cardiovascular Division, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, [6]Department of Medicine, University of Virginia Health System, Charlottesville, VA, United States

Background: Late gadolinium enhancement (LGE) cardiovascular magnetic resonance (CMR) imaging is the gold standard for non-invasive myocardial tissue characterisation. However, accurate segmentation of the left ventricular (LV) myocardium remains a challenge due to limited training data and lack of quality control. This study addresses these issues by leveraging generative adversarial networks (GAN)-generated virtual native enhancement (VNE) images to expand the training set and incorporating an automated quality control-driven (QCD) framework to improve segmentation reliability.
Methods: A dataset comprising 4,716 LGE images (from 1,363 patients with hypertrophic cardiomyopathy and myocardial infarction) was used for development. To generate additional clinically validated data, LGE data were augmented with a GAN-based generator to produce VNE images. LV was contoured on these images manually by clinical observers. To create diverse candidate segmentations, the QCD framework involved multiple U-Nets, which were combined using statistical rank filters. The framework predicted the Dice Similarity Coefficient (DSC) for each candidate segmentation, with the highest predicted DSC indicating the most accurate and reliable result. The performance of the QCD ensemble framework was evaluated on both LGE and VNE test datasets (309 LGE/VNE images from 103 patients), assessing segmentation accuracy (DSC) and quality prediction (mean absolute error (MAE) and binary classification accuracy).
Results: The QCD framework effectively and rapidly segmented the LV myocardium (<1 s per image) on both LGE and VNE images, demonstrating robust performance on both test datasets with similar mean DSC (LGE: 0.845 ± 0.075; VNE: 0.845 ± 0.071; $p = ns$). Incorporating GAN-generated VNE data into the training process consistently led to enhanced performance for both individual models and the overall framework. The quality control mechanism yielded a high performance $(\text{MAE} = 0.043,\ \text{accuracy} = 0.951)$ emphasising the accuracy of the quality

control-driven strategy in predicting segmentation quality in clinical settings. Overall, no statistical difference ($p = ns$) was found when comparing the LGE and VNE test sets across all experiments.

Conclusions: The QCD ensemble framework, leveraging GAN-generated VNE data and an automated quality control mechanism, significantly improved the accuracy and reliability of LGE segmentation, paving the way for enhanced and accountable diagnostic imaging in routine clinical use.

KEYWORDS

data augmentation, generative adversarial networks, quality control, segmentation, late gadolinium enhancement, virtual native enhancement, cardiovascular magnetic resonance

1. Introduction

Late gadolinium enhancement (LGE) is a cardiovascular magnetic resonance (CMR) imaging technique that provides crucial information about the extent and location of myocardial damage, allowing clinicians to make accurate diagnoses and treatment decisions (1). It is considered the imaging gold standard for non-invasive myocardial tissue characterisation in a variety of cardiovascular diseases. LGE can identify areas of scar tissue or fibrosis (2), which are often associated with heart disease, such as myocardial infarction (MI) (3) and hypertrophic cardiomyopathy (4). Its quantification provides important information, such as scar-burden, which can predict adverse clinical outcomes like heart failure and sudden death, and may guide risk-modification strategies, such as the use of implantable cardioverter-defibrillator devices (5).

To quantify the extent and location of myocardial pathology in LGE images, the left ventricular (LV) myocardium must be segmented. Manual contouring by experts has been the conventional method, but it is time-consuming and subjective. Recently, there has been growing development in automated segmentation methods to improve efficiency and reduce inter-observer variability. These methods can be broadly categorised as either model-driven (6) or data-driven (7). Model-driven methods use prior knowledge about the structure of the LV myocardium to guide the segmentation process, while data-driven methods use machine learning algorithms to learn from examples in a training dataset, typically yielding superior results than model-driven methods (8). However, despite progress in automated segmentation techniques, clinical translation has been limited by two major challenges. First, data-driven methods require a large amount of high-quality training data (9), which may not always be available, particularly for rare or heterogeneous diseases. Second, even with sufficient training data, unflagged segmentation errors can still occur (10), leading to inaccurate scar quantification, posing a significant concern for clinical decision-making. Thus, there remains a pressing need for a well-validated, automated quality control (QC) mechanism that can detect and flag segmentation errors in a reliable and efficient manner (11).

To overcome the challenge of data scarcity, or limited access, in medical applications, various approaches have been proposed, such as transfer learning, domain adaptation, and data augmentation (12). Transfer learning and domain adaptation aim to leverage knowledge from pre-existing datasets, while data augmentation methods generate new data by applying transformations to existing data. Among these approaches, data augmentation with synthetic data, using Generative Adversarial Networks (GANs), has gained popularity due to its potential to generate large amounts of diverse and realistic data, which can be particularly useful for limited datasets (13, 14). However, the use of synthetic data for medical applications poses a challenge of clinical validation, as the generated data may not accurately reflect the true biological and pathological variations seen in real-world data (15). Therefore, it is crucial to validate the synthetic data before using it for medical purposes.

Automated approaches for flagging inaccuracies in automatic segmentation have gained increasing attention in recent years (11). Post-analysis QC tools have been recently proposed to assess the reliability of segmentation outputs, which are considered the final indicator of a model's performance. These methods typically act as binary classifiers (16, 17), assigning correct/incorrect labels to a segmentation, or as regressors (18, 19), which attempt to infer well-known validation metrics or uncertainty estimates. While these approaches have been successfully applied to CMR T1 mapping (20) and short-axis cines (21), a QC pipeline for LGE segmentation—an important clinical tool—is still missing.

In this study, we present a novel approach for LGE segmentation that overcomes the challenges of both limited training data and lack of quality control for clinical applications. Our framework leverages the power of GAN-generated data, incorporating virtual native enhancement (VNE) images (22, 23), to further expand the training dataset with clinically-validated data. This emerging contrast-agent-free CMR modality exploits native signals to produce "virtual" LGE images. Additionally, we extend an automated quality control mechanism to flag problematic cases for focused inspection before clinical use. We build upon the quality control-driven (QCD) framework (19, 20), which can predict a confidence metric in absence of ground truth.

2. Materials and methods

2.1. Imaging data

The development dataset of 4,716 LGE images (1,363 patients) was obtained from the following: (1) the multi-centre Hypertrophic Cardiomyopathy Registry study (24) (HCMR, $n = 3{,}286$ images from 1,129 patients, 24 centres); (2) the University of Oxford Centre for Clinical Magnetic Resonance Research clinical service (OCMR, $n = 712$ images from 109 patients), and (3) the Oxford Acute Myocardial Infarction study (25) (OxAMI, $n = 718$ images from 125 patients), with institutional review committee and ethics approvals. Altogether, the 4,716 LGE images comprised of 3,286 LGE images from 1,129 patients with hypertrophic cardiomyopathy, and 1,430 LGE images from 234 patients with MI (255 images from 65 patients with chronic MI; 1,175 LGE images from 169 patients with acute MI). CMR scanning was undertaken in Siemens MR scanners (Siemens Healthcare, Germany) with magnetic field strengths of 1.5T (71% of data) and 3T (29% of data). CMR protocols included cine steady-state free precession imaging, native and post-contrast T1 mapping using the ShMOLLI (Shortened Modified Look-Locker Inversion recovery) sequence (26, 27), and LGE imaging acquired at around 10 min after intravenous administration of 0.1 to 0.2 mmol/kg of a gadolinium-based contrast agent, typically with the phase-sensitive inversion recovery sequence (24). Briefly, the manual quality control involved selection of uncorrupted, paired cines, T1 maps and LGE images, which were manually segmented by experienced trained observers (MKB, YPL and IA), in previous studies (22, 23, 28).

2.2. Data augmentation using a generative adversarial network

The data were augmented with a conditional generative adversarial network (cGAN) approach to generate VNE images (22, 23) from paired short-axis cine and T1 map. These VNE images exploited native components, including native T1 mapping and pre-contrast cine frames throughout the cardiac cycle. This provided image contrast, alterations in myocardial tissue properties, myocardial structure (such as wall thickness/thickness), motion data of the cardiac wall, and more distinct myocardial borders. The deep learning generator processed these inputs to produce VNE images that closely resembled LGE images in terms of structure and contrast. The clinical utility of VNE lies in its ability to generate “virtual” LGE images without the need for gadolinium, enabling faster, lower-cost, and contrast-free CMR scans.

The VNE generator (**Figure 1**) consisted of parallel convolutional neural network streams that processed cine frames and motion-corrected T1 maps (29) individually. Each stream utilised a six-level encoder-decoder U-Net structure (30). The encoder computed image features at various scales, with successive convolutional layers for feature extraction and downsampling at each level, offering a multiscale feature representation. The corresponding decoder fused these multiscale features to generate the final feature maps, with symmetrical upsampling layers and convolutions for sequential combination of the multiscale features. These feature maps from the streams were concatenated and fed into an additional two-level encoder-decoder block, which combined information from the different modalities to create the final VNE image in a late fusion manner. Each encoder-decoder block was followed by a tanh activation function.

In the customised cGAN approach (31), the architecture included two discriminators, D1 and D2, modelled after the VGG16 model (32). Discriminator D1, aimed at verifying the “clarity” of larger images, used an expanded architecture with an input layer accommodating the resultant VNE and the input cine stack, which ensured sharper clearer images. This involved a series of convolutional layers, alternating between feature

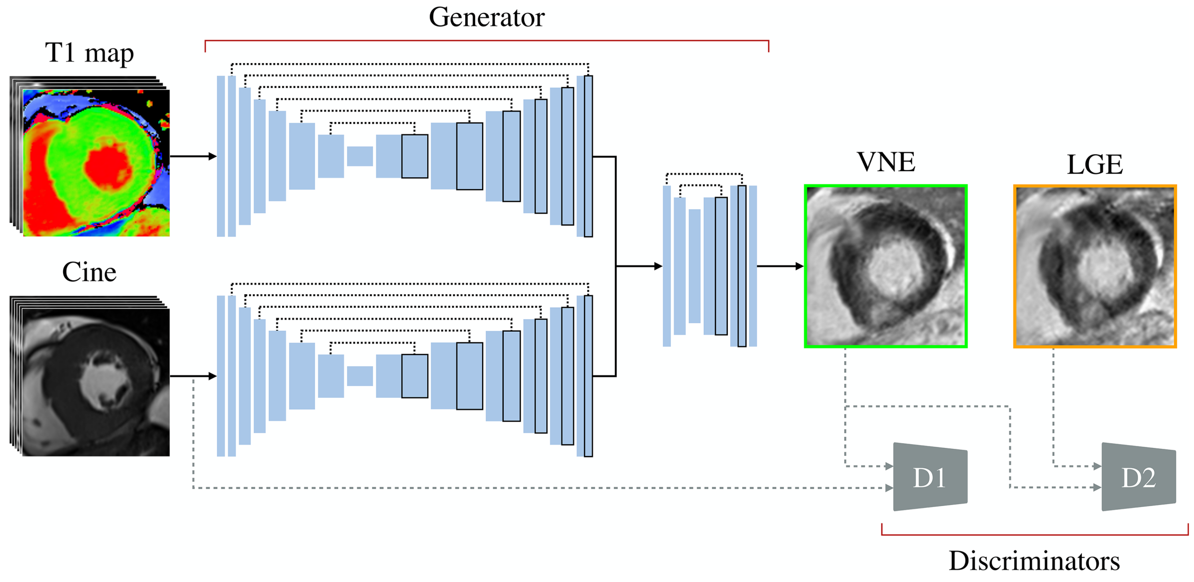

FIGURE 1
Data augmentation framework. A late gadolinium enhancement (LGE) image is augmented by using its paired short-axis cine and T1 map, producing a virtual native enhancement (VNE) image, using a modified conditional generative adversarial network approach. Parallel deep auto-encoders extract features from native signals, which are fused through a shallow autoencoder to derive a VNE image. The discriminators, D1 and D2, during training, are used to enhance the D1 image “clarity” and the image “realness” with perceptual similarity, respectively.

extraction and downsampling, each followed by leaky rectified linear unit activation functions. Discriminator D2, designed to check the "realness" in single-channel images, adopted a similar and more compact structure, processing both the resultant VNE and the paired LGE. The generator's objective was to create VNE images that had a high perceptual similarity (33) to LGE images and were indiscernible from LGE contrast images. The discriminator's objective was to differentiate between VNE and LGE images. After training the neural networks in an adversarial manner, we obtained a trained generator capable of translating native CMR signals into LGE-like representations.

With the previously trained VNE generator (22, 23), we expanded the LGE images in the development data by producing corresponding VNE images (**Figure 2**), in independent datasets. Expansion of the imaging data was successfully carried out for all cases, except for the subset related to acute myocardial infarction, which is awaiting further validation before inclusion. All augmented data were also manually segmented. Through the utilisation of position-matched T1 maps and cine, the derived VNE closely resembled the position-matched LGE; however, in some cases, there were slight differences in slice position between the paired T1/cine and the final LGE, for instance, due to patient movement between the image acquisitions (**Figure 2**, cases 5 and 6). Serendipitously, this introduced increased diversity and realism into the training data, thereby enhancing the robustness of the model.

2.3. Quality control-driven ensemble framework

A quality control-driven ensemble framework (19) (**Figure 3**) was developed to enhance the accuracy and reliability of the segmentation process by leveraging the strengths of multiple convolutional neural networks. This framework utilised various U-Nets (30) with different depths to create a diverse set of candidate segmentations. These segmentations were then combined using statistical rank filters in a pixel-wise fashion (34), further expanding the pool of segmentation candidates and improving robustness.

In the ensemble framework, six U-Nets (30), with depths ranging from 1 to 6 levels, were employed. Each U-Net consisted of an encoder and a decoder. The encoder had convolutional layers followed by dropout layers (35) for regularisation, with increasing dropout rate with each layer to mitigate overfitting. Post-convolution and dropout, a max pooling operation was applied. The decoder mirrored the encoder but used transposed convolutional layers for upscaling. It also utilised skip connections, coupling outputs from the decoder with corresponding encoder layers. The final layer underwent additional convolutions and a softmax activation to produce the final segmentation. The overall process, which was repeated for each depth, allowed the generation of diverse candidate segmentations, contributing to the ensemble's performance.

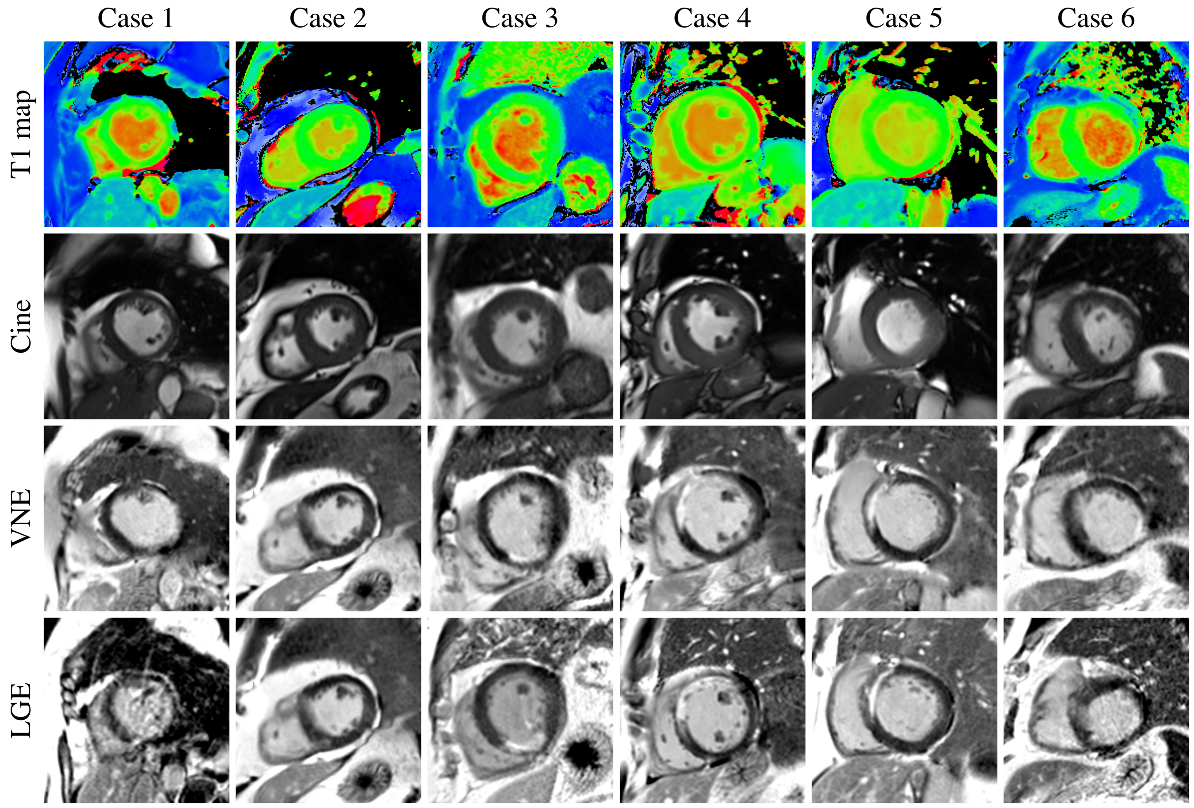

FIGURE 2
The resultant database includes the late gadolinium enhancement (LGE) data and virtual native enhancement (VNE) data, as ways of data augmentation and additional validation.

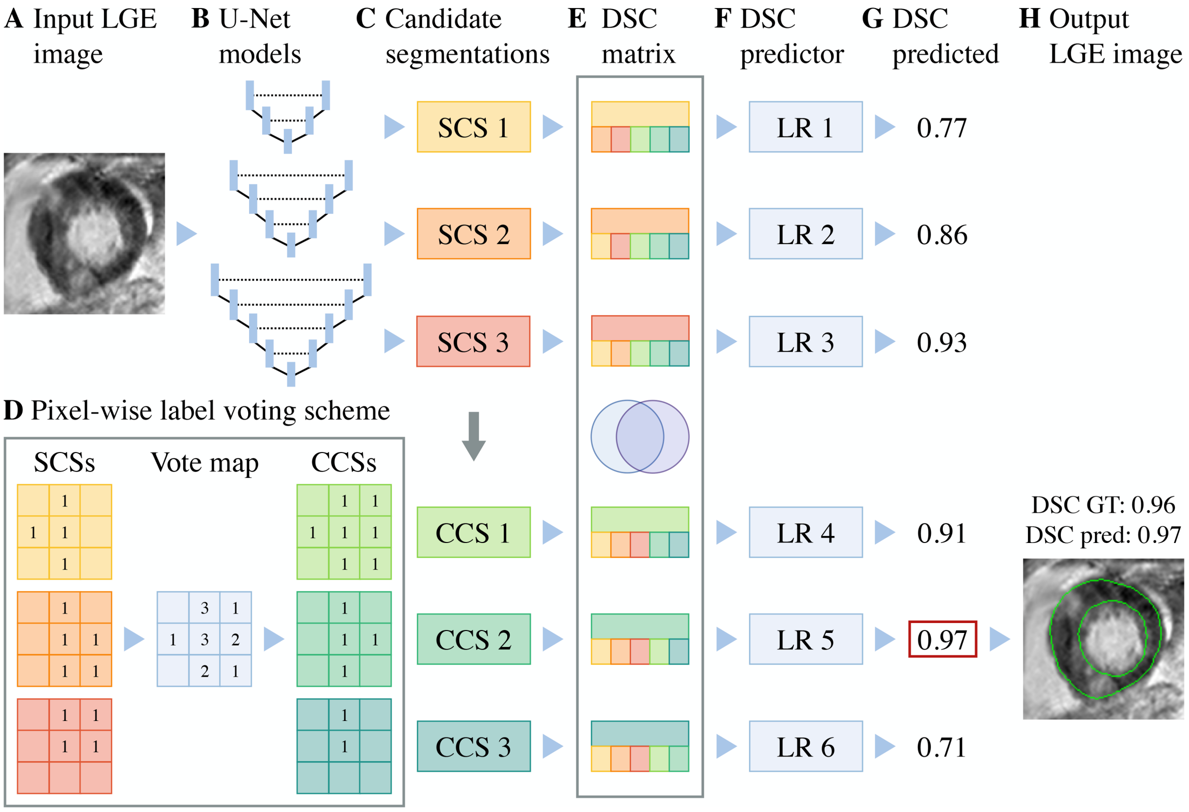

FIGURE 3
Illustrative quality control-driven ensemble framework depicted with 3 (out of 6) U-Nets. (A) A late gadolinium enhancement (LGE) image is processed by (B) an ensemble of independent U-Net segmentation models to produce (C) single candidate segmentations (SCSs). (D) The SCSs are then combined via a pixel-wise label voting scheme to derive combined candidate segmentations (CCSs). (E) An association matrix of Dice Similarity Coefficients (DSC) is generated upon the agreement between SCSs and CCSs. The inter-candidate DSCs are supplied to the (F) linear regressors (LR), and (G) each model outputs the predicted the DSC, in absence of ground truth (GT); finally, (H) the model with the highest predicted DSC and its corresponding automated segmentation output are selected on-the-fly.

The automatic quality scoring mechanism at the core of the framework predicted the Dice Similarity Coefficient (DSC) for each candidate segmentation by exploiting their differences. It calculated the pairwise agreement, or inter-segmentation DSC matrix, between segmentations, capturing the overlap and divergence between different candidates. These DSC matrices were then fed into separate linear regression models for each candidate, with the target being the DSC between the candidate and the ground truth.

For each input image, the framework assigned a predicted DSC, with respect to the ground truth segmentation, to every candidate segmentation, both single and combined. The final segmentation was selected by identifying the candidate with the highest predicted DSC, indicating the most accurate and reliable result. This selection process was performed automatically by the framework, without any manual intervention. This approach effectively emulated a multidisciplinary clinical team, where the consistency among multiple expert opinions served as a marker for the best approach in managing complex cases. By incorporating this quality control-driven strategy, the ensemble framework aimed to improve overall segmentation performance and provide confidence metrics, particularly useful in clinical settings.

2.4. Implementation

The data were augmented with the VNE technology using available co-located short-axis cines and ShMOLLI T1 maps, resulting in 3,541 VNE images. The development dataset was randomly partitioned into: (1) 85% for the training dataset (4,092 LGE images and 2,917 VNE images from 1,158 patients); (2) 7.5% for validation (309 LGE/VNE images from 102 patients); (3) 7.5% for the test dataset (309 LGE/VNE images from 103 patients), per recommended guidelines (36). Image pixel values were scaled from 0 to 1 and zero-padded to 256 × 256. For the segmentation models, the Adam method (37) was used for optimising the categorical cross-entropy loss, with a learning rate of 5×10^{-5} for 200 epochs; an automated early stop was used to avoid overfitting, using the validation set. For the quality prediction models, a linear regressor for each candidate segmentation was fit with the inter-agreement between its corresponding candidate segmentation and the rest. These regressors were trained on the validation set to avoid autocorrelation with the training set. The models were trained and tested on TensorFlow (38) with an NVIDIA GeForce RTX 3090 GPU, taking approximately 11.5 h.

2.5. Evaluation

The performance of the QCD ensemble framework was evaluated for myocardial contours on both LGE and VNE test datasets, and across the main pathologies. The segmentation accuracy was assessed by DSC, comparing the agreement between the optimally selected mask and the ground-truth mask. The predicted segmentation accuracy was assessed in terms of the mean absolute error (MAE) and binary classification accuracy, with a DSC threshold at 0.7 (39). The former was used to measure the difference between the predicted DSC and the observed ground-truth DSC derived from the manual segmentation. The latter assessed whether the segmentations were classified into good (≥0.7) or poor quality (<0.7) to demonstrate the practical usage of the DSC prediction. In the evaluation, false positives occurred when the predicted DSC was ≥0.7 but the real accuracy was <0.7, while false negatives arose when the predicted DSC was <0.7 despite the real accuracy being ≥0.7. The threshold of 0.7 ensured a balance between sensitivity and specificity in the classification of good and bad quality segmentations. As evidenced in our prior works on aortic (19) and myocardial segmentation (20, 40), this threshold provided a standard measure of segmentation quality across different studies. In this context, a DSC of 0.7 implied that 70% of the segmentation correctly overlapped with the ground truth, which was considered to be an acceptable level of accuracy for our applications. A Wilcoxon signed-rank test was conducted using Python to determine if there was a statistically significant difference between the segmentation results obtained on LGE data and VNE data, paired when possible, and within pathology groups. $P < 0.05$ was considered significant. This analysis helped to ascertain the robustness of the model when segmenting both types of images and the potential benefits of incorporating VNE data into the training process.

2.5.1. Comparative analysis

A comparative study was conducted to investigate the pipeline's main components performance and the impact of incorporating VNE data. Firstly, the performance of the deepest, top-performing employed U-Net (depth of 6 levels) and the QCD segmentation framework were assessed to highlight the benefit of a higher segmentation accuracy with a quality

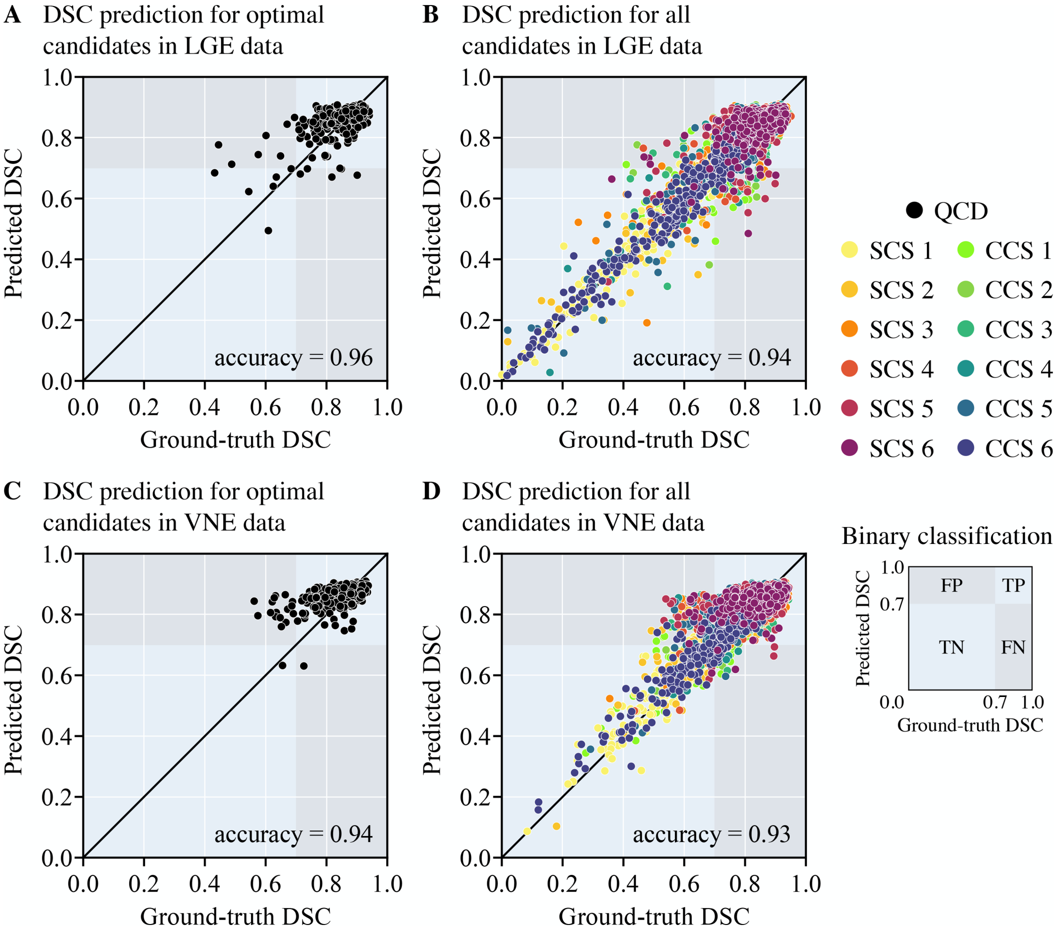

FIGURE 4
Scatter plots of the observed ground-truth Dice Similarity Coefficient (DSC) (x-axis) versus the predicted DSC (y-axis) for myocardial contours in (A,B) late gadolinium enhancement (LGE) and (C,D) virtual native enhancement (VNE) images for the optimal candidates (A,C) and for all single (SCS) and combined segmentation (CCS) models (B,D). The shown overall binary classification accuracy is measured as the proportion of true results (true positive (TP) or true negative (TN)—light blue background), in a population of both true and negative results (false positive (FP) or false negative (FN)—grey background), with a binary threshold of DSC ≥0.7.

predictive capacity. Secondly, transversely, each experiment involved training with LGE data, VNE data, and both combined, to thoroughly evaluate the data augmentation capability of the GAN-generated VNE data. Thirdly, each experiment was also tested on LGE data, VNE data, and both combined, to exhibit the robustness of the proposed method. The segmentation accuracy and the quality prediction accuracy were assessed in all experiments, to compare the differences. Lastly, the extensive data augmentation techniques, proposed in the nnU-Net framework (41) were also implemented, to assess the added benefit of VNE data in the pipeline.

3. Results

3.1. Segmentation and prediction accuracy

The scatter plots (**Figure 4**) reflect the parity between the ground-truth DSC and the predicted DSC for the resultant framework output and for every candidate model output in the test set, being able to accurately predict from underperforming models to highperforming models. The QCD framework successfully and rapidly segmented the LV myocardium on LGE and VNE images. The QCD framework demonstrated robust segmentation performance on both LGE and VNE test datasets, with similar mean DSC (LGE: 0.845 ± 0.075; VNE: 0.845 ± 0.071; $p = ns$). The QCD framework also exhibited robust segmentation performance across the main pathologies (hypertrophic cardiomyopathy: 0.845 ± 0.069; MI: 0.844 ± 0.085; $p = ns$). The mean absolute error (MAE) for the predicted DSC was low at 0.043 ± 0.043, demonstrating the accuracy of the quality control-driven strategy in predicting the segmentation quality. Moreover, using the DSC threshold of 0.7, the binary classification accuracy was high at 0.951, further emphasising the practical usefulness of the proposed ensemble framework in clinical settings. **Figures 5** and **6** show representative test cases of the QCD framework on LGE and VNE images, respectively, for true positive, true negative, false positive and false negative cases.

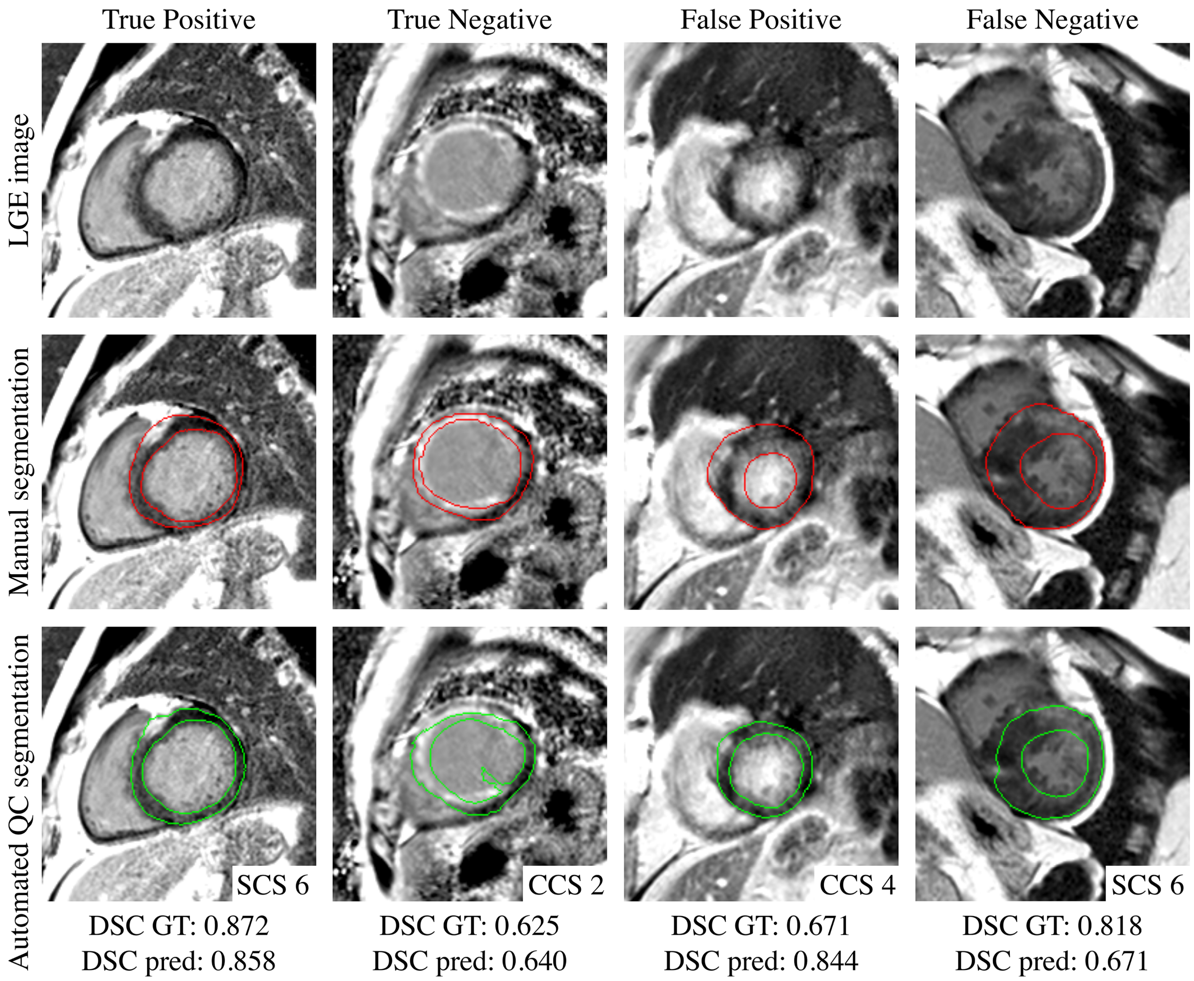

FIGURE 5
Examples of true positive (93.9%), true negative (1.9%), false positive (2.3%) and false negative (1.9%) for predicted quality-controlled (QC) segmentations in late gadolinium enhanced (LGE) images. The left ventricular myocardium is manually segmented in red and automatically segmented in green, from different single (SCS) and combined candidate segmentation (CCS) models. The corresponding observed ground-truth (GT) Dice Similarity Coefficient (DSC) and predicted DSC are provided at the bottom.

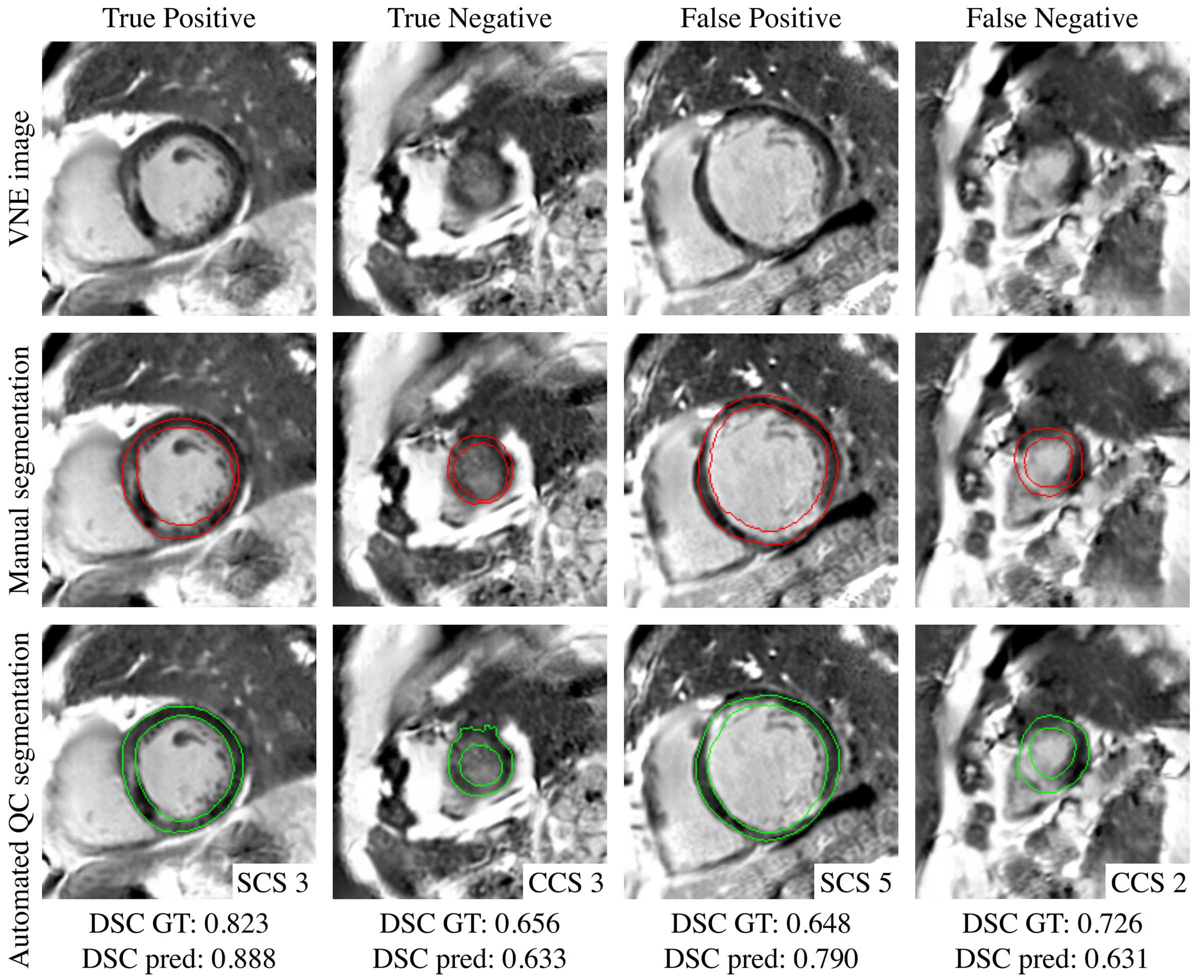

FIGURE 6
Examples of true positive (94.2%), true negative (0.3%), false positive (5.2%) and false negative (0.3%) for predicted quality-controlled (QC) segmentations in virtual native enhancement (VNE) images. The left ventricular myocardium is manually segmented in red and automatically segmented in green, from different single (SCS) and combined candidate segmentation (CCS) models. The corresponding observed ground-truth (GT) Dice Similarity Coefficient (DSC) and predicted DSC are provided at the bottom.

3.2. Comparative analysis

The comparative analysis (Table 1) highlights the contribution of the GAN-generated VNE as data augmentation and the QCD segmentation framework as an automated quality control mechanism. Firstly, the individual performance of the deepest U-Net was on par with the ensemble performance of the QCD segmentation framework, which was also able to estimate the quality of the resultant segmentation. When analysing the results using all datasets for training and testing, the deepest U-Net achieved a segmentation accuracy of 0.845 ± 0.070, similar to the segmentation accuracy of the QCD segmentation framework. Secondly, the performance of the deepest U-Net, trained on only LGE data, completely generalised to the VNE test set, with a DSC of 0.836 ± 0.082 and 0.838 ± 0.075 for the LGE and VNE test sets, respectively. Similarly, the deepest U-Net, trained on only VNE data, with 30% less data, also generalised to the LGE test set, with a DSC of 0.791 ± 0.119 and 0.824 ± 0.084 for the LGE and VNE test sets, respectively. These findings, similar to the ones with the ensemble framework, effectively support the resemblance of VNE to LGE images, and validates the data augmentation approach for virtually yielding the same performance as its counterpart.

Thirdly, including the GAN-generated data consistently improved the performance in both experiments of the individual model and the QCD segmentation framework, in every experiment. For instance, the QCD segmentation framework trained on only LGE data yielded a mean DSC of 0.835 ± 0.082 and 0.838 ± 0.080 for the LGE and VNE test sets, respectively; whereas the QCD segmentation framework trained on both LGE and VNE data yielded a higher segmentation performance of 0.845 ± 0.075 and 0.845 ± 0.071, indicating the benefits of including VNE data in the training process for the segmentation accuracy. The quality predictive capacity also improved when trained and tested in both datasets. Lastly, the individual segmentation performance of the model trained with only LGE data including extensive data augmentation (41) (0.846 ± 0.072) was also surpassed by the same

TABLE 1 Comparative analysis of individual models (U-Net with depth of 6 levels) and quality control-driven (QCD) segmentation framework.

Models and traning data		Test data								
		LGE ($n = 309$)			VNE ($n = 309$)			LGE + VNE ($n = 618$)		
Models	Training set	DSC	MAE	Acc.	DSC	MAE	Acc.	DSC	MAE	Acc.
U-Net	LGE	0.836	—	—	0.838	—	—	0.837	—	—
U-Net	VNE	0.791	—	—	0.824	—	—	0.807	—	—
U-Net	LGE + VNE	0.844	—	—	**0.846**	—	—	0.845	—	—
QCD	LGE	0.835	**0.042**	**0.971**	0.838	0.046	0.922	0.837	0.044	0.947
QCD	VNE	0.799	0.057	0.922	0.833	**0.041**	**0.958**	0.816	0.049	0.940
QCD	**LGE + VNE**	**0.845**	0.042	0.958	0.845	0.043	0.945	**0.845**	**0.043**	**0.951**

The models were trained on late gadolinium enhancement (LGE; $n = 4{,}092$) and/or virtual native enhancement (VNE; $n = 2{,}917$) data and tested on LGE ($n = 309$) and/or VNE ($n = 309$) data, evaluated by their segmentation performance with the mean Dice Similarity Coefficient (DSC) and the quality predictive capacity with the mean absolute error (MAE) and the binary classification accuracy (Acc.). The best results are highlighted in bold.

framework when the VNE data were added (0.851 ± 0.068). Overall, no significant differences were observed when comparing LGE and VNE test sets between the experiments.

4. Discussion

In this study, we demonstrated that the proposed framework, leveraging GAN-generated VNE data and incorporating an automated quality control mechanism, significantly improved the accuracy and reliability of LGE segmentation. The comparative analysis demonstrated the benefits of using VNE data, with generally better image quality and more consistency (22, 23), in the training process and the effectiveness of the ensemble framework in enhancing segmentation performance. Moreover, the proposed framework revealed robustness in both LGE and VNE data. This represents an accountable pipeline for automated segmentation in the gold-standard and emerging contrast-agent-free modalities, paving the way for faster and reliable diagnosis of myocardial damage.

Data scarcity or lack of access remains a persistent challenge in developing robust and reliable deep learning models for medical image segmentation, particularly in the context of LGE segmentation. This limitation stems from the high costs and ethical considerations associated with acquiring, labelling, and sharing large-scale annotated datasets, which often result in insufficient representation of diverse and rare pathological cases (9). As a consequence, models may underperform or fail to generalise well to unseen cases (10), hampering their clinical utility. The application of generating VNE images (22, 23) assured a data augmentation technique with more reliable image contrast. This significantly improves over prior methods of using synthetic LGE images (42, 43), which had not been clinically-validated and were not designed for displaying LGE lesion signals. Addressing data scarcity through the generation of VNE images, as demonstrated in this study, can alleviate this issue by augmenting the already-available LGE data, or potentially substituting the data, leading to a more robust and reliable model that can better handle the complexity of clinical cases and ultimately improve patient care.

The presented work focused on the integration of an automated quality assurance framework, which was developed with a traditional encoder-decoder U-Net architecture (30) with different depths. The quality control-driven strategy provided reliable quality predictions, crucial for clinical decision-making. The adapted regression-based quality prediction scheme enables further exploitation of the diversity of different candidates, with a deterministic approach of assessing the agreement between candidates, proven to be more effective than the emerging Monte Carlo-based quality assurance scheme (40). Newer network architectures and advanced pre-processing schemes could be incorporated to increase such diversity, ranging from existing LGE segmentation approaches (44) to spatial transformation-based pre-processing (45, 46). Future work will cover dataset extension, different candidate models, and scar burden evaluation.

The clinical implications of the proposed QCD ensemble framework are substantial, as it introduces an automated quality control mechanism for the first time in automated LGE segmentation, improving both accuracy and reliability. Moreover, the quality control-driven strategy allows for the identification and refinement of suboptimal segmentations, ensuring the system's efficiency and trustworthiness, paving the way for increased adoption in clinical workflows. This enhancement streamlines the diagnostic process, reduces contouring variability, and bolsters clinician confidence in automated segmentation results, potentially leading to easier scar burden quantification on a routine basis, better-informed treatment decisions and improved patient outcomes.

This work has some limitations. The main goal has been the fundamental task of the myocardial delineation, thus avoiding engaging the generalisation into the known difficulties of scar tissue quantification (47). In particular the diffuse, less structured and scattered lesions require lower segmentation thresholds, and the targets are subject to many methodological choices and biases in ground truth data (48, 49). While our model was evaluated on an extensive international database, enabling potential generalisation to various conditions requiring LGEs, the findings have been primarily concentrated on patients with hypertrophic cardiomyopathy and MI, suggesting the need for future validation across a wider spectrum of pathologies. Lastly, our selection of the DSC threshold of 0.7 for discerning between acceptable and unacceptable segmentations, despite being effective in our prior work (19, 20, 40), might not address all the geometric properties of cardiac structures (10). We expect that proposed methods would generalise well to any single particular threshold or application, yet further research is needed to explore the available choices for the clinical routine use beyond the scope of this work.

5. Conclusion

In conclusion, our study presents a novel approach for automated LGE segmentation that overcomes the challenges of limited training data and lack of quality control, for clinical applications. By leveraging the power of GAN-generated VNE images and incorporating an automated quality control mechanism, we demonstrate the potential for improved automated segmentation performance and reliability. This framework could be seamlessly integrated into clinical studies, providing an efficient, quality-controlled and reliable tool for clinicians in diagnosing and managing patients with myocardial damage.

Data availability statement

The data analysed in this study is subject to the following licenses/restrictions: HCMR, OCMR and OxAMI studies can be approached independently for data sharing upon reasonable request. De-identified clinical imaging data can be shared with interested parties, subject to ethical approval, consent, data-protection governance, data sharing agreements and funding available. Requests to access these datasets should be directed to HCMR, https://hcmregistry.org; OCMR, https://www.rdm.ox.ac.uk; OxAMI, https://oxami.org.uk.

Ethics statement

Ethical approval was not required for the studies involving humans because the data used were retrospectively obtained from subjects of previously ethically-approved research studies. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

RG designed the study and drafted the manuscript. DI and QZ curated the database. RG and DI conducted the experiments. EH and IP assisted in the design of the algorithm. MB, YL and İA performed manual segmentation. WW, RK, CK, SN, HCMR and OxAMI study investigators acquired the original MRI datasets. VF, QZ and SP provided research guidance and conceived the study. All authors contributed to the article and approved the submitted version.

Funding

RG acknowledges support for his DPhil studies from the Clarendon Fund, the Balliol College and the Radcliffe Department of Medicine, University of Oxford. DI is funded by Artificio Inc. and supported by the Research Experience for Peruvian Undergraduates (REPU) Program. IP, MB, VF, and SP acknowledge support from the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre at The Oxford University Hospitals NHS Foundation Trust. VF acknowledges support from the BHF (CH/16/1/32013). MB is supported by a British Heart Foundation (BHF) Clinical Research Training Fellowship (FS/19/65/34692). MB, VF, QZ and SP acknowledge support from the Oxford BHF Centre of Research Excellence (RE/18/3/34214). VF, QZ and SP acknowledge the John Fell Oxford University Press Research Fund. The authors thank Hypertrophic Cardiomyopathy Registry (HCMR) Investigators for the imaging and available ground truth data, which was collected under BHF (project grant PPG/15/71/31731), the National Heart, Lung, and Blood Institute (grant U01HL117006-01A1). The authors also thank the Oxford Acute Myocardial Infarction (OxAMI) study, supported by the British Heart Foundation (CH/16/1/32013) and BHF Centre of Research Excellence, Oxford (grant RG/13/1/30181).

Conflict of interest

VF, QZ, and SP have authorship rights for patent WO2021/044153 (“Enhancement of Medical Images”; granted March 11, 2021). EH, IP, VF, QZ and SP have authorship rights for patent WO/2020/161481 (“Method and Apparatus for Quality Prediction”; granted August 13, 2020). SP has patent authorship rights for US patent US20120078084A1 (“Systems and Methods for Shortened Look Locker Inversion Recovery [Sh-MOLLI] Cardiac Gated Mapping of T1”; granted March 15, 2016). DI was employed by Artificio Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2023.1213290/full#supplementary-material.

References

1. Kim RJ, Wu E, Rafael A, Chen EL, Parker MA, Simonetti O, et al. The use of contrast-enhanced magnetic resonance imaging to identify reversible myocardial dysfunction. *N Engl J Med*. (2000) 343:1445–53. doi: 10.1056/NEJM200011163432003

2. Sebastian K, Roes SD, Klein C, Kokocinski T, de Roos A, Eckart F, et al. Prognostic value of myocardial infarct size, contractile reserve using magnetic resonance imaging. *J Am Coll Cardiol*. (2009) 54:1770–7. doi: 10.1016/j.jacc.2009.07.027

3. Roes SD, Kelle S, Kaandorp TA, Kokocinski T, Poldermans D, Lamb HJ, et al. Comparison of myocardial infarct size assessed with contrast-enhanced magnetic resonance imaging, left ventricular function and volumes to predict mortality in patients with healed myocardial infarction. *Am J Cardiol*. (2007) 100:930–6. doi: 10.1016/j.amjcard.2007.04.029

4. Weng Z, Yao J, Chan RH, He J, Yang X, Zhou Y, et al. Prognostic value of LGE-CMR in HCM: a meta-analysis. *JACC: Cardiovasc Imaging*. (2016) 9:1392–402. doi: 10.1016/j.jcmg.2016.02.031

5. Singh A, Chen W, Patel HN, Alvi N, Kawaji K, Besser SA, et al. Impact of wideband late gadolinium enhancement cardiac magnetic resonance imaging on device-related artifacts in different implantable cardioverter-defibrillator types. *J Magn Reson Imaging*. (2021) 54:1257–65. doi: 10.1002/jmri.27608

6. Karim R, Bhagirath P, Claus P, James Housden R, Chen Z, Karimaghaloo Z, et al. Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late gadolinium enhancement MR images. *Med Image Anal*. (2016) 30:95–107. doi: 10.1016/j.media.2016.01.004

7. Zhuang X, Xu J, Luo X, Chen C, Ouyang C, Rueckert D, et al. Cardiac segmentation on late gadolinium enhancement MRI: a benchmark study from multi-sequence cardiac MR segmentation challenge. *Med Image Anal*. (2022) 81:102528. doi: 10.1016/j.media.2022.102528

8. Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: a survey. *IEEE Trans Pattern Anal Mach Intell*. (2022) 44:3523–42. doi: 10.1109/TPAMI.2021.3059968

9. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. *Radiology*. (2020) 295:4–15. doi: 10.1148/radiol.2020192224

10. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans Med Imaging*. (2018) 37:2514–25. doi: 10.1109/TMI.2018.2837502

11. Galati F, Ourselin S, Zuluaga MA. From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: a review. *Appl Sci*. (2022) 12:3936. doi: 10.3390/app12083936

12. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. *Med Image Anal*. (2020) 63:101693. doi: 10.1016/j.media.2020.101693

13. Garcea F, Serra A, Lamberti F, Morra L. Data augmentation for medical imaging: a systematic literature review. *Comput Biol Med*. (2023) 152:106391. doi: 10.1016/j.compbiomed.2022.106391

14. Al Khalil Y, Amirrajab S, Lorenz C, Weese J, Pluim J, Breeuwer M. On the usability of synthetic data for improving the robustness of deep learning-based segmentation of cardiac magnetic resonance images. *Med Image Anal*. (2023) 84:102688. doi: 10.1016/j.media.2022.102688

15. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng*. (2021) 5:493–7. doi: 10.1038/s41551-021-00751-8

16. Albà X, Lekadir K, Pereañez M, Medrano-Gracia P, Young AA, Frangi AF. Automatic initialization and quality control of large-scale cardiac MRI segmentations. *Med Image Anal*. (2018) 43:129–41. doi: 10.1016/j.media.2017.10.001

17. Sander J, de Vos BD, Išgum I. Automatic segmentation with detection of local segmentation failures in cardiac MRI. *Sci Rep*. (2020) 10:21769. doi: 10.1038/s41598-020-77733-4

18. Valindria VV, Lavdas I, Bai W, Kamnitsas K, Aboagye EO, Rockall AG, et al. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE Trans Med Imaging*. (2017) 36:1597–606. doi: 10.1109/TMI.2017.2665165

19. Hann E, Biasiolli L, Zhang Q, Popescu IA, Werys K, Lukaschuk E, et al. Quality control-driven image segmentation towards reliable automatic image analysis in large-scale cardiovascular magnetic resonance aortic cine imaging. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, et al., editors. *Medical Image Computing and Computer Assisted – MICCAI 2019*. Cham: Springer International Publishing (2019). p. 750–8. Available from: https://doi.org/10.1007/978-3-030-32245-8_83.

20. Hann E, Popescu IA, Zhang Q, Gonzales RA, Barutçu A, Neubauer S, et al. Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping. *Med Image Anal*. (2021) 71:102029. doi: 10.1016/j.media.2021.102029

21. Ruijsink B, Puyol-Antón E, Oksuz I, Sinclair M, Bai W, Schnabel JA, et al. Fully automated, quality-controlled cardiac analysis from CMR. *JACC: Cardiovasc Imaging*. (2020) 13:684–95. doi: 10.1016/j.jcmg.2019.05.030

22. Zhang Q, Burrage MK, Lukaschuk E, Shanmuganathan M, Popescu IA, Nikolaidou C, et al. Toward replacing late gadolinium enhancement with artificial intelligence virtual native enhancement for gadolinium-free cardiovascular magnetic resonance tissue characterization in hypertrophic cardiomyopathy. *Circulation*. (2021) 144:589–99. doi: 10.1161/CIRCULATIONAHA.121.054432

23. Zhang Q, Burrage MK, Shanmuganathan M, Gonzales RA, Lukaschuk E, Thomas KE, et al. Artificial intelligence for contrast-free MRI: scar assessment in myocardial infarction using deep learning–based virtual native enhancement. *Circulation*. (2022) 146:1492–503. doi: 10.1161/CIRCULATIONAHA.122.060137

24. Kramer CM, Appelbaum E, Desai MY, Desvigne-Nickens P, DiMarco JP, Friedrich MG, et al. Hypertrophic cardiomyopathy registry: the rationale and design of an international, observational study of hypertrophic cardiomyopathy. *Am Heart J*. (2015) 170:223–30. doi: 10.1016/j.ahj.2015.05.013

25. Scarsini R, Shanmuganathan M, De Maria GL, Borlotti A, Kotronias RA, Burrage MK, et al. Coronary microvascular dysfunction assessed by pressure wire and CMR after STEMI predicts long-term outcomes. *JACC: Cardiovasc Imaging*. (2021) 14:1948–59. doi: 10.1016/j.jcmg.2021.02.023

26. Piechnik SK, Ferreira VM, Dall'Armellina E, Cochlin LE, Greiser A, Neubauer S, et al. Shortened modified look-locker inversion recovery (ShMOLLI) for clinical myocardial T1-mapping at 1.5 and 3 T within a 9 heartbeat breathhold. *J Cardiovasc Magn Reson*. (2010) 12:69. doi: 10.1186/1532-429X-12-69

27. Zhang Q, Werys K, Popescu IA, Biasiolli L, Ntusi NA, Desai M, et al. Quality assurance of quantitative cardiac T1-mapping in multicenter clinical trials—a T1 phantom program from the hypertrophic cardiomyopathy registry (HCMR) study. *Int J Cardiol*. (2021) 330:251–8. doi: 10.1016/j.ijcard.2021.01.026

28. Carapella V, Puchta H, Lukaschuk E, Marini C, Werys K, Neubauer S, et al. Standardized image post-processing of cardiovascular magnetic resonance T1-mapping reduces variability and improves accuracy and consistency in myocardial tissue characterization. *Int J Cardiol*. (2020) 298:128–34. doi: 10.1016/j.ijcard.2019.08.058

29. Gonzales RA, Zhang Q, Papież BW, Werys K, Lukaschuk E, Popescu IA, et al. MOCOnet: robust motion correction of cardiovascular magnetic resonance T1 mapping using convolutional neural networks. *Front Cardiovasc Med*. (2021) 8:1–9. doi: 10.3389/fcvm.2021.768245

30. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical Image Computing, Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing (2015). p. 234–41. Available from: https://doi.org/10.1007/978-3-319-24574-4˙28.

31. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision, Pattern Recognition (CVPR)*. Honolulu, HI, USA: IEEE (2017). p. 5967–76. Available from: https://doi.org/10.1109/CVPR.2017.632.

32. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations (ICLR 2015)*. Computational, Biological Learning Society (2015)

33. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer. In: Leibe B, Matas J, Sebe N, Welling M, editors. *Computer Vision – ECCV 2016*. Cham: Springer International Publishing (2016). p. 694–711. Available from: https://doi.org/10.1007/978-3-319-46475-6˙43.

34. Li X, Aldridge B, Fisher R, Rees J. Estimating the ground truth from multiple individual segmentations incorporating prior pattern analysis with application to skin lesion segmentation. *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. Chicago, IL, USA: IEEE (2011). p. 1438–41. Available from: https://doi.org/10.1109/ISBI.2011.5872670.

35. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. (2014) 15:1929–58. https://dl.acm.org/doi/10.5555/2627435.2670313

36. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*. (2018) 286:800–9. doi: 10.1148/radiol.2017171920

37. Kingma DP, Ba JLA. A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. San Diego, CA, USA: Computational and Biological Learning Society (2015).

38. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: a system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*. Savannah, GA, USA.: USENIX Association (2016). p. 265–83.

39. Robinson R, Valindria VV, Bai W, Oktay O, Kainz B, Suzuki H, et al. Automated quality control in image segmentation: application to the UK Biobank cardiovascular

magnetic resonance imaging study. *J Cardiovasc Magn Reson*. (2019) 21:18. doi: 10.1186/s12968-019-0523-x

40. Hann E, Gonzales RA, Popescu IA, Zhang Q, Ferreira VM, Piechnik SK. Ensemble of deep convolutional neural networks with Monte Carlo dropout sampling for automated image segmentation quality control and robust deep learning using small datasets. In: Papież BW, Yaqub M, Jiao J, Namburete AIL, Noble JA, editors. *Medical Image Understanding and Analysis*. Cham: Springer International Publishing (2021). p. 280–93. Available from: https://doi.org/10.1007/978-3-030-80432-9˙22.

41. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z

42. Chen C, Ouyang C, Tarroni G, Schlemper J, Qiu H, Bai W, et al. Unsupervised multi-modal style transfer for cardiac MR segmentation. *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*. Cham: Springer International Publishing (2020). p. 209–19. Available from: https://doi.org/10.1007/978-3-030-39074-7˙22.

43. Campello VM, Martín-Isla C, Izquierdo C, Petersen SE, Ballester MAG, Lekadir K, Combining multi-sequence and synthetic images for improved segmentation of late gadolinium enhancement cardiac MRI. *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*. Cham: Springer International Publishing (2020). p. 290–9. Available from: https://doi.org/10.1007/978-3-030-39074-7˙31.

44. Wu Y, Tang Z, Li B, Firmin D, Yang G. Recent advances in fibrosis and scar segmentation from cardiac MRI: a state-of-the-art review and future perspectives. *Front Physiol*. (2021) 12:1–23. doi: 10.3389/fphys.2021.709230

45. Gonzales RA, Seemann F, Lamy J, Arvidsson PM, Heiberg E, Murray V, et al. Automated left atrial time-resolved segmentation in MRI long-axis cine images using active contours. *BMC Med Imaging*. (2021) 21:101. doi: 10.1186/s12880-021-00630-3

46. Davies RH, Augusto JB, Bhuva A, Xue H, Treibel TA, Ye Y, et al. Precision measurement of cardiac structure and function in cardiovascular magnetic resonance using machine learning. *J Cardiovasc Magn Reson*. (2022) 24:16. doi: 10.1186/s12968-022-00846-4

47. Heiberg E, Engblom H, Carlsson M, Erlinge D, Atar D, Aletras AH, et al. Infarct quantification with cardiovascular magnetic resonance using “standard deviation from remote” is unreliable: validation in multi-centre multi-vendor data. *J Cardiovasc Magn Reson*. (2022) 24:53. doi: 10.1186/s12968-022-00888-8

48. Mikami Y, Kolman L, Joncas SX, Stirrat J, Scholl D, Rajchl M, et al. Accuracy and reproducibility of semi-automated late gadolinium enhancement quantification techniques in patients with hypertrophic cardiomyopathy. *J Cardiovasc Magn Reson*. (2014) 16:85. doi: 10.1186/s12968-014-0085-x

49. Flett AS, Hasleton J, Cook C, Hausenloy D, Quarta G, Ariti C, et al. Evaluation of techniques for the quantification of myocardial scar of differing etiology using cardiac magnetic resonance. *JACC: Cardiovasc Imaging*. (2011) 4:150–6. doi: 10.1016/j.jcmg.2010.11.015