MiMICRI: Towards Domain-centered Counterfactual Explanations of Cardiovascular Image Classification Models

Grace Guo
Georgia Institute of Technology
Atlanta, Georgia, USA
gguo31@gatech.edu

Lifu Deng
Cleveland Clinic
Cleveland, Ohio, USA
dengl2@ccf.org

Animesh Tandon
Cleveland Clinic
Cleveland, Ohio, USA
tandona2@ccf.org

Alex Endert
Georgia Institute of Technology
Atlanta, Georgia, USA
endert@gatech.edu

Bum Chul Kwon
IBM Research
Cambridge, Massachusetts, USA
bumchul.kwon@us.ibm.com

ABSTRACT

The recent prevalence of publicly accessible, large medical imaging datasets has led to a proliferation of artificial intelligence (AI) models for cardiovascular image classification and analysis. At the same time, the potentially significant impacts of these models have motivated the development of a range of explainable AI (XAI) methods that aim to explain model predictions given certain image inputs. However, many of these methods are not developed or evaluated with domain experts, and explanations are not contextualized in terms of medical expertise or domain knowledge. In this paper, we propose a novel framework and python library, MiMICRI, that provides domain-centered counterfactual explanations of cardiovascular image classification models. MiMICRI helps users interactively select and replace segments of medical images that correspond to morphological structures. From the counterfactuals generated, users can then assess the influence of each segment on model predictions, and validate the model against known medical facts. We evaluate this library with two medical experts. Our evaluation demonstrates that a domain-centered XAI approach can enhance the interpretability of model explanations, and help experts reason about models in terms of relevant domain knowledge. However, concerns were also surfaced about the clinical plausibility of the counterfactuals generated. We conclude with a discussion on the generalizability and trustworthiness of the MiMICRI framework, as well as the implications of our findings on the development of domain-centered XAI methods for model interpretability in healthcare contexts.

CCS CONCEPTS

• **Human-centered computing** → **Visualization systems and tools**; **Visual analytics**; **Human computer interaction (HCI)**.

KEYWORDS

explainable AI, human-centered AI, interactive visualizations, counterfactual explanation

1 INTRODUCTION

In the recent decade, research studies on cardiovascular imaging have grown significantly due to the prevalence of publicly accessible, large datasets that have been made available to artificial intelligence (AI) researchers [19, 48, 53, 58, 61]. Coupled with advancements in AI and machine learning (ML) methods, increasingly sophisticated models have been developed to support diverse clinical tasks from image segmentation to patient risk assessment and classification [48, 53, 61]. At the same time, the potentially significant legal and ethical consequences of medical AI classification models [10, 44, 65] have led to concerns about their trustworthiness, transparency, and interpretability [3, 64, 71]. To address these concerns, a range of explainable AI (XAI) methods, such as saliency maps [47], image perturbation [37], and example-based approaches [60], have been developed to explain model predictions given certain image inputs [12, 49, 62].

However, these methods often do not contextualize explanations in terms of relevant domain knowledge, thus limiting their interpretability and usefulness to domain experts. Saliency maps, for example, often produce outputs similar to simple edge detection, which can lead to risks of confirmation bias [1]. More seriously, many XAI methods and feature attribution tools are not evaluated with domain experts [57], and do not improve human interpretations of model behavior [6, 9]. Taken together, these limitations have led some researchers to conclude that "most work in explainable artificial intelligence uses only the researchers' intuition of what constitutes a 'good' explanation" [45], highlighting the need for contextualized XAI methods that are developed for and interpretable to domain experts. In the healthcare domain, for example, a cardiologist who wants to understand and evaluate a model trained to classify hypertension in cardiac magnetic resonance images (MRIs)

may want to probe how morphological features, such as the various heart chambers, influence model predictions. They may also want to validate model predictions against known medical facts, such as testing how patient age increases the predicted likelihood of hypertension.

Prior works have found that XAI methods centering user context and domain relevant concepts more effectively support model explanations [17, 29, 33]. In this paper, we build on these existing approaches to develop the MiMICRI (**M**orphological **MI**xing for **I**nteractive **C**ounterfactual **R**ecombined **I**mages) framework for domain-centered counterfactual explanation of cardiovascular image classification models. Counterfactual explanation is an XAI technique that, for a given input instance and an AI model, identifies the minimal perturbations necessary for the outcome predicted by the AI model to change. They help users reason about the causes behind certain outcomes, and what it would take to change that outcome. It has been argued that counterfactuals make for effective explanations because they are intuitive and actionable [13, 20], compliant with legal regulations such as the GDPR[1] [64, 65], and are model-agnostic [65].

MiMICRI relies on state-of-the-art segmentation algorithms to identify the domain-relevant morphological features in an image that can be perturbed to generate counterfactual images. To explain why a particular target image has a certain predicted label, domain experts can first select morphological segments to mask and replace with corresponding segments from other images, creating recombined instances. The same predictive model is then used to generate new labels for the recombined images. If a recombined image has a different label from the original target image, it is then a counterfactual of that target image. Since the framework generates recombined images by replacing *only* selected morphological segments, this allows users to attribute changes in model predictions to the replaced segments, thus providing an explanation of the model based on domain-relevant morphological image features.

To explore the effectiveness of our proposed MiMICRI framework, we implemented components of the framework as a Python package of the same name. We then worked with two healthcare experts to evaluate the interpretability and effectiveness of the counterfactual explanations generated. Overall, experts found that MiMICRI was helpful for reasoning about the relative influence of morphological features on model predictions, validating the model in context of known medical facts, and comparing between patient subgroups. However, they also raised potential concerns around the clinical plausibility of the recombined images due to the structural and physiological interdependence of image segments. In section 6, we discuss the implications of these concerns on the development of domain-centered XAI methods for model interpretability in healthcare domains.

To summarize, the main contributions of our work are the following: i) the MiMICRI visualization framework for counterfactual explanation and inspection of cardiovascular image classification models, ii) the MiMICRI Python visualization package, which provides an implementation of the proposed framework, iii) findings from an evaluation of MiMICRI with two experts, and iv) a discussion of the generalizability, trustworthiness, and implications

---

[1]General Data Protection Regulation

of our findings for developing domain-centered XAI methods that enhance model interpretability in healthcare domains.

2 RELATED WORK

This section examines recent advancements made in the fields of AI, explainable AI, and visual analytics applied to medical imaging data, with a focus on cardiovascular imaging.

2.1 Cardiovascular image analysis using AI/ML/DL Models

Recently, the number of research studies on cardiovascular imaging has grown significantly due to the prevalence of publicly accessible, large datasets and the rise of artificial intelligence (AI) methods. In addition, the rapid developments in the field of computer vision have made it possible to apply neural networks trained with medical imaging data to a variety of tasks such as automated segmentation of cardiac structure, volumetric estimation, disease diagnosis, and outcome prediction (death or cardiac events). Various model architectures, ranging from convolutional neural networks to transformers, have been adapted for the purpose of predicting clinical outcomes based on information extracted from medical imaging datasets. There are various types of medical images that are being used for such purposes, such as Cardiovascular Magnetic Resonance (CMR), echocardiography (ultrasound), computed tomography (CT), and nuclear imaging. Active research in this field has become feasible, thanks in part to the open accessibility of imaging data and their corresponding clinical measurements being made available to AI researchers (e.g. UK Biobank [19, 58]). For more detail, readers may refer to the following reviews [48, 53, 61].

2.2 XAI approaches for medical image analysis

Many XAI approaches have emerged within the field of computer vision and have been adapted for the analysis of medical imaging models. Previous surveys [12, 49, 62] offer comprehensive reviews of these XAI methods. Of these, one of the most popular approaches for cardiac imaging analysis is the use of saliency maps (pixel-attribution maps). These model-based approaches visualize attention by employing various class activation mapping methods, and applying a heatmap to highlight the pixels in an input image that contribute most to the outcome. For instance, Grad-CAM has been employed to identify the areas in echocardiograms of newborns that contribute the most to pulmonary hypertension [47]. Perturbation-based approaches, which are model-agnostic, explain predictions by replacing portions of an input image and displaying the predicted outcome of the altered image. They have been employed to identify critical brain features contributing to autism spectrum disorders [37]. In contrast, example-based approaches explain predictions by presenting a similar sample from the training data. Still other researchers have proposed generating image patches from the training data and displaying prototypical examples that resemble a given image [60].

While these automated approaches can provide reasonable explanations, they may fall short when it comes to explaining why a specific instance is predicted as a certain class by the underlying model. Pixel-attribution methods, such as saliency maps, assign importance to individual pixels, which may not be readily

interpretable to clinicians. Furthermore, these explanations have been found to resemble simple edge-detection [1], and do not pass benchmark tests for interpretability [9]. Feature perturbations have shown some promising results, but it can be challenging to automatically perturb images in a biologically plausible manner. Similarly, example-based approaches may not elucidate how the model makes decisions based on the entire image, as they only display similar samples from prototype patches.

In response to these challenges, prior studies have shown that when working with tabular data, counterfactual explanations performed better than saliency maps and reduced over-reliance on "wrong" AI outputs during clinical decision making [36]. This suggests the potential for counterfactual image generation to also serve as an effective approach for explaining image classification models in the same context. Many studies have proposed semantic counterfactual image generation approaches to explain AI models for general (i.e. non-medical) applications [26, 63, 74]. These frameworks extract semantic regions from a target image and replace a subset of these regions to generate a counterfactual of the query image. However, users often have no control over the image regions replaced, which may not always be relevant or appropriate for a particular domain application. Additionally, in all frameworks, only a single counterfactual is produced to explain a target image. This can limit the effectiveness of the frameworks since having multiple counterfactuals with slightly different perturbations provides a better explanation than a single sample [13, 65]. Furthermore, many existing counterfactual image generation techniques were also not developed for medical domains where there is a need to match and compare patients based on demographic characteristics [70]. In these contexts, users may require more control over the source of the replacement regions in order to ensure that the characteristics of the source and target images/patients are sufficiently similar.

While there exists an approach, GANterfactual [43], which applies a similar segments-based counterfactual generation framework to medical images, GANterfactual automatically computes image segments using the SLIC algorithm [50], resulting in unlabeled segments that do not necessarily map to high-level features (e.g. morphological features). In this paper, we build on existing frameworks to develop a domain-centered counterfactual generation approach for cardiovascular image classification models that provides interpretable and domain relevant explanations for medical experts.

2.3 Visualization for XAI in (medical) image analysis

To date, various visualization systems have been developed to explain AI trained on image data. They help in identifying out-of-distribution samples [11], guiding clustering algorithms [31], diagnosing and refining the layers of deep neural networks [38–40, 66, 68], searching for similar images [8], identifying visual concepts [24], detecting and resolving biases [32], and comparing the performance of generative adversarial nets [67]. Of these visual analytics applications, some were developed specifically for the medical domain [8, 31, 70]. Motivated by a need to explain clinical decision support systems and improve user acceptance of AI-assisted decision making, these prior studies collaborated with physicians and

pathologists to design and build visualization systems that support expert-guided refinement of medical image search results [8] and image classification [31], as well as the identification of key image features that influence AI image analysis [70]. Findings from the CheXplain system [70], in particular, demonstrated the effectiveness of comparative and contrastive examples for explaining model predictions. However, CheXplain selects these contrastive examples from in-data set images instead of generating new counterfactual instances. Furthermore, participants also mentioned that since physiological features may appear different for patients with different characteristics/demographics, there remained a need for tools that help them group patients and compare features by subgroup.

At the same time, a range of visual analytics systems have also been developed that broadly aims to provide counterfactual explanations of machine learning models across domains. In these tools, users were able to guide the counterfactual generation process [13], analyze the importance of various features for model prediction [20, 69, 76], and compare counterfactuals between user-defined subgroups [13, 21, 69, 76]. In evaluation studies, not only were these visual analytics systems found to provide effective explanations of AI models, the ability for subgroup comparison was also crucial for evaluating model fairness [69] and revealing bias [13]. However, the majority of existing systems were developed primarily for tabular data. For example, some functions of the What-If Tool [69], such as data perturbation and editing, could not be used on image data sets. These limitations highlight the challenges of generating counterfactual explanations when working with image data. The concept of a domain-relevant "feature" cannot be captured through pixel perturbations, which complicates the process of generating an interpretable and meaningful counterfactual through minimal edits. In this paper, we address this challenge with MiMICRI, a domain-driven counterfactual explanation framework for cardiovascular image classification models. We describe the details of this framework in the following section.

3 THE MiMICRI FRAMEWORK

The MiMICRI framework (Fig. 1) is developed to explain decisions made by cardiac image classification models using counterfactual explanations grounded in relevant context and domain knowledge. Consider an AI model trained to predict the likelihood of a patient having hypertension from their cardiac MRI data. Domain experts – such as data scientists and healthcare providers – may want to understand how the model is making predictions based on domain-relevant MRI image/video features. These experts may also be interested in validating the model based on known medical knowledge. For example, when doctors evaluate for hypertension in cardiac MRIs, they expect the LV myocardium to be thickened. They may thus want to validate that changing this morphological feature influences model predictions in the expected direction (i.e. thinner LV myocardium decreases the likelihood of predicted hypertension and *vice versa*).

3.1 Counterfactual Generation Criteria

To guide our initial development of MiMICRI, we identified three criteria from prior works that should be satisfied to produce a good counterfactual:
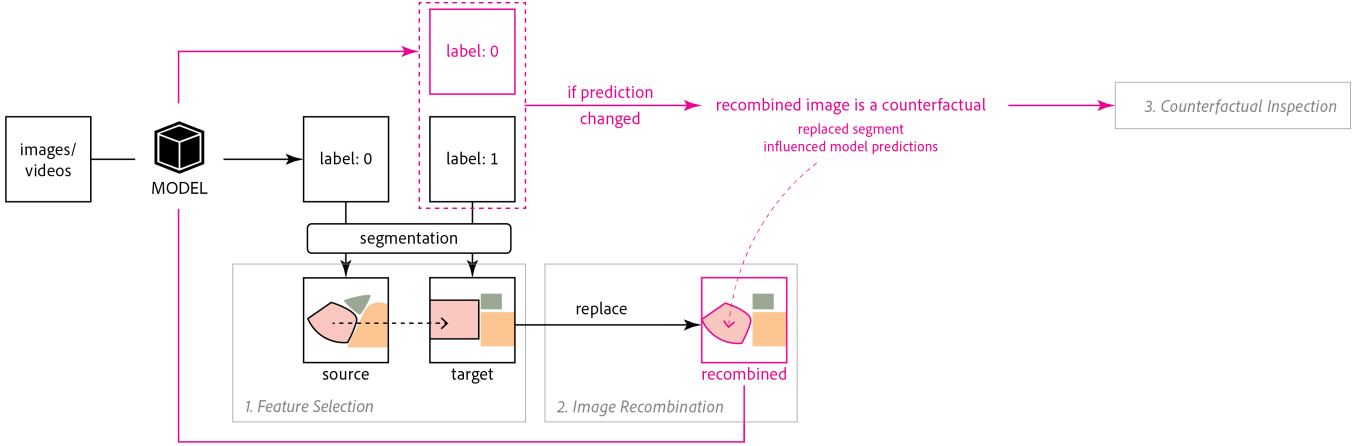
Figure 1: A high-level summary of the MiMICRI framework. To explain a classification MODEL, users can identify domain-relevant semantic image segments in each image in the data set, then replace segments in a target image with corresponding segments from a source image. This creates a recombined image. If the MODEL predicts that the recombined image has an alternate label to the target image, this recombined image is a counterfactual, and we can conclude that the replaced segment changed the MODEL prediction.

(1) **Sparse/Minimal.** A counterfactual should make the minimum amount of changes needed to change the predicted label of an input image or video [13, 26–28, 65, 74, 76]. Users should be able to quickly identify the changes made, and reason about how they affected model predictions.

(2) **Plausible.** Also referred to as post-hoc validity, counterfactuals should be data instances that can realistically occur [13, 27, 28]. For example, completely removing the left ventricle from an MRI will likely change the predicted model output, but this is not informative because such an MRI is unlikely to be obtained naturally from a patient.

(3) **Meaningful.** Since counterfactuals are meant to explain model predictions, this criteria ensures that the counterfactuals generated are sufficiently human-interpretable and explanatory [26].

While these criteria are best-practices that apply to counterfactual generation regardless of data modality, it is not always clear how they should be extended to images and videos in particular. For instance, a counterfactual image that is *sparse/minimal* might not satisfy the *meaningful* criteria. Consider adversarial attack algorithms (for example [16, 18, 25]) that make minor perturbations to image pixels to change the model prediction. While these approaches might meet the *sparsity* criteria, the changes are often undetectable to the human eye, and the resulting images are rarely referred to as counterfactuals since they do not provide *meaningful* explanations of the model [65].

The MiMICRI framework is inspired by existing image data augmentation methods such as CutMix [73], Mixup [75] and Cutout [15] that identify a bounding box in an image that can be masked and/or replaced with pixels from another image, thus creating new images from available data. However, in these prior methods, the resulting outputs are not designed to appear realistic (or *plausible*), often producing images with missing patches [15] or areas that are visibly recombined [73, 75]. MiMICRI extends these prior approaches by using segmentation algorithms trained on expert labels to identify domain-relevant morphological features to mask and replace. In cardiac MRI data, for example, the left ventricle (LV) cavity, LV myocardium, and right ventricle (RV) cavity may be the relevant features to segment and identify.

To explain the predicted label for a particular target image, experts using MiMICRI can select and replace segments in the target image with corresponding segments from source images that have a different predicted label. This replacement process generates new recombined images. If a recombined image is also predicted to have a different label from the original target image, it is then a counterfactual of the target image. Since only relevant morphological segments were replaced, and replacement segments were sampled from real (in-data) patient MRIs, this approach ensures that the counterfactual is *plausible* and *meaningful*. Furthermore, since the rest of the original target image remains unaltered, the changes are also *sparse/minimal*, allowing users to attribute any differences in model predictions to the replaced segments. The MiMICRI framework has four main components: *Image Segmentation (Pre-processing)*, *Feature Selection*, *Image Recombination*, and *Counterfactual Inspection* (Fig. 2). We describe each component in detail in the following sections. Our example uses MiMICRI to explain a model trained to predict hypertension likelihoods from cardiac MRI data.

3.2 Pre-processing: Segmentation

MiMICRI assumes that all images/videos in the data set have been segmented to identify the high-level semantic visual features relevant to the user and the domain. While this pre-processing can be completed manually, it is also possible to train a segmentation algorithm based on expert labels. In our example (Fig. 2), three key morphological features are identified: the LV cavity, the LV myocardium, and the RV cavity.
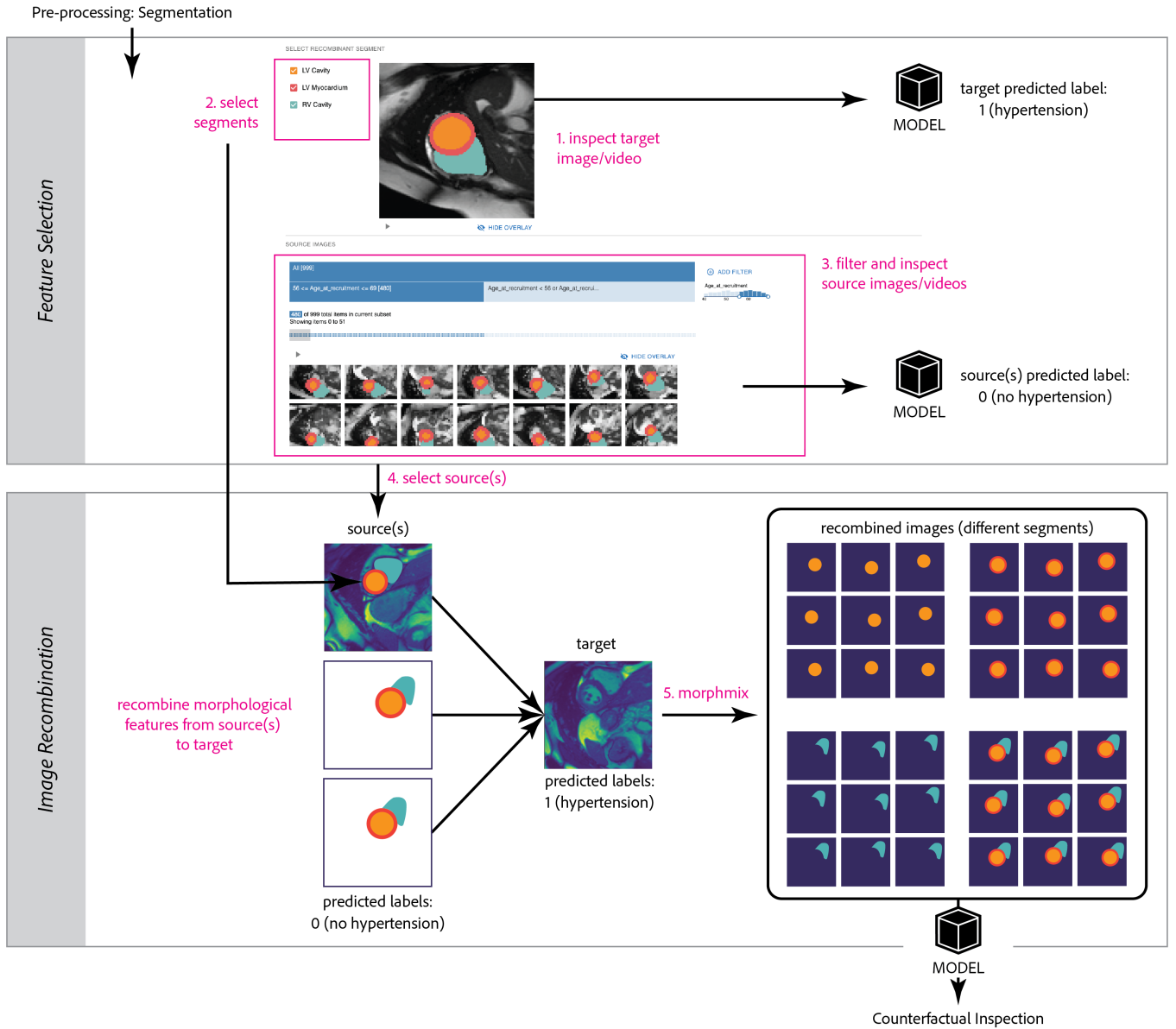
Figure 2: The detailed MiMICRI framework and corresponding visualization modules. *Top:* **Users interactively select source and target images or videos. They can also select combinations of segmented features to be replaced. Selected segmented areas (e.g. circular shapes in orange for LV Cavity) are overlaid on top of MRIs at their corresponding positions.** *Bottom:* **Selected morphological segments from target(s) are masked and replaced with corresponding segments from source(s). We implemented the *MorphMix* method to do this. New predicted labels are generated for the recombined images or videos. Users interactively inspect the model by viewing the counterfactuals generated for each replaced segment.**

3.3 Feature Selection

The framework starts with selecting the images/videos and cardiac segments for counterfactual generation. Users can first select a target cardiac MRI with a certain model prediction (e.g. hypertension) to be explained (Fig. 2, 1). They can also select a set of source MRIs with the opposite predicted label (e.g. no hypertension) (Fig. 2, 3 and 4), and the combination of segments to replace (Fig. 2, 2).

3.4 Image Recombination

Having selected the source image(s) and segment(s) to replace, MiMICRI next generates all possible recombined images (Fig. 2, 5). In this paper, we developed the *MorphMix* method (see Section 4.3) for segment replacement, however, other methods can be considered in future work, such as training a generative adversarial network to inpaint a masked area.

3.5 Counterfactual Inspection

Finally, after generating multiple recombined images, the same hypertension model can be used to predict the disease likelihood of the recombined images. Images with a predicted label different from the original target image would be considered counterfactuals. Since the recombinations are identical to the target in all places except for the replaced segments, we can conclude that any counterfactual predictions (a different label from the original target) must be attributed to the replaced segments. Furthermore, if experts selected multiple different combinations of segments for recombination, they can also compare the relative influence of different morphological features on model prediction.

4 MiMICRI

To implement the steps of the MiMICRI framework, we built a Python visualization package[2] that includes a selector module and the *MorphMix* method. The selector module helps users interactively select source and target images, as well as segments to be replaced. Using the selected images and segments, the *MorphMix* method then generates recombined images by replacing selected morphological segments in the target images with corresponding segments in source images. The recombined images can be fed into the same classification model to generate new predicted labels. In the following sections, we detail the implementation of the MiMICRI package. MiMICRI uses the IPyWidgets[3] framework and is designed to work in JupyterLab. Module front-ends are implemented in React[4], D3 [7] and WebGL[5].

4.1 Data Set, Hypertension Predictive Model, and Image Segmentation

To demonstrate MiMICRI, we trained a video classification model to predict hypertension from input MRIs. Each MRI is a video with 50 image frames. We used a 3D ResNet [22] for the model architecture comprising 50 layers with alternating 3D convolutional and batch normalization layers following the guidelines from the original paper. For the dataset, we randomly split 23,043 patients' cardiac MRI scans of UK Biobank [58] into training (18,434) and test (4,609) datasets. We downsampled and standardized the MRI scans to 50×128×128 (frames×width×height). For training, we selected a batch size of 16 to balance memory usage and computational efficiency. To facilitate convergence, we used the Adam optimizer with an initial learning rate of 0.001, which can be adjusted during training using learning rate schedulers, such as a step decay or cosine annealing schedule. After training for 100 epochs, the model achieved a performance of accuracy 0.87 and auroc 0.65. We saved the weights of the trained model and used it for the demonstration and use case in the following sections. To protect the privacy of patients' healthcare data, we do not describe personally identifiable information of patients. We also exclude any original MRI files in the code repository and in system screenshots. We segment all cardiac MRIs using the state-of-the-art *ukbb_cardiac* library [4, 5].

4.2 Feature Selection: Selector

The selector visualization module (Fig. 3) is designed to help users systematically generate recombined images by selecting 1) the segments to replace, and 2) the source images to be used as replacements. Before using the selector module, users should segment all files to identify domain-relevant morphological features. In our example, we use the LV cavity, the LV myocardium, and the RV cavity segments. Users can instantiate the module using the `Selector()` function. This function accepts as parameters the `target` image to be explained, `sources` with a different label from the target, an `imgReader` function that converts all file paths to a *numpy* array, and a `segMap` object that includes the names of the segments identified. On load, the selector module will display the target image in the top panel (Fig. 3, 1). The bottom panel is where users can filter, explore, and select source images for recombination.

4.2.1 Select image segments for recombination. All segments identified during pre-processing are shown in overlay over the target and source images or videos. Users can manually select the segments to be used in the subsequent MorphMix method (Fig. 3, 2). This allows MiMICRI users to determine the image features that should be replaced based on domain knowledge and prior expertise. Users can also toggle the visibility of the overlay for the target and source panels separately. Once users have selected the segments they want to recombine, they can access the indices of the segments using the `.segments` command in a subsequent JupyterLab notebook cell.

4.2.2 Dynamically filter, explore and select source images for recombination. Using the drop-down menu in the bottom panel of the selector module, users can successively filter the source subset based on their demographic data (Fig. 3, 3) and view the subset of filtered images or videos. The distribution of each variable is visualized as a histogram, and the range of selected values can be modified by dragging the range slider along the *x*-axis of each histogram (Fig. 3, 4). For example, if a target cardiac MRI is from an individual aged 65 years, we may want to select source MRIs from patients with a similar age range to control for any age-related differences. Once a value range is selected, the icicle plot dynamically updates to reflect the cohort size after the new filters are applied (Fig. 3, 5). Below the icicle plot is a unit visualization of all possible source images provided (Fig. 3, 5). Selected source images are colored dark blue and left-aligned such that the unit visualization is visually consistent with the lowest layer of the icicle plot. A gray rectangular brush can be dragged along the horizontal axis so that users can view the selected source images/videos in the tiled display below. By design, ~50 images or videos can be displayed at any time for rendering efficiency. Once users are content with the selected sources, they can access the IDs of the selected items using the `.subset` command in a subsequent JupyterLab cell.

4.3 Image Recombination: MorphMix

After selecting the target, source(s), and segment(s) to replace, the `Selector.morphmix()` function can be used to generate recombined images. In the case of video data, each frame can be processed as a separate image. The MorphMix method first masks the selected morphological segments of interest in an image or video – that is to say, the pixels belonging to the segment(s) are removed (Fig. 4,
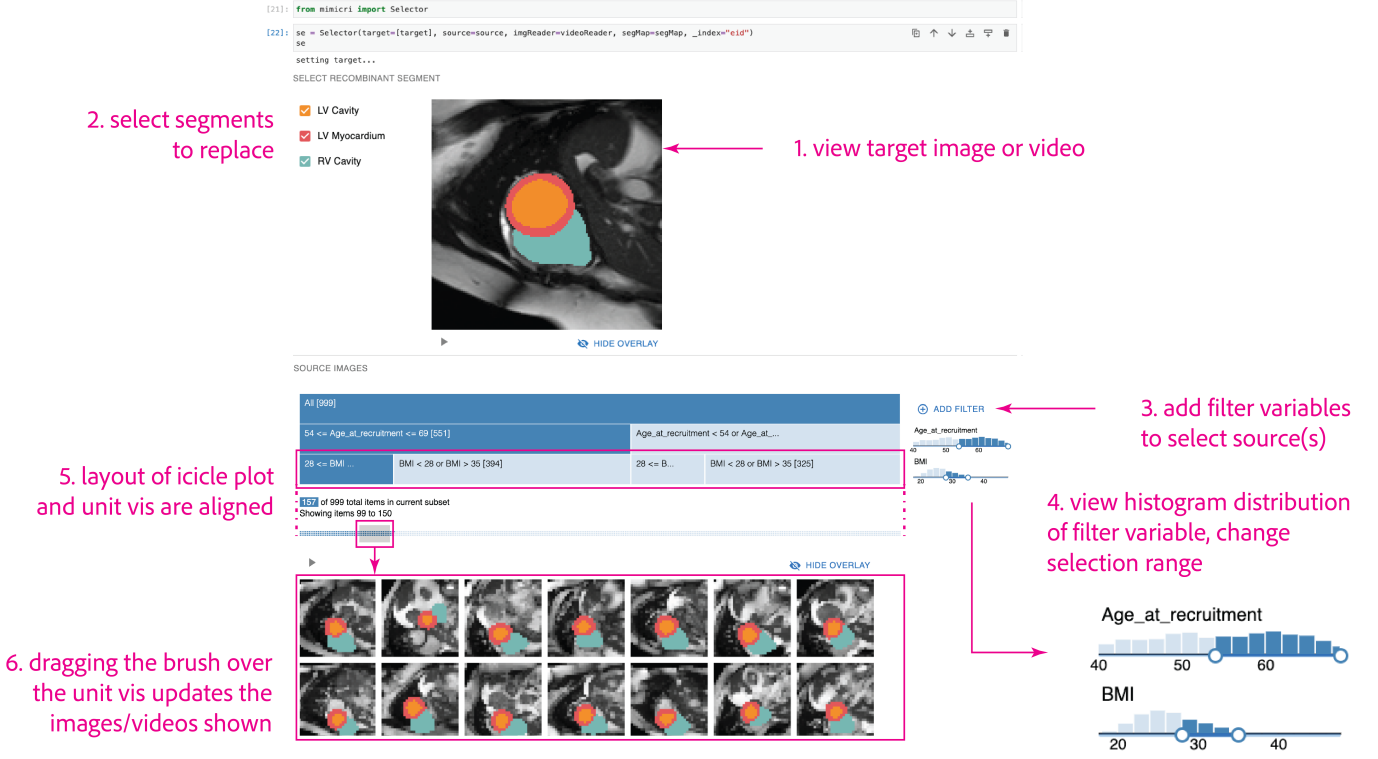
Figure 3: The MiMICRI selector module. In this module, users can 1) view a target image or video, 2) select segments to replace, 3) select source images by demographic by adding filter variables, 4) view and modify the range of selected values for each demographic filter, 5) view the selected subset in an icicle plot and corresponding unit visualization, and 6) view detailed source images or videos by dragging the brush over the unit visualization. In both the top and bottom panel, the visibility of the overlay can be toggled. If the files are videos, the videos can be paused. The panels can be resized.

Middle). More than one segment can be selected during this process, and different sets of segments can be tested combinatorially. The method then replaces the masked target segment(s) with the corresponding segment(s) from the selected source image(s) (or video frame). In our MorphMix implementation, we use a heuristic method to align the centroids of the segment(s) to be replaced. We then start at the centroids and use flood-fill to copy pixels from the source image into the target. This outputs a recombined image with only the segment(s) of interest replaced. The rest of the image pixels will be identical to the original target image (Fig. 4, *Right*).

4.4 Counterfactual Inspection

Finally, new labels can be generated for all recombined images using the original predictive model to identify counterfactuals (i.e. recombined images with a different label from the original target image being explained). In our example, we took the first 100 MRIs in the data set, of which 21 were predicted to have hypertension, and 79 were predicted to have no hypertension. We ran `Selector.morphmix()` using the hypertension group as sources and the no hypertension group as targets, replacing all possible combinations of the three identified cardiac segments. We then repeated the run, switching the source and target groups. In total, we generated 23226 recombined images (21 hypertension × 79
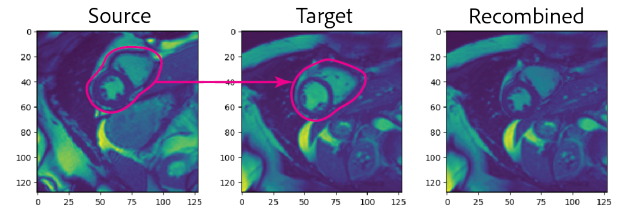

Figure 4: *Left:* **A single frame from a source cardiac MRI.** *Middle:* **A single frame from a target cardiac MRI with all 3 cardiac segments to be masked.** *Right:* **A recombined frame where pixels from the 3 corresponding cardiac segments in the source image were copied into the target image. Note how, with the exception of the replaced segments, the recombined image is identical to the target.**

no hypertension × 2 runs × 7 segment combinations). Using the same hypertension predictive model, we generated labels for all recombined images (Table 1). The results correspond well to expected intuitions, where replacing more segments influenced model predictions to a greater extent, resulting in more counterfactuals generated. This suggests that the hypertension predictive model in

segment(s) replaced	counterfactuals (count)	unchanged (count)	% counterfactuals
LV cavity	520	2798	0.157
LV myocardium	0	3318	0.000
RV cavity	496	2822	0.149
LV cavity + LV myocardium	639	2679	0.193
LV myocardium + RV cavity	762	2556	0.230
LV cavity + RV cavity	471	2847	0.142
LV cavity + LV myocardium + RV cavity	782	2536	0.236

Table 1: Count and proportion of counterfactuals generated from 23226 recombined images. In general, replacing more segments influenced model predictions to a greater extent, resulting in more counterfactuals.

our example has, correctly, learned associations between cardiac features in an MRI and hypertension likelihoods. However, there remain notable exceptions. For example, unlike established medical cases, our results indicate that replacing the LV myocardium alone did not affect model predictions at all. We discuss this further in expert evaluations (Section 5.2).

5 EVALUATION

Since there are few methods for counterfactual generation in healthcare, and, to the best of our knowledge, none for cardiovascular imaging specifically, we evaluate the MiMICRI framework in two ways: first by inspecting the algorithmic segmentation outcomes of the recombined images generated, and secondly, through human validation in collaboration with two clinicians who have expertise viewing and assessing cardiac MRIs.

5.1 Evaluation by Segmentation

Before presenting MiMICRI to experts, we first evaluated the generated recombined images by running the same *ukbb_cardiac* segmentation model on a random set of recombined images with different combinations of source, target, and replaced segments. We then manually inspected the segmentation outputs. As seen from Fig. 6, the segments identified corresponded well to their respective source and target images, providing initial validation that the framework produced recombined images of sufficient similarity to real MRIs for cardiac segments to be identified algorithmically.

5.2 Expert Evaluation

We worked with two healthcare domain experts (2M) to evaluate the MiMICRI framework. E1 is a pediatric cardiologist of 9 years. His work involves performing clinical cardiovascular MRI (CMR) exams, deriving imaging-based biomarkers in congenital heart disease (CHD), as well as studying the use of wearable biosensors for CHD. He uses AI models daily in CMR exams, specifically for image segmentation, but has yet to apply them to patient prognosis as such models are still under development. E2 is a data scientist of 8 years in functional brain imaging, and who has recently transitioned to cardiac imaging. His work mainly involves developing novel ML algorithms related to cardiac MRI acquisition and shape modelling of the heart, as well as supporting data extraction, transformation and loading processes for clinical outcomes research. In prior work, E2 has used convolutional neural networks (VGG16 [56]) and

functional MRIs (a type of brain imaging) to measure differences between young and older adults during visual perception.

In a series of three meetings with the experts, we presented the end-to-end MiMICRI framework (Section 4) using the hypertension prediction model for demonstration. The meetings took place virtually, and lasted about an hour each. Separately, we also provided a set of *MorphMix* recombined cardiac MRIs to the experts for evaluation. The recombined cardiac MRIs shared with the experts varied by source image, target image, and segments replaced. In following sections, we discuss the expert feedback and concerns raised.

5.2.1 MiMICRI's framework and implementation provided more domain-relevant and interpretable explanations of model outputs than current methods. As part of his prior expertise, E2 has used existing explainability tools such as saliency maps [55] and Grad-CAM [51, 52] to inspect and evaluate the models he developed. Compared to the previous methods, E2 found MiMICRI *"less technically demanding"*, as it hides the technical details and allowed users to inspect model performance through the counterfactual explanations generated. In contrast, tools like Grad-CAM expected users *"to have certain knowledge about the structures of those models and the underlying machine learning frameworks"* (E2). Users of MiMICRI could also make domain-relevant manipulations of the counterfactual explanations that are relevant to the clinical scenarios, which is rarely offered by existing explainability tools. More crucially for domain experts, these recombined images helped them interpret and validate the model in context of their domain knowledge. When viewing Figure 4, for instance, E1 quickly reasoned that *"one would hope that you're not looking at abdominal fat... if your AI model is only just looking at subcutaneous fat in the abdomen or in the chest wall, and then it's making a prediction based on that, it probably doesn't matter which part of the heart you put into the new image or not."* Taken together, both experts found that MiMICRI was more effective and interpretable than many existing approaches.

5.2.2 Allowing users to select combinations of segments to replace across multiple source and target images supported more nuanced interpretation of model predictions. During the evaluation, experts were particularly interested in the possibility of replacing specific segments of cardiac MRIs. For example, E1 asked about combining *"the small LV with the dilated RV"* or *"a thick LV with a normal thickness RV"*, going on to explain that *"those are probably going to be more physiologically explicable than taking the heart, the entire heart image, out of one person putting it in the other."* Additionally,

E2 also mentioned that *"an advantage of this toolbox is that it generates positive (counterfactuals) and negative (not counterfactuals) recombined images."* As such, though only some recombined images are counterfactuals, by selecting different combinations of source, target, and morphological segments, users can generate multiple recombined images that, when aggregated (as in Table 1), explain how each morphological feature affects model predictions.

5.2.3 Filtering and creating subgroups from demographic data is important to physicians but less crucial to data scientists. In clinical settings, AI models trained on imaging data may not always incorporate demographic information about the patient as input. As such, E2, a data scientist, found that the filtering and subgrouping features in MiMICRI *"might be unnecessary from a model development standpoint"* if the variable was not used during model training. However, he also acknowledged that *"demographic information of a patient is crucial in clinical settings."* This is confirmed by E1, a physician, who emphasized that when using MiMICRI, *"you'd have to do that within people with similar other clinical characteristics so that you're not biasing your data set and then having [the model] predict off the clinical factors".*

In medical practice, BMI, weight, and abdominal fat, are all potential biological indicators of increased hypertension likelihoods. For a physician using a hypertension predictive model, it is thus necessary to validate that the model provides *"additional value"* (E1) beyond what is known about the patient. These concerns apply to any specific clinical task or clinical prediction where *"there's a whole bunch of confounders that come into play that you also have to account for in the image let alone in the clinical history"* (E1). This also confirms prior works that found a need to compare patients with similar characteristics since physiological features may appear different between subgroups [70].

5.2.4 Results of the MorphMix method are not always clinically plausible since segments are structurally interdependent. While *MorphMix* was able to generate recombined images where the replaced cardiac regions can be identified with high fidelity by our segmentation algorithm, E1 found that recombined cardiac MRIs were sometimes clinically implausible. As he described: *"the LV myocardium and blood pool (cavity) are completely interrelated because one bounds the other, so you can't change one without the other."* This likely also explains the unexpected result in Table 1, where replacing the LV myocardium alone did not generate counterfactuals because *"the structures are interrelated, so you can't just pick the body of the LV out without changing the shape or features of the other objects"* (E1).

This highlights the challenge of recombining image segments when they are structurally interdependent. Even in cases where the segmentation algorithm can accurately identify features in the recombined image, there may still be errors that are apparent to domain experts such as E1 (Fig. 5, *Middle*). This suggests that generating recombined images that better meet the *plausibility* criterion will require alteration to parts of the image surrounding the replaced segment. However, this would come at a trade-off to requirements for *sparse/minimal* changes, since image differences will no longer be limited to the segment being replaced. In future work, a more in-depth study of different methods to mask and replace image segments may better balance the trade-offs between these criteria.

6 DISCUSSION

From expert evaluations, we found that the MiMICRI counterfactual explanations helped users reason about model predictions based on morphological structures and established medical knowledge (Sections 5.2.1 and 5.2.2). These findings provide further support for frameworks proposed in prior works that center user context and domain relevant concepts to provide more effective model explanations [17, 29, 33]. However, our expert evaluations also surfaced concerns about the generalizability, trustworthiness, and plausibility of our framework. In this section, we discuss the implications of these concerns on how domain-centered XAI methods should be designed and developed.

6.1 Generalizability

Overall, while the experts found that the counterfactuals were useful, they also raised concerns about how well the MiMICRI framework applies to related data sets, such as other types of medical imaging. To extend MiMICRI to other medical domains, generalist models, such as the Segment Anything Model (SAM) [30], can be fine-tuned to specific types of medical imaging [54] and diseases [59]. Alternatively, recent advances in transformers for medical images [42], active learning [41], and parameterized approaches [35] can also be explored for segmentation.

At the same time, it must be highlighted that the *MorphMix* method is most appropriate for organs that have *"well defined anatomical structures that you can easily replace"* (E2), such as cardiac and skeletal structures. It may not be as effective in generating recombined images in organs with complex contours, such as cortical foldings in the brain (E2). Similarly, E1 also mentioned that *"there are some organs or images that are more amenable to generating realistic/biologically plausible images, so... human oversight is needed to make the judgement [for when MiMICRI should be used]."* This concern emphasizes the necessity of collaborating with domain experts to determine whether an XAI method would be applicable to a particular task and usage scenario.

6.2 Model versus Explanation Trustworthiness

More crucially, it is necessary to acknowledge that while the goal of many healthcare XAI tools, including MiMICRI, is to increase user trust in AI models [2, 14], simply explaining a model does not necessarily result in greater trust. And nor should it. In our approach, while the recombined counterfactuals helped medical experts reason about how the model made predictions, the counterfactuals were not always clinically plausible (Section 5.2.4). Or as E1 described, *"you can manipulate the image but that's not what a real person looks like."* This implausibility led E1 to go on to comment that *"I'm just worried about the implementation of MorphMix, not the underlying idea of it".* This demonstrates that while explanations can be useful, they may also be another source of error. In our tool, some counterfactuals were valid while some were implausible (examples of both in Fig. 6), and we should take care that the recombined images are never presented as real MRIs. Furthermore, we should also ensure that there is transparency in both model and explanation such that experts can evaluate the trustworthiness of both methods.
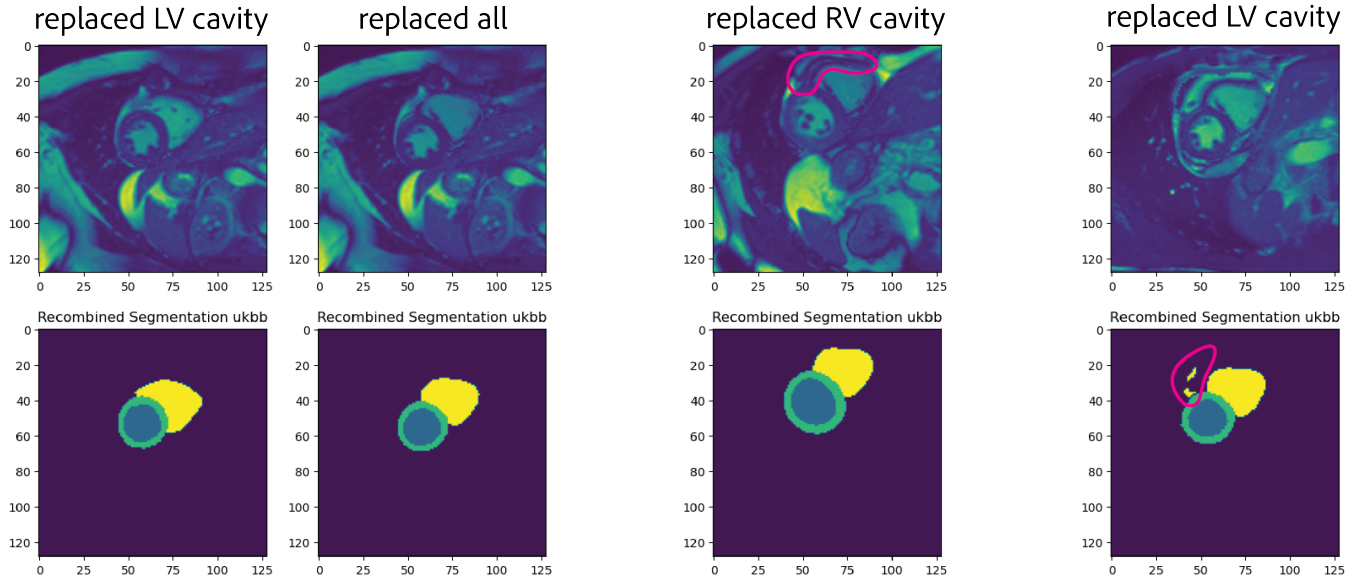
Figure 5: Expert feedback for four recombined images. Original source and target images were omitted for a compact layout. *Left:* Two recombined MRIs and corresponding segmentation that were acceptable to experts. *Middle:* A recombined image with a double wall in the RV. *"Though that may not affect segmentation, it would likely affect any whole-image analysis"* **(E1). *Right:* A particularly egregious example where physiological features were disordered and segmented regions contain artifacts.**

This distinction between model and explanation has been highlighted in prior works by Sperrle et al. [57] and Hohman et al. [23]. Hohman et al., in particular, found it concerning when participants were quick to rationalize explanations without questioning. However, their evaluation was performed with randomly selected data scientists who were not necessarily domain (in their case, the housing market) experts. In contrast, our work suggests that domain experts may be more cautious when evaluating XAI tools. Such critical evaluations may be particularly important in medical domains where misplaced trust in erroneous predictions can result in serious adverse consequences [34, 46, 77]. Future work developing domain-centered XAI methods should thus look beyond just contextualizing explanations based on domain relevant information, but also ensure that both model and explanation are evaluated for trustworthiness with experts.

6.3 Supplement, not Substitute, Expertise

Finally, while the MiMICRI framework can be useful for explaining cardiovascular image classification models, the **explanations must be interpreted in context of domain knowledge, and should not substitute real-world clinical practice.** In our hypertension example (Table 1), we found that replacing the LV myocardium alone generated no counterfactuals. This contradicts known medical facts where LV myocardial thickness is one of the first things clinicians look at to see whether patients have the worst stage of hypertension (E1). In this case, our results should not be taken as new medical claims. Instead, they are more likely caused by the structural interdependence of cardiac segments (Section 5.2.4), and should lead to a careful evaluation of the explanations presented.

Additionally, our process of developing and evaluating MiMICRI also revealed that **it is insufficient for XAI tools to rely solely on established best-practices or algorithmic evaluations.** We had initially determined three criteria of "good" counterfactuals from prior work (Section 4.3), and later validated the recombined images by ensuring that constituent morphological features can be accurately identified using the same segmentation algorithm (Section 5.1). However, concerns were still surfaced during the evaluation about the clinical plausibility of the counterfactuals (Section 5.2.4). Even in cases where the segmentation algorithm accurately re-identified cardiac segments, there were inconsistencies in the recombined MRI that were apparent to clinical experts (Fig. 5, middle). This highlights the gap in understanding between our definition of the *plausibility* criterion and the expectations of domain experts. It also reiterates findings from earlier work that XAI methods sometimes rely on "researchers' intuition of what constitutes a 'good' explanation" [45] and may not meet the needs of intended users. As such, for domain-driven XAI methods to truly enhance the interpretability and trustworthiness of AI models, best-practices and algorithmic evaluation are not enough – they must be developed from an orientation that centers and builds on the knowledge and expectations of domain experts.

6.4 Limitations

While this work aims to provide domain-driven explanations of AI models in healthcare contexts, it is necessary to highlight that there are inherent risks associated with the use of AI in medical applications. In radiology, for example, it has been found that *"the occurrence of AI errors strongly influences treatment outcomes"* regardless of the experience of the radiologist or their familiarity

with AI tools [72]. While some studies have found that unlike other explainability techniques (such as saliency maps), counterfactuals effectively reduce over-reliance on 'wrong' AI outputs during clinical decision making [36], there remains a need for us, as users and developers of XAI tools, to be cautious that the explanations do not lead to over-reliance on the AI.

Furthermore, as we better understand the uneven distribution of disease likelihoods in the population – particularly for minoritized and underserved demographics – it becomes increasingly important to validate model predictions for specific population subgroups to ensure more equitable outcomes. Although our tool was designed for users to view data distributions and create subgroups via the histogram filtering feature, this approach is time consuming, and assumes that users know the demographics to inspect *a priori*. In future work, we plan to explore extensions to our tool that enhance subgroup analysis, such as algorithmic subgroup detection methods. Finally, we also want to highlight that some groups, such as women and children, are underrepresented in AI algorithms and healthcare data sets in general. This fundamental difference cannot be easily mitigated using our proposed framework, and would require more systematic changes in how we collect and curate data sets to ensure fair and equitable subgroup representation in clinical AI tools.

7 CONCLUSION

In this paper, we proposed the MiMICRI framework for conducting domain-driven counterfactual image analysis to understand cardiovascular image classification models. This involves replacing high-level image features from one image into another to probe the relative importance of various morphological features for model predictions. We implemented the framework as a Python visualization package, then evaluated its effectiveness with two medical experts. Our findings highlighted the benefits of a domain-driven counterfactual explanation method, but also surfaced concerns about the generalizability and trustworthiness of our proposed framework. We discussed these implications of our findings on the generalizability and trustworthiness of XAI methods, as well as the need to center and supplement domain expertise when developing tools for enhancing model interpretability in healthcare domains.

ACKNOWLEDGMENTS

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).

[2] AS Albahri, Ali M Duhaim, Mohammed A Fadhel, Alhamzah Alnoor, Noor S Baqer, Laith Alzubaidi, OS Albahri, AH Alamoodi, Jinshuai Bai, Asma Salhi, et al. 2023. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion* (2023).

[3] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. 2021. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences* 11, 11 (2021), 5088.

[4] Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M Lee, Nay Aung, Elena Lukaschuk, Mihir M Sanghvi, et al. 2018. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance* 20, 1 (2018), 1–12.

[5] Wenjia Bai, Hideaki Suzuki, Chen Qin, Giacomo Tarroni, Ozan Oktay, Paul M Matthews, and Daniel Rueckert. 2018. Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*. Springer, 586–594.

[6] Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. 2024. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences* 121, 2 (2024), e2304406120.

[7] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2301–2309.

[8] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proc. Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 4–14. https://doi.org/10.1145/3290605.3300234

[9] Stephen Casper, Tong Bu, Yuxiao Li, Jiawei Li, Kevin Zhang, Kaivalya Hariharan, and Dylan Hadfield-Menell. 2023. Red teaming deep neural networks with feature synthesis tools. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[10] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. 2019. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety* 28, 3 (2019), 231–237.

[11] C. Chen, J. Yuan, Y. Lu, Y. Liu, H. Su, S. Yuan, and S. Liu. 2020. OoDAnalyzer: Interactive Analysis of Out-of-Distribution Samples. *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. https://doi.org/10.1109/TVCG.2020.2973258

[12] Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. 2022. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *npj Digital Medicine* 5, 1 (Oct. 2022), 1–15.

[13] Furui Cheng, Yao Ming, and Huamin Qu. 2020. Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1438–1447.

[14] Christine M Cutillo, Karlie R Sharma, Luca Foschini, Shinjini Kundu, Maxine Mackintosh, Kenneth D Mandl, and MI in Healthcare Workshop Working Group Beck Tyler 1 Collier Elaine 1 Colvis Christine 1 Gersing Kenneth 1 Gordon Valery 1 Jensen Roxanne 8 Shabestari Behrouz 9 Southall Noel 1. 2020. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ digital medicine* 3, 1 (2020), 47.

[15] Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* (2017).

[16] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9185–9193.

[17] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.

[18] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. 2019. Adversarial attacks on medical machine learning. *Science* 363, 6433 (2019), 1287–1289.

[19] Carissa G. Fonseca, Michael Backhaus, David A. Bluemke, Randall D. Britten, Jae Do Chung, Brett R. Cowan, Ivo D. Dinov, J. Paul Finn, Peter J. Hunter, Alan H. Kadish, Daniel C. Lee, Joao A. C. Lima, Pau Medrano-Gracia, Kalyanam Shivkumar, Avan Suinesiaputra, Wenchao Tao, and Alistair A. Young. 2011. The Cardiac Atlas Project–an imaging database for computational modeling and statistical atlases of the heart. *Bioinformatics* 27, 16 (Aug. 2011), 2288–2295.

[20] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. Vice: Visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 531–535.

[21] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2021. AdViCE: Aggregated Visual Counterfactual Explanations for Machine Learning Model Validation. In *2021 IEEE Visualization Conference (VIS)*. IEEE, 31–35.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[23] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on

human factors in computing systems. 1–13.

[24] Jinbin Huang, Aditi Mishra, Bum Chul Kwon, and Chris Bryan. 2023. Concept-Explainer: Interactive Explanation for Deep Neural Networks from a Concept Perspective. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 831–841. https://doi.org/10.1109/TVCG.2022.3209384

[25] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. 2017. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284* (2017).

[26] Paul Jacob, Éloi Zablocki, Hedi Ben-Younes, Mickaël Chen, Patrick Pérez, and Matthieu Cord. 2022. STEEX: steering counterfactual explanations with semantics. In *European Conference on Computer Vision.* Springer, 387–403.

[27] Mark T Keane, Eoin M Kenny, Eoin Delaney, and Barry Smyth. 2021. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. *arXiv preprint arXiv:2103.01035* (2021).

[28] Mark T Keane and Barry Smyth. 2020. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In *Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings 28.* Springer, 163–178.

[29] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning.* PMLR, 2668–2677.

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).

[31] R. Krueger, J. Beyer, W.-D. Jang, N. W. Kim, A. Sokolov, P. K. Sorger, and H. Pfister. 2020. Facetto: Combining Unsupervised and Supervised Learning for Hierarchical Phenotype Analysis in Multi-Channel Image Data. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 227–237. https://doi.org/10.1109/TVCG.2019.2934547

[32] Bum Chul Kwon, Jungsoo Lee, Chaeyeon Chung, Nyoungwoo Lee, Ho-Jin Choi, and Jaegul Choo. 2022. DASH: Visual Analytics for Debiasing Image Classification via User-Driven Synthetic Data Augmentation. In *EuroVis 2022 - Short Papers*, Marco Agus, Wolfgang Aigner, and Thomas Hoellt (Eds.). The Eurographics Association. https://doi.org/10.2312/evs.20221099

[33] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. 2023. Selective explanations: Leveraging human input to align explainable ai. *arXiv preprint arXiv:2301.09656* (2023).

[34] Emily LaRosa and David Danks. 2018. Impacts on trust of healthcare AI. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* 210–215.

[35] Ho Hin Lee, Quan Liu, Shunxing Bao, Qi Yang, Xin Yu, Leon Y Cai, Thomas Z Li, Yuankai Huo, Xenofon Koutsoukos, and Bennett A Landman. 2023. Scaling up 3d kernels with bayesian frequency re-parameterization for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 632–641.

[36] Min Hun Lee and Chong Jun Chew. 2023. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–22.

[37] Xiaoxiao Li, Nicha C. Dvornek, Juntang Zhuang, Pamela Ventola, and James S. Duncan. 2018. Brain Biomarker Interpretation in ASD Using Deep Learning and fMRI. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018 (Lecture Notes in Computer Science)*, Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger (Eds.). Springer International Publishing, Cham, 206–214. https://doi.org/10.1007/978-3-030-00931-1_24

[38] M. Liu, S. Liu, H. Su, K. Cao, and J. Zhu. 2018. Analyzing the Noise Robustness of Deep Neural Networks. In *IEEE Conference on Visual Analytics Science and Technology (VAST'18).* 60–71. https://doi.org/10.1109/VAST.2018.8802509

[39] M. Liu, J. Shi, K. Cao, J. Zhu, and S. Liu. 2018. Analyzing the Training Processes of Deep Generative Models. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 77–87. https://doi.org/10.1109/TVCG.2017.2744938

[40] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu. 2017. Towards Better Analysis of Deep Convolutional Neural Networks. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 91–100. https://doi.org/10.1109/TVCG.2016.2598831

[41] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. 2023. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 22290–22300.

[42] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications* 15, 1 (2024), 654.

[43] Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. 2022. Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in artificial intelligence* 5 (2022), 825565.

[44] D Douglas Miller. 2019. The medical AI insurgency: what physicians must know about data to practice with intelligent machines. *NPJ digital medicine* 2, 1 (2019), 62.

[45] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[46] Thomas P Quinn, Manisha Senadeera, Stephan Jacobs, Simon Coghlan, and Vuong Le. 2021. Trust and medical AI: the challenges we face and the expertise needed to overcome them. *Journal of the American Medical Informatics Association* 28, 4 (2021), 890–894.

[47] Hanna Ragnarsdottir, Laura Manduchi, Holger Michel, Fabian Laumer, Sven Wellmann, Ece Ozkan, and Julia E. Vogt. 2022. Interpretable Prediction of Pulmonary Hypertension in Newborns Using Echocardiograms. In *Pattern Recognition (Lecture Notes in Computer Science)*, Björn Andres, Florian Bernard, Daniel Cremers, Simone Frintrop, Bastian Goldlücke, and Ivo Ihrke (Eds.). Springer International Publishing, Cham, 529–542. https://doi.org/10.1007/978-3-031-16788-1_32

[48] Pouria Rouzrokh, Bardia Khosravi, Sanaz Vahdati, Mana Moassefi, Shahriar Faghani, Elham Mahmoudi, Hamid Chalian, and Bradley J. Erickson. 2023. Machine Learning in Cardiovascular Imaging: A Scoping Review of Published Literature. *Current Radiology Reports* 11, 2 (Feb. 2023), 34–45. https://doi.org/10.1007/s40134-022-00407-8

[49] Ahmed Salih, Ilaria Boscolo Galazzo, Polyxeni Gkontra, Aaron Mark Lee, Karim Lekadir, Zahra Raisi-Estabragh, and Steffen E. Petersen. 2023. Explainable Artificial Intelligence and Cardiac Imaging: Toward More Interpretable Models. *Circulation. Cardiovascular Imaging* 16, 4 (April 2023), e014519.

[50] Ludwig Schallner, Johannes Rabold, Oliver Scholz, and Ute Schmid. 2020. Effect of superpixel aggregation on explanations in lime–a case study with biological data. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I.* Springer, 147–158.

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

[52] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450* (2016).

[53] Maxime Sermesant, Hervé Delingette, Hubert Cochet, Pierre Jaïs, and Nicholas Ayache. 2021. Applications of artificial intelligence in cardiovascular imaging. *Nature Reviews Cardiology* 18, 8 (Aug. 2021), 600–609. https://doi.org/10.1038/s41569-021-00527-2

[54] Peilun Shi, Jianing Qiu, Sai Mu Dalike Abaxi, Hao Wei, Frank P-W Lo, and Wu Yuan. 2023. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics* 13, 11 (2023), 1947.

[55] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[56] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[57] Fabian Sperrle, Mennatallah El-Assady, Grace Guo, Rita Borgo, D Horng Chau, Alex Endert, and Daniel Keim. 2021. A Survey of Human-Centered Evaluations in Human-Centered Machine Learning. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 543–568.

[58] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* 12, 3 (March 2015), e1001779.

[59] Animesh Tandon, Navina Mohan, Cory Jensen, Barbara EU Burkhardt, Vasu Gooty, Daniel A Castellanos, Paige L McKenzie, Riad Abou Zahr, Abhijit Bhattaru, Mubeena Abdulkarim, et al. 2021. Retraining convolutional neural networks for specialized cardiovascular imaging tasks: lessons from tetralogy of fallot. *Pediatric cardiology* 42, 3 (2021), 578–589.

[60] Kazuki Uehara, Masahiro Murakawa, Hirokazu Nosato, and Hidenori Sakanashi. 2020. Prototype-Based Interpretation of Pathological Image Analysis by Convolutional Neural Networks. In *Pattern Recognition (Lecture Notes in Computer Science)*, Shivakumara Palaiahnakote, Gabriella Sanniti di Baja, Liang Wang, and Wei Qi Yan (Eds.). Springer International Publishing, Cham, 640–652. https://doi.org/10.1007/978-3-030-41299-9_50

[61] Marly van Assen, Alexander C Razavi, Seamus P Whelton, and Carlo N De Cecco. 2023. Artificial intelligence in cardiac imaging: where we are and what we want. *European Heart Journal* 44, 7 (Feb. 2023), 541–543. https://doi.org/10.1093/eurheartj/ehac700

[62] Bas H. M. van der Velden, Hugo J. Kuijf, Kenneth G. A. Gilhuijs, and Max A. Viergever. 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis* 79 (July 2022), 102470.

[63] Simon Vandenhende, Dhruv Mahajan, Filip Radenovic, and Deepti Ghadiyaram. 2022. Making Heads or Tails: Towards Semantically Consistent Visual Counterfactuals. In *ECCV 2022*.

[64] Alfredo Vellido. 2020. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications* 32, 24 (2020), 18069–18083.

[65] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[66] J. Wang, L. Gou, H. Shen, and H. Yang. 2019. DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 288–298. https://doi.org/10.1109/TVCG.2018.2864504

[67] J. Wang, L. Gou, H. Yang, and H. Shen. 2018. GANViz: A Visual Analytics Approach to Understand the Adversarial Game. *IEEE Transactions on Visualization and Computer Graphics* 24, 6 (2018), 1905–1917. https://doi.org/10.1109/TVCG.2018.2816223

[68] J. Wang, L. Gou, W. Zhang, H. Yang, and H. Shen. 2019. DeepVID: Deep Visual Interpretation and Diagnosis for Image Classifiers via Knowledge Distillation. *IEEE Transactions on Visualization and Computer Graphics* 25, 6 (2019), 2168–2180. https://doi.org/10.1109/TVCG.2019.2903943

[69] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[70] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proc. Conference on Human Factors in Computing Systems (CHI'20)*. ACM, 1–13. https://doi.org/10.1145/3313831.3376807

[71] Guang Yang, Qinghao Ye, and Jun Xia. 2022. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion* 77 (2022), 29–52.

[72] Feiyang Yu, Alex Moehring, Oishi Banerjee, Tobias Salz, Nikhil Agarwal, and Pranav Rajpurkar. 2024. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine* (2024), 1–13.

[73] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6023–6032.

[74] Mehdi Zemni, Mickaël Chen, Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. 2023. OCTET: Object-aware Counterfactual Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15062–15071.

[75] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).

[76] Huijie Zhang, Jialu Dong, Cheng Lv, Yiming Lin, and Jinghan Bai. 2023. Visual analytics of potential dropout behavior patterns in online learning based on counterfactual explanation. *Journal of Visualization* 26, 3 (2023), 723–741.

[77] Jie Zhang and Zong-ming Zhang. 2023. Ethics and governance of trustworthy medical artificial intelligence. *BMC Medical Informatics and Decision Making* 23, 1 (2023), 7.

A EVALUATION BY SEGMENTATION

To validate the quality of the recombined images, we selected random pairs of source and target images and replaced different combinations of cardiac segments for each pair. We then used the *ukbb_cardiac* library to perform segmentation on the recombined images to ensure that the recombined morphological features can be identified (Section 5.1). Figure 6 shows one example of a pair of source and target images and the recombined images generated.
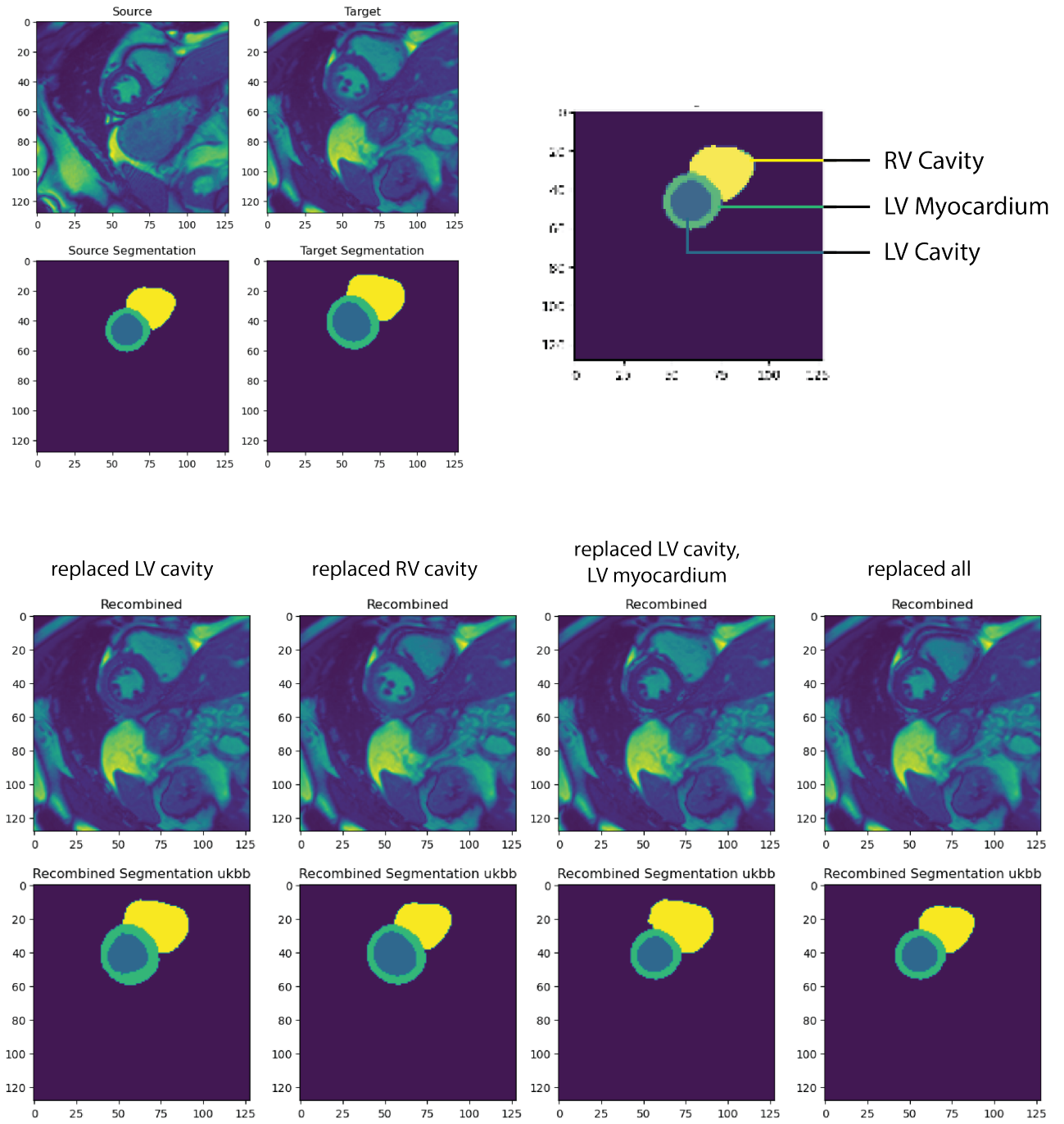
Figure 6: One pair of source and target MRIs and the resulting recombined images. As can be seen, the cardiac segments are re-identified by the segmentation algorithm with a high degree of fidelity. Note that due to space constraints, some infeasible combinations were excluded from this sample (e.g. replacing the LV myocardium alone since it is geometrically impossible to alter the LV myocardium without also altering the LV cavity enclosed within the myocardium).