Automatic scar segmentation on late gadolinium enhancement cardiovascular magnetic resonance images of patients with Tetralogy of Fallot

Laura Ros Freixedes
Automatic scar segmentation on late gadolinium enhancement cardiovascular magnetic resonance images of patients with Tetralogy of Fallot

Laura Ros Freixedes

BACHELOR'S DEGREE FINAL PROJECT / 2017

Universitat Pompeu Fabra
Barcelona

BACHELOR’S DEGREE IN BIOMEDICAL ENGINEERING

Supervisor(s):
Dr. Miguel Ángel González Ballester, Universitat Pompeu Fabra, Barcelona, Spain
Prof. Guang-Zhong Yang, Hamlyn Centre, Imperial College London, London, United Kingdom
## TABLE OF CONTENT

1. INTRODUCTION ............................................................................................................ 1
   1.1 Tetralogy of Fallot and late outcomes prediction: the motivation of the study ................................................................. 1
      1.1.1. Tetralogy of Fallot ........................................................................... 1
      1.1.2. Motivation of the study ................................................................. 4
      1.1.3. Objectives .................................................................................. 5
   1.2. Late gadolinium enhancement MRI principles ........................................ 6
      1.2.1. Magnetic resonance imaging principles ..................................... 6
      1.2.2. Late gadolinium enhancement MRI ........................................... 8
   1.3. State of the art on scar segmentation .................................................... 9

2. METHODS .................................................................................................................. 19
   2.1. Data acquisition and pre-processing .................................................... 19
   2.2. Scar segmentation algorithms ............................................................. 21
      2.2.1. Algorithm 1: gaussian mixture models with expectation maximization algorithm .................................................. 23
          2.2.1.1. Background ........................................................................ 23
          2.2.1.2. Algorithm implementation .............................................. 27
      2.2.2. Algorithm 2: gaussian mixture models with full-width-half-maximum ..................................................... 31
      2.2.3. Algorithm 3: gaussian mixture models with GrowCut .................................................. 32
          2.2.3.1. Background ........................................................................ 32
          2.2.3.2. Algorithm implementation .............................................. 34
      2.2.4. Algorithm 4: 3D automatic GrabCut .......................................... 35
          2.2.4.1. Background ........................................................................ 36
          2.2.4.2. Algorithm implementation .............................................. 39
      2.2.5. Algorithm 5: n-SD .................................................................... 41
   2.3. Statistical analysis ....................................................................................... 43
   2.4. Evaluation metrics .................................................................................... 43
      2.4.1. Dice similarity coefficient ......................................................... 43
      2.4.2. Total volume difference .............................................................. 44

3. RESULTS .................................................................................................................. 45
To Guillem, for always being there
Tetralogy of Fallot (TOF) is the most common form of cyanotic congenital heart disease, and one of the first to be successfully repaired by congenital heart surgeons. Even though the long-term outcome is favourable in most cases, several authors have drawn special attention to the unexpected occurrence of arrhythmias and sudden death originated by the alteration of the electrophysiological function of the heart due to fibrosis and the scars produced in the surgical repair. However, currently there is no reliable way to assess which patients are at risk. Taking on this challenge, the research group led by Babu-Narayan S. is trying to relate fibrosis patterns in MRI late gadolinium enhancement patient images with late outcome to determine the prognosis of those patients who underwent TOF surgical repair. To enable the latter, it is crucial to obtain accurate cardiac scar segmentations from the patient images. Current methodology is time consuming, not reliable, often require manual refinement and the turnover time would never be applicable to clinical workflow. This underlines the need for the development of a reliable algorithm capable of removing observer bias and with clinically acceptable accuracy.

With that objective, we explored several algorithms for automatic scar segmentation and compared its performance: Gaussian Mixture Models, Gaussian Mixture Models with Full-Width-at-Half-Maximum, Gaussian Mixture Models with GrowCut, 3D automatic GrabCut and n-SD.

The quantitative analysis and qualitative analysis of the results based on 13 patient datasets revealed Gaussian Mixture Models with Full-Width-at-half-Maximum as the most promising approach for automatic scar segmentation and discarded the use of the n-SD approach. Nevertheless, further work should be carried out with improved data and larger sample size to draw reliable conclusions.

**Keywords:** scar segmentation; late gadolinium enhancement MRI; Tetralogy of Fallot
Abbreviations used:

bSSFP: balanced steady state free precession
CMR: cardiovascular magnetic resonance
FWHM: full-width-at-half-maximum
EM: Expectation-Maximization
GMM: Gaussian mixture model
LGE: late gadolinium enhancement
LGE-MRI: late gadolinium enhancement magnetic resonance imaging
LV: left ventricle
MR: magnetic resonance
MRI: magnetic resonance imaging
ROI: region of interest
RV: right ventricle
RVOT: right ventricle outflow tract
SD: standard deviation
TOF: tetralogy of Fallot
VSD: ventricular septal defect
VT: ventricular tachycardia
1. INTRODUCTION

1.1 TETRALOGY OF FALLOT AND LATE OUTCOMES PREDICTION: THE MOTIVATION OF THE STUDY

1.1.1 TETRALLOLOGY OF FALLOT

Tetralogy of Fallot (TOF) is the most common form of cyanotic congenital heart disease. It features four heart defects (figure 1):

- Ventricular septal defect (VSD): a hole between the ventricles.
- Pulmonary stenosis: in most cases, obstruction of the outflow from the right ventricle (RV). It commonly coexists with obstructions at other sides and an abnormal pulmonary valve.
- Right ventricular hypertrophy: the thickening of the ventricle walls.
- A rightward deviation of the aorta with biventricular origin of the aortic valve leaflets, known as overriding aorta.

![Comparison of a normal and an abnormal heart](image)

Figure 1. Comparison of a normal and an abnormal heart. Left: Cross section of a normal heart structure and blood flow. Right: Heart defects and blood flow in patients with Tetralogy of Fallot. Source: National Heart, Lung and Blood Institute (NIH) website.

The entity tetralogy of Fallot should be regarded as a spectrum of diseases, all characterised by an anterocephalad deviation of the outlet ventricular septum and similar intracardiac anatomy but highly variable in terms of degree of right ventricular outflow tract (RVOT) obstruction, presence of associated abnormalities, such as
atrioventricular septal defect, and hence outcomes. In the most common form of the disease, the heart has normal segmental anatomical structure, the right ventricular outflow tract is patent at birth without the presence of major aortopulmonary collateral arteries to supply the lungs with blood and no other major associated abnormalities, exist [1], [2].

INCIDENCE
One percent of the population is born with a congenital heart disease. About 3-5% of all infants born with a congenital heart disease have tetralogy of Fallot, corresponding to one in 3600 or 0-28 every 1000 livebirths, with males and females being affected equally [2].

PATHOPHYSIOLOGY
If the ventricular septal defect is large and non-restrictive, there is equalization of the pressures in both ventricles. The direction and magnitude of the flow through the defect highly depends on the severity of the obstruction of the right ventricular outflow tract. When the obstruction is severe, a large right-to-left shunting is present at ventricular level, through the VSD, and this causes severe cyanosis. However, when the right ventricle outflow tract obstruction is mild, patients often have minimal cyanosis which can increase during the first weeks and months of life as pulmonary vascular resistance falls [1], [2].

DIAGNOSIS
In the current era, tetralogy of Fallot is frequently diagnosed during fetal life, which allows better planning of perinatal management, avoiding life-threatening cyanosis in the early new-born period. In the United Kingdom, it may also be found at early checks after birth and diagnosed by echocardiography. Some children will present with symptoms. Childhood presentation depends on the severity of the RVOT obstruction. Adult presentation is uncommon but possible [1], [2].
**MANAGEMENT**

TOF is repaired with primary open-heart surgery unless earlier palliative surgery is applied. Since the first reported intracardiac repair of tetralogy in 1955, the age of patients receiving primary corrective surgery has gradually decreased, with some units advocating surgery at diagnosis, even within the first few days of life. Most centres prefer to operate on children aged 3–9 months. Some centres continue to offer surgical palliation by construction of a systemic-to-pulmonary arterial shunt, balloon dilation, or placement of a stent in the right ventricular outflow, in neonates and young infants, thereby deferring intracardiac repair [2]. Some of the reasons include the low weight of the patients or to delay the repair to allow the pulmonary arteries to grow. However, late repair has been associated to higher late morbidity and mortality [3].

The surgical repair consists in widening the narrowed outflow tract to the pulmonary artery to improve the blood flow to the lungs, and repairing the VSD by covering the hole with a patch to stop the mixing of the oxygen-poor blood between the ventricles. The right ventricular reconstruction includes resection of muscle bundles beneath the pulmonary valve and depending on the anatomy and surgical choices may include disruption of the pulmonary valve annulus and patching of the right ventricular outflow tract and or pulmonary arteries to improve their size [1]. These procedures leave scar tissue around the operated areas due to surgical resection, incision, patching, suturing, or vent insertion in the left ventricle to de-air the heart [4],[5]. This scarred tissue can affect the electrophysiological function of the heart and provides a substrate for arrhythmia. It is easy to envisage how arrhythmia can be originated in the presence of fibrosis in the myocardium, with localized slowing of conduction favouring the development of reentry circuits.

Although advances in cardiothoracic surgery have greatly improved the prognosis for patients who undergo TOF repair and the long-term outcome is favourable in most cases, several authors have drawn special attention to the unexpected occurrence of sudden death from cardiac causes both early and late after apparently successful reparative surgery [3], [6], [7]. Some studies describe ventricular scars as a source of ventricular tachycardia (VT), causing late morbidity and sudden death [8]–[11]. In a clinicopathologic study performed on autopsy specimens of 6 patients who died suddenly after the repair of TOF, arrhythmias were supported as the cause of death [12]. In four of the postoperative patients, sustained VT originated from reentry circuits in the
right ventricular outflow tract was observed, and the histologic examination showed extensive fibrosis of the right ventricular myocardium in the ventriculotomy site, septum and outflow tract.

1.1.2. MOTIVATION OF THE STUDY

Arrhythmias and sudden death can be potentially prevented using catheter ablation. Electrophysiological studies and electroanatomic mappings can be performed to characterize the location of potential critical isthmuses for reentry circuits. An electrophysiological study is performed by positioning invasive electrode catheters in the RV after patient sedation and performing some programmed stimulations to induce ventricular tachycardia and to obtain the QRS morphology. RV mappings are performed with a catheter with recording electrodes to construct a 3D voltage map using an electroanatomic mapping system as in the technique presented in [8], to guide the ablation procedure. Nevertheless, this technique presents some limitations: it cannot prevent sudden death in patients with non-inducible ventricular tachycardia, the size of fibrotic regions identified is not clearly defined and it may not identify narrow bands of fibrosis that also can serve as conduction block [8]. Therefore, there is a need for finding a reliable way for assessing fibrosis and predicting which patients are at risk.

With that aim, Babu-Narayan et al. [4] studied the distribution of cardiac fibrosis in a cohort of TOF patients using LGE-MRI. This is an imaging technique commonly used to detect fibrosis. The sensitivity of this technique can be outlined by the fact that the perioperative apical vent insertion site fibrosis in patient images could even be seen decades after repair surgery.

In [4], the authors reported the ubiquitous occurrence but varying locations and extent of right ventricular myocardial fibrosis in adults late after TOF repair. The locations where the fibrotic tissue was found agreed with the areas where the surgical procedures had been carried out, but suggestive fibrotic tissue was also seen in sites remote from surgical instrumentation, in trabecular and endocardial sites, which may be more vulnerable to ischaemic insult. In this study, it was suggested that these fibrotic areas could be originated by ventricular dilatation and hypertrophy resulting from pulmonary regurgitation or stenosis, which lead to stretching, arrhythmia generation, and adverse
RV remodelling, particularly in vulnerable border zones between normal and fibrotic myocardium. They also found that the extent of right ventricular fibrosis related was associated with adverse clinical markers, including right ventricular dysfunction, objective exercise intolerance, and neurohormonal activation, and with the incidence of clinical arrhythmias.

With these data at hand, the research team from Royal Brompton Hospital led by Sonya Babu-Narayan believed that an extended study on the fibrotic tissue distribution in patients with TOF could in the future contribute to risk stratification and decision making on the timing of arrhythmia intervention or hemodynamic surgery such as pulmonary valve replacement. That is why, currently, they are focusing their efforts on finding fibrosis patterns from the LGE-MRI patient images, to relate them to late outcome and be able to determine the prognosis of those patients that underwent TOF surgical repair. Integrating scar imaging into the diagnostic and treatment procedures electrophysiologists performed invasively in the laboratory could guide the procedures allowing them to be performed more easily, quickly and successfully or altogether avoided if the prognostication from the MRI data in some cases was sufficient.

To enable the latter, it is crucial to obtain clinically acceptable cardiac scar segmentations from the patient images. Currently, these segmentations are performed semi-automatically, using a commercial software based on thresholding (Mimics, Materialise). They take from 2 to 3 hours to segment both the cavities and the scar after a learning curve where they can take longer. Variabilities in time depend on the amount of scar in each patient case. A small amount takes less time, whereas a large amount takes longer. Therefore, the process is time consuming. Moreover, the accuracy of the method is not always reliable, manual refinement is required and the turnover time would never be applicable to clinical workflow. This underlines the need for the development of a reliable algorithm capable of removing observer bias and with clinically acceptable accuracy, as it is key for clinically useful scar quantification to determine patient prognosis.
1.1.3. OBJECTIVES

The purpose of this study was to understand and familiarize with the state-of-the-art techniques used in image segmentation and use them to develop an automatic algorithm capable of segmenting the scars in the LGE-MRI images of patients with Tetralogy of Fallot, with proper reproducibility and reliability.

1.2. LATE GADOLINIUM ENHANCEMENT MRI PRINCIPLES

1.2.1 MAGNETIC RESONANCE IMAGING PRINCIPLES

Magnetic resonance imaging (MRI) is a relatively new imaging modality. In essence, MRI measures a magnetic property of tissue [13].

All atomic nuclei consist of protons and neutrons, with a net positive charge. Certain atomic nuclei possess a property derived from quantum mechanics known as spin, dependent on the number of protons and neutrons. The ‘spin’ can be thought of as angular momentum, which is the momentum associated with the spinning motion of a rotating body. Nevertheless, the nucleus itself does not spin in the classical meaning but by virtue of its constituent parts (moving charges) induces a magnetic moment, generating a local magnetic field - according to Faraday’s law - with north and south poles, like a bar magnet. Some nuclei with an even number of nucleons have zero overall spin, while all nuclei with an odd number of nucleons will have net spin, such as the hydrogen nucleus, $^1$H. Since hydrogen is abundantly available in the human body, MRI focuses on the visualization of hydrogen-containing tissues (muscles, brain, kidney, cerebrospinal fluid, edema, fat, bone marrow, etc.) [14], [15].

Since nuclei normally have random motion, the components of magnetic moment cancel out and there is no overall magnetic field associated with matter. However, with the application of a longitudinal, strong, external magnetic field ($B_0$), non-zero spin nuclei try to align themselves parallel to the external field, as this is the lowest energy state available to them. Although a bar magnet would orientate completely parallel or antiparallel to the field, as the nuclei have an angular momentum due to its rotation, they will precess, like a gyroscope, around the $B_0$ axis. The velocity of rotation around the field direction is described by the Larmor frequency. Larmor frequency is proportional to the field strength by the formula: $\omega_0 = \gamma B_0$, where $\omega_0$ is the angular
frequency, $\gamma$ is the gyromagnetic ratio ($\gamma = 42.6 \text{ MHz/Tesla}$ for hydrogen) and $B_0$ is the strength of the field applied. Since the magnetic moments of the nuclei are no longer randomly aligned, they add together to form a net magnetic field from the tissue parallel to the external field, but much smaller in magnitude [14], [15]. Nevertheless, the technique of cardiac MRI can detect this magnetisation, as explained as follows.

Nuclei that possess spin can be excited within the static magnetic field, $B_0$, by application of a second radiofrequency (RF) magnetic field $B_{\text{RF}}$, perpendicular to $B_0$ and produced by a large superconducting magnet. The RF energy is usually applied as a train of short pulses separated by a constant interval (defined by the repetition time, TR), each lasting microseconds. Electromagnetic energy is transmitted from a coil to the nuclei, exciting them to a higher energy state. The net magnetic moment in this higher energy state moves out of the direction of $B$. In the absence of continued RF pulsation, the system will return to thermal equilibrium. This process is called relaxation. In this phase, the nuclei, liberate the absorbed RF energy in the form of an electromagnetic wave while they transition back to lower energy levels (i.e. parallel to $B_0$). This electromagnetic wave can be detected by a suitably tuned coil of wire, amplified and displayed as the free-induction decay (FID) [14], [15].

There are two types of relaxation, longitudinal and transverse relaxation, and these are described by the time constants, T1 and T2, respectively. The FID has mixed T1 and T2 weighting [16].

T1 is also known as spin-lattice relaxation, whereby the lattice is the surrounding nucleus environment. As relaxation occurs, energy is dissipated into the lattice. This relaxation results in a small increase of the longitudinal component of the magnetization vector (LM) [17]. T1 is the length of time taken for the system to return 63% toward thermal equilibrium following an RF pulse as an exponential function of time. T1 can be manipulated by varying the times between RF pulses, the repetition time (TR). Water and cerebrospinal fluid (CSF) have long T1 values, and thus they appear dark on T1-weighted images, while fat has a short T1 value and appears bright on T1-weighted images [14].

Relaxation processes may also redistribute energy among the nuclei within a spin system, without the whole spin system losing energy. Thus, when a RF pulse is applied,
nuclei align predominantly along the axis of the applied energy. On relaxation, there is dephasing of nuclei orientations as energy is transferred between the nuclei and there is reduction in the resultant field direction, with a more random arrangement of alignments. This is T2, termed transverse relaxation, because it is a measure of how fast the spins exchange energy in the $x$-$y$ plane. T2 is also known as spin-spin relaxation [14].

Different tissues return to their lower energy states at different rates due to their different content in $^1$H (ex: free water vs $^1$H bound to tissue) and so tissues can be differentiated. The received signal can be reconstructed using a mathematical technique called a Fourier transformation and powerful computing to give an MRI image.

1.2.2. LATE GADOLINIUM ENHANCEMENT MRI

Late Gadolinium Enhancement (LGE) MRI is an imaging technique used to test the viability of the myocardium, which refers to the myocardium that is living and contributing to the systolic ejection of the blood. Therefore, this technique is very useful for scar detection, as recommended by many scientific societies. It is based on the differences in volume distribution of Gadolinium (Gd) between normal tissue and that of fibrosis or necrosis due to differences in extracellular space along the myocardium.

Gadolinium is an element which is toxic in its unbound state. Therefore, it is administered bound to large molecules called contrast agents (CA) as a protection. After being injected intravenously, it is mostly excreted by the kidneys, having a half-life of two hours and being almost completely cleared from the mainstream after 24 hours.

The gadolinium contrast medium has a molecular size that allows it to distribute in the extracellular space without entering the myocardial cells in normal conditions.

Nevertheless, under certain pathologies, this extracellular space is increased, and more contrast agent is accumulated in this area. This increased amount of gadolinium can be observed with the T1-weighted imaging, between 10 and 30 minutes after the contrast administration on the patient. This can happen when the myocardial cell membrane is disrupted, under cell necrosis and lysis conditions or due to scar production, as the deposits of collagen make the extracellular space increase.
It is crucial to choose the correct inversion time (TI) to obtain images with enough diagnostic quality. The inversion time is a parameter used in MRI to null the signal generated for specific tissues. The optimal TI is that which results in suppression of normal myocardium (with low signal and, thus, dark image) along with a brighter signal from the myocardial cavity and a very bright signal in the scar tissue. An optimal TI should enable the examiner to differentiate endo-, mid- and epicardial LGE [18].

### 1.3. STATE OF THE ART ON SCAR SEGMENTATION

Scar segmentation consists in the classification of image pixels between scar and non-scar tissue. During the last two decades, many methods have been explored. Although many of them show a good performance, the optimal method remains unclear. In this section, we will make an overview of all the methods that have been developed.

Fixed-thresholding methods were the first to appear and they are the most extensively studied due to their easy implementation. The first one was developed by Kim. R.J. in 1999 [19], and it was applied to infarct segmentation in the left ventricle. This method was based on the standard deviation (SD) of the image pixel intensities. It consisted on thresholding the images to a fixed number of SD from the mean intensity of the healthy myocardium. The number of SD was set to 2.

In 2004, Amado L.C. [20] introduced the full-width-at-half-maximum (FWHM) algorithm, where half of the maximum intensity within a user-selected hyper-enhanced region was established as the fixed intensity threshold. Using this threshold, a region-growing process was employed from user-selected seeds within infarcted regions to obtain the scar segmentations.

Since then, many studies have used SD and FWHM methods [21]–[27]. Although 2-SD was initially advocated by official guidelines, 3-, 4-, 5- and 6- SD have also been used and analysed, with the region in which the statistical measures are computed varying from the nulled myocardium to the blood pool [21]–[24], [26], [27].
Nevertheless, these methods resulted in different mean LGE volumes, which is not suitable for research and prognostic definition as quantification reproducibility is required. As shown in the study of Flett et al., 2011 [23], when comparing the results of 2-,3-,4-,5-,6-SD and FWHM against the manual scar delineation, they saw that the 2-SD technique could double the LGE volume compared with the manual, FWHM and 6- or 5-SD techniques. They also showed that although one might assume that the manual segmentation could be used as the gold-standard, it lacked reproducibility as there was inter-observer and intra-observer variability. As a conclusion of the study, they reported FWHM as the most reproducible technique for scar segmentation in hearts with different cardiac conditions. However, they also mentioned that in other studies 6-SD was reported as a superior technique due to the FWHM quantification uncorrelation with the visual assessment.

In Karim et al, 2013 [28], it was suggested that using a fixed model (SD and FWHM) was not suitable for scar segmentation in the left atrium and the left ventricle, despite all the studies utilising them. The reasons given were that a fixed model cannot handle all the different variabilities encountered in patient images due to the varied internal (size, distribution and heterogeneity of scar) and varied external (resolution, image noise, inversion time, surface coil intensity variation) situations. They also justified it by citing the existence of a study where it was shown that the threshold had to be re-adjusted on various slices to obtain a suitable segmentation [26].

That might be the reason why besides fixed model methods, more complex methods, some incorporating machine learning techniques, were developed to perform scar segmentation during the last years. These new methods were mainly gathered into two challenges organised by important congresses in the medical image analysis field, ISBI 2012 and MICCAI 2012.

These challenges invited participants to compare their own methods using common data so that to be able to benchmark them. In 2013, the results of the evaluation of seven different algorithms for scar segmentation in the left atrium were published [28]. And in 2016, the results of the evaluation of five methods for scar segmentation in the left ventricle [29] were published.
However, in this study, we were interested in scar segmentation in the right ventricle, as it is the cavity where the TOF patients present scars. Nevertheless, no scar segmentation studies had been performed in this area. This task presents added difficulties as the RV segmentation remains challenging due to its variable shape and its ill-defined borders [30], apart from the presence of trabeculae with same grey-levels as the wall myocardium and the papillary muscles [31].

Nevertheless, knowing the methods that have been used in other heart cavities and the results of their evaluation is advisable, as it can give an intuition of which ones are more prone to work in our area of interest. That is why in the next section, we will summarise these methods.

**LEFT ATRIUM FIBROSION AND SCAR SEGMENTATION CHALLENGE (2013)**

This challenge is especially interesting for our task as the thickness of the right ventricle myocardium wall is more similar to the left atrium than to the left ventricle. However, this cavity differs from the right ventricle in the fact that it doesn’t contain trabeculae.

In this challenge, eight algorithms were presented. All the participants were given the atrium endocardium segmentations. Previous methods existed based on automatic threshold computation, clustering and graph-cuts and some of the new methods offered small variations of them. Next, we will briefly explain these methods.

Bai et al. developed an algorithm based on hysteresis thresholding [28]. Hysteresis thresholding consists in the selection of two thresholds for image segmentation. In this method, they thresholded the images after computing a scar probability map based on pixel intensity and distance from the atrium endocardium. They classified the pixels above the higher threshold as scar and also the ones connected to them that were above the lower threshold. That way they ensured coherence in the result.

Hennemuth et al. used region growing, which is an appropriate choice to find groups of connected pixels with similar intensity, and combined it with Gaussian Mixture Models (GMM) to find the seeds for the region growing segmentation [28], [32]. A Gaussian Mixture Model is a parametric probability density function represented as a weighted sum of Gaussian component densities. The different Gaussian components act together to model the overall feature density. Here, GMMs are trained to model each class (in
this case, scar, left atrium and background) and to emit the likelihood of each pixel belonging to each class. Finally, those likelihoods are used for the final decision, by thresholding them.

Lu et al. proposed a modified version of Graph-Cuts and fuzzy c-means [33]. Graph-cuts is a segmentation technique used for foreground/background subtraction in which the goal is to divide the image into exactly two regions whose appearance models are known beforehand. In this case we would have the models for scar and non-scar tissue. An image is represented as an undirected graph, which is a network of nodes connected by edges. In this image representation, the image pixels become the nodes and the edges represent the connection between the pixels, that is to say, the connection between adjacent pixels. Each edge is assigned a non-negative weight cost defined by intensity dissimilarity between adjacent pixels. Suppose there are two extra special vertices, with one being the ‘source’ (called s) and the other being the ‘sink’ (called t), which represent the terminal nodes for each label (i.e. foreground/background, which in our case would be scar/background). Each pixel node is also connected to both s and t nodes by a weighted edge in which the weight is the probability of the node for the label, given the known appearance models. A cut that separates the graph into two subgraphs S and T, is known as an ‘s-t cut’ and it is normally defined as the sum of weight edges between the two sets, which takes into account both the pixel adjacency and the label probability weights. The goal of Graph-Cuts is to find the minimum s-t cut (which may not be unique, because many cuts can have the same cost). This cut gives a partitioning that corresponds to the segmentation.

In this work, they used the c-means fuzzy clustering to compute the s,t edge weights.

Gao et al. [34] firstly used active contours to obtain the epicardial atrial boundary. Active contouring consists on modelling a contour by using a spline (i.e. a free form curve) allowing it to grow flexibly with additional constraints placed by the image and captured by an energy function which defines the final contour shape by energy minimization. After this, they obtained the atrial wall by subtracting the endocardium to the epicardium. Once they had the atrial wall, they modelled the tissue intensities using two GMMs (one for scar and one for non-scar) with the Expectation-Maximization (EM) algorithm to estimate the parameters of the Gaussian components.
Peters et al. applied simple thresholding after manual wall delineation [28]. The same threshold was applied to the whole 3D volume and the delineation of the wall allowed the avoidance of close enhanced cardiac structures that are not scar. However, this method had the disadvantage that it only considered the intensity information and missed the relationship between pixels, by not taking into account whether they were contiguous or not.

Karim et al. used graph-cuts with GMM [35]. First of all, they developed two Gaussian mixture models and used the EM-algorithm to find the optimal parameters of the Gaussian mixture components. The Gaussian mixture models where obtained using different approaches. To train the scar tissue model, they used the leave-one-out approach, that is to say, they used all the patient cases except the one where they wanted to perform the segmentation to train the model. Then they used a Gaussian distribution of the scar intensities in the training images to obtain the model, representing these scars intensities as a ratio of scar to average blood-pool.

To obtain the non-scar model, they just used the target image and used a Gaussian mixture with EM-algorithm to set the component parameters. They kept the number of components variable according to the configuration that fitted best the image intensity distribution. Only the regions that were 3 mm inside and outside the endocardium segmentation where used for this purpose.

From these models, they could obtain the likelihood of each pixel for each label (scar and non-scar), as in the algorithm of Lu et al., and compute the standard Graph-Cuts to obtain the final segmentation.

Cates et al. [26] used a method based on thresholding after manual wall delineation. The threshold was selected according to the intensity histograms of the pixels within the wall. This histograms were expected to have two modules: one for enhanced region and one for non-enhanced. The threshold was set as +2-4 SD off the mean of the lower mode of the histogram. This method had the inconvenient that the threshold needed to be reestimated for some of the slices to improve its performance.

Finally, Perry et al. used unsupervised learning using k-means clustering [28]. Given the number of clusters $k$ and a set of data $n$, this algorithm works iteratively finding the
points \( k \) - called centres - among the data \( n \) that minimize the mean squared distance (or any other predefined distance or error metric) from each data point to the nearest centre. To use these clustering algorithm, they needed to test which was the optimal number of clusters and which was the best feature vector for comparing pixels. These were found by empirical evaluation. The optimal number of clusters was finally set as 4 with normalised voxel intensity was used as the feature vector.

To obtain the final segmentation, they selected the resulting cluster with the highest mean intensity as the scar cluster.

In this evaluation framework, they compared the segmentations obtained with each of the methods to the ones obtained with the n-SD and FWHM methods by using several quantitative metrics that showed how similar the results were to the pseudo ground-truth data. The ground-truth data was obtained by the merging of two manual segmentations performed slide by slide by different observers with same guidelines.

As shown, some methods made assumptions about the intensity distribution of enhanced pixels within atrial myocardium. Modelling the distribution with a statistical distribution such as a Gaussian is a common technique. Prior to modelling, some normalised the atrial myocardium intensities to the easily observable atrial blood pool by taking its average. The evaluation of the algorithms suggested that modelling enhancement and normalising it was not sufficient to obtain good results and that other modes of information should be considered, as some of the presented algorithms did.

The comparison of the results showed that all the methods outperformed the FWHM and n-SD methods. This suggested that a fixed model for scar was not a viable solution and that improvements can be made. This was also supported by the institutions that implemented methods with simple thresholding, as they had to change the threshold for every slice to achieve improvements over the fixed methods.

The results also showed that the segmentation of LA myocardial wall is an important step before segmentation scar to avoid artefact segmentation. The results obtained from the methods that used manual contouring of the LA revealed that a good wall LA segmentation could not only counteract to a great extent the unwilling segmentation of the enhanced aortic wall but also overall improve LGE CMR segmentation.
In 2016, the results of the evaluation of several methods for infarct segmentation in the left ventricle [29] were published. These algorithms were also based on the fact that myocardial infarction is enhanced in LGE CMR images.

This evaluation was done as there is a need of obtaining reliable, fast, reproducible and accurate segmentations of the infarcted regions so that to be able to develop scar maps that could improve the guidance of the ablation procedures in ventricular tachycardia, apart from helping in the interpretation of infarct volume estimation, which is becoming more and more relevant.

In this task, the myocardium segmentation walls were provided. This made the segmentation of non-infarction areas more improbable.

Five new different methods were introduced, which were added to the existing previous methods based on n-SD, FWHM, clustering, Otsu’s thresholding, mixture models and Graph-Cuts.

In the one proposed by Lara et al. [36], a great variety of image processing techniques were used: Otsu’s thresholding, support vector machines (SVM) and level set were used for scar labelling.

Otsu’s Method is a thresholding algorithm that is used to find the threshold $\tau$ that minimizes the within-class variance, which is defined as the weighted sum of variances of two groups of pixels.

SVM is a supervised learning method that can be used for classification.

Level-sets is a method in which given an initial curve or surface, it can move and evolve within the region to be segmented handling topological merging and breaking naturally, working in any number of space dimensions, and without requiring the moving surface be written as a function.

In this proposed method for infarct segmentation, the Otsu’s method was firstly applied in order to compute the optimum threshold to classify scar and healthy tissue. Afterwards, a pixel-connectivity analysis was also performed to overcome the thresholding limitations, such as the ones regarding healthy and scar tissue having overlapping intensities. From the two groups of pixels obtained, several features were extracted which were used for the SVM classification. It must be taken into account that
apart from SVM, other classifiers were tried (linear Bayesian discriminant, k-nearest neighbours and linear perceptron), but SVM was chosen due to its better performance on the training data. After that, the level-set method was used to refine the results, taking the contours of SVM as the initialization curve, and having a limited search area defined by the first step of the method.

Albà et al. [37] presented a 2D segmentation method based on region growing with automatic seed selection. The seed selection was automated and repeated for each slice. A minimum of two seeds for both scar and healthy tissue were selected for each slice, by using a threshold. This threshold was set as the mean of the myocardium intensity plus the standard deviation of the intensities. The two brightest and largest regions above the threshold were selected as the scar seeds, and the two darkest and largest regions below the threshold were selected as the myocardium seeds. After that, region growing was applied. The limitation of this approach was that a maximum of two disconnected regions per class could be obtained in each slice. After the region growing step, some image processing steps were performed to label unlabelled areas and fill holes and gaps.

Karimaghaloo et al. [38] used a variant of SVM known as relevance vector machines (RVM) [39] to prelabel the data to apply Graph-Cuts with conditional random fields (CRF) instead of Markov Random Fields (MRF) to obtain the final segmentation. Therefore, the posterior \( p(Y|X) \) likelihood was estimated by learning a direct map from observations to the class labels in the training images, instead of using Gaussian distributions.

Graph-Cuts were applied two times to perform the segmentation once the image were prelabelled using RVM. In the first time, image intensity information was used. In the second time, the output of the first Graph-Cut attempt was used to obtain the final segmentation. However, more information was used to perform the segmentation. A two-dimensional histogram encoding the distribution of image brightness values in the neighbourhood of a particular reference point was constructed. This is the spin image, which encoded local information around infarct candidates. Then, besides voxel intensity, these spin image features were also used for CRF.

Hennemuth et al. [40] modelled the myocardium intensity of the unseen images using different statistical distribution models (Gaussian, Rician and Rician-Gaussian) and
chose the one that gave less least mean fitting error. The EM algorithm was used to find the optimal fit. Once the mixture model was obtained, a threshold based on the highest mean between the two mixture classes was applied to determine the seeds of the watershed transformation which was used to finally label the scar. Watershed is a classical image segmentation technique where the gradient image is considered as a topographic surface. The high-low-high intensity gradient of scars creates basins in the image. Once points are located inside each basin they can be segmented by following paths of decreasing altitudes on the topography of the gradient descent. Afterward, a connected-component analysis was used to remove small noisy structures.

Karim et al. [35] also presented a method based on Gaussian-mixture models and Graph-Cuts in this challenge. This method was very similar to the one presented for left atrium scar segmentation. The only change was that the number of Gaussian mixture components that was used to model the healthy myocardium was set to 3, instead of keeping it variable.

In this challenge, the pseudo-ground-truth segmentations were obtained by merging volumetric segmentations from three separate observers with same guidelines to perform them. The images belonged to both human and pig hearts.

The results of this study were analysed objectively with quantitative measures and showed that the new algorithms presented a better overlap with the ground-truth than the n-SD fixed-thresholding methods, but found FWHM as an exception. Moreover, it was shown that the algorithms based on region growing and morphology, SVR with conditional random fields and support vector machines with level sets had a better overlapping than the ones based on Graph-Cuts with EM-algorithm and EM-algorithm with watershed transformation. At the same time, it was outstood the use of image post-processing steps in the algorithms based on support vector machines with level sets and region growing and morphology for error correction, as they avoided the inclusion of false positive regions.

Overall, both challenges claimed that fixed threshold methods are not recommended for scar segmentation as thresholds usually require user readjustments due to the fact that the contrast levels on the images are directly dependent on the inversion time selected in
the LGE CMR. A recent review on left atrium scar detection supported this idea as well [41].

However, in the literature computer-aided classification of enhanced tissue in LGE CMR is still an open question, and no algorithm has been deemed clearly better than others. In fact, fixed-thresholding methods are still quite used in clinical research and studies based on them are still being published. One of the possible reasons could be that these methods are easier for clinicians to use, as they can be easily implemented with commercial softwares.
2. METHODS

2.1 DATA ACQUISITION AND PRE-PROCESSING

IMAGE ACQUISITION PROTOCOL

The CMR image acquisition followed the protocol described in detail in previous studies conducted from the same centre [42]. Acquisitions were typically performed in diastole, the exact timing being determined from four chamber cine acquisitions reviewed to note the still time in the cardiac cycle (onset and end). 3D LGE imaging was carried out 5-to-10 minutes after the gadolinium administration. It was performed during free-breathing using diaphragmatic navigators to restrict the respiratory motion to a 5mm acceptance window around the end expiratory pause position. The inversion time (TI), the parameter that nulls the myocardium, was computed dynamically with an algorithm [43]. The difference between dynamic TI to image quality should not be discernible.

DATA SET SELECTION AND PRE-PROCESSING

An initial set of 37 TOF patient cases was provided. All patient files were anonymized and a number from 001 to 037 was assigned to identify each patient. All of them had undergone protocolised cardiovascular magnetic resonance with 3D balanced steady state free precession (3D bSSFP) and 3D LGE CMR. Both sequences of each patient were provided, as well as the semi-automatic segmentation masks of the right ventricle cavity and the scar files. All files were in NIfTI format and compressed in gzip.

The segmentations were obtained by a cardiologist with expertise in LGE MRI, using the commercial software Mimics, from Materialise, based on manual thresholding. The scar segmentation was performed manually or semi-automatically with manual corrections on the LGE sequences. The right ventricular cavity was defined as right ventricular blood pool bounded by the tricuspid and pulmonary valves. However, this segmentation was obtained semi-automatically from the bSSFP sequence and aligned to the LGE sequence, as it was easier to distinguish the myocardial wall from the blood
pool in this sequence type. Alignment between bSSFP and LGE sequences was assumed.

LGE and bSSFP sequences had different voxel size, so that the same happened to the RV cavity and scar segmentations. In order to be able to work with them properly, a reslicing operation was performed so that the voxel size become the same for all the files. This operation was carried out using the function `reslice_nii` of the toolbox `Tools for NIfTI and ANALYZE image` for Matlab that can be downloaded in the file exchange website of Mathworks [44]. Notice that different interpolation methods were required for reslicing the different files. For both MRI sequences, the trilinear interpolation method was used. However, this method could not be used on segmentation files as they contain just logical values, 0 for background and 1 for foreground, and using trilinear interpolation would have given new pixels with intermediate intensity values, which would have made the scar size increase. That is why the nearest neighbour method was used instead.

In some patient cases, the scar segmentations for the same patient were provided in different files, according to their location. As the purpose of this work was not to differentiate them and there was no need to label them separately, all the scar files were merged into one file by reading the `img` field of each `.nii` file, summing them and setting all the pixels with label above 0 as 1.

Out of the total 37 patient cases, 13 were used for algorithm evaluation. Exclusion criteria included low image contrast, non-reliable ground-truth, RV wall labelling instead of cavity labelling, very bad RV labelling or image artefacts. Image artefacts were partly caused by surgical material left in the body after TOF repair, such as sternum wires, which prevented the correct visualisation of the RV cavity and evaluation of its scar. Examples of excluded images can be found in the supplementary figures 1-3.

Imaging quality and resolution of the 13 datasets can be found in table 1. Regarding image quality, images were scored by a professional in CMR imaging into three possible categories: 1-good, 2-average and 3-poor.
Table 1. Image quality score for each patient case. The sign “-” indicates that the score was not provided

<table>
<thead>
<tr>
<th>Patient number</th>
<th>Imaging score</th>
<th>Acquired resolution (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>1</td>
<td>1.48438 x 1.48438 x 2</td>
</tr>
<tr>
<td>002</td>
<td>-</td>
<td>0.7422 x 0.7422 x 2</td>
</tr>
<tr>
<td>003</td>
<td>1</td>
<td>0.7422 x 0.7422 x 2</td>
</tr>
<tr>
<td>004</td>
<td>1</td>
<td>0.7422 x 0.7422 x 2</td>
</tr>
<tr>
<td>005</td>
<td>-</td>
<td>0.7422 x 0.7422 x 2</td>
</tr>
<tr>
<td>006</td>
<td>1</td>
<td>0.7422 x 0.7422 x 2</td>
</tr>
<tr>
<td>007</td>
<td>1</td>
<td>0.7422 x 0.7422 x 2</td>
</tr>
<tr>
<td>008</td>
<td>1</td>
<td>0.7422 x 0.7422 x 2</td>
</tr>
<tr>
<td>009</td>
<td>2</td>
<td>0.742188 x 0.742188 x 2</td>
</tr>
<tr>
<td>010</td>
<td>1</td>
<td>0.742188 x 0.742188 x 2</td>
</tr>
<tr>
<td>011</td>
<td>1</td>
<td>0.742188 x 0.742188 x 2</td>
</tr>
<tr>
<td>012</td>
<td>1</td>
<td>0.742188 x 0.742188 x 2</td>
</tr>
<tr>
<td>013</td>
<td>1</td>
<td>1.4844 x 1.4844 x 2</td>
</tr>
</tbody>
</table>

2.2. SCAR SEGMENTATION ALGORITHMS

Five algorithms were implemented for the data scar segmentation for comparison, based on Gaussian Mixture Models, GrowCut, full-width-at-half-maximum, Graph-Cuts and n-SD. Before describing the algorithm implementation, we will firstly go into the basics of each of them, so that to properly understand the limitations that they offer and the results.

Before starting explaining the algorithms, it is important to remark how the definition region of interest for scar searching and the data normalisation were performed, as it is used in all the algorithms.

REGION OF INTEREST DEFINITION

The first step in scar segmentation consists on defining the region of interest (ROI) in which to look for the scar in the testing image. In this case, we considered that given the pathophysiology of the Tetralogy of Fallot, all the right ventricle should be considered as a region of interest, including the myocardium and the cavity. The reason seems clear considering that scars can either be present in the myocardium wall and in the trabeculae, as explained in the introduction section. However, right ventricle epicardium segmentations were not provided. Knowing that actual epicardium segmentation is challenging due to the small thickness of the wall and its ill-defined borders in LGE.
CMR images, we decided to use the dilated right ventricle cavity segmentation masks to define the patient ROIs.

The dilation process was performed using the *disk* morphological Matlab structuring element. This is a 2D circular structuring element. This choice was performed considering the RV morphology. 3D structuring elements were avoided for speed purposes. Instead, 3D dilation was approximated by performing 2D-slice-by-slice dilations in the 3 directions of the volumetric mask, summing the results of each of them and converting the final mask to logical type. The size of the dilation was independently set for each patient case to make sure that ROIs covered the RV wall and therefore, the scars.

We discarded performing any other pre-processing morphological operations before dilation, such as opening and closing, useful to remove potential mislabelled pixels in the RV masks, as the RV masks presented a lot of unconnected regions that we did not want to remove.

An example of the resulting ROI can be found in figure 2.

Figure 2. Example of ROI definition in a short-axis view image. a) Original image. b) RV cavity segmentation mask. c) Resulting ROI after the morphological operations. d) Overlapping of both RV cavity segmentation and ROI masks.

**DATA NORMALIZATION**

We considered normalisation necessary as intensities in CMR do not directly correspond to tissue types as they do in computed tomography (CT) imaging, and the
range of scar intensity values can vary from one patient to another depending on the image contrast. In this project, image normalization was performed with respect to the maximum and minimum pixel intensity of the image, according to the next formula:

\[
\text{normalized intensity} = \frac{(\text{pixel intensity} - \min(\text{ROI intensities}))}{(\max(\text{ROI intensity}) - \min(\text{ROI intensities}))}
\]  

(1)

In some other algorithms applied to other cavities, such as in [35], normalization was performed with respect to the blood pool. We considered that this approach was not appropriate for this case as there can be enhanced areas in the blood pool as well, due to fibrotic tissue present in the trabeculae. This would introduce error when normalizing the testing data, due to the impossibility of separating the scar areas from the blood pool data before segmenting.

2.2.1. ALGORITHM 1: GAUSSIAN MIXTURE MODELS WITH EXPECTATION MAXIMIZATION ALGORITHM

2.2.1.1. BACKGROUND

GAUSSIAN MIXTURE MODELS

In general, Gaussian Mixture Models are often used in biometric systems, due to their capability of representing a large class of sample distributions by forming smooth approximations to arbitrarily shaped densities.

Gaussian Mixture Models (GMM) [45], are weighted sums of \(M\) Gaussian probability density functions (PDFs) that describe a class. Commonly, they are referred to as Gaussian density components or Gaussian components. The Gaussian PDF of a \(D\)-dimensional space is defined as a matrix of the form:

\[
g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right\}
\]  

(2)

Where \(\mu\) is the mean vector and \(\Sigma\) the covariance matrix, which can be defined as this matrix equality:

\[
\Sigma = E[(x - E[x])(x - E[x])^T]
\]  

(3)
Where $E[\cdot]$ symbolises the expected value.

Then, a Gaussian Mixture Model has the form:

$$p(x|\theta) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i)$$

(4)

Where $x$ is a D-dimensional data vector (containing measurement or features), $w_i$, with $i = 1, ..., M$, are the mixture component weights, and $g(x|\mu_i, \Sigma_i)$ are the component Gaussian densities. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} w_i = 1$.

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices, and mixture weights from all component densities. These components are represented with this notation:

$$\theta = \{ w_1, \mu_1, \Sigma_1, ..., w_M, \mu_M, \Sigma_M \}$$

(5)

MAXIMUM-LIKELIHOOD ESTIMATION

One of the approaches to estimate the GMM component parameters is the maximum-likelihood estimation.

Images are sets of pixels. If we assume that there is a set of pixels $X = \{x_1, ..., x_N\}$ that belong to the same distribution (i.e. to the same class) described by a probability density function $p(y_n; \theta)$, where $\theta$ is the PDF parameter list, then the likelihood function

$$L(X; \theta) = \prod_{n=1}^{N} p(x_n; \theta)$$

(6)

tells the likelihood of the data $X$ given the distribution parameters $\theta$. The goal is to find the parameters $\hat{\theta}$ that maximizes the likelihood:

$$\hat{\theta} = \arg\max_{\theta} L(X; \theta)$$

(7)

Usually, instead of maximizing directly this function, it is maximized its logarithm

$$\mathcal{L}(X; \theta) = \ln L(X; \theta) = \sum_{n=1}^{N} \ln p(x_n; \theta)$$

(8)
called the log-likelihood function which is analytically easier to handle and gives the same solution.

Depending on the $p(x_n; \theta)$ it might be possible to find the maximum analytically by setting the derivatives of the log-likelihood function to zero and solving $\theta$. In fact, it could be done for the Gaussian PDFs, but usually this approach is intractable. That’s why in practice, an iterative approach such as the expectation maximization (EM) algorithm is used.

**EXPECTATION-MAXIMIZATION ALGORITHM**

The Expectation-Maximization algorithm is an iterative method used for calculating the maximum-likelihood distribution estimates from incomplete data (some elements missing in the feature vector, i.e. not having all the pixel intensities that would form the underlying intensity distribution).

The basic idea of the EM algorithm is, beginning with an initial model $\theta$, to estimate a new model $\hat{\theta}$, such that $L(Y; \hat{\theta}) \geq L(Y; \theta)$. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached, normally suitably selected as $T = \|\theta^{i+1} - \theta^i\| \leq \varepsilon$. The log-likelihood is guaranteed to increase on each iteration until it converges.

In detail, the Expectation-Maximization algorithm has two steps: the E-step and the M-step.

Assume that the training data is a set of pixels $X$ that are interpreted as incomplete data. The missing part $Y$ is the knowledge of which component produced each sample $x_n$. For each $x_n$ there is a binary vector $y_n = \{y_{n,1}, ..., y_{n,C}\}$, where $y_{n,c} = 1$, if the sample was produced by the component $C$, or zero otherwise. Then, the complete data likelihood is:

$$\ln L(X,Y; \theta) = \sum_{n=1}^{N} \sum_{c=1}^{C} y_{n,c} \ln(\alpha_c p(x_n|c; \theta))$$

(9)

Then, the E-step computes the conditional expectation of the complete likelihood, given $\gamma$ and the current estimate $\hat{\theta}(\tau)$. The result it is the so-called Q-function:

$$Q(\theta, \theta^i) \equiv E[\ln L(X,Y; \theta)|X, \theta^i] = \ln(X,W; \theta)$$

(10)
where $\theta^i$ is the previous estimate for the distribution parameters and $\theta$ is the variable for a new estimate describing the (full) distribution. $L$ is the likelihood function exposed above. The Q function calculates the likelihood of the data, including the unknown feature $Y$ marginalized with respect to the current estimate of the distribution described by $\theta^i$.

The elements of $W$ are defined as:

$$w_{n,c} \equiv E[y_{n,c}|X, \theta^i] = Pr[y_{n,c} = 1|x_n, \theta^i]$$  \quad (11)

The probability can be calculated with the Bayes law:

$$w_{n,c} = \frac{\alpha^c_i p(x_n|c; \theta^i)}{\sum_{j=1}^{C} \alpha^c_i p(x_n|j; \theta^i)}$$  \quad (12)

where $\alpha^c_i$ is the priori probability (of estimate $\theta^i$) and $w_{n,c}$ is the posteriori probability that $x_n$ was produced by component $c$.

After that, the M-step is applied, which consists on updating the parameter components according to:

$$\hat{\theta}(t + 1) = \arg \max_{\theta} Q(\theta, \hat{\theta})$$  \quad (13)

Then, the resulting iteration formulas are as follows:

$$\alpha^i_{c+1} = \frac{1}{N} \sum_{n=1}^{N} w_{n,c}$$  \quad (14)

$$\mu^i_{c+1} = \frac{\sum_{n=1}^{N} x_n w_{n,c}}{\sum_{n=1}^{N} w_{n,c}}$$  \quad (15)

$$\Sigma^i_{c+1} = \frac{\sum_{n=1}^{N} w_{n,c} (x_n - \mu^i_{c+1})(x_n - \mu^i_{c+1})^T}{\sum_{n=1}^{N} w_{n,c}}$$  \quad (16)

All these parameters are kept together at $\theta^{i+1}$.

Both steps are iteratively computed until the convergence criterion is satisfied.

One of the drawbacks of the EM-based methods is that they are highly dependent on the initialization. In this case, the initial model $\theta^0$ was obtained by using fuzzy c-means, which is a clustering algorithm that generates fuzzy partitions for any set of numerical
data. These partitions are useful for corroborating known substructures or suggesting substructure in unexplored data. The clustering criterion used to aggregate subsets is a generalized least-squares objective function \[46],[47], as the one shown below:

\[
J_m = \sum_{i=1}^{N} \sum_{j=1}^{num\_gauss} \mu_{ij}^m \|x_i - c_j\|^2
\]  

(17)

where \(N\) is the number of pixels that we want to aggregate in clusters, \(num\_gauss\) corresponds to the number of clusters or Gaussian components that we want to obtain, \(\mu_{i,j}\) is the degree of membership of the pixel \(x_i\) in the \(j\)th cluster, \(c_j\) is the centre of the \(j\)th cluster and \(m\) is the fuzzy partition matrix exponent for controlling the degree of fuzzy overlap, with \(m > 1\). Fuzzy overlap refers to how fuzzy the boundaries between clusters are, that is the number of data points that have significant membership in more than one cluster.

### 2.2.1.2. ALGORITHM IMPLEMENTATION

This algorithm was implemented using the Matlab toolbox in [48].

**MODEL GENERATION**

In order to create the Gaussian Mixture Models of the scar and non-scar intensities, we used the leave-one-out approach, that is to say, we used all the patient images of the data set except the unseen one to generate them. This training procedure was implemented as follows:

For each patient case, we loaded the LGE CMR image and both the RV cavity and the scar segmentations. Then, normalisation of the RV pixel intensities was performed as explained before. Afterwards, the scar mask was used to find which pixels inside the RV cavity belonged to scar and which ones to non-scar tissue. That way, two lists were obtained: one containing the intensities of the pixels belonging to scar, called ‘rv_scar_intensity’ and one containing all the intensities of the non-scar pixels, called ‘rv_no_scar_intensity’.
A label was assigned to the pixels of each list. Label 1 was assigned to the scar pixels while label 2 was associated to the non-scar ones. These labels were contained in vectors of same length as the intensity vectors.

All these data were kept in two two-column matrices. The first one, called ‘training_data_total’, contained the pixel intensities of all the RV patient cavities. The scar intensities were kept in the first column while the non-scar ones were kept in the second one. The second matrix, contained the labels corresponding to each pixel in the same manner.

After that, a function called ‘gmmb_create’ (from Matlab toolbox [48]) computed the two Gaussian models, one for the scar tissue and one for the non-scar, by finding the parameters of the Gaussian models that best estimated the intensity densities using the Expectation-Maximization algorithm, as explained in the background section. The number of desired Gaussian components was set to 11 by parameter analysis, as it will be explained. The initialization of the EM algorithm was performed using the Matlab function ‘fcm’, with the parameter $m$ set as 2. This function randomly initializes the cluster membership values $\mu_{i,j}$ and calculates the cluster centres according to

$$c_j = \frac{\sum_{i=1}^{N} \mu_{i,j}^m x_i}{\sum_{i=1}^{N} \mu_{i,j}^m}$$  \hspace{1cm} (18)$$

After that, it updates $\mu_{i,j}$ according to

$$\mu_{i,j} = \frac{1}{\sum_{k=1}^{num\_gauss} \left( \|x_i - c_j\|^2 \|x_i - c_k\|^2 \right)^{\frac{m-1}{2}}}$$  \hspace{1cm} (19)$$

and computes the objective function $J_m$. These steps are performed until $J_m$ improves less than 0.001.

**IMAGE TESTING**

Once the models were generated, the LGE CRM image and the RV cavity segmentation of the unseen image were used. The RV cavity segmentation was pre-processed and dilated as explained in the beginning of section 2.2. The dilation mask was used to define the ROI in which the GMM classification needed to be performed. First, the
pixel intensities of the ROI were normalised. After that, a function called ‘gmm_pdfs’ (from Matlab toolbox [48]), which implements equation 4, was used to compute the probability of each pixel of belonging to each possible label (scar or non-scar).

The output of ‘gmm_pdfs’ was the probability density function of each pixel according to each of the Gaussian models, kept in a two-column matrix. By thresholding these probabilities according to a certain confidence value, a segmentation of the image could be obtained. As probabilities can range from 0 to 1, this was the possible range of confidence values. After the parameter analysis, we decided to set the confidence threshold as 0.55.

All these operations were performed linearizing the image. Therefore, in order to obtain the segmentation mask, the linear vectors needed to be transformed to a 3D arrays again and kept in niftii format using the function ‘save_untouch_nii’ from Matlab Toolbox of [44].

**PARAMETER ANALYSIS**

Choosing the number of GMM components is an important issue. The usual trade-off in model order selection arises: with too many components, the mixture may over-fit the data, while a mixture with too few components may not be flexible enough to approximate the true underlying model [49].

The number of components to model both the scar intensity and the non-scar intensity distributions was set empirically, by means of both a qualitatively and a quantitatively analysis of the results obtained when changing the parameters. For the quantitative analysis, the Dice overlap index (see section 2.4.1) was used to analyse the performance of the algorithm. The number of possible Gaussian mixture components considered were: 3, 4, 5, 7, 9, 11 and 13.

The confidence value at which the classification was done, was also set empirically. This confidence value was equivalent to the probability threshold above which the pixels were classified as scar. As this value was a probability, it could range from 0 to 1 and that is why these are the values that were considered: 0.05, 0.07, 0.1, 0.2, 0.3, 0.4, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9 and 0.95.
In figure 3, it is shown how the Dice overlap indexes changed with respect to the number of Gaussian components used. Overall, the number of Gaussian mixture components selected did not show much influence on the result, as same median Dice indexes were obtained. Although this must seem surprising, similar results were obtained in an evaluation of GMM of colours and texture features for image segmentation [49], which showed that when GMM modelled a single texture class and the segmentation process was supervised, selecting the number of components became less important.

<table>
<thead>
<tr>
<th>Median dice metric results for each number of Gaussian components and confidence threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Gaussian components</td>
</tr>
<tr>
<td>Confidence Threshold</td>
</tr>
<tr>
<td>Dice</td>
</tr>
</tbody>
</table>

Figure 3. Parameter analysis for the GMM based on the dice overlap indices.

Nevertheless, in our case we can see that while the number of Gaussian mixture models selected did not influence at all the result for low confidence threshold values, slightly higher Dice overlap indexes were obtained with higher number of Gaussian components when considering higher confidence thresholds. For this reason, we decided to set the number of Gaussian components to 11, although in other studies on scar segmentation it was shown that keeping the number of Gaussian components between 1 and 5 was enough to model the image intensities [35].
Regarding the confidence threshold, best results were obtained when the confidence threshold was low, between 0.1 and 0.2. However, the lower was the confidence threshold, the bigger the scar segmentation area and the bigger the number of false positives. A qualitative analysis of the ground-truth data indicated that scars were too generous in some patient cases when using these thresholds (figure 4). For this reason, we decided to set the threshold at 0.55, so that to avoid having noisy areas.

Figure 4. Scar segmentations for different confidence threshold values on a short-axis view of the right ventricle (patient 002). It can be observed that for confidence values 0.20 and 0.40, some pixels were mislabelled. For confidence 0.55, no more mislabelled pixels were obtained. However, the higher the threshold, the smaller the area recognized as scar.

### 2.2.2. ALGORITHM 2: GAUSSIAN MIXTURE MODELS WITH FULL-WIDTH-HALF-MAXIMUM

In this algorithm, we applied the GMM as before, using the same parameters. That way, we obtained a prelabelled image which was used as to provide the seeds for the FWHM algorithm. The objective of implementing FWHM was to try to improve the results of
the GMM. As it can be seen in figure 4, there is a need to use high confidence thresholds to avoid including blood pool pixels in our segmentation. This leads to very small scar segmentation areas, which can become bigger by using region growing.

To implement this algorithm, firstly, the unseen image and its RV cavity segmentation were loaded. After that, the training of the GMMs was performed using all the other images to create the Gaussian mixture models for both scar and non-scar intensities. These models were applied onto the unseen image to classify their pixels into scar and non-scar by thresholding the probability density functions obtained from them, as in algorithm 1. The resulting segmentation was used as a seed for the FWHM algorithm.

In FWHM, the seeded region was determined to grow to include all pixels with signal intensity (SI) > 50% of the maximum signal intensity (MX) among the seeded pixels. Therefore, the scar region was defined as all the pixels surrounding the seeds with $SI > (0.5 \cdot MX)$ [20]. 2D region growing was performed using this condition. An extra condition was added to avoid the growth outside of the region of interest.

2.2.3. ALGORITHM 3: GAUSSIAN MIXTURE MODELS WITH GROWCUT

2.2.3.1. BACKGROUND

GrowCut is an image segmentation method based on Cellular Automaton (CA). A cellular automaton is generally a computer algorithm that is discrete in time and space and operates in a lattice of sites, in the case of image segmentation, on pixels or voxels. In GrowCut, images are segmented using a small number of user-labelled pixels as seeds. These seeds act like bacteria that grow and want to occupy the neighbouring pixels by attacking them with a predefined force. The strongest ones win the battles and give their label to the neighbouring pixels, obtaining that way the final segmentation [50].

A (bi-dimensional, deterministic) cellular automaton (CA) is a triplet $A = (S, N, \delta)$, where $S$ is a nonempty set, called the state set, $N$ is the neighbourhood system and $\delta: S^N \rightarrow S$ is the local transition function (rule). This function defines the rules of calculating the cell’s state at $t + 1$ time step, given the states of the neighbourhood cells.
at previous time step $t$. The argument of $\delta$ indicates the state of the neighbourhood cells at a given time, while its value the central cell state at the next time.

Commonly used neighbourhoods are the von Neumann neighbourhood and the Moore neighbourhood. Both use norms to define them. For a pixel $p \in \mathbb{Z}^n$, its neighbourhood $qen$ would be defined as:

- Neumann neighbourhood: $N(p) = \{ q \in \mathbb{Z}^n : \|p - q\|_1 := \sum_{i=1}^{n} |p_i - q_i| = 1 \}$; \hspace{1cm} (20)
- Moore neighbourhood: $N(p) = \{ q \in \mathbb{Z}^n : \|p - q\|_{\infty} := \max_{i=1,n} |p_i - q_i| = 1 \}$; \hspace{1cm} (21)

The cell state $S_p$ is defined as a triplet $(l_p, \theta_p, \vec{c}_p)$, where $l_p$ is the label of the current cell, $\theta_p$ is the strength of the cell and $\vec{c}_p$ is the feature image vector. The strength is assumed to be $\theta_p \in [0,1]$.

An unlabelled bi-dimensional image of $k \times m$ pixels is considered a particular configuration of the cellular automaton, where the cellular space $P$ is defined by the $k \times m$ set by the image, and initial states for $\forall p \in P$ are set to:

$$l_p = 0, \theta_p = 0, \vec{c}_p = I_p;$$ \hspace{1cm} (22)

where $I_p$ is the intensity of pixel $p$ in a grayscale image.

The final goal of the segmentation is to assign each pixel one of the possible $K$ possible labels, in our case, 2 (as we want to distinguish between scar and non-scar tissue).

When the user starts the segmentation by specifying the segmentation seeds, the seeded cells’ labels are set accordingly while their strength is set to the seed strength value. This sets the initial state of the cellular automaton.

At iteration $t + 1$ cell labels $l_p^{t+1}$ and strengths $\theta_p^{t+1}$ are updated as follows:

Each cell (or pixel $p$) is evaluated. First, their cell state at time $t$ ($S_p^t$) is copied to initialize the new state $S_p^{t+1}$. After this, the neighbours try to attack the current cell, and the values of the new state can be updated according to the next rules in code 1:
for $\forall q \in N(p)$

\[
\text{if } g \left( \| \vec{C}_p - \vec{C}_q \|_2 \right) \cdot \theta_q^t > \theta_p^t \\
\quad l_{p}^{t+1} = l_{q}^{t} \\
\quad \theta_{p}^{t+1} = g \left( \| \vec{C}_p - \vec{C}_q \|_2 \right) \cdot \theta_q^t 
\]

end
end

Where $g \left( \| \vec{C}_p - \vec{C}_q \|_2 \right)$ is a simple monotonous decreasing function bounded to [0, 1]:

\[
g \left( \| \vec{C}_p - \vec{C}_q \|_2 \right) = 1 - \frac{\| \vec{C}_p - \vec{C}_q \|_2}{\max \| \vec{C} \|_2} \tag{23}
\]

That way, we can see that the attack force depends on both the attacker cell’s strength $\theta_q^t$ and the feature vectors $\vec{C}_p$ and $\vec{C}_q$. If the attack force is greater than the current cell’s strength $\theta_p^t$, the cell is conquered and its label and strength are changed for the ones of the attacker. That way, the strongest pixels occupy the neighbouring sites and gradually spread over the image, until a stable configuration is achieved.

### 2.2.3.2. ALGORITHM IMPLEMENTATION

In this algorithm, we applied the GMM as before, using same parameters to obtain the seeds for the GrowCut algorithm, as in algorithm 2. However, in this case, the seeds were defined in a different way:

To initialize the GrowCut algorithm, images were seeded with 3 labels:

- 1 for scar pixels
- -1 for non-scar pixels
- 0 for pixels where it was uncertain whether they were scar or not.

In the region of interest, pixels were labelled with label 1 when the probability of these pixels being scar was above a confidence value. Pixels were labelled with label -1 when the probability of this pixels being no-scar was above a certain confidence value. This confidence value was set to 0.55 for both cases. The threshold choice was done based on
the parameter analysis performed in section 2.2.1.2. The remaining pixels inside the ROI where scar was uncertain were labelled with label 0. All the pixels outside the region of interest were assigned label -1. An example can be found in figure 5.

![Figure 5](image_url)

**Figure 5. Short-axis RV phantom image to illustrate the GrowCut pre-labelling.** The red line defines the ROI. All the pixels outside the ROI are labelled -1. Inside the ROI, those pixels with intensities belonging to scar are labelled with 1. Regions where it is certain that there is no scar are labelled with -1. Otherwise, the label is 0.

After that, the GrowCut algorithm, as explained, was applied. A 2D version of the algorithm was used for speed purposes and can be found in [51]. A 3D approximation was obtained by applying the algorithm in the three axes of the image. The result of this algorithm was kept for posterior analysis.

### 2.2.4. ALGORITHM 4: 3D AUTOMATIC GRABCUT

In this work, we also pretended to implement an algorithm based on Graph-Cuts for scar segmentation. The reason is that many studies had used Graph-Cuts for scar segmentation before, with promising results, and we wanted to see if it could also work for our data set. For this reason, we wanted to implement a 3D version of the GrabCut algorithm with the use of the GMMs from algorithm 1 to initialise the algorithm.

Unfortunately, the full development of this algorithm could not be performed during the scheduled period. An initial implementation of the 3D automatic GrabCut showed that
the obtained segmentations where exactly the same as the resulting from the GMM. For this reason, its results will not be presented in the results section.

The 2D implementation of the automatic GrabCut algorithm already showed very few scar pixel labelling changes with respect to the GMM scar segmentations. We hypothesize that as 3D segmentations encourage coherence in the results and normally segmented areas reduce their size with respect to the 2D procedure, it might be possible that as changes were already small, there were no changes anymore. Future work should involve a deep exploration of the algorithm so that to determine where the problem remains or whether the results of the GMM are indeed of such a very high quality that this algorithm adds no modifications.

In the next section, we will explain the basics of the algorithm and how it was implemented.

2.2.4.1. BACKGROUND

GrabCut is a variant of Graph-cuts which allows foreground/background subtraction based on a user selection of a region of the image by drawing a bounding box. Two colour histograms are created, one using the pixels inside the box and another using the pixels outside the box. With these two models, the minimum s-t cut of the graph constructed from the image pixel is found. New color histograms are then created using the assignment resulting from the minimum cut, and the process is repeated until convergence. Once the foreground has been separated from the background, it can be cut out of the original background, composited onto a new background and transformed to logical type to obtain the segmentation mask [52].

Before starting to explain our method in detail, let us refresh the basics of Graph-cuts.

In Graph-cuts, an image is represented as an undirected graph. A graph \( G = (v, e, w) \) is a network of nodes (vertices \( v \)) connected by edges (\( e \)). In this image representation, the image pixels \( p \in P \) become the nodes and the edges represent the connection between the pixels, i.e. the connection between adjacent pixels. Each edge \( e \in e \) is assigned a non-negative weight cost \( w_e \) defined by the \( w \) cost function. Suppose there are two extra special vertices, with one being the ‘source’ (called \( s \)) and the other being the ‘sink’ (called \( t \)), which represent the terminal nodes for each label (i.e. foreground/background, which in our case would be scar/non-scar). Then \( v = P \cup ...
\{S, T\}. A cut that separates the graph into two subgraphs \(S\) and \(T\), with \(s \in S\) and \(t \in T\), is known as an ‘s-t cut’ and it is normally defined as the sum of weight edges between the two sets:

\[
c(S, T) = \sum_{u \in S, v \in T, (u, v) \in \varepsilon} w(u, v)
\]

The goal of Graph-cuts is to find the minimum s-t cut (which may not be unique, because many cuts can have the same cost). This cut gives a partitioning that corresponds to an underlying image or volume.

Following the example of Boykov and Jolly, 2001 [53] and Rother, Kolmogorov and Blake, 2004 [52], the image is linearized and an array of grey values \(z = (z_1, \ldots, z_p, \ldots, z_P)\) is obtained, indexed by the (single) index \(n\). \(P\) represents the total number of pixels in the image. The segmentation of the image is expressed as an array \(\alpha = (\alpha_1, \ldots, \alpha_p, \ldots, \alpha_P)\) with same length as the image array, with values ranging \(\alpha_n \in \{0, 1\}\) , with 0 for background and 1 for foreground. Another vector \(k = \{k_1, \ldots, k_n, \ldots, k_N\}\) is introduced, with \(k_n \in \{1, \ldots, K\}\), assigning to each pixel, a unique GMM component, either from the scar model or the non-scar one, according to the \(\alpha_n\) value. This is done for optimization purposes.

The set of edges \(\varepsilon\) that made up the graph would consist of two types of undirected edges: \(n\)-links (neighbouring) and \(t\)-links (terminal-links, i.e. the ones that connect each pixel to \(s\) and \(t\)). Each pixel \(p\) has two links connecting it to each of the terminal nodes \(\{p, s\}\) and \(\{p, t\}\). Each pair of adjacent pixels \(p\) and \(q\) \(((p, q))\) in \(N\), where \(N\) is the set of pairs of neighbouring pixels, are connected by an n-link. Therefore, \(\varepsilon = N \cup_{p \in P} \{(p, S), (p, T)\}\).

That way we can define different weights for the edges in \(\varepsilon\) according to the next formulas:

\[
\begin{align*}
\{p, q\} &= \gamma [\alpha_p \neq \alpha_q] e^{-\beta \|z_p - z_q\|^2} \\
\{p, S\} &= D(\alpha_p, k_n, \theta_s, z_p) = -\log(p(z_p | \alpha_p, k_p, \theta_s)) - \log(\pi(\alpha_p, k_p)) \\
\{p, T\} &= D(\alpha_p, k_n, \theta_T, z_p) = -\log(p(z_p | \alpha_p, k_p, \theta_T)) - \log(\pi(\alpha_p, k_p))
\end{align*}
\]

Where \(p(\cdot)\) is a Gaussian probability distribution and \(\pi(\cdot)\) are mixture weighting coefficients and \(\theta_i = \{\pi(\alpha, k), \mu_\ell(\alpha, k), \Sigma_\ell(\alpha, k)\}, \alpha = \{0, 1\}, k = 1 \ldots K\) for \(\theta_i = \)
\{\theta_s, \theta_t\} \), are the parameters for the two GMM models (one for scar and one for non-scar), as introduced in the GMM section. Note the change of nomenclature. Here, \(\pi\) are the weights of the Gaussian components, \(\mu\) are the means and \(\Sigma\) the covariance matrices of the \(2K\) Gaussian components. \(k_p\) represents the Gaussian component that has been assigned to pixel \(p\).

Using these weights, an energy function \(E\) can be built to define the cost of the minimum cut \(c(S, T)\). This function is designed so that its minimum corresponds to a good segmentation, as it is guided by the edges weights defined above, which take into account the grey-level histograms for scar and non-scar and the coherence along the segmentation. Then, the \(E\) function is captured by a ‘Gibbs’ energy of the form:

\[
E(\alpha, \theta, z) = U(\alpha, k, \theta, z) + V(\alpha, z)
\]

Where the data term \(U\), known as unary term, evaluates the fit of each Gaussian mixture model to each pixel, and it is defined to be as in [52]:

\[
U(\alpha, k, \theta, z) = \sum_{p \in P} D(\alpha_p, k_n, \theta_i, z_p) \text{ for } \theta_i = \{\theta_s, \theta_t\}
\]

And the smoothness term \(V(\alpha, z)\), known as pairwise term, can be written as in [53] and not as in [52] as we are working with monochromatic images instead of coloured images:

\[
V(\alpha, z) = \gamma \sum_{p, q \in N} [\alpha_p \neq \alpha_q] e^{-\beta(z_p - z_q)^2}
\]

This energy term encourages coherence in regions of similar grey-level. The constant \(\beta\) can be \(\geq 0\). If \(\beta = 0\), it encourages smoothness everywhere, to a degree determined by the constant \(\gamma\). However, it is better to choose \(\beta > 0\) as it relaxes the tendency of smoothness in regions of high contrast, as proposed by [53]. That’s why, following the indications of [52], \(\beta\) is defined as:

\[
\beta = \left(2\langle(z_p - z_q)^2\rangle\right)^{-1}
\]

Where \(\langle \cdot \rangle\) denotes expectation over an image sample.

Once the energy model \(E\) is fully defined, the segmentation can be estimated as a global minimum:

\[
\hat{\alpha} = \arg\min_\alpha E(\alpha, \theta, z)
\]
Minimisation is performed by using the max flow algorithm, a standard minimum cut algorithm proposed by Yuri Boykov and Vladimir Kolmogorov [54].

Loosely speaking, the max flow algorithms find the minimum cut by finding the maximum flow from the source $s$ to the sink $t$, where the maximum flow is the maximum ‘amount of water’ that can be sent from $s$ to $t$ by interpreting the edges as directed ‘pipes’ with capacities equal to edge weights. The maximum flow saturates a set of edges in a graph dividing the nodes into two disjoint parts $\{S, T\}$ corresponding to a minimum cut. Thus, the min-cut and max flow problems are equivalent. In fact, the maximum flow value is equal to the cost of the minimum cut.

The max flow algorithms can mostly belong to two different groups: Goldberg-Tarjan style ‘push-relabel’ methods and the Ford-Fulkerson style methods based on ‘augmenting paths’. We will give an intuition of how the second group methods work.

Standard augmenting paths based algorithms work by pushing flow along non-saturated paths from the source to the sink until the maximum flow in the graph $G$ is reached. A typical augmenting path stores information about the distribution of the current $s \rightarrow t$ flow $f$ among the edges of $G$ using a residual graph $G_f$. The topology of $G_f$ is the same as $G$ but capacity of an edge in $G_f$ reflects the residual capacity of the same edge in $G$ given the amount of flow already in the edge. At the initialization, there is no flow from the source to the sink ($f = 0$) and edge capacities in the residual graph $G_0$ are equal to the original capacities in $G$. At each new iteration, the algorithm finds the shortest path $s \rightarrow t$ along non-saturated edges of the residual graph. If a path is found, then the algorithm augmentates it by pushing the maximum possible flow $df$ that saturates at least one of the edges in the path. The residual capacities of edges in the path are reduced by $df$ while the residual capacities of the reverse edges are increased by $df$. Each augmentation increases the total flow from source to sink by $f = f + df$. The maximum flow is reached when $s \rightarrow t$ path crosses at least one saturated edge in the residual graph $G_f$.

In GrabCut, energy minimization is performed iteratively. The advantage of this method is that it allows automatic refinement of the segmentations. At each iteration, the newly labelled pixels are used to refine the GMM parameters. The end of the iterations can be
established with a threshold, stopping them when the Energy function is lower than a certain number.

### 2.2.4.2. ALGORITHM IMPLEMENTATION

The implementation of the GrabCut algorithm with GMM was based on a 2D GrabCut Matlab source-code found in [55], which was a direct implementation of the algorithm described in the background section, for foreground subtraction in colour images [52]. The code modifications performed can be summarised as follows:

- Code adaptation to grey-scale images
- Implementation of the 3D extension
- Introduction of the RV cavity dilation as the search area for the scar
- Introduction of the scar and non-scar intensity Gaussian mixture trained models to compute the unary term.

These modifications will be more detailed below:

Firstly, the user selection of a rectangular region containing the foreground, i.e. the scar, was automatized by replacing it by the dilation of the RV cavity segmentation as the region of interest to find the scar.

After that, instead of using k-means to generate two Gaussian mixture models (one for foreground and one for background) after the user selection, as the source-code initially did, we directly used the trained GMMs computed as in algorithm 1.

Next, we computed the pairwise terms. The original code considered a 4-pixel neighbourhood and computed the pairwise between one pixel and the pixels that were up, down, on the left and on the right of this pixel. However, as we were willing to implement the 3D extension of the code, we also had to compute the pairwise term for the neighbours in other image planes. That way, we were considering images as 6-connected voxel graphs, as shown in figure 6.
Figure 6. 2D and 3D pixel adjacency graphs. a) A 2D image with 4x4 pixels. b) A 4-connected pixel adjacency graph. c) A volume image with 3x3x3 voxels. d) A 6-connected voxel adjacency graph.

Notice that the extension to the third dimensions also needed to be performed to compute the \( \beta \) parameter.

When evaluating the existence of the neighbours, a condition was added to assure that they were found inside the region of interest. If neighbours existed but where located beyond the limits of the search area, their pairwise term was not computed.

The unary term for each pixel was computed using the trained GMMs of algorithm 1 for the first iteration.

The energy minimization process was performed using the source-code that was found in the original code. We considered that no changes were needed as we always worked with linearized images. However, this assumption might be incorrect, as results were not as expected.

The parameter \( \gamma \) was set as 20. We tried the performance of the code for several \( \gamma \) values, ranging from 0.5 to 1000, but the results were the same in all cases.

2.2.5. ALGORITHM 5: n-SD

As n-SD algorithms has been extensively used in clinical research, we also implemented them for comparison purposes.

In the n-SD method, a fixed number of standard deviations from a mean signal within a non-enhanced region is computed. However, as reported in [42], there is no agreement regarding how this procedure must be performed. For example, in scar segmentation in the left atrium, different studies made use of three and four SDs above the mean intensity of a region of the left atrium blood pool, two SD above the mean intensity of a healthy left atrial wall, six SDs above the mean intensity of a region of the left
ventricular wall. The reason of the existence of all these methods is that the mean intensity of enhanced region varies in unpredictable ways due to the complex response to scan parameters and patient physiology, rendering the selection of a signal intensity threshold not so straightforward. That's why many published techniques that seem to perform very well for a set of images might not work well for another set of images. Some of the discrepancy reasons involve the variability in the CMR acquisition parameters, such as the optimal timing of imaging after contrast enhancement administration, the choice of ideal dosage of contrast agent and the selection of the best inversion time.

For this work, we decided to use the pixel intensities of the region of the left ventricle wall in contact with the right ventricle to find the statistical measures to compute the threshold. Since Tetralogy of Fallot patients do not only present scar in the wall but within the cavity due to the trabeculae, the RV cavity segmentation mask could not be used for this purpose, as there can be some enhanced regions present in it. Also, using healthy RV myocardium intensities was not feasible since RV walls cannot be easily identified in this cohort of patients.

As segmentations for the myocardium were not provided, they were performed manually. Nevertheless, manual segmentations are time consuming. Due to the limited time to perform this project, we decided to just segment 10 slices for patient case. Segmentations were performed on the short-axis view, by using the ITK-snap software. An example of a segmentation can be found in figure 7.

![Figure 7. LV myocardium segmentation for mean and SD computation (patient 008). Left: original LGE MR image. Right: segmentation of the LV myocardium.](image_url)

Regarding the number of SD to define the threshold, we decided to use 2-SD, 3-SD, 4-SD, 5-SD, 6-SD, 7-SD, 8-SD, 9-SD, 10-SD and 15-SD and see with the results which one gave the best performance.
The leave-one-out approach was also applied with this methodology to compute the scar segmentation in each patient case. For each patient, the segmenting threshold was set by computing the statistical measurements using the LV myocardium segmentations of the remaining 12 patients.

2.3. STATISTICAL ANALYSIS

The Wilcoxon rank sum test was used for algorithm performance comparison. A p-value < 0.05 was considered significant.

2.4. EVALUATION METRICS

In this work, two quantification metrics were used to evaluate the performance of the algorithms: the Dice similarity coefficient and the total volume difference.

2.4.1 DICE SIMILARITY COEFFICIENT

The dice overlap index was used as a quantitative measure of the algorithm performance. The dice score is a quantification measure used to compute the overlap between the ground truth segmentation and the obtained segmentation. It gives the proportion of true positives in the segmentation. Its formula is given by:

$$d = \frac{2|X \cap Y|}{|X| + |Y|}$$

Where $X$ is the segmented region in the ground-truth data and $Y$ is the new segmentation.

This was implemented by computing the number of non-zero common elements in both segmentations matrices, by using the & and nnz Matlab commands, multiplying it by two and dividing it by the sum of the non-zero elements of each of the segmentations.
2.4.2. TOTAL VOLUME ERROR

The total volume error between the obtained segmentations and pseudo ground truth is a volume-based metric based on the formula:

$$\delta V = |V_T - V_G|$$  \hspace{1cm} (34)

where $V_T$ is the volume of scar in the segmentation and $V_G$ is the volume of scar in consensus segmentation.
3. RESULTS

In this section, the results from our evaluation are presented with figures and tables.

For the quantitative analysis, the scar segmentation masks obtained semi-automatically by the cardiologists were used as pseudo-ground truth. Figure 8 shows the Dice overlap scores for all the algorithms in the 13 patient cases. The median Dice overlap for each algorithm are shown in table 2, together with the median total volume differences between the pseudo-ground truth and the algorithm segmentations produced.

Table 2. Median Dice overlap indexes and median volume differences for each algorithm.

| Algorithm       | Median Dice overlap index | Median | |V| (ml) |
|-----------------|---------------------------|--------|-----------------------|
| GMM             | 47.5057                   | 14.5633|
| GMM + FWHM      | 57.6963                   | 4.2649 |
| GMM + GrowCut   | 48.4161                   | 13.4623|
| 2-SD            | 15.0543                   | 210.8621|
| 3-SD            | 18.4213                   | 179.2052|
| 4-SD            | 21.6212                   | 144.9705|
| 5-SD            | 29.5089                   | 79.3496 |
| 6-SD            | 36.0850                   | 36.2574 |
| 7-SD            | 40.6079                   | 30.1855 |
| 8-SD            | 45.8333                   | 24.2138 |
| 9-SD            | 48.0441                   | 14.1389 |
| 10-SD           | 41.8707                   | 11.9897 |
| 15-SD           | 10.8492                   | 26.1021 |
The results for each patient individually can be shown in figure 9. Regarding the n-SD methods, as showing the results for all the n-SD threshold values would make the figure look messy, we decided to show the individual results for the n-SD that resulted in the best median Dice index, in this case, 9-SD. The individual total volume differences are shown in supplementary table 1 and 2.

![Dice overlap indexes for each patient case](image)

Figure 9. Patient dice overlap indexes for each method.

As it can be observed in figure 9, 9-SD gave very low Dice results for some of the patients. For this, in table 3, we listed the number of SDs that defined the best threshold for scar segmentation in each patient based on the Dice result (column n-SD for max dice). High variability of this number can be observed among the different patient cases. In table 3, the statistical measures that were used for each patient to compute the threshold are also displayed.
Table 3. Statistical measures for n-SD threshold definition and selection of the methods that give best quantitative results.

<table>
<thead>
<tr>
<th>Patient case</th>
<th>Mean</th>
<th>SD</th>
<th>n-SD for Max Dice</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>12.4388</td>
<td>4.2810</td>
<td>10-SD</td>
</tr>
<tr>
<td>002</td>
<td>28.2316</td>
<td>14.7998</td>
<td>5-SD</td>
</tr>
<tr>
<td>003</td>
<td>9.8942</td>
<td>4.2237</td>
<td>10-SD</td>
</tr>
<tr>
<td>004</td>
<td>12.0761</td>
<td>3.7323</td>
<td>10-SD</td>
</tr>
<tr>
<td>005</td>
<td>20.6304</td>
<td>4.6199</td>
<td>6-SD</td>
</tr>
<tr>
<td>006</td>
<td>13.2602</td>
<td>5.2601</td>
<td>9-SD</td>
</tr>
<tr>
<td>007</td>
<td>5.0624</td>
<td>2.5018</td>
<td>8-SD</td>
</tr>
<tr>
<td>008</td>
<td>17.6149</td>
<td>5.3456</td>
<td>8-SD</td>
</tr>
<tr>
<td>009</td>
<td>15.6224</td>
<td>4.4863</td>
<td>7-SD</td>
</tr>
<tr>
<td>010</td>
<td>11.3184</td>
<td>3.8235</td>
<td>9-SD</td>
</tr>
<tr>
<td>011</td>
<td>14.0815</td>
<td>4.2894</td>
<td>4-SD</td>
</tr>
<tr>
<td>012</td>
<td>9.2232</td>
<td>4.2612</td>
<td>10-SD</td>
</tr>
<tr>
<td>013</td>
<td>12.6924</td>
<td>5.2393</td>
<td>8-SD</td>
</tr>
</tbody>
</table>

Median Dice overlap indexes for the 9-SD case and the best-dice-n-SD case are shown in table 4, for comparison purposes.

Table 4. Median Dice overlap scores for the 9-SD, best-dice-n-SD method.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median Dice overlap index</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-SD</td>
<td>48.0441</td>
</tr>
<tr>
<td>Best-dice-n-SD</td>
<td>58.8973</td>
</tr>
</tbody>
</table>

All dice overlap indexes obtained by the different methods for each patient case can be found in figure 10.

Figure 10. Dice overlap scores for all the methods and patients including the best-dice-n-SD.
Example segmentations for a single slice with the different methods can be shown in figure 11.

Figure 11. Scar segmentation results for the different algorithms in a short axis view of patient 001. Clockwise from top-left: original LGE scan, pseudo-ground-truth scar segmentation mask, GMM+FWHM, 15-SD, GMM+growcut, GMM.

The Wilcoxon rank test was performed to find whether the differences in Dice metrics among the different algorithms were significant. All the p-values are summarised in table 5. The results showed that the differences in results between GMM+FWHM and GMM (p<0.05) were statistically significant. Differences in results between the best-dice-n-SD and all the methods except GMM+FWHM were also significant.

Table 5. p-values resulting from the Wilcoxon rank test for each pair of algorithms.

<table>
<thead>
<tr>
<th>Method</th>
<th>GMM</th>
<th>GMM+FWHM</th>
<th>GMM+GrowCut</th>
<th>Best-dice-n-SD</th>
<th>9-SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>-</td>
<td>0.005619</td>
<td>0.411924</td>
<td>0.001031</td>
<td>0.719522</td>
</tr>
<tr>
<td>GMM+FWHM</td>
<td>0.005619</td>
<td>-</td>
<td>0.051329</td>
<td>0.572684</td>
<td>0.040239</td>
</tr>
<tr>
<td>GMM+GrowCut</td>
<td>0.411924</td>
<td>0.051329</td>
<td>-</td>
<td>0.005619</td>
<td>0.681618</td>
</tr>
<tr>
<td>Best-dice-n-SD</td>
<td>0.001031</td>
<td>0.572684</td>
<td>0.005619</td>
<td>-</td>
<td>0.007639</td>
</tr>
<tr>
<td>9-SD</td>
<td>0.719522</td>
<td>0.040239</td>
<td>0.681618</td>
<td>0.007639</td>
<td>-</td>
</tr>
</tbody>
</table>
4. DISCUSSION

In this section, the limitations of the study and the conclusions drawn after the algorithm performance comparison will be exposed.

Before starting the discussion, we would like point out that to counteract the fact that the Dice index might not always represent well the performance of an algorithm as we can see when analysing the results qualitatively, we recommend to combine and read the results all together with the total volume error measure, as proposed in the evaluation framework of the challenges [28], [29]. With this volume-based metric, a truer picture of the segmentation can be given.

DATA LIMITATIONS

We can see that the Dice overlap results obtained when using all the methods are quite small, especially if we compare them with the median Dice scores obtained with the challenge algorithms (around 80% in the left atrium post-ablation scar segmentation algorithms) [28], [29]. Let us start the discussion by explaining some of the reasons that arise from the data used.

In first place, we must bear in mind that the RV cavity segmentation masks were obtained semi-automatically by a method based on thresholding on the 3D bSSFP MRI sequences, under the assumption that the 3D LGE and the 3D bSSFP sequences were aligned. However, although this assumption was partially correct, there were in fact misalignments that affected the ROI definition (figure 12). Because of that, when performing the dilation of the RV cavity mask, some surrounding tissues with similar intensity as scar, such as the surrounding fat, were included in the search area, introducing false positive scar segmentations in some patient cases. Moreover, as the RV masks were obtained with thresholding, they were not smooth and many times they were made up of several unconnected regions. This also affected our ROI definition and introduced errors.
Figure 12. Example of a short-axis slice of patient 005. Left: original LGE MR image. Right: LGE MR image with the original RV cavity mask overlaid. It can be observed that the mask is moved to the left, covering areas outside of the RV chamber.

Secondly, the RV cavity mask dilation to cover the wall was performed with a fixed-size in all the directions. However, the part of the wall in contact with the LV is thicker than the rest. As surgical scars can be found in the LV wall due to the VSD repair, it was necessary to assure that the dilation covered them. This made dilation masks too big for other areas, increasing the number of non-scar bright regions included in the ROI, which led to false positives in few cases.

As it can be seen from these two first points, obtaining a good segmentation of the RV epicardium is an important step before performing scar segmentation to avoid the inclusion of surrounding non-scar bright structures in our ROI. As shown in the left atrium scar segmentation challenge [28], best results were obtained when the epicardium segmentations were manually defined. Nevertheless, due to the time limitation to perform this project, we could not refine the RV masks, as the manual procedure is very time consuming. In future work, the RV masks should be improved for ROI definition. However, an automatic and accurate method for ROI definition should be developed to reduce the burden of time.

Thirdly, we must consider that segmentation masks provided as pseudo-ground truth were also obtained semi-automatically with a method based on thresholding in many cases, with some posterior manual corrections performed by experts. When performing a detailed analysis of the scars, we detected some areas labelled as scar that were more likely to be image noise. This small segmentation errors could have introduced errors in our GMM training, affecting the final results as well.

Furthermore, in general, the pseudo-ground-truth segmentations over-estimated the scar tissue, as they included areas where cardiologists could not be that certain whether they were scar or not (figure 13). With our methodology, we obtained thinner scar
segmentations that covered just the areas where you could be more certain that there was scar.

![Figure 13. Scar segmentation on a short-axis slice of patient 002. From left to right: Original LGE CMR image, pseudo-ground-truth scar mask on image, scar mask obtained with GMM+FWHM.](image)

In LGE CMR, it is often challenging to establish ground-truth on scar regions. A limitation of this work was that the pseudo-ground truth data was obtained by just one experienced observer, which made the algorithm evaluation difficult as the pseudo-ground truth used was not robust to interobserver variability. For instance, in images with poor contrast enhancement ratio, observers may differ in their opinion of the level of enhancement that is likely to be scar. Ideally, the pseudo-ground truth should have been obtained by at least two different observers and combined to obtain a consensus segmentation, as performed in the challenges [28], [29].

The last important issue is that LGE MRI also enhances the pulmonary valve, which should not be labelled as scar unless the valve itself presents fibrosis. Although in many cases the valve was located outside of the ROI and there was not risk of segmenting it as scar, there were some others in which it was inside the ROI and it was segmented as scar (figure 14), introducing false positives in our segmentations. With the improvement of the RV masks, this problem should be partially solved. However, future algorithms should include a valve classifier, which could be obtained considering positional information, to avoid labelling it as scar.

![Figure 14. Example of valve enhancement in a short axis view slice (patient 004). From left to right: original LGE MR image, pseudo-ground-truth scar mask (red), scar mask obtained with GMM (green) and scar mask obtained with GMM (green) with the ROI mask overlaid (blue). Arrow indicates the where the pulmonary valve is.](image)
EVALUATED ALGORITHMS

Moving to the comparison of the results obtained with the different methods, several conclusions can be drawn.

Firstly, it can be concluded that it was not possible to define a single fixed-threshold that guaranteed good results for all the cases of our data set. If we look at the results obtained for 9-SD (figure 9), we see that although good results were obtained for some cases, for some others the threshold was just too high and no segmentation was performed. In fact, the optimal n-SD threshold for each patient case differed a lot, as it can be shown in table 3. Some of the discrepancy reasons might involve the physiological variability in patients and/or the variability in the CMR acquisition parameters, such as the optimal timing of imaging after contrast enhancement administration, the choice of ideal dosage of contrast agent and the selection of the best inversion time, as explained before.

However, the fact that just 10 slices were used to compute the mean and SD of the nulled myocardium must be considered. As explained in [42], the extent of the area evaluated for the statistical measures computation can imply a significant difference in the final threshold. To give an intuition, figure 15 shows how the mean and standard deviation vary depending on the analysed region size. These results made us think that a proper implementation of the n-SD must involve the evaluation of the intensity mean and SD using the healthy myocardium of all the slices per patient case rather than just a few of them, so that to consider all the variability.

![Intensity Mean ± SD](image)

Figure 15. Example of regional variation in the reference region statistics. The region considered is marked in red. By choosing a larger area, the mean and SD of the signal intensity increases.

Some studies have suggested employing a patient-specific (instead of fixed) threshold level to solve the variability problem [56]. However, as the aim of the project was the automation of the scar segmentation process, we considered that this method was not
suitable for scar segmentation in our cohort of patients, where variability is abundant, and we discarded its use. This conclusion agrees with the challenges, which claimed that fixed models cannot handle variability and, therefore, they are not suitable for scar segmentation [28], [29].

Regarding the GMM-based methods, we first must consider two limitations before analysing the results:

The parameter analysis was performed on the dataset on which the testing was performed. This could have introduced a bias towards good results. Ideally, we would have used a separate set for parameter tuning, and therefore divided our data into training set, validation set and testing set. Given that we are just using 13 cases, this approach would have left very few cases for testing, which would have not allowed us to draw conclusions. A better approach would have been the use of cross-validation. In any case, given that the number of Gaussian components does not have much influence on the results, parameters should not change much when performing a proper parameter analysis.

Another limitation of our evaluation framework is the size of the image database. In this study, only patient cases with high quality images and reliable segmentations were considered. Although it was sufficient for most of our purposes, further studies with increased data size should be performed including poor quality images or patient datasets from other centres to see how robust the algorithms are to different protocols and acquisition parameters and for study completeness.

However, considering the median Dice score obtained, the median volume difference, and the statistical significance, GMM + FWHM seemed the most powerful approach for scar segmentation in patients with tetralogy of Fallot. Considering the conclusions drawn in the left atrium scar segmentation challenge after the algorithm evaluation [9], our hypothesis is that even though it is claimed that intensity enhancement modelling and normalisation, such the ones performed with GMM, might not be sufficient for accurate scar segmentation given the dynamic range of scar intensities and the limitations of our data set, with the application of other algorithms afterwards, great improvements in performance can be achieved. Using the GMM method to obtain the scar seeds for the FWHM algorithm, which includes region growing that keeps the pixel
connectivity, seems a promising approach. Nevertheless, we noticed that qualitative differences among the scar segmentations obtained with the different methods in each patient case were not very high and that further studies involving a larger sample size and corrected data should be performed to draw reliable conclusions.

In addition, we observed that there were few cases in which we were not able to detect much of the pseudo-ground-truth scar. We saw that in those cases there was less intensity contrast between the scar and non-scar regions than in others. Homogeneity in image contrast among all the patient cases is crucial for the proper functioning of the GMM (i.e. for proper modelling of the scar and non-scar intensities), which is one of the limitations of our approach.

Furthermore, by analysing the results for each patient case, we could also notice that even though we detected the scars from surgical repair of the VSD and the widening of the RVOT in all the cases, we missed the scars located in the trabeculae in some cases. The reason is that VSD and RVOT scars were brighter than the scars located in the trabeculae, so they were more confidently labelled as scar by the GMM and are easier to detect and differentiate from the blood pool. An example of a missed trabeculae scar is shown in figure 16.

![Figure 16. Scar segmentation on a short axis view slice (Patient 008). a) Original LGE CMR image. b) pseudo-ground-truth scar segmentation mask (red). c) Scar segmentation obtained with GMM+FWHM (green). d) Both scar masks overlaid. Arrow indicates scar in the trabeculae. Notice that it is not detected by our algorithm.](image-url)
Finally, we would like to end the discussion saying that the results of the segmentations were shown to the cardiologists of Royal Brompton Hospital. Their positive feedback was a proof that even though further work must be done to automatize the scar segmentation procedure and improvements must be performed in many levels, we are undoubtedly moving to the right direction.

However, as there is still room for improvement, it is probable that there will be new algorithmic developments in scar segmentation in the near future. One of the possibilities is the appearance of algorithms based on deep learning and neural networks. In fact, deep learning has already been applied to segment the endo- and epicardium of the RV with promising results [57]. Therefore, it is obvious the potential of these emerging techniques for TOF scar segmentation: they could not only be used for the scar segmentation but also for the automatic selection of the ROI. Nevertheless, for the emergence of completely new techniques in scar segmentation, advancements in the image acquisition technique will also be required.
5. CONCLUSIONS

Although scar segmentation can be an easy task for the human visual system, it remains a challenging task for computers.

In this project, we have explored several algorithms to segment the scars in patients with tetralogy of Fallot and we have analysed their performance.

The results showed that fixed-thresholding methods such as n-SD are not suitable for scar segmentation in our data set, as a global threshold cannot be defined due to the high variability among the patient images.

GMM with FWHM turned out to be the most promising approach, giving the best results quantitatively and qualitatively. However, further studies should be performed with improved data and larger sample size to overcome the limitations and draw reliable conclusions.
REFERENCES


[34] Y. Gao, B. Gholami, R. S. MacLeod, J. Blauer, W. M. Haddad, and A. R. Tannenbaum, “Segmentation of the endocardial wall of the left atrium using local region-based active contours and statistical shape learning,” 2010, p. 76234Z.


[52] C. Rother, V. Kolmogorov, and A. Blake, “GrabCut — Interactive Foreground Extraction using Iterated Graph Cuts.”


[56] R. Karim et al., “A Method to Standardize Quantification of Left Atrial Scar

ACKNOWLEDGEMENTS

I would like to offer my deep gratitude to Prof. Guang Zhong Yang for welcoming me in such an outstanding institution as it is the Hamlyn Center and for letting me feel as one of their researchers.

I would also like to express my gratitude to Dr. Sonya Babu-Nayaran for giving me the opportunity of getting involved in such a challenging project. Many thanks for the data and all the information.

I would particularly like to acknowledge all the assistance provided by the PhD Yun Gu. Thanks to him, I learnt more than I could imagine when I was starting the project.

I would also like to thank Dr. Miguel Ángel González Ballester, for his suggestions, comments and encouragement.

Special thanks to all the people who gave me support during this project, particularly, to Andreas Schmitz and my family. Their advices were more valuable than one could imagine.

Last but not least, I would like to express my deepest heartfelt appreciation to Guillem, for his love and for being always there.
Supplementary figure 1. Example of low contrast patient images. a) Left: short-axis view heart slide where scar presence is assumed. Right: pseudo-ground-truth scar segmentation mask overlaid on the slide. b) Left: four-chamber view heart slide where scar presence is assumed. Right: pseudo-ground-truth scar segmentation mask overlaid on the slide.


**Supplementary table 1. Total volume differences for each patient and GMM-based method.**

| Patient case | $|8V|$ GMM (ml) | $|8V|$ GMM + FWHM (ml) | $|8V|$ GMM + GrowCut (ml) |
|--------------|---------------|----------------------|-------------------------|
| 001          | 23.0981       | 11.3299              | 22.5211                 |
| 002          | 16.7677       | 3.1733               | 15.2963                 |
| 003          | 25.3192       | 20.3777              | 24.8879                 |
| 004          | 4.2453        | 4.2649               | 4.0936                  |
| 005          | 32.7995       | 6.3340               | 29.6920                 |
| 006          | 14.5633       | 1.4706               | 13.4623                 |
| 007          | 11.3139       | 1.7298               | 11.6896                 |
| 008          | 15.0175       | 9.7657               | 14.5093                 |
| 009          | 5.2093        | 0.3062               | 4.7122                  |
| 010          | 9.3765        | 2.7849               | 8.5208                  |
| 011          | 6.7743        | 29.7967              | 1.4755                  |
| 012          | 8.5977        | 3.4211               | 7.9992                  |
| 013          | 27.3184       | 20.1254              | 26.7930                 |
### Supplementary table 2. Total volume differences for each patient and n-SD based method.

| Patient case | $|\Delta V|$ 2-SD (ml) | $|\Delta V|$ 3-SD (ml) | $|\Delta V|$ 4-SD (ml) | $|\Delta V|$ 5-SD (ml) | $|\Delta V|$ 6-SD (ml) | $|\Delta V|$ 7-SD (ml) | $|\Delta V|$ 8-SD (ml) | $|\Delta V|$ 9-SD (ml) | $|\Delta V|$ 10-SD (ml) | $|\Delta V|$ 15-SD (ml) |
|--------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 002          | 210.8621       | 166.4561       | 110.0044       | 86.2181        | 63.0721        | 44.8021        | 28.5151        | 13.9861       | 4.9521        | 32.1499       |
| 004          | 200.5322       | 179.2052       | 153.4062       | 131.3122       | 95.8832        | 57.3865        | 39.4868        | 17.4488       | 7.3021        | 10.4268       |
| 005          | 153.4210       | 123.9510       | 80.1800        | 48.6029        | 14.9521        | 30.1855        | 7.5266         | 46.3281       | 49.4597       | 53.9164       |
| 006          | 505.2200       | 439.9480       | 350.7670       | 265.4370       | 157.1630       | 57.7181        | 30.5911        | 0.9186        | 11.4235       | 28.3318       |
| 007          | 226.5195       | 196.4795       | 144.9705       | 104.5315       | 47.6139        | 21.2685        | 8.9108         | 11.3139       | 19.1757       | 26.1021       |
| 009          | 144.8844       | 104.7644       | 64.4404        | 24.6352        | 9.5167         | 1.0212         | 2.5053         | 5.2093        | 7.5323        | 11.2882       |
| 010          | 149.5824       | 121.2754       | 89.1904        | 56.4200        | 31.0522        | 15.6537        | 17.6639        | 1.1930        | 3.9554        | 15.2068       |
| 012          | 299.6962       | 254.0032       | 213.1392       | 170.8262       | 122.9492       | 65.7123        | 46.8808        | 22.2672       | 11.5265       | 10.2780       |