Using deep learning to segment breast and fibroglandular tissue in MRI volumes

Mehmet Ufuk Dalmış,[a] Geert Litjens, Katharina Holland, Arnaud Setio, Ritse Mann, Nico Karssemeijer, and Albert Gubern-Mérida
Radboud University Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen, The Netherlands

(Received 9 August 2016; revised 19 December 2016; accepted for publication 20 December 2016; published 13 February 2017)

Purpose: Automated segmentation of breast and fibroglandular tissue (FGT) is required for various computer-aided applications of breast MRI. Traditional image analysis and computer vision techniques, such atlas, template matching, or, edge and surface detection, have been applied to solve this task. However, applicability of these methods is usually limited by the characteristics of the images used in the study datasets, while breast MRI varies with respect to the different MRI protocols used, in addition to the variability in breast shapes. All this variability, in addition to various MRI artifacts, makes it a challenging task to develop a robust breast and FGT segmentation method using traditional approaches. Therefore, in this study, we investigated the use of a deep-learning approach known as "U-net."
Materials and methods: We used a dataset of 66 breast MRI's randomly selected from our scientific archive, which includes five different MRI acquisition protocols and breasts from four breast density categories in a balanced distribution. To prepare reference segmentations, we manually segmented breast and FGT for all images using an in-house developed workstation. We experimented with the application of U-net in two different ways for breast and FGT segmentation. In the first method, following the same pipeline used in traditional approaches, we trained two consecutive (2C) U-nets: first for segmenting the breast in the whole MRI volume and the second for segmenting FGT inside the segmented breast. In the second method, we used a single 3-class (3C) U-net, which performs both tasks simultaneously by segmenting the volume into three regions: nonbreast, fat inside the breast, and FGT inside the breast. For comparison, we applied two existing and published methods to our dataset: an atlas-based method and a sheetness-based method. We used Dice Similarity Coefficient (DSC) to measure the performances of the automated methods, with respect to the manual segmentations. Additionally, we computed Pearson's correlation between the breast density values computed based on manual and automated segmentations.
Results: The average DSC values for breast segmentation were 0.933, 0.944, 0.863, and 0.848 obtained from 3C U-net, 2C U-nets, atlas-based method, and sheetness-based method, respectively. The average DSC values for FGT segmentation obtained from 3C U-net, 2C U-nets, and atlas-based methods were 0.850, 0.811, and 0.671, respectively. The correlation between breast density values based on 3C U-net and manual segmentations was 0.974. This value was significantly higher than 0.957 as obtained from 2C U-nets ($P < 0.0001$, Steiger's Z-test with Bonferoni correction) and 0.938 as obtained from atlas-based method ($P = 0.0016$).
Conclusions: In conclusion, we applied a deep-learning method, U-net, for segmenting breast and FGT in MRI in a dataset that includes a variety of MRI protocols and breast densities. Our results showed that U-net-based methods significantly outperformed the existing algorithms and resulted in significantly more accurate breast density computation. [https://doi.org/10.1002/mp.12079]

Key words: breast segmentation, deep learning, MRI

1. INTRODUCTION

Automatic segmentation of breast and fibroglandular tissue (FGT) is a key step in automated analysis of breast MRI for several clinically relevant applications. One example is computer-aided detection (CADe) systems which use breast segmentation as an initial step to determine the region to search for lesions.[1–3] Automated quantification of background parenchymal enhancement (BPE) could be considered as another example to the applications which require FGT segmentation, as BPE is evaluated within FGT.[4] More importantly, volumetric measurement of breast density in MRI requires segmentation of breast and FGT.[5–9] Breast density, measured by the amount of FGT relative to the breast volume, is a strong predictor of breast cancer risk.[10–12] Although breast density is often measured based on mammograms, these are two-dimensional projection images which may lead to inaccuracy in estimation of breast density, due to the tissue superimposition.[13] T1-weighted images in breast MRI provide three-dimensional (3D) information with a strong

contrast between fat and fibroglandular tissues within the breast, making it ideal for evaluating breast density. For an objective evaluation of breast density based on MRI, automatic segmentation of breast and FGT is required.

Breast segmentation consists of breast-air and breast-pectoral muscle separation, where the latter is usually considered to be a challenging problem. There have been several studies investigating breast segmentation in MRI. Nie et al.[5] developed an algorithm to segment breast and FGT, using B-spline curve fitting for pectoral muscle-breast separation, which required initial inputs from the user. Milenkovich et al.[14] proposed a fully automated method using edge maps obtained by applying a tunable Gabor filter, and they reported 0.96 for the average Dice similarity coefficient[15] (DSC) value for 52 MRI's. Koenig et al.[16] developed a method which uses nipple detection as a prior step. Martel et al.[17] used Poisson reconstruction method to define breast surface using automatically detected edges, achieving a DSC score of 0.90 for 332 Dixon images and 0.96 for 8 T1-weighted images. Similarly, Gallego-Ortiz et al.[18] suggested a method that aims to construct breast surface, and they used a breast atlas. They reported a DSC value of 0.88 for a large dataset consisting of 409 MRI's. Atlas-based methods were also employed by Khalvati et al.[19] (obtaining average DSC scores of 0.94 for Dixon and 0.87 for T1w images), Lin et al.[20], and Gubern-Mérida et al.[9,21] (obtaining an average DSC of 0.94 in 50 MRI's). Pectoral muscle-breast boundary separation in the latter method was based on automatic detection of sternum as a landmark which is not always clearly visible in all patients. Wang et al.[22] suggested a method which uses sheet-like appearance of pectoral muscle boundary and does not require detection of a landmark. Giannini et al.[23] used gradient characteristic of pectoral muscle to separate it from breast. Ivanovska et al.[8] suggested a level-set method to simultaneously correct bias field and segment breast, which achieved an average DSC of 0.96 in 37 MRIs.

Several studies have also investigated FGT segmentation. Although MRI provides a high contrast between fat and FGT in breast, intensity inhomogeneities introduce the main difficulty in FGT segmentation. Adaptive fuzzy c-means (FCM) was used by Nie et al.[5] for FGT segmentation. Gubern-Mérida et al.[9] used N4 bias-field correction[24] as a prior step to their pipeline and then applied a gaussian mixture model for segmenting FGT, which resulted in an average DSC of 0.80 in 50 MRI's. At a later study, we extended this work by adding an additional N4 bias-field correction step applied within the computed breast mask for each breast[4] and obtained a DSC value of 0.81 on 20 breast MRI's. In another study, simultaneous FCM and bias-field correction methods were applied,[8] which achieved an average DSC of 0.83 on 37 cases. Wu et al.[6] suggested use of FGT atlas as a refinement step after FCM and reported 0.77 as average DSC value. As postprocessing steps, Gubern-Mérida et al.[9] and Ravazi et al.[25] also proposed methods for removing skin folds from the segmentation, an imaging artifact that mimics FGT. An average DSC of 0.84 was reported in the latter work.

Although the studies summarized above reported satisfactory results within their datasets, applicability of these methods is usually limited by the characteristics of the images used in the study datasets. However, breast MRI varies with respect to different MRI protocols used. Even in a single hospital, a variability would be expected in MRI data across years, as protocols are changed from time to time due to the improvements in acquisition or MRI units. In addition to the variability in MRI protocols, there is also variability in breast shapes, sizes, densities, and pectoral muscle shapes. Another problem is the MRI artifacts such as intensity inhomogeneities, ghosting, or aliasing effects. Skin folds may also occur which mimic the appearance of FGT. For each of such artifacts, a separate algorithm or filter is needed to be designed and included in the segmentation process. Even then, strength and presentation of these artifacts vary with respect to breast shapes, acquisition protocols, or patient movements during the acquisition. All this variability and artifacts make it a challenging task to develop a robust and widely applicable breast and FGT segmentation method using traditional approaches. Therefore, we decided to investigate the use of deep-learning methods for breast and FGT segmentation as an alternative to the traditional methods. The main advantage of the deep-learning methods lies in their ability to learn relevant features and models directly from examples, rather than requiring design of features or filters specific for each problem. Although it has a longer history, deep convolutional neural networks have attracted a considerable attention in the last few years due to the groundbreaking success it demonstrated in various fields such as image and speech recognition, natural language understanding, and lately, in medical image analysis field. In this study, we investigated two deep-learning approaches based on the U-net architecture[26] for breast and FGT segmentation in MRI. To our knowledge, this is the first study that applies deep learning for breast and FGT segmentation in MRI.

One advantage of U-net is that it is possible to use entire images of arbitrary sizes, without dividing them into patches. This results in a large receptive field that the network uses while classifying each voxel, which is important in segmentation of large structures like the breast. To represent the variety mentioned above, we used MRI scans obtained by different acquisition protocols and breasts from different breast density categories in our dataset. We reported segmentation performances of the trained U-nets for each beast density category and for each MRI protocol, as well as the overall performance. For comparison, we also applied two different existing methods[9,22] on our datasets.

2. MATERIALS AND METHODS

2.A. Patient population and DCE-MRI acquisition

At the Radboud University Medical Center, Nijmegen, the Netherlands, breast MRI is performed for screening women with intermediate or high risk for developing breast cancer; for preoperative staging in women with an invasive lobular carcinoma, an invasive carcinoma under the age of

50, indeterminate tumor size, tumors larger than 3 cm, locally advanced carcinoma treated with neo-adjuvant chemotherapy; for troubleshooting in women with findings that cannot be resolved by biopsy (BIRADS 0); and for evaluation in women with lymph-node metastasis with an unknown primary tumor. For this study, we randomly selected 66 breast MRI examinations from different women (ages 25–75) that had no history of breast cancer. To cover a large variability of MRI protocols and breast sizes, we intentionally included scans acquired in a long time span (from 2000 to 2015) from women at different breast density categories. For 53 of the patients, breast density scores were measured on the mammograms that were acquired within 6 month prior to the MRI, with the Volpara software method (version 1.5.0; Volpara Health Technologies, Wellington, New Zealand). Accordingly, 10, 14, 14, and 15 women were categorized in breast density categories 1, 2 ,3, and 4, respectively (1 is the least and 4 is the most dense). Volpara scores were not available for the remaining 13 cases. For some of these patients, there were no mammography studies close to the date when the MRI was scanned (mammograms were acquired in another period of the year or, for women younger than the age of 35, they were not acquired at all). For the other patients, only processed mammograms were available, while Volpara requires raw mammograms to compute density.

In our study dataset, there was a variability in MRI acquisition parameters, as we selected images from an image archive that covers a large time period. Twenty-seven of the 66 MRI's were acquired in 1.5 Tesla (T) MRI units, while 3T was used in the rest. The MRI's acquired in 3T also varied in acquisition parameters. We have grouped these acquisition parameters into five protocols and details are given in Table I. Note that the symbols X, Y, and Z denote the directions corresponding to the directions orthogonal to the sagittal, coronal, and axial planes, respectively. All of the images used in this study were non-fat-suppressed T1-weighted images.

TABLE I. DCE-MRI acquisition parameters.

	Protocol 1	Protocol 2	Protocol 3	Protocol 4	Protocol 5
Number of MRI's	27	17	4	13	5
Field strength (tesla)	1.5	3	3	3	3
Orientation	Coronal	Coronal	Axial	Axial	Axial
Flip angle	20	13	10	20	15
Echo time (ms)	4	2.36–2.41	1.71	2.06	1.71
Repetition time (ms)	7.8	7.35	4.56	5.03	5.5
Resolution-X (mm)	0.66	0.88–1.0	0.8	0.8	0.8
Resolution-Y (mm)	1.3	1	0.8	0.8	0.8
Resolution-Z (mm)	0.66	0.88-1.0	1.0	1.0	1.0
Field Of view-X (mm)	340	340-380	360	360	360
Field Of view-Y (mm)	156	160	360	360	360
Field Of view-Z (mm)	170	170-190	176	160	176

2.B. Manual segmentation

For each MRI, we manually generated reference breast and FGT segmentations, using an inhouse-developed workstation. These manual segmentations were performed by a trained biomedical engineer and were revised and validated by a breast radiologist with 9 yr of expertise in breast MRI.

For manual breast segmentation, we manually delineated each breast by drawing contours in several axial slices. The workstation automatically interpolated between these contours and generated mask of the whole breast volume. The annotator could interactively add more contours between slices when the result of the interpolation was not satisfactory. The interpolation algorithm uses a spline surface function out of the path points of the contours and scans the missing slices in between by a marching squares algorithm.

For FGT segmentation, we used manual thresholding to select FGT voxels. First, to reduce the effect of bias field, we applied N4 bias-field correction using the breast segmentation mask obtained in the previous step. Then, we manually selected a threshold value for the whole breast to select FGT voxels. Finally, we applied additional manual corrections where necessary. These manual corrections included modifying the threshold value for individual slices, and manually excluding some of the regions by drawing contours.

2.C. Data preprocessing

We initially re-oriented all coronal acquisitions into axial orientation. As two breasts of a patient are often symmetric, we approached the segmentation problem as a segmentation task for a single breast. Therefore, we divided each breast MRI scan into two breast volumes, each one including one breast (right and left). This could be performed safely by dividing the volume from the middle, as the use of a dedicated breast coil and positioning of the patients in MRI units ensure that right and left breasts always appear in their respective right and left halves in the MRI scan. After this division, we mirrored the left breasts so that they appear similar to the right breasts. This action was performed to facilitate the learning process, thus the network only needed to learn the shape variations of the right side. The segmentation task was then performed for both right and mirrored left breasts. We did not apply any further preprocessing to the data.

2.D. Deep-learning network architectures and U-net

Deep-learning methods are defined by LeCun et al.[27] as *"representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one*

level (starting with the raw input) into a representation at a higher, slightly more abstract level.". The main idea behind this approach is that it is possible to implement very complex functions using convolutional neural networks with several layers, each having nonlinear output layers. Each of these layers transform the representation of the data (image) at their input to a higher level of representation. The advantage of this approach is, the weights in the network are learned by examples through iterations of gradient-descent based algorithms, leaving no need for engineering features for specific problems.

A patch-based approach is commonly employed in deep learning, where images are divided into small patches of certain sizes and provided to the network. However, this approach limits context information provided to the network, as the receptive field of these networks is limited to the patch size. Increasing the patch size to increase contextual information is not always the best solution because training might then become computationally infeasible. The advantage of convolutional networks with U-net[26] architecture is that it is possible to use entire images of arbitrary sizes, without dividing them into patches. This results in a large receptive field that network uses while classifying each voxel, which is important in segmentation of large structures like the breast. Therefore, we selected this architecture to investigate the use of deep-learning methods for breast and FGT segmentation.

The name of the "U-net" stands for the "U"-shape of the network as seen in Fig. 1. This is a fully convolutional network, which consists of convolution and max-pooling layers at the descending, or in other words, the initial part of the U. This part can be seen as down-sampling stage, as at each max-pooling layer, the input image size is divided by the size of the max-pooling kernel size. At the later part of the network, or at the ascending part of the U, up-sampling operations are performed which are also implemented by convolutions, where kernel weights are learned during training. The arrows between the two parts of the U show the incorporation of the information available at the down-sampling steps into the up-sampling operations performed in the ascending part of the network. In this way, the fine-detail information captured in descending part of the network is used at the ascending part.

In this study, we used four down-sampling and four up-sampling steps. In each down-sampling step, we used two convolutional layers with a kernel size of 3 × 3, each followed by a rectified-linear unit (ReLu) for nonlinearity, and

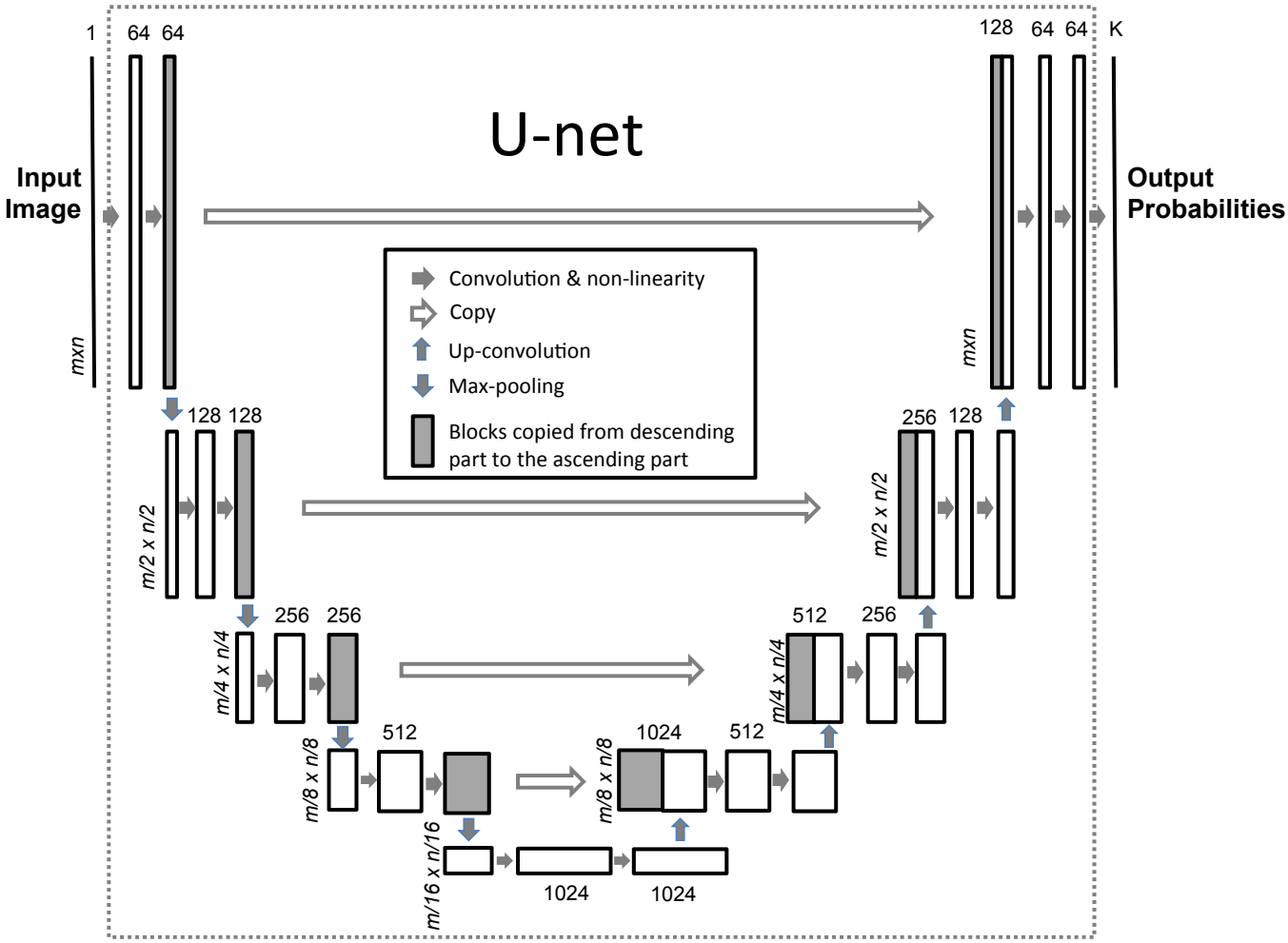

FIG. 1. Deep-learning network with U-net architecture used in this study. K is the number of classes. [Color figure can be viewed at wileyonlinelibrary.com]

finally a max-pooling layer with 2×2 kernel size for down-sampling. We used Glorot-uniform[28] initialization for weights of the network and RMSProp[29] for gradient-descent optimization. In the up-sampling steps, we up-convolved the output of the previous step, by initially upscaling the image by a factor of two using nearest neighbor interpolation, then convolving it with a convolutional layer with a kernel size of 2×2 and a ReLu layer. Then, this was concatenated with the output of the corresponding down-sampling layer. Finally, two convolutional layers, each followed by a ReLu, were applied to this concatenated image. After the up-sampling steps, there is a final convolutional and a following sigmoid unit, which produces probability outputs for each class. Different than the U-net described by Ronneberger et al.,[26] we used padded convolutions in this network, which allowed us to process the whole image at once.

2.E. Segmenting breast and FGT with U-net

The final goal of the segmentation task in this study is to get three-class labels within a given MRI: nonbreast (L_{nb}), fat tissue inside breast (L_{fat}) and FGT inside breast (L_{FGT}). The combination of fat and FGT voxels constitutes the whole breast ($L_{breast} = L_{fat} \cup L_{FGT}$). Traditional approaches usually perform this segmentation task in two steps, first segmenting the breast, and at the second step, segmenting FGT within the obtained breast mask. It would be possible to follow a similar approach using two consecutive (2C) U-nets, performing separate but consecutive tasks. On the other hand, it is also possible to approach the problem as a 3-class (3C) problem and train a single U-net with 3-class outputs. In this study, we investigated both approaches, which are illustrated in Fig. 2. In both approaches, we used all two-dimensional axial slices as samples provided to the network in random order. As we apply four max-pooling layers in total, the sizes of the input images need to be factors of 16 at both dimensions. Therefore, we padded the images with zero values to meet this criteria before feeding them into the network and we cropped the output of the network back to the original size of the images. The final segmentation was determined by selecting the class for each voxel having highest corresponding probability and selecting the largest connected component within the resulting breast masks. Finally, we combined the output segmentations for all slices of right and left breasts to get the final segmentation in the whole MRI. We used all axial slices of the MR volumes, both in training and testing.

2.E.1. Two consecutive U-nets (2C U-nets)

In this approach, we trained a 2-class U-net for breast segmentation and a subsequent 2-class U-net for FGT/fat segmentation within the breast. To get the final result, the segmentation output of the FGT segmentation was masked with the segmentation output of the breast segmentation. While training the second network, the loss function was computed only in the breast region defined by the automated breast segmentation which was produced by the first network. The reason of not using the manual breast segmentations during this training stage was to make the FGT segmentation network to learn to exclude nonrelevant voxels from FGT

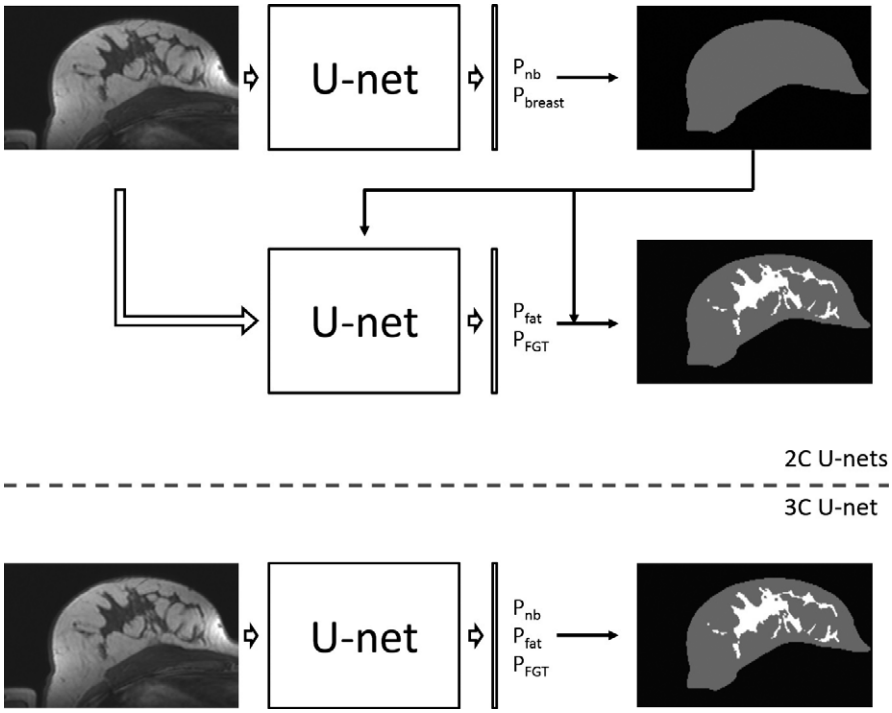

FIG. 2. Two different approaches for applying U-net to breast and FGT segmentation. The upper figure shows 2C U-nets, where two consecutive U-nets are used. The figure below illustrates the other approach, a single U-net with 3-class outputs. P_{nb}, P_{breast}, P_{fat} and P_{FGT} denote the probability values of voxels to belong to nonbreast, breast, fat, and FGT, respectively.

segmentation whenever the first network over-segments the breast.

2.E.2. Three-class single U-net (3C U-net)

In this approach, the output of the last up-sampling layer of the U-net was convolved by three filters, followed by a sigmoid unit, which outputs the probability values of voxels to belong to one of the three classes: P_{nb} for nonbreast, P_{fat} for fat tissue in the breast and P_{FGT} for FGT tissue in the breast. For the cost function, we computed the mean of the negative log-likelihood values of the voxels within the whole slice. However, as FGT/fat separation occurs within breasts and breasts only occupy a small portion of the whole MRI volumes (between one-third to one-twentieth of MRI volumes, depending on the field of view of the acquisition and breast shape), the loss caused by breast/nonbreast separation would dominate the loss caused by FGT/fat separation inside breast, which results in under-training for FGT segmentation. To prevent this, we weighted the loss computed inside the breast by a factor of 10.

2.F. Existing methods for breast and FGT segmentation

We applied two existing breast segmentation methods to our dataset, one of which also has an FGT segmentation step.

The atlas-based method by Gubern-Mérida et al.[9] initially applies a bias-field correction algorithm and normalizes intensity values to reduce intra- and interimage intensity variability. Second, the breasts are segmented using spatial information encoded in a multiprobabilistic atlas[3] and the sternum as an anatomical landmark. FGT segmentation is performed on each breast independently: remaining intensity inhomogeneities are corrected using N4 and subsequently FCM is applied to select FGT voxels inside the breast.[4] Finally, a skin fold removal step is applied.

The sheetness-based breast segmentation method by Wang et al.[22] uses second-order derivatives based on the Hessian matrix to enhance voxels that are part of sheet-like structures. Breast-pectoral and breast-air surfaces are identified using this information and the breast mask is obtained as a combination of both. This method does not include an FGT segmentation step.

2.G. Experiments and evaluation

We used threefold cross-validation to train and test the U-net methods. We had breast MRI's from 66 patients, which corresponded to MRI volumes of 132 breasts in total. Each breast volume consisted of 160 to 260 axial slices in MRI, depending on the acquisition protocol. The separation of the dataset into folds was random; however, we made sure to keep a balanced distribution of breasts from different density categories in training and test sets. We also took care that both breasts of the same patient were always placed in the same set. In each fold, MRI's of five patients in the training set, therefore 10 breasts, were excluded from the training set to be used as a validation set. Therefore we had 39, 5, and 22 MRI scans in training, validation, and test sets, respectively, in each fold. Performances on the validation sets during training were measured using DSC values. When the performance did not improve any further, the training was stopped, and the network that corresponds to the highest performance value in the validation set was selected as the final network for the fold. This final network was applied to the test set. Furthermore, as networks with the U-net architecture are claimed to be trainable with fewer number of images compared to other deep-learning algorithms,[26] we investigated this using the proposed pipeline. Following the same cross-validation strategy as in the previous experiment, we trained the 3C U-net with 5, 10, and 20 training volumes of each fold. We plotted the resulting DSC values for breast and FGT segmentations.

We used DSC to measure the overlap between automated and manual segmentations of the whole MRI volumes. To compare performances of different segmentation algorithms, we applied paired *t*-test to the DSC values obtained for each MRI. We applied multiple test correction to the *P*-values using Bonferoni correction for nine tests (six tests for the four methods in breast segmentation, and three tests for the three methods in FGT segmentation). We also reported DSC results per density category, for the cases where breast density scores as determined by Volpara on a mammogram were available, as well as DSC results for each MRI protocol used in the acquisitions. Furthermore, to complement DSC values on the whole dataset, five additional performance metrics were also computed: average Hausdorff distance[30] (H), average of the highest 5% of H, maximum H, sensitivity, and specificity.

As one of the relevant clinical applications of breast and FGT segmentation is to provide volumetric breast density, defined as ratio of FGT volume to the breast volume, we also measured performance of the FGT segmentation methods in this respect. Pearson's correlation of the volumetric breast density estimates obtained from manual segmentations to the ones obtained from automated segmentations were computed. The correlation values were compared using Steiger's Z-test for two dependent correlations with one variable in common,[31,32] with Bonferoni correction for three tests. We also conducted a Bland–Altman analysis to investigate any bias in breast density measurements and how differences between measurements are distributed. We provided the plots for the Bland–Altman analysis and we computed limits of agreement (LOA), coefficient of variation (CV), standard deviation of the mean values, and sum of squared errors (SSE) for each automated segmentation method.

3. RESULTS

3.A. DSC values for breast segmentation and FGT segmentation

DSC values of breast and FGT segmentations for each method with respect to the manual segmentations are given in Tables II and III, respectively. DSC values obtained for

TABLE II. DSC values for breast segmentation: overall, per breast density category and per MRI acquisition protocol.

All images (66)		2C U-nets 0.944 (0.026)	3C U-net 0.933 (0.028)	Atlas-based 0.863 (0.087)	Sheetness-based 0.848 (0.071)
Per density	Category 1 (10)	0.953 (0.012)	0.946 (0.018)	0.881 (0.045)	0.874 (0.016)
	Category 2 (14)	0.937 (0.017)	0.948 (0.012)	0.889 (0.033)	0.870 (0.027)
	Category 3 (14)	0.938 (0.029)	0.937 (0.019)	0.873 (0.073)	0.858 (0.117)
	Category 4 (15)	0.906 (0.039)	0.921 (0.03)	0.782 (0.134)	0.775 (0.116)
Per protocol	Protocol 1 (27)	0.949 (0.024)	0.939 (0.029)	0.88 (0.06)	0.849 (0.084)
	Protocol 2 (17)	0.930 (0.029)	0.928 (0.026)	0.812 (0.13)	0.826 (0.074)
	Protocol 3 (4)	0.947 (0.011)	0.936 (0.022)	0.863 (0.036)	0.850 (0.03)
	Protocol 4 (13)	0.937 (0.021)	0.926 (0.033)	0.89 (0.035)	0.864 (0.054)
	Protocol 5 (5)	0.940 (0.008)	0.938 (0.013)	0.874 (0.04)	0.878 (0.02)

TABLE III. DSC values for FGT segmentation: overall, per breast density category and per MRI acquisition protocol.

All images (66)		2C U-nets 0.811 (0.11)	3C U-net 0.850 (0.086)	Atlas-based 0.671 (0.207)
Per density	Category 1 (10)	0.665 (0.172)	0.748 (0.117)	0.386 (0.140)
	Category 2 (14)	0.785 (0.09)	0.825 (0.073)	0.659 (0.129)
	Category 3 (14)	0.877 (0.044)	0.90 (0.038)	0.792 (0.125)
	Category 4 (15)	0.849 (0.073)	0.870 (0.061)	0.724 (0.185)
Per protocol	Protocol 1 (27)	0.792 (0.137)	0.845 (0.1)	0.676 (0.19)
	Protocol 2 (17)	0.823 (0.124)	0.857 (0.088)	0.683 (0.202)
	Protocol 3 (4)	0.795 (0.086)	0.83 (0.076)	0.625 (0.265)
	Protocol 4 (13)	0.835 (0.074)	0.86 (0.067)	0.689 (0.218)
	Protocol 5 (5)	0.823 (0.06)	0.843 (0.073)	0.58 (0.302)

different breast density categories and different MRI protocols are also provided in the same tables. Comparing overall results of the two different U-net methods, we see that 2C U-nets performed better by 1.1% in breast segmentation compared to 3C U-net (P =0.0055). On the other hand, 3C U-net achieved a higher DSC value in FGT segmentation ($P < 0.0001$). Atlas-based and sheetness-based methods performed significantly worse compared to U-net methods ($P < 0.0001$ in all comparisons). The difference between DSC values of atlas-based and sheetness-based methods was not statistically significant ($P > 0.1$).

Regarding the MRI acquisition protocols, the most remarkable drop in segmentation performance was observed in breast segmentation with atlas-based method for Protocol 2 (Table II). The average DSC for the MRI volumes obtained with this acquisition protocol was 0.812 (0.13), while it was 0.88 (0.05) for all other MRI volumes obtained with other acquisition protocols. The comparison between the two averages produced a P-value of 0.061 when a t-test for independent samples was applied.

To train the second U-net of the 2C U-nets approach, we used automated breast masks generated from the first U-net and a DSC value of 0.811 (0.11) for FGT segmentation was obtained. This approach was chosen over using manual breast masks during training to mimic the testing conditions. However, we did not observe a significant difference compared to using manual breast segmentations during training (DSC of 0.808 (0.16), $P = 0.73$ with paired t-test).

Additional performance measures for breast and FGT segmentations are given in Table IV. The differences between performances of the methods measured with these metrics are comparable to the differences based on DSC measurements. Sensitivity is an exception to this. In particular, sensitivity of the atlas-based method for FGT segmentation is significantly higher than those obtained using U-net methods ($P < 0.001$ with paired t-test for both U-net methods). However, atlas-based method also had a significantly lower specificity compared to the same methods ($P < 0.001$ with paired t-test for both U-net methods). While the difference in specificity values might look small, one should note that FGT regions occupy a small volume pulling the specificity values close to 1. For the same reason, small differences in FGT segmentation have a large effect on sensitivity values.

The changes in segmentation performances of the 3C U-net when trained with 5, 10, 20, and 39 volumes of each fold are illustrated in Fig. 3. The DSC values obtained using 20 and 39 training volumes were almost identical for breast segmentation (only 0.002 higher with 39 training volumes and $P = 0.49$ with paired t-test). For the same numbers of training volumes, the difference between DSC values obtained for FGT segmentation was 0.023, which was statistically significant ($P = 0.002$ with paired t-test).

3.B. Visual examples

In this section, we provide examples from our segmentation results. Figure 4 shows a case for which segmentation was performed properly by all methods. Other examples demonstrate how different variations and artifacts in MRI volumes may affect segmentation algorithms.

The example given in Fig. 5 corresponds to a dense breast (density category 4) in which the breast-pectoral muscle boundary is less visible due to the presence of dense tissue.

Figure 6 shows an example of how segmentation algorithms were affected by magnetic field inhomogeneities. The strong bias field visible in the image caused the sheetness-based method to misinterpret the breast-pectoral muscle

TABLE IV. Additional performance metrics for breast and FGT segmentation.

		2C U-nets	3C U-net	Atlas-based	Sheetness-based
Breast	Mean H (mm)	2.9 (1.1)	2.9 (0.1)	5.7 (3.3)	6.0 (2.1)
	5% H (mm)	11.2 (4.9)	11.3 (5.2)	21.5 (8.7)	23.7 (9.7)
	Max H (mm)	34.3 (15.1)	30.5 (9.2)	40.0 (12.6)	49.2 (15.3)
	Sensitivity (%)	93.7 (3.7)	97.6 (2.0)	91.6 (7.5)	93.6 (8.7)
	Specificity (%)	99 (0.5)	98.1 (1.0)	96.1 (3.7)	95.0 (3.4)
FGT	Mean H (mm)	3.1 (1.8)	2.6 (0.1)	7.4 (4.6)	-
	5% H (mm)	11.6 (7.8)	9.7 (5.2)	26.7 (12.4)	-
	Max H (mm)	49.3 (15.0)	45.0 (9.2)	51.3 (14.3)	-
	Sensitivity (%)	79.5 (15.2)	84.0 (12.7)	94.0 (9.7)	-
	Specificity (%)	98.8 (1.3)	99.0 (1.2)	96.8 (2.1)	-

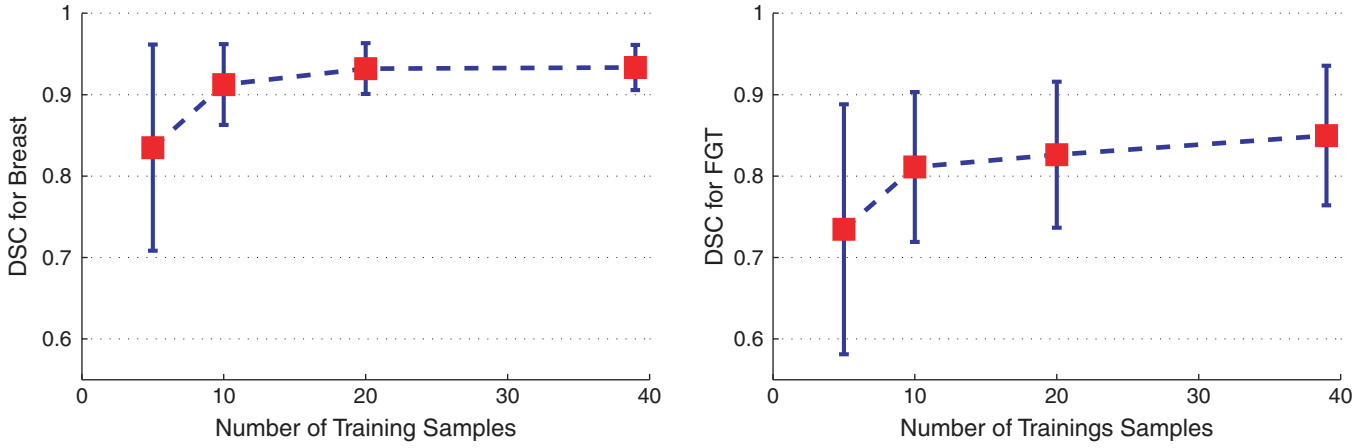

FIG. 3. DSC values obtained when 3C U-net was trained with varying number of volumes in each fold. [Color figure can be viewed at wileyonlinelibrary.com]

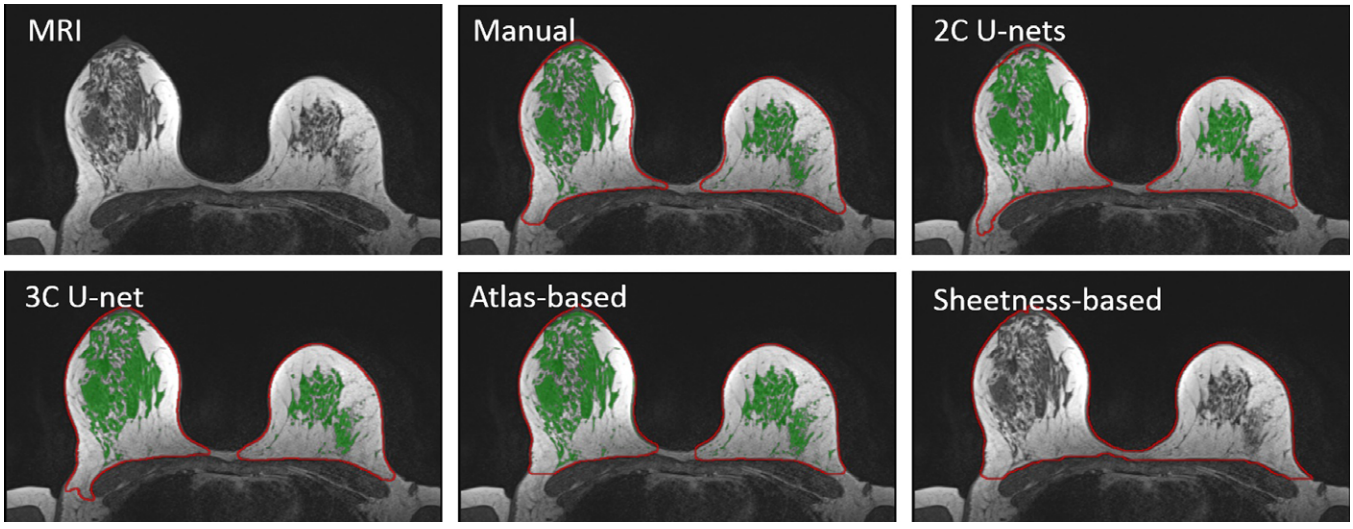

FIG. 4. A breast MRI, which was segmented accurately by all automated methods. [Color figure can be viewed at wileyonlinelibrary.com]

boundary at the right breast. Regarding the atlas-based segmentation approach, although breast segmentation was not affected by bias field, most of the fat voxels were wrongly classified as FGT because of the same artifact.

In the example shown in Fig. 7, ghosting artifacts severely affected atlas-based segmentation. Additionally, unusual breast and body shape affected both atlas-based and sheetness-based methods by making them cut the breast boundary at a higher

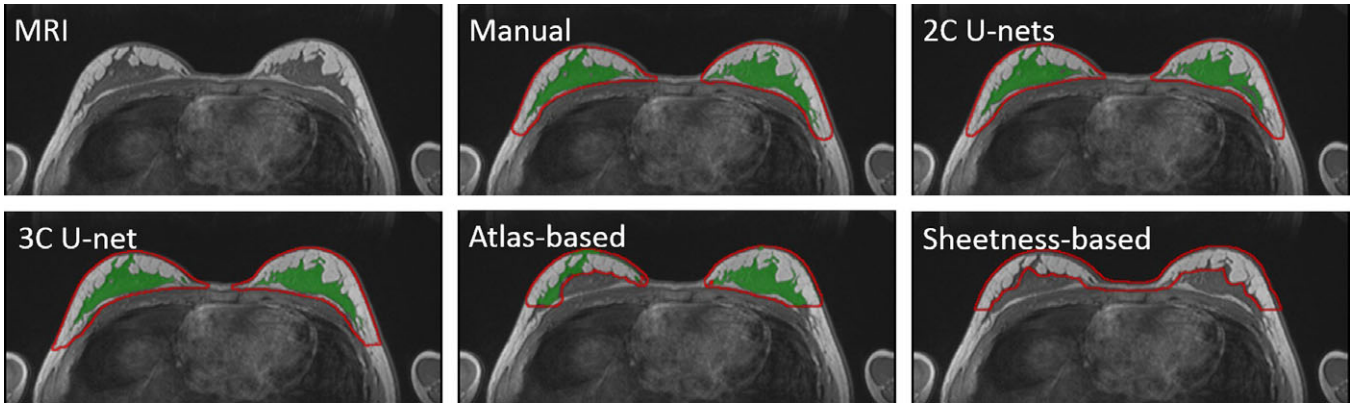

FIG. 5. Segmentation for a dense breast (category 4). [Color figure can be viewed at wileyonlinelibrary.com]

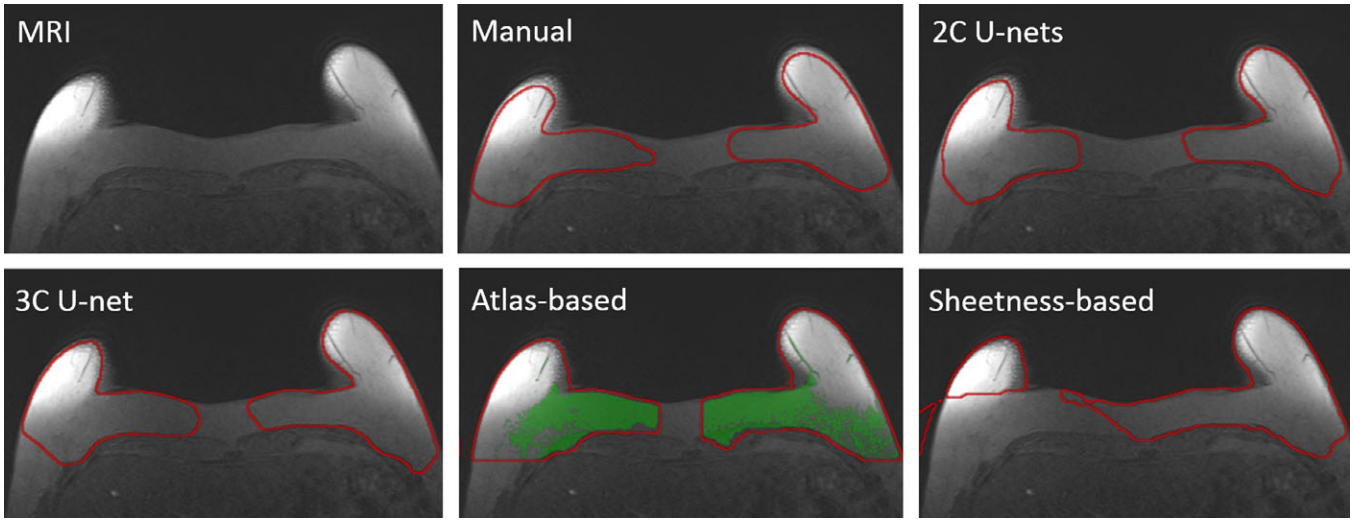

FIG. 6. An example to illustrate the effect of a bias field on segmentation. [Color figure can be viewed at wileyonlinelibrary.com]

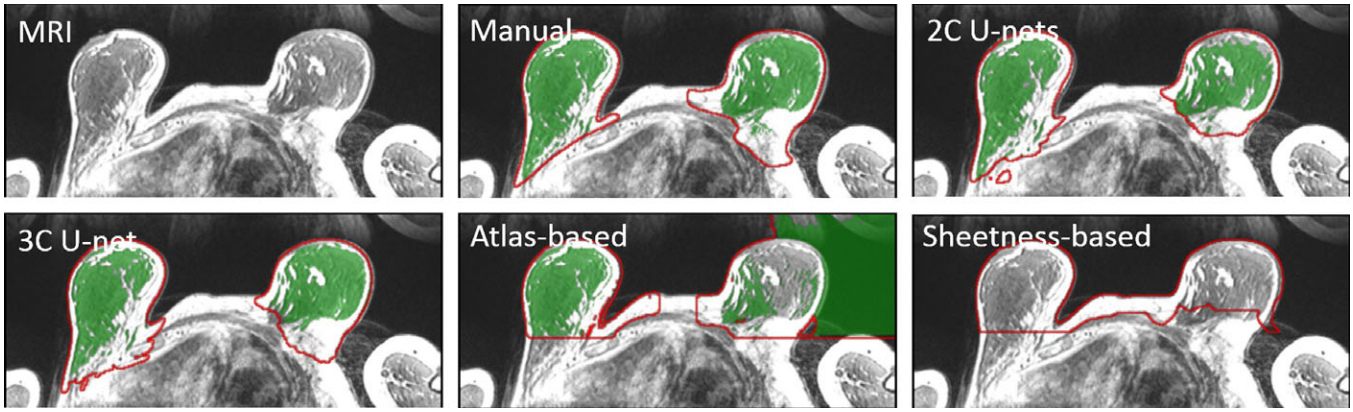

FIG. 7. An example to illustrate the effect of ghosting artifact on segmentation. [Color figure can be viewed at wileyonlinelibrary.com]

level than desired. U-net-based methods were not affected by ghosting artifacts, and breast-body boundaries were affected minimally. In the other example given in Fig. 8, the shape of the breast was usual compared to the previous case, but strong ghosting effects still caused problems in both atlas-based and sheetness-based methods for segmenting the breast.

In most of the cases, the image quality is low in caudal and cranial ends of the MRI volume, which causes

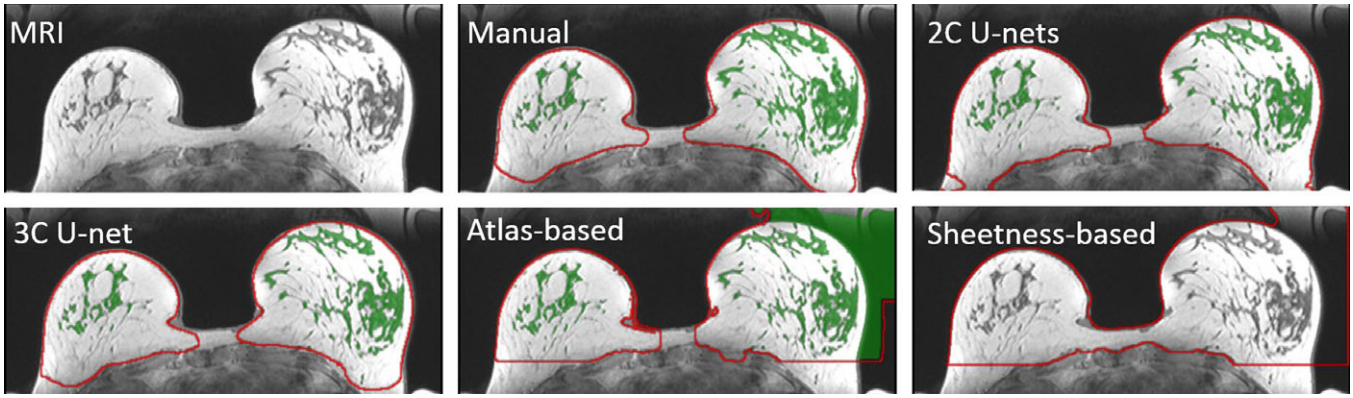

FIG. 8. An example to illustrate the effect of ghosting artifact on segmentation. [Color figure can be viewed at wileyonlinelibrary.com]

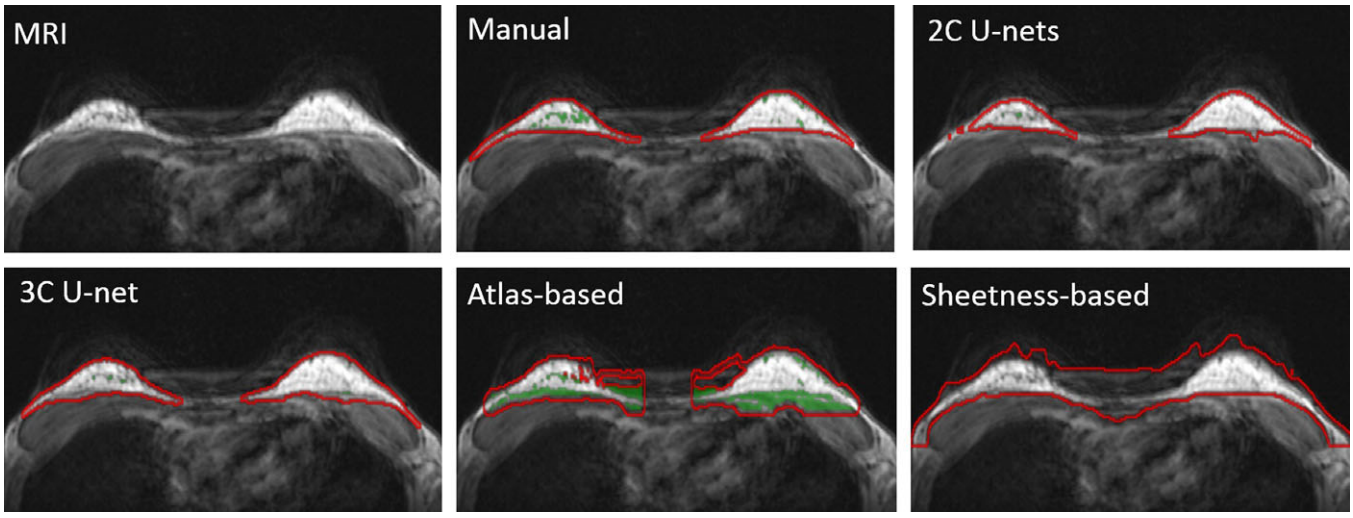

FIG. 9. An example to illustrate the effect of low image quality at the ends of the MRI volume. [Color figure can be viewed at wileyonlinelibrary.com]

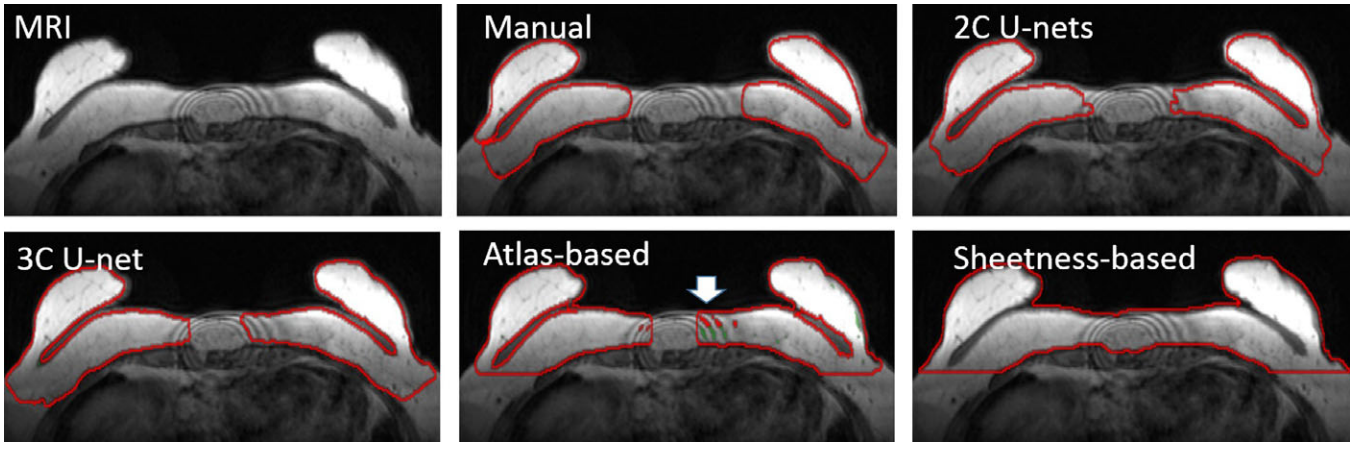

FIG. 10. An example where skin folds of the breast and zebra artifacts are present in the image. [Color figure can be viewed at wileyonlinelibrary.com]

difficulties for segmentation algorithms. This is demonstrated in the example given in Fig. 9, where U-net-based methods showed relatively more robust performance compared to atlas-based and sheetness-based methods.

Figure 10 illustrates an example where skin folds are present. Not only U-net based methods, but also atlas-based method was able to exclude skin folds in this example. This example also shows an MRI artifact known as "zebra" or

"Moire" artifact,[33] which had a slight effect on the atlas-based method by making darker pixels classified as FGT (see area pointed by white arrow in Fig. 10).

3.C. Measurement of breast density

Correlation of the breast density values obtained from manual segmentations to the values obtained from 2C U-nets, 3C U-net, and atlas-based methods were 0.957, 0.974, and 0.938, respectively. In this metric, 3C U-net method had a significantly better performance compared to 2C U-nets and atlas-based methods ($P < 0.0001$ and $P = 0.0016$, respectively). The difference between the correlation values obtained from atlas-based and 2C U-nets methods was not statistically significant ($P > 0.1$). Bland–Altman plots in Fig. 11 show that there was a positive bias of 3.8 % in breast density measurements obtained from the atlas-based method. The biases in 2C U-nets and 3C U-net methods were −0.8 % and −1.5 %, respectively. According to the same plots, the values for LOA were smallest in 3C U-net method (5.88) and largest in atlas-based method (7.84).

Both of the U-net methods strongly disagreed with manual segmentation in one case (indicated by arrows in Fig. 11) in breast density measurement. Manual segmentation resulted in a density value of 42%, while the values obtained from 2C U-nets and 3C U-net methods were 22% and 23%, respectively. An example slice from this case is provided in Fig. 12.

4. DISCUSSION

In this study, we investigated the use of deep-learning methods, in particular U-net, for breast and FGT segmentation. We explored two different approaches. The first method does breast and FGT segmentation in two consecutive steps using 2 U-nets (2C U-nets). The second method performs this simultaneously in a 3-class U-net (3C U-net). We collected a challenging dataset that covers a large time period including variations in MRI acquisition protocol, in addition to

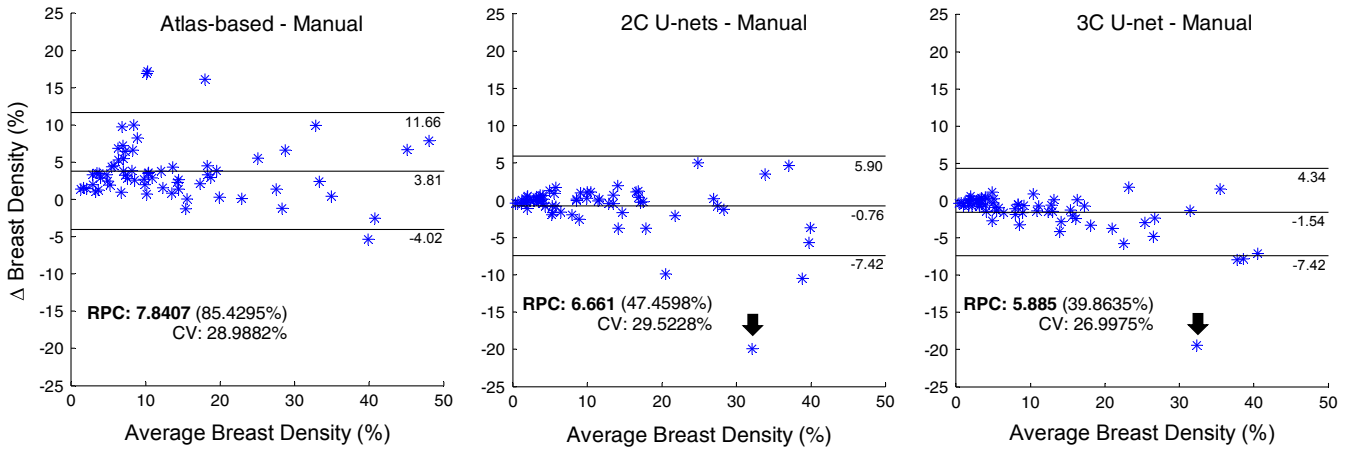

FIG. 11. Bland–Altman plots for breast density values obtained from automated segmentations with respect to the values obtained from manual segmentations. Solid lines show the average values of the differences in breast density measurements. Dashed lines show the values at a distance of 1.96 times the standard deviation to the mean value. Arrows point the case with the most disagreement between estimations using manual and U-net segmentations. This case is shown in Fig. 12. [Color figure can be viewed at wileyonlinelibrary.com]

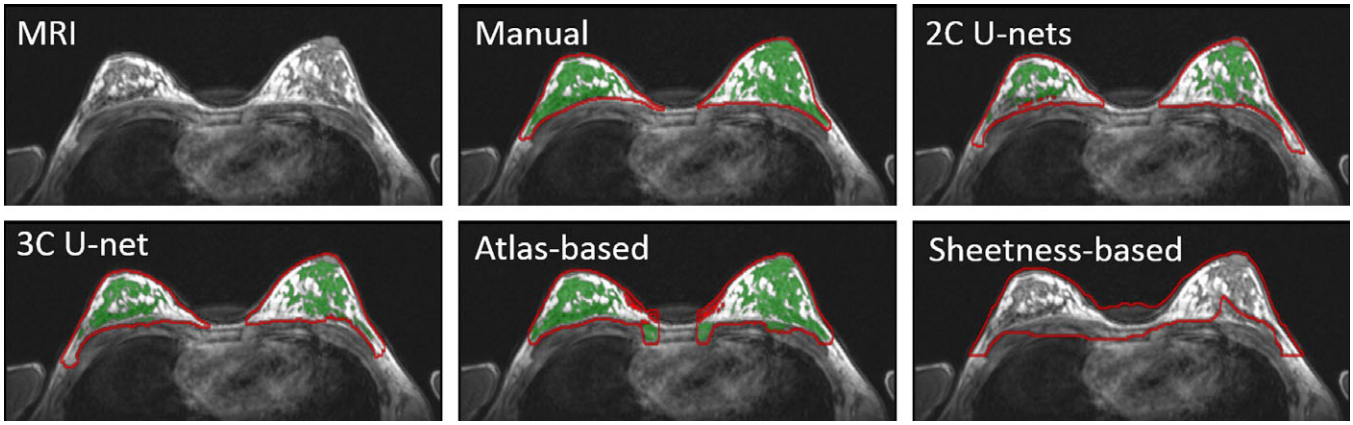

FIG. 12. A slice from the MRI for which the strongest disagreement occurred between the density values obtained from U-net segmentations and manual segmentation. [Color figure can be viewed at wileyonlinelibrary.com]

variations in breast density. The two presented deep-learning-based methods were applied to this dataset as well as two other previously published algorithms for comparison.

One of the relevant clinical applications of breast and FGT segmentation is to compute breast density. Our results show that 3-class U-net performed significantly better than two-stage U-nets method and atlas-based method in determining breast density, by achieving a correlation value of 0.974 with respect to the density values obtained by manual segmentations. Bland–Altman analysis showed that the LOA was largest for the atlas-based method, which means that the average difference between the density values obtained from manual and atlas-based segmentations tended to be larger than when computed with the other methods. This finding is in accordance with the correlation values computed. According to the same analysis, atlas-based method had a tendency to overestimate breast density, while a slight tendency in the opposite direction was observed in U-net-based methods. This effect was the strongest in the case shown in Fig. 12. It can be seen that FGT was under-segmented by U-net-based methods in this case. Additionally, breast region was also defined as larger by the U-net methods compared to the manual segmentation. These together led to the least accurate breast density estimation with U-net in our dataset.

According to the DSC values we obtained, U-net based methods outperformed both of the existing methods by a large margin as given in Tables II and III. This holds for all breast density categories as given in the same tables. In the examples given in section 3.B., we demonstrated how different variations in breast images caused problems for traditional segmentation techniques, while U-net-based methods were relatively more stable against these variations. U-net-based methods were less affected by the obscured pectoral muscle boundaries in dense breasts, compared to the traditional approaches, which is also reflected in DSC values. U-net methods were also minimally affected by MRI artifacts like ghosting effects, while these artifacts troubled both atlas-based and sheetness-based segmentation methods. Another important observation is that U-net-based methods were minimally affected by intensity inhomogeneities, although no bias-field correction was applied as a prior step. This indicates that U-net was able to learn the bias field in our dataset. According to our observations in the segmentation results, U-net-based methods were relatively more robust to low image quality, which often occurs at caudal and cranial ends of an MRI volume. Furthermore, U-net approaches were also capable of learning to exclude skin folds from the segmentations they output. Another advantage of using U-net over traditional approaches was regarding the definition of the extent of the breast. Traditional approaches determine breast area using a distance criteria to anatomical landmarks, such as the sternum, cutting the resulting segmentation at certain locations. However, as seen in examples given in Figs. 7 and 5, this does not generalize well to all breasts. We observed that U-net-based methods had the flexibility to learn to mimic the human annotator using the training examples.

We observed that the performances of U-net- and sheetness-based algorithms were relatively stable across different MRI acquisition protocols, while the most remarkable difference occurred when the atlas-based method was used for breast segmentation with MRI scans obtained from the Protocol 2. The DSC value in this set was 0.812, while it was 0.88 for other MRI protocols. This might be related to the strong ghosting artifacts, as both of the examples given in Figs. 7 and 8 are images obtained with this protocol and we have not observed a similar case with the other protocols. However, the difference between the DSC values in this set and the DSC values obtained from other MRI protocols did not reach statistical significance level ($P = 0.061$). Therefore, further studies are needed to reach a final conclusion on how different MRI protocols and systems affect such segmentation algorithms.

Comparing the two U-net-based methods, we found that 2C U-nets method performed better in breast segmentation by 1.1% in DSC values. This result may be expected, as in the former approach the whole network is dedicated to the breast segmentation task, while in the latter approach, the same network is performing two tasks simultaneously. However, the performance of the 3C U-net was better in FGT segmentation. The poorer performance of the two-stage approach in FGT segmentation might be related to the propagation of the errors of the breast segmentation stage to the FGT segmentation stage. Although breast segmentations obtained from two-stage approach are slightly better than those obtained from the 3-class network, when an under-segmentation occurs in this first stage, there is no way to recover this error in the second stage because the final result is masked by the result of breast segmentation. Such under-segmentation errors directly cause FGT to be under-segmented as well. There is no such error-propagation problem in 3-class network, as breast segmentation and FGT segmentation is learned and applied simultaneously by the same network in this approach. This is also reflected, as explained above, in the superior performance of 3C U-net in determining breast density.

We used 39 training volumes in each cross-validation fold. However, we observed that breast segmentation did not improve any further when we increased the number of training volumes from 20 to 39. On the other hand, a significant DSC increase was observed for FGT segmentation when the number of volumes used for training was incremented from 20 to 39. This result might be related to the fact that FGT segmentation is more vulnerable to a variety of MRI artifacts, thus it benefited more from increasing the number of examples.

Most of the U-net parameters used in this study are based on the parameter values reported by Ronneberger et al.[26]. However, we have introduced some modifications such as padded convolutions and the weighting factor in the loss function of the 3C U-net, which was empirically determined. Furthermore, in this study, rather than splitting the study dataset into single training, validation and tests sets, we followed a threefold cross-validation strategy to evaluate the presented methods. Following this approach, we were able

to report the performance of the studied algorithms for the whole study dataset. We observed that a common training procedure with equal hyperparameters, in particular the learning rate, did not always lead to the most optimal solution. Therefore, we tuned the learning rate for each fold based on the changes in performance on the corresponding validation set. The final networks were selected based on performance on the validation sets and finally applied to the "unseen" data (i.e., test sets). Hyperparameter tuning is a known problem and research topic that is currently being investigated.[34–36]

Our study had some limitations. Although we used MRI scans obtained by different MRI protocols in this study, all scans were obtained in MRI units of Siemens, in the same hospital, and all of them were non-fat-suppressed images. In the future, we aim at applying and evaluating the presented methods on a multicenter and multivendor dataset. Furthermore, we did not investigate how inter-reader differences affect manual segmentations. Training deep convolutional networks (such as U-nets) require a decent amount of data, but preparing manual breast segmentation for a case is a tedious and time-consuming task. Note that the full process for manually segmenting breast and FGT may take more than 45 min per MRI volume. For the sake of being able to generate a large dataset, we used single annotation per case in this study; therefore, we were not able to investigate inter-reader variability. Lastly, although we are dealing with volumetric data, we used a two-dimensional approach. To improve the algorithms, in the future, we will explore the use of 3D convolutions. We expect this to be beneficial especially for the breast segmentation task in which depth information might be important.

In conclusion, we investigated using deep-learning methods for breast and FGT segmentation in a challenging dataset that includes many variations in terms of MRI acquisition and breast density. We based our methods on U-net architecture, and we compared them to two of the existing methods in terms of their segmentation performances. U-net-based deep-learning methods outperformed two of the traditional methods in our dataset in both tasks: breast segmentation and breast density computation.

ACKNOWLEDGMENTS

This work was funded by the European 7th Framework Program grant VPH-PRISM (FP7-ICT-2011-9, 601040).

DISCLOSURES

Nico Karssemeijer is co-founder and shareholder of Volpara Solutions Ltd. (Wellington, New Zealand), co-founder and shareholder of QView Medical Inc (Los Altos, CA), and co-founder and shareholder of QView Medical ScreenPoint Medical BV (Nijmegen, the Netherlands).

[a]Author to whom correspondence should be addressed. Electronic mail: mehmet.dalmis@radboudumc.nl

REFERENCES

1. Behrens S, Laue H, Althaus M, et al. Computer assistance for MR based diagnosis of breast cancer: present and future challenges. *Comput Med Imaging Graph*. 2007;31:236–247.
2. Yao J, Chen J, Chow C. Breast tumor analysis in dynamic contrast enhanced MRI using texture features and wavelet transform. *IEEE J Sel Top Signal Process*. 2009;3:94–100.
3. Gubern-MTrida A, Martf R, Melendez J, et al. Automated localization of breast cancer in DCE-MRI. *Med Image Analy.* 2015;20:265–274.
4. Dalmis MU, Gubern-MTrida A, Borelli C, Vreemann S, Mann RM, Karsse-meijer N. A fully automated system for quantification of background parenchymal enhancement in breast DCE-MRI. In: *SPIE Medical Imaging*. International Society for Optics and Photonics; 2016:97850L–97850L.
5. Nie K, Chen J-H, Chan S, et al. Development of a quantitative method for analysis of breast density based on three-dimensional breast MRI. *Med Phys*. 2008;35:5253–5262.
6. Wu S, Weinstein SP, Conant EF, Kontos D. Automated fibroglandular tissue segmentation and volumetric density estimation in breast MRI using an atlas-aided fuzzy c-means method. *Med Phys*. 2013; 40:122302.
7. Baltzer HL, Alonzo-Proulx O, Mainprize JG, et al. MRI volumetric analysis of breast fibroglandular tissue to assess risk of the spared nipple in BRCA1 and BRCA2 mutation carriers. *Ann Surg Oncol*. 2014;21: 1583–1588.
8. Ivanovska T, Laqua R, Wang L, Liebscher V, Völzke H, Hegenscheid K. A level set based framework for quantitative evaluation of breast tissue density from MRI data. *PloS One*. 2014;9:e112709.
9. Gubern-Mérida A, Kallenberg M, Mann RM, Marti R, Karssemeijer N. Breast segmentation and density estimation in breast MRI: a fully automatic framework. *IEEE J Biomed Health Inform*. 2015;19:349–357.
10. Boyd NF, Martin LJ, Yaffe MJ, Minkin S. Mammographic density and breast cancer risk: current understanding and future prospects. *Breast Cancer Res*. 2011;13:1.
11. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiol Biomarkers Prev.* 2006;15:1159–1169.
12. Vachon CM, Van Gils CH, Sellers TA, et al. Mammographic density, breast cancer risk and risk prediction. *Breast Cancer Res*. 2007;9:1.
13. Gubern-Mérida A, Kallenberg M, Platel B, Mann RM, Martí R, Karssemeijer N. Volumetric breast density estimation from full-field digital mammograms: a validation study. *PLoS One*. 2014;9:e85952.
14. Milenković J, Chambers O, Mušič MM, Tasič JF. Automated breast-region segmentation in the axial breast MR images. *Comput Biol Med*. 2015;62:55–64.
15. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index 1: Scientific reports. *Acad Radiol*. 2004;11:178–189.
16. Koenig M, Laue H, Boehler T, Peitgen H-O. Automatic segmentation of relevant structures in DCE MR mammograms. In: *Medical Imaging*. International Society for Optics and Photonics; 2007:5141S–65141S.
17. Martel AL, Gallego-Ortiz C, Lu Y. Breast segmentation in MRI using Poisson surface reconstruction initialized with random forest edge detection. In: *SPIE Medical Imaging*. International Society for Optics and Photonics; 2016:97841B–97841B.
18. Ortiz CG, Martel A. Automatic atlas-based segmentation of the breast in MRI for 3D breast volume computation. *Med Phys*. 2012;39:5835–5848.
19. Khalvati F, Gallego-Ortiz C, Balasingham S, Martel AL. Automated segmentation of breast in 3-D MR images using a robust atlas. *IEEE Trans Med Imaging*. 2015;34:116–125.
20. Lin M, Chen J-H, Wang X, Chan S, Chen S, Su M-Y. Template-based automatic breast segmentation on MRI by excluding the chest region. *Med Phys*. 2013;40:122301.
21. Gubern-MTrida A, Kallenberg M, Martf R, Karssemeijer N. Segmentation of the pectoral muscle in breast MRI using atlas-based approaches. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2012:371–378.

22. Wang L, Platel B, Ivanovskaya T, Harz M, Hahn HK. Fully automatic breast segmentation in 3D breast MRI. In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2012: 1024–1027.
23. Giannini V, Vignati A, Morra L, et al. A fully automatic algorithm for segmentation of the breasts in DCE-MR images. In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE; 2010:3146–3149.
24. Tustison NJ, Avants BB, Cook P, et al. N4ITK: improved N3 bias correction. *Med Imaging IEEE Trans On*. 2010;29:1310–1320.
25. Razavi M, Wang L, Gubern-MTrida A, et al. Towards accurate segmentation of fibroglandular tissue in breast MRI using fuzzy c-means and skin folds removal. In: *International Conference on Image Analysis and Processing*. Springer; 2015:528–536.
26. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2015:234–241.
27. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
28. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *Aistats*. 2010;9:249–256.
29. Tieleman T, Hinton G. RMSprop gradient optimization.
30. Dubuisson M-P, Jain AK. A modified hausdorff distance for object matching. In: *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*. Vol. 1. IEEE; 1994:566–568.
31. Lee IA, Preacher KJ. Calculation for the test of the difference between two dependent correlations with one variable in common [computer software]. http://quantpsy.org/corrtest/corrtest2.htm 2013, [Online; accessed 2016-06-29].
32. Steiger JH. Tests for comparing elements of a correlation matrix. *Psychological Bull*. 1980;87:245.
33. Harvey JA, Hendrick RE, Coll JM, Nicholson BT, Burkholder BT, Cohen MA. Breast mr imaging artifacts: how to recognize and fix them 1. *Radiographics*. 2007;27:S131–S145.
34. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res*. 2012;13:281–305.
35. Klein A, Falkner S, Bartels S, Hennig P, Hutter F. Fast bayesian optimization of machine learning hyperparameters on large datasets. arXiv preprint arXiv:1605.07079 2016.
36. Andrychowicz M, Denil M, Gomez S, et al. Learning to learn by gradient descent by gradient descent. arXiv preprint arXiv:1606.04474 2016.