CONSTRUCTOR UNIVERSITY

Efficient Breast Parenchyma Segmentation in MRI through Active Learning and Sparse Annotation

by

Gajaria Anand

Master Thesis in Data Engineering

Supervisor(s):

Prof. Dr.-Ing. Markus Wenzel, Prof. Dr. Stefan Kettemann

Submission: August 29, 2023

Statutory Declaration

Family Name, Given/First Name	Anand, Gajaria
Matriculation number	20332598
Kind of thesis submitted	Master Thesis

English: Declaration of Authorship

I hereby declare that the thesis submitted was created and written solely by myself without any external support. Any sources, direct or indirect, are marked as such. I am aware of the fact that the contents of the thesis in digital form may be revised with regard to usage of unauthorized aid as well as whether the whole or parts of it may be identified as plagiarism. I do agree my work to be entered into a database for it to be compared with existing sources, where it will remain in order to enable further comparisons with future theses. This does not grant any rights of reproduction and usage, however.

This document was neither presented to any other examination board nor has it been published.

German: Erklärung der Autorenschaft (Urheberschaft)

Ich erkläre hiermit, dass die vorliegende Arbeit ohne fremde Hilfe ausschließlich von mir erstellt und geschrieben worden ist. Jedwede verwendeten Quellen, direkter oder indirekter Art, sind als solche kenntlich gemacht worden. Mir ist die Tatsache bewusst, dass der Inhalt der Thesis in digitaler Form geprüft werden kann im Hinblick darauf, ob es sich ganz oder in Teilen um ein Plagiat handelt. Ich bin damit einverstanden, dass meine Arbeit in einer Datenbank eingegeben werden kann, um mit bereits bestehenden Quellen verglichen zu werden und dort auch verbleibt, um mit zukünftigen Arbeiten verglichen werden zu können. Dies berechtigt jedoch nicht zur Verwendung oder Vervielfältigung.

Diese Arbeit wurde noch keiner anderen Prüfungsbehörde vorgelegt noch wurde sie bisher veröffentlicht.

29.08.2023

Date, Signature

Abstract

The thesis aims to seamlessly integrate active learning coupled with sparse annotation into the segmentation process of breast imaging, specifically focusing on segmenting fibroglandular tissue. This workflow endeavors not only to help annotators in reducing the annotation efforts but also tries to enhance the efficiency of the segmentation model. Active learning is an iterative process that optimizes the learning efficiency of a model. It makes this possible by selectively querying the most informative samples on each iteration that can be used in successive iteration. Sparse annotation allows annotators to focus only on rectifying substantial errors on the selected samples. This results in an efficient and swift annotation process. Analysing and Reporting the amount of fibroglandular tissue present in the breast plays an important role in the breast cancer risk assessment. The quantitative assessment of its presence is done by using image processing and deep learning techniques. However, a conventional deep learning method often demands a high volume of annotated data for optimal performance. Furthermore, collecting a large set of high-quality labelled data is difficult in medical imaging because, there are some challenges such as subjectivity, complexity, time consumption and potential inconsistencies associated with the manual annotation process. The workflow proposed in this study seeks to address the challenges encountered by the annotators during the breast imaging annotation process, and strives to develop a proficient model while minimizing the reliance on extensive labeled data. The results of this approach were compared to a baseline model, trained on full set of training data. Using the active learning approach a similar dice score of 0.64 as the baseline model, was achieved in five active learning iterations.

Contents

List of Figures

List of Tables

List of Abbreviations

1. **MRI:** Magnetic Resonance Imaging
2. **FGT:** Fibroglandular tissue
3. **ROI:** Region of Interest
4. **CCE:** Categorical Cross Entropy
5. **VLEC:** Volume of Largest Error Component
6. **D-VLEC:** Dilated Volume of Largest Error Component
7. **mm:** Millimetre
8. **ml:** Millilitre
9. **ASD:** Average Surface Distance
10. **AL:** Active Learning
11. **VD:** Volume Difference

1 Introduction

Breast density is a well-known factor that influences the probability of developing breast cancer. Generally, the measurement of its density is carried out on a 2D mammography. There are various volumetric analysis tools available that can attempt to predict the volume of dense tissue from the 2D projections. It has been observed that they fail to accurately measure the density in women with dense breasts. This happens because, the glandular and fatty tissue present in the breast, intersect in the 2D projection. As a result, it becomes challenging to distinguish them [18].

Breast MRI has the capability to provide high-resolution 3D images with distinctive tissue contrast. It has also been observed in various studies that, there is a similarity between the proportion of fibroglandular tissue calculated in MRI when compared with mammographic breast imaging. From this, it can be said that, MRI can possibly be used in quantifying the percentage of fibroglandular tissue (FGT%) present in the breast, and may also produce better results [15].

Fibroglandular tissue segmentation is the fundamental step required to quantify the density of the tissue. However, this process is challenging in several aspects. First, there is no fixed location of this tissue, it can be in varying amounts and appearances. Second, There is no such information or clues present in the MRI that can help in recognizing the tissue [18] .

In recent years, deep learning algorithms based on U-Net have been widely utilized to develop an accurate FGT segmentation model, that can aid in addressing the problems discussed above. However, there are some limitations associated with the conventional approach of training this deep learning models. First, the demand for a large volume of fully annotated data to achieve reliable results. Second, manual annotation of medical images is a complex, costly, and time-consuming process. Only experts can annotate the complex instances in these images effectively, and often there are too many instances to annotate [16], resulting in having less amount of fully and precisely annotated data. Due to this, the performance of the deep learning model is negatively affected [7].

The purpose of this study is to integrate active learning with sparse annotation into the breast imaging segmentation process. This workflow not only helps to reduce the annotator's efforts in annotating fibroglandular tissue, but also tries to tackle the demand for extensive labeled data.

2 Background

The proliferation of breast cancer has emerged as a significant healthcare concern globally. It is the most common type of malignancy in women [6]. Approximately 12% of women suffer from this disease during their life in the USA and European countries [5]. There is no definite reason for the occurrence of breast cancer. However early detection of this cancer can play a pivotal role in the treatment and control of the disease.

Fibroglandular Tissue: Explanation and Importance in Breast Health Assessment

Fibroglandular tissue is an essential element within the breast structure. It is a blend of fibrous connective tissue (the stroma) and the functional(or glandular) epithelial cells that form the lining of the breast ducts, commonly known as the parenchyama [4].

The fibrous tissue provides structural support to the breast, imparting stability and shape. This eventually contributes to the overall firmness of the breast. The Glandular Tissue is responsible for milk production. This tissue includes milk ducts and lobules. The balance between fibrous and glandular tissue varies among individuals and changes over time.

The quantification of fibroglandular tissue within the breast plays an important role in assessing the risk of breast cancer. Numerous studies have found a positive correlation between fibroglandular tissue density and breast cancer risk. Women with fibroglandular tissue density exceeding 60% of total breast volume are estimated to have a higher risk of developing breast cancer. Whereas Women constituting less than 5% of total breast volume appear to have reduced risk [14]. This insight into the relationship between fibroglandular tissue density and breast cancer risk assists in tailoring the screening and diagnostic approaches, which can enhance the possibility of early detection and optimal care.

Different Methods of Breast Imaging

- Mamography:

  Description: A mammogram is a 2D image used to identify morphologically suspicious findings in breast cancer. These findings can include masses, asymmetrical clarifications and deformed breast areas. In this approach 2D radio graphic images are produced by penetrating low-energy X-rays through the tissues [6]. Detecting fibroglandular tissue in a mammogram is difficult because of overlapping tissues and high attenuation of X-Ray by fibroglandular tissue [10]..

- Magnetic Resonance imaging(MRI):

  Description: MR Imaging of the breast involves the use of a strong magnetic field and radio frequency pulses. It has the capability to produce high-resolution 3D images which can help in a thorough evaluation of fibroglandular tissue distribution and its interaction with surrounding structures. However, the disadvantages of using MRI for breast imaging is its high cost and scanning time [10].

3 Problem Statement and Motivation of Research

The task of annotating data in the medical field is a pivotal issue that reverberates across the healthcare domain. Manual annotation of anatomical structures is a fundamental step in training a deep learning algorithm for segmentation tasks. However, this process is labor-intensive and time-consuming. The possible reasons behind this are the intricacies of the nature of the data and the size of the data. The aim is to address these challenges, paving the way for harnessing the complete capabilities of medical imaging data which can ultimately contribute to enhancing medical analysis and diagnoses. This can be achieved by leveraging the principles of active learning and sparse annotation. Active learning on the one hand by selecting the most informative samples optimizes the training process and on the other hand, sparse annotation further streamlines the process by guiding the annotators to focus on correcting the most critical errors. This combined approach not only enhances the annotation efficiency but also contributes to improving the accuracy of the deep learning model. The integration of these methods can improve the segmentation task of fibroglandular tissue present in the breast. The primary idea is to develop a comprehensive end-to-end active learning architecture that not only streamlines the intense task of fibroglandular tissue annotation but also simultaneously enhances the accuracy of segmentation outcomes. In summary, The relationship between active learning and annotation refinement holds the potential to make the process of fibroglandular tissue segmentation more efficient, paving the way for more precise breast cancer risk assessment and improved medical decision-making.

4 Methodology

4.1 Data Description

The imaging process in breast care involves clinical institutions using different types of medical imaging devices provided by a variety of vendors to obtain crucial diagnostic data. These scanners are designed to capture detailed images of the breast tissue using advanced imaging techniques[11]. Two of the most widely used modalities in breast imaging are mammography and magnetic resonance imaging (MRI).

The data used in this study is a proprietary dataset provided by four clinical institutions employing scanners from two distinct vendors, namely Siemens Healthcare and GE Medical Systems. It consists of 3D Volumetric breast MRI scans that were taken with dynamic contrast enhancement, however, in this study, only pre-contrast time point is considered. Overall, the dataset holds 52 3D MRI scans with corresponding reference fibroglandular tissue masks. The reference breast masks were created by a radiological technologist as per the guidance and instructions given by a radiologist. They perform annotation on breast images by carefully scrutinizing the region of interest (ROI). For the current experiment, the region of interest is the outline of fibroglandular tissue present in the given breast image volume. The dataset was divided into three subsets: 32 randomly selected images were used for training, 7 for validation, and 13 as model evaluation set.

This study endeavors to carry out experiments with 2D images, 2D image slices were extracted from the 3D images coming from the source data pool. This detailed operation of how this process is carried out and its utilization is explained in the section 5.

Images	Resolution	Voxel Size [mm^3]	Vendor
20	352 * 352 * 160	1.02 * 1.02 * 1.2	Siemens
15	512 * 512 * 46	0.8 * 0.8 * 2.5	GE
10	896 * 896 * 112	0.4 * 0.4 * 1.8	Siemens
3	512 * 512 * 46	0.7 * 0.7 * 2	GE
1	384 * 384 * 35	0.9 * 0.9 * 5.6	Siemens
1	512 * 512 * 55	0.6 * 0.6 * 3.6	Siemens
1	512 * 512 * 46	0.8 * 0.8 * 2.5	GE
1	512 * 512 * 46	0.6 * 0.6 * 2.5	GE

Table 1: Summary of the dataset from the two institutions

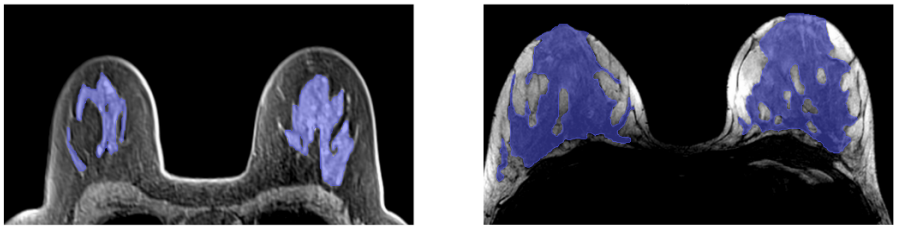


Figure 1: Two example images with corresponding Reference masks

The images above are taken from the training dataset, where the blue coloured contours on the breast image represent the ROI.

4.2 Active learning Workflow Overview

The performance of the deep learning model depends on the amount of available labeled data. However, annotating a substantial pool of medical images presents a series of difficulties to the annotators. One of the prominent challenge is, from a large pool of unannotated images, the annotator has to select which images to annotate, as it is impossible to annotate all the available images. This is due to the inherent complexities and unique attributes of medical images [19]. To address these issues that are associated with the traditional approach of training a deep learning model, the integration of Active learning emerges as a promising solution [3].

Active learning is a machine learning approach that iteratively tries to select the most informative samples in each training iteration. By leveraging this strategy, this study aims to improve the model's performance using fewer but more insightful data points. Based on this principle a workflow has been designed which aligns with the characteristics of the data and the specific goal that I am trying to achieve.

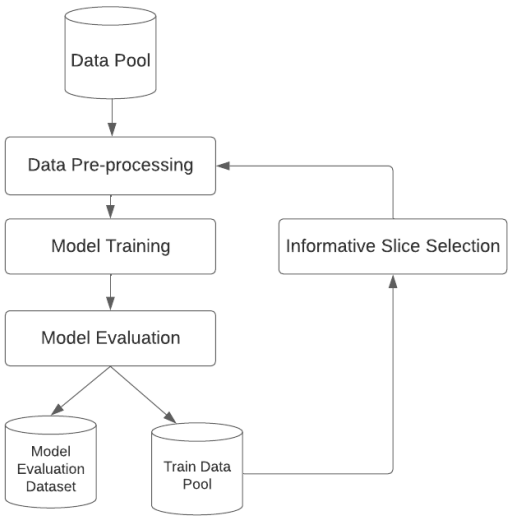


Figure 2: Basic Active learning Workflow

As can be observed from the Figure 2, the flowchart commences with the data pool component, which consists of all the 3D MRI Images. Comprehensive information regarding this component is described in the subsection 4.1. Following to this, the data pre-processing step comes into play. This step retrieves the images from the data pool and splits them into train, validation and model evaluation. Additionally, it also does some crucial pre-processing on the data which will be elaborated in the section 5.

Once the data is prepared, it undergoes the first iteration of the active learning. During this phase, a 2D U-Net model is trained using the training data and validated by employing the validation data. The subsequent component is model evaluation. In this step, the model evaluation data and the training data are evaluated based on the predefined evaluation metrics. Following to that the data used in the previous training iteration is filtered out and the remaining data is passed to the succeeding component.

The heart of this study lies in the informative slice selection component, where various methods are employed to identify and select the most erroneously predicted image slices with respect to the reference standard. These selected slices, combined with the previously used training image slices are utilized in training the model for the successive Iteration. A detailed explanation of these methods is described in the subsection 5.2. Out of these methods, few of them are based on the concept of sparse annotation. The basic principle of sparse annotation can be understood from the subsection 4.3.

The entire process constitutes an iterative approach, wherein each run aims to identify the most insightful image slices that can help in improving the model's performance. Consequently, the efforts of the annotators are significantly reduced, as they only need to annotate the most critical region in the selected image slice instead of the whole image slice.

In total five active learning iterations were conducted using each of these methods. The performance of each active learning approach was compared with a baseline model. This baseline model was trained using the full training data, and keeping the same model configurations as used while performing active learning.

4.3 Sparse annotation

Sparse annotation refers to an annotating strategy where only a subset of the whole image is manually annotated or labeled, while the rest of the data is left unannotated or unlabeled. This method is used mostly in deep learning tasks of image segmentation and object detection. The rationale behind using sparse annotation is that the annotator does not have to correct every small unimportant errors, but can focus only on correcting the large significant errors. This can reduce the time and effort associated with manual labeling and eventually assist in achieving effective model training [19]. In the context of my study, the slices selected on the basis of the largest error component, are assigned a label, while the rest of the image is left unlabeled. The details on how this strategy is implemented in this study can be found in the section 5.

4.4 2D U-Net Model Architecture

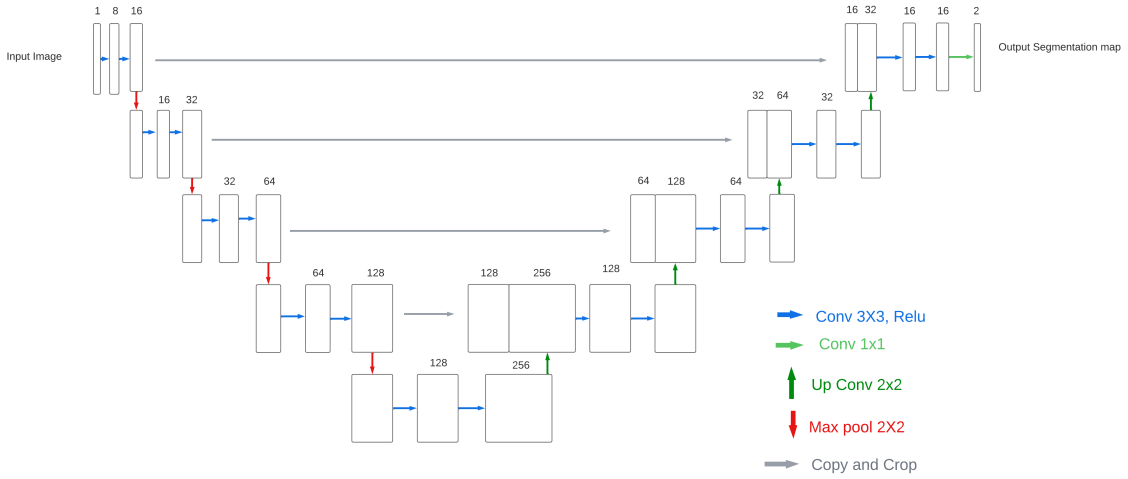


Figure 3: U-Net architecture[13]

The diagram shows multi-channel feature maps with their channel number indicated on top of each feature map.

As illustrated in the above Figure 3, a 2D U-Net architecture adopts an encoder-decoder design. The leftmost part of the architecture applies a series of operations to compress the input data[13]. This part consists of two consecutive 3X3 padded convolutions, followed by a rectified linear unit (ReLU) activation. A 2X2 max pooling operation with a stride of two is also performed on this part for downsampling. At each downsampling step, the number of feature channels are doubled, which facilitates the extraction of higher-level features[13].

Conversely, the rightmost part of the architecture performs the upsampling of the feature map using 2X2 up-convolution. This operation halves the number of feature channels at each step[13]. Resulting in the enhancement of the information flow while also preserving the spatial details. This expanding path concatenates the feature map with the corresponding feature map from the above-mentioned contracting part. As padded convolution is used in the contracting part, the size of the feature map remains of the same size. As a result cropping of the feature map is not required.

Within the contracting and expanding part, two 3X3 convolutions, each followed by a ReLU activation, are performed. These operations are performed to further enhance feature representations[13].

At the final layer, a point-wise convolution is employed to map each feature vector to the desired number of classes. This makes it possible for the U-Net to generate segmentation maps with high precision and accuracy. In total, the architecture consists of 5 levels, allowing for the extraction of intricate features and facilitating the model's ability to capture complex patterns in the input data[13].

To ensure that images of different sizes can be effectively processed by a 2D U-Net model, a "unify sizes" preprocessing step is applied. This function takes the maximum extent of

the X and Y dimensions of the image. Then it crops the image and resizes it to a specific dimension. In this study, the image is resized to have dimensions that are multiple of 16.

The architecture also incorporates batch normalization, applied after each convolution layer. This helps in accelerating the convergence of the model. Furthermore, a dropout rate of 0.25 is used to prevent overfitting and improve generalization.

4.5 Loss functions used in model training

4.5.1 Dice Loss

The Dice loss is a loss function widely used in training neural networks for image segmentation tasks. It is derived from the Dice coefficient, an explanation of the Dice coefficient can be seen in the subsection 4.7, and is used to optimize the neural network's performance [8].

The formula to calculate Dice Loss is given by:

$$\text{Dice Loss} = 1 - \frac{2 \sum_i p_i \cdot g_i}{\sum_i p_i^2 + \sum_i g_i^2}$$

Where:

$$p_i : \text{predicted probability of class } i$$
$$g_i : \text{ground truth probability of class } i$$

As discussed in subsection 4.7, optimal Dice values tend to approach 1, while effective model training requires low loss values for proper weight adjustments during backpropagation. Consequently, a loss value nearing 0 signifies favorable model performance. The equation above suggests a negative relationship between Dice score and Dice loss. As the dice loss reaches its peak, it aligns with a loss of 0, indicating potentially optimal model performance.

4.5.2 Categorical Cross-Entropy Loss

The Categorical Cross-Entropy Loss also known as Softmax Cross-Entropy Loss, assists in quantifying the difference between between predicted and ground truth pixel-wise class probabilities.

The CCE is calculated as follows:

$$L(y, \hat{y}) = -\sum_i y_i \cdot \log(\hat{y}_i)$$

Here:

- y represents the true probability distribution of the classes.

- $\hat{y}$ represents the predicted probability distribution of the classes from the model's output.
- y_i and $\hat{y}_i$ are the corresponding elements of the true and predicted distributions for class i.

In this study, a combination of Dice loss and CCE are used to train the segmentation model effectively. The influence of both of these loss function is kept equal. The rationale behind giving equal importance is, to encourage a balance between accurate segmentation and proper class probability distribution prediction.

The combined loss function can be defined as follows:

$$\text{Combined Loss} = \text{CCE Loss} + \alpha \times \text{Dice Loss}$$

Here:

- α is a balancing factor that controls the trade-off between the two loss components.

4.6 Weight based model training

When slice selection methods related to sparse annotation were employed, at that time, along with the original image slice and reference mask, a weight mask was also utilized. This weight mask helps the model to focus on specific areas where the weight is assigned as 1. During the training process, the model calculates the loss based on the weight assigned to that voxel. This means that, when the value of the weight is 0 the model gives zero importance to the learning from that voxels.

The combined loss with weighted mask is calculated as follows:

$$\text{Combined Loss} = \frac{1}{N} \sum_{i=1}^{N} w_i \left(\alpha \times \text{Dice Loss}_i + (1 - \alpha) \times \text{CCE Loss}_i\right)$$

Where:

$$\begin{aligned} N &: \text{Total number of voxels in the image or batch} \\ w_i &: \text{Weight assigned to voxel } i \text{ (0 or 1)} \\ \alpha &: \text{Weight coefficient between Dice Loss and CCE Loss} \\ \text{Dice Loss}_i &: \text{Dice Loss for voxel } i \\ \text{CCE Loss}_i &: \text{CCE Loss for voxel } i \end{aligned}$$

4.7 Evaluation Metrics

There are various evaluation metrics that have been used in this study. These metrics help in analysing the model performance as well as to find the most informative image slices for the successive run in the active learning cycle.

The metrics used are Dice score, Average surface distance, Volume difference and the volume of the largest connected component. These are commonly used in medical image segmentation tasks. This section will cover a brief explanation of all these metrics.

4.7.1 Dice Score

The Dice score is also known as the Dice coefficient or Dice Similarity coefficient. It is a commonly used metric to evaluate the performance of an image segmentation algorithm. It measures the overlap between the predicted mask and the reference mask[12].

The formula to calculate the Dice score is given by:

$$\text{Dice coefficient} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{1}$$

Where:

- A contains the elements of the predicted segmentation.
- B contains the elements of the ground truth segmentation.
- $|A \cap B|$ represents the size of the intersection between sets A and B.
- $|A|$ and $|B|$ represent the sizes of sets A and B respectively.

The Dice score ranges from 0 to 1. A score of 1 indicates perfect overlap between the predicted and reference mask, whereas a score of 0 indicates no overlap at all.

In medical image segmentation, the Dice score is a majorly used evaluation metric, as it provides a measure of the algorithm's accuracy in capturing the true extent of the segmented anatomical structures. In this study, dice score is also used as one of the methods to select the most informative slices for the successive active learning runs.

4.7.2 Average Surface Distance

The Average Surface Distance(ASD) measures the average distance between the points on the surface of the predicted mask and the nearest points on the surface of the reference mask[12].

$$\text{Average Surface Distance}(A, B) = \frac{1}{N_A + N_B} \left(\sum_{a \in A} \min_{b \in B} d(a, b) + \sum_{b \in B} \min_{a \in A} d(a, b) \right) \tag{2}$$

Where:

- $A = \{a_1, a_2, \ldots, a_{N_A}\}$ is the set of points representing Surface A with N_A points.
- $B = \{b_1, b_2, \ldots, b_{N_B}\}$ is the set of points representing Surface B with N_B points.
- $d(a, b)$ represents the distance between points a and b.

- min denotes the minimum operation.

Similar to the Hausdorff distance, the smaller the Average surface Distance is, the higher is the level of agreement and accuracy between the predicted and the reference mask.

4.7.3 Volume Difference

Volume difference is another essential metric used in medical image segmentation to evaluate the accuracy of the algorithm. It tries to find the level of discrepancy in volume between the segmented region in the predicted mask and the corresponding region in the reference mask[17].

The formula to calculate the volumetric difference between two datasets A and B is given by:

$$\text{Volumetric Difference} = |V_A - V_B| \tag{3}$$

Where:

- V_A is the volume of set A.
- V_B is the volume of set B.
- $|\cdot|$ denotes the absolute value.

4.7.4 Volume of the Largest Error Component

The volume of the largest Error Component measures the volume of the region where the model is making most of the error. By analyzing this, we can pinpoint areas in the predictions where the model needs to improve its performance. In the context of image segmentation, a connected component represents a group of voxels spatially connected and shares the same label. By quantifying the volume of the largest error component, we can gain insights into the regions that require further attention and learning by the model. This area or region is considered to be the most informative for the model. The unit of the volume is in milliliters.

(a) Model Prediction (b) Largest Error Component

Figure 4: In the above images blue contours represent reference mask and green contours represent Largest error component.

4.8 Framework and Modules used for Image Processing

MeVisLab is a powerful and Component-based framework for image processing research and development. The specific focus area of this platform is on medical imaging. It includes advanced modules for segmentation, volumetry and quantitative morphological and functional analysis. It is developed by MeVis Medical Solutions AG in close coordination with the research institute Fraunhofer MEVIS[1].

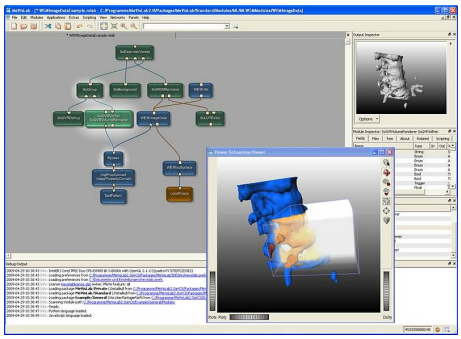


Figure 5: Example Image of MeVisLab Framework[1]

The image represents an example view of the MeVisLab Framework and how various modules can be connected to form an image processing network.

The image represents an example view of the MeVisLab Framework and how various modules can be connected to form an image processing network. There are multiple modules from MeVisLab that have been utilized in this study to do Image processing tasks at various stages. The list and a brief explanation of the modules can be seen below.

- **Bounding box module**: This module scans the input image and calculates the bounding box of all the voxels that are in the provided grey level range.
- **SubImage module**: This module extracts subimages from its input image.
- **FastMorphology module**: This module performs standard morphological operations such as erosion, dilation, opening and closing.
- **ConnectedComponentMacro module**:This module performs a connected component analysis on 2D/3D images and returns an output image with largest connected component
- **Arithmetic2 module**: This module performs basic arithmetic operations on given two images. The resulting output is the processed image based to the selected function.
- **CompareMasks module**: Given Predicted mask and reference mask as an Input to this module. It can compute various similarity scores like Dice Score, Volumetric Overlap, Volume difference and more

5 Implementation

5.1 Detailed Active Learning Workflow Description

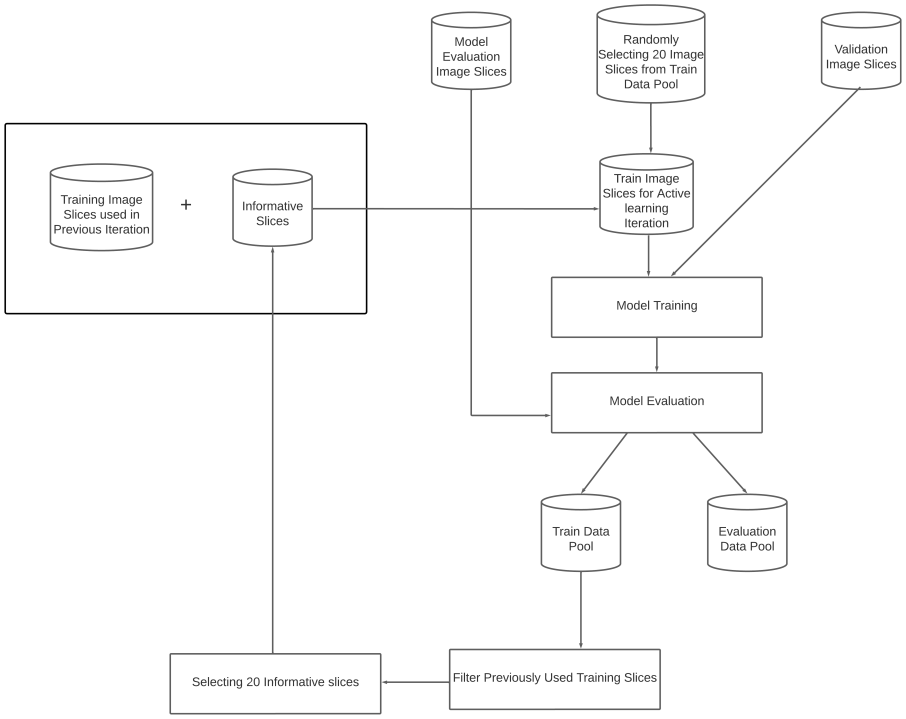


Figure 6: Detailed Active learning Workflow

The above diagram describes a detailed workflow of the Active learning process used.

In this section, A detailed implementation of the proposed architecture is explained. Beginning with the acquisition of the data from the data pool, the successive stages encompass data pre-processing, model training, performance evaluation, and selection of the most informative slices. This comprehensive walk-through targets to provide a detailed insight into the operational intricacy of the architecture.

According to the description provided in the subsection 4.1 , All the data is fetched from the designated data pool. This pool encapsulates all the breast images as well as their corresponding reference masks.

The collected data is then divided into three groups training, validation, and model evaluation. A bounding box is created over the interested area and the unwanted region around the breast is cropped out . After this step, 2D slices are extracted from all these 3D images. From the training dataset, 20 random image slices are chosen to serve as the initial training images for the first iteration of the active learning model training. For validating the model all the 792 image slices from the validation data are consumed. The validation data is kept constant for every iteration. This consistent validation set is used to evaluate the model's performance at each iteration, ensuring that the comparison between iterations is fair and meaningful. A brief description of the obtained dataset can be seen in the below Table 2

Dataset	3D Images	2D slices	Selected Slices per Active Learning Iteration
Train	32	3347	20
Validation	7	792	
Model Evaluation	13	1267	

Table 2: Obtained Dataset description

The 2D U-Net model described in the subsection 4.4, adopts the 'combineloss' function, synergizing two distinct loss functions - The Dice Loss and the Categorical Cross Entropy Loss. The weight parameter with a value of 0.5, used in this 'combineloss' function, equally distributes the influence between the two loss functions. The model was trained for 25000 iterations, 'CosineAnnealingLR' learning rate scheduler was utilized, and Adam optimizer was employed for optimization. Leveraging the above-mentioned architecture and configurations, model training process was carried out.

Parameter	Setting
Loss Function	Dice+CCE
Iterations	25000
Learning rate	CosineAnnealingLR
Optimizer	Adam optimizer
Batch size	100
Dropout rate	0.25
Activation Function	ReLU

Table 3: Model Parameter and Settings

To measure the model's efficacy on each of the image slices of both train and model evaluation datasets, the evaluation metrics described in the subsection 4.7 were used. The compare mask module from the MeVisLab assists in calculating these similarity scores which gives a quantitative measure of the model's performance.

After the model evaluates each slice from the train data pool and the model evaluation data pool, the slices from the train data that were part of the previous training iteration are filtered out. The rationale behind filtering out the slices used in the previous training iteration is to ensure that the model focuses on learning from new, and potentially more informative data. From the remaining training slices, the 20 slices with the largest errors, with respect to the chosen error metric in the experiment are selected. These selected slices along with the previous training slices are utilized for the next iteration of the active learning. The rationale for integrating previously employed training images with new ones in successive iterations is twofold: not only does the model learn from new instances, but it also endeavors to rectify errors made in previous training images. It is important to note that, Here I am not trying to apply the human-in-the-loop process directly. Instead, I am assuming that the image slices from the train data pool which were not used in the training process are without reference masks. When image slices from this pool are selected for the next iteration, the pre-existing reference mask is employed, mirroring the assumption of a human-informed correction.

The selection of these slices on which the model has made highest error, is performed using diverse methodologies. Initially, the Dice score was used as selection criteria, followed by assessing the volume of the largest error component and the third approach tried was by dilating the volume of the largest error component and then selecting the slices based on that. The detailed implementation of these methods is explained in the subsection 5.2.

The whole process was carried out for five Active learning iterations for each method. The changes in the number of training slices per iteration can be understood by the given below Figure 7. For each method, the initial active learning iteration employed the same set of image slices from the training data pool. This idea was deliberately leveraged to establish a common starting point among these methods, allowing for a comparative analysis of their performance and facilitating insightful observations at the conclusion of this study.

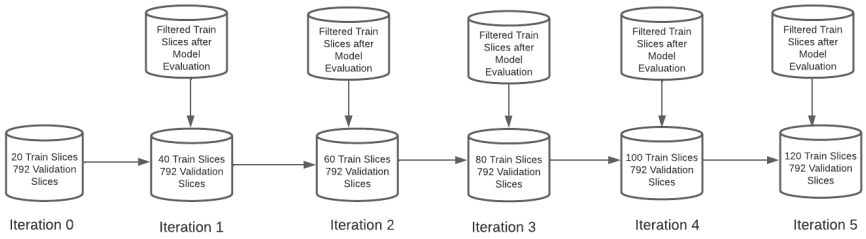


Figure 7: Data Flow Throughout The Active Learning Process

The above diagram describes how the train slices from previous run are concatenated with the 20 most erroneously predicted slices for the next iteration.

5.2 Methods Used For Selecting Informative Image Slices

5.2.1 Dice Score as Selection Criteria

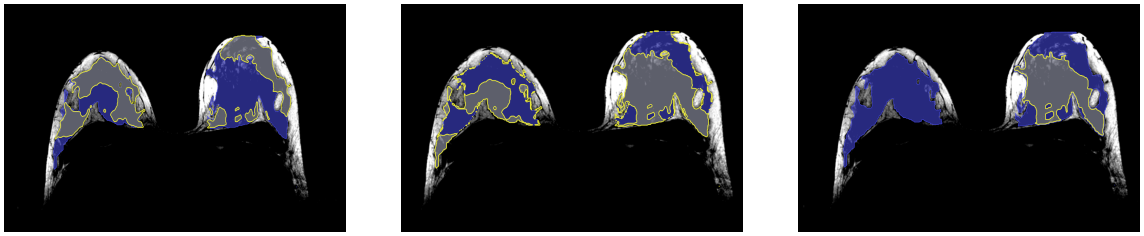
In this method, once the model has evaluated all the image slices, they are then subsequently arranged in ascending order according to their Dice scores. Afterward, the 20 slices with the lowest Dice score, and which were not used in the previous training iteration, are selected. These chosen slices are concatenated with the image slices utilized in the previous iteration, and are then employed for the subsequent active learning iteration.

The reason behind choosing the Dice score as a selection criterion is because of its ability to highlight the image slices where the model failed to perform well, and also because it is the most common error metric used in medical image segmentation. These slices with high prediction errors, indicate the model's learning gaps, thereby emphasizing the need for further improvement and attention [2].

Initially, this simpler method was implemented to observe how this approach performs on the data used in this study. However, It was noticed that there were some challenges associated with this approach, which are discussed in the subsequent section 6. As a result, more sophisticated methods were employed to tackle these issues.

5.2.2 Volume of the Largest Error Component

The volume of the largest error component is obtained by implementing several steps. Initially, for each image slice, a comparison is made between the predicted mask and the reference mask using an XOR operation. This is conducted through a dedicated Arithmetic Module in MeVisLab. This operation highlights the area of mismatch between the predicted and the reference masks. Subsequently, the FastMorphology module of MeVisLab is used to perform a Morphological Erosion operation on the resulting binary image. The idea of using an Erosion operation is that it helps in refining the analysis of discrepancies between the predicted and reference mask. This operation filters out minute differences and focuses on the larger areas of disagreement. Finally, the ConnectedComponentMacro module of MevisLab is used to find the volume of the largest error component.



(a) Model Predictions (b) XOR operation (c) Largest error component

Figure 8: Visual representation to find Largest error component

The blue contours in the graph represents reference mask, and the yellow contours represents predicted mask in (a). The yellow contours in (b) and (c) represent the area of mismatch and Largest error component respectively.

After calculating the volume of error for each slice, they are organized in descending order based on the volume of the largest error component. Afterward, The slices from the previous active learning iteration are filtered out, and then 20 slices with the largest volume of error component, from the remaining set are selected for the successive iteration.

Unlike the Dice score method, where both the image slice and its corresponding mask slice were used for training the model, this method includes an additional weight mask in the model training. The weight mask contains two weights: ”0 and 1”. The Largest error component gets a weight of 1, while the rest of the image gets a weight of 0. This helps in reducing the efforts of the annotator which I am stimulating here.

Here, In this study, as I am trying to integrate a sparse annotation approach to the active learning approach, the weight mask indirectly helps in implementing it. As per the concept explained in the subsection 4.3, this study assumes that a human annotator has corrected the identified largest error component and has labeled it as 0 or 1 depending on if there was parenchyma tissue present or not. A weight of 1 is assigned to all the voxels the annotator corrected. With respect to this study, the corrected part is a direct reference to the region of the actual reference mask.

(a) Original Image (b) Reference mask (c) Weight mask

Figure 9: Example Image of The image,mask and weight slices passed to the model for training.

The above image demonstrates the training data that will be passed to the model when sparse annotation is used.

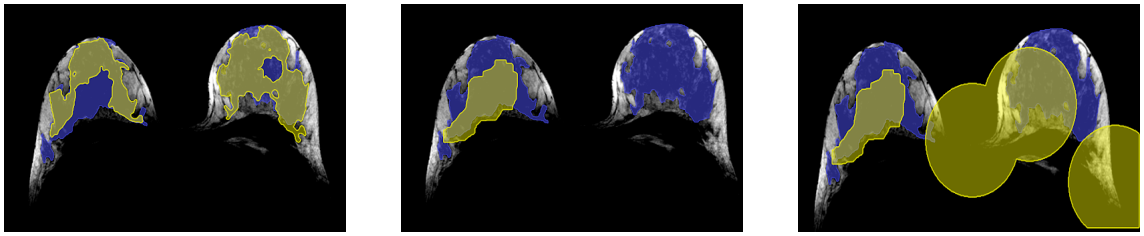
One important point to note: The slices that were selected for the initial active learning iteration, would not have a corrected weight mask. As a result, for those slices, a weight mask was generated where all the voxels were assigned a weight of 1. This signifies that the model should focus on the whole image.

There was an interesting issue that was encountered in this approach which is discussed in section 6. This issue was addressed using a similar kind of approach, albeit incorporating some additional modifications. The detailed description of this method is explained in the below subsubsection 5.2.3.

5.2.3 Dilated Volume of the largest connected component with Circular Blobs

The procedure to find the largest error component is similar to the procedure described in the above subsubsection 5.2.2. However, some additional changes were implemented on the weight slice. The output image from the module ConnectedComponentMacro is dilated with the help of the FastMorphology module. The rationale behind expanding the error component was to capture adjacent regions that could provide valuable context, Instead of just focusing on the core area of disagreement.

In addition to dilating the error component, Three circular Blobs were randomly created on the weight mask. The radius of these circles was formulated in such a way that the ratio of the combined volume of these circles to the volume of parenchyma present inside the largest connected error component is 9:1. This addition of circular blobs inside the weight mask contributed majorly in solving the issue faced while using the previous method. In this strategy it assumed that the annotator would also draw random blobs or scribbles in the background along with correcting the predicted segmentation in the largest connected error component.



(a) Model Predictions (b) Dilated Largest error component (c) Largest error component with Blobs

Figure 10: Visual representation to find the Largest error component with Blobs

The blue contours in the graph represents reference mask, and the yellow contours represents predicted mask in (a). The yellow contours in (b) and (c) represent the Dilated volume of Largest error component and Largest error component with Blobs respectively.

The underlying concept of the formula provided below lies in determining a radius value for circular blobs. This radius is chosen in a way that the combined volumes of all the circles are equivalent to nine times the volume of the parenchyma within the largest error component. The multiplication of the voxel size in the z dimension with the circle's area serves the purpose of making a comparison with the volume of the largest component. The voxel size is measured in millimeters, while the volume of the largest connected component is quantified in milliliters. To facilitate the conversion from millimeters to milliliters, the volume of the circular blob is divided by 1000.

The radius is calculated using the formula:

$$\text{radius} = \sqrt{\frac{9 \times 1000 \times \text{vol\_para}}{3 \times \pi \times \text{z\_component}}}$$

Where:

- vol_para = volume of parenchyma in largest error component (millilitres).
- z_component = voxel size in z dimension (millimeters).

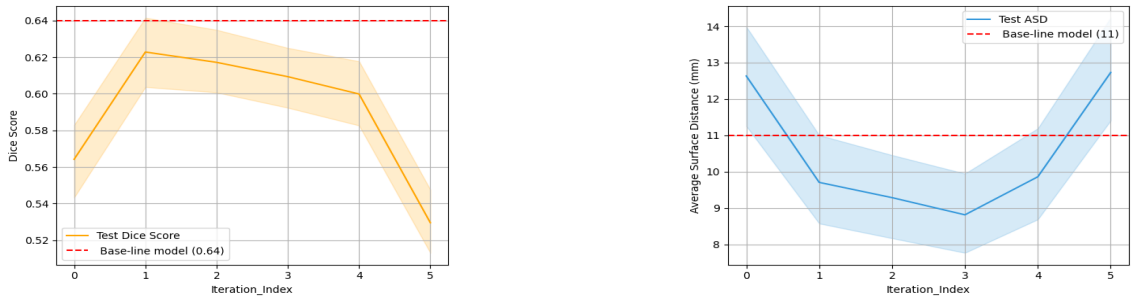
(a) Original Image (b) Reference mask (c) Largest error component with Blobs

The above images demonstrates the training data that will be passed to the model when sparse annotation(with blobs) is used.

6 Result and Observations

This section presents a comprehensive analysis of the implemented approaches detailed in the previous section 5. The evaluation will provide an insight on, if an active learning methodology coupled with sparse annotation can effectively address the hurdles linked with the traditional deep learning approach, and can provide an efficient fibroglandular tissue segmentation model. These challenges encompass the substantial need for extensive annotated datasets, and the constraints inherent in the manual annotation process in medical imaging. The fundamental assessment is carried out on the basis of the Dice score and Average surface Distance. Additionally, observations based on, the difference in the volume between predicted and reference segmentation are also carried out. These observations assists in understanding the problems that are associated with the "Dice score" based selection method, and the "Largest Error component without dilation" based method. The Dice scores and the Average surface distance measure obtained during each active learning iteration are compared with a base-line model which was trained on the full training dataset. It is worth noting that in-total 120 image slices were used in active learning based model training, and 3347 image slices for training the base-line model. The model evaluation data was kept fixed for evaluating both these approaches.

6.1 Analysis of Dice score based selection Method



(a) Obtained Dice score on evaluation data

(b) Obtained Average surface distance (ASD) on evaluation data

Figure 12: The above graph shows the performance of the model obtained in each active learning iteration.

The x-axis in the graph represents the iteration number, and the y-axis represents the value of dice score and Average Surface Distance for that iteration. The red dotted line in the both the graphs is the Dice score and ASD measure obtained from the base-line model.

The line graphs in Figure 12 depicts the dynamic progression of Dice score and Average surface distance across all the active learning iterations.From the Figure 12 (a), it can be observed that initially the model demonstrated an upward trend in the Dice score, but in the subsequent iterations, the upward trajectory gradually gave way to a declining trend. From this shift it can be inferred that, while the model initially captured essential learning's from the informative slices, but in the subsequent iterations it became challenging for the model to fetch essential new information. Overall, throughout the active learning process

the Dice score of the model remained below the base-line model performance. A similar trend can be observed from the Figure 12 (b), where in the beginning , when the model was able to capture essential learning's, the average surface distance between predicted mask and the reference mask decreased, however in the successive runs as the performance of the model started declining, the Average surface distance again started increasing.The possible reasons behind this method not performing well will be presented in the following subsubsection 6.3.1

6.1.1 Reference volume based Analysis

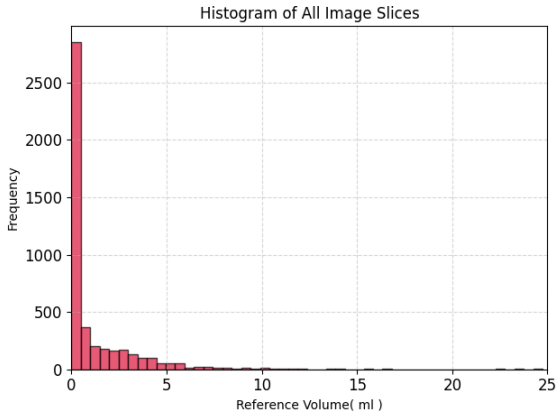


Figure 13: Reference volume Distribution of All Image Slices

The X-axis in the above diagram refers to the Reference volume, and Y-axis represents the frequency.

The Histogram shown in Figure 13 depicts the distribution of frequency of Reference volume (milliliters), which corresponds to the volume of Parenchyma present in the overall dataset. On one hand, It can be observed that the distribution is noticeably right-skewed. This indicates that a significant proportion of image slices possess minimal reference volume. However, on the other hand, it can also be seen that there are considerable amount of slices having varied reference volumes. This implies that to get an efficient segmentation model using an active learning approach, the methods that are employed for the informative slice selection should be able to fetch informative slices across the entire distribution, and not get influenced by data bias.

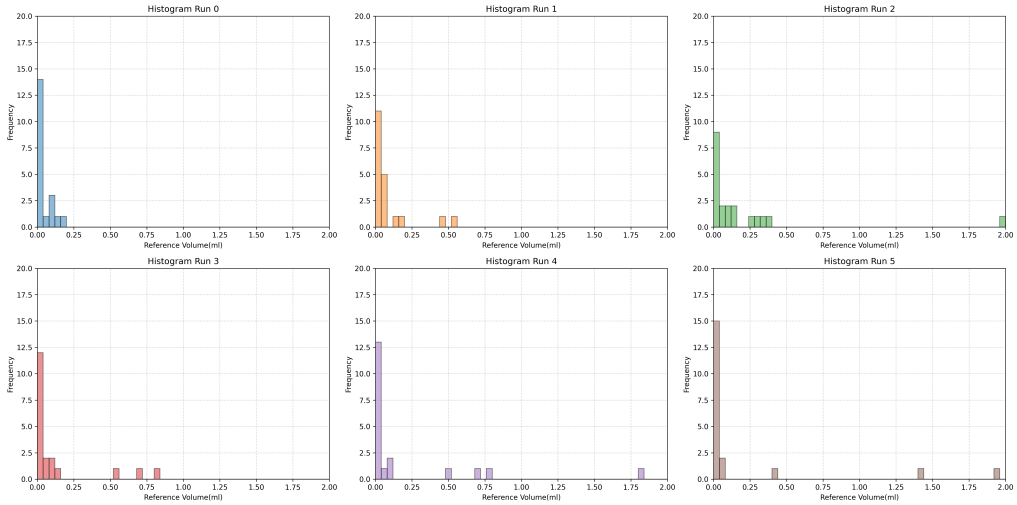


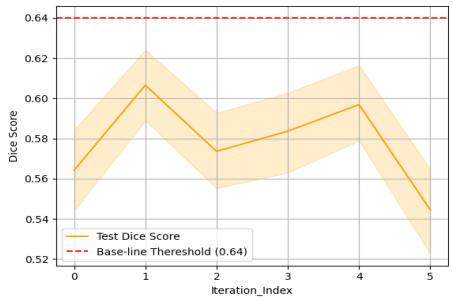
Figure 14: Reference volume Distribution of Selected Training Image Slices From Dice Score based selection method

The X-axis in the above diagram refers to the Reference volume, and Y-axis represents the frequency of the Reference volume.

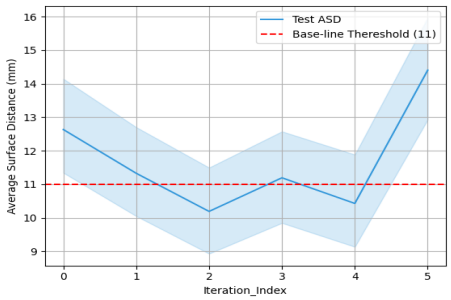
The Histograms in Figure 14 are achieved by plotting the reference volume of the 20 selected training slices employed in each active learning iteration. These slices were selected using the approach described in subsubsection 5.2.1. The height of each bar represents the frequency of the reference volume falling within specific bins. On observing the spread of these histograms, we can identify that all the training slices that were selected using this method were having similar and very minor reference volumes.

On comparing the distributions from Figure 14 with the population distribution from Figure 13. It becomes evident that this method struggles to comprehensively capture diverse information present in the dataset. The possible reason behind this could be, the weakness of the Dice score to have smaller values for structure with smaller volumes [9]. This deficiency directly impacts the model's ability to consistently improve its performance. As a result, it can be said that this simpler Dice score based approach was not able to perform well on the dataset used in this study.

6.2 Analysis of Volume of Largest Error component based Method



(a) Obtained Dice score on evaluation data



(b) Obtained Average surface distance (ASD) on evaluation data

Figure 15: The above graph shows the performance of the model obtained in each active learning iteration.

The X-axis in the graph represents the iteration number, and the Y-axis represents the value of dice score and Average Surface Distance for that iteration. The red dotted line in both the graphs is the Dice score and ASD measure obtained from the base-line model.

The line graph in Figure 15 (a) and (b), shows the variation in the Dice score throughout the active learning process. It can be observed that, As the iterations progressed, the Dice score exhibited significant fluctuations. This indicates an unstable and unreliable model performance. As can be seen from Figure 15 (b), a similar kind of fluctuating trend was also noticed in the Average surface distance graph. It is evident from its graph that, Initially the model struggled to accurately predict the segmentation boundaries, however in the successive iterations it actually performed better than the base-line model. Eventually in the last iteration average surface distance values drastically spiked upward, suggesting a substantial misalignment between the predicted and the reference segmentation. From this, it can be said that this informative slice selection method failed to choose slices, that can actually contribute in enhancing the model's performance. The possible reasons behind this method's under-performance are explained in the following subsubsection 6.3.1.

6.2.1 Reference volume based Analysis

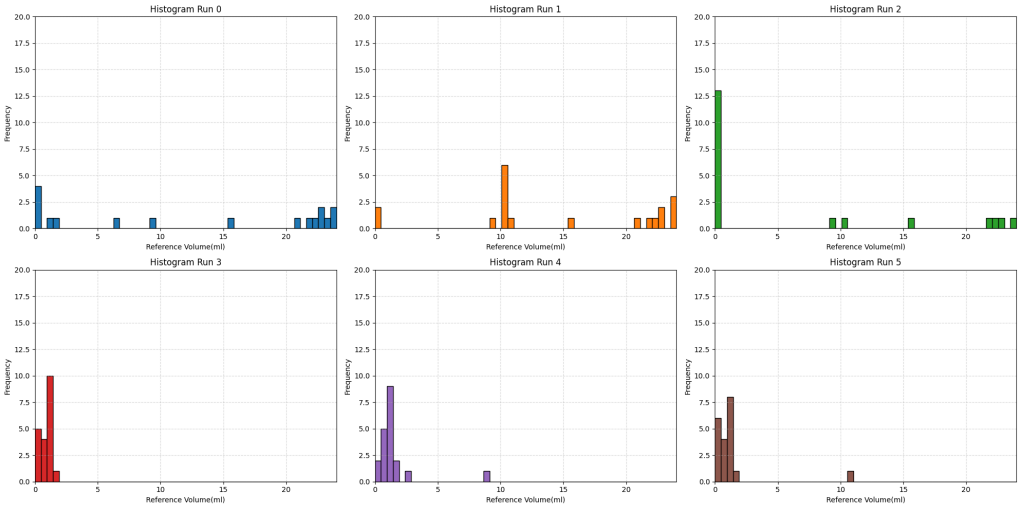
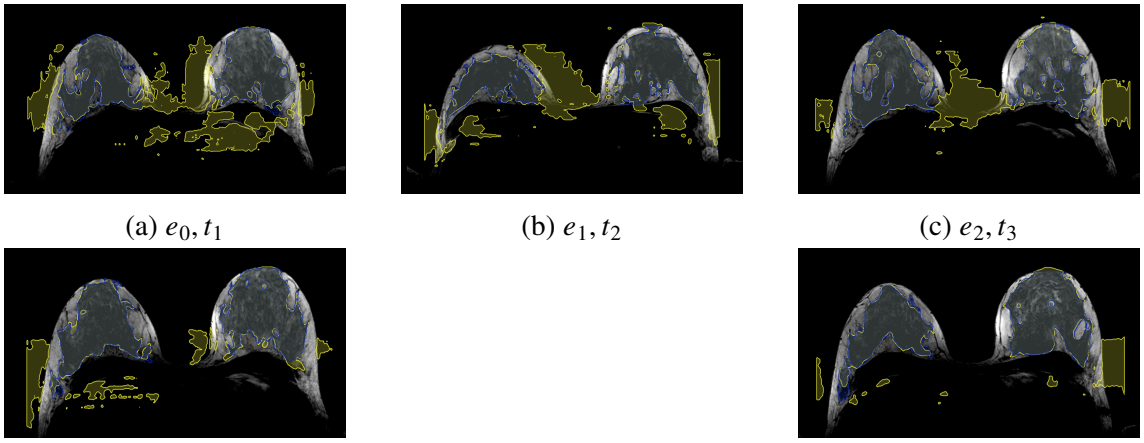


Figure 16: Reference volume Distribution of Selected Training Image Slices From Volume of largest error component based selection method

The X-axis in the above diagram refers to the Reference volume, and Y-axis represents the frequency of the Reference volume.

The Histograms shown in the Figure 16 are formed by plotting the reference volume of training slices used in each active learning iteration. These slices were selected based on the method described in subsubsection 5.2.2. It can be clearly seen that, in the first three iterations, this method was able to fetch Informative slices across the entire distribution. This can be verified by visualising the population distribution from Figure 13. Whereas in the last three iterations, the majority of the selected informative slices were having reference volumes in the range of 0 ml to 5 ml. As a result, it can be said that, Unlike the Dice score based method, this method was able to fetch image slices of varied reference volume. However, On observing Figure 15 (a) and (b), It can be deduced that, although this method was able to fetch slices with varied reference volume, it was still underperforming.

There was a reason associated behind this under performance. Notably, a significant level of over-segmentation was noticed in the informative slices chosen during each iteration. This was causing the slice selection process to select slices originating from the same 3D MRI image, that these over-segmented slices belong to. As a result, although the slices that were being selected in each iteration were having varied reference volumes, but as they belonged to the same MRI image, redundant information was being passed to the model in each active learning iteration. A potential explanation for this over-segmentation could be the imbalanced distribution of foreground and background information within the weight mask.



(a) e_0, t_1 (b) e_1, t_2 (c) e_2, t_3

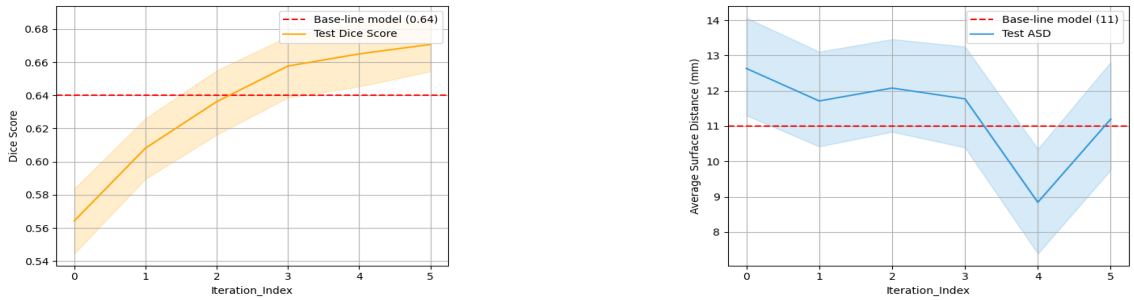
(d) e_3, t_4 (e) e_4, t_5

Figure 17: Over-segmentation in training slices

In the images shown above, the blue contours represent the reference mask, the yellow contours represent the predicted mask, and the label e_i, t_{i+1} indicates that the slice was selected as the most erroneously predicted slice by iteration i and was used as training slice in iteration $i + 1$ respectively

The images shown in the Figure 17 are some of the instances that were used in training the model during the active learning process, and these slices belong to the same MRI image. It can be clearly seen that these instances have very similar anatomical structures. Such kind of patterns were observed in all the training slices that were used across the active learning iterations. As a result, due to this redundancy in information, the model was not able to learn efficiently and improve its performance.

6.3 Analysis of Dilated volume of Largest Error component based Method



(a) Obtained Dice score on evaluation data

(b) Obtained Average surface distance (ASD) on evaluation data

Figure 18: The above graph shows the performance of the model obtained in each active learning iteration.

The X-axis in the graph represents the iteration number, and the Y-axis represents the value of the Dice score and Average Surface Distance for that iteration. The red dotted line in both the graphs is the Dice score and ASD measure obtained from the baseline model.

The line graph in Figure 18 (a) and (b), shows a consistent pattern of progress. It can be observed from Figure 18 (a) that, the dice score consistently showcased an upward trend throughout the active learning process. This signifies a continuous enhancement in the model's performance. Additionally, it can also be observed that in the last three iterations, the model even performed better than the base-line model. A similar trend was also seen in the Average surface distance graph shown in Figure 18 (b), where the model showed a slow, but consistent progress. In the fourth iteration, the ASD of the model was significantly less than the base-line model. This is a clear indication of the model being able to predict the boundaries of anatomical structure considerably well. In the last iteration, the ASD of the model slightly increased, but still, it was very near to the base-line model ASD. Overall, these trends reflect the method's ability to effectively select informative slices, which eventually helps the model to enhance its performance.

6.3.1 Reference volume based Analysis

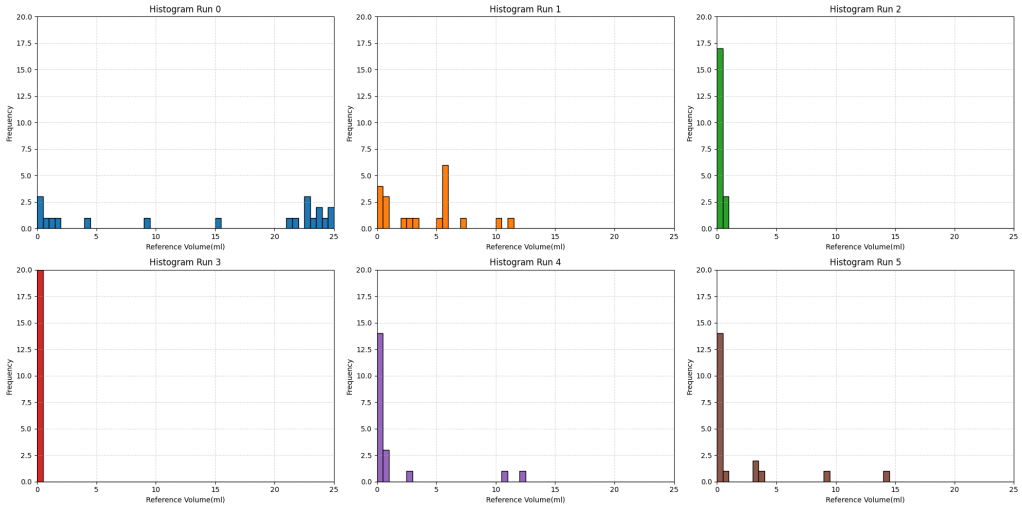


Figure 19: Reference volume Distribution of Selected Training Image Slices From Dilated Volume of largest error component based selection method

The X-axis in the above diagram refers to the Reference volume, and Y-axis represents the frequency of the Reference volume.

The Histograms shown in Figure 19 are generated by plotting the reference volume of training slices selected by the method described in subsubsection 5.2.3. By examining the graphs it can be understood that, initially, the majority of the selected slices had considerably high reference volume, in the subsequent iterations the slices had nearly zero reference volume, and eventually, there was a blend of both slices with zero and high reference volumes. From this, it can said that the slices selected by this model were able to provide sufficient and efficient information. Resulting in a reliable and enhanced model performance.

6.4 Comparative analysis of Sparse annotation based methods

This section will provide a brief analysis of the over-segmentation problem that was discussed in the subsubsection 6.2.1.Furthermore, it aims to offer insights into, why the Dilated volume of the largest connected component method demonstrated better performance compared to the one without dilation.

6.4.1 Over-segmentation analysis

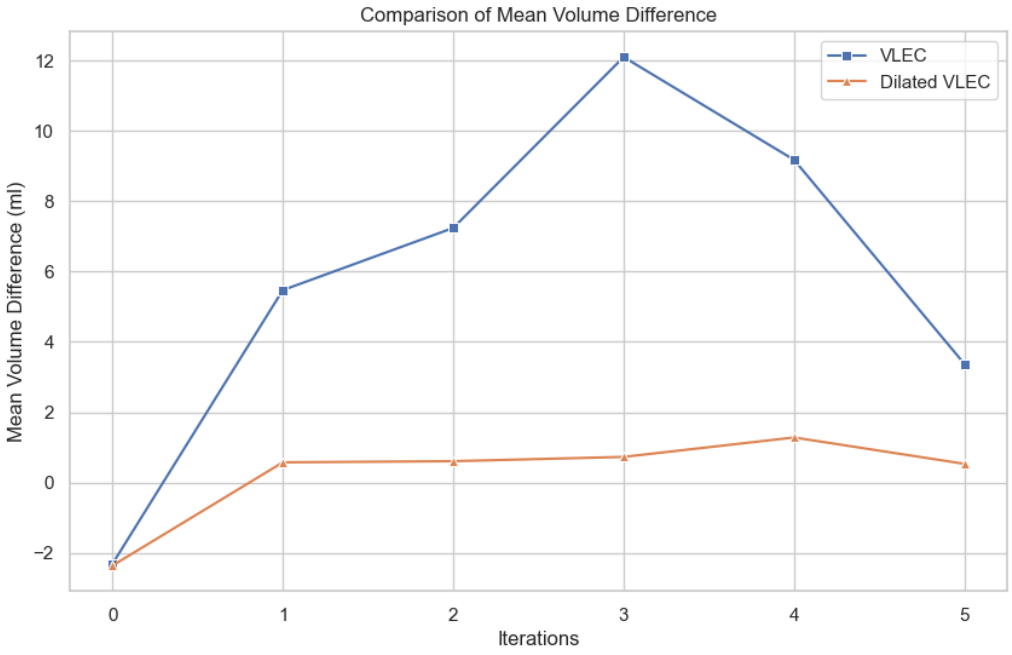


Figure 20: Line graph representing the mean volume difference observed in each iteration

The X-axis in the above diagram refers to the iteration number, and Y-axis represents the mean volume difference(ml).

As discussed in the subsubsection 4.7.3, if the value of volume difference is greater than 0 then, it indicates the presence of over-segmentation. The line graph shown in the Figure 20 is formed by taking the mean of the volume difference of all the informative slices selected by each method per iteration. As it can be observed from this graph, the mean volume difference of ”volume of largest connected component method (VLEC)” increases significantly in every iteration except the last two. This indicates that throughout the active learning process, this method was only able to select over-segmented slices that were fetched from a particular set of MRI images. Additionally, it can also be observed that the Dilated volume of the largest error component was not facing this issue. There were two reasons behind this, First, capturing adjacent regions around the largest error component through dilation operation, helped in providing additional valuable information, and Second, by creating random circular blobs the problem of imbalance between the foreground and background class was tackled.

6.4.2 Diversity of selected images

In this section, the analysis is carried out using three metrics. Namely, Average number of unique images in each run, Percentage of images present in run before, and Average distance between the slices. Before getting into the analysis part. A brief explanation of these analytical metrics are given below.

Average number of unique images in each run.

This metric explains that, From how many unique images, these methods were able to select informative slices. It is calculated by taking the average of unique images present in each iteration. Higher the number of unique images, higher are the chances of efficient information gain.

Percentage of images present in run before.

This metric explains the percentage of images repeated in the i^{th} iteration when compared with the $i - 1^{\text{th}}$ iteration. This gives an overview of, how many images are repeated in each successive run. Higher the percentage, lower are the chances of model performing well, due to redundant information being carry forward in the successive iterations.

Average distance between the slices.

It has been observed that, in MRI images, neighboring slices often display similar anatomical structures. These similarities arises from the smooth variations of tissues and structure across the volume. As a result, this leads to consistent patterns in adjacent slices.The computation of the average distance involves evaluating the difference between the slice indices of the i^{th} and $i + 1^{\text{th}}$ iterations. Notably, this calculation is constrained to slices originating from the same image within both the i^{th} and $i + 1^{\text{th}}$ iterations. This method unveils the degree of information variation inherent in each $i + 1^{\text{th}}$ iteration, with a higher average distance signifying a greater likelihood of diverse information content within the training slices of the $i + 1^{\text{th}}$ iteration compared to the i^{th} iteration.

Observations

VLEC	D-VLEC
5	3

(a)

Iteration	VLEC	D-VLEC
Run2 vs Run1	2	2
Run3 vs Run2	3	7
Run4 vs Run3	11	4
Run5 vs Run4	0	5

(b)

Iteration	VLEC	D-VLEC
Run2 vs Run1	100%	20%
Run3 vs Run2	66%	50%
Run4 vs Run3	100%	50%
Run5 vs Run4	0%	20%

(c)

Figure 21: (a): Mean number of unique images in each run,(b): Mean distance between the slices, (c): Percentage of images present in run before

The data presented in Figure 21 tables offers a fundamental statistical overview of the aforementioned metrics. Analyzing the data depicted in Figure 21 (a), it becomes apparent that the method based on the dilated volume of the largest connected component exhibited a higher count of unique images throughout the active learning process. This trend is further highlighted in Figure 21 (b), which indicates that the average distance between slices in successive iterations was greater in the dilated volume method, implying a broader coverage of distinct slices. Moving to Figure 21 (c), it's evident that the volume of the largest connected component approach predominantly drew around 66% of its slices from images already utilized in the previous runs, in contrast to the other method, which had an average percentage of 30%. Overall, this analysis underscores that the volume of the largest connected component method was plagued by a redundant inflow of information during active learning, contributing to its under-performance. Conversely, the dilated volume of the largest error component approach exhibited comparatively lower redundancy, thus fostering the model's effective performance enhancement.

7 Conclusion

In conclusion, this research aimed to harness the potential of active learning combined with sparse annotation to tackle the challenges inherent in traditional deep learning methods for medical image segmentation. Through rigorous experimentation and analysis, it became evident that the proposed approach holds promise in overcoming the limitations of extensive annotated data requirements and the time-consuming manual annotation process. The observed trends in the evaluation metrics, particularly the Dice score and Average Surface Distance, provided valuable insights into the model's performance.

One of the most remarkable achievements is the realization that, the active learning approach achieved comparable accuracy to the baseline model with a substantially reduced dataset of just 120 slices, in contrast to the 3347 slices used by the baseline model. Furthermore, the novel concept of utilizing the dilated volume of the largest error component-based method showcased its potential not only in elevating model performance, but also in potentially streamlining the laborious manual annotation of complex anatomical structures, such as the Fibroglandular tissue.

In essence, this study sheds light on the immense potential of the proposed workflow. Beyond its proven ability to achieve accurate and efficient medical image segmentation, this approach addresses the challenges of manual annotation in medical imaging. By strategically combining active learning with sparse annotation, the workflow not only streamlines the arduous annotation process but also demonstrates its capacity to yield a highly proficient model with minimal reliance on annotated data. This dual impact, on both improving annotation efficiency and enhancing model performance, underscores the transformative capabilities of this novel approach, promising to reshape the landscape of medical image analysis and pave the way for more effective and streamlined diagnostic tools.

8 Future Work

In order to build upon the foundations laid by this study, several directions for future research emerge. Firstly, to thoroughly evaluate the robustness of our approach, it is imperative to extend the active runs to encompass a broader range of iterations. This will provide a more comprehensive understanding of how the model's performance evolves and stabilizes over a longer training period. Additionally, exploring the impact of increasing the number of training slices per iteration could yield valuable insights. This adjustment could potentially expedite convergence rates and enhance the model's ability to capture intricate anatomical features. Moreover, the integration of uncertainty-based active learning holds substantial promise. By incorporating metrics that quantify the uncertainty of predictions, we could devise a selection strategy that actively targets the most informative slices, thus further enhancing the efficiency of the annotation process. This uncertainty-based approach can significantly increase the scalability of the model, as data lacking annotation certainty can help identify regions where the model's predictions are less confident. Annotators can then rectify these regions, incorporating them into model training to improve its overall performance. Finally, an intriguing avenue for investigation is the application of a true human-in-the-loop approach. This would involve collaborating with domain experts to provide real-time feedback during the active learning process, infusing the model with valuable insights that mirror the expertise of skilled practitioners. These explorations hold the potential to refine and amplify the efficiency of our proposed methodology, contributing to a more adaptive and responsive segmentation model.

9 Acknowledgement

I would like to express my sincere appreciation to the individuals who have been pivotal in my journey of conceptualizing, implementing, and composing my Master's thesis. First and foremost, my heartfelt gratitude goes to Dr. Markus Wenzel, my primary supervisor, for entrusting me with the opportunity to delve into a compelling and pertinent subject matter. His unwavering guidance and constructive input were instrumental in steering my research towards fruition. I am also deeply thankful to Kai Gießler, whose insightful perspectives consistently enriched my understanding and resolutions to intricate challenges.

I extend my acknowledgment to my peer, Anurag Sundar, whose engaging discussions have provided valuable perspectives across various facets of my thesis. Gratitude is also owed to the larger Fraunhofer MEVIS community, which fostered an environment of continuous learning and intellectual exploration. The invaluable feedback I received from multiple members at MEVIS stands as a testament to their commitment to nurturing growth and excellence.

In a realm extending beyond academia, my parents and elder sister stand as unwavering pillars of support throughout my life journey, including the duration of this thesis. The culmination of my thesis owes its depth and quality to these remarkable individuals. Their collective efforts have been essential in equipping me with the tools and inspiration to effectively realize and articulate my research. This transformative journey has been a profound learning experience, underscored by the steadfast encouragement of these exceptional individuals.

References

[1] 2023 MeVisLab (c) MeVis Medical Solutions AG. "MeVisLab Framework Information". In: (2023). URL: https://www.mevislab.de/mevislab.

[2] Karl G Baum et al. "Application of the Dice Similarity Coefficient (DSC) for failure detection of a fully-automated atlas based knee mri segmentation method". In: *ISMRM Annual Meeting*. 2010, pp. 1–7.

[3] Mélanie Gaillochet, Christian Desrosiers, and Hervé Lombaert. "TAAL: Test-time augmentation for active learning in medical image segmentation". In: *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*. Springer. 2022, pp. 43–53.

[4] Neeti B Goel et al. "Fibrous lesions of the breast: imaging-pathologic correlation". In: *Radiographics* 25.6 (2005), pp. 1547–1559.

[5] Dirk Grosenick et al. "Review of optical breast imaging and spectroscopy". In: *Journal of biomedical optics* 21.9 (2016), pp. 091311–091311.

[6] Sepideh Iranmakani et al. "A review of various modalities in breast imaging: technical aspects and clinical outcomes". In: *Egyptian Journal of Radiology and Nuclear Medicine* 51.1 (2020), pp. 1–22.

[7] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[8] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 fourth international conference on 3D vision (3DV)*. Ieee. 2016, pp. 565–571.

[9] D Müller, I Soto-Rey, and F Kramer. "Towards a Guideline for Evaluation Metrics in Medical Image Segmentation. arXiv 2022". In: *arXiv preprint arXiv:2202.05273* ().

[10] Mario Mustra, Mislav Grgic, and Jelena Bozek. "Application of gabor filters for detection of dense tissue in mammograms". In: *Proceedings ELMAR-2012*. IEEE. 2012, pp. 67–70.

[11] Inside Radiology. "General Imaging process in Breast Care". In: (2023). URL: https://www.insideradiology.com.au/breast-mri/.

[12] Annika Reinke et al. "Common limitations of image processing metrics: A picture story". In: *arXiv preprint arXiv:2104.05642* (2021).

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[14] Jun Wei et al. "Correlation between mammographic density and volumetric fibroglandular tissue estimated on breast MR images". In: *Medical physics* 31.4 (2004), pp. 933–942.

[15] Shandong Wu et al. “Fully-automated fibroglandular tissue segmentation in breast MRI”. In: *Breast Imaging: 11th International Workshop, IWDM 2012, Philadelphia, PA, USA, July 8-11, 2012. Proceedings 11*. Springer. 2012, pp. 244–251.

[16] Lin Yang et al. “Suggestive annotation: A deep active learning framework for biomedical image segmentation”. In: *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. Springer. 2017, pp. 399–407.

[17] Varduhi Yeghiazaryan and Irina Voiculescu. “Family of boundary overlap metrics for the evaluation of medical image segmentation”. In: *Journal of Medical Imaging* 5.1 (2018), pp. 015006–015006.

[18] Yang Zhang et al. “Automatic breast and fibroglandular tissue segmentation in breast MRI using deep learning by a fully-convolutional residual neural network U-net”. In: *Academic radiology* 26.11 (2019), pp. 1526–1535.

[19] Zhenxi Zhang et al. “A sparse annotation strategy based on attention-guided active learning for 3D medical image segmentation”. In: *arXiv preprint arXiv:1906.07367* (2019).