FIT3161 Computer Science Project

Project Design

Developing a New Method for Diagnosis the Breast Cancer in Mammography Images

Segmentation of Tumour and Extracting Useful Features From Mammography Images

Team ID: MA_21
Team Members: Daniel Kee
Jaclyn Neoh
Tan Sook Mun
Project Supervisor: Dr Golnoush

Table of Contents

Introduction

According to the national cancer institute, breast cancer is one of the most common cancers for women. (National Cancer Institute,n.d.) Regular screening is encouraged for early detection of Breast cancer. The reason for this is because early detection and mitigation can help avoid the patient regressing into late-stage cancer where it may deteriorate their health or even cause death. The common practice of detecting tumours in mammography images is done manually by a radiologist. However, this method leaves room for human error. The radiologist may develop fatigue or lack attention to detail after many repetitions. With these problems, there is high demand for Computer-Aided Diagnosis (CAD) (Berber,2013) to reduce mistakes and human error. As advanced as the diagnosis of breast cancer in mammography images that we have in hospitals now, there are still some cases where tumours might be missed in the diagnosis. Our project objective is to address this problem, by proposing an accurate algorithm for segmenting abnormalities in mammography images. We hope to be able to detect tumours in the mammography image more accurately, down to the least noticeable tumour.

The main objective of the project is to develop a new method for the diagnosis of breast cancer in mammography images. In this project, there are three different stages. Firstly, is to remove the pectoral muscle from the mammography. Next, the breast tumour will be segmented out and its features extracted. Lastly, is the classification process whereby using machine learning it will be decided whether the tumour is benign or malignant. Our team will be working on the second stage, where we will have to propose an accurate segmentation algorithm to segment out the abnormal tumour from normal breast tissue in the mammography image and extract any useful features. The end result of this project would be a diagnosing system that takes in mammography images with the pectoral muscle removed. Then it will perform segmentation and feature algorithms onto the mammography. Lastly, it outputs outstanding features on that image that would potentially be a tumour.

1.0 Representation of Design

1.1 Operational Framework

Below is an illustration of our project as an overview:

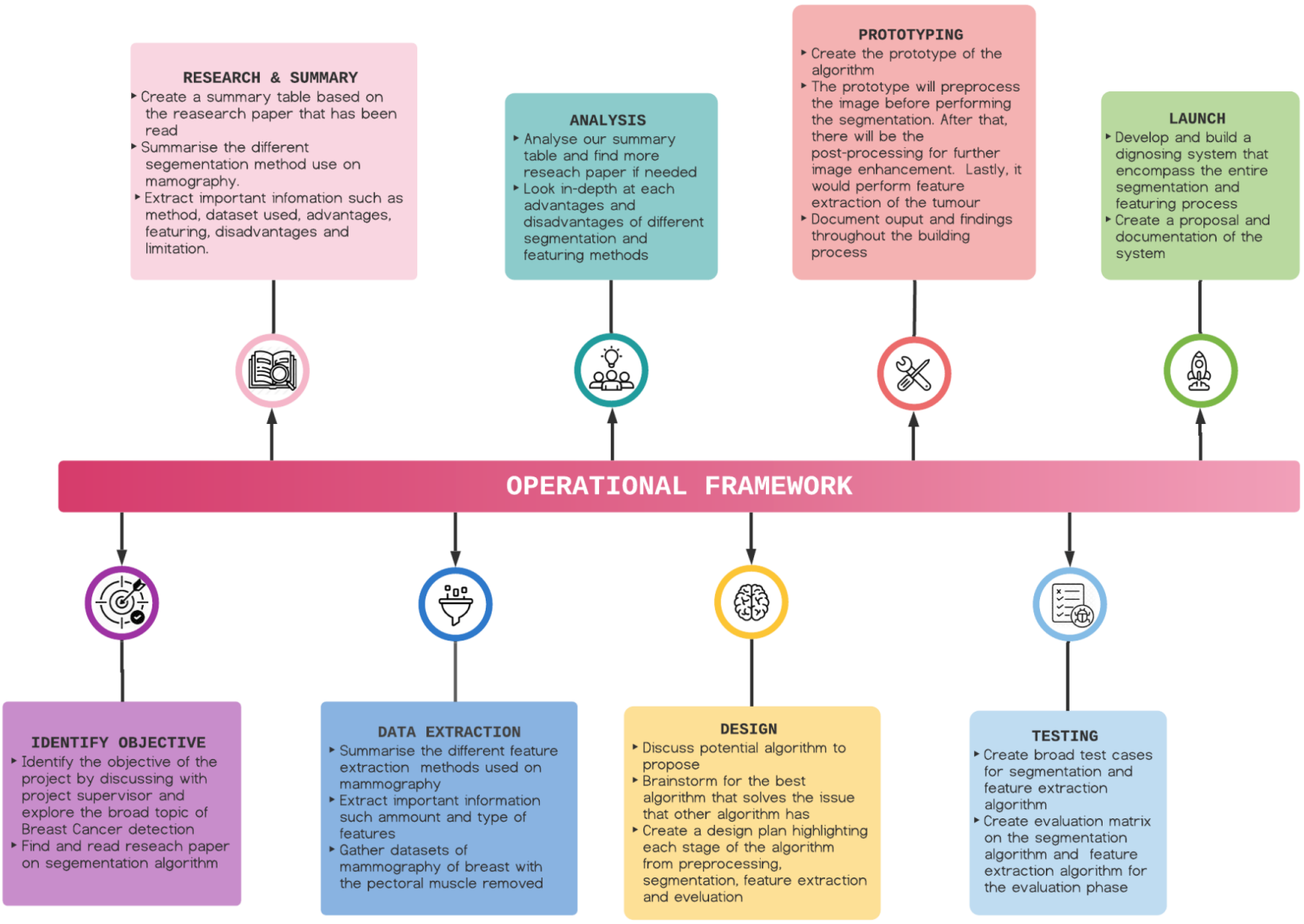


Figure 1-Operational Framework

The diagram above presents the operational framework of our project. The first task that every project needs is to identify the objective of the project. This is to ensure we do not stray from the objective and deliver what we are supposed to. The bulk of the journey consists of the research phase. This is represented by research & summary, data extraction, and analysis. During these stages, we research different segmentation and featuring algorithms. We deeply analyse and summarise our findings. This is very important because we will need this information to move to our design phase. The design, prototyping and testing stages represent the implementation phase. Finally, we will launch our product which is a diagnosing system that implements our proposed segmentation and featuring methods

1.2 Algorithm Framework

Below is an illustration of our algorithm process:

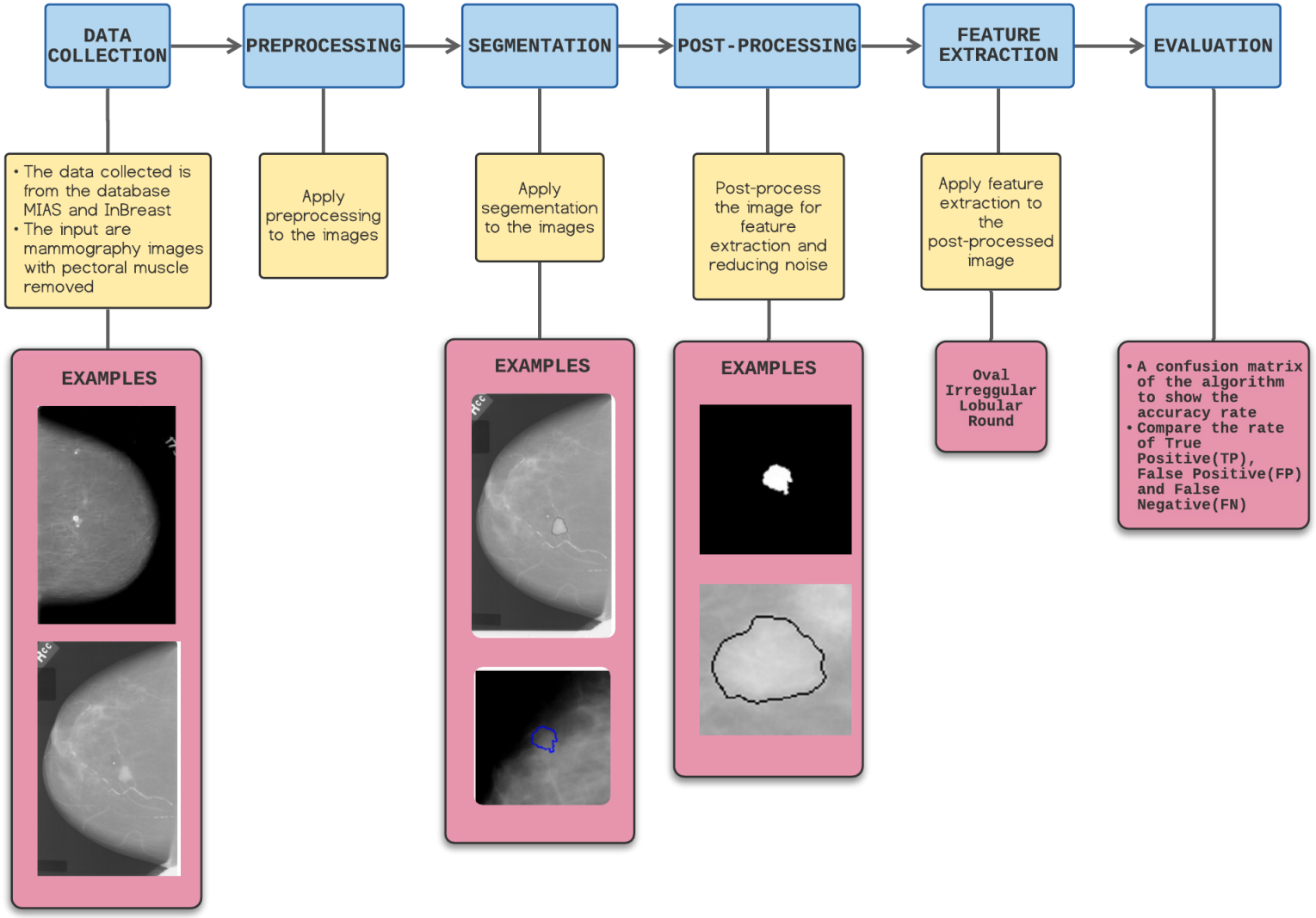


Figure 2-Algorithm Framework

The diagram above represents our algorithm framework. The first stage is collecting mammography of the breast that has the pectoral muscle removed. Next, we preprocess the mammography so it is easier to segment out the tumour. Based on the research we had done, this is an important step because it increases the accuracy rate. The algorithm then performs segmentation. The above shows some examples of the mammography output. Before extracting the feature, it is important to post-process the mammography image. This is to ease the feature extracting process and decrease noise in the image. Lastly, we perform feature extraction onto the image. It will extract features of the tumour and output its attributes. We then evaluate our algorithm to ensure it is accurate. We create a confusion matrix of the algorithm and compare the accuracy rates.

1.3 Algorithm Summary Table

Below is the summary of different algorithms :

Table 1 - Algorithm Summary Table

ALGORITHM SUMMARY TABLE			
Segmentation Methods	Watershed segmentation	Conditional Generative Adversarial Network (cGAN	Level-set segmentation
Approach	The objective of the Watershed segmentation method is to identify the catchment basins that represent the border between the breast tissue and tumour.	This segmentation algorithm has a preprocessing phase. Before applying the segmentation, it will apply Single Shot Detector (SSD) to locate and produce the coordinate of the tumour. During the segmentation, it computes the proper coordinates of the tumour to crop the image.	Level-set segmentation is a method that enhances the region according to an energy field. Basically, this method starts with an initial region and grows the region while minimizing region energy.
Advantage	The advantage of the Watershed segmentation method is that it has the lowest false positive rate. With this, we can infer that the method is accurate and consistent as doing the diagnosis manually	The Conditional Generative Adversarial Network (cGAN method gives better results based on the cropping of the image. It gives better results such as higher accuracy when it is cropped in a tight frame compared to other methods.	The advantages of the level set segmentation are that it has high accuracy and a true positive rate among other algorithms.
Disadvantage	The disadvantage of the watershed segmentation method is that it has a low true positive rate and accuracy value. sWith this we can infer that this method could not cover the mass area as accurately as other methods.	The disadvantage of this method is that it is not able to detect when there are two tumours in the loose framing and one tumour is incomplete and the other complete. It is not able to properly segment the complete tumour and it will fail to segment the incomplete one.	The disadvantage of this method is that the false positive rate of level set-algorithms are higher than other known algorithms.

Based on the table above, we have a few possible methods that we could implement for our project. This table lists out the overview of each algorithm, the approach it implements, and what are its advantages and disadvantages if we were to implement it. The table above

represents just a few of the many algorithms we had studied and analysed. Based on the table above it is clear that different approaches yield different results. Each method has its own advantages and limitations. Our project plan is to improve segmentation methods by inferring from these different methods and proposing a method that provides a better accuracy rate compared to other methods.
For our software prototyping, we choose to show watershed algorithms. This is because it has the lowest false positive rate. Also with these results, we can infer that the method is as accurate and consistent as doing the diagnosis manually. However for our project, we will not be implementing this method. It is to show our proof of concept and general segmentation process.

2.0 Hardware & Software Specification

2.1 Hardware and Software Requirements

Table 2 - Hardware Specification

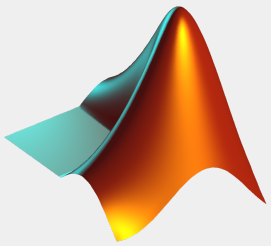
Hardware Specification	
Operating System	Windows 10 Home
CPU	Intel i5
GPU	GTX 1080
RAM	8GB
External Hardware	Hard Disk 1Tb

Table 3 - Software Specification

Software Specification	
Programming Language	MATLAB 2020B
Software libraries	Image Processing Toolbox
Collaborative tool	GitLab
Project Management Tool	Notion

2.2 Software Justification

Table 4 - Software Justification

Chosen Tool	Alternatives
Matlab  MATLAB is a programming language that is a proprietary multi-paradigm. MATLAB is a very popular language as it can be used for mathematics and computation. It allows plotting of functions and data, matrix manipulations, implementation of algorithms, and creation of user interfaces. In terms of framework, library or tools, you are only required to install the Image Processing Toolbox which is really easy to do. Matlab is widely used for Image Processing. In Monash, Matlab is taught during the Image Processing unit. Two of the members are currently taking up this unit and have been using Matlab since the start of the semester. The syntax is rather easy to understand and the language is easy to use. The main reason why we chose to use Matlab is that the team is already familiar with using Matlab for Image Processing despite only learning it this semester.	**Python**  Python is a very versatile language as it can be used for multiple purposes. It is a programming language that all three of us are comfortable with as we have used it in prior units. The syntax itself is very simple. This are the following tools/ library/ framework that could be used when using python for Image Processing: - TenserFlow - OpenCV - Theano - Keras The main reason why we did not go for Python is mainly due to the fact that even though all three of us used Python before, we had never tried Python for Image Processing. There are also a lot of frameworks that can be chosen to be used if we decide to go with Python and it will cost us some time to explore each of them to discover which frameworks fits us the best. **C++**

	 C++ is also a very versatile programming language. It supports object-oriented, generic, and functional features as well as facilities for low-level memory manipulation. C++ is a language that neither of us used before. The syntax of C++ is not as simple as Python or Matlab. These are the following tools/ library/ framework that could be used when using C++ for Image Processing: - Tensorflow - OpenCV Neither of us used C++ much less use C++ for Image Processing. The main reason why we didn't go with it is due to the lack of experience with this language. In terms of syntax, it is also not that easy to code compared to Matlab and Python.
GitLab  GitLab is a tool that can be used to help sync up all the team's code though a Git-repository manager. We've decided to go with Gitlab as all of us have had some experience using it before from our previous assignments and internship.	**Matlab Drive**  MATLAB Drive is a cloud-based storage location and it can be used to help sync up the team's code. With MATLAB Drive, you only have 5GB worth of space. We felt like space is limited and we have not used Matlab Drive before.

Since we have decided to use Matlab, we don't need to have an additional framework, tool or library other than the Image Processing Toolbox to code the segmentation algorithm, which is easy to download. In terms of designing and building our graphical user interface(GUI), we will be using the guide environment that is already in the standard package of MathWorks. It is easy to use and doesn't require us to download additional toolboxes. In terms of deploying the application, we will be using deploytool which requires

us to download MATLAB Compiler. MATLAB Compiler is used to build standalone executables and web apps from MATLAB programs and it requires installation.

2.3 Project Management Tool Justification

Table 5 - Project Management Tool Justification

Chosen Tool	Alternatives
Notion  Notion is a tool that has comprehensive functionality such as notes, databases, kanban boards, wikis, calendars and reminders. This tool can be accessed both through an app and the website which gives us easy access and edit. This tool allows all of us to edit and add to the project flow. Unlike other tools that have a specific use, this tool has a kanban board, timeline, sprint cycle to help keep track of our progress. The tool also provides a comprehensive To-Do list that allows us to set the status, priority, assignments and more. You can even sort the task by different attributes and set it as a list or a kanban board. This tool also has a wiki tab that enables us to write documents such as (code reviews, how-to guide and etc) for other members to read. With all these functionalities stated above, it is a clear reason we choose this tool above all others. This tool enables us to create a project timeline for us to ensure we are progressing on the right track. The to-do list is very comprehensive so that we can separate project tasks and unit assignment tasks. We can also assign tasks to other members and it is much easier to update each other on what to do.	**Trello**  Trello is a project management tool that helps keep you and your team on track and has build-in collaboration tools. You can use it to create tasks really easily by writing the task in cards, adding extra details of the task onto the card itself. It is also really user-friendly as it is so easy to use Trello. Unfortunately, the reason why we did not go with Trello is because there are not much functionality and features to it like for example reporting functionality and time-tracking features which is a feature that we need. **Monday.com** 

Monday.com's interface is very appealing. They have a lot of templates for you to choose from based on the task you have. It is also very flexible and easy to use. Not only that, but as the workload increases, you can also increase the number of boards.

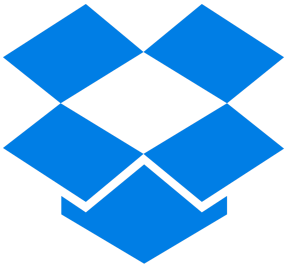
With that being said, there is not much functionality to it. There is also a lack of recurring tasks and the mobile app version of Monday.com is not the easiest to use.

Kanban Tool

Kanban Tool is a good project management tool, mainly for teams that are using the agile methodology as it offers seamless time-tracking functionality and allows you to visualize your workflow. It is really easy to use it and it also has other features like reporting, notifications and team management features that make collaboration and task tracking easy.

Our biggest problem with Kanban Tool is the fact that there's only a free plan that is up for only two users and this plan only supports two project boards. If we pay $15 per month, we do have unlimited boards but we still cannot unlock features like time tracking, reporting, user management, and process automation.

Asana

	Asana is a project management tool that focuses on collaboration and it is very flexible on how projects can be structured. Asana supports the productivity and collaboration of a team really well. Asana's interface is rather simple and is not as appealing compared to other project management tools. There is also not much functionality to it. Their system is not that easy to use as it is really rigid.
Google Drive  Google Drive is a non-self hosted cloud service that helps sync your documents. It is a very common tool to store files and other documentation. It takes close to little effort to set up a shared drive for the time to use. The reason why we chose Google Drive is that we are very familiar with it as we have been using Google Drive not only since the start of the semester but at the very beginning of our educational journey at Monash University. With our Monash student account, we have unlimited storage, and it is pretty secure as it offers two-factor authentication and encrypts your data when it's in transit	**Dropbox**  Dropbox is one of the most famous self-hosting cloud services. They have amazing security, 2GB worth of space without needing us to pay extra money. The reason why we did not go through with Dropbox other than the fact that the 3 of us have more experience with Google Drive but also, it is way harder to set it up. Also, another big factor is the amount of space we have. Comparing 2GB versus Google Drive's 15GB, we felt like Google Drive is the best option for us.
Whatsapp  Whatsapp is a great communication tool we like to use. It is super easy to create a	**Slack**  Slack is also a great alternative to consider as a communication tool. It supports multiple channels which makes it easy to

group for discussion. With our frequent meetings and the use of Notion, we just need an easy channel to pass information to each other. Hence why we are using Whatsapp	discuss certain topics based on the channel. With that being said, with the frequent meetings and the use of Notion, using Slack is not necessary at all. Setting up a group in Slack will also take up some time as neither of us has experience creating a group via Slack. **Email**  Email is also another option of communication. Email is not as effective as Slack nor Whatsapp. It is also more convenient to use Whatsapp over email as we already have each other's email.

3.0 Proof of Concept

3.1 Overview of Segmentation Process

One of the datasets we use is the MIAS dataset. Its full name is The Mammographic Image Analysis Society, “an organisation of UK research groups interested in the understanding of mammograms and has generated a database of digital mammograms”. “Films (are) taken from the UK National Breast Screening Programme”. It has a total of 322 mammography images with 69 images with Benign tumor and 55 images with Malignant tumor.

Table 6 - Breakdown of Dataset

Number of Images that has tumor		Number of Images that doesn’t have tumor
Benign Tumor	Malignant Tumor	
69	55	198

The following code and images uses MATLAB as its programming platform and uses the watershed method to perform segmentation on the mammography image. Note that this is not the final product of the project, but a guide or overview of how the whole process of the software/algorithm works.
Note that these images have been preprocessed (imbinarized function in MATLAB) only for the purpose of showing the overview of the segmentation process below.

The image from the MIAS dataset used for this overview is as follows:

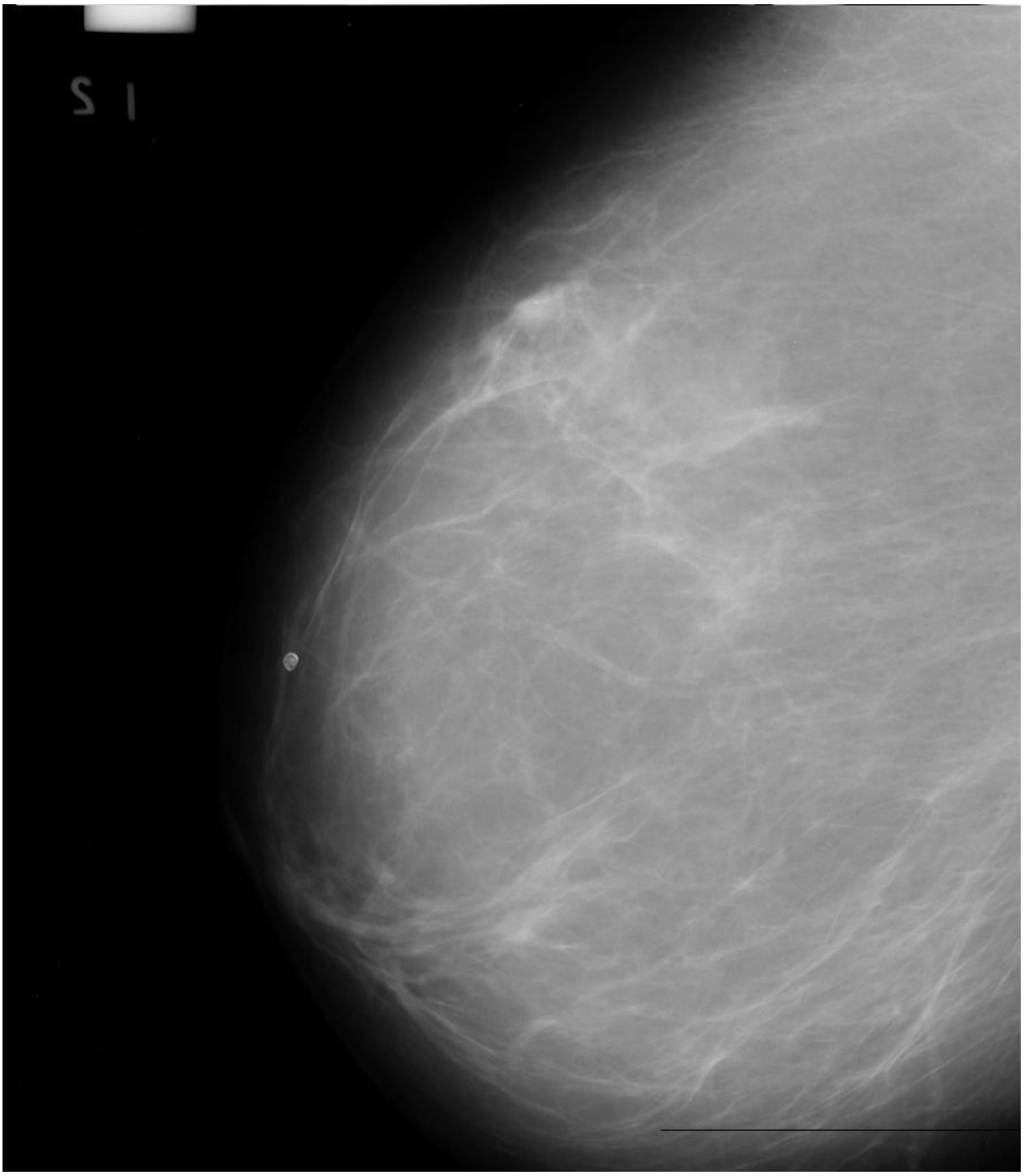


Figure 3

The following code binarizes the mammography image with a gaussian adaptive threshold of the image(T1). T2 and T3 are the adaptive thresholds of the median and mean of the image respectively.

```
I = imread('mdb258.jpg');
I = double(I);
T1 = adaptthresh(I, 'Statistic','gaussian');
T2 = adaptthresh(I, 'Statistic', 'median');
T3 = mean2(I);
bw = imbinarize(I, T1);
figure,subplot (2,2,1),imshow(I, []), title ("Original image gaussian A.T.");
subplot (2,2,2),imshow(bw, []), title ("Binary image gaussian A.T.");
D = bwdist(~bw);
subplot (2,2,3),imshow(D, []), title ("Distance image gaussian A.T.");
L = watershed(D);
subplot (2,2,4),imshow(L, []), title ("Watershed transform gaussian A.T.");
I(L == 0) = 0;
figure, imshow(I, []), title ("segmented image gaussian A.T.");
```

Code 1

The result of the above code is as follows:

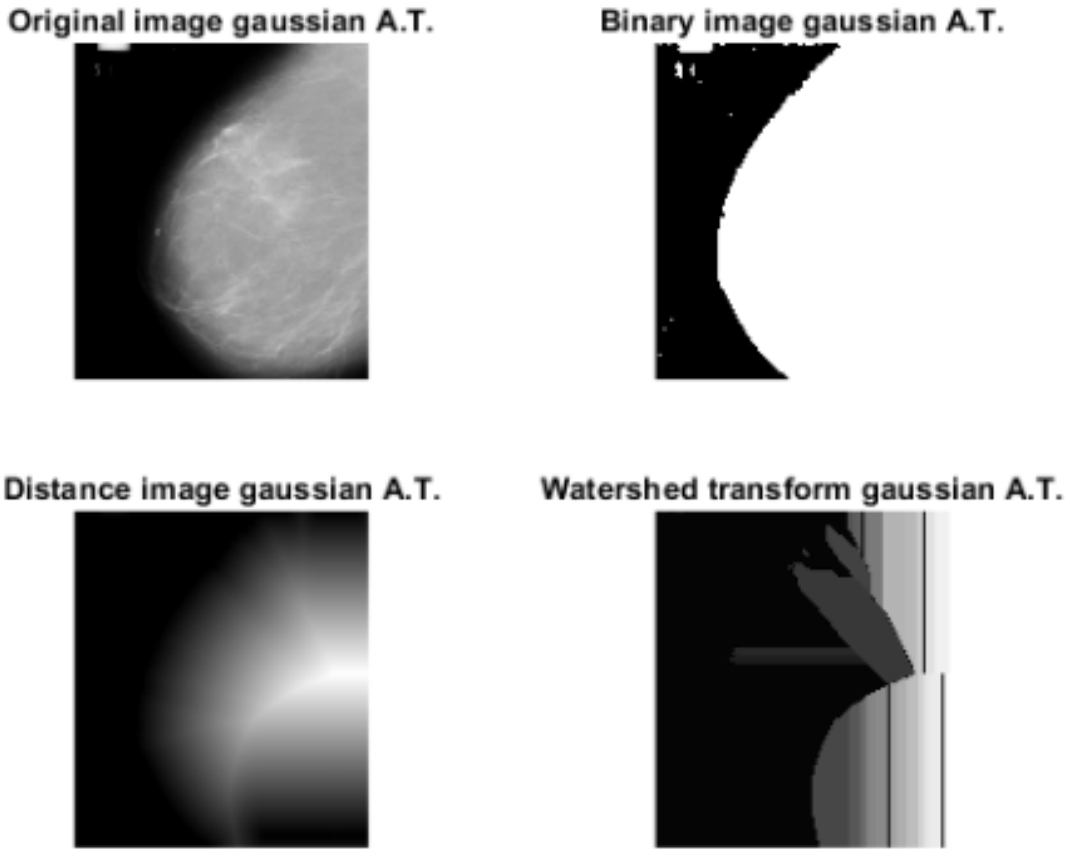


Figure 4
Top left: original image
Top right: binarized image with gaussian adaptive threshold
Bottom left: Distance Image
Bottom right: Watershed algorithm applied on image

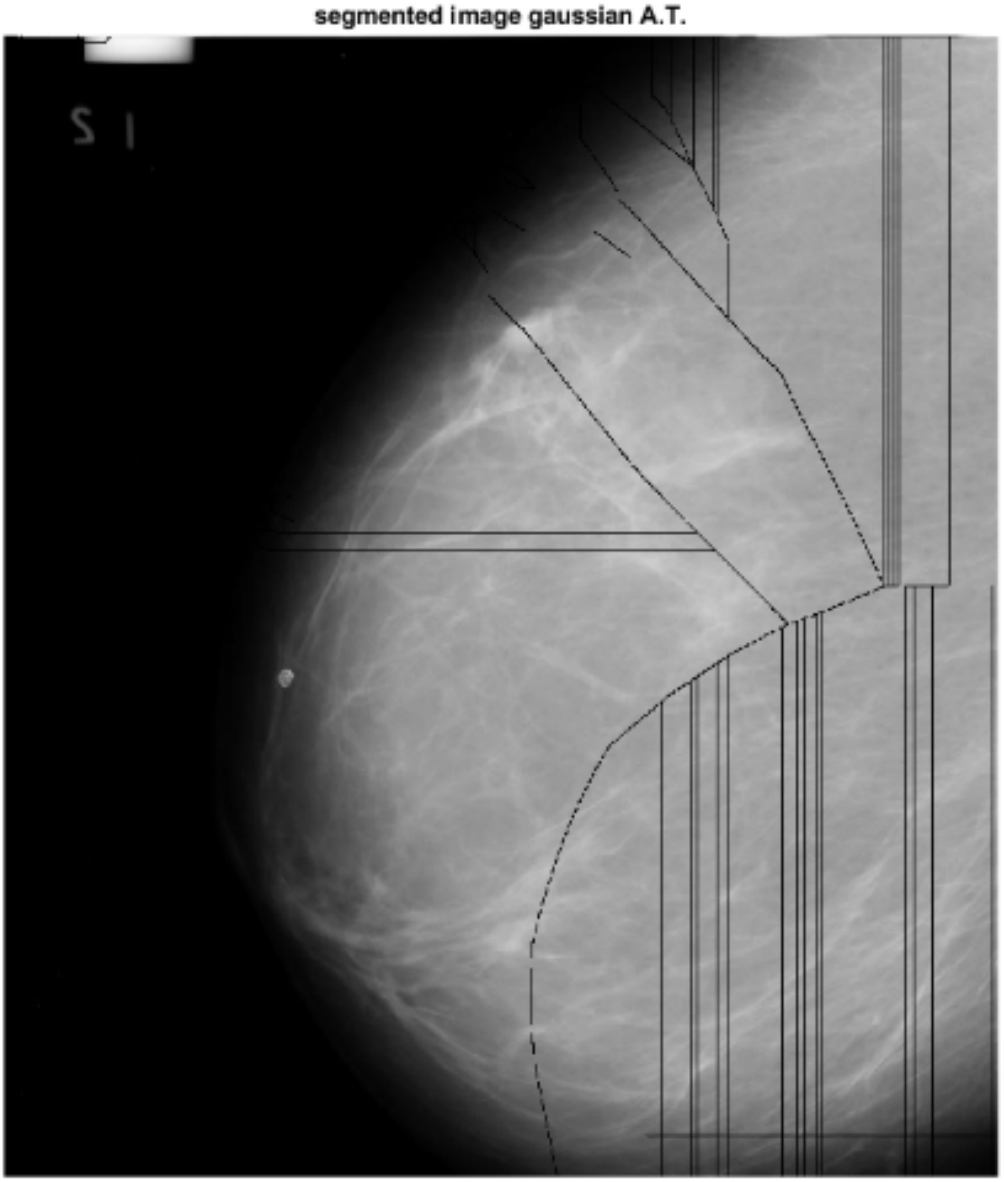


Figure 5

The following results follow the same method and code as above, but the threshold is different, in this case, the median adaptive threshold is used.

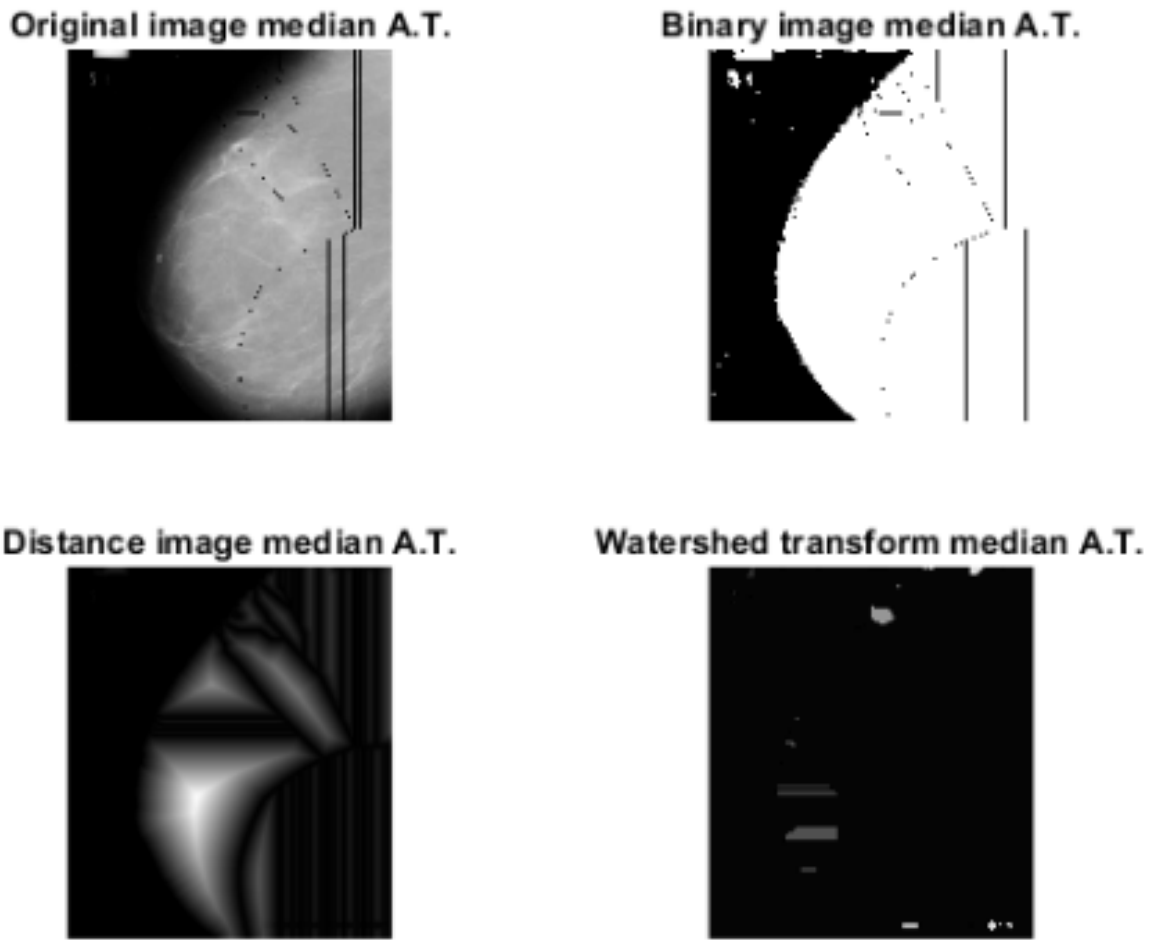


Figure 6
Top left: original image
Top right: binarized image with median adaptive threshold
Bottom left: Distance Image
Bottom right: Watershed algorithm applied on image

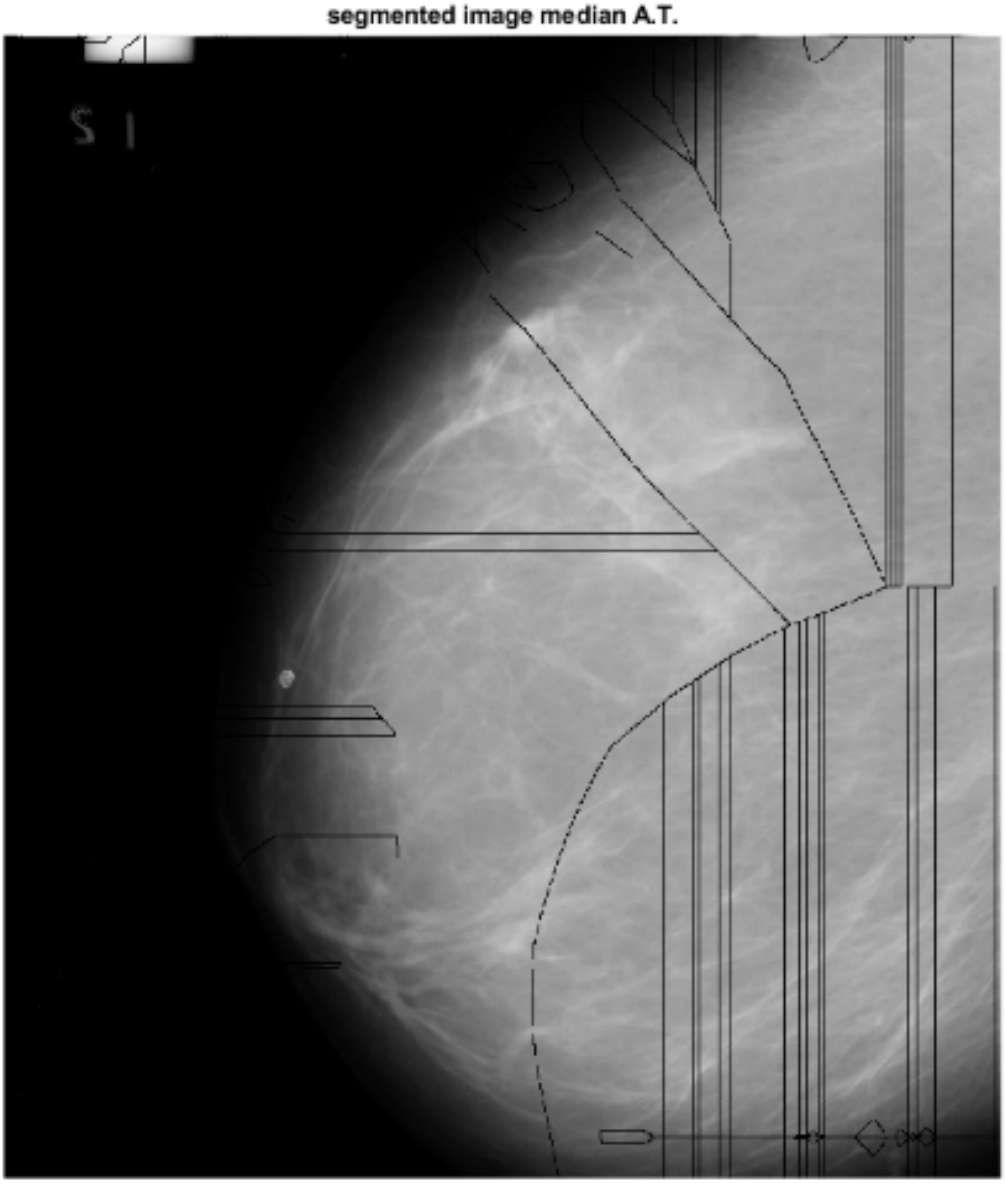


Figure 7

This time, a mean adaptive threshold was used in this watershed algorithm.

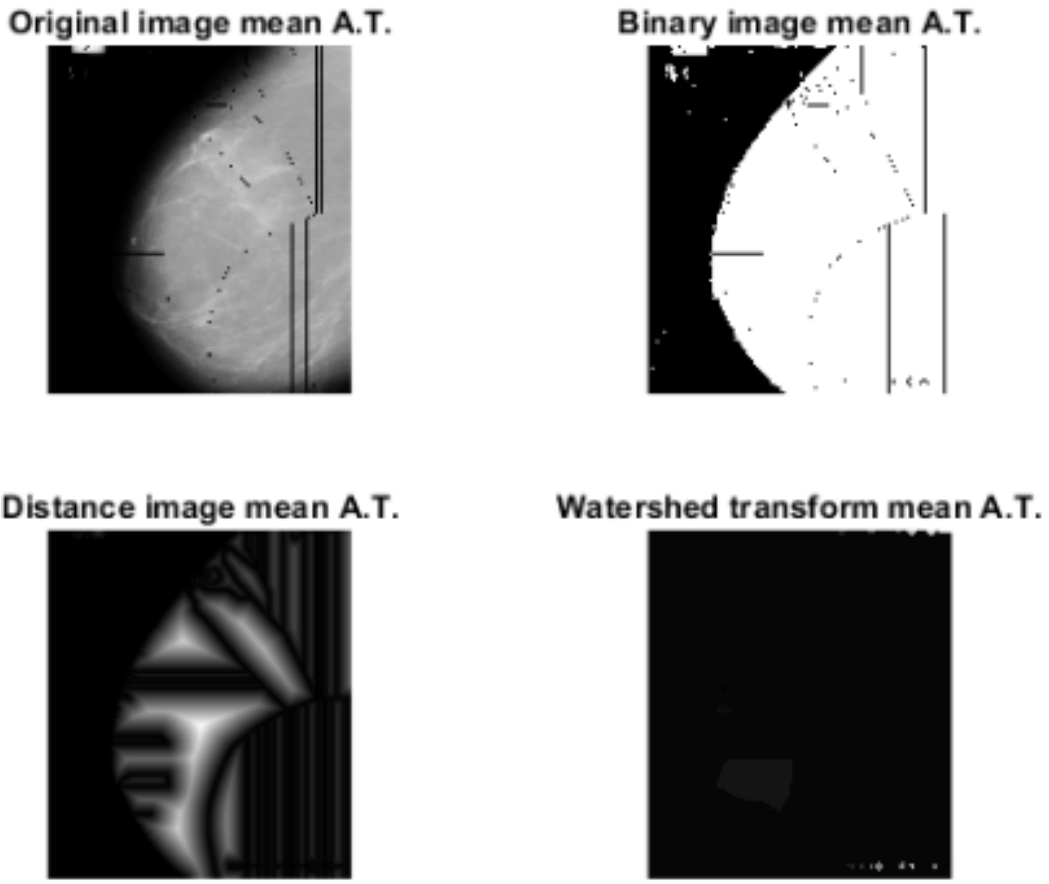


Figure 8
Top left: original image
Top right: binarized image with mean adaptive threshold
Bottom left: Distance Image
Bottom right: Watershed algorithm applied on image

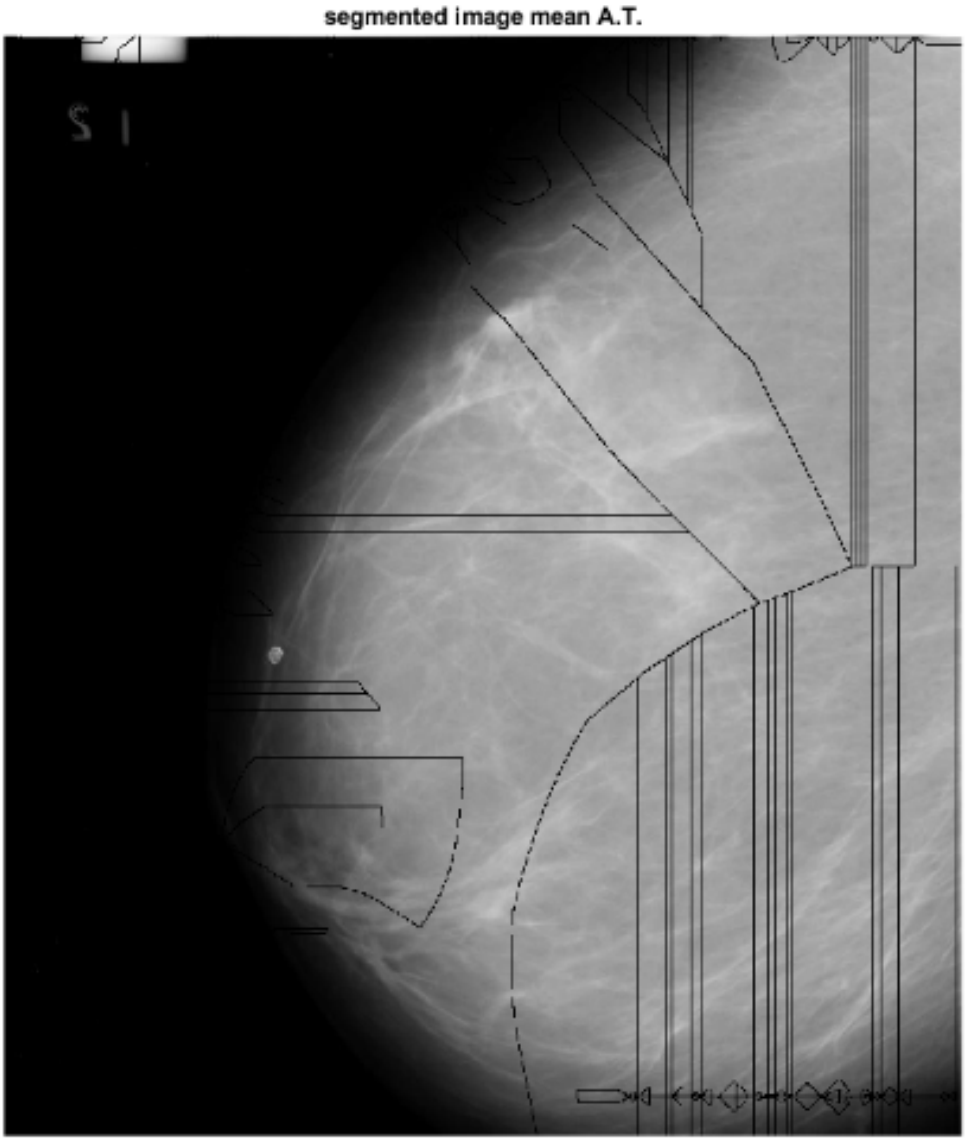


Figure 9

The whole procedure above, again, is just an overview of how the project would proceed from its start to end. This proof of concept only uses a dataset from a single source, whereas our project will be using datasets from different types of sources. Also, the algorithm used above, as mentioned before, is using a simplified version of the watershed method. Our project will be implementing an algorithm that is a bit more complex in nature.

3.2 Overview of User Interface of Software

The following is a rough process of what a user of our software would expect when they want to process a mammography image.

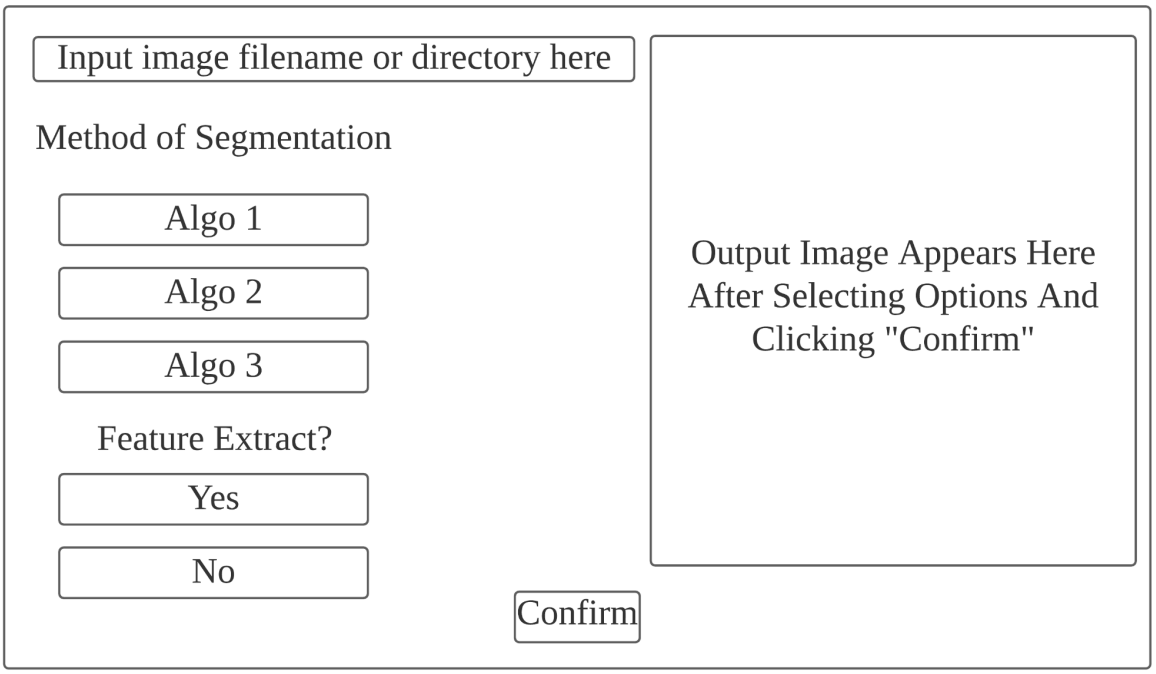


Figure 10 - User Interface

The user would enter the global directory of the image location into the input box. Then, the user would then select which algorithm they want the segmentation to be from the choices above (the diagram above shows that there are 3 choices, but depending on our final outcome of our project, the choices will be more or less than shown). Furthermore, there will also be a choice of whether the user desires to extract features from the mammography image - in this case it is to display the area of the tumor. Finally, the output image will be shown on the right hand side of the interface.

Reference

Berber, T., Alpkocak, A., Balci, P., & Dicle, O. (2013). Breast mass contour segmentation algorithm in digital mammograms. Computer methods and programs in biomedicine, 110(2), 150-159.

Domínguez, A. R., & Nandi, A. K. (2009). Toward breast cancer diagnosis based on automated segmentation of masses in mammograms. *Pattern Recognition*, *42*(6), 1138-1148

Pereira, D. C., Ramos, R. P., & Do Nascimento, M. Z. (2014). Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm. *Computer methods and programs in biomedicine*, *114*(1), 88-101.

Singh, V. K., Rashwan, H. A., Romani, S., Akram, F., Pandey, N., Sarker, M. M. K., ... & Torrents-Barrena, J. (2020). Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network. Expert Systems with Applications, 139, 112855.

Tsochatzidis, L., Koutla, P., Costaridou, L., & Pratikakis, I. (2021). Integrating segmentation information into CNN for breast cancer diagnosis of mammographic masses. *Computer Methods and Programs in Biomedicine*, *200*, 105913.

https://www.cancer.gov/types/common-cancers(retrieved 2021)

https://neptune.ai/blog/best-image-processing-tools-used-in-machine-learning(retrieved 2021)

https://kissflow.com/project/best-project-management-tools/ (retrieved 2021)

https://en.wikipedia.org/wiki/GitLab (retrieved 2021)

https://au.mathworks.com/products/matlab-drive.html (retrieved 2021)

https://en.wikipedia.org/wiki/MATLAB (retrieved 2021)

https://en.wikipedia.org/wiki/Python_(programming_language) (retrieved 2021)

https://en.wikipedia.org/wiki/C%2B%2B (retrieved 2021)

https://www.jotform.com/blog/dropbox-vs-google-drive/ (retrieved 2021)

https://au.mathworks.com/help/matlab/creating_guis/about-the-simple-guide-gui-example.html (retrieved 2021)