Visualisation and Topological Aspects of Higher Dimensional Data

Final Report for CS39440 Major Project

Author: Samuel Jackson (slj11@aber.ac.uk)
Supervisor: Prof. Reyer Zwiggelaar (rrz@aber.ac.uk)

April 24, 2015
Version: 1.1 (Draft)

This report was submitted as partial fulfilment of a MEng degree in Software Engineering (G601)

Department of Computer Science
Aberystwyth University
Aberystwyth
Ceredigion
SY23 3DB
Wales, UK

Declaration of originality

In signing below, I confirm that:

- This submission is my own work, except where clearly indicated.
- I understand that there are severe penalties for plagiarism and other unfair practice, which can lead to loss of marks or even the withholding of a degree.
- I have read the sections on unfair practice in the Students' Examinations Handbook and the relevant sections of the current Student Handbook of the Department of Computer Science.
- I understand and agree to abide by the University's regulations governing these issues.

Signature .....

Date .....

Consent to share this work

In signing below, I hereby agree to this dissertation being made available to other students and academic staff of the Aberystwyth Computer Science Department.

Signature .....

Date .....

Acknowledgements

I am grateful to...

I’d like to thank...

Abstract

Include an abstract for your project. This should be no more than 300 words.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Background & Objectives

1.1 Mammography

Breast cancer is the leading cause of death among women and is the most common form of cancer found in women [34]. Early screening of breast cancer using mammography has been shown to reduce the mortality rate of women [27, 35].

Mammography is the analysis of female breast tissue through the use of X-ray radiology with the goal of producing high resolution images of the structure within the female breast. The composition of the parenchymal patterns and tissue density revealed by in a mammographic evaluation can be used in the early detection of breast cancer.

Qualitatively speaking the composition of breast tissue can be split into four distinct categories. These are Nodular densities (corresponding to Terminal Ductal Lobular Units (TDLUs), linear densities (corresponding to ducts, vessels, and fibrous strands), homogeneous, structureless densities (corresponding to fibrous supporting tissue), and radiolucent areas (corresponding to adipose tissue) [38]. Typical markers used in the detection of cancer can are the presence of clusters of micro-calcifications, masses, architectural distortions, breast density and parenchymal patterns [24, 33].

There are two main projections used in X-ray mammography. These are known as craniocaudal (CC) and mediolateral oblique (MLO). Craniocaudal view is the ”bottom-up” view of the breast and is the best method for visualising the medial aspect of breast tissue [15]. The mediolateral oblique projection is a ”side-on” view of the breast that provides maximum visualisation of the breast tissue in its entirety but is limited in its ability to visualise the inner breast tissue [15].

1.1.1 Mammogram Acquisition

The difference between high risk and low risk breasts is inherently small in mammogram due to the low level tissue contrast between the two classes [23]. Because of this most essential part of mammographic imaging to ensure that a high contrast resolution is achieved in order to successfully capture the fine detail of the internal structures.

Mammopgraphic imaging has historically been carried using film but is now more commonly performed using digital mammography. Regardless of the medium the general technique for ac-

quisition remains the same. The breast is placed on a plate between the X-ray tube and the detector. The breast is compressed from its normal conical shape onto the plate. This improves imaging by helping to ensure the X-ray attenuation through the breast tissue is as uniform as possible.

Traditional film acquisition is hampered due to the film's sigmoid shaped response to x-ray exposure [23]. This can lead to under or over exposure of the film which in turn leads to poor contrast. Digital mammography does not suffer from this issue because the response curve is essentially linear.

1.1.2 Risk Assessment

Mammograms provide a non-invasive means to assess the risk of a patient developing cancer. Several different systems have been developed to aid the classification of mammographic risk based on the parenchymal patterns visible through X-ray mammography.

1.1.2.1 Wolfe

The earliest attempt to classify mammographic risk using parenchymal patterns was suggested by Wolfe [43]. Wolfe proposed a classification system which split patients into four categories depending on the relative visible density of fat, ducts and connective tissue. The four categories are described, in order of lowest to highest risk, in ref. [43] as:

- **N1** - Breast is mostly composed of fat with no visible ducts and very little amounts of dysplasia present.
- **P1** - The parenchyma is primarily composed of fat with up to one quarter of the breast density being composed of visible ducts in the anterior position which may extend into a quadrant.
- **P2** - Breast indicates prominent duct pattern beyond one quarter of the breast that can occupy the entire parenchyma.
- **DY** - Breast is characterised by a severe increase in breast density and often appear as homogenous, missing the duct pattern present in P2 breasts.

1.1.2.2 Boyd

Boyd et al. [6] proposed a quantitive assessment of risk based on increasing classes of mammographic density, known as the six class categories (SCC). These classes are based on the proportion of dense tissue relative to the area of the breast. The six classes are:

- 0%
- >0 to $<10\%$
- 10 to $<25\%$
- 25 to $<50\%$

- 50 to <75%
- $\geq$ 75%

1.1.2.3 Tabár

Tabár et al. [17] proposed a classification scheme which classifies a breast based on the percentage presence of the four building blocks of breast composition [17, 38]. The description of each of the five patterns is given as:

- **Pattern I** - Breast corresponding to pattern I exhibit scalloped contours and cooper's ligaments with evenly scattered TDLU's.
- **Pattern II** - Complete fatty replacement of both
- **Pattern III** - Prominent retroareolar duct pattern and fatty involution.
- **Pattern IV** - Extensive linear and nodular densities present throughout the parenchyma.
- **Pattern V** - Homogeneous, structureless fibrosis with a convex contour.

1.1.2.4 BI-RADS

The Breast Imaging Report and Data System (BI-RADS) [4, 13] was developed by the American College of Radiology (ACR) in an attempt to standardise the lexicon used to describe mammography reports during standard screening. BI-RADS classifies the breast into four categories based on density [4].

1. Fatty Breast (<10% of dense tissue)
2. Fibroglandular (10 - 49% of dense tissue)
3. Heterogeneously dense (49 - 90% of dense tissue)
4. Homogeneously dense (>90% of dense tissue)

A radiologist will then classify the breast according to one of 7 categories after interpretation [4]. These are one of:

- Incomplete. Additional evaluation needed.
- Normal.
- Typically benign.
- Probably benign. A shorter interval follow-up is recommended.
- Suspicious Abnormality. Biopsy considered.
- Highly suggestive of malignancy. Biopsy should be performed.
- Histologically proven malignancy.

1.2 Features

Features are higher level descriptive abstractions computed from lower level structure such as areas of high intensity, edges, and corners present within an image. Features are the result of computing a descriptive property about an image from the intensity information contained within it. Hundreds of different types of features have been proposed [10, 16]. Broadly speaking, image features can be split into three distinct categories: shape (or morphological), intensity, and texture features. A brief review of some examples of each type are given in this section.

1.2.1 Shape Features

Shape features are features which detected within an image based on the morphological properties of a region of interest (ROI) within an image. The ROI could either be a suspicious artefact present in the mammogram, such as a malignant or benign tumour or cluster of micro-calcifications [10], or it could simply be a property present in the parenchymal structure of the mammogram (which is the approach used in this project) such as regions of high density tissue [9] or prominent linear structures [45].

Typical example of features extracted from shapes of an ROI are the area covered by the shape, the circularity or rectangularity of the shape, and measures based on the perimeter of the shape [30]. Additional features based on the normalised radial length (NRL) [22] such as boundary roughness, mean, entropy and area ratio [30, 31], and statistics based on normalised chord length [44] such as the first four statistical moments of the resulting chord distribution [14].

1.2.2 Intensity Features

Intensity features are simple descriptive statistics based on the grey-level histogram of an image. Examples of such features are the mean, standard deviation/variance, skew, and kurtosis of the grey-level histogram of an ROI [10, 11].

1.2.3 Texture Features

Texture features are measurements of an image based on the repetition of patterns over an ROI [28]. Features based on the texture of an image are highly desirable because mammograms are obtained using a single medium of acquisition and the spatial distribution of features can be found in a single band [16].

A set of commonly used set of texture features can be derived from the grey-level co-occurrence matrices (GLCM) [18]. Grey-level co-occurrence matrices are used to describe the positions of pixels having similar grey-level values [28]. Pixel pairs within the GLCM are defined over a range of distances between pixels (often simply one) and orientations (often 0° , 45°, 90°, and 135°). From GLCMs computed for each direction and orientation features representing texture context information such as contrast, homogeneity, energy and entropy [10, 18, 28] can be calculated.

Another approach to texture features uses the grey level difference statistics of a of an image vector. This takes the form of the absolute difference between pairs of grey levels within an image or between the average of local neighbourhoods. Like GLCM features, this technique is

parameterised by the distances between the two image patches and the orientation of the distances used as an offset. First order statistics and measures of spread (such as entropy, contrast and angular second moment) can then the derived [42].

One final approach to mammogram texture analysis is the use of a grey level run length statistics. This technique counts the number of consecutive occurrences of pixels with the same grey level value in a given orientation. Features are based on weighting longer or shorter runs as being more significant can then be created from the run length distribution [10, 42].

1.3 Dimensionality Reduction

Collections of features extracted from a set of images form a feature matrix (also known as as feature space) where each row in the matrix corresponds to a single image and each column corresponds to a single feature extracted from that image. Each entry in the matrix contains the value of a specific feature detected for that image. The number of columns in a feature matrix m is known as the dimensionality of feature matrix. The goal of a feature matrix is that it encapsulates key information about what we are aiming to measure, allowing us to make inferences about what the data is telling us.

The field of dimensionality reduction is concerned with reducing the number of dimensions that a dataset has to only preserve the most important ones. In this way dimensionality reduction can be viewed as a feature selection method, discarding irrelevant or noisy dimensions in favour of those which best represent the data.

1.3.1 Curse of Dimensionality

For some applications the dimensionality of a feature matrix may be quite small. Indeed it is trivial to ascertain the relationships between the features in a matrix where m is equal to 2 or 3 by pure visualisation. However this is not the case when the feature matrix contains higher dimensional data where m may be in the order of 100s or 1000s of features.

As the number of features (dimensionality) increases so does the volume of the feature space in which the data points are contained. This causes issues with any technique the requires statistical significance because a fixed number of data points will indefinitely become sparsely distributed as the dimensionality of the space increases. This is known as the curse of dimensionality [5].

This causes havoc with algorithms (such as classifiers) that depend on distance metrics such as Euclidean distance. The effect is such that in a very high dimensional space the distance pairs of data points becomes negligible in computations such as the nearest neighbour algorithm. All data points are effectively equally far away so the choice essentially become one of randomness [12].

However, often it is not the case that all of dimensions of a feature space are required in order to obtain a meaningful representation. Quite often the the data points in a feature space correspond to a lower dimensional manifold that is embedded within the higher dimensional feature space. Dimensionality reduction (also referred to as feature selection) techniques can be used to disregard noisy or irrelevant dimensions while still retaining meaningful information present in the higher dimensional space.

1.3.2 Linear

Linear approaches to dimensionality reduction make the assumption that the data in question lies approximately on a linear subspace within the higher dimensional representation. The core idea behind linear dimensionality reduction algorithms is to find the basis vectors representing the lower dimensional subspace.

1.3.2.1 Principle Components Analysis

Principle components analysis aims to preserve the maximal covariance of the data [37]. First the covariance matrix of the feature space is computed via:

$$C = \frac{1}{n}\Sigma \mathbf{x}_i \mathbf{x}_i^T \tag{1}$$

Where C is the co-variance matrix and $\mathbf{x}$ is the feature matrix. Eigendecomposition is then performed on the covariance matrix and to obtain the basis vectors for the subspace. Taking the first n basis vectors creates a linear projection of the data onto a lower dimensional representation.

1.3.3 Non Linear

Often data points within the feature space do not lie on a linear subspace. Instead they are embedded in a non-linear manifold within a higher dimensional space. In this case taking a linear projection such as that produced by PCA will result in a distorted results with many parts of data overlapping with one another. However, linear techniques such as PCA are often useful as a preprocessing step to remove highly redundant or noisy dimensions before using a non-linear dimensionality reduction method.

1.3.3.1 t-Distributed Stochastic Neighbour Embedding

t-Distributed Stochastic Neighbour Embedding (t-SNE) [40] is a non-linear dimensionality reduction technique which aims to preserve the distance between very similar data points. t-SNE models the high and low dimensional representations as joint probability distributions modelled using Student's t-distributions. The algorithm optimises the mapping by minimising the Kullback-Leibler divergence between the two distributions. In this way it models the probability that a higher dimensional point p would choose q as its neighbour. t-SNE can be used both with Euclidean distance or with other distance metrics.

1.4 Visualisation

Traditionally plots of related low dimensional variables can be visualised directly using two or three dimensional scatterplots. However, direct visualisation of data with more than two or three dimensions is impossible with traditional methods. While dimensionality reduction techniques can be used to reduce the dimensionality of data down to two or three dimensions, it is unlikely

that two or three dimensions are enough to accurately capture the structure of a complex higher dimensional manifold.

While a higher dimensional dataset cannot be directly visualised in the traditional sense, several different visualisation techniques have been developed that can be used to examine the structure present in higher dimensional data. This sections provides a brief review of several of these techniques.

1.4.1 Scatterplot matrix

A scatterplot matrices are possibly the simplest form of higher dimensional plot. Traditional scatterplots show the correlation between two variables. A scatterplot matrix is an $n \times n$ matrix of scatterplots where each individual plot shows the correlation between two variables i.e. the plot in row i and column j will show the plot of variables X_i and X_j

1.4.2 Parallel Coordinates

Parallel coordinate plots [19] show each variable as a single axis along the bottom of the graph. A single point on the plot is represented as a piecewise line across each of the axes. The position at which a point intersects an axis is the value of that the n-dimensional data point has for that variable.

There are some limitations to parallel coordinate plots. The principle limitation is that the effectiveness of the visualisation is dependant on the ordering of the axes. Different orderings will produce different visualisations. Sometimes the data naturally has a specific ordering (such as time-series data), however this is often not the case. Tatu et al. [39] experimented with using the Hough transform to automatically infer "good" axis orderings. Johansson et al. [20] investigated ranking axis orderings using a combination of clustering, correlation and outlier features on the data. The scaling of each axis is also an important factor in the visualisation, however this can be solved by rescaling the data before visualisation.

1.4.3 Andrews Plot

An Andrews plot defines each n-dimensional data point as a Fourier series with n terms. This visualisation is essentially the same as a parallel coordinates plot with each of the data points interpolated with Fourier interpolation. As such it suffers from the same limitations as parallel coordinates.

1.4.4 RadViz

RadViz [25] is a technique for visualising data points on a plane. RadViz models each n dimensional data point as a single point on a 2D plane surrounded by a unit circle on which each of the n features are placed equal distance apart from one another. Each data point is virtually "connected" to each of the points on the unit circle via a "spring". All of the springs are in equilibrium with one another and the final resting point for a point in the visualisation corresponds to the total strength exerted over point by each spring.

1.5 Quality Measures

TODO

1.6 Analysis

This project aims to investigate the effect of mapping the feature space of both real and synthetic mammograms to a lower dimensional representation. This will involve the extraction of a variety of different shape, intensity, and texture features to produce a high dimensional feature space. We will then use dimensionality reduction techniques to produce a lower dimensional mapping and visualise the results using an appropriate technique.

The hypothesis that we are aiming to test in the project is: Does the lower dimensional mappings of the higher dimensional feature space of synthetic mammograms match those of real patients? The high level questions that this project aims to answer are:

- Are features which are important to real mammogram comparable to those of synthetic mammograms?
- If not, why are they different and does this relate to the author's own limitations?
- Can the knowledge gained be used to influence the perception of what features are important in a real mammogram?
- Could it be used to suggest how to build new mammogram models?

The technical work of the project will be concerned with implementing some existing approaches to feature extraction and dimensionality reduction in mammographic image analysis, but applying it to synthetic mammograms in order to evaluate the similarities and differences.

There are four core components of the methodology that need to be considered before implementation. The first is a decision on the number and type of features to be extracted from the images. Ideally the method used in this project should incorporate approaches that cover all three categories discussed in section 1.2. This is so that a good variation in the discriminative properties of an image will be incorporated into the system. Focussing on only one technique for features considerably reduces the information that could be used to discriminate between images.

The second is the choice of dimensionality reduction techniques used to produce a lower dimensional mapping. Many different methods have been proposed for dimensionality reduction. It is highly likely relationship between variables in the feature space will not be a linear subspace. For this reason the chosen technique will almost certainly be a non-linear dimensionality reduction technique. However, it may be the case that the a linear dimensionality reduction technique such as PCA can be used to remove extremely redundant dimensions and reduce noise as a preprocessing step to a non-linear dimensionality reduction technique.

Thirdly the lower dimensional mapping must be visualised. This may be done either directly if the mapping is to two or three dimensions. Alternatively the a higher dimensional plot such as those discussed in section 1.4 could be used to higher examine the dimensional spaces. The choice of visualisation techniques will depend heavily on the choice of the two prior components.

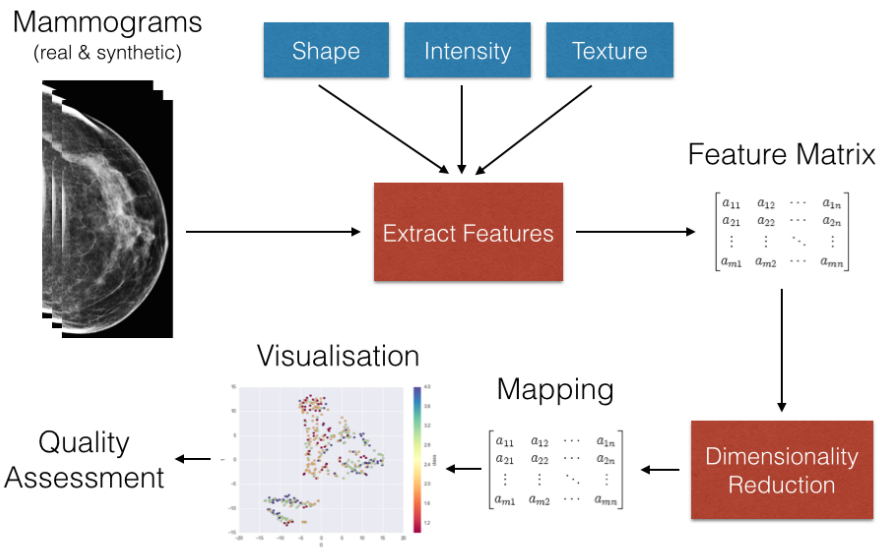

Figure 1.1: Conceptual overview of the image analysis pipeline to be produced as part of this project.

It is likely that the project will use a combination of different visualisation strategies depending on what specific aspect of the dataset is being examined. One of the issues with higher dimensional data is that it is impossible to produce visualisation that accurately capture all aspects of the data at once. Whatever technique is used will only ever show a projection of the feature space at best.

The final component of the system will be a form of quality measurement. While visual examination of the mapping is a necessity for qualitatively interpreting how the mapping is derived from the feature space, it is important to gain a quantitive measurement of the mapping. Quantitive measures can be used to check parameters for the dimensionality reduction algorithm are well configured and compare can compare between different combinations of features.

A technical decision must also be made regarding the choice of technologies and programming languages to be used as part of the project. All four components of the system are likely to be non-trivial to implement. Where possible existing implementations of parts of the four major components of this project should be used. The reasons for this are threefold. 1) using existing implementations should ensure that the development time is kept to a minimum, 2) they are likely to be better tested for validity, and 3)

1.7 Research Method

The research method used in this project will be an experimental one. There will be an initial exploratory phase in which features, dimensionality reduction algorithms, visualisations and quality measures are tried and tested. Modifications will be made to existing approaches if necessary. Once an adequate implementation of both features, dimensionality reduction techniques, visualisation, and quality analysis has been achieved I will carry out the experiment with a subset of the real and synthetic mammogram datasets. I will use the largest possible subset from the real and synthetic datasets that I can as part of the experiment in order ensure the best possible representa-

tion of the feature space without over representing the subjects.

1.8 Development Methodology

The methodology for the implementation of the project will draw core on ideas from the agile development methodology. This software development methodology is the only one which makes sense for a research orientated project. Research orientated projects are by definition exploratory in nature and the results of the project cannot be stated in concrete at the start of the project. This rules out methodologies like waterfall and feature driven because in both systems the end result of development must be clearly stated up front, with little room for manoeuvre.

Agile approaches embrace change. This is a desirable attribute in a project where we can define the high level end goals of development but where the specifics must be flexible towards finding an approach that works based on analysis of the results from initial prototyping.

As this project is an individual endeavour the full power of an agile approach cannot be fully realised because some of the core principles rely interactions between multiple individuals (such as daily stand-up meetings and paired programming from XP).

There are however many other benefits to an agile approach that are relevant to an individual project. Test driven development (TDD) is a concept that is highly relevant. In software development TDD is used to ensure that you have confidence to make changes and verification that what you have written works. This is especially relevant and desirable in research projects because it not only gives you verification that the software is working, but can also be used to verify that the output of the program and therefore results are correct.

Short iterations and small releases that incrementally add value is another concept that is relevant to an individual project. This provides a measurable indicator of progress throughout the project. In this project I will aim to produce weekly iterations where I will select several stories to work on throughout the week. I will time the start of an iteration to coincide with the meeting with my postdoctoral supervisor who will effectively act as a “customer” to the project. Work for the week will be decided based on discussions during these meetings so this naturally defines an anchor for iteration beginnings and endings.

Programmatic idioms associated with XP are also of value to an individual project. Simplicity (especially the concept of YAGNI) and heavy refactoring are likely to feature heavily in this project. Simplicity ensures that we keep the focus of the project limited to the goals of the research question without adding superfluous code that doesn’t contribute to answering the questions outlined the problem analysis. Refactoring will allow the design of the project to be incremental. We can start with an initial basic design and rewrite and modify the structure of the code when it is needed in order to accommodate problems encountered or other change. In this way a good design (where verification of correctness is controlled by TDD) becomes a natural byproduct of development.

Chapter 2

Experiment Methods

2.1 Overview

The technical outcome of this project was to produce an image analysis pipeline. Broadly speaking the pipeline can be broken into four distinct components. These are feature detection and extraction, dimensionality reduction, quality evaluation, and visualisation.

2.2 Techniques

2.2.1 Preprocessing

Very little preprocessing of the images has been used in this project. Each of the mammograms used has an accompanying binary mask which is used to remove the background and pectoral muscle (see section 2.3). The skin around the edge of the breast can cause issues with the blob and line detection due to the intense response during image acquisition. As we are only interested in structure within the parenchyma binary erosion is performed on the breast masks using a disk shaped kernel with a radius of 30 pixels.

2.2.2 Features

In this project we have used three different types of image features to build a feature space from the mammogram datasets. We have used two different types of shape features one to detect blobs and one to detect linear structure. These features intrinsically define a ROI which has some parenchyma pattern of interest. From the ROIs defined by these features we can extract intensity features (based on the image histogram of the ROI) and texture features (in the case of this project GLCM features).

2.2.2.1 Blob Features

The blob detection was achieved by following an approach similar to that described by Chen et al. [8,9]. This is a multi-scale approach based on building a Laplacian of Gaussian (LoG) pyramid

over ten different scales.

To obtain a multi-scale view but retain comparative performance (due to the very large size of mammographic images) instead of increasing the size of the kernel the sigma of the Gaussian is fixed to 8.0 and the image is smoothed and downscaled by a factor or $\sqrt{2}$ instead. Once the image has been convolved with the LoG kernel, the resulting image is finally upscaled to full size before peak detection.

Each of the mammographic images in the real dataset come with a set of breast masks which segment the tissue of the breast from the result of the image (such as the pectoral muscle). This helps to ensure that we are detecting the only within the parenchyma but causes issues due to a large edge response produced from the LoG kernel around the edge of the breast. In Chen's thesis this effect was dealt with by means of a "deformable" convolution. The image is convolved with a standard image kernel when the area of the kernel is entirely within the mask area. When the kernel being convolved falls outside of the mask the filter kernel is modified so that it returns zero outside of the mask and the LoG response normalised by the number of nonzero pixels otherwise.

For each scale image produced the local maxima are detected using maximum filter and a conservative threshold. The resulting position of the peak defines the location of the blob detected in the image while the effective sigma of the Gaussian for the scale of the image σki (where i is the scale and k is the downscale factor equal to $\sqrt{2}$) is the radius of the blob.

This procedure returns a very large number of responses with many overlapping detections. Since we are only interested in blob that best characterises the detected peak, a blob merging strategy is employed to remove redundant blobs. As we are only concerned with blobs that characterise patterns within the parenchyma all blobs whose radius falls outside of the bounds of the image mask are removed.

Next a thresholding technique is used to remove blobs that fall below a certain level of intensity. The area of the image covered by the blob is categorised into 9 clusters. The average intensity of each cluster is computed and the top 5 clusters are selected. The threshold used is defined to be the average intensity of the top five most intense clusters across all blobs less the standard deviation of those same clusters.

In Chen's thesis the clustering is performed using the Fuzzy c-means algorithm while in my implementation I have used the k-means algorithm. The result of this is that the algorithm requires more clusters (5 compared to only the top 3 clusters in Chen's thesis) to achieve comparable results.

After these operations the number of blobs detected is significantly reduced but there are still a large number of blobs which significantly overlap one another in high density regions. To achieve a better representation of the distribution of high density tissue in the mammogram blobs are merged according to how they interact one another. The intersections are classified into one of three categories:

- External: $d \geq r_A + r_B$
- Intersection: $r_A - r_B < d < r_A + r_B$
- Internal: $d \leq r_A + r_B$

Where d is the distance between the two blobs and the r_A and r_B are the radii of blobs A and B. The above definitions assume that $r_A \geq r_B$.

Merging proceeds as follows: if a blob is external then it remains retained. If a blob is internal to a larger (coarser) blob it will be removed. If the blobs intersect one another and they are closely located ($d \leq r_A + \alpha r_B$ for $0 \leq \alpha \leq 1$, assuming $r_A \geq r_B$). In the experiments documented in the report overlap parameter α used was 0.01.

For each of the blobs detected, the coordinates of the detected blob and the radius (i.e. the sigma of the Gaussian associated with the scale the blob was detected at) are saved to a comma separated value (CSV) file. This file was then loaded into the IPython notebook system for analysis. Using the analysis python module developed as part of the general pipeline framework the following features were calculated from the radius and position of the blobs for each image:

- Number of blobs detected
- Average radius
- Standard deviation of the radius
- Min & max radius
- Small/medium/large radius count: For the radius was binned into three separate equally sized bins across the range scales used to detect blobs.
- Density: the average distance between this blob and the k nearest blobs. In all experiments k was set to 4.
- 25/50/75 percentiles
- Count of blobs above the mean.

2.2.2.2 Line Features

Along with shape features based on blobs of high intensity a shape feature based on finding linear structure within a mammogram was used. Linear structure aims to try and characterise the ductal shapes visible in a typical mammogram. For the implementation of this feature we follow the work of Zwiggelaar et al. [45] and use an orientated bins method to pick out ROIs which may otherwise be missed using just blob features alone.

The orientated bins method proposed in ref. [45] filters the local neighbourhood of an image by dividing the neighbourhood window into n angular bins with an angular resolution of $\frac{2\pi}{n}$. The line strength of the local neighbourhood is given by computing the difference between the maximum average intensity of opposing bins and the average intensity of the local neighbourhood as a whole. In the case of a well defined line only one set of opposing bins would give a high response. The orientation of the image is given by the orientation of the maximum bin.

Once the line image for has been generated the results can be enhanced by applying non-maximal suppression [36] to the line image to strengthen the detected structure. After suppression the image is once thresholded using a conservative value to remove noise and the image is converted to a binary image. A morphological closing operation is used to improve the connectivity of the line shapes detected from the image. Any points which are within an 8-connectivity of one another are counted as being part of the same structure.

The resulting shape features can be measured by standard first order statistics to produce descriptive features about the blobs and lines detected from the mammogram. As with blob features the detections were exported to a CSV file and then loaded into a IPython notebook for further analysis. Features created from the distribution of areas detected by the line features were:

- Number of lines detected
- Average area
- Standard deviation of the areas
- Min & max area
- 25/50/75 percentiles
- Count of areas above the mean.

2.2.2.3 Intensity

The shape features detected using orientated bins and multi-scale blobs define regions of interest across the breast. From these ROIs the patch of the image which is covered by the area or radius of the shape feature can be extracted. The histogram of the intensity values of this image patch provide can provide additional discriminative information about ROI. Descriptive statistics derived from the histogram of the ROI were computed. The final features derived were:

- Number of intensity values
- Mean
- Standard deviation.
- Min & max values
- 25/50/75 percentiles
- Skew
- Kurtosis

2.2.2.4 Texture

As with intensity features, texture features can be extracted from the patches defined by the shapes features as well. In this project we have only used texture features derived from the grey-level co-occurance matrix [18]. The properties computed from the co-occurrence matrices were homogeneity, dissimilarity, energy, and contrast. The definitions of each are given as the following:

Homogeneity:

$$\sum_{i,j=0}^{levels-1} \frac{P_{i,j}}{1+(i-j)^2} \tag{1}$$

Dissimilarity:

$$\sum_{i,j=0}^{levels-1} P_{i,j}|i-j| \tag{2}$$

Energy (or the square root of the angular second moment):

$$\sqrt{\sum_{i,j=0}^{levels-1} P_{i,j}^2} \tag{3}$$

Contrast:

$$\sum_{i,j=0}^{levels-1} P_{i,j}(i-j)^2 \tag{4}$$

The experiments performed in chapter 3.1.3 all used a distance of 1 and a combination of eight different orientations of angles 0.0, 22.5, 45.0, 67.5, 90.0, 112.5, 135.0, 157.5 degrees.

2.2.3 Dimensionality Reduction

In this project the t-SNE as the dimensionality reduction algorithm with which to produce the lower dimensional representations from the higher dimensional feature spaces. The implementation we have used as part of this project is the standard algorithm available through the scikit learn library.

Before running the dimensionality reduction algorithm the input feature matrix is standardised by removing the mean from each feature and scaling to standard variance. This ensures that the all of the features will roughly on the same scale as one another and therefore that the distance metric used to compute neighbours shouldn't be heavily weighted in favour of feature orders of magnitude larger than the others.

t-SNE was chosen as the primary dimensionality reduction algorithm for use in this project because 1) it generally produces reasonably good visual representations when reducing data to 2 or 3 dimensions and 2) it aims to preserve the local neighbourhood to produce representations where points close in the higher dimensional space will be close in the visualisation. This is useful because it can be used to visually show whether the points corresponding to synthetic mammograms appear close to the real mammogram images.

Unfortunately, a major limitation of the t-SNE algorithm is that it make no attempt to preserve the global structure of the underlying manifold in higher dimensional space. For this reason the mappings produced for t-SNE cannot be used to infer anything about the global structure of the manifold.

2.2.4 Visualisation

The visualisation aspects of this project have largely been handled by the inbuilt functionality that available in the matplotlib and pandas Python libraries. However I have also implemented some additional custom visualisation techniques for examining what the images look like for each point in the lower dimensional mapping.

2.2.4.1 Visualisation of Images from Mapping

In order to examine the how images change across the lower dimensional mapping and to try and understand why images are grouped closely together I created a small utility that would read in the mapping of the feature space and display a scatter plot of the mapping. When hovering over each point in the visualisation the image corresponding to that data point is displayed to the right of the scatter plot.

2.2.4.2 Median Image Plot

Similarly to the previous visualisation this visualisation takes a projection of the feature space and creates a two dimensional histogram of the points. From each of the resulting bins the median point in both the x and y directions if found. The image which corresponds to this point is selected to be used as part of the visualisation. Each of the images is stitched together to form a matrix of images in the same shape as the lower dimensional mapping.

2.3 Datasets

In this project we are using two different datasets. One consisting of mammograms for taken from a collection of real patients. The other dataset is a collection of artificially generated breast phantoms generously created by the University of Pennsylvania.

2.3.1 Real Data

The dataset of real mammograms was taken from a private dataset captured using a Hologic full field digital mammography system. The dataset contains images of 90 unique patients each with a craniocaudal and mediolateral oblique view of both the left and right breasts, resulting in 360 total images. Each image in the dataset also has a corresponding binary breast mask. This mask is used to segment the background and the pectoral muscle from the breast parenchyma.

2.3.2 Synthetic Data

The synthetic breast phantoms were generated by the University of Pennsylvania using the techniques outlined by Bakic et al. [1–3]. Their simulation system consists of three major components: a breast tissue model, a compression model, and an acquisition model. Adipose tissue compartments within the breast are modelled using thin shells in areas of primarily adipose tissue and as blobs in areas of predominantly fibroglandular tissue. Ductal lobes are also simulated by the model using a randomly generated tree. As the phantoms have not been formally assigned a BI-RADS by an expert radiologist (as in the case of the real mammogram dataset) the ground truth for the risk associated with a particular mammogram is given by it's volumetric breast density (VBD). This information was supplied in the meta-data produced alongside the phantom mammograms.

2.4 Implementation

This section provides a brief overview of the implementation details used in the project. All of the methods discussed in the preceding sections were implemented in Python. The technical output of the project is a Python library and command line tool which is used to create the pipeline discussed in the first part of this chapter.

2.4.1 Python Package

The majority of the components used in the project are built upon the top of the scipy stack [21]. Two major additional libraries (which also rely on the scipy stack) which have been used heavily in the project are the scikit image [41] and scikit learn [29] projects.

The library created as part of this project forms a complete python package. The package consists of five top level modules and a collection of submodules implementing specific functionality relating to the features detected by the system. The description of each of the modules is as follows:

- **reduction**: The reduction module implements functions performing multi-processed feature extraction from a dataset.
- **analysis**: The analysis module implements commonly used functions used to analyse the images after feature extraction has been performed. These functions are typically called directly from the python interpreter to in a IPython notebook session.
- **plotting**: This module provides a collection of custom, convenience plotting functions that largely depend on the matplotlib library.
- **io_tools**: The io_tools module implements functions for iterating over directories of images and there corresponding masks as well as image loadings and preprocessing functions.
- **utils**: The utils module contains a collection of miscellaneous helper functions used in various places throughout the package.

An additional sub-package contains the modules used by the reduction module to perform feature extraction these modules include:

- **blobs**: Contains the code implementing blob detection using the approach outlined in section 1.2. This also uses an additional private module which provides the code for the graph built as part of the blob merging scheme.
- **linear_structure** Contains the code implementing line feature detection using the approach outlined in section 1.2. This also uses a couple of additional private modules within the sub-package which provide the code for non-maximal suppression and the orientated bins feature.
- **texture** Contains the code for computing the GLCM and texture features from ROIs.
- **intensity** Contains the code for computing first order statistics from ROIs.

The last part of the library is the deformable convolution module. This implements the deformable convolution approach outlined by Chen et al. [9]. Due to the performance considerations associated with convolution I chose not to write this module directly in Python but to implement it as a C function which is compiled using the Python C API.

2.4.2 Command Line Interface

The command line tools offered by the program are built using the Click library [32]. The CLI provides a thin wrapper to the some of the higher level library functions available in the package. The most important commands offered by the CLI are those concerned with image processing to collect features from the images. These functions involve iterating over a folder of images and applying the feature extraction techniques outlined in section 1.2. These operations can take in the order of several hours to complete and so it is useful to have them exposed on the command so they can be run and left until complete. These commands for image reduction also add the ability to automatically dump the output of the reduction to a CSV file named by the user upon completion.

Other useful commands included in the CLI interface are the ability to run and plot the output of shape feature detection from a single image and the running the t-SNE algorithm on a feature dataset from the command line.

Chapter 3

Results and Conclusions

This chapter outlines the results of the experiments performed with the system using the methods outlined in the preceding chapter. This section is split into two sections. The first section examines the effect of mapping the higher dimensional feature space to a lower dimensional representation using the methods described in section 2.2.2. The second section presents an investigation into the quality of the mapping.

3.1 Comparison of Real and Synthetic Datasets

For each of the experiments in this section the full set of 360 images of both left, right and medio-lateral oblique and craniocaudal views. As the full set of phantoms mammograms represents only a small number of cases (6 in total) if the full set was to be used each phantom would be over represented. To combat features were extracted for all phantoms and then a random phantom was selected from each case to be representative of the case as a whole.

3.1.1 Blob features

Figure 3.1.1 shows the result of applying the t-SNE algorithm to the feature matrix of extracted blob radii using the techniques outlined in the methodology to reduce the feature space to two dimensions. Real mammograms are represented by dots and phantom mammograms are represented by triangles. Each data point is coloured by either its BI-RADS risk class (in the case of the reals) or by the volumetric breast density (VBD) (in the case of the phantom mammograms). Figure 3.1.1 shows same features but mapped to 3 dimensions using t-SNE. For both the 2D and 3D cases parameters used for t-SNE were a learning rate of 300 and perplexity of 40.

For each of these plots it can be seen that there is some general clustering according to BI-RADS risk class, although separation of risk classes are far from optimum. The distribution of data points shown in figures 3.1.1 and 3.1.1 are heavily influenced by features relating to the relative proportion of blobs detected across the range of scales. Figure 3.1.1 shows the number of blobs detected for each scale for each risk class. While the distribution of blobs primarily look the same across all scales, the number of small blobs detected is markedly larger in low risk mammograms.

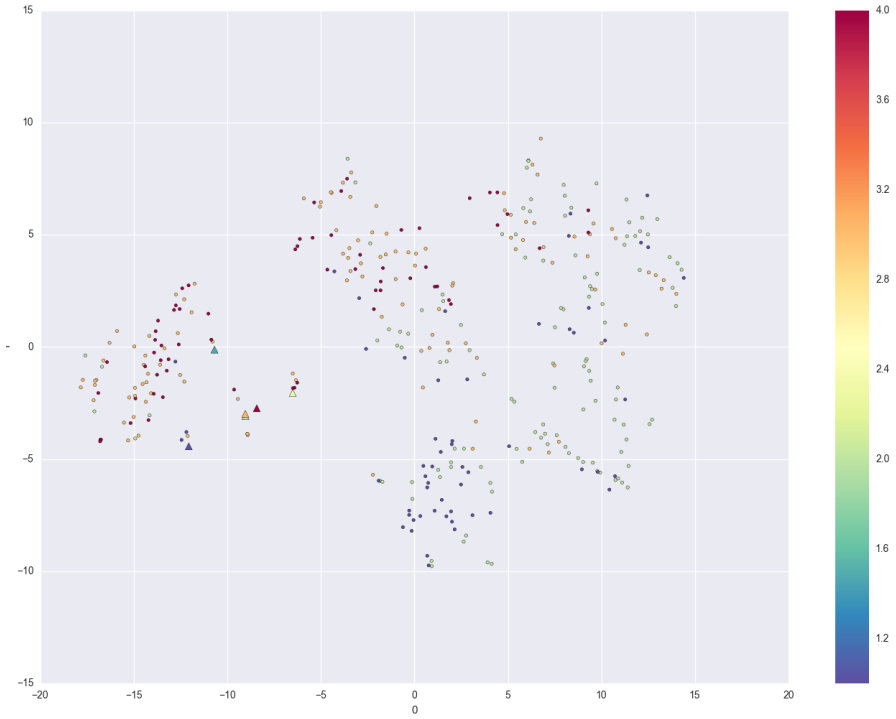

Figure 3.1: The 2D visualisation produced by t-SNE from the blob features extracted from real and phantom mammograms. real mammograms are dots and phantoms are triangles.

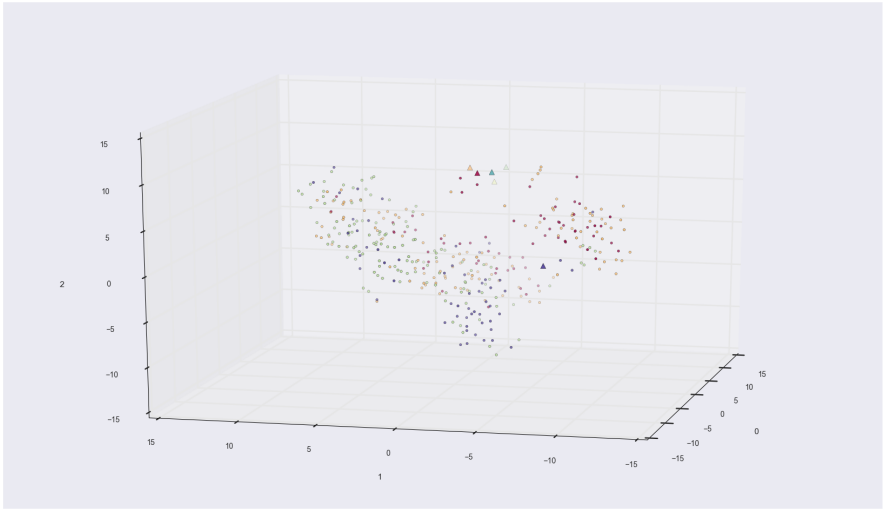

Figure 3.2: The 3D visualisation produced by t-SNE from the blob features extracted from real and phantom mammograms. real mammograms are dots and phantoms are triangles.

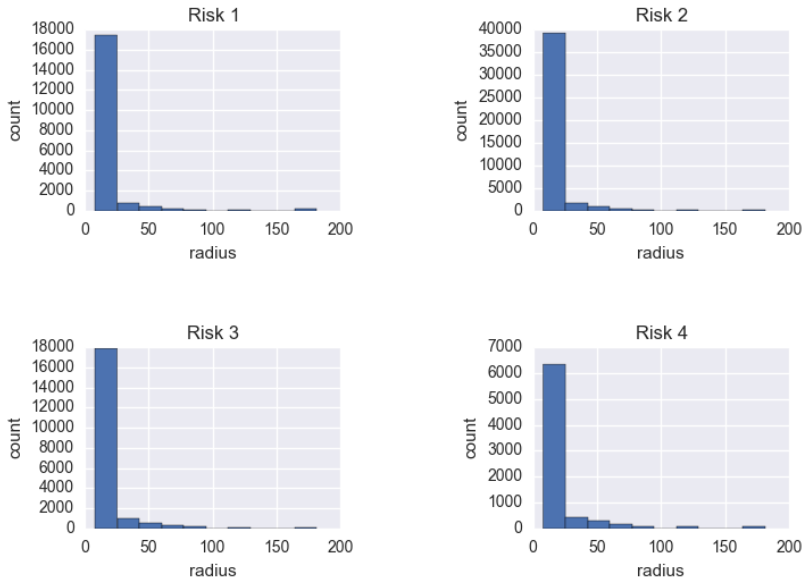

Figure 3.3: Histograms showing the number of blobs at each scale detected across all risk classes for real mammograms.

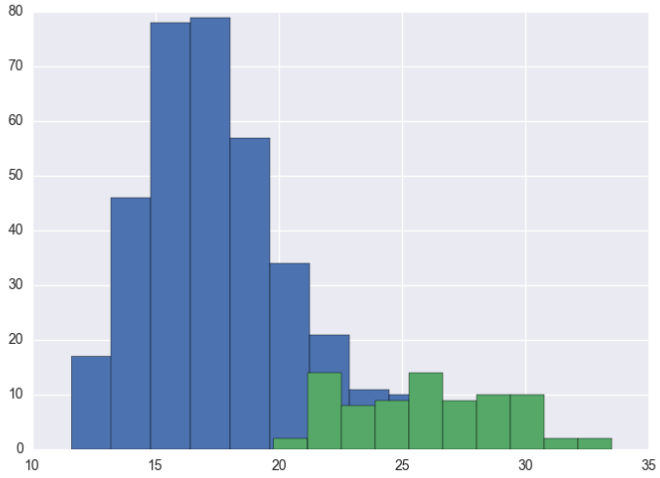

Figure 3.4: Histograms showing the average blob size for real mammograms (blue) and all phantom mammograms (green)

In general the breast phantoms, regardless of the their VBD, are clustered towards mammograms of a higher risk class. This suggests that there are a lower number of smaller blobs present across all VBDs for phantoms. As can be seen from figure 3.1.1 the average number of blobs is shifted compared to the real mammograms.

3.1.2 Line features

3.1.3 Intensity & Texture Features

Texture and intensity features proved proved the least successful for the comparison of the real and phantom datasets. The intensity distribution of the phantom mammograms are nothing like a real mammogram. Because of this difference the results show that the they are clearly in different spaces.

3.2 Quality Evaluation of Mapping

Chapter 4

Critical Evaluation

4.1 Evaluation of the Project

4.2 Future Work

Appendices

Appendix A

Third-Party Code and Libraries

Appendix B

Code samples

Annotated Bibliography

[1] P. R. Bakic, M. Albert, D. Brzakovic, and A. D. Maidment, "Mammogram synthesis using a 3d simulation. i. breast tissue model and image acquisition simulation," *Medical physics*, vol. 29, no. 9, pp. 2131–2139, 2002.

[2] ——, "Mammogram synthesis using a 3d simulation. ii. evaluation of synthetic mammogram texture," *Medical physics*, vol. 29, no. 9, pp. 2140–2151, 2002.

[3] ——, "Mammogram synthesis using a three-dimensional simulation. iii. modeling and evaluation of the breast ductal network," *Medical physics*, vol. 30, no. 7, pp. 1914–1925, 2003.

[4] C. Balleyguier, S. Ayadi, K. Van Nguyen, D. Vanel, C. Dromain, and R. Sigal, "Birads classification in mammography," *European journal of radiology*, vol. 61, no. 2, pp. 192–194, 2007.

[5] R. Bellman and R. E. Kalaba, *Dynamic programming and modern control theory*. Academic Press New York, 1965.

[6] N. Boyd, J. Byng, R. Jong, E. Fishell, L. Little, A. Miller, G. Lockwood, D. Tritchler, and M. J. Yaffe, "Quantitative classification of mammographic densities and breast cancer risk: results from the canadian national breast screening study," *Journal of the National Cancer Institute*, vol. 87, no. 9, pp. 670–675, 1995.

[7] I. Buciu and A. Gacsadi, "Gabor wavelet based features for medical image analysis and classification," in *Applied Sciences in Biomedical and Communication Technologies, 2009. ISABEL 2009. 2nd International Symposium on*. IEEE, 2009, pp. 1–4.

[8] Z. Chen, "Mammographic image analysis: Risk assessment and microcalcification classi cation aspects," 2013.

[9] Z. Chen, L. Wang, E. Denton, and R. Zwiggelaar, "A multiscale blob representation of mammographic parenchymal patterns and mammographic risk assessment," in *Computer Analysis of Images and Patterns*. Springer, 2013, pp. 346–353.

[10] H. Cheng, X. Shi, R. Min, L. Hu, X. Cai, and H. Du, "Approaches for automated detection and classification of masses in mammograms," *Pattern recognition*, vol. 39, no. 4, pp. 646–668, 2006.

[11] I. Christoyianni, E. Dermatas, and G. Kokkinakis, "Fast detection of masses in computer-aided mammography," *Signal Processing Magazine, IEEE*, vol. 17, no. 1, pp. 54–64, 2000.

[12] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[13] C. J. D'orsi, A. C. of Radiology, A. C. of Radiology, B.-R. Committee, *et al.*, *Illustrated Breast Imaging Reporting and Data System:(illustrated BI-RADS)*. American College of Radiology, 1998.

[14] N. El-Faramawy, R. Rangayyan, J. Desautels, and O. Alim, "Shape factors for analysis of breast tumors in mammograms," in *Electrical and Computer Engineering, 1996. Canadian Conference on*, vol. 1. IEEE, 1996, pp. 355–358.

[15] U. Fischer, F. Baum, and S. Luftner-Nagel, *Breast imaging*. Thieme, 2008.

[16] K. Ganesan, U. Acharya, C. K. Chua, L. C. Min, K. Abraham, and K. Ng, "Computer-aided breast cancer detection using mammograms: A review," *Biomedical Engineering, IEEE Reviews in*, vol. 6, pp. 77–98, 2013.

[17] I. T. Gram, E. Funkhouser, and L. Tabár, "The tabar classification of mammographic parenchymal patterns," *European journal of radiology*, vol. 24, no. 2, pp. 131–136, 1997.

[18] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973.

[19] A. Inselberg and B. Dimsdale, "Parallel coordinates," in *Human-Machine Interactive Systems*. Springer, 1991, pp. 199–233.

[20] S. Johansson and J. Johansson, "Interactive dimensionality reduction through user-defined combinations of quality metrics," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 993–1000, 2009.

[21] E. Jones, T. Oliphant, and P. Peterson, "{SciPy}: Open source scientific tools for {Python}," 2014.

[22] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," *Medical Imaging, IEEE Transactions on*, vol. 12, no. 4, pp. 664–669, 1993.

[23] D. B. Kopans and D. Kopans, *Breast imaging*. Lippincott-Raven Philadelphia, 1998.

[24] V. A. McCormack and I. dos Santos Silva, "Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis," *Cancer Epidemiology Biomarkers & Prevention*, vol. 15, no. 6, pp. 1159–1169, 2006.

[25] L. Novakova and O. Stepankova, "Radviz and identification of clusters in multidimensional data," in *Information Visualisation, 2009 13th International Conference*. IEEE, 2009, pp. 104–109.

[26] A. C. of Radiology. BI-RADS Committee and A. C. of Radiology, *Breast imaging reporting and data system*. American College of Radiology, 1998.

[27] I. U. P. on Breast Cancer Screening *et al.*, "The benefits and harms of breast cancer screening: an independent review," *The Lancet*, vol. 380, no. 9855, pp. 1778–1786, 2012.

[28] J. R. Parker, *Algorithms for image processing and computer vision*. John Wiley & Sons, 2010.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] N. Petrick, H.-P. Chan, B. Sahiner, and M. A. Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," *Medical Physics*, vol. 26, no. 8, pp. 1642–1654, 1999.

[31] N. Petrick, H.-P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification," *Medical Physics*, vol. 23, no. 10, pp. 1685–1696, 1996.

[32] A. Ronacher. (2015) Click library. [Online]. Available: http://click.pocoo.org/4/

[33] M. P. Sampat, M. K. Markey, A. C. Bovik, *et al.*, "Computer-aided detection and diagnosis in mammography," *Handbook of image and video processing*, vol. 2, no. 1, pp. 1195–1217, 2005.

[34] R. Siegel, J. Ma, Z. Zou, and A. Jemal, "Cancer statistics, 2014," *CA: a cancer journal for clinicians*, vol. 64, no. 1, pp. 9–29, 2014.

[35] R. A. Smith, D. Manassaram-Baptiste, D. Brooks, V. Cokkinides, M. Doroshenk, D. Saslow, R. C. Wender, and O. W. Brawley, "Cancer screening in the united states, 2014: a review of current american cancer society guidelines and current issues in cancer screening," *CA: a cancer journal for clinicians*, vol. 64, no. 1, pp. 30–51, 2014.

[36] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[37] H. Strange and R. Zwiggelaar, *Open Problems in Spectral Dimensionality Reduction*. Springer, 2014.

[38] L. Tabár, T. Tot, and P. B. Dean, *Breast cancer: the art and science of early detection with mammography: perception, interpretation, histopathologic correlation*. Thieme, 2005.

[39] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. Keim, "Combining automated analysis and visualization techniques for effective exploration of high-dimensional data," in *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*. IEEE, 2009, pp. 59–66.

[40] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[41] S. Van Der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "scikit-image: image processing in python," *PeerJ*, vol. 2, p. e453, 2014.

[42] J. S. Weszka, C. R. Dyer, and A. Rosenfeld, "A comparative study of texture measures for terrain classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 4, pp. 269–285, 1976.

[43] J. N. Wolfe, "Breast patterns as an index of risk for developing breast cancer," *American Journal of Roentgenology*, vol. 126, no. 6, pp. 1130–1137, 1976.

[44] Z. You and A. K. Jain, "Performance evaluation of shape matching via chord length distribution," *Computer vision, graphics, and image processing*, vol. 28, no. 2, pp. 185–198, 1984.

[45] R. Zwiggelaar, T. C. Parr, and C. J. Taylor, "Finding orientated line patterns in digital mammographic images." in *BMVC*, 1996, pp. 1–10.