Summer 8-18-2023

Characterization of Mammographic Dense Tissue Sub-Types Using Classification Algorithms On Metric Space Technique Output Functions

Madison Raza
University of Maine, razamadison@gmail.com

CHARACTERIZATION OF MAMMOGRAPHIC DENSE TISSUE SUB-TYPES USING CLASSIFICATION ALGORITHMS ON METRIC SPACE TECHNIQUE OUTPUT FUNCTIONS.

By

Madison Raza

B.S. in Biomedical Engineering, University of Maine 2022

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

(in Biomedical Engineering)

The Graduate School

The University of Maine

August 2023

Advisory Committee:

Andre Khalil, Professor of Biomedical Engineering, Advisor

Michael Mason, Professor of Biomedical Engineering

David Bradley, Professor of Mathematics and Statistics

CHARACTERIZATION OF MAMMOGRAPHIC DENSE TISSUE SUB-TYPES USING CLASSIFICATION ALGORITHMS ON METRIC SPACE TECHNIQUE OUTPUT FUNCTIONS.

By Madison Raza

Thesis Advisor: Prof. Andre Khalil

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Science
(in Biomedical Engineering)
August 2023

In recent years breast cancer has become the leading cause of global cancer incidence. One of the most common forms of screening is through the use of digital x-ray screening mammography. Risk assessment models which help predict a patient's risk of developing breast cancer rely mainly on patient history and qualitative breast density assessment from screening. The 2D wavelet transform maxima modulus (2D WTMM) method uses a sliding window approach to quantify the spatial organization of underlying mammographic tissue according to Hurst-exponent ranges (H) as fatty ($H \leq 0.45$), healthy dense ($H \geq 0.55$) and risky dense ($0.45 < H < 0.55$) resulting in grey-scale maps composed of H pixel values in the shape of mammograms.

The metric space technique (MST) is a method for quantifying 2D maps as a 1D output function where characteristics of an image are measured across threshold values and plotted. The MST was run on 89 tumorous patients (71 cancer, 18 benign) of the Perm data set which is comprised of H-value maps of mediolateral oblique (MLO) and craniocaudal (CC) views. Of thirty possible metrics, six are concluded to have statistically significant differences between cancer and benign categories. These six metrics were used to train univariate and multivariate general linear models (GLM) and k-nearest neighbor

(KNN) models. The univariate KNN models outperformed the univariate GLM models which resulted in acceptable discrimination and high specificity, but low sensitivity and balanced accuracy. The multivariate KNN achieved the highest area under the curve of the receiver operator curve (ROC AUC) of 0.71 indicating acceptable discriminatory capactity of the model. A modification to the MST is suggested which would address a dilution effect introduced through the examination of fatty, risky dense, and healthy dense AUC as discrete regions. Further work on feature selection and work with a larger, more balanced data set is necessary to validate these results.

DEDICATION

To my Mom and Michael.

ACKNOWLEDGEMENTS

I would like to thank my thesis committee, Prof. Andre Khalil, Prof. Michael Mason, and Prof. David Bradley for their time and attention to edits. I would like to acknowledge my principal advisor Prof. Andre Khalil for the opportunity to learn and grow in a space that would not have been possible without his support and mentorship. Finally, I would like to thank the Bioscience Association of Maine and the National Cancer Institute for funding this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

BACKGROUND

1.1 Biological Basis of Breast Cancer and Scope of Impact

In 2020 breast cancer became the leading cause of global cancer incidence at 11.7 percent for all cancer cases. Among women, for every four cancer cases one is breast cancer, and for every six cancer deaths one is attributed to breast cancer.[1] The visualization and screening of breast tissue and generation of a body of knowledge of risk factors has allowed for reduced breast cancer mortality. Multi-modal screening techniques could result in earlier detection of breast cancer and improved patient outcome.

Tissue visualization instrumentation has allowed for the screening, assessment, and diagnoses of human disease. X-rays are a common form of tissue visualization in health care and are one of the most common forms of breast cancer screening. Different tissue types have different radiological density based on their atomic number and material density, meaning they absorb different amounts of emitted energy. Mammography involves emitting x-rays through breast tissue, where a detector is placed on the other side of the tissue to collect energy not absorbed by the tissue.[2] Images are often taken in craniocaudal (CC) view and mediolateral oblique (MLO) view when performing x-ray film or digital mammography. Digital breast tomosynthesis (DBT) results in a 3D mammogram created from an array of x-rays at different angles that are stitched together digitally. Screening mammography is used to monitor changes in breast tissue that may indicate signs of cancer in patients who are otherwise not presenting with symptoms.[2,3] Screening of breast tissue composition is an important factor in identification and early diagnosis of breast cancers.

Breast tissue is composed primarily of fibroglandular or dense tissue, and fatty tissue. The components of fibroglandular tissue include fibrous connective tissue (stroma) and glandular tissue (glandular epithelium). Fibrous connective tissue is composed of the collagenous extracellular matrix (ECM) and fibroblasts and functions in supporting and

protecting glandular tissue within the breast.[2,4] The ECM not only provides supportive architecture, it is necessary for differentiation, proliferation, polarity, and maintenance of healthy cells.[5] The ECM is primarily composed of collagen and elastin. ECM proteins affect breast density and stromal alterations which may play parts in initial changes in tissue microenvironment and tumor development.[6,7] Generally, glandular tissue makes and releases substances into ducts and body fluids. Breast glandular tissues are the lobes (glands) and ducts. Lobes produce milk and are composed of clusters of smaller lobules. Ducts carry milk from the lobes to the nipple.[2,4] Breast tissue type correlates to pixel intensity, or pixel brightness when viewing mammography images. Fibroglandular tissue appears white on mammography images and fatty tissue appears dark on mammography images.[6,8] The visual distinction between tissue types allows for metrics to be assigned to mammography images.

Mammographic density (MD) is the appearance of fat versus epithelium and stroma on a mammogram.[7,8] Breast density has been described by the American College of Radiology through use of the Breast Imaging and Reporting System (BI-RADS) which describes categories of density and risk. Categories which describe density are coined category A through D. Category A describes almost entirely fatty breast composition , category B describes scattered areas of fibroglandular density, category C describes heterogeneous density which may obscure small masses, and category D describes extremely dense breast composition which may lower mammography sensitivity.[9] Approximately 43 percent of women aged 40 to 74 years of age have either heterogeneously or extremely dense breasts, where almost half (45 percent) of these women are under 50 years of age.[10] An estimated 1 in 8 breast cancers are not detected in screening mammography.[11] Breast density is associated with increased risk of false negative screening due to lack of contrast between cancer and fatty tissue causing difficulty in detection. Qualitative metrics relating to breast tissue allow for health care professionals to better understand breast composition and how that composition may be related to the overall health of the patient. Mammographic

percent density (MPD) quantifies the proportion of dense breast tissue in a mammogram through the use of computer-aided detection. Mammographic percent density is not yet widely adopted as part of the breast cancer screening process.[12,13]

Risk factors that increase a woman's risk for breast cancer also include hormonal factors such as age at menarche, menopause status, and parity status; demographic factors such as race and ethnicity; and factors of family history and genetics.[1,5] The recommended age to begin annual screening by the American College of Radiology and the American Cancer Society is 40 years. Screening mammography is proven to reduce the mortality from breast cancer in the US and abroad. Screening and adjuvant therapy resulted in a decreased mortality rate between 1975 and 2000 in the United States based on a model of estimated mortality trends from the University of Wisconsin-Madison. The model predicted an increase in mortality rate by 30 percent for scenarios with no screening and no adjuvant treatment, while screening alone reduced the rate of death by approximately 20 percent.[14,15] In a Dutch cohort study following women aged 49-63 years of age on January 1, 1995 who were invited for breast cancer screening the effects of one additional screening were shown to reduce cumulative risk of dying from breast cancer by approximately 10 percent over 17 years.[16]

1.2 Risk Assessment Models and Screening Tools

Current breast cancer risk assessment models most widely adopted in practice are the BRCAPRO,[17] Gail,[18] Tyrer-Cuzick,[19] Claus,[20] and BCSC models.[21] Except for the Tyrer-Cuzick and BCSC models, current risk assessment models rely on patient history.[22] Breast density as a semi-qualitative metric is reported in the Tyrer-Cuzick model and the BCSC model, where the Tyrer-Cuzick model only adopted in the breast density metric in 2019.[23,24] Risk assessment model parameters are summarized in Table 1.1[22] The BRCAPRO, Gail, Tyrer-Cuzick, Claus, and BCSC models have good calibration and modest discrimination, with increasing discriminatory capacity for models that include

Table 1.1 Risk assessment model parameters.

Model Parameters	BRCAPRO	Gail	Tyrer-Cuzick	BCSC	Claus
Age, y	Any	35-90	19-85	35-74	20-79
First menarche, y	0	1	1	0	0
First live birth, y	0	1	1	0	0
Menopause, y	0	0	1	0	0
Parity	0	0	1	0	0
HRT	0	0	1	0	0
Race/ethnicity	1	1	1	1	0
Ashkenazi Jewish ancestry	1	0	1	0	0
Prior biopsies	0	1	1	1	0
Atypical hyperplasia	0	1	1	1	0
LCIS	0	1	1	1	0
Relatives/degree	Any	First	First-Third	First	First/Second
Age at onset	1	0	1	0	1
Bilateral breast cancer	1	0	1	0	0
Ovarian cancer	1	0	1	0	1
Breast density	0	0	1	1	0

BCSC = Breast Cancer Surveillance Consortium. HRT = hormone replacement therapy. LCIS = lobular carcinoma in situ. 0 = not included in model. 1 = included in model.

available breast density information.[22] Combining current risk assessment models with quantitative tools and machine learning could increase the discriminatory capacity of risk assessment models.

Classification tools are being investigated to explore a multitude of avenues for detection of breast cancer with the goal of improving early diagnoses and accuracy of diagnoses. Features of cancer cells have been selected using Gray Wolf Optimization algorithms and modeled with support-vector machines resulting in high AUC criterion.[25] A hybrid approach was used with statistical preprocessing eliminating insignificant features before the use of discrete particle swarm optimization for classification of pathology slides from the Wisconsin breast cancer data set. The hybrid model achieved approximately 99 percent accuracy, 100 percent sensitivity, and 98 percent specificity.[26] Features including age, body-mass index (BMI), and metrics of the blood product have been modelled using a

Naive Bayesian classifier, decision tree, k-nearest neighbours (KNN), and artificial neural networks (ANN). The ANN had the highest classification accuracy between healthy and non-healthy samples at 80 percent, followed by KNN at 77 percent, Naive Bayesian at 73 percent, and decision tree model at 71 percent.[27] The performance of feature engineering and prediction models for discrimination of mammograms have been evaluated, where different types of models were used as base learners to be build a meta-learner based on the first level predictors which produce a final prediction. A hybrid multi-layer model achieved better accuracy than a single model on prediction of cancer mammograms.[28] Traditional patient history based risk assessment models could benefit from the use of classification tools in screening to improve accuracy and early diagnoses.

1.3 Spatial Organization and the 2D Wavelet Transform Modulus Maxima Method

Spatial organization broadly describes the way in which things in space are arranged. In a biological image analysis context, spatial organization describes the organization of tissue structures within an image. There are different ways of quantifying how tissue is spatially arranged such as through the use of co-occurrence features, the fractal dimension, or spectral features.[29] In the work presented screening digital mammograms were preprocessed using a spectral feature technique called the 2D wavelet transform modulus maxima method (2D WTMM). The mammographic breast tissue was processed with a sliding window approach, where a 360 x 360 pixel sub-region was incremented by 32 pixels across the image resulting in a grid of overlapping sub-regions. The 2D WTMM method is a multi-spectral approach which acts as a mathematical microscope, the result of which is the assignment of a Hurst-exponent value (H) to each sub-region of the mammographic breast tissue.[30–32]

The Hurst exponent value is used to describe the roughness of a surface. The Hurst range is generally between zero and one, and can be broken into three categories.

Anti-correlated behavior exists when H is less than 0.5 and correlated behavior exists when H is greater than 0.5. Anti-correlated Hurst-values indicate a surface which is organized in its roughness, whereas correlated Hurst-values indicate a surface which is organized in its smoothness. The point in the middle when H equals 0.5 indicates uncorrelated behavior, or unorganized and random surface roughness.[33] When Hurst-value ranges are applied as a preprocessing step to mammograms, the result is a uniform grid of Hurst-values in the shape of the mammogram which describe the surface roughness of the sub-regions of tissue.

CHAPTER 2

IMAGE ANALYSIS AND THE METRIC SPACE TECHNIQUE

2.1 Data Set Background

The Perm data set includes screening mammography of histologically verified images with benign or cancerous lesions and the associated image of the opposite breast. There is a total of 104 patients with evidence of tumor, 81 are cancerous, and 23 are benign. Cancer and benign categories can be further broken down into sub types. There are two cancer sub-types: invasive lobular cancer (ILC), and invasive ductal carcinoma (IDC). There are two benign sub types: fibroadenoma (Fib_a) and fibrocystic mastopathy (Fib_m). The number of records broken down by sub-type is reported in Table 2.1. Generally, each patient had images taken from 2 views, mediolateral oblique (MLO) and craniocaudal (CC), of the tumorous and opposite breast.

Table 2.1 Perm data set tumor populations

Malignant	Tumor Type	Patients
•	ILC	43
•	IDC	38
	Fib_a	12
	Fib_m	11

ILC, invasive lobular carcinoma; IDC, invasive ductal carcinoma; Fib_a, fibroadenoma; Fib_m, fibrocystic mastopathy; "•", indicates malignancy.

Tumorous images with a record of all four views were used in the presented study resulting in 89 total patients in the study population, 71 had cancerous tumors and 18 had benign tumors. Each mammogram was divided into overlapping sub-regions of size 360 x 360 pixels, where only the central 256 x 256 window was kept, and a sliding window incremented by 32 pixels was analyzed from the top left to the bottom right. The 2D WTMM method was used to compute the Hurst-exponent values for each sub-region. Each sub-region can be classified based on the Hurst-exponent as follows: where $H \leq 0.45$ is

associated with fatty breast tissue, $0.45 < H < 0.55$ is associated with disrupted tissue and tumor-associated loss of homeostasis, and $H \geq 0.55$ is associated with dense breast tissue.[30–32] The Hurst-exponent ranges are then color coordinated as seen in Figure 2.1, where blue corresponds to fatty tissue, red healthy dense, and yellow risky dense.

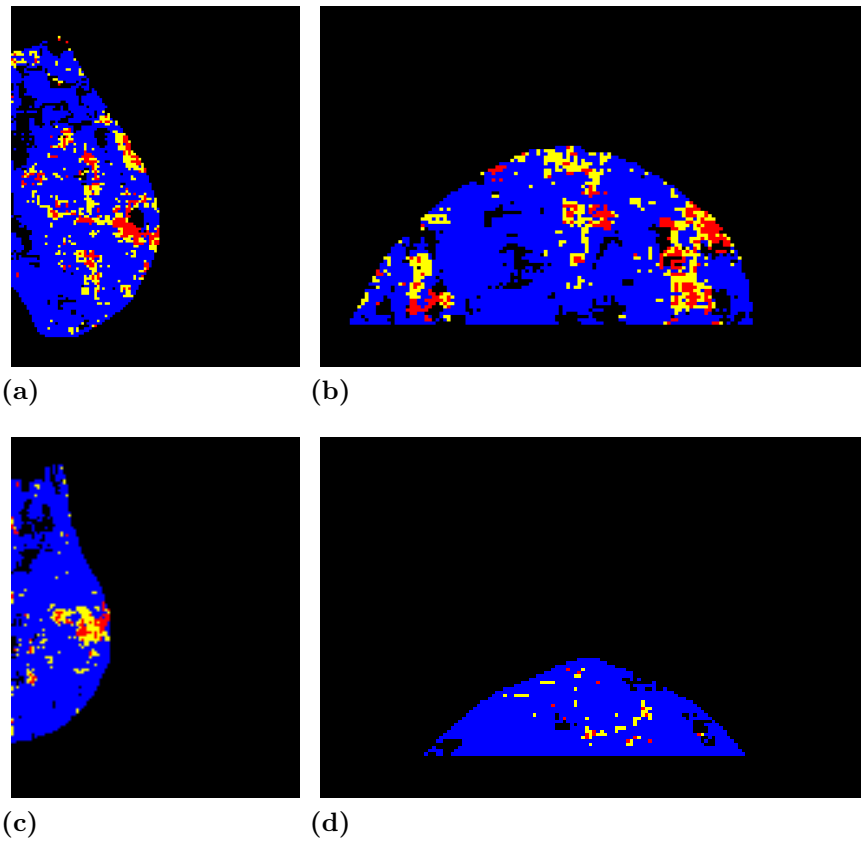


Figure 2.1 RGB Hurst-value heat maps in MLO Tumorous and CC tumorous view (a) Cancerous MLO view. (b) Cancerous CC view. (c) Benign MLO view. (d) Benign CC view.

The images in Figure 2.1 were selected as representative images as the visual difference of the area of Hurst-values in the dense tissue range between the cancerous and benign images are distinct. The MLO cancerous image and CC cancerous image, Figure 2.1 (a) and (b) respectively, appear to have a higher proportional area of Hurst-values which correspond to dense tissue when compared to the MLO benign image and CC benign image, Figure 2.1 (c) and (d) respectively.

Grey-scale images were constructed from the associated Hurst-exponents of screening digital mammograms. The grey-scale Hurst-value maps presented in this work represent -0.2 $< H \leq$ 1.0. Grey-scale representative images from the same patient views as Figure 2.1 can be seen in Figure 2.2.

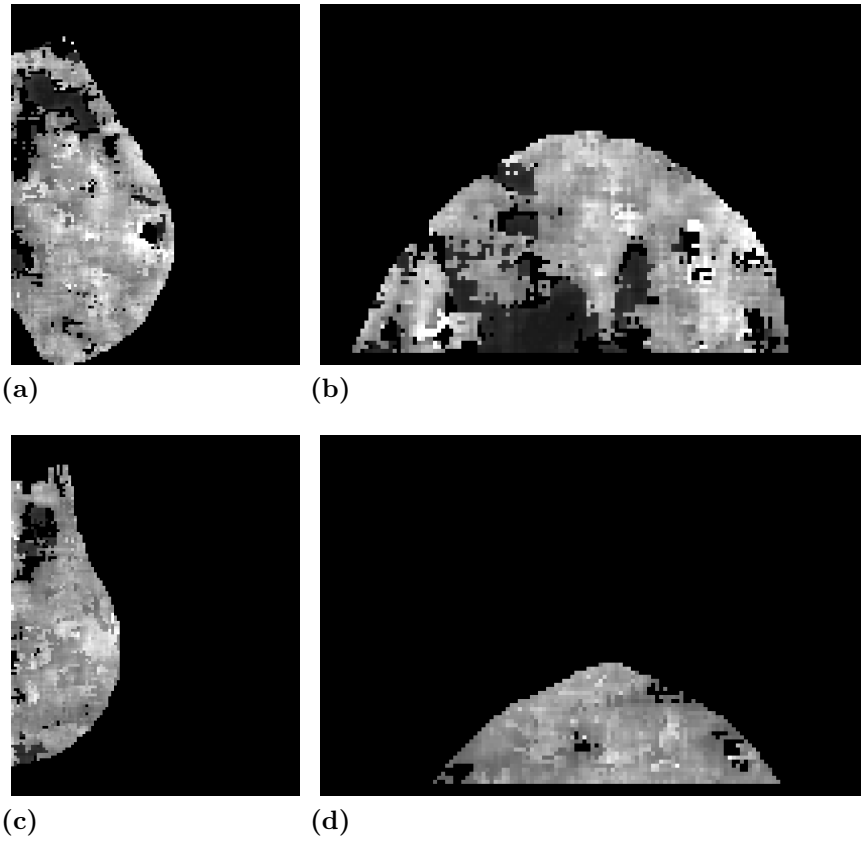


Figure 2.2 Grey-scale Hurst-value heat maps in MLO tumorous and CC tumorous view (a) Cancerous MLO view. (b) Cancerous CC view. (c) Benign MLO view. (d) Benign CC view.

2.2 The Metric Space Technique

The metric space technique (MST) was developed to quantitatively describe the difference between astrophysical maps.[34,35] The MST functions by quantifying 2D characteristics of a map as a 1D output function allowing for the difference between two signals to be investigated. The MST had been used in the study of astrophysics to explore the morphological features of hydrogen clouds,[36,37] and to explore the multi-scale

morphological features of galaxy point distributions.[38] The MST had also been investigated in biological contexts to quantify the Down syndrome cell adhesion molecule (DSCAM) as a generalized modulator of cell adhesion during neural development in mouse retina,[39] and to investigate the spatial relationship of collagen and nerves in adipose tissue.[40]

2.2.1 Mathematical Formalization

In this context, the MST is the tool used in a data mining exercise to classify Hurst-value maps of cancerous versus benign mammograms. Four output functions of the MST were explored in their original form, the distribution of area, the distribution of density, the distribution of components.[34,35] As a note, the use of the symbolic $\sum$ as both a variable and an operator has been continued in this work for precedence sake due to its use in previous works,[34,35,38] it can be assumed by context when $\sum$ is symbolic of a variable or operator.

The MST defines reduced maps σ_m and σ_v, respectively for distribution of density and distribution of area. The reduced intensity map is defined as the remaining pixel density of the map at a reference density (threshold value),

$$\sigma_m(x; \sum) = \sigma(x)\ \Theta[\sigma(x) - \sum] \tag{2.1}$$

where Θ is a step function and $\sum$ a threshold value. The fraction m of the reduced map

$$m(\sigma; \sum) = \frac{\int d^2x\, \sigma_m(x; \sum)}{\int d^2x\, \sigma_m(x)} \tag{2.2}$$

is the total pixel density of the map at a pixel density greater than the threshold over the pixel density of the whole intensity of the map integrated over the domain D.

As with the reduced intensity map, the reduced area map is defined where the remaining

$$\sigma_v(x; \sum) = \Theta[\sigma(x) - \sum] \tag{2.3}$$

area of the map at a threshold value. The fraction v of the reduced map is defined where

$$v(\sigma; \sum) = \frac{\int d^2x\, \sigma_v(x; \sum)}{\int d^2x} \tag{2.4}$$

the total area of the map at a pixel density greater than the threshold of the whole map integrated over the domain D. For use with mammography images that have been processed with the 2D WTMM method the area of the whole map corresponds to regions of breast tissue in Hurst-value maps, and does not include background or pectoral muscle.

The distribution of components is a count of components at each threshold, where a component meets a user-defined minimum size in any direction and had two step connectivity. Firstly, a reduced space must be defined:

$$X_{\sum}^{+} = \{x \in D \big| \sigma(x) > \sum\}. \tag{2.5}$$

The distribution of components, n, is then the number of components at each threshold,

$$n(\sigma; \sum) \tag{2.6}$$

where a component is an island of 2-step connected pixels which meet a minimum diameter. The distribution of components only explores the quantity of components at each threshold, but does not give information the shape of the components at each threshold.

The filament index describes a component's morphology, and like the distribution of components only describes components which meet a user-defined minimum size. The filament index is described as,

$$F_j = \frac{P_j D_j}{4A_j} \tag{2.7}$$

where P_j is the perimeter of the j th component and D_j the diameter of the j th component.[38] This form of the filament index had been investigated to better characterize non-convex objects and was in the work presented.[38] Diameter is defined as the maximum length between any two points of the component where,

$$D(A) = max\{|x - y| \;\big|\; x, y \in A\} \tag{2.8}$$

x and y represent points in a Cartesian space in the element A. As F_j, the filament index of the jth component, becomes greater than one the component morphology becomes more filamentous or thread-like.

The distribution of filament indices is quantified as the summation of filament indices of components by weighting values,

$$f(\sigma; \sum) = \frac{1}{n(\sigma; \sum)} \sum_j w_j F_j \tag{2.9}$$

w_j, over the number components as a function of threshold value. The unweighted version of the the filament index defines, $w_j = 1$, whereas the weighted version weights based on the area of the component, $w_j = A_j / < A >$, where A_j is the area of the jth component at the threshold and $< A >$ is the average area of components at the threshold. The distribution of filament indices allows for the degree to which components deviate from a circle to be compared between two maps.

2.2.2 The Metric Space Technique in Practice

A three by three representation of a Hurst-value map can bee seen in Figure 2.3, where the pixel values range between $0.0 \leq H \leq 1.0$ The image was thresholded between $0.0 \leq H \leq 1.0$ with a step value of $\Theta = 0.1$ resulting in 11 bins. The minimum size of a component was defined as a set of connected pixels greater than 2 pixels in any direction. Pixels were considered to be connected if they were two-step neighbors.

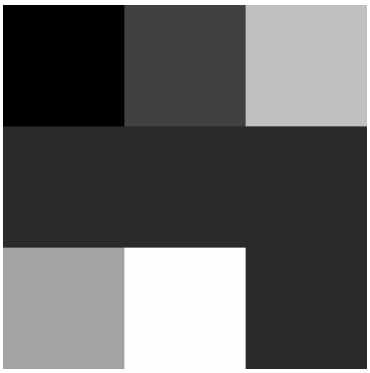


Figure 2.3 Three by three grey-scale image of a Hurst-value map

At the first threshold, where $\sum = 0.0$, the whole image is considered. At threshold, $\sum = 0.1$, any pixels greater than 0.1 are considered. The progression of the thresholded image can be seen in Table 2.2. It is notable that at threshold $\sum = 0.2$, the island of pixels in the

Table 2.2 A three by three grey-scale image of a Hurst-value map over the threshold range

$\sum$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1										
2										

1: Grey-scale images at each threshold, where black is masked regions. *2*: colored images highlighting components, where a component is a minimum of 2 pixels in size in any direction and blue is masked regions.

upper right corner is more than two-step neighbors away from the island of pixels in the bottom left corner resulting in 2 components at this threshold value. At $\sum = 0.2$ both islands of pixels meet the minimum size definition to be considered a component, whereas at $\sum > 0.7$ the remaining pixels in the image do not meet the size requirements to be defined as a component though, the pixels area and density are still considered with the distribution of area and distribution of density output functions. The reduced area and density maps, number of components, and unweighted summed filament index of Figure 2.3 over all threshold values are reported in Table 2.2. The reduced maps would then be

Table 2.3 Reduced maps, components, and unweighted $\sum_j F_j$ over thresholds of 3 by 3 example image

$\sum$	$\sigma_v(x;\sum)$	$\sigma_m(x;\sum)$	$n(\sigma;\sum)$	$\sum_j F_j$
0.0	9	3.6	1	1.000
0.1	9	3.6	1	1.125
0.2	4	2.8	2	3.000
0.3	3	2.5	1	1.500
0.4	3	2.5	1	1.500
0.5	3	2.5	1	1.500
0.6	3	2.5	1	1.500
0.7	2	1.8	0	0
0.8	1	1	0	0
0.9	1	1	0	0
1.0	0	0	0	0

normalized by the map at $\sum = 0$ and the summed unweighted filaments index would be averaged by the number of components at each threshold, $n(\sigma; \sum)$. Output function curves for distribution of area, distribution of density, distribution of components, and distribution of unweighted filaments are then plotted against the threshold range.

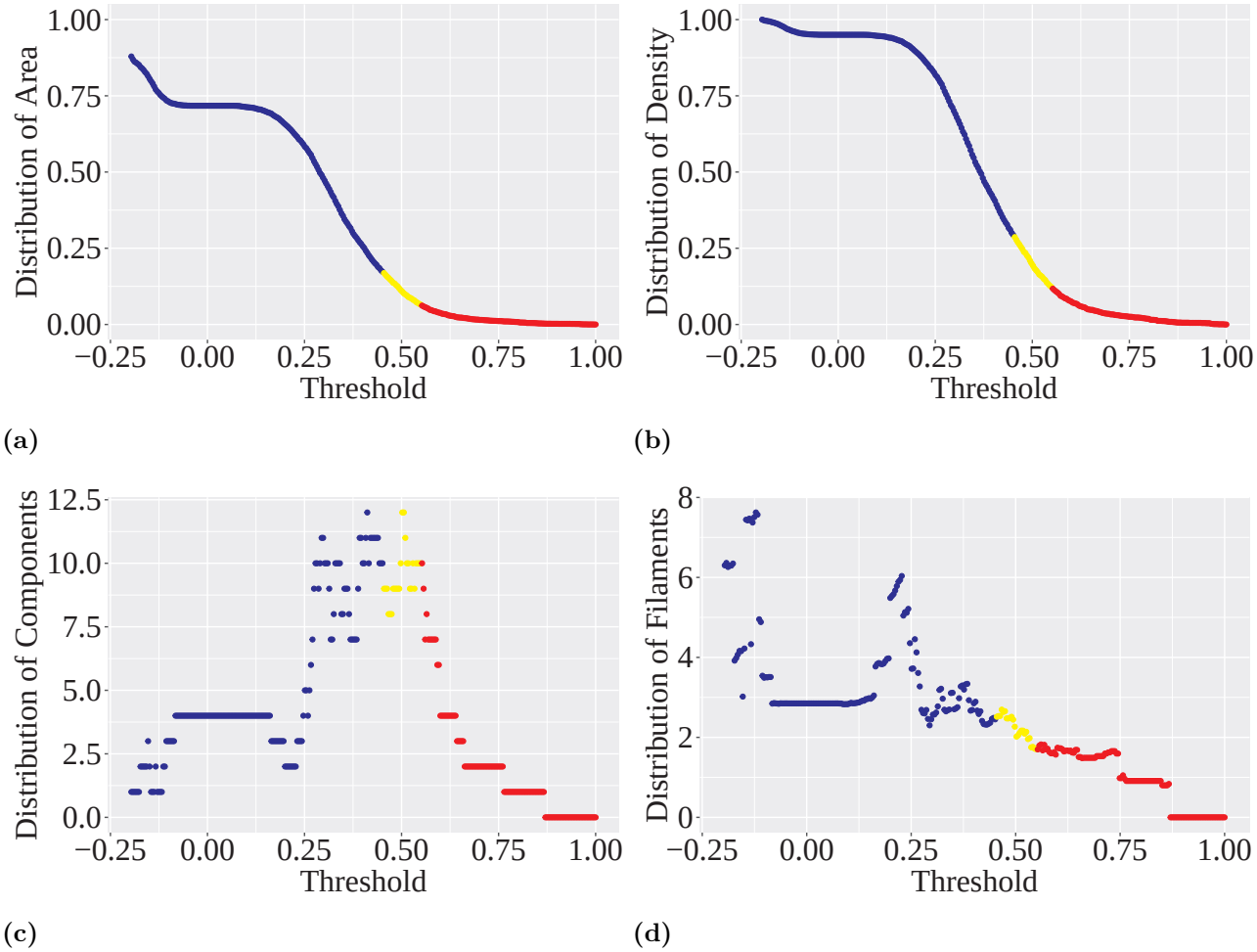


Figure 2.4 Output function curves of MLO cancerous patient
The x-axis is the threshold range and the y-axis is the MST output function. The threshold range is broken into regions correlating to Hurst tissue characterization. (a) Distribution of area. (b) Distribution of density. (c)-(d) Distribution of components and distribution of filaments, respectively, where minimum component size is 8 pixels.

The process outlined above results in five output function curves for each patient. The output function curves of Figure 2.2 (a) cancerous MLO view, are seen in Figure 2.4 where (a) is distribution of area, (b) distribution of density, (c) distribution of components, and

(d) unweighted distribution of filaments. The grey-scale Hurst-value maps were thresholded between H= 0.10392156862 and H= 1.3 with a step value of Θ = 0.00392156862 resulting in 306 bins, where pixel values within the map have been scaled by an addition of 0.3 to ensure continuous, decreasing output functions. A high degree of precision was necessary to limit improper rounding in the (Python) environment. The area under the curve (AUC) of discrete threshold ranges correlating to Hurst-exponent tissue characterization are taken resulting in thee metrics per output function curve. The three metrics per output function are: the fatty AUC region with thresholds of $H \leq 0.45$, the risky dense AUC region with thresholds of $0.45 < H < 0.55$, and the healthy dense AUC region with thresholds of $H \geq 0.55$. This results in 15 metrics per singular patient view, and 30 metrics per patient between tumorous views.

CHAPTER 3

METRIC SPACE OUTPUT FUNCTION ANALYSIS

3.1 Percent Yellow Squares

Percent yellow squares is a previously explored metric which counts the number or pixels within the risky dense region of the RGB Hurst-value heat map, Figure 2.1, as a percentage of the whole RGB heat map.[32] Percent yellow squares is an alternative method to the MST which quantifies the relationships between Hurst-value tissue categories and the statistical chance that an image with certain characteristics is cancerous. A percent yellow squares analysis was run on the Perm data set. Tumorous MLO and CC views were investigated for cancerous versus benign images and the statistical differences were quantified using the Wilcoxon rank-sum test (Wilcox test), where a p-value < 0.05 indicated statistically significance. The percent yellow squares method output is summarized in Figure 3.1 for cancer versus benign images, where (a) was MLO images and (b) was CC images.

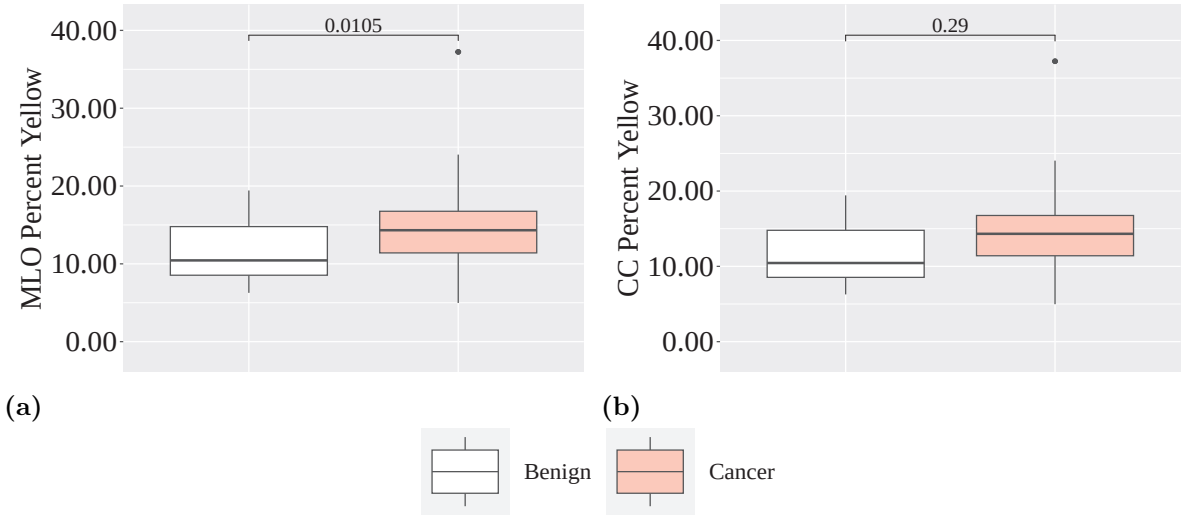


Figure 3.1 MLO and CC Percent Yellow Squares, Benign versus Cancer
The percentage of yellow squares in the RGB Hurst-value heat map was calculated. (a) MLO imaged, where $p \sim 0.01$. (b) CC imaged, no significance.

In both MLO and CC views the median percent yellow squares is higher for cancer than for benign, but only statistically significant differences exist in MLO with $p \sim 0.01$. Percent yellow squares statistics can be used as a quasi-benchmark when evaluating the MST as it is a previously tested method which had shown statistical significance in the MLO region when evaluating the Perm data set.[32] The general trend between these and previous works persist where the percent yellow squares had statistical significance in the MLO view between benign and cancerous images and there are no statistically significant differences in the CC view, the difference occurs in that the significance reported in this work for MLO percent yellow squares was higher than previously reported.[41] The percent yellow squares p-values reported here are different than in previous literature as the fitting methods within the 2D WTMM method have been altered resulting in different Hurst-exponent value outputs.

3.2 Distribution of Area and Distribution of Density

The results reported in this section encompass tumorous MLO and CC area AUC metrics for MST output functions which are not functions of component size. The distribution of area quantifies the reduced area of the breast region over total area of the breast region over the range of threshold values, where the total area of the breast region was the masked portion of the whole Hurst-value map which does not include background pixels. The distribution of density quantifies the reduced density of the breast region over the total density of the breast region, where reduced density was the summed pixels values of the Hurst-value map at thresholds and total density was the summed density of the masked portion of the whole Hurst-value map which does not include background pixels.

The patient curve, Figure 2.2, for each output function was broken up into three discrete regions corresponding to tissue characteristic of the Hurst-value range and the area under the curve of each region was taken. As each point on an output function curve represents the value of the output function at a threshold value, the points found in the

fatty region of the curve do not only reflect pixel values which correspond to fatty Hurst-values. The fatty region of the curve reflects the whole area and density of pixels which correspond to dense tissue over all threshold values while the area and density of pixels which correspond to fatty tissue degrades the threshold region. Points on the output function curve that lie in the risky dense region reflect the whole area and density of pixels which correspond with healthy dense tissue while pixels which correspond with risky dense tissue degrade over the threshold region. In the healthy dense region of the output function curve only pixels which are associated with healthy dense Hurst-values exist. The inclusion of pixels which are categorized as different tissues based on their Hurst-value within threshold regions for AUC metrics results in a dilution effect where the most amount of data lies at the beginning of the curve and the least at the ends. Thus, the AUC of the fatty region was orders of magnitude larger than the AUC of dense regions.

The dilution effect can be seen in Figure 3.2 which reports distribution of area for each tissue region and the Wilcox test of cancerous versus benign data, where (a) and (b) are

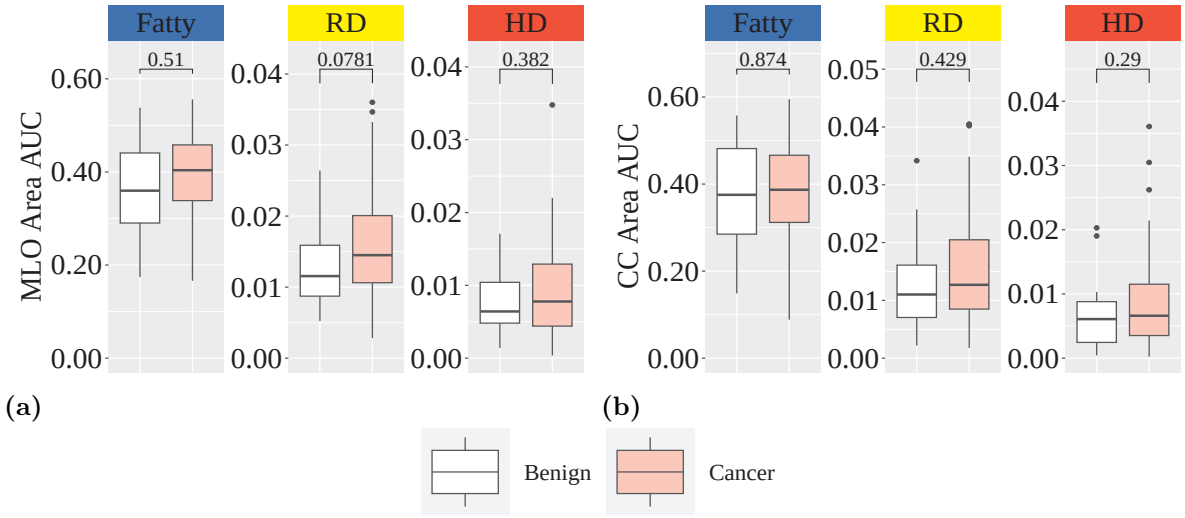


Figure 3.2 Tumorous MLO and CC distribution of area AUC in fatty and dense regions. The reduced area map over the total area map was calculated per threshold. The AUC of cancer and benign patient curves were taken in fatty, risky dense (RD), and healthy dense (HD) regions. (a) Tumorous MLO images, no regions of significance. (b) Tumorous CC images, no regions of significance.

MLO and CC, respectively. The AUC in both views trends downward across regions where the fatty region AUC across patients was an order of magnitude larger than the dense regions AUC across patients. This is due to breaking up the area under the curve into discrete regions where the most data lies in the fatty AUC region and the least in the dense AUC regions, so much so that the fatty AUC region is an order of magnitude larger. The cancerous median was marginally higher than the benign median in all regions and views, but this trend had inconclusive significance.

The distribution of density AUC statistics, Figure 3.3, follow the same trend as the distribution of area AUC in that the fatty region AUC was an order of magnitude larger than the dense regions and the cancerous median was higher than the benign median in all regions and views, with significance in the fatty MLO region where $p \sim 0.01$. The

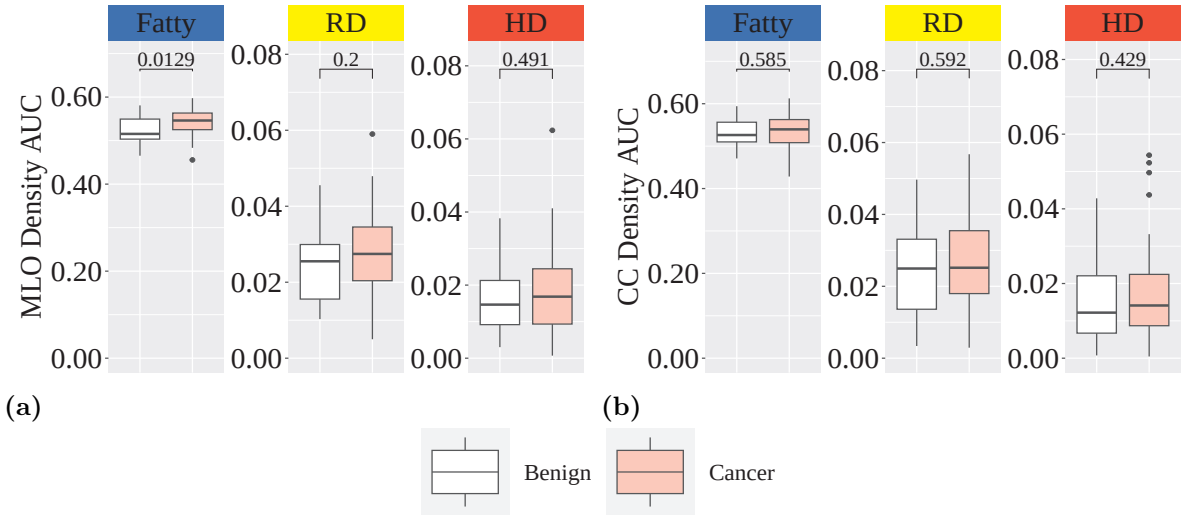


Figure 3.3 Tumorous MLO and CC distribution of density AUC in fatty and dense regions. The reduced density map over the total density map was calculated per threshold. The AUC of cancer and benign patient curves were taken in fatty, risky dense (RD), and healthy dense (HD) regions. (a) Tumorous MLO images, where $p \sim 0.01$ in fatty. (b) Tumorous CC Images, with no regions of significance.

distribution of density differs from the distribution of area as the distribution of density quantifies an extra dimension over the distribution of area, which is that the distribution of density accounts for how high the pixel were to be if it were a column and the distribution

of area counts the 2D surface. As such, the distribution of area and distribution of density are affected by dilution at different scales.

3.3 Distribution of Components

Distribution of components is a function of the user-defined minimum component size. A component is an island of connected pixels where neighboring pixels meet the definition of 2-step neighbors. Distribution of components is a count of the number of components within the breast region of a Hurst-value map over threshold values that meets the user-defined minimum component size. A data mining exercise was undertaken to determine which minimum component size achieved the best statistically significant results, cancerous versus benign AUC metrics in both views were assessed in fatty and dense regions for minimum component definitions between one and fifty pixels. A minimum component definition was considered best if the Wilcox p-value was less than 0.05 and was the minimum p-value of all component sizes in its respective tissue region. The primary y-axis reports AUC of the region, and the secondary y-axis reports the p-value associated with the Wilcox test of cancerous versus benign records where the horizontal black line corresponds to $p = 0.05$ on the secondary axis. The minimum component study for MLO distribution of components is reported in Figure 3.4, where (a)-(c) represent the fatty, risky, and healthy dense AUC regions, respectively.

The cancer median AUC tends to be higher than the benign median AUC for fatty and dense regions with a minimum component definition less than ten. The median AUC of cancer and benign appear to approach convergence as the minimum component definition becomes greater than ten with the median AUC trending towards zero. Neither the fatty or healthy dense region achieved significance at any of the 50 minimum component definitions. The risky dense region had a minimum significant $p \sim 0.02$, where the minimum component size was eight pixels. MLO distribution of components at a minimum component size of eight for fatty and dense regions, Figure 3.5, shows a granular view of

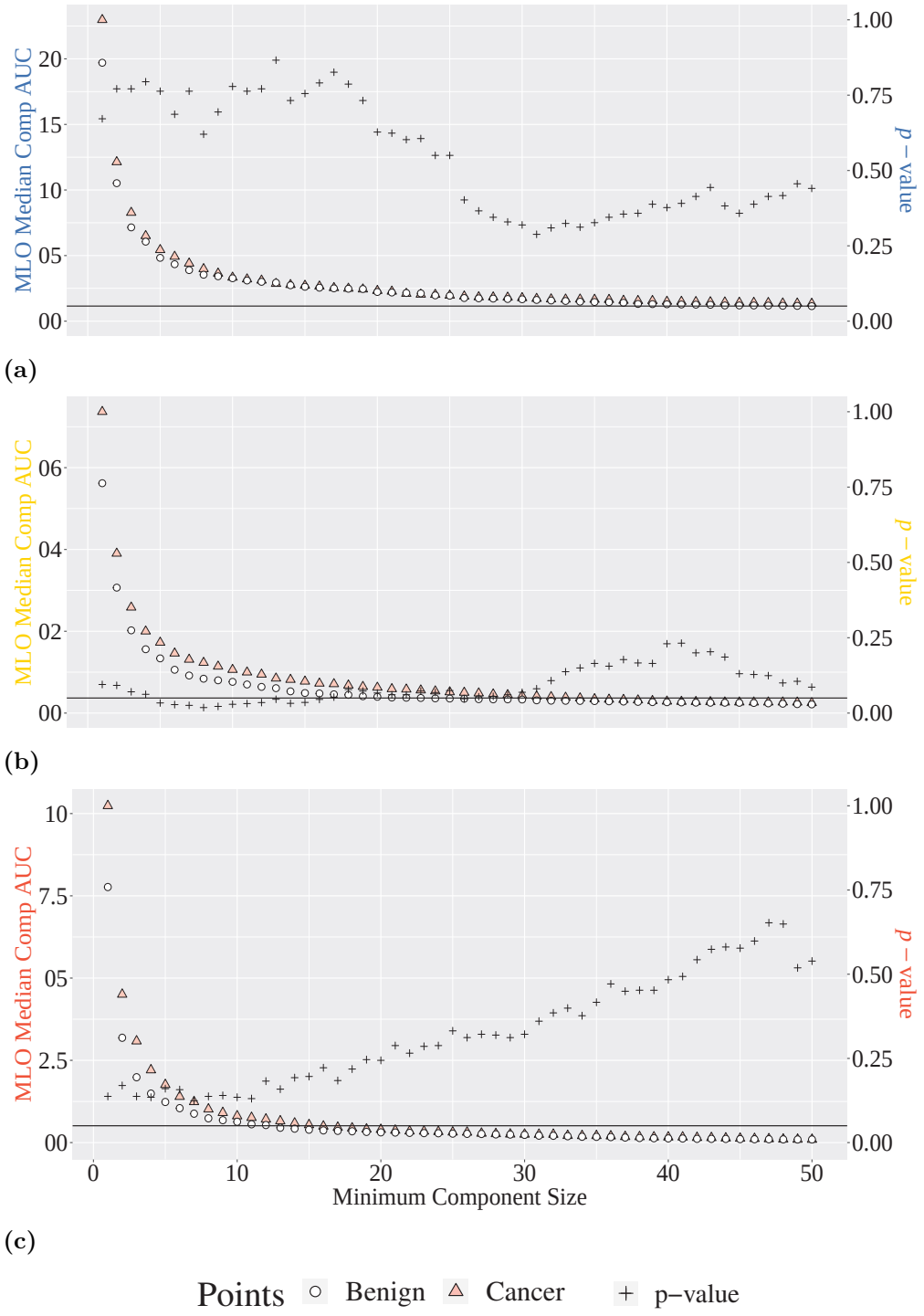


Figure 3.4 Minimum component study: MLO distribution of components. Two-step connected pixels meeting a minimum size definition were counted at each threshold. The median AUC of cancer and benign images in fatty and dense regions for minimum component size definitions [1,50] pixels and Wilcox p-values at each component size were calculated.The primary y-axis is the median AUC, while the secondary y-axis is the p-value scale with a horizontal line at $p = 0.05$. (a) Fatty. (b) Risky dense. (c) Healthy dense.

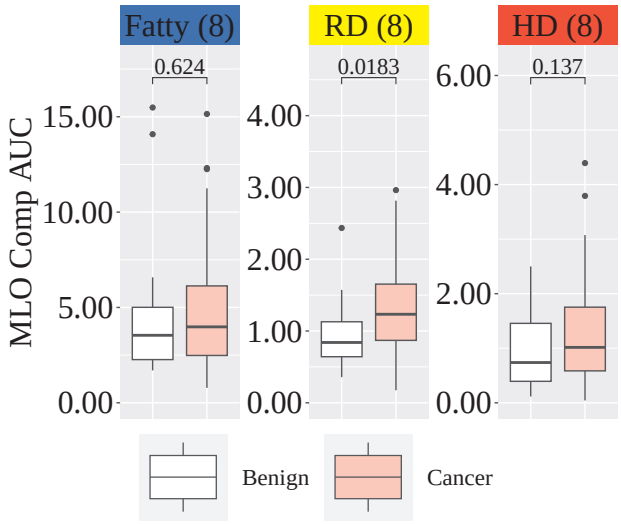


Figure 3.5 Tumorous MLO distribution of component AUC in fatty and dense regions, minimum component size of 8 pixels.
The minima p-value of the MLO distribution of components study was significant in the risky dense region at a minimum component definition of 8 pixels, where $p \sim 0.02$. Corresponding tissue regions were reported regardless of significance.

the same trends in Figure 3.4. The median AUC was marginally higher in fatty and healthy dense, but the difference was inconclusive. While the cancer median AUC was significantly higher than the benign median AUC in the risky dense region.

The minimum component study for CC distribution of components is reported in Figure 3.6, where (a)-(c) represent the fatty, risky dense, and healthy dense regions respectively. The primary y-axis reports AUC of the region, and the secondary y-axis reports the p-value associated with the Wilcox test of cancerous versus benign records where the horizontal black line corresponds to $p = 0.05$ on the secondary axis. While the cancerous median AUC tend to be higher than the benign median AUC where at higher minimum component definitions the median AUCs asymptotically approach zero as in the corresponding MLO study, there was no statistically significant differences in the CC distribution of components minimum component definition study.

As with distribution of area and distribution of density, the fatty region of threshold values includes pixels which are fatty and dense, while the risky dense threshold region includes pixels for both dense sub-types, and the healthy dense threshold region only had

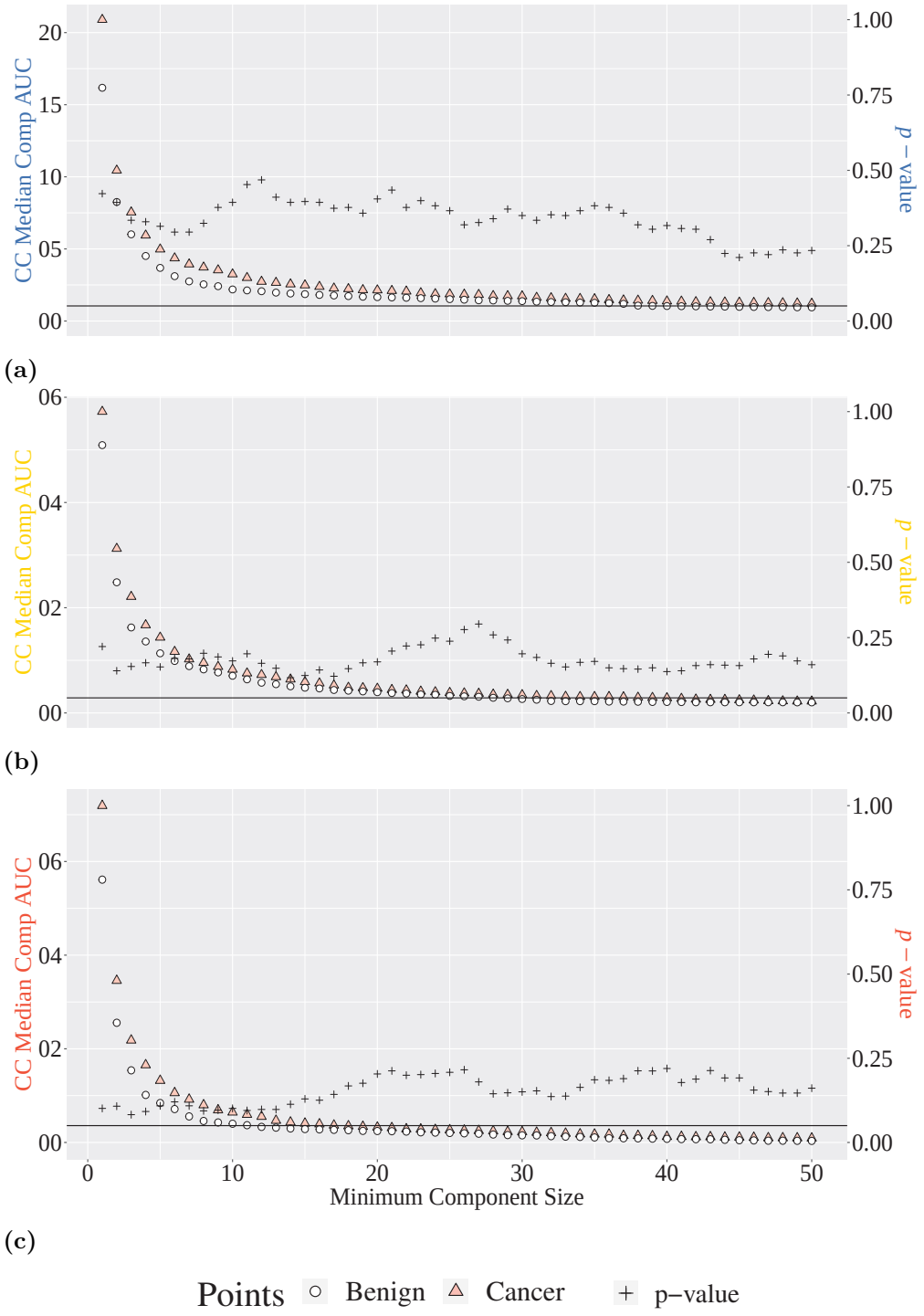


Figure 3.6 Minimum component study: CC distribution of components.
Two-step connected pixels meeting a minimum size definition were counted at each threshold. The median AUC of cancer and benign images in fatty and dense regions for minimum component size definitions [1,50] pixels and Wilcox p-values at each component size were calculated. The primary y-axis is the median AUC, while the secondary y-axis is the p-value scale with a horizontal line at $p = 0.05$. (a) Fatty. (B) Risky dense. (C) Healthy dense.

healthy dense pixels. This means a single component in the fatty threshold region could have pixels which fall into all three tissue categories, fatty, risky dense, and healthy dense, based on their assigned Hurst-values. A single component in the risky dense threshold region could have pixel values which fall into dense tissue categories. As these threshold regions are not exclusive to tissue categories, a dilution effect may occur. As the composition of the Hurst-value maps tends to be primarily fatty categorized pixels, more data was present within images in the fatty threshold region resulting in high AUC values in fatty regions and lower AUC values in dense regions. There was no conclusive trend relating fatty and dense tissue due to this dilution affect, but distribution of components does offer clues to trends regarding dense tissue. When setting a minimum component size of eight, cancerous images tend to have more components than benign images when analyzing threshold regions that only include pixels in the dense tissue categories healthy and risky dense, meaning there are generally more islands of connected dense regions in cancerous images.

3.4 Distribution of Filaments

The distribution of filaments can be calculated in unweighted and weighted form. The unweighted form averages the summed filament index at a threshold value by the number of components at that threshold value, while the weighted form introduces a weighting factor based on the area of the j th component at a threshold over the average area of components at that threshold. As the distribution of filaments quantifies average component morphology at a threshold, it is a function of the minimum defined component size. In the next two sub-sections the unweighted and weighted distribution of filaments fatty and dense AUC regions were investigated over minimum component definitions between one and fifty pixels for cancerous versus benign records in both views. As with the distribution of components component study, the the distribution of filaments component

study considered the minimum significant p-value to be the best minimum component size definition for that AUC region.

3.4.1 Unweighted Distribution of Filaments

The unweighted distribution of filaments had a weighting factor of $w_j = 1$, and quantifies the average component morphology at a threshold regardless of the area of components. As more pixels are present within the image at earlier threshold values it is expected that at earlier threshold values and components will be larger and fewer, at mid-threshold values there is expected to be a higher number of smaller components, and at later threshold value there is expected to be a small number of small components. The distribution of filament MLO component study is reported in Figure 3.7. There was no clear trend relating to median AUC of cancer and benign as the minimum component size definition was increased in fatty or dense AUC regions of the distribution of filament metric. The MLO distribution of filament component study yielded no statistically significant differences between cancer and benign records and inconclusive results.

Compared to the MLO, the CC distribution of filament component study had clear trend. Seen in Figure 3.8, as the median cancer AUC became greater than the median benign AUC the p-values decreased. This trend can be seen in Figure 3.8 (b) which represents the riksy dense AUC regions. At minimum component sizes of less than 25 pixels median benign AUC were equal to or greater than median cancer AUC and p-values were high. Starting at the minimum component size of 25 pixels there was a turning point, where median cancer AUC became greater than median benign AUC and the p-values trend downward and asymptotically reach $p \sim 0.25$ at a minimum component size of 35 pixels. While this trend does not exhibit statistically significant differences between benign and cancerous records, it does correspond with the trends in Figure 3.8 (a) and (c) where a median cancer AUC greater than a median benign AUC resulted in low and statistically significant p-values.

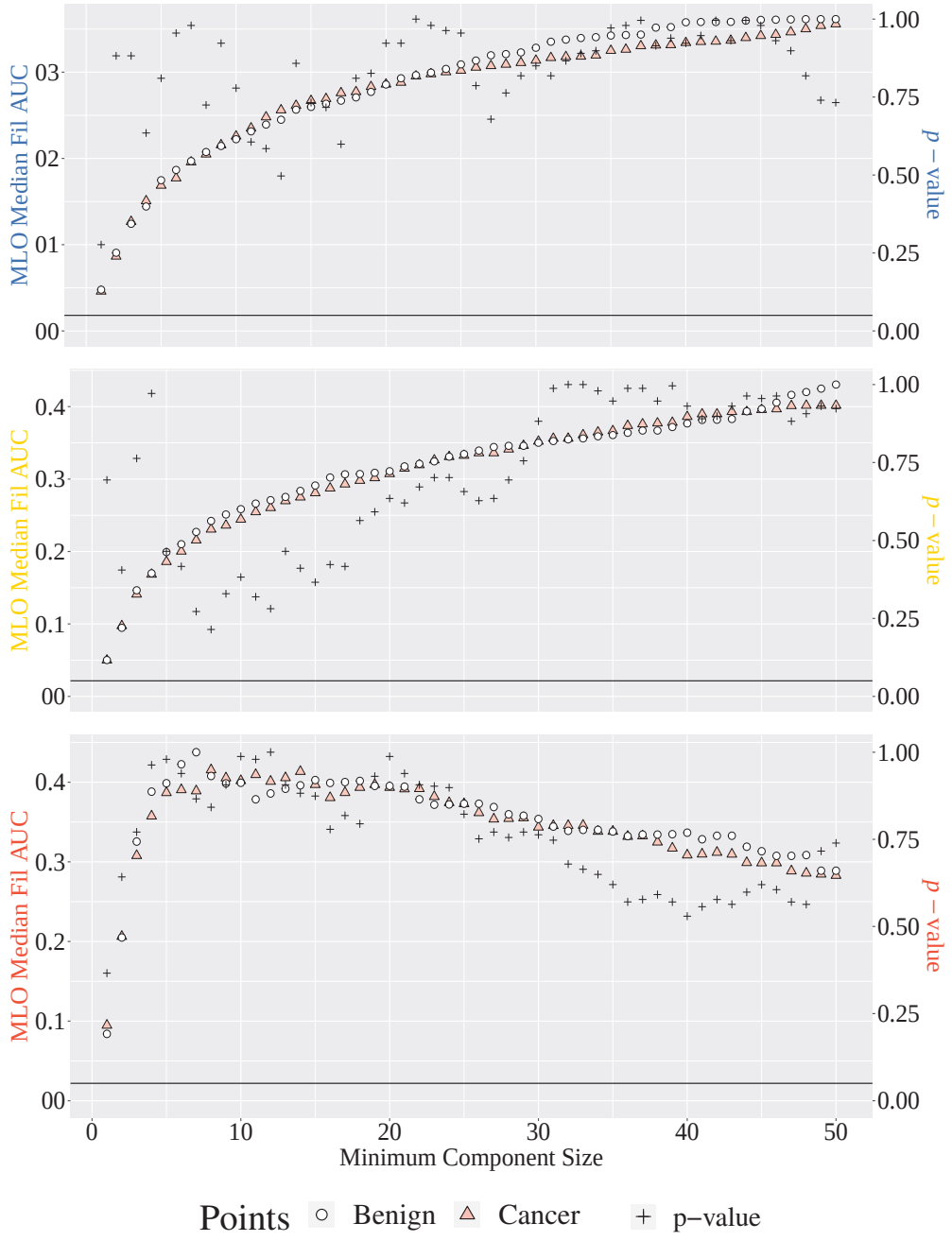


Figure 3.7 Minimum component study: MLO distribution of unweighted filaments
The quantitative morphology of each two-step connected component meeting a minimum size definition were summed and averaged by the number of components per threshold. The median AUC of cancer and benign images in fatty and dense regions for minimum component size definitions [1,50] pixels and Wilcox p-values at each component size were calculated. The primary y-axis is the median AUC, while the secondary y-axis is the p-values scale with a horizontal line at $p = 0.05$. (a) Fatty. (b) Risky dense. (c) Healthy dense.

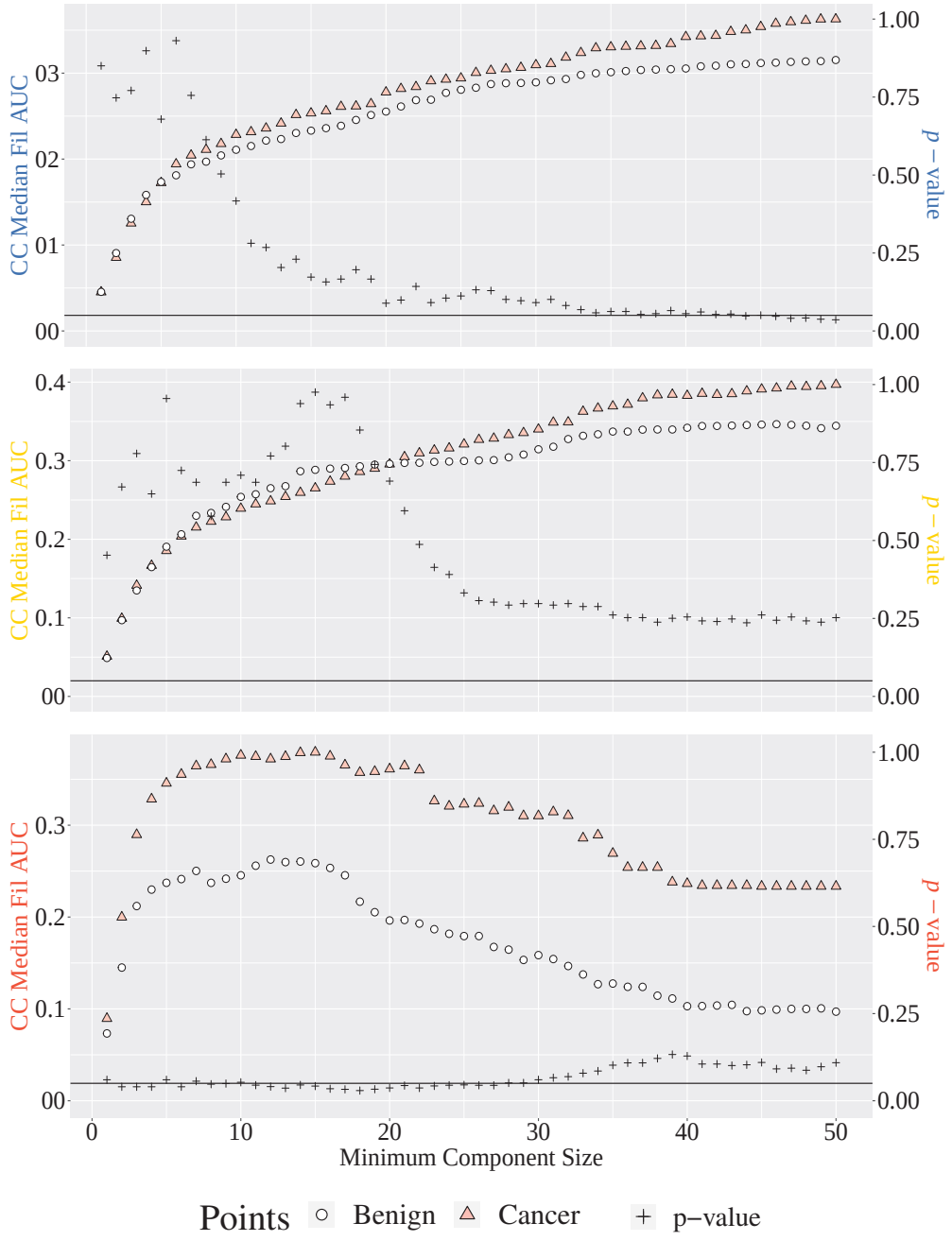


Figure 3.8 Minimum component study: CC distribution of unweighted filaments
The quantitative morphology of each two-step connected component meeting a minimum size definition were summed and averaged by the number of components per threshold. The median AUC of cancer and benign images in fatty and dense regions for minimum component size definitions [1,50] pixels and Wilcox p-values at each component size were calculated. The primary y-axis is the median AUC, while the secondary y-axis is the p-values scale with a horizontal line at $p = 0.05$. (a) Fatty. (b) Risky dense. (c) Healthy dense.

The fatty region and the healthy dense region, Figure 3.8 (a) and (b) respectively, reach statistical significance at different minimum component definitions. The fatty AUC region hovered around significance at a minimum component definition of greater than 35 pixels, whereas the healthy dense AUC region hovered around significance at a minimum component definition of less than 30 pixels. This trend could have a physical link, as the fatty AUC region corresponds to images on the early end of threshold range where more pixels and data are present within the image. Since there are more pixels at earlier threshold values the components tend to be larger. The healthy dense AUC region corresponds to images at the latter end of the threshold range where there are less pixels and components tend to be smaller. In the healthy dense AUC region there is information lost when setting the minimum component size greater than 30 pixels, as the amount of data within the image is reduced, components are less likely to meet the size requirement. The opposite effect may be true in the fatty AUC region, where the image is saturated with data and the minimum component size of greater than 35 acts as a filter.

The best minimum component definitions of the fatty and healthy dense AUC regions are represented in Figure 3.9 (a) and (b), respectively. The fatty AUC region of the CC distribution of filaments component study reached a minimum p-value at the minimum component size definition of 50 pixels, where $p \sim 0.04$. The healthy dense AUC region of the CC distribution of filament component study reached a minimum p-value at the minimum component size definition of 18 pixels, where $p \sim 0.03$. All AUC regions were reported per best minimum component definition regardless of significance. At a minimum component definition of 50 pixels in the fatty AUC region, median cancer AUC was higher than median benign AUC. The same can be said for the healthy dense AUC region at a minimum component definition of 18, the median cancer AUC was higher than the median benign AUC. Overall, the unweighted distribution of filaments component study yielded inconclusive results in the risky dense AUC region. Though the study did find that at high minimum component size definitions in the fatty AUC region and at low minimum

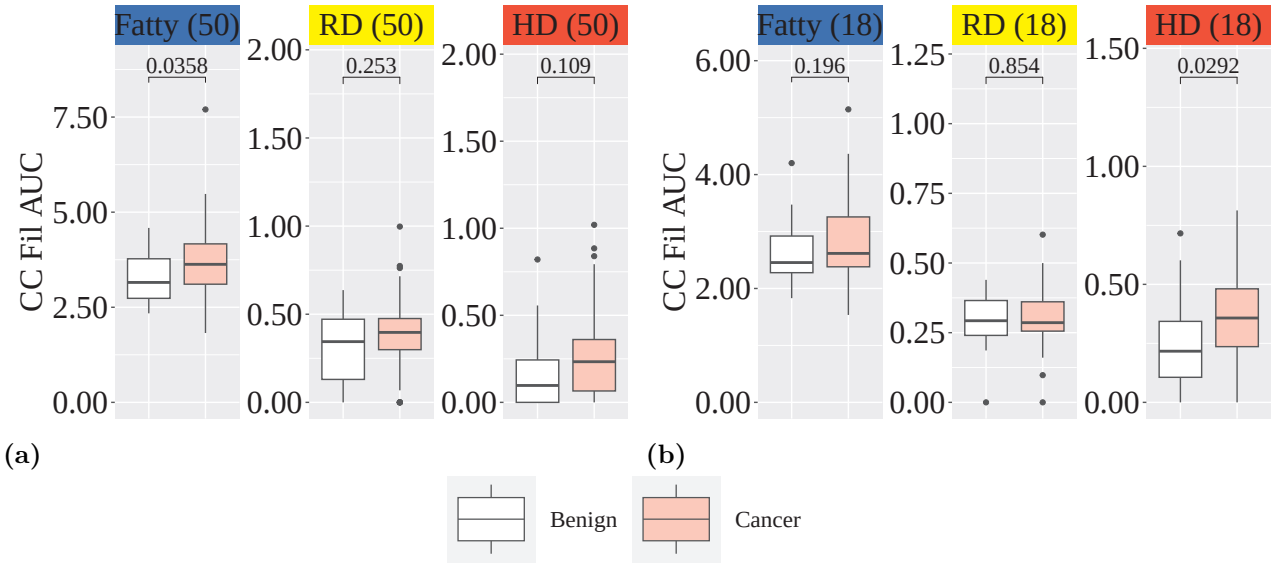


Figure 3.9 Cancer versus benign CC distribution of unweighted filaments AUC. Minimum p-values were significant in the fatty and healthy dense regions at different minimum component values of the CC distribution of filaments study. Corresponding tissue regions were reported regardless of significance for both minima. (a) Minimum component size is defined as greater than 50 pixels. Significance is seen in the fatty region, where $p \sim 0.04$. (b) Minimum component size is defined as greater than 18 pixels. Significance is seen in the healthy dense region, where $p \sim 0.03$.

component size definition in the healthy dense AUC cancer cases tend to have higher median AUC than benign cases, indicating that the distribution of filaments was higher across threshold regions and more likely to be filamentous than benign cases.

3.4.2 Weighted Distribution of Filaments

Weighted distribution of filaments is weighted by average area with a weighting factor of $w_j = A_j / < A >$. A minimum component size study was run for weighted distribution of filaments to identify the best minimum component. The MLO tumorous study can seen in Figure 3.10. No clear trend exists in the fatty dense AUC region with the majority of p-values at 0.75 or above. In the healthy dense AUC region median benign AUC was higher than median cancer AUC across all minimum component definitions. As the minimum component definition increases, the p-values trend downward and reach a minimum at a

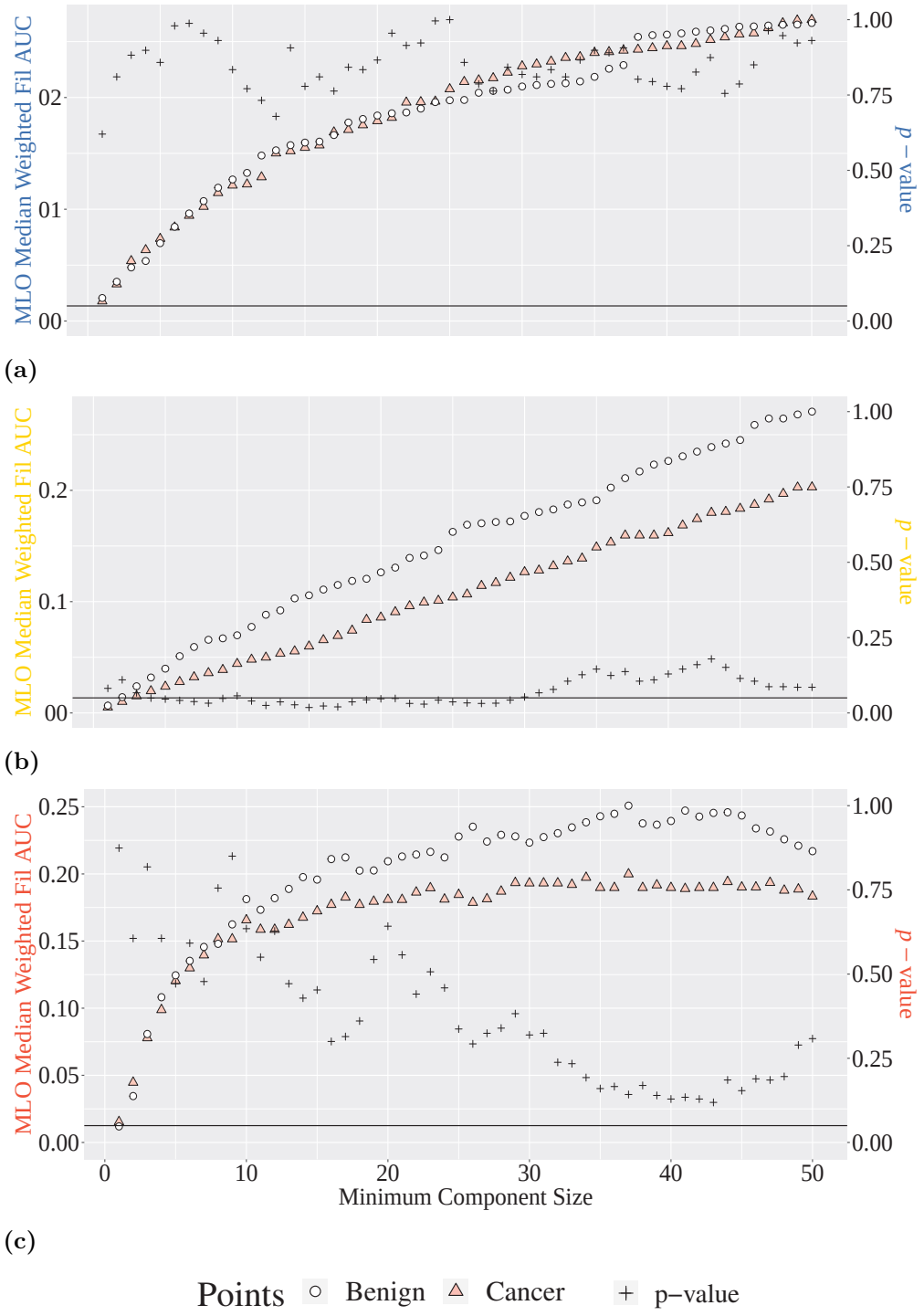


Figure 3.10 Minimum component study: MLO distribution of weighted filaments
The quantitative morphology of each two-step connected component meeting a minimum size definition were summed, weighted, and averaged by the number of components per threshold. The median AUC of cancer and benign images in fatty and dense regions for minimum component size definitions [1,50] pixels and Wilcox p-values at each component size were calculated. The primary y-axis is the median AUC, while the secondary y-axis is the p-values scale with a horizontal line at $p = 0.05$. (a) Fatty. (b) Risky dense. (c) Healthy dense.

minimum component definition of 44 pixels after which p-values trend upward. While there was a clear trend in the healthy dense region AUC, there was no statistical significance.

In the risky dense AUC region across all minimum component definitions the median benign AUC was higher than the median cancer AUC and the p-values are low. The p-values hover around 0.05 for a minimum component definitions less than 30 pixels, with a minimum $p \sim 0.02$ at a minimum component definition of 15 pixels. As the median benign AUC was higher than the median cancer AUC, the weighted average filament index across the risky dense region was generally higher in benign cases than cancer cases, indicating that benign cases are more filamentous across risky dense thresholds. The minimum component size definition with the most significant p-value for the tumorous MLO distribution of weighted filament risky dense AUC region is reported in Figure 3.11, where all AUC regions are reported regardless of significance. The median benign AUC was

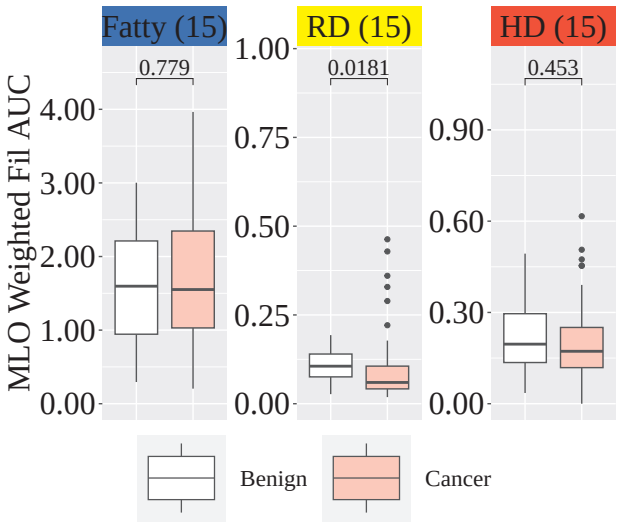


Figure 3.11 Cancer versus benign MLO weighted distribution of filaments AUC. The minima p-value of the MLO distribution of weighted filaments study was significant in the risky dense region at a minimum component definition of 15 pixels, where $p \sim 0.02$. Corresponding tissue regions were reported regardless of significance.

higher than the median cancer AUC in the dense AUC regions, though only the risky dense region showed statistically significant differences.

The CC tumorous weighted distribution of filaments study, Figure 3.12, had inconclusive trend and no significance in the fatty or risky dense AUC regions. In the

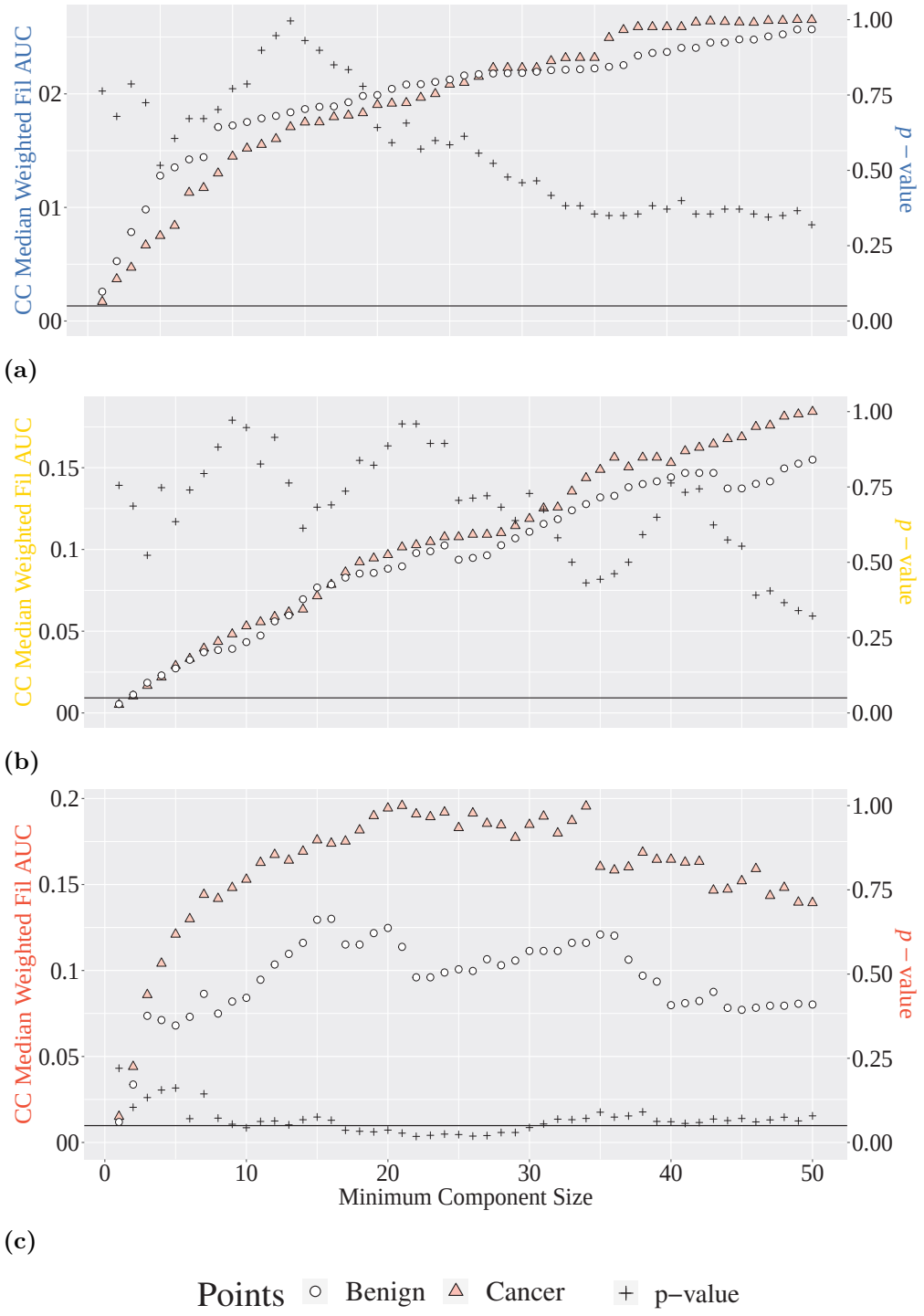


Figure 3.12 Minimum component study: CC weighted distribution of filaments
The quantitative morphology of each two-step connected component meeting a minimum size definition were summed, weighted, and averaged by the number of components per threshold. The median AUC of cancer and benign images in fatty and dense regions for minimum component size definitions [1,50] pixels and Wilcox p-values at each component size were calculated. The primary y-axis is the median AUC, while the secondary y-axis is the p-values scale with a horizontal line at $p = 0.05$. (a) Fatty. (b) Risky dense. (c) Healthy dense.

healthy dense AUC region median cancer AUC was higher than median benign AUC across all tested minimum component definitions. The minimum component definition range of 16 to 30 pixels reached significance of $p < 0.05$, with a minimum at a minimum component definition of 22 pixels where $p \sim 0.02$. The minimum significant component definition for the CC healthy dense AUC region is reported in Figure 3.13, where all tissue characteristic AUC regions are reported regardless of significance. The macroscopic trend across the minimum component study seen in Figure 3.12 can be seen in the healthy dense AUC region of Figure 3.13 where the median cancer AUC was higher than the median benign AUC with inconclusive trend in the fatty and risky dense AUC regions. The trend seen in the CC healthy dense AUC region was the same between weighted and unweighted filament indices, where median cancer AUC was higher than median benign AUC with minimum significant values close to the middle of the tested minimum component definition range. Unweighted and weighted CC healthy dense median AUC share a range of minimum component definition which are significant of 17 to 27 pixels. The reported median AUC of

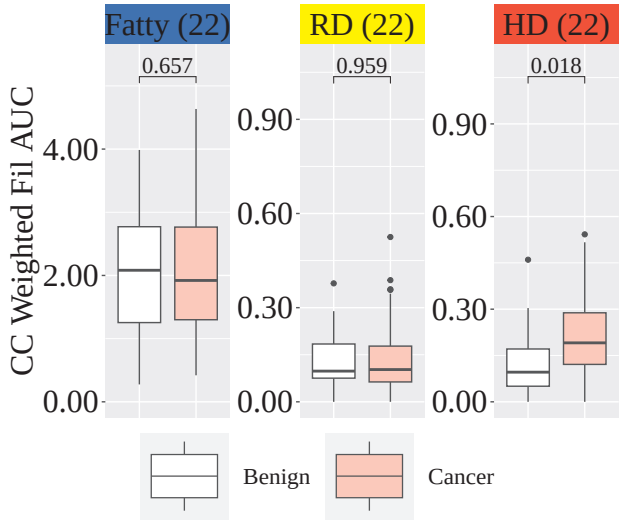


Figure 3.13 Cancer versus benign CC distribution of weighted filaments AUC. Minimum p-values were significant in the risky dense and healthy dense at different minimum component values of the MLO and CC distribution of weighted filaments study, respectively. Corresponding tissue regions were reported regardless of significance for both minima. (a) Minimum component size is defined as greater than 15 pixels. Significance is seen in the risky dense region, where $p \sim 0.02$. (b) Minimum component size is defined as greater than 22 pixels. Significance is seen in the healthy dense region, where $p \sim 0.02$.

the CC healthy dense region qualitatively appears to be less than the reported median AUC of the MLO healthy dense region, indicating that across healthy dense thresholds the average filament indices was generally less filamentous when weighting for area. In the future, investigating the differences between views quantitatively could provide further insights.

CHAPTER 4
STATISTICAL AND MACHINE LEARNING APPROACHES TO CLASSIFICATION

4.1 Univariate Models

Statistical and machine learning models have been widely used in biomedical applications for the detection of breast cancer. Features of cancer cells in pathology slides are being used to train classification models with high AUC criterion, accuracy, sensitivity, and specificity.[25,26] Traditional risk assessment approaches such as age and body-mass index (BMI) are being combined with metrics from the blood product of patients to train classification models with a high of 80 percent accuracy.[27] Hybrid multi-layer models which combine a multitude of machine learning have approaches improved accuracy of detection of cancerous mammograms.[28]

General linear models (GLM) and k-nearest neighbor (KNN) models were explored for classification of tumorous patients. It is important to note that GLMs are a classification algorithm which relies on statistical predictions whereas KNNs are machine learning algorithms. Predictors were chosen based on the MST output functions AUC regions which based on the Wilcox test had the minimum p-value. The *R* library *Caret* was used for model configuration, training, and testing.[42,43] Models were trained to maximize ROC AUC statistics.

Two methods of k-fold cross validation were used in training, leave-one-out cross validation (LOOCV) and 10 fold cross-validation with 3 repeats (repeated CV). The training and testing split for LOOCV was 88 training cases to 1 testing case, where the split for repeated CV was 63 training to 26 testing representing a 70/30 split. The first approach taken with model training was with GLMs and repeated CV. This model configuration yielded indiscriminate results causing a shift in method. Leave-one-out cross validation was chosen because it is a good candidate to be used with imbalanced data,[44]

and KNNs were chosen because they are a common form of supervised model which is well suited for few input variables.[45]

Several statistical methods were used in the assessment of classification models. The receiver operator curve (ROC) was used to assess the classification performance at different classification thresholds, where sensitivity was on the y-axis and specificity on the x-axis. In this case, the classification threshold indicates where to split the response of the model into a binary classifier of cancerous or benign records. The ROC AUC measured the ability of a model to discriminate between binary classifiers.[46] Sensitivity, specificity, and balanced accuracy were also assessed for each model configuration. Sensitivity is the measure of the classification of cases which were cancer as cancer,

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4.1}$$

where TP are true positive classifications and FN are false negative classifications.[47,48] Specificity is the measure of the classification of cases which were benign as benign,

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.2}$$

where TN are true negative classifications and FP are false positive classifications.[47,48] Balanced accuracy is the mean of sensitivity and specificity and was assessed instead of accuracy as the data set is imbalanced.[48]

In Figure 4.1, the receiver operator curve of GLMs of a single predictor are estimated. Each subplot contains two curves, one which used LOOCV methods and one which used repeated CV methods. The predictors used for each model correspond to the alphabetically categorized subplots: (a) corresponds to MLO fatty distribution of density AUC metrics in Figure 3.2, (b) to MLO risky distribution of component AUC metrics in Figure 3.5 with a minimum component size definition of 8 pixels, (c) to MLO risky weighted distribution of filament AUC metrics in Figure **??** with a minimum component size definition of 15 pixels, (d) to CC dense distribution of filament AUC metrics in Figure **??** with a minimum component size definition of 22 pixels, (e) to CC fatty distribution of

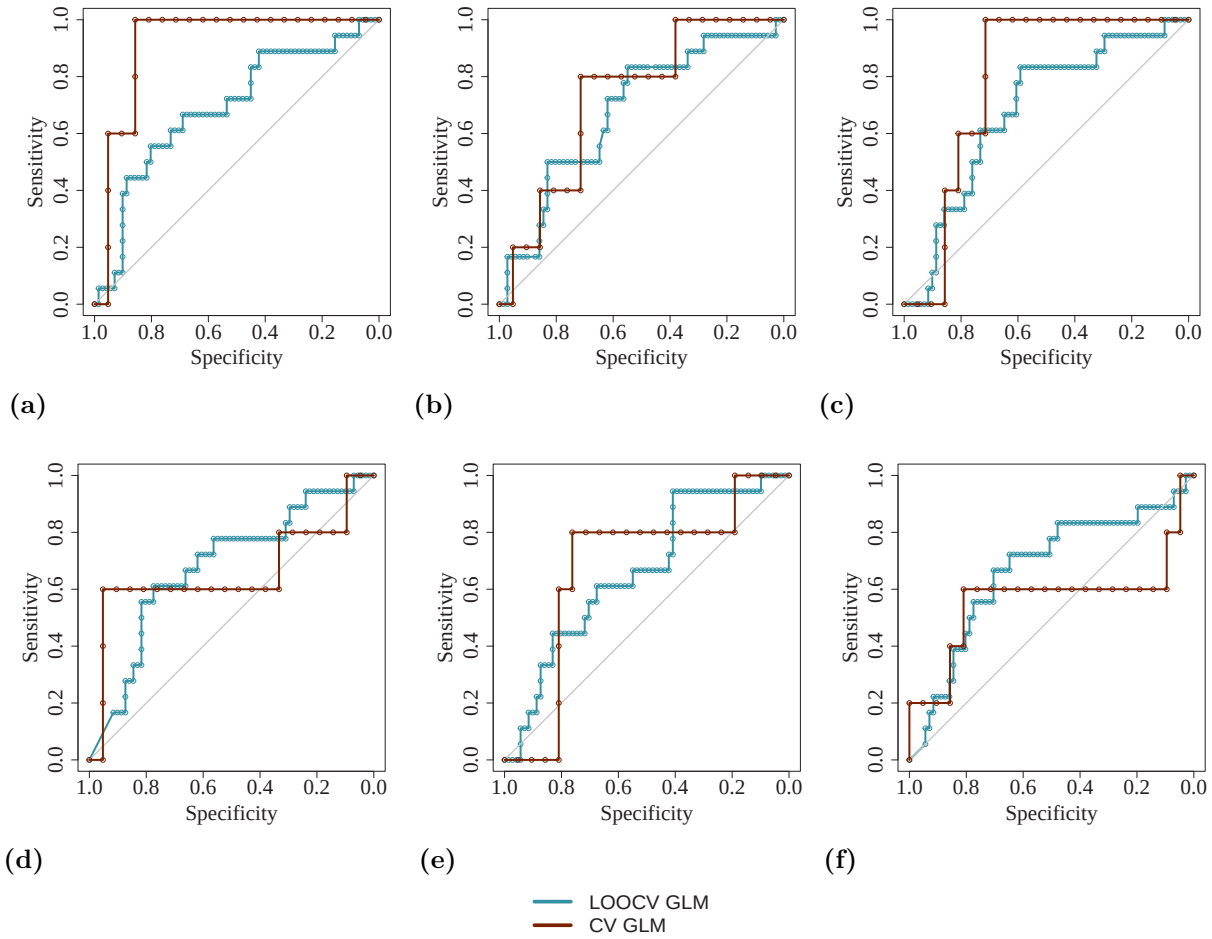


Figure 4.1 GLM ROC of Single Predictors
Output function AUC were predictors, functions of component size were defined by best minimum component size. CV methods: LOOCV, and 10 fold CV with 3 repeats (70/30). (a) MLO fatty distribution of density. (b) MLO risky distribution of component, 8. (c) MLO risky weighted distribution of filament, 15. (d) CC dense weighted distribution of filament, 22. (e) CC fatty distribution of filament, 50. (f) CC dense distribution of filament, 18.

filament AUC metrics in Figure 3.9 with a minimum component size definition of 50 pixels, and (f) to CC dense distribution of filament AUC metrics in Figure 3.9 with a minimum component size definition of 18 pixels.

Quantitatively, a well performing model will have sensitivity and specificity close to one and an ROC AUC close to one. Qualitatively this looks like the ROC curve lying to the left of the identity line, to which the GLM ROC generally conforms. Guidelines for ROC AUC defines ROC AUC = 0.5 as a model with no discriminatory capacity, between 0.5 <

ROC AUC < 0.7 as considered poor discrimination, between $0.7 \leq$ ROC < 0.8 as acceptable discrimination, between $0.8 \leq$ ROC < 0.9 was excellent discrimination, and ROC AUC ≥ 0.9 as outstanding discrimination. If a model ROC AUC is less than 0.5 it performs worse than a random classifier.[46,49]

Quantitative statistics for all tested single predictor GLM are summarized in Table 4.1. All models scored ROC AUC between 0.6 and 0.8 with poor or acceptable discriminatory

Table 4.1 ROC AUC, sensitivity, specificity, and balanced accuracy of single predictor general linear models.

Predictor	ROC AUC	Sens	Spec	Balanced Acc
(a) LOOCV	0.664	0	0.986	0.493
(a) Repeated CV	0.603	0	0.987	0.5
(b) LOOCV	0.643	0	1	0.5
(b) Repeated CV	0.657	0	0.993	0.5
(c) LOOCV	0.450	0	1	0.5
(c) Repeated CV	0.630	0	0.993	0.5
(d) LOOCV	0.636	0	1	0.5
(d) Repeated CV	0.712	0	1	0.5
(e) LOOCV	0.612	0	1	0.5
(e) Repeated CV	0.713	0	0.987	0.5
(f) LOOCV	0.606	0	1	0.5
(f) Repeated CV	0.753	0	0.940	0.5

Predictor associated with Figure 4.1 (a)-(f).

capacity, except for MLO risky weighted distribution of filaments at a minimum component size of 15 which scored ROC AUC equal to 0.45. While the ROC of single predictor GLM indicate potentially promising results, final model testing resulted in zero sensitivity and balanced accuracy of approximately 0.5 in all model configurations. Balanced accuracy was chosen particularly for these tests because in the GLMs standard accuracy was approximately 80 percent. While this number was initially promising, it became clear when viewing the confusion matrices that the models would classify all cases as cancer resulting in misclassification of all benign cases. Due to the imbalance in the data, standard accuracy was not a metric which provided value to the evaluation of these models. Poor

testing statistics between all GLM configurations necessitated shift in methods. All single predictor GLM configurations were duplicated and run with a k-nearest neighbor model.

In Figure 4.2, the receiver operator curve of KNN models of a single predictor are reported. The ROC of KNN models which used LOOCV appear smoother than the ROC of GLM models which used LOOCV. The KNN LOOCV also had a higher area under than the repeated CV method. The statistics pertaining to each final KNN model are reported in Table 4.2. All KNN models had ROC AUC greater than 0.5 meaning all meet some level

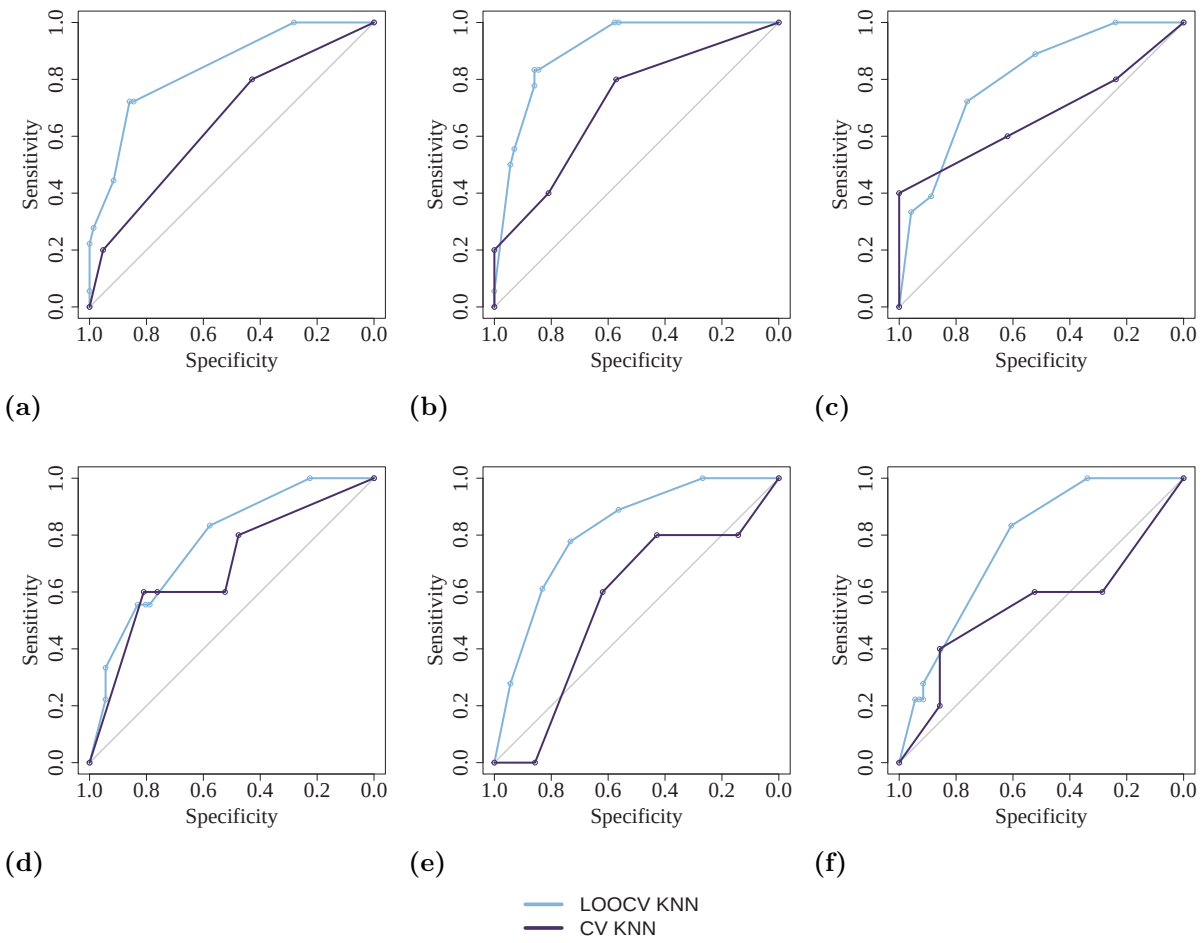


Figure 4.2 KNN ROC of Single Predictors
Output function AUC were predictors, functions of component were size defined by best minimum component size. CV methods: LOOCV, and 10 fold CV with 3 repeats (70/30). (a) MLO fatty distribution of density. (b) MLO risky distribution of component, 8. (c) MLO risky weighted distribution of filament, 15. (d) CC dense weighted distribution of filament, 22. (e) CC fatty distribution of filament, 50. (f) CC dense distribution of filament, 18.

Table 4.2 ROC AUC, sensitivity, specificity, and balanced accuracy of single predictor k-nearest neighbor models.

Predictor	ROC AUC	k	Sens	Spec	Balanced Acc
(a) LOOCV	0.605	7	0.278	0.986	0.632
(a) Repeated CV	0.513	5	0	0.973	0.576
(b) LOOCV	0.705	5	0.278	0.901	0.59
(b) Repeated CV	0.675	9	0.283	0.960	0.5
(c) LOOCV	0.640	9	0.056	0.915	0.486
(c) Repeated CV	0.627	5	0.050	0.940	0.7
(d) LOOCV	0.571	9	0.056	0.944	0.5
(d) Repeated CV	0.620	9	0	1	0.5
(e) LOOCV	0.527	7	0	0.972	0.486
(e) Repeated CV	0.583	7	0	0.953	0.5
(f) LOOCV	0.655	9	0.056	0.901	0.465
(f) Repeated CV	0.690	9	0.017	0.947	0.5

Predictor associated with Figure 4.2 (a)-(f).

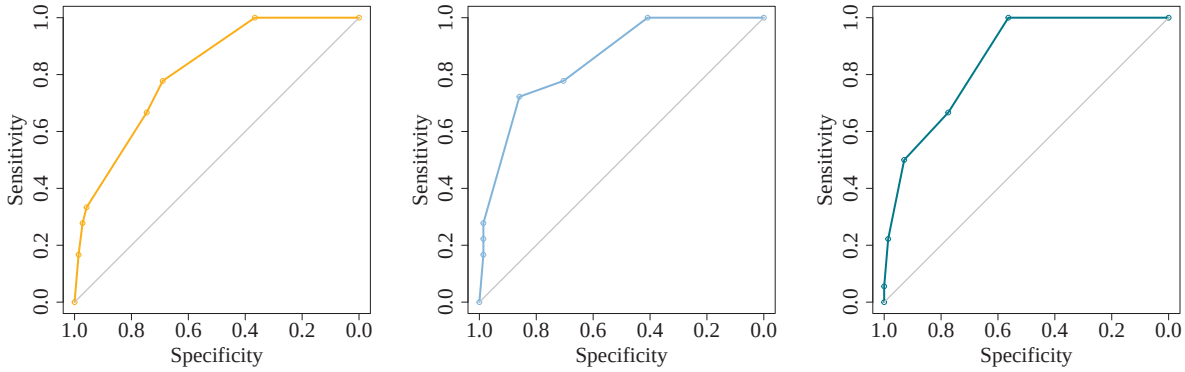
of discrimination. The maximum ROC AUC was achieved by LOOCV with the predictor of MLO risky dense distribution of component AUC at a minimum component size of 8 pixels, where ROC AUC = 0.705 and $k = 5$. The sensitivity of this model was 0.278, the specificity was 0.901, and the balanced accuracy was 0.59. These results are a marked improvement from the analogous GLM statistics. While specificity went down by approximately 10 percent, ROC AUC, sensitivity and balance accuracy improved.

The model with the highest sensitivity overall uses repeated CV and MLO risky dense distribution of component AUC at a minimum component size of 8 pixels as a predictor, but had a 0.5 balanced accuracy. The model with the highest specificity using LOOCV was tied by two models, the previously discussed MLO risky dense distribution of component AUC at a minimum component size of 8 pixels, and MLO fatty distribution of density AUC. The KNN statistics for MLO fatty distribution of density AUC as a predictor using LOOCV improve over the GLM statistics. While ROC AUC goes down from 0.664 to 0.604, sensitivity goes up from 0 to 0.278, balanced accuracy goes up from 0.493 to 0.632, and specificity remains the same as the highest of the KNN models.

The single predictor KNN models mark an improvement over the single predictor GLMs. ROC AUC between the algorithms are comparable indicating a comparable discriminatory capacity. Where the KNN models improve over the GLM models are in their increased sensitivity and balanced accuracy. KNN models using LOOCV were favored for multivariate testing.

4.2 Multivariate KNN Models

Multivariate KNN models using LOOCV were testing with two to three predictors. Three configurations of predictors were chosen, Figure 4.3, (a) was MLO risky dense



(a) **(b)** **(c)**

Figure 4.3 Multivariate KNN ROC
Output function AUC were predictors, functions of component were size defined by best minimum component size. LOOCV was the only cross validation method investigated. (a) MLO Risky Dense Distribution of Components AUC (8) + MLO Fatty Distribution of Density AUC. (b) MLO Risky Dense Distribution of Components AUC (8) + MLO Risky Dense Weighted Distribution of Filaments (15). (c) MLO Risky Dense Distribution of Components AUC (8) + CC Healthy Dense Weighted Distribution of Filaments AUC (22) + MLO Fatty Distribution of Density AUC.

distribution of components AUC at a minimum component size definition of 8 pixels and MLO Fatty Distribution of Density AUC, (b) was MLO risky dense distribution of components AUC at a minimum component size definition of 8 pixel and MLO risky dense weighted distribution of filaments AUC at a minimum component size definition of 15 pixel, and (c) was MLO risky dense distribution of components AUC at a minimum component

Table 4.3 ROC AUC, sensitivity, specificity, and balanced accuracy of multivariate k-nearest neighbor models.

Predictor	ROC AUC	k	Sens	Spec	Balanced Acc
(a) LOOCV	0.695	9	0.167	0.972	0.569
(b) LOOCV	0.702	5	0.278	0.901	0.59
(c) LOOCV	0.712	7	0.278	0.986	0.623

Predictor associated with Figure 4.3 (a)-(c).

size definition of 8 pixel, CC healthy dense weighted distribution of filaments AUC at a minimum component size definition of 22 pixels, and MLO fatty distribution of density AUC. Multivariate KNN model statistics are reported in Table 4.3. Of the multivariate models, model (c) had the highest ROC AUC of 0.712, the highest specificity of 0.986, and the highest balanced accuracy of 0.623. Model (c) tied model (b) with a specificity of 0.278. Model (c) achieved the highest ROC AUC of all KNN models, with comparable sensitivity, specificity, and balanced accuracy to the best performing single predictor KNN models. While the KNN LOOCV multivariate model (c) achieved the highest ROC AUC of all models with acceptable discrimination between classifiers, the model was still lacking with a low balanced accuracy and sensitivity. Further exploration into feature choice and repeat modelling with a more balanced data set is necessary to improve model performance.

CHAPTER 5

DISCUSSION

Early breast cancer detection requires robust, multi-modal techniques. The metric space technique was assessed for use with grey-scale Hurst-value maps of screening digital mammograms. The total population included 89 patients, 71 of which were cancer cases and 18 of which were benign. Tumorous MLO and CC views were investigated using five MST output functions, the distribution of area, the distribution of density, the distribution of components, and the unweighted and weighted distribution of filaments.[34,35,38] Each output function resulted in a patient curve where the x-axis was the threshold (Hurst-value range) and the y-axis was . The threshold range of the patient curve can be divided into 3 discrete AUC regions according to the Hurst-value tissue categories where $H \leq 0.45$ is associated with fatty breast tissue, $0.45 < H < 0.55$ is associated with disrupted tissue and tumor-associated loss of homeostasis, and $H \geq 0.55$ is associated with dense breast tissue. The percent yellow squares metric, which is not related to the MST, has been previously validated and was used as a quasi-benchmark statistic.[32] Statistically significant differences were found in MLO views for cancerous versus benign patients where the median cancer percent yellow was higher than the median benign percent yellow, where $p \sim 0.01$. No statistically significant differences were found in the CC views of the percent yellow metric. Minimum significant values in the MLO metric space studies achieved comparable results to the MLO percent yellow squares study across 3 metrics. Additionally the MST improves over the percent yellow squares technique by achieving statistically significant differences in the CC view across 3 metrics.

Output functions can be divided into two categories, those which are and are not functions of minimum component size. For output functions which are a function of minimum component size a data mining exercise was undertaken to determine the best minimum component size for each output function and AUC region. Distribution of area

and density are not functions of component size and so were run once for each view. The MLO and CC distribution of area resulted in no regions of significance. This was due to the quantity of pixels which are in the fatty region compared to the dense regions. In the dense regions dividing the reduced area map by the total area map results in a dilution effect as the reduced map was sufficiently smaller than the total map which reduced the appearance of trend when breaking the area under the curve into discrete regions. A proposed solution to this dilution effect would be to 're-set' the total map at the start of each region, where at H = -0.2 the total map includes fatty and dense pixels, at $H = 0.45$ the total map includes all risky dense and healthy dense pixels, and at $H = 0.55$ the total map contains only healthy dense pixels. The only region of significance found in distribution of density was in the MLO fatty AUC region, where $p \sim 0.01$. As with the distribution of area, the distribution of density was susceptible to a dilution effect by the proportion of fatty pixels within the total intensity map. The change in method of 're-setting' the total intensity map at the start of each tissue region could offer a new way to explore the distributions of area and density that is more suited to the method of breaking up the continuous function into discrete regions.

Distribution of components and distribution of filaments are functions of minimum component size. The distribution of components study only had significance in the MLO risky dense AUC region, $p \sim 0.02$, with a minimum component size of 8 pixels. The median cancer AUC was significantly higher than the median benign AUC in the risky dense region indicating there are generally more dense components in MLO cancer images than in MLO benign.

Distribution of unweighted filaments yielded inconclusive results in the MLO component study, and found significance in the CC fatty and healthy dense AUC region at different minimum component size definitions. In the CC fatty AUC region the minimum component definition was 50 pixels where $p \sim 0.04$, and in the CC healthy dense AUC region the minimum component definition was 18 pixels where $p \sim 0.03$. Large scale trends

in the CC unweighted distribution of component study showed that in the fatty region restricting minimum component sizes to greater than 35 pixels caused yielded p-values which were approximately 0.05, but did not always meet significance. Conversely, in the healthy dense region restricting minimum component sizes to less than 30 yielded p-values which were approximately 0.05. This was due to a filtering effect where an abundance of area in the fatty threshold region means components will generally be larger in area, by defining a smaller component size noise could be introduced obscuring trend. The opposite was true in the healthy dense region where the area that was left in the thresholded image was not likely to meet large minimum component size definitions, and therefore information was lost at minimum component size definitions greater than 30 pixels. Exploring an expanded range of minimum component size definitions in the CC fatty AUC region to greater than 50 pixels is recommended to explore statistical differences that may exists at minimum component sizes greater than explored in this work.

In the CC healthy dense AUC region of the weighted distribution of filaments minimum significance, $p \sim 0.02$ was achieved with a minimum component definition of 22 pixels where the median cancer AUC was higher than the minimum benign AUC. The trend seen in the CC healthy dense AUC region of unweighted distribution of filaments was not seen in the CC healthy dense AUC region of weighted filaments. Where in the unweighted case the AUC region was generally significant when the minimum component size definition was less than 30 pixels, in the weighted case significance was generally seen when the minimum component size definition was in the range of 16 to 30 pixels. In the MLO risky dense AUC region there was a reversal. The median benign AUC was greater than the median cancer AUC, wherein all other significant AUC regions the median cancer AUC was greater than the median benign AUC.

Of thirty possible metrics, six were found to have AUC ranges with minimum significant values. Models were trained, tested, and quantified using specificity, sensitivity, and balanced accuracy. Output functions which contained regions of significance were used

to train six univariate GLMs resulting in poor balanced accuracy. A shift in method led to the use of KNNs which improved accuracy results. The best performing model was the multivariate KNN whose predictors were MLO risky dense distribution of components AUC at a minimum component size definition of 8 pixel, CC healthy dense weighted distribution of filaments AUC at a minimum component size definition of 22 pixels, and MLO fatty distribution of density AUC. The best model achieved ROC AUC of 0.712, sensitivity of 0.278, specificity of 0.986, and balanced accuracy of 0.623.

Analysis of the Perm data set with the MST resulted in metrics which can discriminate between cancer and benign patients, but with need to improve sensitivity and balanced accuracy. Implementation of changes to the MST to make it more suitable to discrete AUC regions, especially in distribution of area and density, could improve reported ranges of statistical significance or reveal new trends. Methodology of selection of features and predictors in statistical and machine learning models should be improved upon, where ideally feature selection and weighting should be done statistically as a pre-processing step or as a function of the model itself. In the work presented there are trends which were not explored due to statistical insignificance. If these trends were explored information which could improve model discrimination may, regardless of significance. Further work may be done to validate the results reported with testing on a larger more balanced data set.

REFERENCES

[1] Sung, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **71**, 209–249 (2021).

[2] Yaffe, M. & Boyd, N. Quantitative image analysis for estimation of breast cancer risk. In *Handbook of Medical Image Processing and Analysis* (Elsevier Science Technology, 2008), 2 edn.

[3] Bassett, L. & Conner, K. Section 7: Imaging. In *Holland-Frei Cancer Medicine* (BC Decker, 2003), 6 edn.

[4] Memorial Sloan Kettering Cancer Center. Anatomy of the breast. `https://www.mskcc.org/cancer-care/types/breast/anatomy-breast` (2023). [Accessed: 16th March, 2023].

[5] Li, T. *et al.* The association of measured breast tissue characteristics with mammographic density and other risk factors for breast cancer. *Cancer epidemiology, Biomarkers and Prevention* **14**, 343–349 (2005).

[6] Nazari, S. & Mukherjee, P. An overview of mammographic density and its association with breast cancer. *Breast Cancer* **25**, 259–267 (2018).

[7] Sherratt, M., McConnell, J. & Streuli, C. Raised mammographic density: Causative mechanisms and biological consequences. *Breast Cancer Research* **18** (2016).

[8] Boyd, N. *et al.* Mammographic breast density as an intermediate phenotype for breast cancer. *The Lancet Oncology* **6**, 798–808 (2005).

[9] Sickles, EA, D'Orsi CJ, Bassett LW, et al. ACR BI-RADS© Mammography. In *ACR BI-RADS© Atlas, Breast Imaging Reporting and Data System* (American College of Radiology, 2013).

[10] Sprague, B. *et al.* Prevalence of mammographically dense breasts in the united states. *Journal of the National Cancer Institute* **106** (2014).

[11] American Cancer Society. Mammograms. `https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and early-detection/mammograms.html` (2022). [Accessed: 16th March, 2023].

[12] Vachon, C. *et al.* Mammographic density, breast cancer risk and risk prediction. *Breast Cancer Research* **9** (2007).

[13] Vachon, C. *et al.* Longitudinal trends in mammographic percent density and breast cancer risk. *Cancer Epidemiology Biomarkers and Prevention* **16** (2007).

[14] Berry, D. *et al.* Effect of screening and adjuvant therapy on mortality from breast cancer. *The New England Journal of Medicine* **353**, 1784–1792 (2005).

[15] Cancer intervention and surveillance modeling netwok: Cisnet model registry uwbcs (wisconsin). `https://resources.cisnet.cancer.gov/registry/packages/uwbcs-wisconsin/basics` (2013). [Accessed: 16th March, 2023].

[16] Van Ourti, T., O'Donnell, O., Koç, H., Fracheboud, J. & de Koning, H. Effect of screening mammography on breast cancer mortality: Quasi-experimental evidence from rollout of the dutch population-based program with 17-year follow-up of a cohort. *International Journal of Cancer* **146**, 2201–2208 (2020).

[17] Mazzola, E., Blackford, A., Parmigiani, G. & Biswas, S. Recent enhancements to the genetic risk prediction model BRCAPRO. *Cancer Informatics* **14**, 147–157 (2015).

[18] Gail, M. *et al.* Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. *JNCI: Journal of the National Cancer Institute* **81**, 1879–1886 (1989).

[19] Tyrer, J., Duffy, S. & Cuzick, J. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in Medicine* **23**, 1111–1130 (2004).

[20] Claus, E., Risch, N. & Thompson, W. D. Autosomal dominant inheritance of earlyonset breast cancer. implications for risk prediction. *Cancer* **73**, 643–651 (1994).

[21] Ballard-Barbash, R. *et al.* Breast cancer surveillance consortium: A national mammography screening and outcomes database. *American Journal of Roentgenology* **169**, 1001–1008 (1997).

[22] McCarthy, A. *et al.* Performance of breast cancer risk-assessment models in a large mammography cohort. *Journal of the National Cancer Institute* **112**, 489–497 (2020).

[23] Brentnall, A., Cuzick, J., Buist, D. & Aiello Bowles, E. Long-term accuracy of breast cancer risk assessment combining classic risk factors and breast density. *JAMA Oncology* **4** (2018).

[24] Brentnall, A. *et al.* A case-control study to add volumetric or clinical mammographic density into the tyrer-cuzick breast cancer risk model. *Journal of Breast Imaging* **1**, 99–106 (2019).

[25] Reza Kamel, S., YaghoubZadeh, R. & Kheirabadi, M. Improving the performance of support-vector machine by selecting the best features by gray wolf algorithm to increase the accuracy of diagnosis of breast cancer. *Journal of Big Data* **6** (2019).

[26] Chang Yeh, W., Wen Chang, W. & Ying Chung, Y. A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. *Expert Systems with Applications* **36**, 8204–8211 (2009).

[27] Ghani, M., Alam, T. & Jaskani, F. Comparison of classification models for early prediction of breast cancer. *International Conference on Innovative Computing* (2019).

[28] Han, L. & Yin, Z. A hybrid breast cancer classification algorithm based on meta-learning and artificial neural networks. *Frontiers in Oncology* **12** (2022).

[29] Gastounioti, A., Conant, E. & Kontos, D. Beyond breast density: A review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. *Breast Cancer Research* **18** (2016).

[30] Gerasimova-Chechkina, E. *et al.* Combining multifractal analyses of digital mammograms and infrared thermograms to assist in early breast cancer diagnosis. *AIP Conference Proceedings* **1760** (2016).

[31] Marin, Z. *et al.* Mammographic evidence of microenvironment changes in tumorous breasts:. *Medical Physics* **44** (2017).

[32] Gerasimova-Chechkina, E. *et al.* Loss of mammographic tissue homeostasis in invasive lobular and ductal breast carcinomas vs. benign lesions. *Front. Physiol* **12**, 660883 (2021).

[33] Carbone, A. Algorithm to estimate the hurst exponent of high-dimensional fractals. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **76** (2007).

[34] Adams, F. C. A topological/geometrical approach to the study of astrophysical maps. *The Astrophysical Journal* **387**, 572–590 (1992).

[35] Adams, F. C. & Wiseman, J. J. Formal results regarding metric space techniques for the study of astrophysical maps. *The Astrophysical Journal* **435** (1994).

[36] Khalil, A., Joncas, G. & Nekka, F. Morphological analysis of h i features. i. metric space technique. *Astrophysical Journal* **601**, 352–364 (2004).

[37] Robitaille, J. F., Joncas, G. & Khalil, A. Morphological analysis of h i features - iii. metric space technique revisited. *Monthly Notices of the Royal Astronomical Society* **405** (2010).

[38] Wu, Y., Batuski, D. J. & Khalil, A. Multi-scale morphological analysis of sdss dr5 survey using the metric space technique. *Astrophysical Journal* **707**, 1160–1167 (2009).

[39] Garrett, A. M., Khalil, A., Walton, D. O. & Burgess, R. W. Dscam promotes self-avoidance in the developing mouse retina by masking the functions of cadherin superfamily members. *Proceedings of the National Academy of Sciences of the United States of America* **115** (2018).

[40] Breeding, P. W., Blaszkiewicz, M., Townsend, K., Khalil, A. & Tilbury, K. B. Exploratory investigation of the spatial relationships of collagen and nerves in subcutaneous white adipose tissue (scWAT) using 2-photon microscopy. In Periasamy, A., So, P. T. C. & König, K. (eds.) *Multiphoton Microscopy in the Biomedical Sciences XIX*, vol. 10882, 1088218. International Society for Optics and Photonics (SPIE, 2019).

[41] White, B. Hybrid power spectral and wavelet image roughness analysis. `https://digitalcommons.library.umaine.edu/etd/3757` (2023). Electronic Theses and Dissertations. 3757. [Accessed: 14th July, 2023].

[42] Kuhn, M. Building predictive models in r using the caret package. *Journal of Statistical Software* **28**, 1–26 (2008).

[43] Kuhn, M. & Johnson, K. *Applied Predictive Modeling* (Springer New York, 2013).

[44] Wang, G., Teoh, J. Y. C., Lu, J. & Choi, K. S. Least squares support vector machines with fast leave-one-out auc optimization on imbalanced prostate cancer data. *International Journal of Machine Learning and Cybernetics* **11** (2020).

[45] Bzdok, D., Krzywinski, M., Smith, M. & Altman, N. Machine learning: supervised methods. *Nature Methods* 5–6 (2018).

[46] Mandrekar, J. Receiver operating characteristic curve in diagnostic test assessment elsevier enhanced reader. *Journal of Thoracic Oncology* **5**, 1315–1316 (2010).

[47] Trevethan, R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health* **5** (2017).

[48] Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, 3121–3124 (2010).

[49] Hosmer, D. W., Jr., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression* (John Wiley Sons, Inc., Hoboken, New Jersey, 2013), 3 edn.

BIOGRAPHY OF THE AUTHOR

Madison Raza is a candidate for the Master of Science degree in Biomedical Engineering from the University of Maine, August 2023.