MULTISCALE AND META-ANALYTIC APPROACHES TO INFERENCE IN CLINICAL HEALTHCARE DATA

A Thesis
Presented to
The Academic Faculty

by

Erin Kinzel Hamilton

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
Wallace H. Coulter Department of Biomedical Engineering

Georgia Institute of Technology
May 2013

MULTISCALE AND META-ANALYTIC APPROACHES TO INFERENCE IN CLINICAL HEALTHCARE DATA

Approved by:

Brani Vidakovic, PhD, Advisor
Department of Biomedical Engineering
Georgia Institute of Technology

Paul Griffin, PhD, Advisor
Department of Industrial and Manufacturing Engineering
Pennsylvania State University

Susan Griffin, PhD
Division of Oral Health
National Center for Chronic Disease Prevention and Health Promotion
Centers for Disease Control and Prevention

Melissa Kemp, PhD
Department of Biomedical Engineering
Georgia Institute of Technology

David Goldsman, PhD
School of Industrial & Systems Engineering
Georgia Institute of Technology

Date Approved: February 2013

For my family – past, present, and future.

"This is my command – be strong and courageous! Do not be afraid or discouraged. For the Lord your God is with you wherever you go."

–Joshua 1:9

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to Dr. Brani Vidakovic for his constant guidance and support throughout the course of this work. I would not have been able to complete this thesis without his instruction, enthusiasm, and extreme patience. He is truly one of the kindest people I have ever met, and I am very fortunate to have had the opportunity to work under the advisement of someone so intelligent and so kind to those he works with. I would also like to thank Dr. Paul Griffin for his constant willingness to provide valuable advice and guidance along the way, with constant patience and extreme kindness. I am grateful to both of these men for their willingness to help and their continued availability during my time as a student at Georgia Tech.

I would like to thank Dr. Julie Jacko for seeing potential in me and for encouraging me to pursue things further than I had ever envisioned. I am so grateful for the doors of opportunity she opened for me during her time of advisement and guiding me through the first years of pursuing my Ph.D. I would not be where I am now without the confidence she had in me, and for that I am forever grateful.

I would also like to express my gratitute to my thesis committee members, Dr. David Goldsman, Dr. Susan Griffin, and Dr. Melissa Kemp for their time, their interest, and their productive questions and comments along the way.

I am thankful to my CDC-ORISE fellowship supervisors, Laurie Barker and Susan Griffin. They provided me with a wonderful environment and a tremendous amount of support during my time at the CDC, and they both helped prepare and encourage me for future endeavors. I am extremely grateful for their patience, support, and time.

I would also like to thank my fellow graduate students who have added so much to my years here at Georgia Tech. I would particularly like to thank Dr. Mahima Ashok who made lab a very happy place to be during her time here at Georgia Tech, has remained a support and encouragement to me from the outside world, and continues to be an amazing friend.

I have an amazing set of girlfriends who have each come into my life just when they were needed, who have continued to stand faithfully by my side throughout my years of graduate school. I am grateful to God for bringing each of them into my life and for their individual strengths that have helped me grow. I am grateful to these women for their unfailing encouragement, support, prayers, and love throughout the various stages of pursuing my Ph.D.

I would like to thank my dear, precious family, my very first and forever teachers, who have supported and encouraged me in all of my pursuits. I would not have made it this far without the foundation that they gave me, followed by their constant willingness to provide wisdom, support, prayers, and love.

And finally, I am so grateful to my patient and loving husband Lee, who has been by my side through it all, since the beginning of this Ph.D. – my strength, support, and constant source of joy. I could not have chosen a better match and partner in life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The field of medicine is regularly faced with the challenge of utilizing information that is complicated or difficult to characterize. Physicians often must use their best judgment in reaching decisions or recommendations for treatment in the clinical setting. The goal of this thesis is to use innovative statistical tools in tackling three specific challenges of this nature from current healthcare applications.

The first aim focuses on developing a novel approach to meta-analysis when combining binary data from multiple studies of paired design, particularly in cases of high heterogeneity between studies. The challenge is in properly accounting for heterogeneity when dealing with a low or moderate number of studies, and with a rarely occurring outcome. The proposed approach uses a Rasch model for translating data from multiple paired studies into a unified structure that allows for properly handling variability associated with both pair effects and study effects. Analysis is then performed using a Bayesian hierarchical structure, which accounts for heterogeneity in a direct way within the variances of the separate generating distributions for each model parameter. This approach is applied to the debated topic within the dental community of the comparative effectiveness of materials used for pit-and-fissure sealants.

The second and third aims of this research both have applications in early detection of breast cancer. The interpretation of a mammogram is often difficult since signs of early disease are often minuscule, and the appearance of even normal tissue can be highly variable and complex. Physicians often have to consider many important pieces of the whole picture when trying to assess next steps. The final two aims focus on improving the interpretation of findings in mammograms to aid in early cancer

detection.

When dealing with high frequency and irregular data, as is seen in most medical images, the behaviors of these complex structures are often difficult or impossible to quantify by standard modeling techniques. But a commonly occurring phenomenon in high-frequency data is that of regular scaling. The second aim in this thesis is to develop and evaluate a wavelet-based scaling estimator that reduces the information in a mammogram down to an informative and low-dimensional quantification of the innate scaling behavior, optimized for use in classifying the tissue as cancerous or non-cancerous. The specific demands for this estimator are that it be robust with respect to distributional assumptions on the data, and with respect to outlier levels in the frequency domain representation of the data.

The final aim in this research focuses on enhancing the visualization of micro-calcifications that are too small to capture well on screening mammograms. Using scale-mixing discrete wavelet transform methods, the existing detail information contained in a very small and course image will be used to impute scaled details at finer levels. These "informed" finer details will then be used to produce an image of much higher resolution than the original, improving the visualization of the object. The goal is to also produce a confidence area for the true location of the shape's borders, allowing for more accurate feature assessment. Through the more accurate assessment of these very small shapes, physicians may be more confident in deciding next steps.

CHAPTER I

META-ANALYSIS OF PAIRED BINARY DATA BY A RASCH-TYPE BAYESIAN HIERARCHICAL MODEL

1.1 Introduction

Meta-analysis refers to the quantitative synthesis of evidence from several research studies for a better understanding of a treatment and its effects. The general aim in a meta-analysis is to more powerfully assess the true effect size. In addition, meta-analysis can be useful for resolving inconsistent results from several related but independent studies, reconciling the complete evidence and estimating an average effect [37].

Many clinical experiments result in data in the form of matched-pairs. For example, crossover trials in drug efficacy, or split-mouth designs in dentistry produce matched-pairs tabular data. Because the control and test groups involve the same individual, or appropriately matched individuals, this design controls for many confounding factors. The advantage of such designs is that they account for the variability between the subjects or between the pairs. Typically, when they are feasible, such designs are preferred because they require a smaller sample size compared to parallel (or unpaired) designs to achieve the same inferential power.

The idea of representing paired data in the form of parallelized binary response tables is time-honored and now textbook material in epidemiology. For example, Kahn and Sempos (1989) combine 0-1 tables, and Mantel-Haenszel theory to assess the risk ratio in a matched-pair table [21]. In this type of representation, a paired observation (paired combination of events/non-events) is represented as a single row in a matrix with two columns – one for each of two binary responses. Ghosh et

al. (2000) represented paired tables from clinical trials comparing two treatments via a binary response table, and modeled response probabilities by a Rasch model [16]. Since the Rasch model is traditionally used in educational assessments, one can informally link the treatments in this case to questions on an exam. If a matched-pair table has n entries, then one may imagine n students each answering two questions, in which the correct answer is coded by 1, and the incorrect answer is coded by 0. Then the paired contingency table of size n corresponds to $2n$ answers that can be fitted into the Rasch paradigm. The basic Rasch setup models the probability of a correct answer to a particular item via a logit function that depends on the item difficulty and the responder's ability. This idea is extended in this paper to accommodate combining several different studies into a Rasch-type meta-analysis where in addition to pair and treatment effects, the model can account for the differences between the studies.

Our approach is Bayesian. Conducting a meta-analysis in a Bayesian fashion is conceptually straight-forward because the methodology assumes existence of a meta-model from which the individual models corresponding to particular studies are generated. Thus the meta-analysis translates to a hierarchical Bayesian inference, and in particular, to an inference about the meta-model which is set at the highest level of the hierarchy. The parameters of the meta-model represent the effects attributable to the pair, the study, and typically of most interest, the treatments. The Bayesian approach allows for the fusion of studies and at the same time estimates the heterogeneity and properly accounts for the variability among the studies naturally. This is in contrast to the classical methods in which the heterogeneity assessment directs the choice of methodology.

1.2 Background

1.2.1 Analysis of Paired Tables

A type of randomized controlled trial commonly used in a variety of settings including education, psychology, dentistry, ophthalmology, and pharmacology trials is a matched-pair design where interventions are applied to the same patient or to patients matched with respect to one or more covariates whose influence is to be controlled. For example, in randomized split-mouth trials comparing the effectiveness of tooth-specific interventions to prevent decay, one tooth in a subject is randomly selected to receive treatment A while the contralateral tooth in the same subject receives treatment B. Another example is a cross-over trial testing the efficacy of drugs. In this design, a patient is randomly administered treatment A or B in the first time period and then after an appropriate washout time interval, administered the remaining treatment in the second time period. The tooth location in the split-mouth design is analogous to order in time in the cross-over design. The pair (i.e. matched subjects, pair of teeth, pair of eyes) forms the unit of randomization for assignment of the treatment. Because the control and test groups are matched, this design controls for many confounding factors. Thus differences in outcomes between test and control groups are likely attributable to the treatment.

The matched-pair design for comparing two treatments can be represented as in Table 1. The sample size n relates to the number of paired observations and cell counts y_{ab} represent the number of pairs for which one of the four combinations of events/non-events was observed. Here $a = 0, 1$ is the response to Treatment A and $b = 0, 1$ is the response to Treatment B.

A split-mouth study design affects the method of analysis since the pairs are not independent as they would be when dealing with separate cases and controls. The

Table 1: Matched-pair design table

		Treatment B		
		Event	Non-Event	TOTAL
Treatment A	Event	y_{11}	y_{10}	$y_{1\cdot}$
	Non-Event	y_{01}	y_{00}	$y_{0\cdot}$
TOTAL		$y_{\cdot 1}$	$y_{\cdot 0}$	n

analysis must account for the fact that the data are "paired" by the same observational unit. The statistical methods used for meta-analysis of paired binary data must then be appropriate for combining these data from multiple studies, while accounting for pair effects. There are methods in literature concerning analysis for individual studies with paired binary data, but none that specifically address the combination of this type of data from multiple studies [24, 31, 45, 19, 41, 33].

We now give some background on the Rasch model, and discuss the link between this model and paired tables.

1.2.2 Rasch Model

The Rasch Model is related to Item Response Theory, which traditionally is used for analyzing data from psychological or educational assessments. The goal of item response models is to account for various parameters affecting the outcomes such as an individual's abilities, attitudes, or other traits, as well as parameters concerning the item itself, such as item difficulty. The formal structure of a Rasch model permits algebraic separation of parameters. Its defining feature is *invariant comparison*, which is the ability to compare particular parameters independently of the others. For example, items may be compared independently of the particular individuals that were used for the comparison, and independently of which other items are being compared.

As an illustration, in the educational context the items would be questions on a test, and the model would contain a parameter for item difficulty. Then in a class of students taking a test consisting of multiple items (or questions), the probability of

subject i giving a correct response to item j may be modeled using a binary logistic regression,

$$r_{ij} \sim \mathcal{B}ernoulli\,(p_{ij})\,, \tag{1}$$
$$\text{logit}(p_{ij}) = \alpha_i - \delta_j,$$

where α_i represents the ability of the subject i, and δ_j represents the difficulty of the item j. Observed r_{ij} is the binary response: 1 if the answer is correct, and 0 if the answer is wrong. When comparing student ability, item difficulty is typically considered a nuisance parameter. This model allows the analysis to take into account the differences in the items themselves to more accurately assess the ability of the individual, rather than assuming each item to be of similar difficulty. At first glance, the model looks unidentifiable. However, we are not interested in absolute values for α_i and δ_j, rather in a comparative order of α_i's and δ_j's.

Although originating in education and psychometrics, Rasch models are increasingly being used in a range of areas because of their general applicability. The model requires a specific structure in the response data, and the equations model the relationships we expect to obtain from that structured data.

Ghosh, et al. [16] set out to analyze binary matched pairs data via a hierarchical Bayesian model, and did so by introducing the use of a Rasch model to represent a paired data table. In this context, they restructured the data from the table of paired event counts into a matrix of individual binary responses, as shown in Figure 1. The likelihood of a certain binary response was modeled using parameters of treatment effect (α_j) and pair effect (θ_i). If r_{ij} – the entry in the (i,j)th matrix position – is the binary response of the jth observation within the ith pair, then $p_{ij} = P(r_{ij} = 1)$. These probabilities are modeled as $p_{ij} = F(\theta_i + \alpha_j)$, where F is a cumulative distribution function, usually normal, logistic, or extreme value. The parameter θ_i for pair i is considered a nuisance parameter, while α_j represents the

effect of the jth treatment. Through this analysis, Ghosh et al. were able to describe the behavior of the posterior mean of $\alpha_1 - \alpha_2$ and the posterior probability that Treatment 1 is better than Treatment 2.

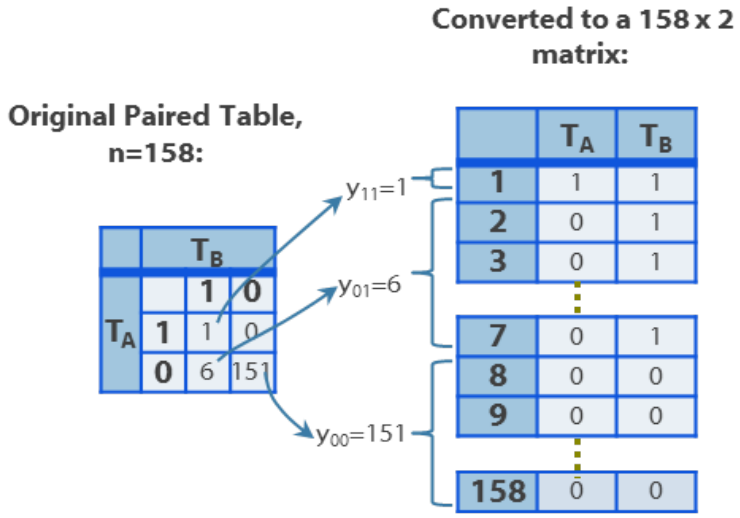


Figure 1: An example of a Rasch-model representation for a paired table: Cell counts y_{ab} represent the number of pairs falling into each of four observed event/non-event combinations. This is converted to an $n \times 2$ matrix where r_{ij} is the response in the ith pair to the jth treatment and is in the (i, j)th matrix position.

In the next section we describe the idea of a Bayesian hierarchical model for meta-analysis, and list some of the advantages of using Bayesian methods.

1.2.3 Bayesian Meta-Analysis

The term "meta-analysis" refers to statistical methods of combining evidence from multiple sources. The general aim of a meta-analysis is to more powerfully estimate the true effect size by merging information from several studies, rather than using a single study under a single set of assumptions and conditions. A meta-analysis provides more statistical power to detect significant effects than analysis based on only one study [37]. In addition to the increase in analytical power, meta-analyses can also be useful for resolving inconsistent outcomes from multiple studies. When several related but independent studies have conflicting conclusions, a meta-analysis

can be used to reconcile the complete evidence and estimate an average effect [37].

Bayesian methods are increasingly being used in health care research partly due to their ability to overcome some of the difficulties met by other more classical methods. A key difference between classical and Bayesian methods is in how they interpret unknown parameters of interest. Classical methods assume there is one true value of a particular parameter and this value can be estimated from the observed data. Bayesians consider the model parameters to be random variables whose conditional distributions depend on observed data. An unknown parameter is treated as a random variable that is generated from an underlying distribution with typically unknown parameters of its own (called hyperparameters). The likelihood function then defines the plausibility of the observed data, conditional on the model parameters.

A clear advantage of the Bayesian approach is that we can incorporate all available pre-experimental information in a coherent way. In addition, the Bayesian paradigm accounts for possible sources of variability in the model. This is useful in the setting of meta-analysis, particularly when the heterogeneity between studies is significant. Bayesian methods can easily handle the question of between-study heterogeneity by accounting for this in the variance of the generating distributions of model parameters. In the classical approach, one usually tests for homogeneity of the studies, and the results of this test will inform whether a fixed-effect model or random-effects model is recommended. Bayesian methods simplify the procedure by accounting for these heterogeneities naturally, whether they are there or not.

Bayesian methods also lend a very natural approach to use for both superiority and non-inferiority testing. Rather than the inherent asymmetry associated with classical hypothesis tests, the Bayesian test amounts to a comparison of posterior probabilities of two competing hypotheses of the true value of a parameter falling within two non-overlapping regions. In standard classical testing for a difference, if the null hypothesis (no difference) cannot be rejected, this still does not indicate

the probability of the null hypothesis actually being true. A different set-up would be needed that takes the null hypothesis to be that there *is* some level of difference between treatments. With Bayesian testing, conclusions are theoretically much more straightforward because probability statements can be directly made regarding the competing hypotheses. It is not limited to a binary conclusion of rejecting or failing to reject the null hypothesis or the p-values, but instead provides a natural assessment of the probabilities of each hypothesis being true.

1.3 Rasch-type Bayesian Hierarchical Model

The staple of our work is the extension of the Rasch model, described in Section 1.2.2, into the context of combining data from multiple paired data tables. This type of model provides a natural mechanism for combining multiple paired tables and modeling in a Bayesian hierarchical structure, however, it has not been used in the context of a meta-analysis previously.

The basic idea of hierarchical modeling is to think of the lowest-level units as organized into a hierarchy of successively higher-level units. Keeping with the educational example, this can be seen as students in classes, and classes in schools. We can then describe outcomes for an individual student as a sum of effects for the individual student, for his class, and for the school. Each of these effects can be regarded as one of an exchangeable collection of effects (e.g. all school-level effects) drawn from a common distribution. Once the model is specified, inferences can be drawn from available data for the population means at any level (school, class, etc.).

In the same manner that Ghosh et al. converted a paired table into a binary logistic model, we can take a series of paired tables containing data from individual studies and combine them into one single array containing all of the binary responses for each individual within each study. Then s paired tables of respective sample sizes $(n_1, \ldots, n_s)$ are restructured into a $s \times n_{max} \times 2$ array, where the binary value r_{uij} –

the response to treatment j in the ith pair in study u – is in the position (u, i, j).

Let $p_{uij} = P(r_{uij} = 1)$, where $u = 1, \ldots, s$, $i = 1, \ldots, n$, and $j = 1, 2$. Then r_{uij} can be modeled as

$$r_{uij} \sim \mathcal{B}ernoulli\,(p_{uij}) \tag{2}$$
$$\text{logit}(p_{uij}) = \beta_u + \theta_{ui} - \alpha_j,$$

where β_u is the effect of study u, θ_{ui} is the effect of pair i within study u, and α_j is the effect of treatment j.

By modeling each individual response in this way, and leveraging the Bayesian hierarchical structure, we can account for the possible sources of heterogeneity at each level (the pair level, the study level, or the overall meta-analysis level) via a posterior distribution of the appropriate parameter. In addition, the model allows for "borrowing of strength" among the studies.

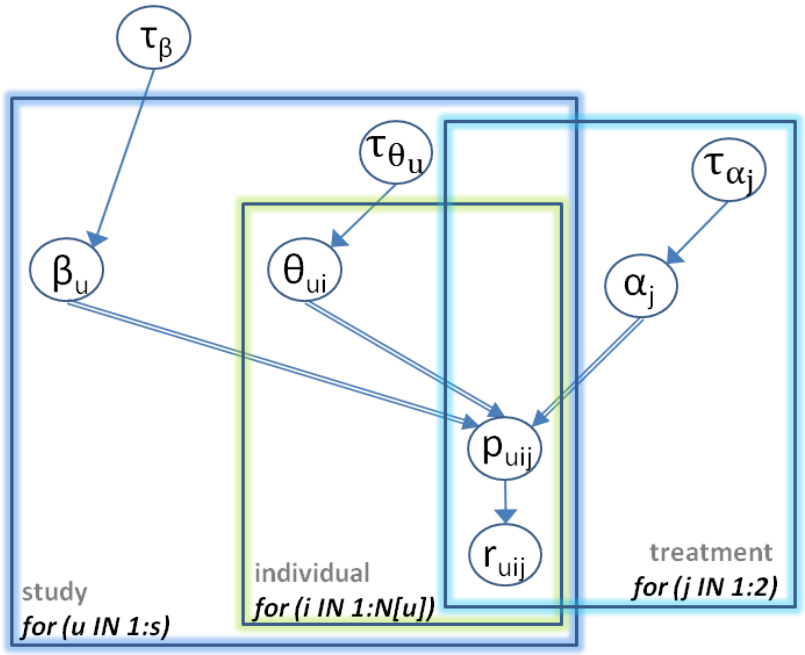


Figure 2: Graphical representation of the hierarchical Bayes model

The model given in (2) is illustrated as a Bayesian graphical model in Figure 2. This visual representation shows that structurally participants are pooled into

studies and studies are pooled into a meta-analysis. The pair effects (θ_{ui}) have their own generating distribution within each individual study. They are treated as nuisance parameters, so are bound to sum to zero at each individual study level. The study effects are also treated as nuisance parameters, so are bound to sum to zero at the meta-analysis level. By treating both pair effects and study effects as nuisance parameters, we are left with the overall treatment effect at the final meta-analysis level.

To clarify how each parameter contributes to the response probabilities at each level, each level is described, beginning at the pair level. The probability of a particular response in pair i to treatment j in study u is modeled as

$$\text{logit}(p_{uij}) = \beta_u + \theta_{ui} - \alpha_j. \tag{3}$$

Since the pair effects (θ_{ui}) in (3) are treated as nuisance parameters, they are modeled in such a way to not individually contribute to the response probabilities once we move up to the higher level of the model. Thus, the probability of a particular response to treatment j in study u is

$$\text{logit}(p_{uj}) = \beta_u - \alpha_j. \tag{4}$$

And since the study effects (β_u) in (4) are also treated as nuisance parameters, they are modeled in such a way to not individually contribute to the response probabilities once we move up to the highest meta-analysis level. However the β_u's are important to assess the heterogeneity in the meta-analysis. The pooled probability of a particular response to treatment j is

$$\text{logit}(p_j) = -\alpha_j. \tag{5}$$

Thus, we are left we the overall treatment effect at the final meta-analysis level in (5).

1.3.1 Prior Settings

A Bayesian model is fully specified when all parameters within the model are assigned prior distributions. For pair and study effects, standard non-informative prior settings are used. We first assume that each pair effect θ_{ui} is sampled from a study-specific normal distribution with a mean of 0, and variance $1/\tau_{\theta_u}$, where τ_{θ_u} is the precision of the distribution. The exception to this is θ_{u1}, which is used to set sum-to-zero constraints on all pair effects within each study for identifiability of the model.

$$\theta_{ui} \sim \mathcal{N}ormal\left(0, \frac{1}{\tau_{\theta_u}}\right),\ i = 2, 3, \ldots, n_u; \qquad \theta_{u1} = -\sum_{i=2}^{n_u} \theta_{ui}, u = 1, 2, \ldots, s.$$

Here $1/\tau_{\theta_u}$ represents the between-pair variation within study u, with pair effects averaging to zero within study u. We adopt a non-informative hyperprior on τ_{θ_u} as $\tau_{\theta_u} \sim \mathcal{G}amma(0.001, 0.001)$.

Similar to the treatment of pair effects within each study, we assume that each study effect β_u is sampled from a normal distribution with a mean of 0, and variance $1/\tau_\beta$, with the exception of β_1, which is used to set sum-to-zero constraints on all study effects to maintain identifiability of the model.

$$\beta_u \sim \mathcal{N}ormal\left(0, \frac{1}{\tau_\beta}\right),\ u = 2, 3, \ldots, s; \qquad \beta_1 = -\sum_{u=2}^{s} \beta_u.$$

Here $1/\tau_\beta$ represents the between-study variation, with all study effects averaging to zero. We adopt a non-informative hyperprior on τ_β as $\tau_\beta \sim \mathcal{G}amma(0.001, 0.001)$.

More care is needed for treatment effects because this parameter provides a particular opportunity to incorporate any available prior information, an advantage of the Bayes approach. This is manifested in the prior chosen to place on α_j's. To begin, we first set a model that keeps a completely non-informative setup, where each treatment effect α_j is assumed to be sampled from a prior normal distribution with a mean of 0, and variance $1/\tau_{\alpha_j}$,

$$\alpha_j \sim \mathcal{N}ormal\left(0, \frac{1}{\tau_{\alpha_j}}\right),$$

which incorporates non-informativity at two levels: means of zero and non-informative precision. We adopt a non-informative hyperprior on τ_{α_j} of $\tau_{\alpha_j} \sim \mathcal{G}amma(0.001, 0.001)$. Here the zero mean on α_j implies both treatments have a 50% probability of a particular outcome *a priori*, which is also a non-informative selection of location of this prior. For these reasons, we will call the model with this prior the Non-Informed Bayesian Rasch model (NBR), although the distribution may not be non-informative "on the average".

As mentioned, an alternative approach besides the NBR model is to use some informative priors, because sometimes we do know that a 50% probability of a certain outcome is not representative of the data we are working with. For example, in data used to compare the probability of tooth decay after treatment with two different dental sealant materials, the probability of decay after either treatment is actually much closer to the range of only 0% to 15%. So we can use a more informative prior on the α_j's that actually provides a more representative model *a priori*. In this case, each treatment effect (α_j) is assumed to be sampled from a prior normal distribution with some postulated mean $\hat{\alpha}$, and variance $1/\tau_{\alpha_j}$,

$$\alpha_j \sim \mathcal{N}ormal\left(\hat{\alpha}, \frac{1}{\tau_{\alpha_j}}\right).$$

In this case, we again adopt a hyperprior on the precision τ_{α_j} as $\tau_{\alpha_j} \sim \mathcal{G}amma(0.001, 0.001)$. The $\hat{\alpha}$ mean on α_j is calibrated by $\exp\{-\hat{\alpha}\}/(1+\exp\{-\hat{\alpha}\})$, which is the probability of the outcome of interest, *a priori*. This model with more informative priors on α_j (more informed in regards to location, not by scale) is called here the Informed Bayesian Rasch model (IBR).

Then, in either case (NBR or IBR), the pooled probability of a particular response to treatment j is modeled as $\text{logit}(p_j) = -\alpha_j$, where p_j is the probability of a particular response after treatment j (as shown in Section 1.3). The posterior median of RR and its 95% credible interval (CI) are adopted as summary measures of the meta-analysis to compare the performance of treatments, where $RR = p_2/p_1$. For a fixed table the

estimator of RR is defined as $\widehat{RR} = \hat{p}_2/\hat{p}_1$, where $\hat{p}_1$ and $\hat{p}_2$ are estimators of p_1 and p_2.

1.4 Simulations

In this section we demonstrate how our methodology performs in assessing the overall RR in data simulated from a known distribution with known RR. Data in 2×2 paired contingency tables are simulated by two different methods, intended to vary the level of heterogeneity among the simulated studies:

1. From a multinomial distribution with parameter vector $\mathbf{p} = [p_1, p_2, p_3, p_4]$, a data vector $\mathbf{x} = [x_1, x_2, x_3, x_4]$ is drawn, where the coordinates are the entries of the table (row-wise). This drawing of $\mathbf{x}$ is repeated for the desired number of studies to pool. In this case, RR is known to be equal to $(p_1 + p_2)/(p_1 + p_3)$.

2. From a Dirichlet distribution with parameter vector $\mathbf{a} = [a_1, a_2, a_3, a_4]$, a parameter vector $\mathbf{p} = [p_1, p_2, p_3, p_4]$ is drawn. Then from a multinomial distribution with that parameter vector $\mathbf{p}$, a data vector $\mathbf{x} = [x_1, x_2, x_3, x_4]$ is drawn. This drawing of $\mathbf{p}$ and then $\mathbf{x}$ is repeated for the desired number of studies to pool. In this case, RR is known to be equal to $(a_1 + a_2)/(a_1 + a_3)$.

Tables 2-4 each show results from ten different simulation runs in which eight studies were drawn from known distributions with known RR, as described above. These eight studies were then pooled within each run. The I^2 statistic (the proportion of total variability explained by heterogeneity) was not preassigned, but calculated once the tables were generated. Simulation runs are labeled 1 through 10 within each table and are ordered by increasing heterogeneity. In addition to the proposed methods (NBR and IBR), we consider the DerSimonian-Laird (DSL) method, which is a common random effects model for meta-analysis.

The performance of each method is summarized according to accuracy of $\widehat{RR}$ as well as adequacy if CI's. Typically, when comparing two interventions in medicine,

an important criteria in determining whether the interventions can be considered different is whether or not the CI of RR contains 1. If the CI does not contain 1 (and the effect measure is sufficiently high to be considered clinically significant), then the treatment effects are considered different. Otherwise, if the CI does contain 1, then treatments cannot be considered different. Thus, the results will include some discussion of false positives (actual RR is equal to 1, but the CI of RR does not contain 1, thus resulting in an erroneous conclusion of a difference between treatments) and false negatives (actual RR is not equal to 1, but the CI of RR contains 1, thus resulting in an erroneous failure to conclude a difference between treatments).

Table 2 shows results from ten different simulation runs in which studies were drawn from original distributions with $\mathbf{p} = [.01, .02, .02, .95]$ or $\mathbf{a} = [1, 2, 2, 95]$. In this setup, there is a 3% probability of the particular outcome of interest, and $RR = 1$. For each run, the estimate that came closest to the true RR is bolded and underlined. While all three methods performed well in this setup, the frequency of being the most accurate in the point estimate of RR is 7/10 for IBR, 2/10 for DSL, and 1/10 for NBR. The number of false positives (CI not containing 1) was not high. The single case is shown with the CI in italicized text, using DSL, and with 0% heterogeneity. And finally, as heterogeneity increases, it is important to notice that the increase in the width of the CI of the Bayesian models is not quite as large as that for DSL.

For a very similar setup, but with different RR, Table 3 shows results from ten runs done with studies drawn from distributions with $\mathbf{p} = [.01, .03, .01, .95]$ or $\mathbf{a} = [1, 3, 1, 95]$. In this setup, there is again a 3% probability of the particular outcome of interest, but this time $RR = 2$. Again, all three methods performed well, IBR and NBR both having the highest frequency of being the most accurate (each 4/10), followed by DSL (2/10). In this case, since the true RR differs from 1, we discuss false negatives (CI containing 1). These cases are shown with an asterisk (*) next

Table 2: Results from estimation of $\widehat{RR}$ in ten different simulation runs, pooling eight studies each. Data were pulled from original distributions with $\mathbf{p} = [.01, .02, .02, .95]$ or $\mathbf{a} = [1, 2, 2, 95]$. There is a 3% probability of the particular outcome of interest, and $RR = 1$

	DSL $\widehat{RR}$	**NBR** $\widehat{RR}$	**IBR** $\widehat{RR}$
Simulation 1 I^2 = **0%**	1.38 *(1.02, 1.87)*	1.43 (1.00, 2.06)	**1.25** (0.94, 1.83)
Simulation 2 I^2 = **26%**	0.91 (0.61, 1.35)	0.88 (0.61, 1.27)	**0.94** (0.68, 1.22)
Simulation 3 I^2 = **40%**	0.93 (0.46, 1.90)	**1.00** (0.66, 1.50)	0.99 (0.64, 1.53)
Simulation 4 I^2 = **53%**	1.14 (0.73, 1.80)	1.18 (0.83, 1.67)	**1.09** (0.84, 1.49)
Simulation 5 I^2 = **54%**	**1.11** (0.73, 1.68)	1.27 (0.94, 1.70)	1.16 (0.92, 1.54)
Simulation 6 I^2 = **64%**	**0.99** (0.62, 1.58)	0.80 (0.56, 1.14)	0.86 (0.60, 1.12)
Simulation 7 I^2 = **64%**	1.23 (0.59, 2.55)	1.32 (0.89, 1.94)	**1.18** (0.90, 1.73)
Simulation 8 I^2 = **67%**	0.79 (0.42, 1.49)	0.73 (0.51, 1.03)	**0.81** (0.56, 1.07)
Simulation 9 I^2 = **73%**	1.23 (0.63, 2.42)	1.18 (0.84, 1.64)	**1.08** (0.86, 1.45)
Simulation 10 I^2 = **75%**	0.68 (0.32, 1.45)	0.74 (0.53, 1.02)	**0.82** (0.59, 1.06)
Average Point Estimation	**1.04**	**1.05**	**1.02**

to the CI. DSL had a 6/10 frequency of a false negative, IBR a 1/10 frequency, and NBR a 0/10 frequency. As mentioned above, as heterogeneity increases, the increase in the width of the CI in the Bayesian models is not quite as large as that for DSL. The effects of these wider CI's can now be seen in the extent of false negatives this can produce, showing the importance of the better performance of the Bayes-Rasch methods in accounting for study heterogeneity within the study effect parameter. Finally, we can see in Simulation 8 that IBR had one case where the CI (italicized) did not contain the true RR of 2, although this was a very small discrepancy (upper bound of 1.99). However, both this discrepancy and this one case of a false negative

for IBR can be attributed to the bias in the point estimation of $\widehat{RR}$ for this run, so that the CI surrounding the estimate is shifted.

Table 3: Results from estimation of $\widehat{RR}$ in ten different simulation runs, pooling eight studies each. Data were pulled from original distributions with $\mathbf{p} = [.01, .03, .01, .95]$ or $\mathbf{a} = [1, 3, 1, 95]$. There is a 3% probability of the particular outcome of interest, and $RR = 2$

	DSL $\widehat{RR}$	**NBR** $\widehat{RR}$	**IBR** $\widehat{RR}$
Simulation 1 I^2 = **0%**	1.96 (1.44, 2.68)	**2.01** (1.40, 2.94)	1.84 (1.28, 2.66)
Simulation 2 I^2 = **15%**	2.30 (1.61, 3.27)	2.35 (1.65, 3.40)	**2.26** (1.59, 3.25)
Simulation 3 I^2 = **25%**	**2.60** (1.74, 3.89)	2.89 (1.90, 4.42)	2.69 (1.79, 4.23)
Simulation 4 I^2 = **56%**	1.71 (1.06, 3.89)	**2.01** (1.43, 2.84)	1.85 (1.31, 2.68)
Simulation 5 I^2 = **58%**	1.37 *(0.81, 2.31)	**1.96** (1.28, 3.12)	1.94 (1.26, 3.01)
Simulation 6 I^2 = **60%**	1.58 *(0.86, 2.91)	2.36 (1.58, 3.57)	**2.19** (1.48, 3.39)
Simulation 7 I^2 = **64%**	1.95 *(0.98, 3.86)	2.07 (1.35, 3.27)	**2.01** (1.31, 3.14)
Simulation 8 I^2 = **70%**	**1.60** *(0.88, 2.91)	1.51 (1.04, 2.20)	1.32 *(0.97, 1.99)*
Simulation 9 I^2 = **73%**	1.39 *(0.78, 2.49)	**1.92** (1.35, 2.77)	1.79 (1.25, 2.59)
Simulation 10 I^2 = **82%**	2.20 *(0.99, 4.88)	2.11 (1.49, 2.98)	**2.03** (1.47, 2.89)
Average Point Estimation	**1.87**	**2.12**	**1.99**

And finally, Table 4 shows results from ten runs done with studies drawn from distributions with $\mathbf{p} = [.05, .10, .05, .80]$ or $\mathbf{a} = [5, 10, 5, 80]$. In this setup, there is a 12.5% probability of the particular outcome of interest, and $RR = 1.5$. As with other setups, all three methods performed well. Here NBR had the highest frequency of being the most accurate (6/10), followed by IBR and DSL, each with 2/10. There were two false negative from DSL in Simulation 8 and 10 where heterogeneity is high.

Table 4: Results from estimation of $\widehat{RR}$ in ten different simulation runs, pooling eight studies each. Data were pulled from original distributions with $\mathbf{p} = [.05, .10, .05, .80]$ or $\mathbf{a} = [5, 10, 5, 80]$. There is a 12.5% probability of the particular outcome of interest, and $RR = 1.5$

	DSL $\widehat{RR}$	**NBR** $\widehat{RR}$	**IBR** $\widehat{RR}$
Simulation 1 I^2 = **0%**	1.38 (1.23, 1.54)	1.52 (1.27, 1.82)	**1.52** (1.28, 1.81)
Simulation 2 I^2 = **24%**	1.42 (1.19, 1.70)	**1.52** (1.26, 1.82)	1.47 (1.22, 1.75)
Simulation 3 I^2 = **25%**	1.38 (1.15, 1.67)	**1.42** (1.20, 1.70)	1.37 (1.15, 1.65)
Simulation 4 I^2 = **34%**	1.62 (1.36, 1.92)	1.61 (1.35, 1.92)	**1.54** (1.28, 1.86)
Simulation 5 I^2 = **36%**	**1.45** (1.16, 1.81)	1.44 (1.19, 1.75)	1.36 (1.12, 1.66)
Simulation 6 I^2 = **55%**	1.33 (1.02, 1.74)	**1.41** (1.19, 1.68)	1.34 (1.12, 1.61)
Simulation 7 I^2 = **62%**	**1.59** (1.28, 1.97)	1.73 (1.46, 2.06)	1.68 (1.42, 2.00)
Simulation 8 I^2 = **65%**	1.24 *(0.94, 1.64)	**1.32** (1.12, 1.57)	1.26 (1.06, 1.50)
Simulation 9 I^2 = **68%**	1.34 (1.09, 1.65)	**1.42** (1.21, 1.68)	1.40 (1.19, 1.64)
Simulation 10 I^2 = **81%**	1.20 *(0.82, 1.76)	**1.50** (1.21, 1.85)	1.41 (1.14, 1.76)
Average Point Estimation	**1.40**	**1.49**	**1.44**

All methods performed well in simulations combining eight studies, with very few cases of the CI not containing the true RR. Overall, both IBR and NBR seem to out-perform DSL in both accuracy and in less widening of the CI bounds as the heterogeneity increases. The combination of these performance measures translates into less tendency in both IBR and NBR to produce false negatives when the true RR differs from 1. This could also be important in cases where the true RR does equal 1, or is very close, and the question is that of equality or non-inferiority. In such cases it is helpful for confidence bounds to not be so wide as to go beyond the

acceptable bounds of non-inferiority (assuming that the point estimate is accurate), resulting in an erroneous failure to conclude equality or non-inferiority.

To assess what might happen in meta-analyses with a higher number of studies, we ran further simulations, increasing the number of studies combined. Tables 5 through 8 each show results from sets of five different simulation runs in which a higher number of studies are drawn from known distributions, as previously described, and pooled. Tables 5 and 6 combine fifteen studies, while Tables 7 and 8 combine thirty studies. Simulation runs are labeled 1 through 5 within each table, ordered by increasing heterogeneity. Again, the I^2 statistic was not preassigned, but calculated once the tables were generated.

Table 5: Results from estimation of $\widehat{RR}$ in five different simulation runs, pooling fifteen studies each. Data were pulled from original distributions with $\mathbf{a} = [10, 20, 20, 950]$. There is a 3% probability of the particular outcome of interest, and $RR = 1$

	DSL $\widehat{RR}$	**NBR** $\widehat{RR}$	**IBR** $\widehat{RR}$
Simulation 1 I^2 = **17%**	**1.00** (0.79, 1.27)	0.97 (0.75, 1.25)	0.98 (0.78, 1.20)
Simulation 2 I^2 = **29%**	0.97 (0.77, 1.22)	0.96 (0.76, 1.23)	**0.98** (0.79, 1.19)
Simulation 3 I^2 = **33%**	0.88 (0.65, 1.19)	0.91 (0.68, 1.22)	**0.94** (0.72, 1.18)
Simulation 4 I^2 = **44%**	1.02 (0.74, 1.38)	1.00 (0.77, 1.31)	**1.00** (0.80, 1.26)
Simulation 5 I^2 = **54%**	1.15 (0.86, 1.55)	1.14 (0.89, 1.44)	**1.07** (0.90, 1.33)
Average Point Estimation	**1.00**	**1.00**	**0.99**

Table 6: Results from estimation of $\widehat{RR}$ in five different simulation runs, pooling fifteen studies each. Data were pulled from original distributions with $\mathbf{a} = [10, 30, 10, 950]$. There is a 3% probability of the particular outcome of interest, and $RR = 2$

	DSL $\widehat{RR}$	**NBR** $\widehat{RR}$	**IBR** $\widehat{RR}$
Simulation 1 I^2 = **15%**	**1.92** (1.50, 2.45)	2.21 (1.67, 2.92)	2.11 (1.62, 2.80)
Simulation 2 I^2 = **34%**	1.92 (1.50, 2.45)	2.07 (1.62, 2.65)	**1.98** (1.54, 2.55)
Simulation 3 I^2 = **38%**	1.74 (1.29, 2.35)	**2.02** (1.54, 2.72)	1.92 (1.45, 2.56)
Simulation 4 I^2 = **43%**	1.85 (1.32, 2.59)	**1.93** (1.45, 2.58)	1.80 (1.33, 2.43)
Simulation 5 I^2 = **61%**	1.72 (1.25, 2.35)	**1.87** (1.48, 2.37)	1.82 (1.45, 2.30)
Average Point Estimation	**1.83**	**2.02**	**1.93**

Table 7: Results from estimation of $\widehat{RR}$ in five different simulation runs, pooling thirty studies each. Data were pulled from original distributions with $\mathbf{a} = [10, 20, 20, 950]$. There is a 3% probability of the particular outcome of interest, and $RR = 1$

	DSL $\widehat{RR}$	**NBR** $\widehat{RR}$	**IBR** $\widehat{RR}$
Simulation 1 I^2 = **33%**	**0.92** (0.77, 1.11)	0.90 (0.75, 1.08)	0.91 (0.77, 1.08)
Simulation 2 I^2 = **34%**	**0.87** (0.72, 1.06)	0.85 (0.70, 1.02)	0.85 (0.71, 1.02)
Simulation 3 I^2 = **36%**	**0.93** (0.77, 1.12)	0.87 (0.73, 1.04)	0.86 (0.71, 1.05)
Simulation 4 I^2 = **40%**	**1.11** (0.93, 1.31)	1.14 (0.97, 1.35)	1.15 (0.97, 1.36)
Simulation 5 I^2 = **44%**	**1.08** (0.88, 1.32)	1.11 (0.93, 1.32)	1.09 (0.93, 1.29)
Average Point Estimation	**1.00**	**0.99**	**0.99**

From these tables one can see that as the number of pooled studies grows larger, the performance for each method is very comparable. All perform well and provide correct conclusions, with the CI containing 1 or not containing 1, as appropriate. There also does not seem to be as drastic of a difference in CI size as the heterogeneity

Table 8: Results from estimation of $\widehat{RR}$ in five different simulation runs, pooling thirty studies each. Data were pulled from original distributions with $\mathbf{a} = [10, 30, 10, 950]$. There is a 3% probability of the particular outcome of interest, and $RR = 2$

	DSL $\widehat{RR}$	**NBR** $\widehat{RR}$	**IBR** $\widehat{RR}$
Simulation 1 I^2 = **34%**	1.77 (1.44, 2.18)	**1.91** (1.56, 2.35)	1.86 (1.53, 2.27)
Simulation 2 I^2 = **35%**	1.72 (1.40, 2.12)	1.91 (1.56, 2.35)	**1.92** (1.57, 2.34)
Simulation 3 I^2 = **35%**	1.90 (1.54, 2.19)	2.13 (1.75, 2.62)	**2.09** (1.73, 2.56)
Simulation 4 I^2 = **39%**	1.69 (1.43, 2.00)	1.93 (1.60, 2.34)	**1.93** (1.60, 2.33)
Simulation 5 I^2 = **43%**	1.68 (1.38, 2.03)	**1.89** (1.59, 2.26)	1.85 (1.57, 2.21)
Average Point Estimation	**1.75**	**1.95**	**1.93**

increases. This could be attributable to the performance of the methods themselves, or to the fact that as the number of studies increases, it is more difficult to simulate higher heterogeneity. We were only able to reach I^2 values up to 0.5-0.6 with fifteen studies, and up to around 0.4 with thirty studies. But this phenomenon is also present in real-world data, with it being very rare to find an extremely high I^2 value resulting from a very large number of studies. Thus the advantages of these new Bayesian methods are more evident in settings with high heterogeneity among a lower (and more typical) number of studies.

1.5 Application to Data from Dental Sealant Trials

A recent topic of debate in the dental community is concerning the comparative effectiveness of materials used for pit-and-fissure sealants. Sealants are coatings applied on the grooves (or pits and fissures) of primarily molar teeth to prevent the growth of bacteria that promote decay on these surfaces. It has been well proven that children with sealants on their molar teeth are less likely to have dental decay (i.e., caries) in these teeth than children without sealants [40, 23, 1, 10, 2, 17]. The standard of

care, with strong evidence for its effectiveness in caries prevention is resin-based (RB) sealants [23, 29]; however careful technique and moisture handling during application are vital for their effectiveness [6]. Glass Ionomer (GI) sealants have been suggested as an alternative with less sensitivity to moisture in application, which are thus easier to use, particularly in less controlled environments. However evidence to conclude GI's comparative effectiveness to RB is lacking [6]. There is an ongoing question as to whether GI can be considered as effective in caries prevention as RB sealants.

Compared to no sealant, there is stronger evidence for RB's effectiveness than for GI. There have been a number of clinical trials over the years comparing the two, as well as systematic reviews and meta-analyses which have tried to summarize and clarify existing evidence [43, 44, 29, 7, 2, 4, 51, 30]. But the lack of conclusive information remains. The challenge begins with the fact that there are conflicting results from clinical trials, with some favoring RB and others favoring GI. In 1996, Simonsen's narrative assessment of the literature was that retention (a previously accepted proxy for effectiveness) for RB is better than for GI, but differences in caries prevention remain equivocal [43]. However, he later changed his assessment in light of two new studies, concluding that RB is not only superior in terms of retention, but also in caries prevention [44]. In 2003, Mejare et al's discussion of existing literature concluded that there is incomplete evidence that sealing with GI has caries-preventive effect [29]. However, in 2006, Beiruti et al's review concluded that Simonsen's original 1996 conclusion that the materials were equivocal still held true, despite the increase of comparison studies. This was based on "no obvious pattern" in the studies included in their review, and thus no evidence of one being better than the other. They also stated that the absence of homogeneity within the studies prevented a quantitative analysis [7].

In 2008, both the American Dental Association (ADA) and the Canadian Dental

Association (CDA) published recommendations favoring the use of RB, but not excluding GI from consideration. The CDA's recommendation was that RB should be preferred, until a time when GI's with better retention rates are developed [4]. The ADA also stated RB as the first choice of material for dental sealants, but allowed for the use of GI as a short-term prevention strategy when concern about moisture control may compromise the placement of RB [6]. In the same year, a Cochrane Review was published [2], in which eight studies comparing GI to RB fit the inclusion criteria. These authors stated that the results of the studies were too divergent to allow for attempting a quantitative meta-analysis and concluded that more research is needed to clarify the relative effectiveness of the materials.

Yengopal et al published a 2009 study attempting to provide more objective assessment by expanding the literature review and quantifying the outcomes through meta-analysis [51]. This resulted in no evidence that either material was superior to the other in caries prevention. Based on this lack of evidence for a difference (and an implied assumption of sufficient power to detect a difference), the authors concluded that both materials appear equally suitable for use as dental sealants. But in a 2011 update performed by the same authors, RB was then found to be more effective than GI 3 years after placement. They concluded that further high quality randomized control trials are needed to conclusively answer the question whether caries occurrence in teeth sealed with either GI or RB is the same [30].

We apply our methodology to data from this context. In addition to these trials, inclusion of trials that were conducted as true split mouth designs, but reported results in a parallel manner was proposed by Barker et al [5] by Bayesian paired cell recovery. Trials recovered by this method are also included in the analysis. Thirteen studies are included, which collected data over several different years of follow-up. Table 9 shows studies included with their respective years of follow-up.

For each year of follow-up, we compare the resulting estimated $\widehat{RR}$'s and their

Table 9: Studies included in meta-analysis of dental sealant material

Study	1 Yr follow-up	2 Yrs follow-up	3 Yrs follow-up	4+ Yrs follow-up
Mills 1993	n = 59	n = 59		
Forss 1994/1998		n = 151		n = 97
Arrow 1995			n = 412	
Karlzen-Reuterv 1995	n = 74	n = 74	n = 74	
Sipahier 1995	n = 86			
Raadal 1996	n = 136	n = 136	n = 132	
Rock 1996	n = 158	n = 132	n = 130	
Williams 1996		n = 295		n = 225
Morrow 1997		n = 35		
Poulsen 2001		n = 203	n = 206	
Ganesh 2006	n = 100	n = 100		
Kervanto-Seppala 2008			n = 657	
Baseggio 2010	n = 640	n = 640	n = 628	
Number of Studies	**s = 7**	**s = 10**	**s = 7**	**s = 2**
Total Subjects	**N = 1253**	**N = 1825**	**N = 2239**	**N = 322**
	I^2 = 0%	I^2 = 60%	I^2 = 88%	I^2 = 0%

CI's to results from two different methods from literature. The first method included here for comparison is DSL, as mentioned in the previous section. The second method for comparison is a more traditional method of Bayesian meta-analysis in which prior distributions are placed on the relative cell probabilities instead of prior distributions being placed on predictive parameters in a logistic regression, as with the Rasch model setup. We label this method for comparison as MB. Tables 10-14 show data from each study and results from these analyses.

Table 10 shows that in Year 1 neither material can be shown to be superior to the other, using any method. But there is an extremely low occurrence of caries in this group, at <1%. This lack of caries could either be because the interventions are equally effective at 1 year, or because the population was low risk, possibly because they were still very young and the teeth were still fairly new. For further investigation, we ran the analysis again, this time excluding the three studies where no caries occurred at all, to see if a difference might be evident in only the studies where caries were

Table 10: Results from Year 1 Analysis
($I^2 = 0\%$)

Study	Both sound(+)	GI+, RB-	GI-, RB+	Both carious(-)	RR (95% CI)
Ganesh 2006	100	0	0	0	1.000 (0.141, 7.099)
Mills 1993	59	0	0	0	1.000 (0.141, 7.099)
Karlzen-Reuterv 1995	73	1	0	0	0.500 (0.070, 3.550)
Rock 1996	151	0	6	1	7.000 (1.140, 42.971)
Raadal 1996	133	0	3	0	4.000 (0.563, 28.397)
Sipahier 1995	79	1	2	4	1.200 (0.781, 2.634)
Baseggio 2010	640	0	0	0	1.000 (0.141, 7.099)
DSL Pooled					**1.319 (0.852, 2.040)**
MB Pooled					**1.402 (0.653, 3.216)**
NBR Pooled					**2.516 (0.999, 6.907)**
IBR Pooled					**2.248 (0.921, 6.076)**

Table 11: Results from Year 1 Analysis, excluding studies where there was no disease in either group
($I^2 = 45\%$)

Study	Both sound(+)	GI+, RB-	GI-, RB+	Both carious(-)	RR (95% CI)
Karlzen-Reuterv 1995	73	1	0	0	0.500 (0.070, 3.550)
Rock 1996	151	0	6	1	7.000 (1.140, 42.971)
Raadal 1996	133	0	3	0	4.000 (0.563, 28.397)
Sipahier 1995	79	1	2	4	1.200 (0.781, 2.634)
DSL Pooled					**1.787 (0.666, 4.795)**
MB Pooled					**1.690 (0.572, 6.966)**
NBR Pooled					**2.367 (1.048, 6.169)**
IBR Pooled					**1.749 (0.867, 4.684)**

present. Results from this analysis are shown in Table 11. In this case, the NBR method does show RB to be superior to GI, since the CI now does not contain 1. The other three methods all are still unable to show any difference between the two methods.

Table 12 shows results from Year 2 data. Year 2, with ten studies, combined the highest number of studies of all of the follow-up years. In each case, the conclusion is that RB is superior to GI. These studies have heterogeneity of 60%, and you can see that in the cases of NBR and IBR CI's are tighter. This is evidence that the NBR and IBR methods handle this heterogeneity explicitly, allowing for more precision in

Table 12: Results from Year 2 Analysis ($I^2 = 60\%$)

Study	Both sound(+)	GI+, RB-	GI-, RB+	Both carious(-)	RR (95% CI)
Ganesh 2006	100	0	0	0	1.000 (0.141, 7.099)
Mills 1993	59	0	0	0	1.000 (0.141, 7.099)
Poulsen 2001	191	2	9	1	3.333 (1.017, 10.922)
Forss 1994	142	2	2	5	1.000 (0.571, 1.751)
Karlzen-Reuterv 1995	72	1	0	1	0.500 (0.125, 1.999)
Raadal 1996	131	0	5	0	6.000 (0.845, 42.596)
Rock 1996	116	0	14	2	8.000 (2.188, 29.249)
Williams 1996	273	1	16	5	3.500 (1.704, 7.190)
Morrow 1997	29	0	5	1	6.000 (1.003, 35.909)
Baseggio 2010	569	14	51	15	2.850 (1.785, 4.551)
DSL Pooled					**2.310(1.355, 3.940)**
MB Pooled					**2.271(1.170, 4.326)**
NBR Pooled					**3.178(2.232, 4.641)**
IBR Pooled					**3.151(2.223, 4.612)**

estimation of the treatment effect.

Table 13 shows results from Year 3. In this case methods DSL and MB cannot prove a difference between the two methods, while both NBR and IBR show RB to be superior to GI. However, the point estimates are not that different between the four methods. So the difference in conclusion is just attributed to the tighter CI's with NBR and IBR. This illustrates well the point of this new method. Year 3 had 88% heterogeneity, which is accounted for explicitly in the NBR and IBR models. In the other two, the heterogeneity is still evident in the CI's, likely resulting in false negatives.

And finally, results from Year 4 are given in Table 14. Results here are very similar between the four methods, and conclusions would not differ regardless of the method chosen (neither material can be shown to be superior to the other). One thing to note is that even though the heterogeneity is 0% in this case (so it wouldn't be expected to see much narrower CI's with the new methods), CI's are still not as wide for NBR and IBR methods as they are in the more traditional Bayesian method (MB).

Table 15 gives a summary of the results from all meta-analyses and all methods.

Table 13: Results from Year 3 Analysis

(I^2 = 88%)

Study	Both sound(+)	GI+, RB-	GI-, RB+	Both carious(-)	RR (95% CI)
Arrow 1995	378	28	3	3	0.194 (0.087, 0.431)
Kervanto-Seppala 2008	625	5	25	2	3.857 (1.767, 8.422)
Poulsen 2001	156	6	37	7	3.385 (1.978, 5.793)
Karlzen-Reuterv 1995	70	3	1	0	0.333 (0.035, 3.205)
Raadal 1996	122	0	10	0	11.000 (1.549, 78.093)
Rock 1996	105	1	21	3	6.000 (2.348, 15.333)
Baseggio 2010	473	29	99	27	2.250 (1.728, 2.930)
DSL Pooled					**2.058 (0.917, 4.617)**
MB Pooled					**2.022 (0.630, 5.963)**
NBR Pooled					**2.237 (1.789, 2.836)**
IBR Pooled					**2.164 (1.727, 2.733)**

Table 14: Results from Year 4+ Analysis

(I^2 = 0%)

Study	Both sound(+)	GI+, RB-	GI-, RB+	Both carious(-)	RR (95% CI)
Forss 1998	66	8	15	8	1.438 (0.881, 2.346)
Williams 1996	189	11	17	8	1.375 (0.791, 2.390)
DSL Pooled					**1.410 (0.977, 2.034)**
MB Pooled					**1.351 (0.507, 3.607)**
NBR Pooled					**1.435 (0.932, 2.223)**
IBR Pooled					**1.214 (0.891, 1.901)**

Table 15: Summary of all Meta-Analysis Results

		DSL	MB	NBR	IBR
Year 1	($I^2 = 0\%$)	No difference	No difference	No difference	No difference
Year 1*	($I^2 = 45\%$)	No difference	No difference	**RB Superior**	No difference
Year 2	($I^2 = 60\%$)	**RB Superior**	**RB Superior**	**RB Superior**	**RB Superior**
Year 3	($I^2 = 88\%$)	No difference	No difference	**RB Superior**	**RB Superior**
Year 4	($I^2 = 0\%$)	No difference	No difference	No difference	No difference

DSL and MB methods only found a difference (showed RB to be superior to GI) in Year 2. This was the year that combined data from ten different studies, the highest number of all follow-up years. As seen in simulations, all methods do the best in settings that combined higher numbers of studies, so this is not a surprising result in agreement by all methods. NBR method was able to show a difference in Years 1-3. From simulation results, we see that the Bayes-Rasch methods tend to show their strength in settings with higher heterogeneity among a lower number of studies, which is evident among these data sets. No method showed any difference in Year 4, which could be because there was in fact no difference to be found, or because there is a very small amount of data, combining only two studies, and thus low power to detect a difference.

1.6 Discussion & Conclusions

In this research we utilized a novel Rasch model representation of paired contingency tables to conduct their meta-analysis. The meta-analysis of paired tables was facilitated by a hierarchical Bayesian model which fuses information from different studies in such a way that contributions of treatments, studies, and individual participants are modeled in an explicit manner. Extensive simulations were conducted to compare the proposed methodology with existing methods. Simulations showed that the Rasch-type fusion of studies is competitive with existing methods, and often estimates the effects of treatments more precisely. In cases of a lower number of studies and higher heterogeneity this method can also perform well while often producing

tighter CI's than other methods, since the study effect parameter, β, accounts for heterogeneity between studies, so that it does not leak so heavily into the treatment effect, α. This is particularly important both in testing hypotheses of superiority, and also hypotheses of non-inferiority or equivalence. When testing hypotheses of superiority, a tighter CI will lessen the chance of a false negative because it is less likely that the CI will contain 1 when in fact the true RR does not equal 1. When testing a hypothesis of non-inferiority or equivalence, this is important because the entire CI must fit within pre-assigned equivalence bounds to prove equivalence.

The use of Bayesian methods is also beneficial when testing hypotheses of non-inferiority or equivalence because conclusions are not limited to a dichotomous rejection or failed rejection of the null hypothesis. Bayesian methods allow one to express results in probability statements regarding the competing hypotheses. This method is also able to handle zero event cell counts without adjustment, an advantage over the competitors. And finally, as was demonstrated in settings for the IBR setup, this method allows for selection of priors guided by the particular application, for more informed prior settings when information is available.

CHAPTER II

DIAGNOSTIC CLASSIFICATION OF DIGITAL MAMMOGRAMS BY WAVELET-BASED SPECTRAL TOOLS

2.1 Introduction

Breast cancer (BC) is one of the most common forms of cancer among women, claiming over 40,000 female lives in the US alone in the year 2007. But since its causes are still not fully understood, prevention is far from being a primary strategy in reducing this number. Early detection is still the best method for improving BC prognosis. Finding the cancer early also means less invasive options for both specific diagnosis and for treatment.

Mammography is currently the best method for early BC detection, but the interpretation of these images can be difficult. Signs of cancer are often missed, while suspicious findings often need to be clarified through additional procedures. If the amount of clear information obtained from the screening images can be increased, the confidence a physician has in approaching next steps could also be improved. In the end, this translates into improved prognoses while also reducing the number of unnecessary procedures or surgical operations.

The aim of this study is to develop and evaluate a wavelet-based data scaling estimator that is optimal for use in the classification of tissue in mammograms as cancerous or non-cancerous for diagnostic purposes. The specific demands for this estimator are that it be robust with respect to distributional assumptions on the data, and with respect to outlier levels in the frequency domain representation of the data.

The diagnostic use of information contained in the background tissue of images is

a novel concept, since most tools are focused on finding irregular shapes and calcifications. This technique allows for the use of information from the entire image—not only artifacts of interest.

The ambiguities in current diagnostic methods often result in additional procedures, extra costs, or missed cancers. With reasonable misclassification errors, this could be a promising new and informative indicator with potential for improving current screening techniques as an additional tool for physicians.

2.2 Background

When dealing with high frequency and irregular data, which is commonly found in real-life settings, the irregular behaviors of these complex structures are often difficult or impossible to quantify by standard modeling techniques. But a commonly occurring phenomenon, in both naturally occurring and human-made high-frequency data, is that of regular scaling. When data is transformed to the frequency domain and observations are inspected at different scales, there is a regular relationship between the behavior at each scale. Examples of this have been found in a variety of systems and processes including economics (stock market, exchange rate fluctuations), telecommunications (internet data), physics (hydrology, turbulence), geosciences (wind and rainfall patterns), and several applications in biology and medicine (DNA sequences, heart rate variability, auditory nerve-spike trains). Summaries of how data scale can be very informative and low-dimensional descriptors of otherwise difficult-to-quantify data structures.

The phenomenon of scaling has been demonstrated in many medical images, leading to the diagnostic use of tools capable of quantifying statistical similarity of data patterns at various scales. The particular application of this type of measure to breast cancer diagnostics is motivated by the perceived potential for high impact. Despite a reduction in the number of breast cancer cases, it still continues to be a major health

concern among women. Breast cancer is one of the most common forms of cancer among women in the United States, second only to non-melanoma skin cancer. The National Cancer Institute estimates that 1 in 8 women born today will be diagnosed with breast cancer during her lifetime [3]. A national objective has been set by the U.S Department of Health and Human Services to reduce the female breast cancer death rate from 22.9 (per 100,000 females) in 2007 down to 20.6 by the year 2020 – a 10% improvement [27]. One of the most important tools toward that goal is advanced precision of screening technologies. The causes of breast cancer are still unclear, meaning prevention is still far from being a primary solution. Early detection remains the best strategy for improving prognosis and also leads to less invasive options for both specific diagnosis and treatment.

But early detection still has its difficulties. Mammography is currently the best method for detecting a breast cancer early, before the malignant tissue is substantial enough to feel or cause symptoms. However, the radiological interpretation of mammogram images is a difficult task since the appearance of even normal tissue is highly variable and complex, and signs of early disease are often small or indistinct. Reading a mammogram image is a skill that physicians develop over time, and confidently stating whether findings are cancerous or not is often quite difficult. Suspicious findings are commonly clarified by follow-up images, ultrasound, or MRI. On the other hand, it has been estimated that $10-30\%$ of cancers which could have been detected are missed [26]. Thus, improving the accuracy of interpretation in mammographic screening is an important goal toward enhancing early detection and improving prognoses while also reducing the number of unnecessary procedures or surgical operations.

This use of scaling estimators in breast cancer diagnostics will bring novel information into use since this modality will include information contained in the background

tissue of images. Most of the references found in literature dealing with breast cancer detection methods are based on microcalcifications [49], [34], [22], [14]. Only recently has information contained in the background come into consideration [36]. This classifying measure based on background tissue would be a new tool for use in combination with existing clinical diagnostic tools, thus improving the power of non-invasive diagnostic techniques.

The standard measure of regular scaling is the Hurst exponent. This measure can also be connected to measures of long memory, dimension, and fractality in signals and images and is viewed as an informative summary. Many techniques for estimating the Hurst exponent exist, and assessing the accuracy of these estimations can be complicated. Estimators can have strengths in certain settings and fall short in others, depending on the nature of the data and the task intended for the use of the estimate. The focus of this work is particularly on dealing with data that may violate most underlying assumptions regarding the distribution of the variances at each level, which may affect scaling estimation, and thus its accuracy and usefulness.

All measures used in this research are based on wavelet theory, which continues to grow in its importance for image processing techniques [11], [36], [50]. In this context wavelet transforms are powerful tools because of their innate ability to model statistical similarity of signals and images at different scales. We will next briefly introduce wavelets, wavelet-based spectra, and the concept of scaling.

2.2.1 The Discrete Wavelet Transform

Among the various techniques developed to make complex data information more accessible and manageable for analysis, wavelet transforms have been shown to be particularly useful. The wavelet transform decomposes a signal into many different scales or frequency bands. Then the innate regularity of a complex data structure can be quantified through summary measures obtained in the wavelet domain, resulting

in informative and low-dimensional descriptors. We now give a brief overview of the discrete wavelet transform, and its extension into the 2-dimensional case.

The discrete wavelet transform (DWT) of a function $\{X(t), t \in \mathbb{Z}\}$ represents this function in terms of shifted and dilated versions of a wavelet (or *mother*) function $\psi(t)$ and shifted versions of a scaling (or *father*) function $\phi(t)$. For specific choices of the scaling functions and wavelets, an orthonormal basis can be formed from the atoms

$$\begin{aligned} \psi_{j,k}(t) &= 2^{j/2}\ \psi(2^j t - k) \\ \phi_{j,k}(t) &= 2^{j/2}\ \phi(2^j t - k), \quad \forall j, k. \end{aligned}$$

Then $X(t)$ can be represented by wavelets as

$$X(t) = \sum_k c_{J_0,k}\phi_{J_0,k}(t) + \sum_{j=J_0}^{\infty}\sum_k d_{j,k}\psi_{j,k}(t), \tag{6}$$

where

$$\begin{aligned} d_{j,k} &= \int X(t)\ \psi_{j,k}(t)dt \quad \text{and} \\ c_{j,k} &= \int X(t)\ \phi_{j,k}(t)dt \end{aligned}$$

are detail and scaling coefficients, respectively. Here, J_0 is the coarsest scale or lowest resolution of the transform, and larger values of j correspond to higher resolutions. For a detailed introduction to wavelet theory, the reader is referred to [9], [25], or [48]. The detail coefficients $d_{j,k}$ in (16) are what we eventually use to assess the "energy" at each level and thus the energy scaling between levels, which will be described further in section 2.2.2.

Data sets can be easily and quickly transformed by the DWT through coding the data by the wavelet coefficients. When dealing with functions that are given by their sampled values, it is customary to set the sampled values to be "smooth" coefficients at the highest resolution level $j = J$. The subsequent "detail" levels obtained through DWT are denoted by d_j, corresponding to $j = J-1, J-2, \ldots, J_0$.

Many signals arising in practical applications (astronomy, geophysics, economics, etc.) are multidimensional. The DWT is easily generalized to the multidimensional case. Since our application of interest uses the wavelet transforms of medical images, the generalization shown here is for the 2-dimensional case. The 2-dimensional wavelet basis functions are constructed via translations and dilations of a tensor product of univariate wavelet and scaling functions:

$$\begin{aligned} \phi(t_1,t_2) &= \phi(t_1)\phi(t_2) \\ \psi^h(t_1,t_2) &= \phi(t_1)\psi(t_2) \\ \psi^v(t_1,t_2) &= \psi(t_1)\phi(t_2) \\ \psi^d(t_1,t_2) &= \psi(t_1)\psi(t_2). \end{aligned} \tag{7}$$

The symbols h, v, d in (18) stand for horizontal, vertical and diagonal directions, respectively, since the atoms capture image features in the corresponding directions.

Consider the wavelet atoms

$$\phi_{j,\mathbf{k}}(\mathbf{t}) = 2^j \ \phi(2^j t_1 - k_1, 2^j t_2 - k_2) \tag{8}$$

$$\psi^i_{j,\mathbf{k}}(\mathbf{t}) = 2^j \ \psi^i(2^j t_1 - k_1, 2^j t_2 - k_2), \tag{9}$$

for $i = h, v, d$, where $j \in \mathbb{Z}$, $\mathbf{t} = (t_1, t_2) \in \mathbb{R}^2$, and $\mathbf{k} = (k_1, k_2) \in \mathbb{Z}^2$. Then, any function $X \in \mathcal{L}_2(\mathbb{R}^2)$ can be represented as

$$X(\mathbf{t}) = \sum_{\mathbf{k}} c_{J_0\mathbf{k}} \phi_{J_0,\mathbf{k}}(\mathbf{t}) + \sum_{j \geq J_0} \sum_{\mathbf{k}} \sum_i d^i_{j,\mathbf{k}} \psi^i_{j,\mathbf{k}}(\mathbf{t}),$$

where the detail coefficients are given by

$$d^i_{j,\mathbf{k}} = 2^j \int X(\mathbf{t}) \ \psi^i(2^j \mathbf{t} - \mathbf{k}) d\mathbf{t}.$$

Since this transformation is linear, a fast DWT can be achieved by matrix multiplication, similar to a Fast Fourier transform. See [48] (pp 115-116, 153-159) for the construction of these wavelet matrices, both in the 1-dimensional and 2-dimensional cases.

2.2.2 Scaling and Wavelet-Based Spectra

The methodology used to analyze scaling is based on the analysis of autocovariances, or correlations between observations as a function of the time separation between them. The variance of a signal in its original domain corresponds to its "energy" in the frequency domain. The term "energy" is an informal name for the squared coefficients in frequency-domain representations of signals and images such as (16). Thus, the correlation between time-separated observations in the original domain corresponds to the scaling of energy in the frequency/scale domains. But the frequency-domain representation allows for more concise means of describing the distribution of that energy (or variance) over a range of frequencies. This introduces the idea of energy spectra as a tool for characterizing the scaling behavior of data. We now describe how this spectra can be represented using wavelet-based methods, and then extend these methods into the 2-dimensional case.

The Hurst exponent ($H \in [0,1]$) is the standard measure of regular scaling, and the key descriptor that it is our eventual goal to estimate and utilize. A stochastic process $\{X(t),\ t \in \mathbb{R}\}$ is self-similar with scaling exponent H if, for any $\lambda \in \mathbb{R}^+$,

$$X(\lambda t) \stackrel{d}{=} \lambda^H X(t), \tag{10}$$

where $\stackrel{d}{=}$ denotes equality of all joint finite-dimensional distributions, throughout this paper. For a fixed level j, it can be shown (*Flandrin late 1980's reference*) that under $\mathcal{L}_2$ normalization,

$$d_{jk} \stackrel{d}{=} 2^{-j(H+1/2)} d_{0,k}.$$

If, in addition, $X(t)$ has stationary increments, then $E(d_{0k}) = 0$ and $E\left(d_{0k}^2\right) = E\left(d_{00}^2\right)$. Therefore,

$$E\left(d_{jk}^2\right) \propto 2^{-j(2H+1)}. \tag{11}$$

By taking logarithms on both sides of (11), we obtain the basis for estimating H, the

wavelet spectrum, which is defined as

$$S(j) = \log_2\left(Ed_{jk}^2\right) = -(2H+1)j + C. \tag{12}$$

For a more rigorous description of the wavelet spectra in one dimension, see [47].

Nicolis et al. [36] generalized the definition of traditional wavelet spectra to 2-dimensions. In this generalization, three different hierarchies $i = \{h, v, d\}$ (horizontal, vertical and diagonal directions) constitute the detail spaces, as in (8). The natural definition of the wavelet spectra then involves power spectrum corresponding to each of those three hierarchies. The expected value of the detail coefficients of a random process with stationary zero-mean Gaussian increments, in 2-dimensions, will verify that

$$E\left[\left|d_{j,\mathbf{k}}^i\right|^2\right] = c_i 2^{-(2H+2)j}, \tag{13}$$

for some constant c_i depending on the wavelets ψ^i in (18), but not on the scale j. By taking logarithms on both sides of the equation (13), we obtain the 2-dimensional wavelet-based spectra

$$S^i(j) = \log_2 E\left[\left|d_{j,\mathbf{k}}^i\right|^2\right] = -(2H+2)j + C_i, \tag{14}$$

by which the H for 2-dimensions is estimated.

While (12) and (14) give the basis for estimating H in 1-dimension and 2-dimesions, respectively, specific methods for this estimation continue to be investigated and improved upon. This is the motivation behind the current investigation, to find an optimal estimator for the context of classification of tissue in mammogram images.

2.3 Scaling Estimators

Two current and well-known scaling estimation methods will be included in this analysis for the purpose of comparing their performance in the classification task with that of newly introduced scaling estimators. These estimators from current literature are introduced next, followed by a novel estimator derived with the goal in

mind of robustness in the context of data that violate distributional assumptions at different levels.

2.3.1 Ordinary least squares regression (OLS)

Using ordinary least squares regression (OLS), directional Hurst exponents (diagonal (H_d), horizontal (H_h), and vertical (H_v)) can be estimated from the slopes of the linear equations in (14). The empirical counterpart to this is an OLS regression defined on pairs

$$\left(j, \log_2 \overline{\left|d^i_{j,\mathbf{k}}\right|^2}\right), \quad i = h, v, d. \tag{15}$$

where $\overline{\left|d^i_{j,\mathbf{k}}\right|^2}$ is an empirical counterpart of $E\left[\left|d^i_{j,\mathbf{k}}\right|^2\right]$. The slope of the regression would estimate H, i.e., $H = -(slope + 2)/2$. This method is in prevalent use for both estimation of H's and classification by H's.

In the presence of normally distributed errors and homoscedasticity, OLS estimation is typically the method of choice. OLS assumes the errors of prediction (deviations from the point $\log_2 \overline{\left|d^i_{j,\mathbf{k}}\right|^2}$) are normally distributed, with a common error variance at all levels. These assumptions are frequently untenable in practice, and violations of these assumptions in the data can heavily influence estimates using OLS regression [35].

2.3.2 Abry-Veitch weighted regression (AV)

Veitch and Abry [47] improved the OLS method to a weighted linear regression, to solve the issue of heteroscedasticity. Since the variances of the $\log_2 \overline{\left|d^i_{j,\mathbf{k}}\right|^2}$ can vary with j, this method weights each level by the inverse of the variance of that level, where

$$\mathbb{V}\mathbf{ar}\left(\log_2 \overline{\left|d^i_{j,\mathbf{k}}\right|^2}\right) \simeq \frac{2}{n_j \ln 2}.$$

H's are then estimated from the slopes of these weighted linear regressions using weights

$$w_j \propto \frac{n_j \ln 2}{2}.$$

We will denote this estimation method as AV.

Although AV accounts for the differences in variances at each level, this method still assumes that the errors are normally distributed at each level. But outlier levels in the observed data (seen as a bump or a hockey stick effect in the wavelet spectra) are common in real-world data, which are often a manifestation of a violation of this assumption. This can still influence estimates using AV regression.

2.3.3 Theil-type weighted regression (TT)

An estimation method is desired that performs well without regard to the nature of the distribution of errors, allowing for estimation that is robust even under non-ideal conditions. Theil [46] proposed a robust measure for the slope of a regression line passing through all sample data points by using information from all possible pairwise slopes between each pair combination of points, weighted in such a way as to reduce the undue influence that outliers can have on estimates. This type of estimate is robust with respect to possible outlier levels and free of distributional assumptions.

A new Theil-type estimator is introduced to the context of our current task. The slopes of the linear equations in (14) will be assessed as a weighted average of all pairwise slopes, $s_{i,j}$ between levels i and j. There have been several proposed weighting schemes for this type of estimator [20], [42], [39], [8]. A weight specifically designed for use in our context is derived with the rationale that each pairwise slope is weighted by an inverse of the variance of the estimated slope for that pair. We will denote this estimation method as TT.

Theorem. *Let* $(j, e_j), j = J_0, ..., J-2, J-1$ *be pairs in which* j *is the multiresolution level and* $e_j = \log_2 \overline{|d_{j,\cdot}|^2}$ *is the log of the average energy in the* j*th level.*

The optimal robust (Theil-type) slope b in the linear regression

$$e_j = b \cdot j + a$$

is the weighted average of pairwise slopes

$$s_{ij} = \frac{e_j - e_i}{j - i}$$

between levels i and j where $i < j$, with weights

$$w_{ij} \propto (i-j)^2 \times HA\left(2^{2i}, 2^{2j}\right),$$

where HA is the harmonic average.

That is, the estimate of the overall slope is

$$b = \sum_{i,j} w_{ij} s_{ij} / \sum_{i,j} w_{ij}.$$

Proof of Theorem. Let $d_j = d_{j\mathbf{k}}$ be an arbitrary (wrt $\mathbf{k}$) wavelet coefficient from the jth level of the decomposition of the m-dimensional fractional Brownian motion $B_H(\omega, \mathbf{t}), \mathbf{t} \in \mathbb{R}^m$,

$$d_j = \int_{\mathbb{R}^m} B_H(\omega, \mathbf{t}) \psi^*_{j\mathbf{k}}(\mathbf{t}) d\mathbf{t},$$

for some fixed $\mathbf{k} = (k_1, \ldots, k_m)$. Here $\psi^*_{j\mathbf{k}}(\mathbf{t}) = \prod_{i=1}^m \psi^*_{jk_i}(t_i)$ where ψ^* is either ψ or ϕ, but in the product there is at least one ψ. It is well known that

$$d_j \stackrel{d}{=} 2^{-(H+m/2)j}\ d_0,$$

where d_0 is a coefficient from the level $j = 0$, and $\stackrel{d}{=}$ means equality in distributions [36].

Coefficient d_j is a random variable with

$$\mathbb{E}\, d_j = 0, \quad \mathbb{V}\mathbf{ar}\, d_j = \mathbb{E}\, d_j^2 = 2^{-(2H+m)j}\ \sigma^2,$$

where $\sigma^2 = \mathbb{V}\mathbf{ar}\, d_0$.

The fBm $B_H(\omega, \mathbf{t})$ is a Gaussian m-dimensional field, thus

$$d_j \sim \mathcal{N}(0, 2^{-(2H+m)j}\sigma^2).$$

The coefficients d_j within the level j are typically considered approximately independent. The covariance decays with the distance between the coefficients and the rate of decay depends on H and N - the number of vanishing moments for the wavelet ψ. Flandrin [15], Tewfik & Kim showed that for $m = 1$,

$$\mathbb{E}\, d_{jk_1} d_{jk_2} \le C|k_1 - k_2|^{2(H-N)},$$

where C depends on j. Although, for small $|k_1 - k_2|$ this covariance may not be small, it decays to 0 as long as $N > H$. To ensure short memory of $d_{jk},\ k \in \mathbb{Z}$, the convergence of

$$\sum_k \mathbb{E}\,|d_{jk_1} d_{jk_2}|$$

is needed, for which it is required that $N > H + 1/2$.

The rescaled "energy"

$$\frac{2^{(2H+m)j}}{\sigma^2} d_j^2 \sim \chi_1^2$$

while, assuming the independence of d_{jk}'s,

$$\frac{2^{(2H+m)j}}{\sigma^2} \sum_{\mathbf{k} \in j\text{th level}} d_{j\mathbf{k}}^2 = \frac{2^{(2H+2m)j}}{\sigma^2}\ \overline{d_j^2}$$

has $\chi^2_{2^{mj}}$ distribution. Here, $\overline{d_j^2}$ is the average energy in jth level.

Thus,

$$\overline{d_j^2} \stackrel{d}{=} 2^{-(2H+2m)j}\sigma^2 \chi^2_{2^{mj}}.$$

From this,

$$\mathbb{E}\,\overline{d_j^2} = \sigma^2 2^{-(2H+2m)j}\mathbb{E}\,\chi^2_{2^{mj}} = 2^{-(2H+m)j}\sigma^2.$$

and

$$\mathbb{V}\mathbf{ar}\,\overline{d_j^2} = \sigma^4 2^{-(4H+4m)j} \times 2 \cdot 2^{mj} = 2^{-4Hj-3mj+1}\sigma^4.$$

Recall that if X has $\mathbb{E}\,X$ and $\mathbb{V}\mathbf{ar}\,X$ finite and φ is a function with finite second derivative at $\mathbb{E}\,X$, then

$$\mathbb{E}\,\varphi(X) \approx \varphi(\mathbb{E}\,X) + \frac{1}{2}\varphi''(\mathbb{E}\,X) \cdot \mathbb{V}\mathbf{ar}\,X.$$

and

$$\mathbb{V}\mathbf{ar}\,\varphi(X) \approx (\varphi'(\mathbb{E}\,X))^2 \mathbb{V}\mathbf{ar}\,X.$$

When φ is logarithm for base 2, then

$$\begin{aligned}
\mathbb{E}\,\log_2 \overline{d_j^2} &= \log_2 \mathbb{E}\,\overline{d_j^2} + \frac{1}{2\log 2}\left(-\frac{\mathbb{V}\mathbf{ar}\,\overline{d_j^2}}{\left(\mathbb{E}\,\overline{d_j^2}\right)^2}\right) \\
&= \log_2\left(2^{-(2H+m)j}\sigma^2\right) - \frac{1}{2\log 2}2^{-mj+1} \\
&= -(2H+m)j - \frac{1}{2^{mj}\log 2} + \log_2 \sigma^2.
\end{aligned}$$

Note that $-\frac{1}{2^{mj}\log 2}$ is the Abry-Veitch bias term, and it is free of H and σ^2. This bias is a second order approximation. Veitch and Abry show that the exact bias involves digamma function Ψ, and in this context is

$$\frac{\Psi(2^{mj-1})}{\log 2} - \log\left(2^{mj-1}\right).$$

Also,

$$\begin{aligned}
\mathbb{V}\mathbf{ar}\,\log_2 \overline{d_j^2} &= \left(\frac{1}{\sigma^2\ 2^{-(2H+m)j}\ \log 2}\right)^2 \\
&\qquad \times \sigma^4 \cdot 2^{-4Hj-3mj+1} \\
&= \frac{2}{2^{mj}(\log 2)^2}.
\end{aligned}$$

Finally,

$$\mathbb{V}\mathbf{ar}\left(\frac{\log_2 \overline{d_j^2} - \log_2 \overline{d_i^2}}{j-i}\right) = \frac{2}{(\log 2)^2} \cdot \frac{1/2^{mj} + 1/2^{mi}}{(j-i)^2}.$$

Since weights w_{ij} are inverse-proportional to the variance, then

$$w_{ij} \propto (i-j)^2 \times HA(2^{mi}, 2^{mj}).$$

where HA is the harmonic average.

□

2.4 Results

2.4.1 Performance in 1-Dimensional Estimation of H

2.4.1.1 Estimation of H in simulated data of known H

We first assess how well these estimators perform in estimating H in simulated data where H is known. Using $MATLAB©$ software, we simulated fractional Brownian motion for a range of known H's (0.3-0.7), and performed the DWT using four different wavelet filters (Haar, Daubechies 4 tap, Symmlet 8 tap, Coiflet 6 tap) to obtain the wavelet spectrum. The estimators were then used to calculate H. This process was done with 500 repetitions at each setting, so the reported prediction errors for each estimator are averaged over 500 runs. Table 16 shows the results of these simulations. Cells highlighted in green show those with the lowest mean-square-error (MSE), which takes into account both the bias of the estimates as well as the variance. The estimates that are underlined are those where the bias was lowest. Across all H's, and for all wavelet filters, estimator TT performed the best with regard to both MSE and bias alone.

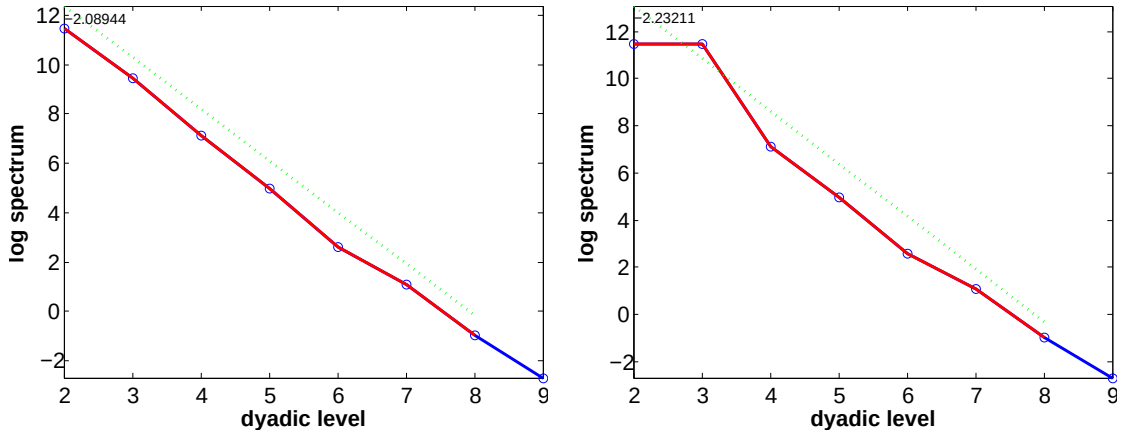
2.4.1.2 Estimation of H in simulated data of known H, contaminated

We next assessed how well these estimators would perform in estimating H in simulated data of known H, but this time with a contamination introduced. This is to simulate more real-world type data where spectrum cannot be dependable to follow a very clean slope, but may sometimes have some sort of anomaly present. In real life, there can be instability in low levels of details in decomposition, especially at very

Table 16: Results from estimations of H in simulated 1-dimensional data with known H. Cells in green show those estimates with lowest MSE. Underlined estimates are those where the bias was lowest.

		Haar				Daubechies 4				Coiflet 6				Symmlet 8			
H		OLS	AV	TT		OLS	AV	TT		OLS	AV	TT		OLS	AV	TT	
0.3	H	0.24	0.22	0.25	0.23	0.28	0.25	0.28	0.27	0.28	0.24	0.28	0.27	0.26	0.22	0.26	0.25
	MSE	0.008	0.009	0.005	0.007	0.005	0.005	0.003	0.004	0.005	0.005	0.001	0.004	0.006	0.007	0.004	0.006
0.4	H	0.36	0.34	0.37	0.35	0.39	0.37	0.40	0.38	0.39	0.34	0.38	0.37	0.36	0.34	0.37	0.36
	MSE	0.006	0.005	0.004	0.005	0.005	0.003	0.003	0.004	0.005	0.005	0.003	0.004	0.006	0.005	0.003	0.005
0.5	H	0.45	0.45	0.47	0.46	0.47	0.46	0.49	0.47	0.48	0.41	0.46	0.45	0.48	0.45	0.49	0.47
	MSE	0.006	0.005	0.003	0.005	0.005	0.003	0.002	0.004	0.004	0.011	0.004	0.006	0.005	0.004	0.003	0.004
0.6	H	0.55	0.55	0.57	0.56	0.53	0.52	0.54	0.53	0.54	0.43	0.50	0.49	0.56	0.51	0.55	0.54
	MSE	0.006	0.005	0.004	0.005	0.009	0.009	0.003	0.007	0.008	0.034	0.014	0.019	0.006	0.011	0.005	0.007
0.7	H	0.65	0.65	0.67	0.66	0.56	0.54	0.57	0.56	0.56	0.43	0.51	0.50	0.62	0.57	0.62	0.61
	MSE	0.006	0.005	0.003	0.005	0.027	0.030	0.023	0.027	0.024	0.085	0.044	0.051	0.011	0.024	0.011	0.015

low levels where there are very few wavelet coefficients. At each run, a "bump" was introduced into the spectra at the third level. This was accomplished by replacing the log-energy value at the third level by repeating that from the second level. An example of this contamination can be seen in Figure 3.



(a) (b)

Figure 3: Example of introduction of contamination to log-energy spectrum. (a) Original uncontaminated spectrum, (b) Spectrum with contamination introduced in the third level

The process was exactly the same as the previous section, simulating a range of H's, using four different wavelet filters, and performing 500 repetitions for each

setting. Results of these simulations are shown in Table 17. Again, cells highlighted in green show those with the lowest MSE, while estimates that are underlined are those where the bias was lowest. Overall, estimator TT again performed the best, with the lowest MSE in 10/20 cases, and with lowest bias in 10/20 cases. AV performed second-best with regard to MSE (9/20 cases), while OLS performed second-best with regard to bias alone (7/20 cases). It is important to note here that both AV and OLS assume Normal models, while TT is free of any distributional assumptions. The data simulated here are originally produced from a Gaussian process, so it is not surprising that AV and OLS will still perform fairly well since the data fit their distributional assumptions. However, in a real-world setting where data really may not follow the assumed Normality, it is expected that TT would further out-perform the others.

Table 17: Results from estimations of H in simulated 1-dimensional data with known H, but with contamination introduced in the third level. Cells in green show those estimates with lowest MSE. Underlined estimates are those where the bias was lowest.

		Haar				Daubechies 4				Coiflet 6				Symmlet 8			
H		OLS	AV	TT		OLS	AV	TT		OLS	AV	TT		OLS	AV	TT	
0.3	H	0.29	0.25	0.29	0.28	0.30	0.26	0.31	0.29	0.32	0.25	0.31	0.30	0.32	0.27	0.32	0.30
	MSE	0.013	0.006	0.005	0.008	0.017	0.005	0.005	0.009	0.021	0.005	0.005	0.010	0.021	0.005	0.007	0.011
0.4	H	0.41	0.37	0.42	0.40	0.40	0.37	0.42	0.40	0.42	0.36	0.42	0.40	0.44	0.38	0.44	0.42
	MSE	0.010	0.004	0.005	0.006	0.021	0.004	0.006	0.010	0.026	0.005	0.006	0.012	0.020	0.003	0.007	0.010
0.5	H	0.53	0.49	0.53	0.52	0.49	0.47	0.51	0.49	0.52	0.43	0.51	0.49	0.52	0.47	0.53	0.51
	MSE	0.012	0.003	0.005	0.007	0.016	0.004	0.004	0.008	0.018	0.007	0.004	0.010	0.019	0.003	0.005	0.009
0.6	H	0.64	0.60	0.64	0.63	0.55	0.53	0.57	0.55	0.59	0.47	0.56	0.54	0.59	0.52	0.59	0.57
	MSE	0.012	0.003	0.006	0.007	0.020	0.008	0.005	0.011	0.016	0.023	0.006	0.015	0.017	0.009	0.004	0.010
0.7	H	0.74	0.71	0.75	0.73	0.58	0.55	0.60	0.58	0.64	0.49	0.59	0.57	0.63	0.54	0.62	0.60
	MSE	0.012	0.004	0.008	0.008	0.034	0.027	0.017	0.026	0.017	0.057	0.019	0.031	0.021	0.031	0.011	0.021

2.4.2 Description of Mammography Data

A collection of digitized mammograms for analysis was obtained from the University of South Florida's Digital Database for Screening Mammography (DDSM) [28]. The DDSM is described in detail in [18]. Images from this database containing suspicious areas are accompanied by pixel-level "ground truth" information relating locations of

suspicious regions to what was assessed and verified through biopsy. We selected 105 normal cases (controls) from volumes normal-01, and 72 cancer cases from volumes cancer-01 and cancer-02. Each case study contains four mammograms (two for each breast: the craniocaudal (CC) and mediolateral oblique (MLO) projections) from a screening exam. We will consider only the CC projections, using the right breast image for all normal controls, and the cancerous breast (right or left) image for cancer cases. A subimage of size 1024 × 1024 was taken from each mammogram image for analysis. An example of an image and its subimage is provided in Fig. 4.

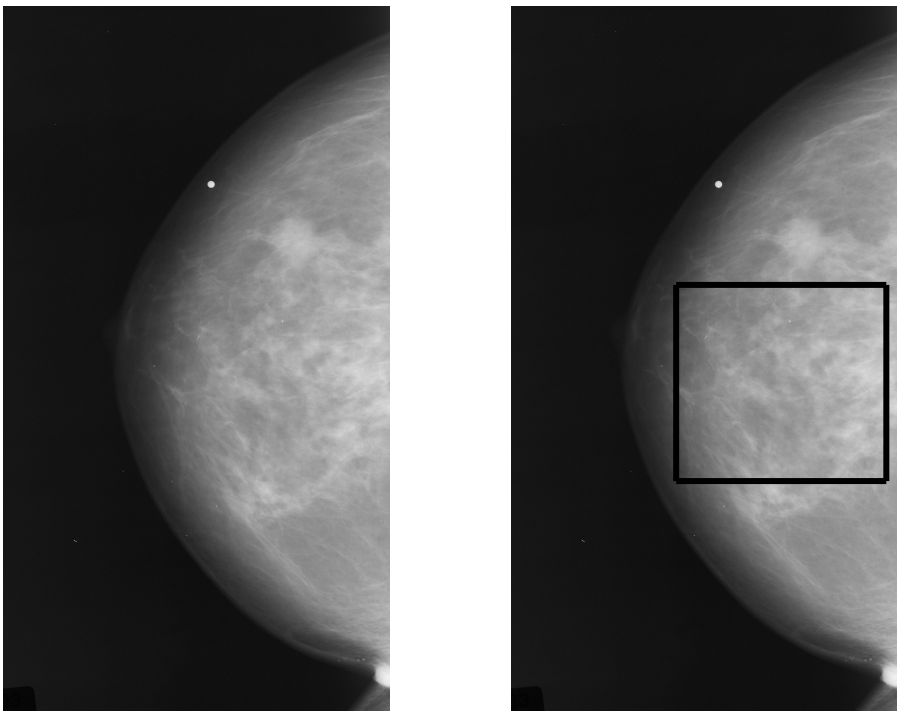


Figure 4: *Left panel*: right CC mammogram corresponding to a cancer case. *Right panel*: subimage of size 1024×1024 to be considered for the analysis.

2.4.3 Estimation of Scaling

For every subimage, we performed the DWT using four different wavelet filters (Haar, Daubechies 4 tap, Symmlet 8 tap, Coiflet 6 tap), ensuring the filter choice does not cause results to favor any estimator over the others. We also tried three different

level ranges: 2 to 6, 2 to 8, and 5 to 8. Detailed results are provided here for only transforms using Daubechies 4 (since this basis is the most local), and for the slopes that involve levels 5 to 8.

After each transform, we used the estimation methods described above (OLS, AV, and TT) to compute the estimated directional Hurst exponents (H_d, H_h, and H_v).

2.4.4 Classification

Multiple classification methods were used for each individual estimator, to inform the tradeoffs between model simplicity versus power:

- Binary logistic regression was performed using each individual directional Hurst exponent (H_d), (H_h), and (H_v); each paired combination (H_d,H_h),(H_d,H_v), and (H_h,H_v); and the combination of all three (H_d,H_h,H_v).

- Both linear and quadratic classification methods were implemented using pair combinations (H_d, H_h), (H_d, H_v), and (H_h, H_v).

In each case, we randomly selected 66% of the data as a training set to fit the classifier and used the remaining 34% of the data to test performance. The random selection of training and testing data was repeated 10,000 times, so the reported prediction errors for each estimator are averaged over 10,000 runs. Performance of each estimator was then compared in terms of sensitivity, specificity, and overall misclassification rate.

Binary logistic regression was first performed for each estimator, using each individual scaling estimate alone (H_d, H_h, and H_v). The binary logistic regression including only H_d as the predictor was the most parsimonious classifying model that still gave useful correct classification rates. The estimated density of H_d's obtained using the AV estimator are shown in Figure 5. Figure 6 is the logistic regression curve (in red) fitted over scores $b_0 + b_1 H_d$. Dots represent cancer cases at level 1, and controls at level 0.

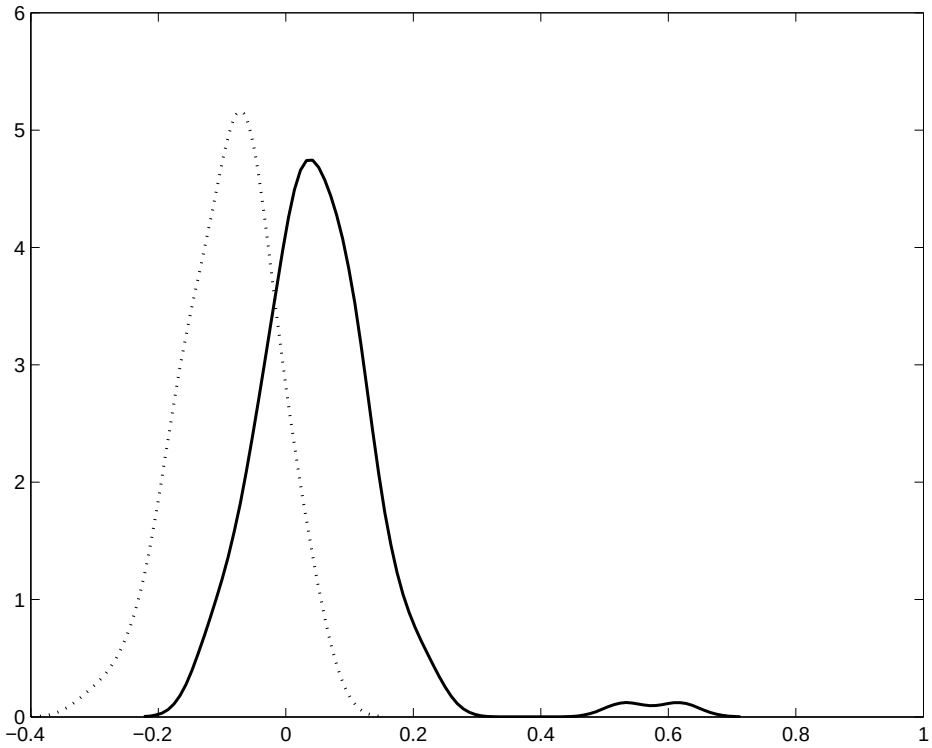


Figure 5: Estimated density of H_d obtained from 105 controls (*solid line*) and 72 cancer cases (*dotted line*). The estimated H's are empirical and flat spectra can cause H to be negative.

Figure 7 shows a ROC curve of H_d (by AV) in differentiating between controls and cancer cases. The diagonal line represents a test with a sensitivity of 50% and a specificity of 50%. This shows the ROC curve lying significantly to the left of the diagonal, where the combination of sensitivity and specificity are highest. The area under the ROC curve, which is proportional to the diagnostic accuracy of the test, is 0.8820. The most distant point from the diagonal (maximum Youden index), which is typically an acceptable compromise between sensitivity and specificity, in this case gives a sensitivity of 84.7% and specificity of 79%.

Table 18 summarizes the results of the classification based only on H_d, for each estimation method. The first column provides the area under the ROC curve (AUC), while the last three columns provide $1 - Sensitivity$, $1 - Specificity$, and Error (misclassification) rate achieved. The best classification rates in this case were achieved

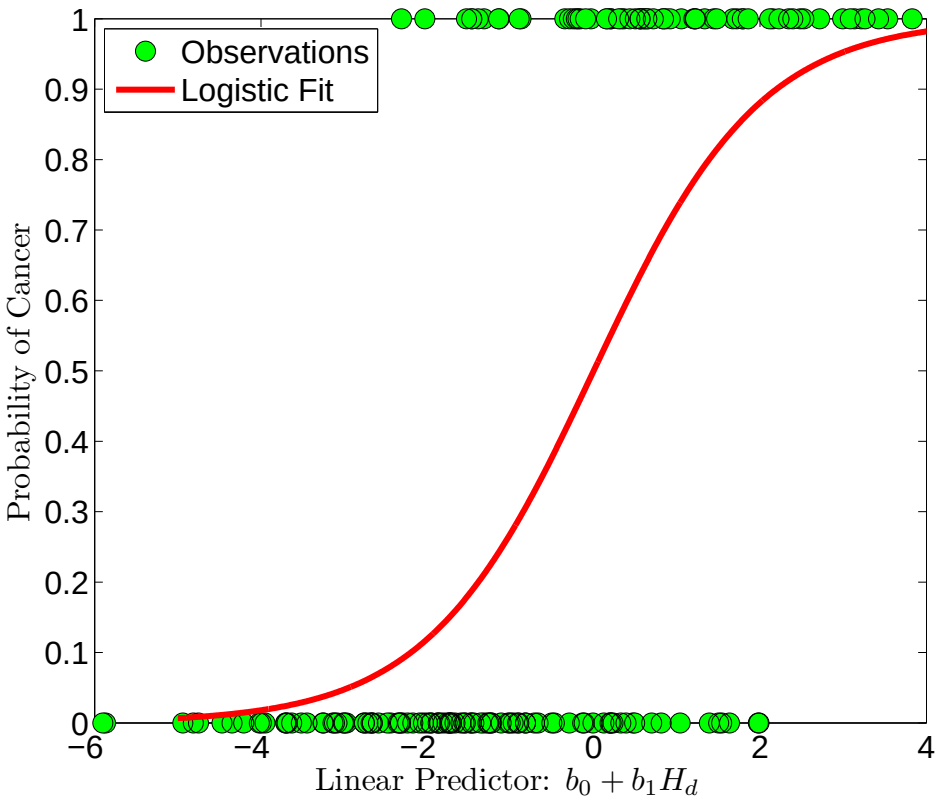


Figure 6: Logistic regression: $\text{logit}(p) = -0.8927 - 22.7722 \cdot H_d$, where H_d is the Abry-Veitch estimator.

with the AV estimator, where the classification error was only 20.11%. TT estimator gave the next best results, with 27.96% error. OLS was the worst performer, with 48.42% error.

Table 18: Results of classification by logistic regression using H_d.

Method	**AUC**	**1-Se**	**1-Sp**	**Error**
OLS	0.5906	0.4658	0.4966	0.4842
AV	0.8821	0.1790	0.2161	0.2011
TT	0.8072	0.2580	0.2942	0.2796

Binary logistic regression was then performed using paired directional H's. Table 19 summarizes the results of the classification based on the pair combination (H_d, H_h), for each estimation method. The best classification rates were again achieved with the AV estimator, where the classification error was only 12.11%. TT estimator gave the next best results, with 16.09% error. OLS was again the worst performer,

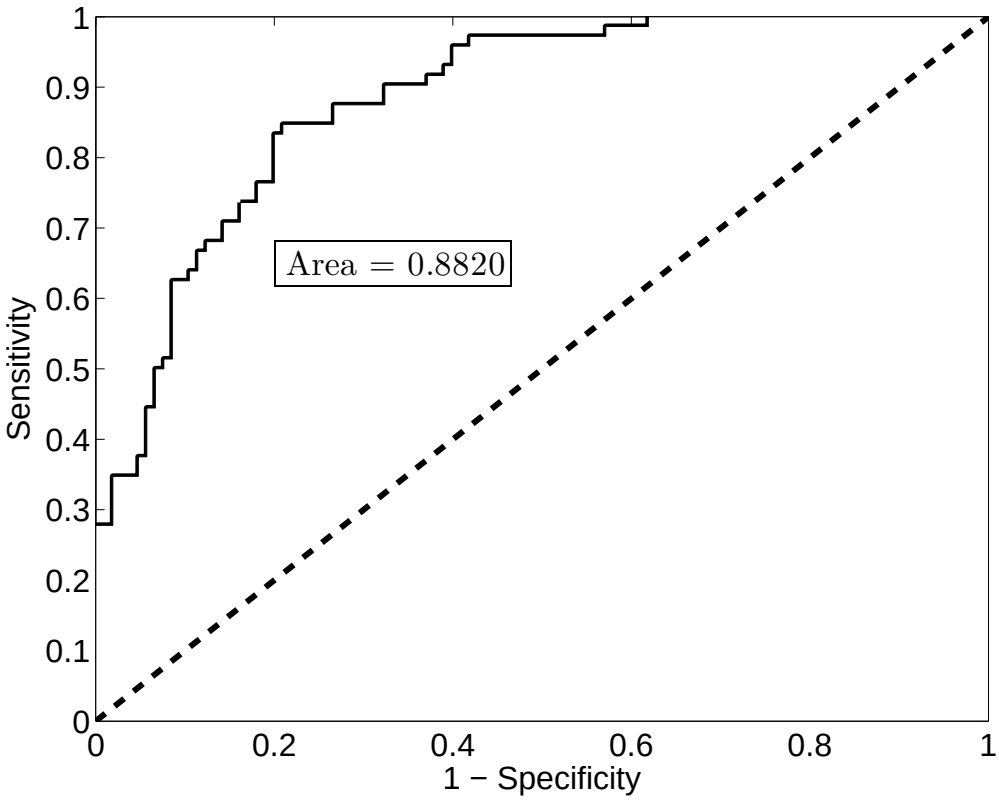


Figure 7: ROC curve for the logistic regression: $\text{logit}(p) = -0.8927 - 22.7722 \cdot H_d$, where the most distant point from the diagonal (Youden index) is achieved at $H_d = -0.0240$ for which Sensitivity was 84.7% and Specificity 79%.

with 34.69% error. The respective overall performances in this case were very similar to the pair combination, (H_d, H_v). Results from the performance of pair (H_h, H_v) were not comparable to those of other combinations and were thus dropped from consideration.

Table 19: Results of classification by logistic regression using (H_d, H_h).

Method	**AUC**	**1-Se**	**1-Sp**	**Error**
OLS	0.7360	0.1914	0.4521	0.3469
AV	0.9451	0.1396	0.1086	0.1211
TT	0.9099	0.1883	0.1424	0.1609

The final binary logistic regression in the preliminary analysis was performed for each estimator using the combination of all three directional H's (H_d, H_h, and H_v).

Table 20 summarizes these results. The respective order of performance is the same in this case, with AV resulting in 12.38% error, TT with 14.55%, and OLS with 32.05%.

Table 20: Results of classification by logistic regression using (H_d, H_v, H_h).

Method	AUC	1-Se	1-Sp	Error
OLS	0.7610	0.2199	0.3886	0.3205
AV	0.9560	0.1246	0.1233	0.1238
TT	0.9242	0.1572	0.1376	0.1455

Next, both linear and quadratic classification methods were then implemented using pair combinations $(H_d, H_h), (H_v, H_h), and (H_d, H_v)$. In both linear and quadratic cases, classifiers based on (H_d, H_v) were comparable in performance to those based on (H_d, H_h). But the remaining combination (H_h, H_v) again gave suboptimal results and was thus dropped from consideration.

Figure 8 shows a scatter plot of cases plotted by H_h versus H_d, illustrating the differentiation between controls and cancer cases. Table 21 summarizes the results of linear and quadratic classifications based on pairs (H_d, H_h), for each estimation method. The performance of linear and quadratic classifiers were comparable to each other, as well as to that of classification using binary logistic regression. In the results based on pair (H_d, H_h), the best classification rates were again achieved with the AV estimator, with 11.3% error in the linear case and 12.43% error for quadratic. TT again was ranked next in performance, with 16.78% and 17.91% errors. OLS was again the worst performer, still resulting in over a third of cases being misclassified.

Table 21: Results of linear and quadratic classification based on pair (H_d, H_h).

	Method	1-Se	1- Sp	Error
Linear	OLS	0.3301	0.3430	0.3377
	AV	0.1152	0.1114	0.1130
	TT	0.1496	0.1802	0.1678
Quadratic	OLS	0.2477	0.4144	0.3466
	AV	0.1282	0.1216	0.1243
	TT	0.1511	0.1982	0.1791

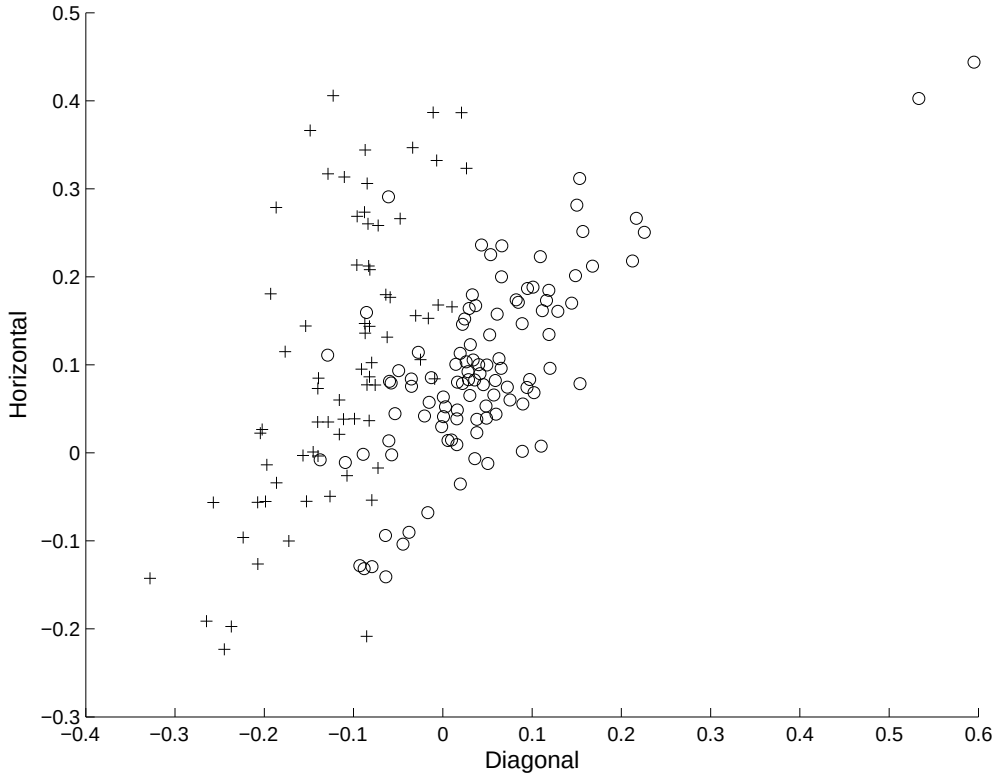


Figure 8: Scatter plot of H_h versus H_d. *Circles* denote controls, and *crosses* denote cancer cases.

Regardless of the number or combination of H's used in classification, the overall performance of the classifiers remained the same, with AV producing the lowest overall error rates, followed by TT, and with OLS performing the worst of all estimators. Results were consistent for a range of wavelets and level choices. It should be noted that these are not necessarily global phenomena, rather specific observations in mammogram classification.

2.5 Enhanced Scaling Estimator

It is important to note that estimation of H and classification by H are two different tasks, and optimal estimators in one context may not perform well in the other. This is especially true for real-life images for which theoretical models are just an approximation. After exploratory simulation, we found that slopes based on the finer

levels in wavelet decompositions are more critical for classification purposes. We then devised a multiplier, $2^{(i+j)}$, to enhance the TT weight by more heavily emphasizing the fine detail levels. This forms an additional alternative weight, which we denoted as the Enhanced Theil-type (ETT). This estimator of the overall slope (by which we estimate H) then is the weighted average of pairwise slopes between levels i and j where $i < j$,

$$\sum_{i,j} w_{ij} s_{ij} / \sum_{i,j} w_{ij},$$

with weights

$$w_{ij} \propto 2^{(i+j)} (i-j)^2 \times HA\left(2^{2i}, 2^{2j}\right).$$

Results for this modified weight are shown in Table 22. When comparing these results with those of other estimators, ETT outperforms the others in the classification task in every case. While the binary logistic regression improves with each additional predictor added, you can see that linear and quadratic classification with two predictors seem to perform particularly well for the ETT estimator, even outperforming binary logistic regression with three predictors. The best case scenario is given with linear classification using two predictors, (H_d,H_h), resulting in an error rate as low as 9.12%.

Table 22: Results of classifications using ETT estimator

	Method	**1-Se**	**1- Sp**	**Error**
(H_d)	binary logistic regression	0.1758	0.1601	0.1664
(H_d,H_h)	binary logistic regression	0.1088	0.1195	0.1151
	linear classification	0.0910	0.0913	0.0912
	quadratic classification	0.1042	0.0853	0.0930
(H_d,H_h,H_v)	binary logistic regression	0.1019	0.0945	0.0975

Note that there could be some trade-offs between sensitivity and specificity with different estimators and different combinations of directional H's in all cases. For example, quadratic classification using (H_d,H_h) estimated by ETT actually gives

the lowest $1 - Sp$ of 8.53%. Although these differences are very slight, a trade-off that slightly increases the overall error rate (by lowering the sensitivity) might be justified to raise the specificity and possibly counteract the adverse effects of anxiety, discomfort, and costs associated with a false positive.

2.6 Discussion & Conclusions

In this chapter we presented two novel wavelet-based estimators of scaling and investigated their diagnostic performance in classification of digital mammograms as cancer vs. non-cancer. TT is a newly defined Theil-type robust estimator, with the optimal weights for pairwise slopes depending on the harmonic average between sample sizes in each level. This estimator is free of distributional assumptions and robust with respect to outlier levels. Its modifications ETT is motivated by the specific application of diagnostic mammography, demonstrating that as the weighted average of pairwise slopes, this method allows for the modification of weights for further optimization in a particular context.

Performance of these estimators in the task of classification was also compared to that of two existing scaling estimators, OLS and AV. We found that ETT, AV, and TT estimators provided the best classification rates, in that respective order, for a range of wavelets and level choices. The standard wavelet-based OLS estimator did not perform well and our recommendation is that this estimator should not be used in tasks of classification. The overall misclassification rate of the new weights proposed in this paper was lower than the ordinary least squares estimate in all settings. It should be noted that these are not necessarily global phenomena, rather specific observations in mammography image classification.

Diagonal spectra (H_d) was found to be the most discriminatory and little power is lost if only this spectra is used. But, although H_d itself is strongly discriminatory and the most parsimonious classifying model, the use of H_d in combination with H_h

(or H_v) does perform better than H_d alone. Further, the results of classification using all three spectra (H_d, H_h, H_v) did perform slightly better than that with only one or two spectra. This implies that each wavelet spectra has some level of power to differentiate between normal and malignant cases.

The diagnostic use of information contained in the background of images is a novel concept that allows the use of information from the entire image, rather than focusing primarily on irregular shapes, masses, and calcifications. A meaningful implication of this research is the improvement of both sensitivity and specificity of current clinical diagnostic tests for breast cancer. The ambiguities involved in current diagnostic methods often result in extra costs, painful additional procedures, or missed cancers. With this tool, reasonable misclassification errors are achieved, and a promising new indicator may be added as an additional tool for physicians in current screening techniques.

CHAPTER III

ENHANCEMENT OF DIGITAL MAMMOGRAMS BY WAVELET-BASED SUB-PIXEL IMAGE INTERPOLATION

3.1 Introduction

Breast cancer prognosis is greatly improved when it is detected early. When assessing objects such as microcalcifications, factors such as their number, spacial arrangement, and their individual morphological features are very important for determining the possibility of malignancy. If the amount of clear information obtained from screening images can be increased while artifacts are still very small, the patient will likely have a better prognosis and less invasive options for both specific diagnosis and for treatment.

However, it can be very difficult to assess the morphological features of a single microcalcification through mammography alone when it is too small to capture well on an image. Calcifications are often microscopic and seen only by the pathologist, but the visible ones may be as small as 0.2 mm. When the goal is early detection, and assessing findings before they get any larger, it is of great importance to be able to get an accurate view of even the tiniest findings. This makes magnification and detail assessment on a very small scale of extreme importance.

The main objective of this research is to enhance the visualization and the assessment of morphological features of microcalcifications that are too small to capture well on a mammogram. The key is to produce an image of higher resolution that portrays as accurate a picture of the true shape's form as possible. Once this image is produced, in addition to better visualization of the shape, diagnostic methods such

as shape analysis and disease classification can be performed with greater precision.

Using scale-mixing discrete wavelet transform methods, the existing detail information contained in a course image can be used to interpolate scaled details at finer levels. These "informed" finer details can then be used in a method that involves the inverse transform of the original image plus the interpolated details. Through this process we are able to produce an average image of much higher resolution than the original, improving visualization, and producing a confidence area for the true location of the shape's borders, allowing for more accurate feature assessment and shape analysis.

This technique will allow for the visualization and assessment of information that is otherwise too small to clearly determine from standard mammogram images. With the ambiguities in current diagnostic methods, screenings often result in additional procedures, extra costs, or missed cancers. This could be a promising new enhancement technique with potential for improving current screening as an additional tool for physicians, leading to less missed cancers and/or less unnecessary follow-up procedures.

3.2 Background

3.2.1 Microcalcifications in Breast Cancer Detection

Although there are several objects and features in a mammographic image that are critical for proper diagnosis, this research focuses on improving the visualization of microcalcifications. About half of the cancers detected by mammography appear as a cluster of microcalcifications. However, microcalcifications are very common, seen in up to 86% of mammograms, and are usually benign occurrences. They are basically specks of calcium (residue) that may be found in any area of rapidly dividing cells. But when many are seen in a cluster, they may indicate a small cancer. In this case, factors such as their number, spacial arrangement, and their individual morphological

features become very important for determining the possibility of malignancy. They are usually associated with masses, and are sometimes associated with densities or asymmetric breast tissue, all of which may be benign or malignant. Although microcalcifications are most often associated with benign processes, their frequent presence in the processes of tissue growth makes them very useful conveyors of information when trying to assess findings as a whole and determine next steps for diagnosis. The exact details of their morphology could often be a telling sign of benign or malignant conditions, and thus allow a physician to be more confident in deciding on next steps.

Calcifications associated with benign conditions are usually larger, fewer in number, widely dispersed, and round. Calcifications associated with malignancy are usually smaller, more numerous, clustered, and variously shaped. In the middle are hard-to-tell calcifications which are often labeled indeterminate. The fact that microcalcifications associated with malignancy are more commonly smaller in size and more irregularly shaped adds another dimension of difficulty to early cancer detection. It can be very difficult to assess the morphological features of a single microcalcification through mammography alone when it is too small to capture well on an image. But the conditions that favor the appearance of microcalcifications within a malignant process also happen to control their size. Calcifications are often microscopic and seen only by the pathologist, but the visible ones may be as small as 0.2 mm. The usual ones, mostly seen on mammograms, are not larger than 0.5 mm. When the goal is early detection, and assessing findings before they get any larger, it is of great importance to be able to get an accurate view of even the tiniest findings. This makes magnification and detail assessment on a very small scale of extreme importance. The main goal of this research is to generate a procedure for enhancing digital mammograms based on scale-mixing discrete wavelet transform methods.

We will briefly describe the traditional method of 2-D discrete wavelet transformation (referred to hereafter as *traditional DWT*) and its use in previous applications to wavelet-based image interpolation. Then we will propose a novel method of wavelet-based image interpolation, utilizing a more powerful transformation that uses scale-mixing between the directional hierarchies. This novel scale-mixing approach produces a more precise translation of the limited information within a course image into one of higher resolution.

3.2.2 Review of Traditional 2-D Discrete Wavelet Transform

As described in the previous chapter, a wavelet transform decomposes a signal into many different scales or frequency bands by expressing it in terms of shifted and dilated versions of a wavelet function $\psi(t)$ and shifted versions of a scaling function $\phi(t)$. As a quick reminder, for specific choices of the scaling functions and wavelets, an orthonormal basis can be formed from the atoms

$$\begin{aligned} \psi_{j,k}(t) &= 2^{j/2}\ \psi(2^j t - k) \\ \phi_{j,k}(t) &= 2^{j/2}\ \phi(2^j t - k), \quad j,k \in \mathbb{Z}. \end{aligned}$$

Then, $X(t)$ can be represented by wavelets as

$$X(t) = \sum_k c_{J_0,k}\phi_{J_0,k}(t) + \sum_{j=J_0}^{\infty}\sum_k d_{j,k}\psi_{j,k}(t),$$

where

$$d_{j,k} = \int X(t)\ \psi_{j,k}(t)\ dt, \qquad c_{j,k} = \int X(t)\ \phi_{j,k}(t)\ dt, \tag{16}$$

are detail and scaling coefficients respectively.

For the purposes of highlighting the procedural differences in the traditional DWT and the scale-mixing DWT (introduced in the next section), we will now briefly go over how these transforms are implemented in practice. Calculating wavelet expansions directly is a computationally expensive task. Also, most interesting wavelets are

without a closed form. Mallat connected quadrature-mirror filtering and pyramidal algorithms from signal processing theory with wavelets. He demonstrated that DWT can be calculated very rapidly via cascade-like algorithms. As mentioned previously, the DWT can be achieved by matrix multiplication since it is a linear transformation. However, in practice, one performs the DWT without exhibiting the matrix W explicitly, but by using fast filtering algorithms based on quadrature mirror filters which are uniquely determined by the wavelet of choice and fast Mallat's algorithm [25]. We start with the use of these fast filtering algorithms in the one-dimensional case.

Let the length of a data-vector $\mathbf{y}$ be $n = 2^J$. Suppose that the vector $\mathbf{y}$ is wavelet-transformed to a vector $\mathbf{d}$. This vector $\mathbf{d}$ can be written as another vector of length 2^J:

$$\mathbf{d} = (\mathcal{H}^l\mathbf{y}, \mathcal{G}\mathcal{H}^{l-1}\mathbf{y}, \ldots, \mathcal{G}\mathcal{H}^2\mathbf{y}, \mathcal{G}\mathcal{H}\mathbf{y}, \mathcal{G}\mathbf{y}), \tag{17}$$

where l is any fixed number between 1 and $J = \log_2 n$. The operators $\mathcal{H}$ and $\mathcal{G}$ are defined by high- and low-pass filters corresponding to the wavelet of choice. For all commonly used wavelet bases, the taps of these filters are readily available in the literature or in standard wavelet software packages.

The elements of $\mathbf{d}$ are the wavelet coefficients. The sub-vectors described in (17) correspond to detail levels in the level-wise organized decomposition. In general, the jth detail level in the wavelet decomposition of $\mathbf{y}$ contains 2^j elements. This linear and orthogonal transform can be fully described by an $n \times n$ orthogonal matrix W, as with all linear transformations.

Now, return to the data-vector $\mathbf{y}$ of length $n = 2^J$. Denote by H_k a matrix of size $(2^{J-k} \times 2^{J-k+1})$, $k = 1, \ldots,$ with elements consistent with the operator $\mathcal{H}$. Define a matrix G_k in the same way, but with elements consistent with the operator $\mathcal{G}$. For the data-vector $\mathbf{y}$, the following matrix equation (representing a J-step discrete wavelet transformation) gives the connection between $\mathbf{y}$ and the wavelet coefficients $\mathbf{d}$ as in

(16):

$$\mathbf{d} = W_J \cdot \mathbf{y},$$

where W_J is defined iteratively,

$$W_1 = \begin{bmatrix} H_1 \\ G_1 \end{bmatrix}, \quad W_2 = \begin{bmatrix} \begin{bmatrix} H_2 \\ G_2 \end{bmatrix} \cdot H_1 \\ G_1 \end{bmatrix},$$

$$W_3 = \begin{bmatrix} \begin{bmatrix} \begin{bmatrix} H_3 \\ G_3 \end{bmatrix} \cdot H_2 \\ G_2 \end{bmatrix} \cdot H_1 \\ G_1 \end{bmatrix}, \ldots$$

This procedure is readily generalized to the 2-dimensional case. As a reminder, the 2-D wavelet basis functions are constructed via translations and dilations of a tensor product of the univariate wavelet and scaling functions:

$$\begin{aligned} \phi(t_1, t_2) &= \phi(t_1)\phi(t_2) \\ \psi^h(t_1, t_2) &= \phi(t_1)\psi(t_2) \\ \psi^v(t_1, t_2) &= \psi(t_1)\phi(t_2) \\ \psi^d(t_1, t_2) &= \psi(t_1)\psi(t_2), \end{aligned} \tag{18}$$

where h, v, and d denote the atoms capturing image features in the horizontal, vertical, and diagonal directions, respectively. The traditional 2-D DWT utilizes these basis functions, which lead to three directional spectra defined by hierarchies of detail coefficients.

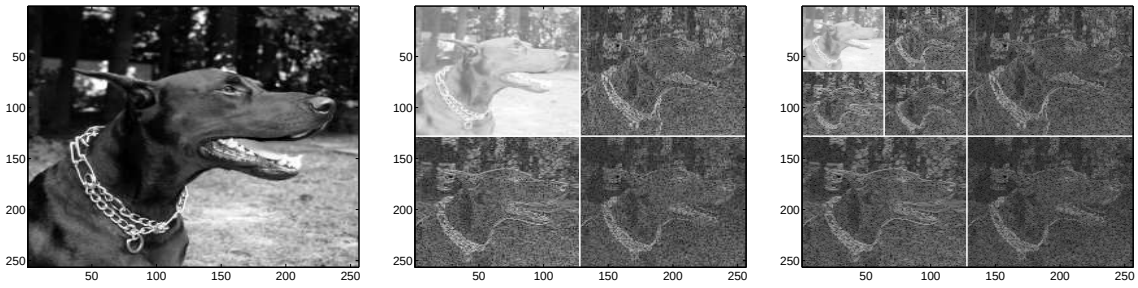
Procedurally, the traditional method of the 2-D DWT is performed by applying the univariate transform on rows and columns of a 2-D object (image), transforming each dimension at a time. During one iteration of a 1-D wavelet transform, two subvectors are produced that are each $\frac{1}{2}$ the size of the previous level: the "estimate"

or "average" of the previous level's information, and the "details" lost in estimating that average. Thus the traditional method that extends this into 2-D has that same effect on each dimension. One step of the decomposing algorithm proceeds as follows:

> Consider an image A, which is also expressible as a $2^n \times 2^n$ matrix comprised of pixel values. The process of wavelet decomposition begins by applying the wavelet low-pass filter $\mathcal{H}$ and high-pass filter $\mathcal{G}$ to the rows of matrix A. This step produces two matrices $\mathcal{H}_r A$ and $\mathcal{G}_r A$, both of dimension $2^n \times 2^{n-1}$ (the subscripts r denote that the filters are applied on rows of the matrix A). Next, apply the filters $\mathcal{H}$ and $\mathcal{G}$ to the columns of matrices $\mathcal{H}_r A$ and $\mathcal{G}_r A$ obtained from step one, producing matrices $\mathcal{H}_c\mathcal{H}_r A, \mathcal{G}_c\mathcal{H}_r A, \mathcal{H}_c\mathcal{G}_r A$ and $\mathcal{G}_c\mathcal{G}_r A$ of dimension $2^{n-1} \times 2^{n-1}$. The matrix $\mathcal{H}_c\mathcal{H}_r A$ is an average representation of the original image, while the matrices $\mathcal{G}_c\mathcal{H}_r A, \mathcal{H}_c\mathcal{G}_r A$ and $\mathcal{G}_c\mathcal{G}_r A$ contain detailed features of image A.

As can be seen in Figure 9(b), during one complete iteration of the traditional 2-D wavelet transform, 4 sub-matrices are produced that are each $\frac{1}{4}$ the size of the previous level (because each dimension was reduced in size by $\frac{1}{2}$): the "estimate" or "average" of the previous level's information, and 3 different sets of "details" lost in estimating that average (each corresponding to a direction of horizontal, vertical, or diagonal). The process is done iteratively, reducing the representation of the signal to $\frac{1}{4}$ its size with every iteration. Figure 9(c) shows an image after two iterations. To produce images with more reduced details, one may repeat the process using the *average* matrix $\mathcal{H}_c\mathcal{H}_r A$ in place of A.

The traditional method described here extends the 1-D wavelet atoms shown in (16) utilizing the 2-D wavelet basis functions shown in (18) as follows. Consider the



(a) (b) (c)

Figure 9: Traditional 2-D Wavelet Transformation. (a) original image; (b) traditional DWT after one iteration; (c) traditional DWT after two iterations.

wavelet atoms:

$$\phi_{j,\mathbf{k}}(\mathbf{t}) = 2^j \ \phi(2^j t_1 - k_1, 2^j t_2 - k_2) \tag{19}$$

$$\psi^i_{j,\mathbf{k}}(\mathbf{t}) = 2^j \ \psi^i(2^j t_1 - k_1, 2^j t_2 - k_2), \tag{20}$$

for $i = h, v, d, \quad j \in \mathbb{Z}$, $\mathbf{t} = (t_1, t_2) \in \mathbb{R}^2$, and $\mathbf{k} = (k_1, k_2) \in \mathbb{Z}^2$. Then, any function $X \in \mathcal{L}_2(\mathbb{R}^2)$ can be represented as

$$X(\mathbf{t}) = \sum_{\mathbf{k}} c_{J_0\mathbf{k}} \phi_{J_0,\mathbf{k}}(\mathbf{t}) + \sum_{j \geq J_0} \sum_{\mathbf{k}} \sum_{i} d^i_{j,\mathbf{k}} \psi^i_{j,\mathbf{k}}(\mathbf{t}), \tag{21}$$

where the wavelet coefficients are given by

$$d^i_{j,\mathbf{k}} = 2^j \int X(\mathbf{t}) \ \psi^i(2^j \mathbf{t} - \mathbf{k}) \ d\mathbf{t}.$$

3.2.3 The Scale-Mixing 2-D Discrete Wavelet Transformation

In this section, we generalize the form of the 2-D wavelet transform in a different way and show that the new transform will be capable of interfacing different scales in assessing the energy distribution of the image. As can be seen in Figure 10, different methods of generalizing the 2-D wavelet transform lead to different types of tessellations (or tiling) of the squared image. For example, if instead of (19) and (20), the wavelet atoms are defined in a way that allows the indexing of each scale

within each dimension, then several hierarchies of details can be utilized. These new wavelet atoms are define as:

$$\phi_{(j_1,j_2),\mathbf{k}}(\mathbf{t}) = 2^{(j_1+j_2)/2}\ \phi(2^{j_1}t_1 - k_1, 2^{j_2}t_2 - k_2) \tag{22}$$

$$\psi^i_{(j_1,j_2),\mathbf{k}}(\mathbf{t}) = 2^{(j_1+j_2)/2}\ \psi^i(2^{j_1}t_1 - k_1, 2^{j_2}t_2 - k_2), \tag{23}$$

where i is one of h, v, or d, as in (18) and $(j_1, j_2) \in \mathbb{Z}^2$. Then for $X \in \mathcal{L}_2(\mathbb{R}^2)$

$$\begin{aligned} X(\mathbf{t}) &= \sum_{\mathbf{k}} c_{(J_0,J_0),\mathbf{k}}\ \phi_{(J_0,J_0),\mathbf{k}}(\mathbf{t}) \\ &+ \sum_{j\geq J_0}\sum_{\mathbf{k}} d_{(J_0,j),\mathbf{k}}\ \psi^h_{(J_0,j),\mathbf{k}}(\mathbf{t}) \\ &+ \sum_{j\geq J_0}\sum_{\mathbf{k}} d_{(j,J_0),\mathbf{k}}\ \psi^v_{(j,J_0),\mathbf{k}}(\mathbf{t}) \\ &+ \sum_{j_1,j_2\geq J_0}\sum_{\mathbf{k}} d_{(j_1,j_2),\mathbf{k}}\ \psi^d_{(j_1,j_2),\mathbf{k}}(\mathbf{t}), \end{aligned}$$

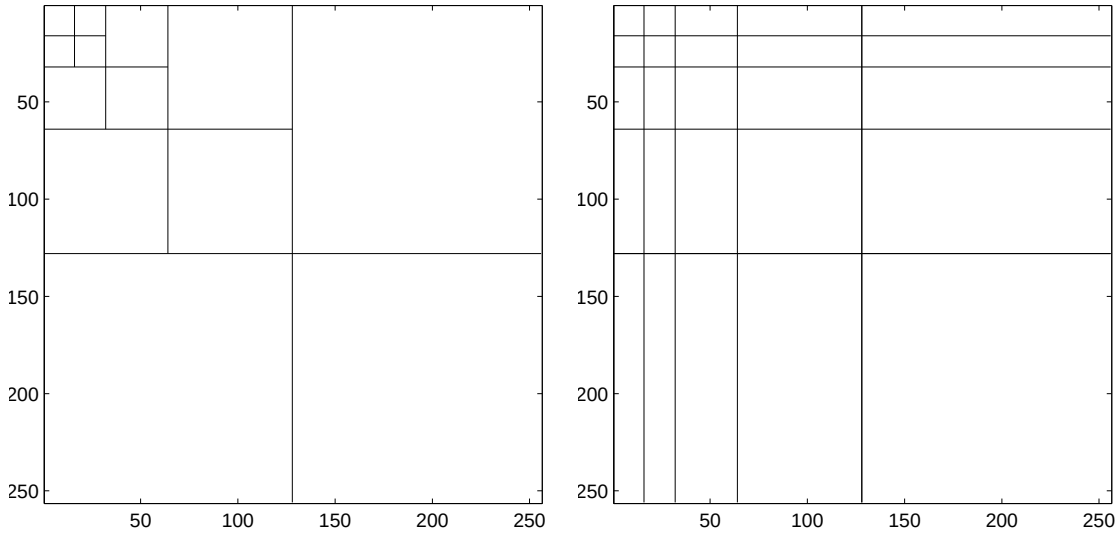
and a new 2-D wavelet transform, called throughout this paper *scale-mixing DWT* is obtained. The new scale-mixing detail coefficients are defined as

$$\begin{aligned} d_{(J_0,j),\mathbf{k}} &= 2^{(J_0+j)/2}\int X(\mathbf{t})\ \psi^h(2^{J_0}t_1 - k_1, 2^j t_2 - k_2)\ dt_1\ dt_2, \\ d_{(j,J_0),\mathbf{k}} &= 2^{(j+J_0)/2}\int X(\mathbf{t})\ \psi^v(2^j t_1 - k_1, 2^{J_0}t_2 - k_2)\ dt_1\ dt_2, \\ d_{(j_1,j_2),\mathbf{k}} &= 2^{(j_1+j_2)/2}\int X(\mathbf{t})\ \psi^d(2^{j_1}t_1 - k_1, 2^{j_2}t_2 - k_2)\ dt_1\ dt_2. \end{aligned} \tag{24}$$

Similar to the traditional one- and two-dimensional cases, the scale-mixing detail coefficients are linked to the original image (2-D time series) through a matrix equation. Suppose that a $2^n \times 2^n$ image (matrix) A is to be transformed into the wavelet domain. If the rows of A are transformed by a one-dimensional transform given by the $2^n \times 2^n$ wavelet matrix W, then the object WA' represents a matrix in which the columns are transformed rows of A. If the same is repeated on the rows of WA' the result is

$$B = W(WA')' = WAW'. \tag{25}$$

Matrix B will be called scale mixing transform of matrix A. It represents a finite-dimensional implementation of (24) for signal $X(\mathbf{t})$ sampled in a form of matrix A. The tessellation induced by the transform in (25) is shown in Figure 10(b).



(a) (b)

Figure 10: Tessellations for 2-D wavelet transforms. (a) Traditional 2-D transform of depth 4; (b) Scale-mixing wavelet transform of depth 4.

The scale-mixing 2-D transform is operationally appealing. The images are usually of moderate size and constructing appropriate W is computationally fast. Since W is orthogonal, the inverse transform is straightforward,

$$A = W'BW.$$

Unlike the traditional 2-D wavelet transform in which extension to rectangular matrices substantially complicates the algorithm, the corresponding scale-mixing 2-D wavelet transforms are straightforward. Since the wavelet transform is applied on the rows first, and then on the columns after (rather than iterating between rows and columns), one can handle not only the rectangular images, but also different bases in W and W', multiple transforms $W_1W_2AW_2'W_1'$, and so on.

By inspecting the tessellation in Figure 10(b), several hierarchies of detail spaces can be identified. The diagonal hierarchy interfaces coefficients with the same component scales and coincides with the diagonal hierarchy in the traditional 2-D spectra.

Just above and below the diagonal hierarchy are hierarchies of detail spaces that interface the scales that differ by 1. For example, the hierarchy above the diagonal, the scales along x-direction are interfaced by the next coarser scale along y-direction. For the hierarchy below the diagonal, roles of x and y are interchanged. Figure 11 shows three hierarchies of detail coefficients: the diagonal hierarchy (circles) and the hierarchies in which dyadic scales differ by 1 (triangles and squares).

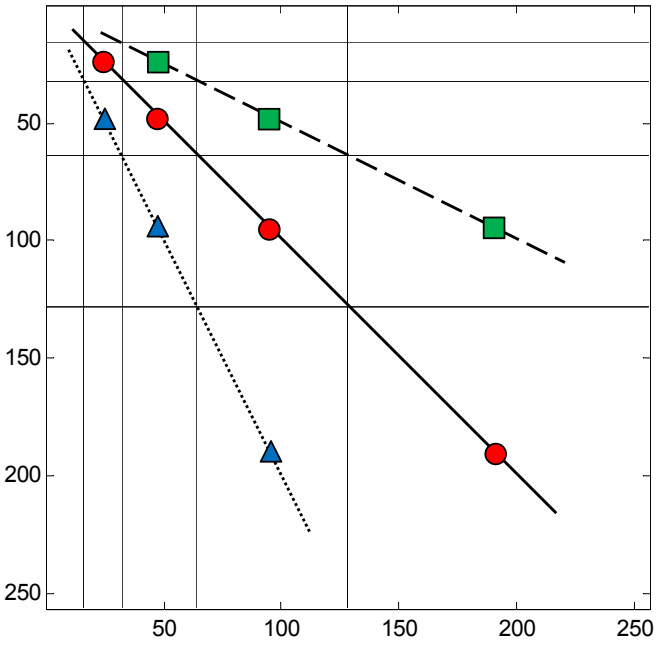


Figure 11: Three detail-space hierarchies generating the scale-mixing 2-D transform, where (j_1, j_2) is indexed as $(j, j+s)$, $s \in \mathbb{Z}$. Circles correspond to $s = 0$, triangles to $s = 1$, and squares to $s = -1$.

The scale-mixing 2-D wavelet transform is typically more compressive than the traditional 2-D wavelet transform, which is a desired property when dimension reduction applications (denoising, compression) are of interest. Informally, if the transform is of depth 2, in scale-mixing transform 9/16 of coefficients correspond to the differencing filters (ψ) in two dimensions while for the traditional transform this proportion is 5/16. The rest of the coefficients correspond to the atoms containing at least one scaling function (ϕ). Taking this into the context of image enhancement, when recovering an object by the reverse-transform, the coefficients that correspond to a scaling

function (corresponding to atoms with ϕ) will contribute to the overall shape of the recovered object. Since the goal in the current application is to maintain the true shape of the object while enhancing only the fine details, it is beneficial to be able to interpolate information into the coefficient spaces that are not affecting the overall shape. Since there are much fewer of these ϕ coefficients in the scale-mixing DWT, it is more appealing for this application than the traditional DWT.

This transform has been applied to the context of environmental time-evolving spatial phenomena by Ramirez et al. [38], but this research provides more theoretical considerations and translation to the field of medical diagnostics.

3.2.4 Data

The images used for this analysis are also obtained from the University of South Florida's Digital Database for Screening Mammography (DDSM) [28], which is described in detail in [18]. Images from this database containing suspicious areas are accompanied by pixel-level "ground truth" information relating locations of suspicious regions to what was assessed and verified through biopsy. We selected a set of 10 images containing microcalcifications confirmed to be malignant, and a set of 8 images containing microcalcifications confirmed to be benign for enhancement. All original images were of size 64×64. From these images, a total of 16 cancerous calcifications and 16 benign calcifications were obtained for further diagnostic analysis.

3.2.5 Basic Image Interpolation Procedure

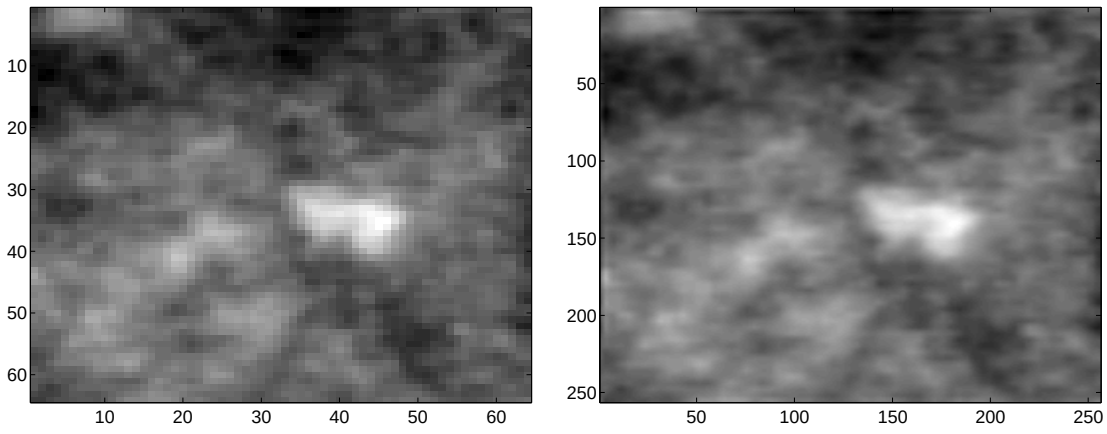
We will first describe a simple enhancement using the inverse scale-mixing DWT without any utilization of the detail spaces for adding further information. In this simple case, these coefficient spaces (besides the area containing the "degraded" image) will all be filled with zeros. When the inverse transform is performed, the higher resolution image is interpolated, and a higher level of detail can be seen. The procedure is as follows:

1. Start with an empty matrix (all entries zero) with dimensions equal to those of the desired final image. For example, if your course image is of size 64 × 64 and you wish to enhance it by two levels, then your final image will be of dimensions 256 × 256. So you will begin with a matrix of those dimensions with all entries zero.

2. The coarse image from a digital mammogram is inserted into matrix containing zeros, in the position where the "degraded" image would be (upper left). The rest of the matrix maintains the zeros.

3. The inverse transform is performed by the desired number of levels.

This process increases the resolution of the degraded, pixelized image and contains 4^k times the number of pixels in the original image, where k is the number of levels enhanced. Figure 12 shows the result of applying the simple enhancement by two levels to an image of a malignant case, using the inverse scale-mixing DWT without any utilization of the detail spaces for adding further information, but only zeros in the detail spaces.



(a) (b)

Figure 12: Results of applying simple scale-mixing wavelet interpolation on an image of a malignant calcification: (a) Original course image of size 64 × 64; (b) 2 level enhanced image of size 256 × 256.

3.2.6 Utilizing Detail Spaces

It is natural to propose utilization of detail spaces to further enhance the information in the interpolated image. As has been shown in previous sections, we propose to utilize the self-similarity and energy scaling of wavelet decompositions in building informative detail spaces.

By taking the original 64×64 course image and performing an initial scale-mixing DWT to reduce the image and obtain several levels of details, we can then use the innate relationship described by the the 2-dimensional wavelet-based spectra,

$$\log_2 E\left[\left|d_{(j,j+s);\mathbf{k}}\right|^2\right] = -(2H+2)j + \log_2 V_{\psi,s}(H), \tag{26}$$

where $V_{\psi,s}(H)$ is a constant depending on ψ, H, and s, but not on the scale j, to impute details at higher levels beyond the original 64×64 course image. By predicting the log-energies at the higher levels, we then can have some idea of their likely true behavior. The use of this information for further image enhancement is described in the next section.

3.3 Methods

3.3.1 Image Interpolation using Imputed Details and Stochastic Resonance

Stochastic resonance is a phenomenon that occurs when an appropriate measure of information is maximized in the presence of a non-zero level of stochastic input noise, and the system resonates at a particular noise level. The idea of adding noise to a system in order to improve the quality of measurements seems counterintuitive, since systems are usually constructed to reduce noise as much as possible and provide the most precise measurement of the signal of interest. But numerous experiments have demonstrated that, in both biological and non-biological systems, the addition of noise can actually improve the probability of detecting the signal. This is the idea of stochastic resonance. This concept can be imitated in the visual system by squinting

one's eyes or moving away from the image. This allows the observer's visual system to average the pixel intensities over areas [32].

We use the predicted behavior described in the previous section to follow the concept behind this idea of stochastic resonance. Since the optimal noise intensity is known approximately, we iteratively produce random gaussian noise of appropriate mean and standard deviation to fill the higher detail spaces with imputed information, and then perform the inverse wavelet transform. This converts the original 64×64 image into a larger image with imputed details. By performing this process iteratively, an indication of the true shape of the image is produced. Since with each iteration, the added noise is random, each produced image with individual random noise is different. Then, multiple noisy images are averaged. Thus, the error generated in one noisy image is minimized by the averaging of different noisy images.

This is carried out for each image by the following process:

We first assess the scaling behavior within the original 64×64 image by reducing original image down by 4 levels and using wavelet-based spectra to find innate relationship between levels. This can be seen in Figure 13, where the detail spaces chosen for use are highlighted in green.

After this scaling behavior is found, we then go through the following steps iteratively (1,000 iterations), using $MATLAB©$ software (code given in Appendix A):

- Start with an empty (all entries zero) matrix of size 256×256 (the desired size of the final image). Use the scaling behavior to impute details at higher levels by producing random gaussian noise of appropriate mean and standard deviation, replacing the zeros in those levels. This can be seen in Figure 14(a), where the filled detail spaces are shown in blue.

- Insert the original 64×64 image into the matrix in the position where a "degraded" image would be (upper left). This can be seen in Figure 14(b).

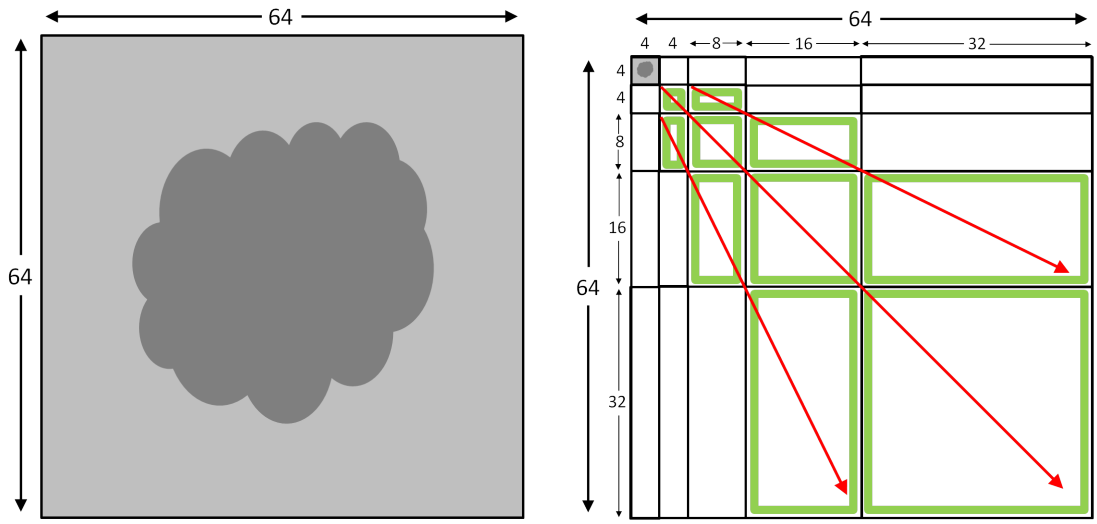


Figure 13: (a) An example of a mock original 64 × 64 image, and (b) an example of the image degraded by 4 levels to assess the innate scaling behavior within the image. Detail spaces used to project further level details are shown in green.

- Perform the reverse transform, enhancing the original 64 × 64 image by two levels. This produces a larger interpolated image, with imputed details, of size 256 × 256, as seen in Figure 14(c).

- We then perform edge detection on this interpolated image, using standard edge detection functions within $MATLAB$.

At each iteration, the following information is continually accumulated:

- Interpolated image information is accumulated within a 256 × 256 matrix. This means that each time an image is interpolated, the information is added to this same matrix, accumulating the total of the information for all 1,000 iterations.

- Detected edge information is accumulated within a 256 × 256 matrix. This means that each time an iteration is performed, the edge detection information is added to this same matrix, accumulating the total of the information for all 1,000 iterations.

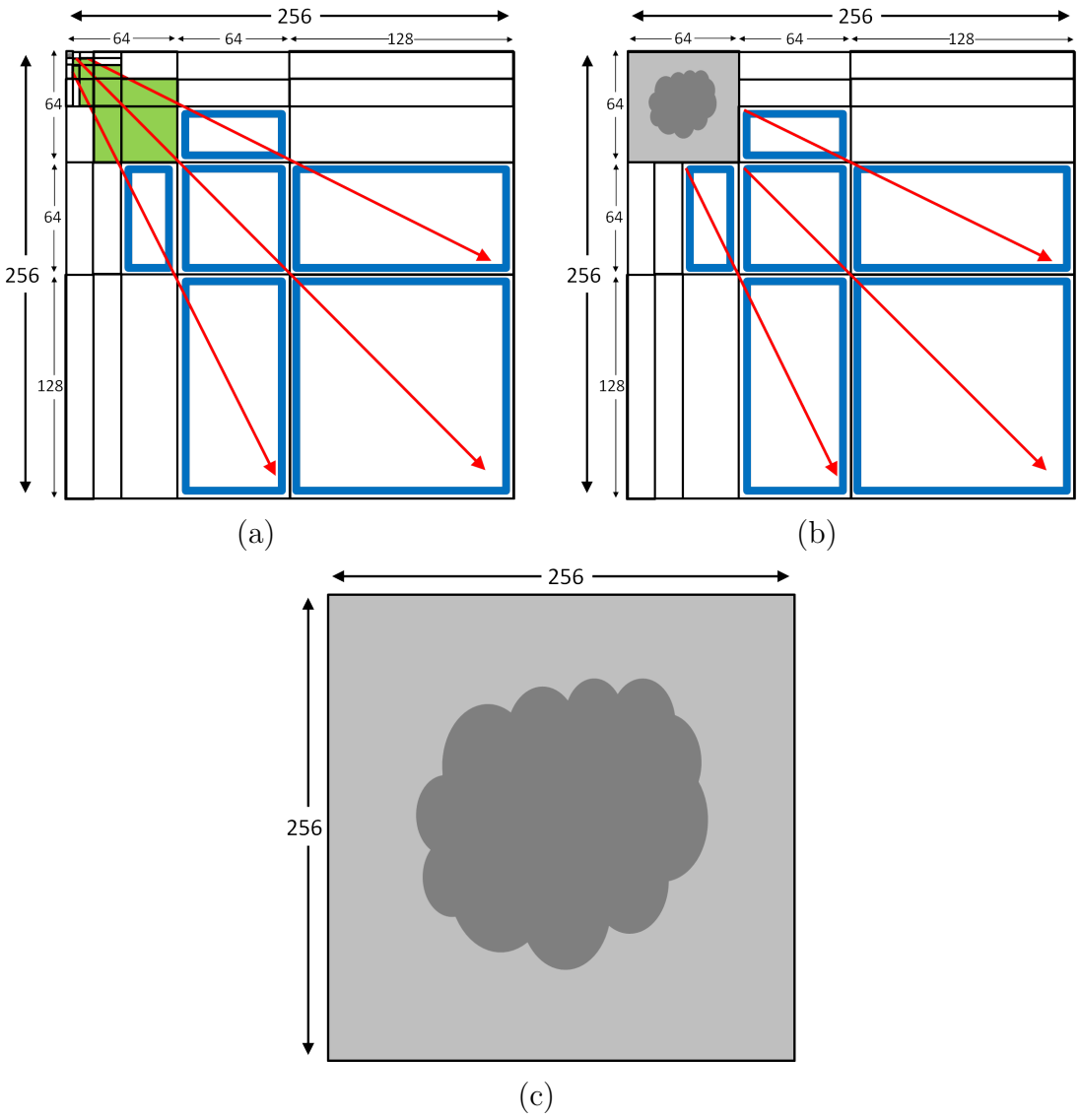


Figure 14: (a) Mock example of the degraded original 64 × 64 image placed in the upper left area of a 256 × 256 matrix, and the innate scale behavior projected to impute 2 more levels of details (new detail spaces shown in blue). (b) The 256 × 256 matrix with the newly imputed detail levels shown in blue, and the original 64 × 64 image placed in the upper left-hand corner where a "degraded" image would typically be. (c) The final 256 × 256 image resulting from a 2-level reverse transform with the imputed details.

After all iterations are performed, an average of the image information is found by dividing the total image matrix by 1,000, resulting in the final average image. Similarly, an average of the edge detection information is found by dividing the total edge matrix by 1,000, resulting in the final average edge detected.

3.3.2 Atlas of Characteristic Cases

Figures 15-23 show the results of applying this enhancement technique with imputed details for two additional levels, producing images of size 256 × 256. As described above, results from the average image and average edge detection are shown. The result shows that the proposed method is effective for improving image quality.

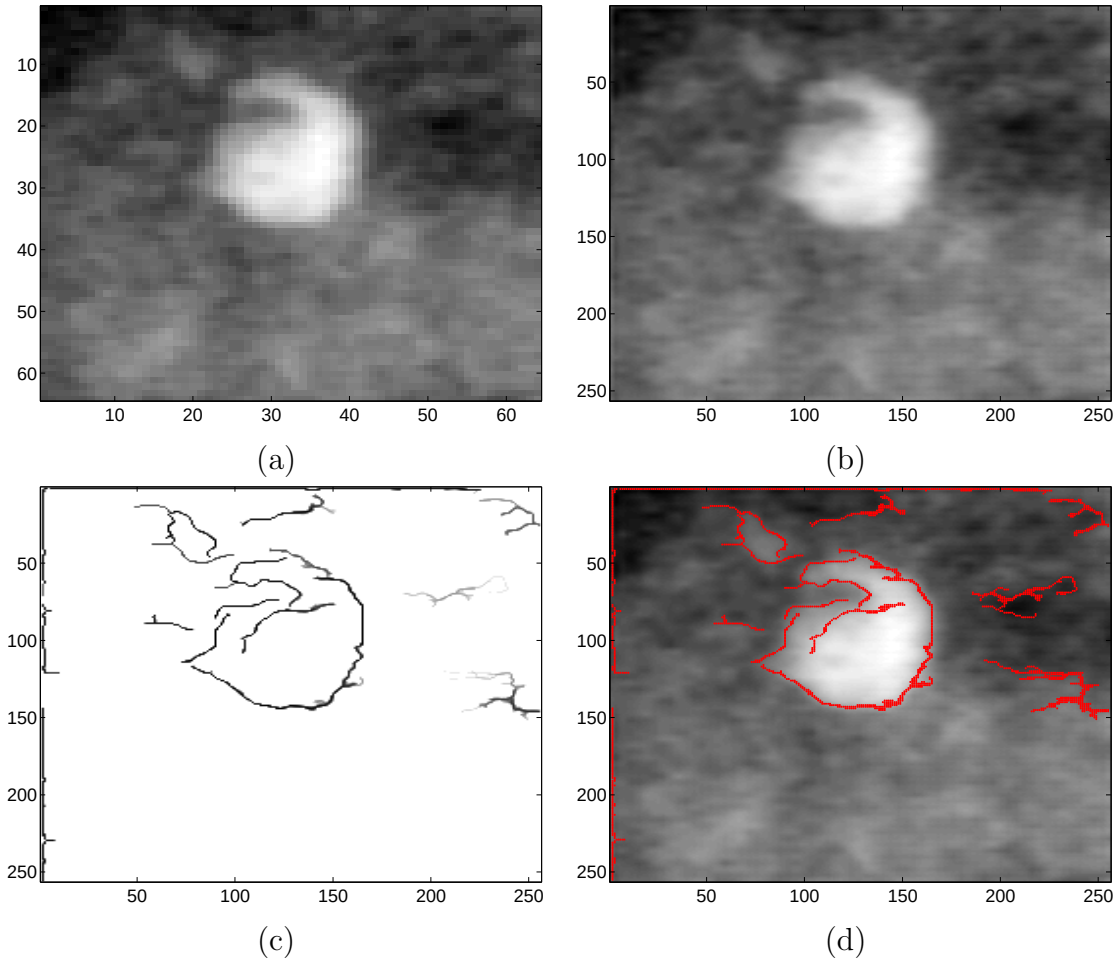


Figure 15: Results of applying the scale-mixing inverse DWT after interpolating two additional levels of details. (a) the original image of size 64 × 64; (b) averaged image after 1000 iterations of adding imputed random gaussian noise; (c) average edge detected after 1000 iterations; (d) average edge detection overlayed onto the averaged image.

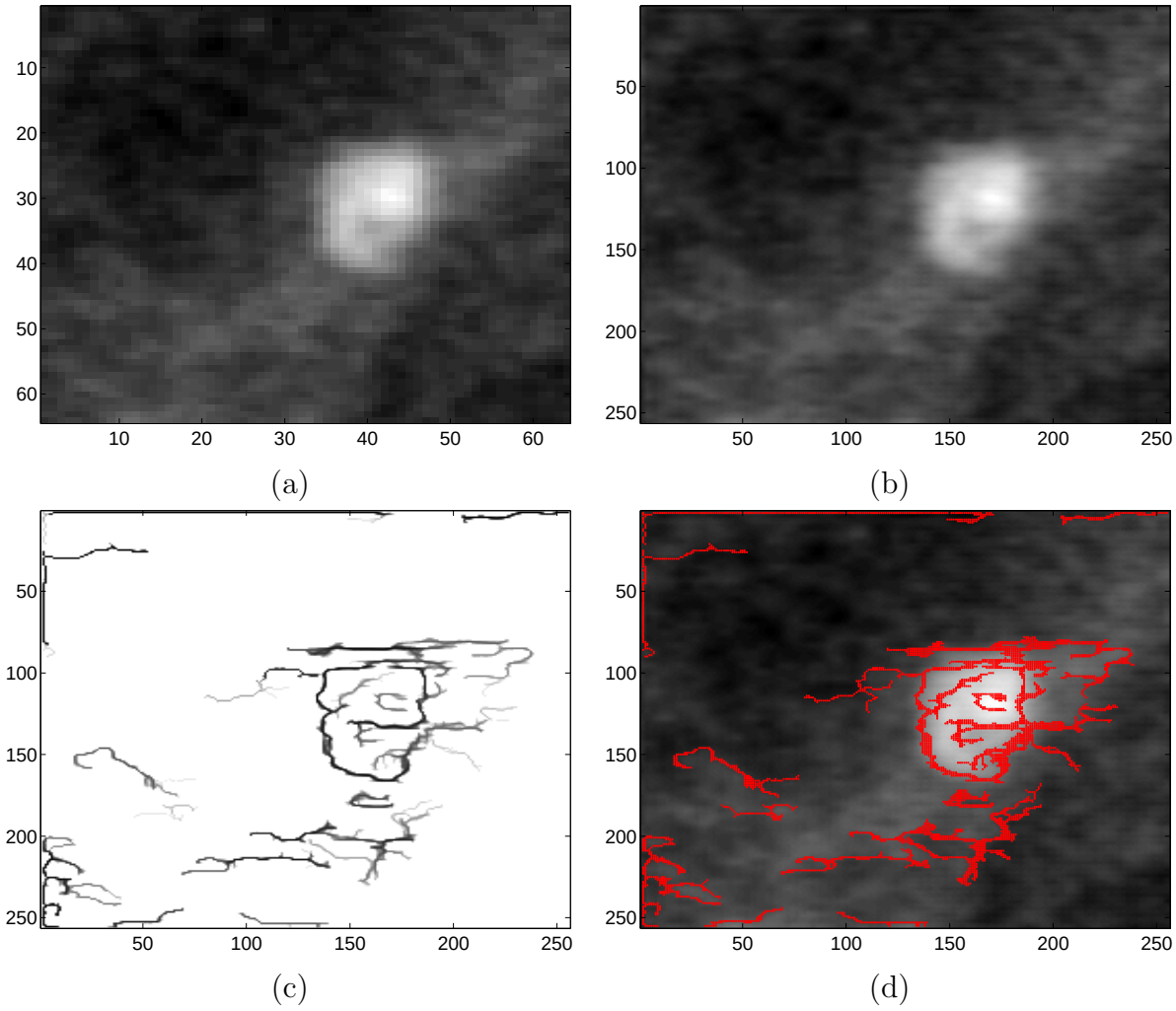


Figure 16: Results of applying the scale-mixing inverse DWT after interpolating two additional levels of details. (a) the original image of size 64 × 64; (b) averaged image after 1000 iterations of adding imputed random gaussian noise; (c) average edge detected after 1000 iterations; (d) average edge detection overlayed onto the averaged image.

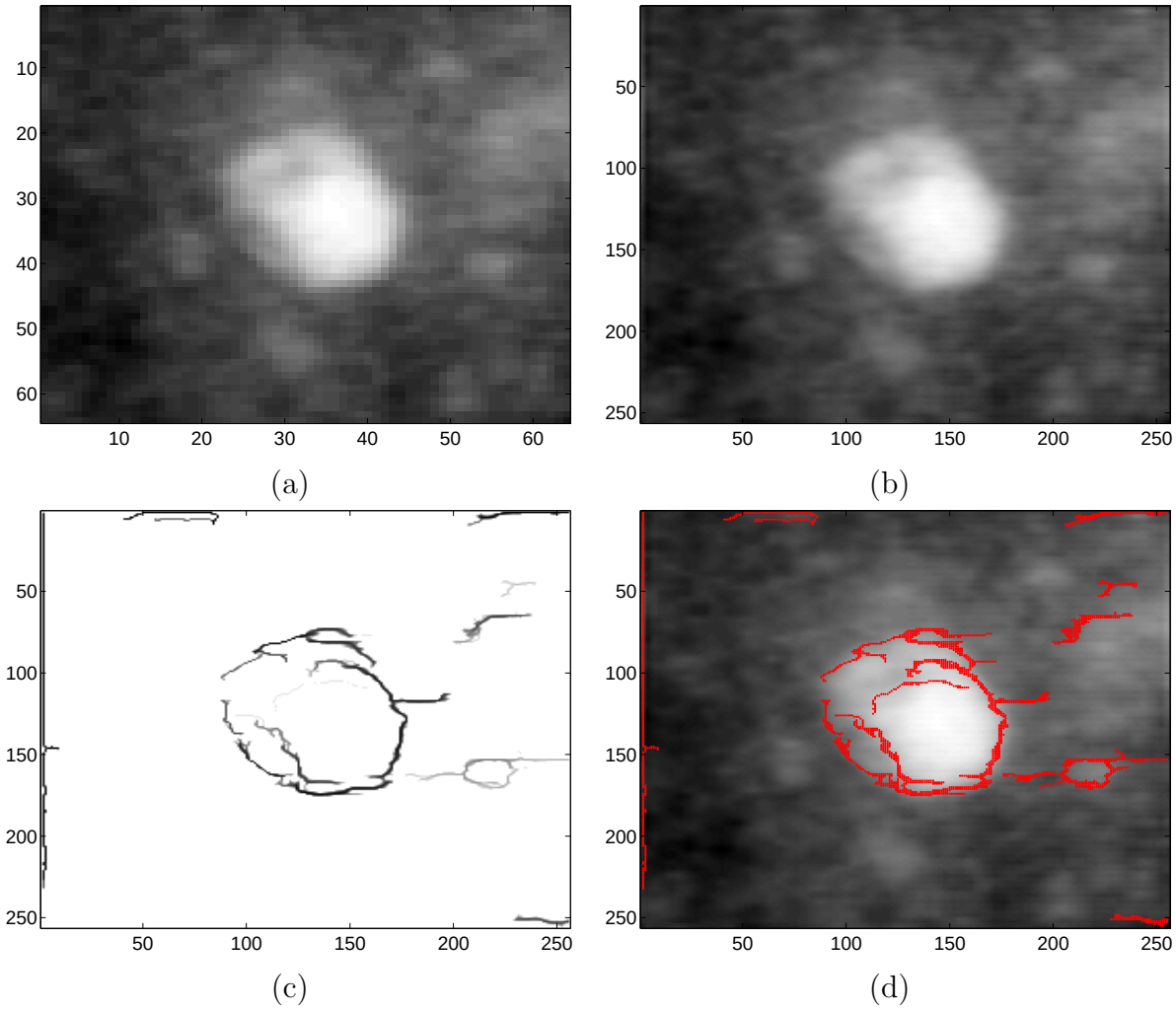


Figure 17: Results of applying the scale-mixing inverse DWT after interpolating two additional levels of details. (a) the original image of size 64 × 64; (b) averaged image after 1000 iterations of adding imputed random gaussian noise; (c) average edge detected after 1000 iterations; (d) average edge detection overlayed onto the averaged image.

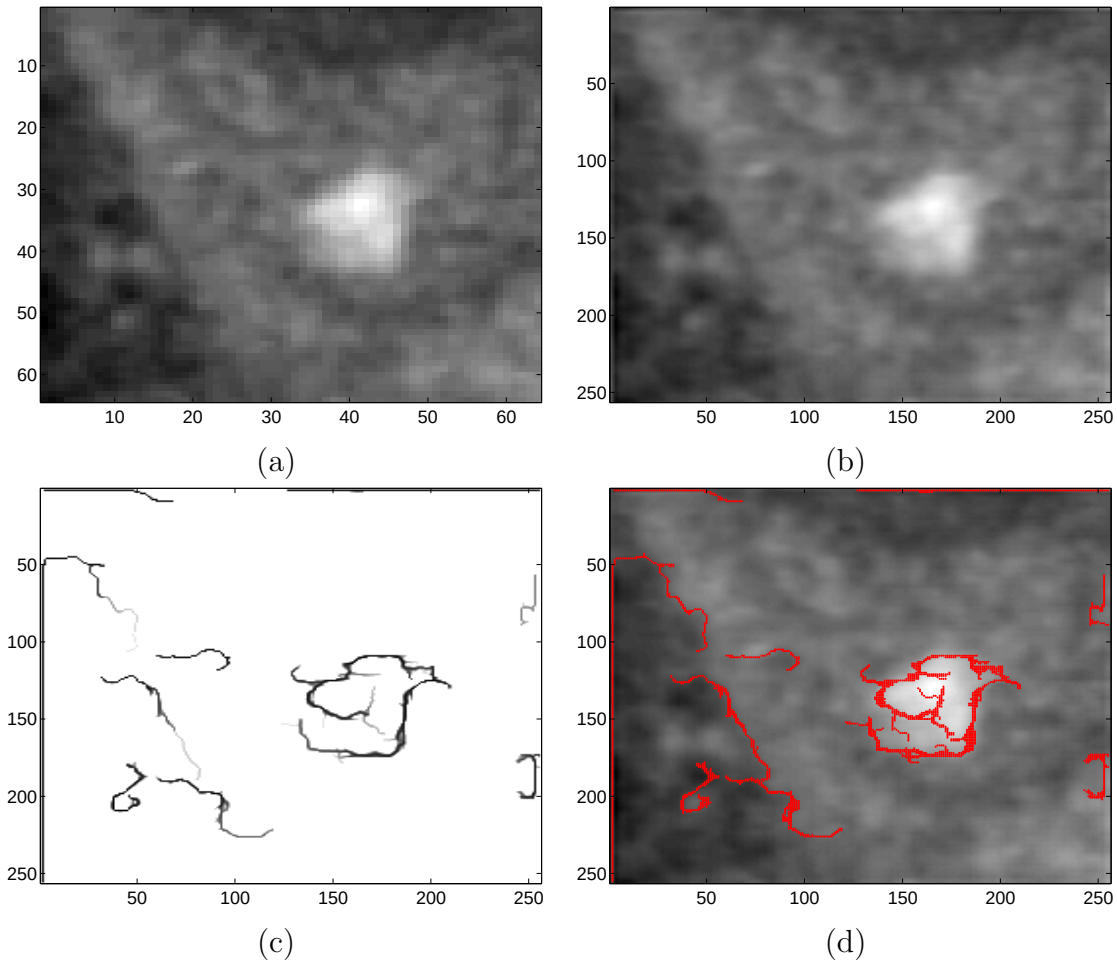


Figure 18: Results of applying the scale-mixing inverse DWT after interpolating two additional levels of details. (a) the original image of size 64 × 64; (b) averaged image after 1000 iterations of adding imputed random gaussian noise; (c) average edge detected after 1000 iterations; (d) average edge detection overlayed onto the averaged image.

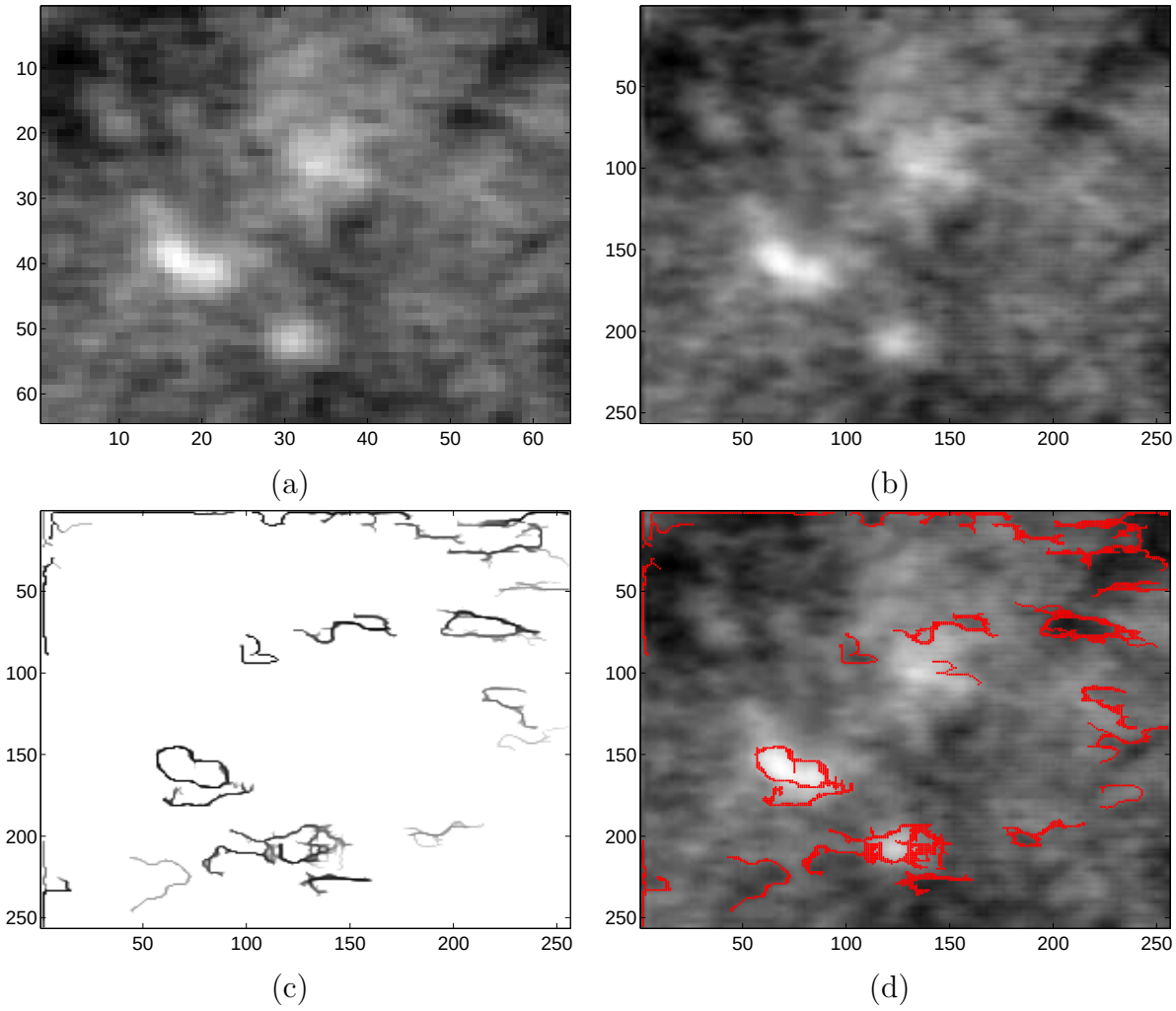


Figure 19: Results of applying the scale-mixing inverse DWT after interpolating two additional levels of details. (a) the original image of size 64 × 64; (b) averaged image after 1000 iterations of adding imputed random gaussian noise; (c) average edge detected after 1000 iterations; (d) average edge detection overlayed onto the averaged image.

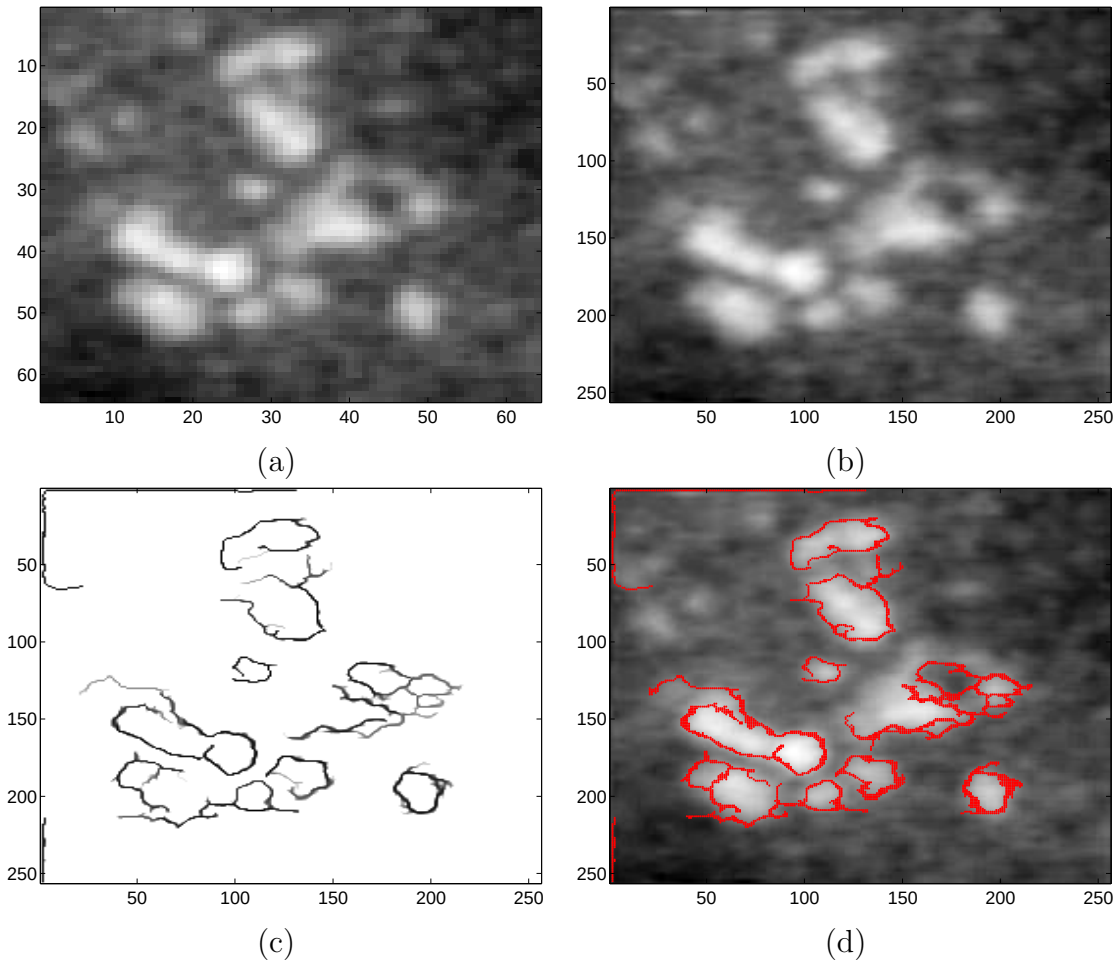


Figure 20: Results of applying the scale-mixing inverse DWT after interpolating two additional levels of details. (a) the original image of size 64×64; (b) averaged image after 1000 iterations of adding imputed random gaussian noise; (c) average edge detected after 1000 iterations; (d) average edge detection overlayed onto the averaged image.

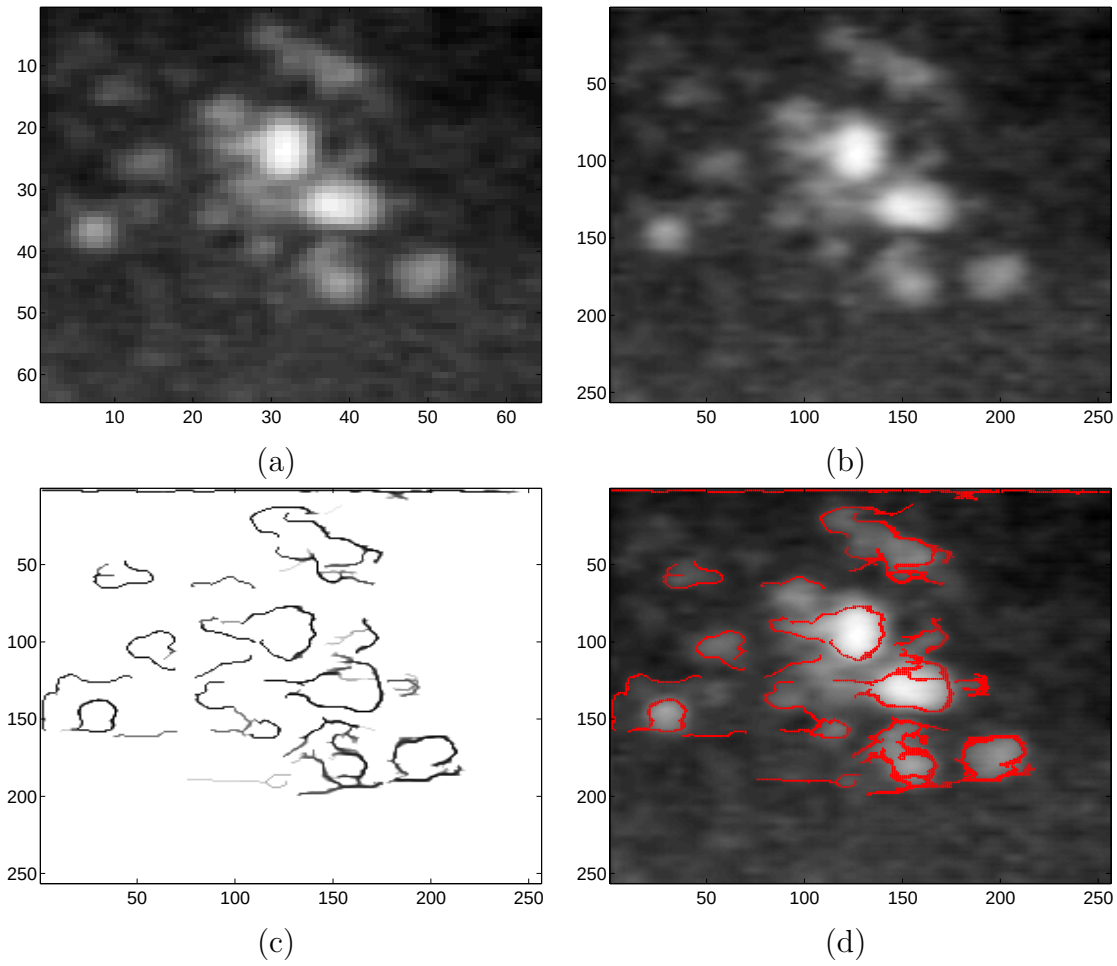


Figure 21: Results of applying the scale-mixing inverse DWT after interpolating two additional levels of details. (a) the original image of size 64×64; (b) averaged image after 1000 iterations of adding imputed random gaussian noise; (c) average edge detected after 1000 iterations; (d) average edge detection overlayed onto the averaged image.

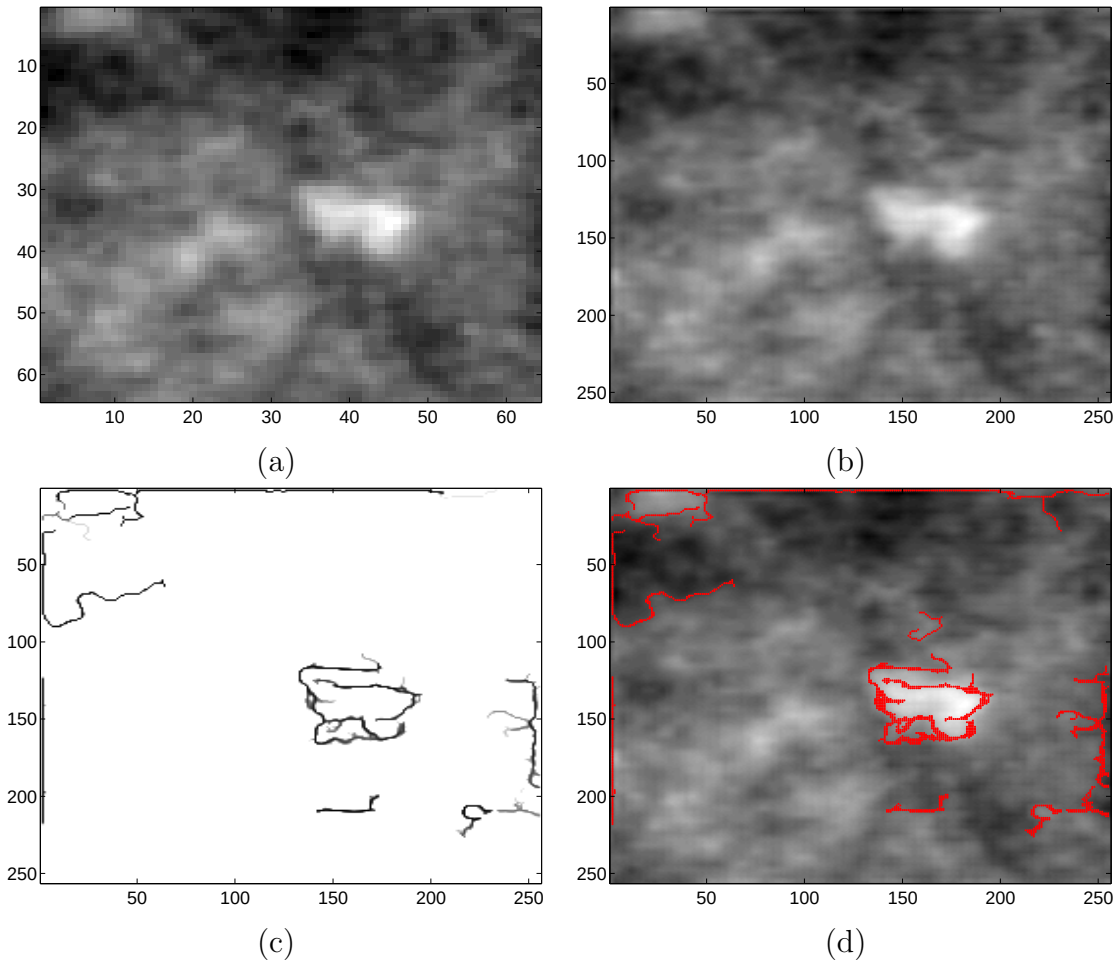


Figure 22: Results of applying the scale-mixing inverse DWT after interpolating two additional levels of details. (a) the original image of size 64 × 64; (b) averaged image after 1000 iterations of adding imputed random gaussian noise; (c) average edge detected after 1000 iterations; (d) average edge detection overlayed onto the averaged image.

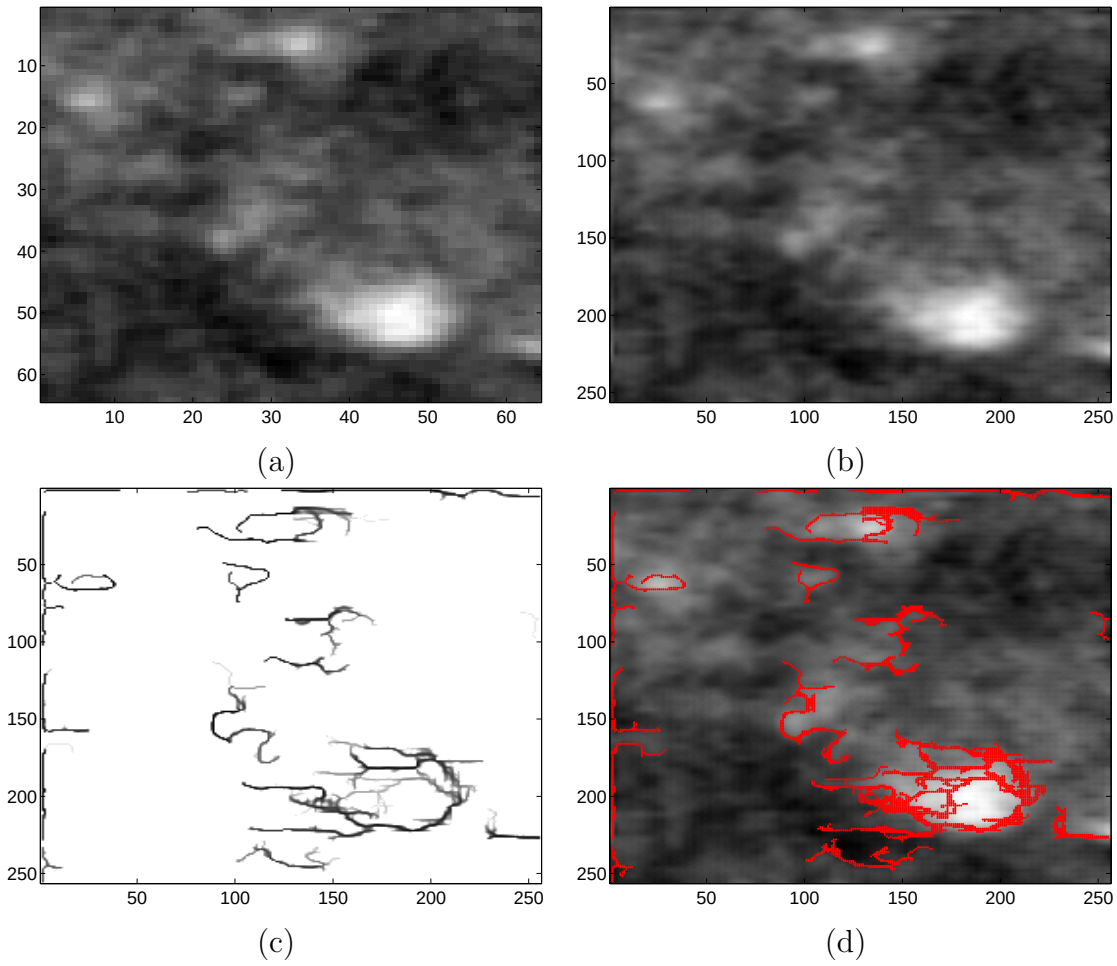


Figure 23: Results of applying the scale-mixing inverse DWT after interpolating two additional levels of details. (a) the original image of size 64×64; (b) averaged image after 1000 iterations of adding imputed random gaussian noise; (c) average edge detected after 1000 iterations; (d) average edge detection overlayed onto the averaged image.

3.3.3 Quantifying Results

In theory, if this method were to be placed into clinical practice, in addition to the enhanced ease of visualization, it would also be helpful to be able to quantify the results for more direct use in diagnostics. While there are many different methods of quantifying characteristics of calcifications in automated computer-automated detection (CAD) systems, some of them are very complex and very intricately programmed. For the purposes of proving the current method, within the scope of this thesis, we adopt a measure that is simple to implement, but gets at the important shape features of calcifications, taking into account the size and the irregularity of the shape. This measurement adopted here is the ratio of the shape border over the total shape area. For this measure, one would expect to see higher numbers for smaller and more irregular shapes (typically characteristic of cancer), and smaller numbers for more regular shapes (more typical of benign calcifications).

This measurement was assessed by placing a grid over each edge detected shape. This grid is made of squares of size 3 pixels × 3 pixels. The areas of the grid that overlap edges were then shaded green. Figure 24 shows some examples of this grid method. Let s_e represent the number of green squares and s_a represent the number of squares falling within the region enclosed by the green squares. For a particular case i then, the shape ratio, r_i is calculated by

$$r_i = \frac{s_e}{s_e + s_a}.$$

3.3.4 Diagnostic Methodology

The purpose of this method of image enhancement is for better visualization and diagnostic classification based on a shape's true form. To show the feasibility of this method, we will use the ratio described above to assess if there can be any differentiation between the shapes estimated for cancerous calcifications versus the

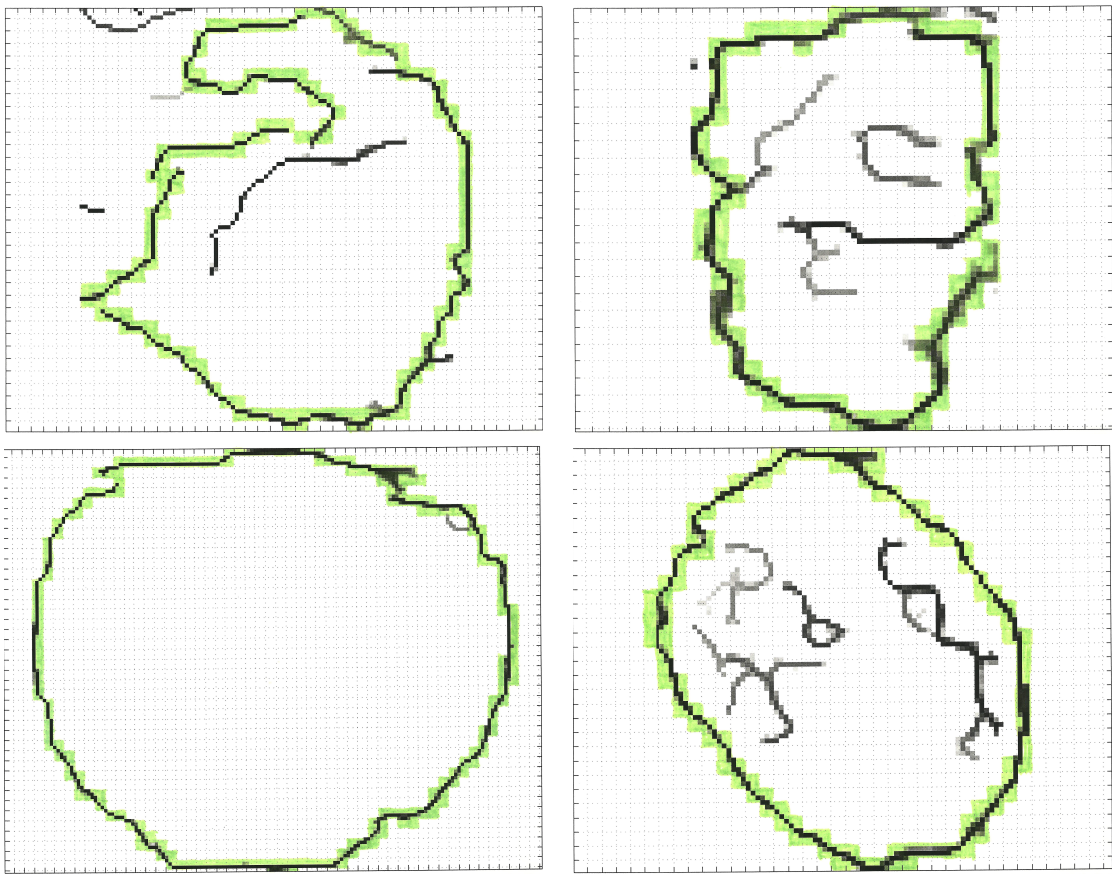


Figure 24: Examples of grid method for assessing ratio of the shape border over the total shape area.

shapes estimated for benign calcifications. Since there are only 16 images of cancerous calcifications and 16 images of benign calcifications, a resampling method called bootstrapping will be used to estimate the sampling distribution of the means of these two groups.

Resampling procedures in statistics are computer intensive methods that use an observed sample to produce many surrogate samples. Bootstrapping is arguably the most popular resampling methodology, made systematic by Brad Efron [12, 13]. One forms surrogate samples called *bootstrap samples* by sampling with replacement from the original sample. The bootstrap samples are of the same size as the original sample. If the original sample is $X_1, X_2, \ldots, X_n$ then $X_1^{*b}, X_2^{*b}, \ldots, X_n^{*b}$ is the bth bootstrap sample. Since the sampling is with replacement, some observations from the original

sample may not be selected in the bootstrap resample, while some may be selected more than once.

If the sample $X_1, X_2, \ldots, X_n$ produces statistic $\hat{\theta}$ for estimating population parameter θ, then each of B a bootstrap re-samples $X_1^{*b}, X_2^{*b}, \ldots, X_n^{*b}$, $b = 1, \ldots, B$, produces the counterpart statistic $\hat{\theta}_b^*$.

Original Sample	$X_1, X_2, \ldots X_n$	$\longrightarrow$	$\hat{\theta}$
Bootstrap Samples	$X_1^{*1}, X_2^{*1}, \ldots, X_n^{*1}$	$\longrightarrow$	$\hat{\theta}_1^*$
	$X_1^{*2}, X_2^{*2}, \ldots, X_n^{*2}$	$\longrightarrow$	$\hat{\theta}_2^*$
	$\ldots$		
	$X_1^{*B}, X_2^{*B}, \ldots, X_n^{*B}$	$\longrightarrow$	$\hat{\theta}_B^*$

If B is large, the ensemble of $\hat{\theta}_b^*$s approximates the sampling distribution of $\hat{\theta}$. In our case, the statistic $\hat{\theta}$ for each sample is the mean, which is used to estimate the population mean, θ. The table directly below shows the original ratios for each group, cancerous or benign.

Original Sample, Cancer	0.6364, 0.3118, 0.4868, 0.1896, 0.2197, 0.2559, 0.2712, 0.4417, 0.4183, 0.4923, 0.4183, 0.4015, 0.4419, 0.5364, 0.5882, 0.3245	$\longrightarrow$	$\overline{r_c} = 0.4022$
Original Sample, Benign	0.2274, 0.3533, 0.3003, 0.2466, 0.3871, 0.4396, 0.1149, 0.1290, 0.2000, 0.5000, 0.4800, 0.3378, 0.2678, 0.1834, 0.4375, 0.3620	$\longrightarrow$	$\overline{r_b} = 0.3104$

MATLAB code for all procedures in this section, including bootstrapping, is given in Appendix A. By performing the bootstrapping procedure on each sample separately (steps shown in figure 25), we are able to approximate the sampling distribution of both $\overline{r_b}$ and $\overline{r_c}$. Figure 26 shows the approximated sampling distributions for the ratio

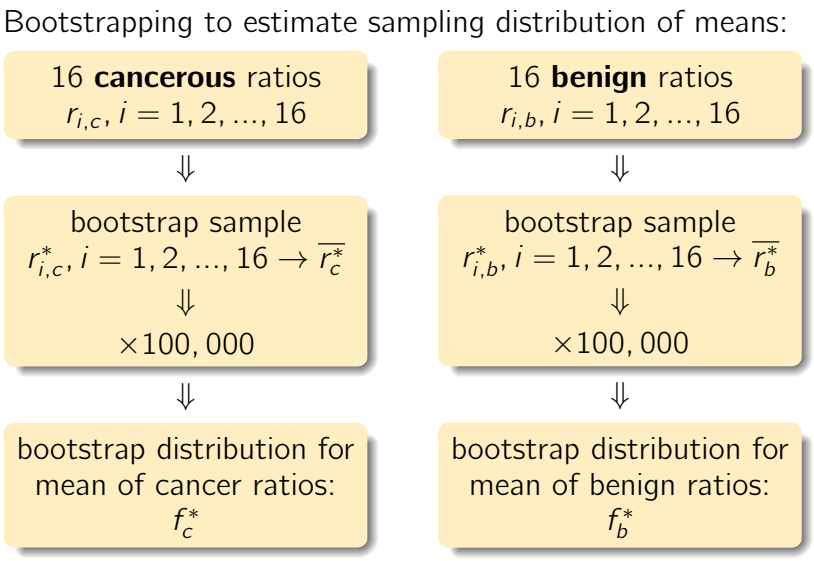


Figure 25: Bootstrapping procedure performed using each sample set of ratios (16 benign, 16 cancer) to approximate the sampling distributions for the ratio means of each.

means, with benign controls (f^*_b) in blue and cancer cases (f^*_c) in red, after running the bootstrap method with 100,000 repetitions.

Once the approximated sampling distributions, f^*_b and f^*_c, are obtained, they then lead to classification threshold (λ) setting. In practice, by setting a particular value for λ, we may then classify new cases whose ratios are found to be below λ to be considered benign, and those above λ to be possible cancer for which one would call back for further diagnostic testing.

$$r \leq \lambda \quad \rightarrow \quad -$$

$$r > \lambda \quad \rightarrow \quad +$$

We now go through several scenarios of threshold setting, using the 200,000 bootstrap sample means (100,000 cancer, 100,000 benign). In each scenario, as always in binary classification, there are four possible outcomes: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Table 23 summarizes

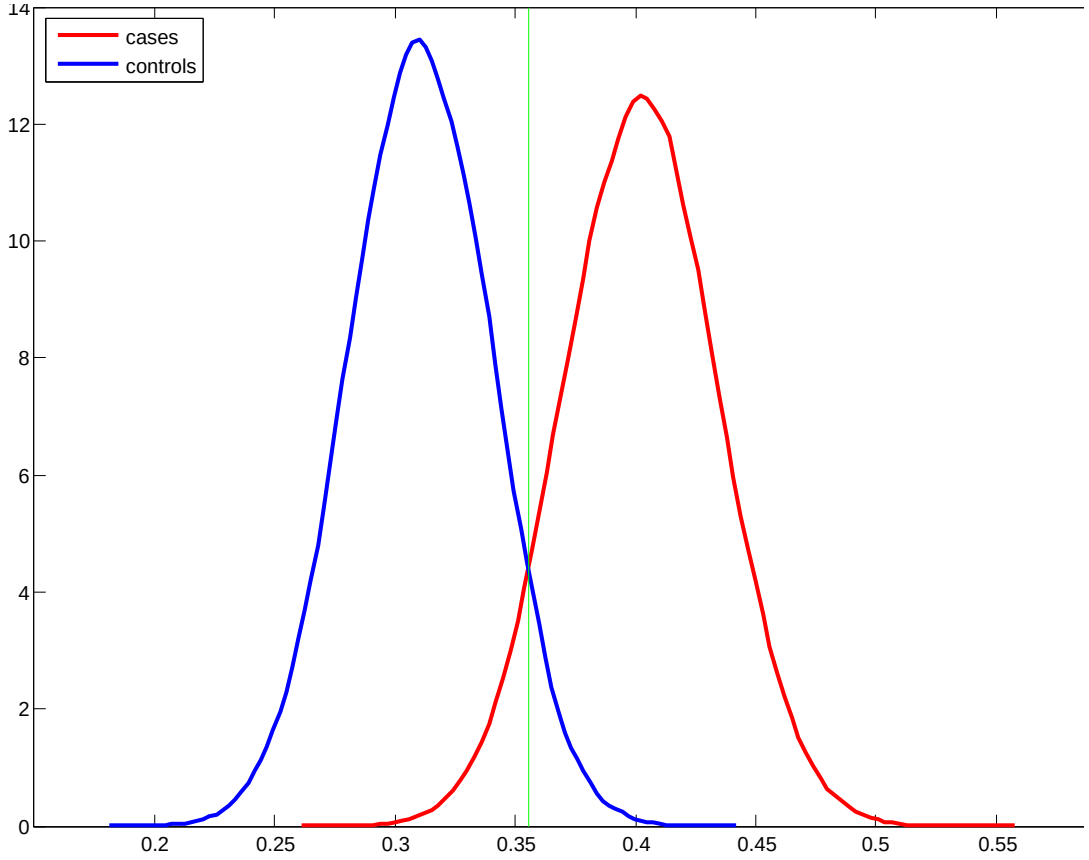


Figure 26: Results of approximated sampling distributions after running 100,000 bootstrap repetitions. Benign controls are shown in blue and cancer cases in red. The green line shows the diagnostic threshold set at the intersection between the two sampling distributions ($\lambda = 0.3555$).

these outcomes with their associated terminology. The number of positive instances is $N_p = TP + FN$. Similarly $N_n = TN + FP$ is the number of negative instances. So for each scenario of setting a value for λ, a resulting table of classification outcomes is formed.

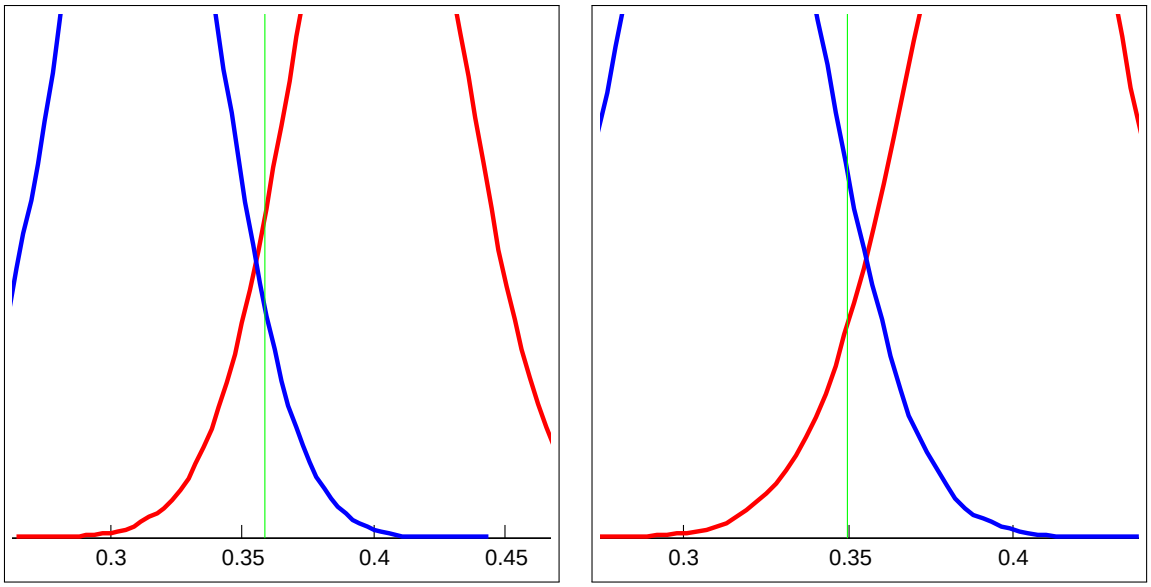
Table 23: Binary Classification Outcomes

		Predicted		
		Cancer	Benign	
True	Cancer	TP	FN	N_p
	Benign	FP	TN	N_n
		100,000	100,000	Total

The first strategy in setting the threshold could be to set it at the point where the two sampling distributions overlap. This is accomplished by setting $\lambda = 0.3555$, which results in an alpha α (probability of a false positive) of 0.0628 and a β (compliment of the power) of 0.0720. This threshold is shown as a green line in Figure 26.

Another strategy in setting the threshold could be to control α at 0.05. Since $\alpha = P(H_1 \mid H_0)$, controlling α at 0.05 means selecting the threshold that keeps the probability of a false positive below 5%. If we adopt this method, the threshold is set at $\lambda = 0.3588$. Figure 27(a) shows the intersection area of the density curves, adopting this λ. In this case, $\beta = 0.0861$, which means a test power of 90%.

If we choose to control the power of the test, this would be accomplished by controlling β. When using this method, one typically sets the power to 80% or 90% since a power that is too high could often result in an unreasonably high number of false positives, possibly causing unneeded stress and anxiety to patients. In this case, since we just went through the scenario of minimizing false positives, we will now show the other extreme of a very strongly powered test. But typically tests would not have such strong power. Here we show the scenario of 95% power (which means $\beta = 0.05$. To accomplish this, the threshold is set at $\lambda = 0.3500$, as shown in Figure 27(b). This results in an α (false positive rate) of 0.0893.



(a) (b)

Figure 27: Different thresholding scenarios. (a) Threshold set to control $\alpha = 0.05$ ($\lambda = 0.3588$). (b) Threshold set to control power at 95%, or $\beta = 0.05$ ($\lambda = 0.3500$).

A final method shown here for threshold setting is by the maximization of the F-measure, which is a common tool for assessing the performance of various classification tools. From the counts in Table 23, the following statistics are derived:

$$\text{Se} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$
$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

where Se is the sensitivity (also referred to as recall, or true positive rate), and PPV is the positive predictive value. The F-measure combines the sensitivity and positive predictive value into a single utility function which is defined as the harmonic mean of the two:

$$F = \frac{2}{1/\text{Se} + 1/\text{PPV}}.$$

To set the threshold using the F-measure, we would find the point that maximizes this measure. Figure 28 shows the plot of the F-measure results along the threshold range of 0.35-0.36. As seen in this figure, the F-score is maximized equally at two

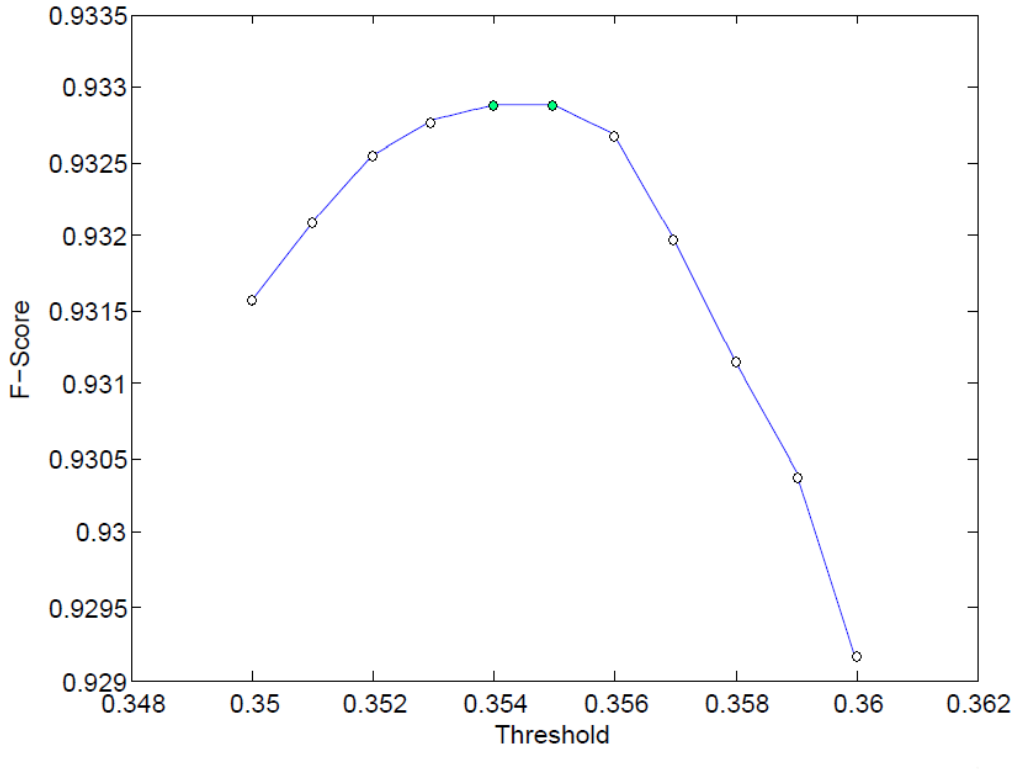


Figure 28: F-score for thresholds ranging $\lambda = 0.35$ to $\lambda = 0.36$. There are two maximum tests at thresholds $\lambda = 0.354$ and $\lambda = 0.355$.

different threshold points, $\lambda = 0.354$ and $\lambda = 0.355$. When $\lambda = 0.354$, this results in $\alpha = 0.0691$ and $\beta = 0.0630$. When $\lambda = 0.355$, this results in $\alpha = 0.0646$ and $\beta = 0.0686$.

Figure 29 shows a ROC curve, plotting the sensitivity (the true positive rate) against the compliment of the specificity (the false positive rate), along the same threshold range of 0.35-0.36. The ROC curve consideration is in agreement with the F-measure values, since the most distant points from the diagonal (which is typically an acceptable compromise between sensitivity and specificity) are also the points associated with thresholds $\lambda = 0.354$ and $\lambda = 0.355$.

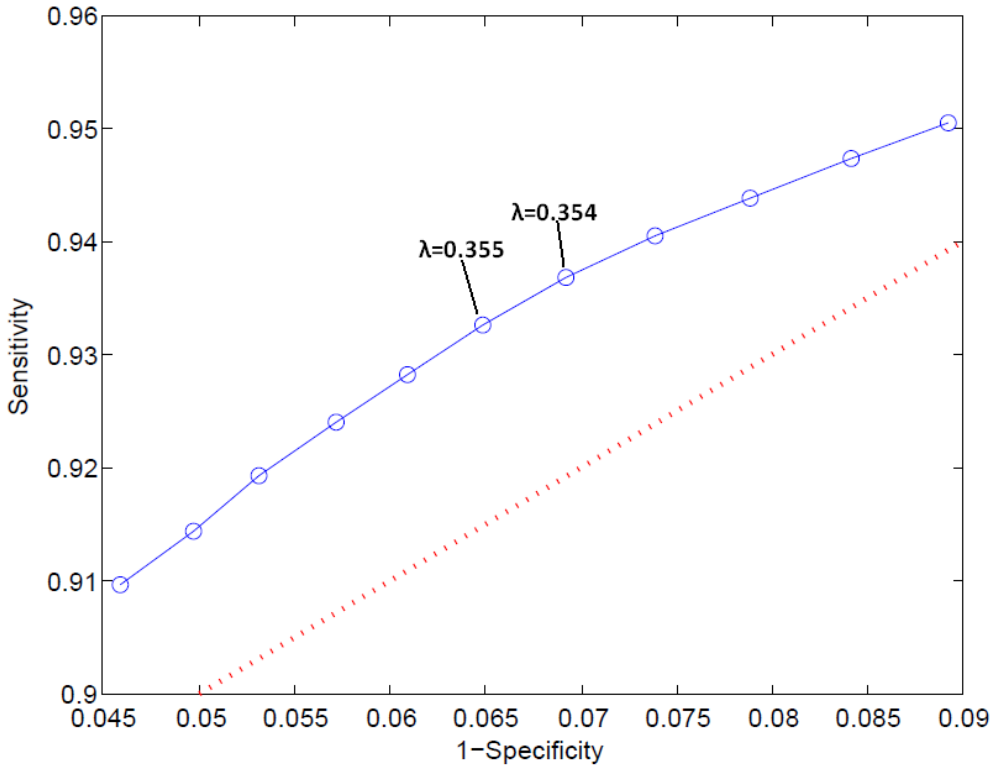


Figure 29: Part of the ROC curve that corresponds to thresholds ranging $\lambda = 0.35$ to $\lambda = 0.36$, with the diagonal shown by the dotted red line. Tests corresponding to thresholds $\lambda = 0.354$ and $\lambda = 0.355$ are equally good with respect to ROC criteria (furthest from the diagonal).

3.4 Discussion & Conclusions

The proposed methodology utilizes a variety of state-of-the-art techniques for the purposes of image enhancement to facilitate the analysis and diagnostic classification of mammograms. First, it uses a novel 2-D wavelet transform used in subpixel enhancement that utilizes fluxes of energy between different scales. This scale-mixing transform is more compressive and computationally simpler. We also introduced the idea of a sort of wavelet-based stochastic resonance, or wavelet-based bootstrap, allowing for the production of many surrogate images to facilitate in estimating the true form of the image. After edge detection and shape analysis, we also included a bootstrap-type diagnostic classifier into the context of microcalcifications.

The demonstration of this method has shown that in a setting where a threshold is chosen to find a balance between sensitivity and precision, we are able to produce α, β in the range of 0.06-0.07. We have shown feasibility for both visualization of calcifications and quantification of calcification characteristics. This feasibility was meant to show proof of the concept. In further research, one could dive more extensively into individual portions of this project, creating more precise edge detection techniques, further shape analysis, and/or add a much large number of images, likely further improving the results and proof of this method.

APPENDIX A

MATLAB CODE

Wavelet Transform, Detail Imputing, Image Interpolation, and Edge Detection

```
OrigIm=ca_a_1131_right_cc; %image to be enhanced

% display original image
figure(1);
colormap gray
imagesc(OrigIm)

% create an empty matrix for manipulating image; place image in it
B = zeros(256,256);
B(1:64,1:64) = OrigIm;

% make wavelet filter.
wf = MakeONFilterExt('Symmlet',4);

% make wavelet transformation matrix of approp. size and depth.
W = Wavmat(wf,256,2,4); %depth = 3

% enhance the image in the matrix of zeros just for comparison; display it
A = W'* B * W;
figure(2);
colormap gray
imagesc(A)


% --------ITERATIVELY CREATE NOISE TO OBTAIN THE APPROXIMATE SHAPE---------
% perform # of desired iterations to sample random normal noise (enhancing
% image and performing edge detection each time); return accumulated image
% info and accumulation of all edges; display it.

EdgeCI = zeros(size(B)); % a matrix to collect all iterations edge data
TotalImage = zeros(size(B)); % a matrix to collect all iterations of image data
k = 1000; % number of iterations
k2=k;
```

```
%% FIND SCALING BEHAVIOR OF ORIGINAL IMAGE
%perform wavelet transformation, 4 levels
h = MakeONFilterExt('Symmlet',4);
W64 = Wavmat(h, 64, 4, 4);
vcrsIm = W64*OrigIm*W64';

%locate detail spaces
d22=vcrsIm(dyad(2),dyad(2));
d23=vcrsIm(dyad(2),dyad(3));
d32=vcrsIm(dyad(3),dyad(2));
d33=vcrsIm(dyad(3),dyad(3));
d34=vcrsIm(dyad(3),dyad(4));
d43=vcrsIm(dyad(4),dyad(3));
d44=vcrsIm(dyad(4),dyad(4));
d45=vcrsIm(dyad(4),dyad(5));
d55=vcrsIm(dyad(5),dyad(5));
d54=vcrsIm(dyad(5),dyad(4));

%square
sqd22=d22.^2;
sqd23=d23.^2;
sqd32=d32.^2;
sqd33=d33.^2;
sqd34=d34.^2;
sqd44=d44.^2;
sqd43=d43.^2;
sqd45=d45.^2;
sqd55=d55.^2;
sqd54=d54.^2;

%find mean
msqd22=mean2(sqd22);
msqd23=mean2(sqd23);
msqd32=mean2(sqd32);
msqd33=mean2(sqd33);
msqd34=mean2(sqd34);
msqd44=mean2(sqd44);
msqd43=mean2(sqd43);
msqd45=mean2(sqd45);
msqd55=mean2(sqd55);
msqd54=mean2(sqd54);

%log of the mean square
lmsqd22=log2(msqd22);
```

```
lmsqd23=log2(msqd23);
lmsqd32=log2(msqd32);
lmsqd33=log2(msqd33);
lmsqd34=log2(msqd34);
lmsqd44=log2(msqd44);
lmsqd43=log2(msqd43);
lmsqd45=log2(msqd45);
lmsqd55=log2(msqd55);
lmsqd54=log2(msqd54);

%find energy spectra
x1=[2.5,3.5,4.5];
x2=[2,3,4,5];
x3=[2.5,3.5,4.5];
y1=[lmsqd23,lmsqd34,lmsqd45];
y2=[lmsqd22,lmsqd33,lmsqd44, lmsqd55];
y3=[lmsqd32,lmsqd43,lmsqd54];
[aa1, bb1]=polyfit(x1, y1, 1);
[aa2, bb2]=polyfit(x2, y2, 1);
[aa3, bb3]=polyfit(x3, y3, 1);
slope1=aa1(1);
slope2=aa2(1);
slope3=aa3(1);
int1=aa1(2);
int2=aa2(2);
int3=aa3(2);

%PROJECT SPECTRA INTO HIGHER DETAIL SPACES
Elmsqd67=slope2*6.5+int1;
Elmsqd56=slope2*5.5+int1;
Elmsqd66=slope3*6+int2;
Elmsqd77=slope3*7+int2;
Elmsqd76=slope4*6.5+int3;
Elmsqd65=slope4*5.5+int3;

%find e^(log energies) to get back to mean energy
Emsqd67=exp(Elmsqd67);
Emsqd56=exp(Elmsqd56);
Emsqd66=exp(Elmsqd66);
Emsqd77=exp(Elmsqd77);
Emsqd76=exp(Elmsqd76);
Emsqd65=exp(Elmsqd65);

%Do exact same process above, but this time looking for the variances of
%the squared means.
```

```
vsqd22=var(sqd22(:));
vsqd23=var(sqd23(:));
vsqd32=var(sqd32(:));
vsqd33=var(sqd33(:));
vsqd34=var(sqd34(:));
vsqd44=var(sqd44(:));
vsqd43=var(sqd43(:));
vsqd45=var(sqd45(:));
vsqd55=var(sqd55(:));
vsqd54=var(sqd54(:));

lvsqd22=log2(vsqd22);
lvsqd23=log2(vsqd23);
lvsqd32=log2(vsqd32);
lvsqd33=log2(vsqd33);
lvsqd34=log2(vsqd34);
lvsqd43=log2(vsqd43);
lvsqd44=log2(vsqd44);
lvsqd45=log2(vsqd45);
lvsqd54=log2(vsqd54);
lvsqd55=log2(vsqd55);

yv2=[lvsqd23,lvsqd34,lvsqd45];
yv3=[lvsqd22,lvsqd33,lvsqd44,lvsqd55];
yv4=[lvsqd32,lvsqd43,lvsqd54];
[av2, bv2]=polyfit(x2, yv2, 1);
[av3, bv3]=polyfit(x3, yv3, 1);
[av4, bv4]=polyfit(x4, yv4, 1);
slopev2 = av2(1);
slopev3 = av3(1);
slopev4 = av4(1);
intv2=av2(2);
intv3=av3(2);
intv4=av4(2);

Elvsqd56=slopev2*5.5+intv2;
Elvsqd67=slopev2*6.5+intv2;
Elvsqd77=slopev3*7+intv3;
Elvsqd66=slopev3*6+intv3;
Elvsqd65=slopev4*5.5+intv4;
Elvsqd76=slopev4*6.5+intv4;

Evsqd67=exp(Elvsqd67);
Evsqd56=exp(Elvsqd56);
Evsqd66=exp(Elvsqd66);
```

```
Evsqd77=exp(Elvsqd77);
Evsqd76=exp(Elvsqd76);
Evsqd65=exp(Elvsqd65);

%find standard deviation from the variances
Esdsqd67=sqrt(Evsqd67);
Esdsqd56=sqrt(Evsqd56);
Esdsqd66=sqrt(Evsqd66);
Esdsqd77=sqrt(Evsqd77);
Esdsqd76=sqrt(Evsqd76);
Esdsqd65=sqrt(Evsqd65);


% perform the resampling and image enhancement the desired # of times
while k > 0

    % fill detail spaces with artificial noise of appropriate mean and
    % standard deviation (these are still using squared means and standard
    % deviations)
    temp77 = Emsqd77 + Esdsqd77.*randn(128,128);
    temp67 = Emsqd67 + Esdsqd67.*randn(64,128);
    temp76 = Emsqd76 + Esdsqd76.*randn(128,64);
    temp56 = Emsqd56 + Esdsqd56.*randn(32,64);
    temp66 = Emsqd66 + Esdsqd66.*randn(64,64);
    temp65 = Emsqd65 + Esdsqd65.*randn(64,32);

    % turn any negative values into zeros
    temp77(temp77<0)=0;
    temp67(temp67<0)=0;
    temp76(temp76<0)=0;
    temp56(temp56<0)=0;
    temp66(temp66<0)=0;
    temp65(temp65<0)=0;

    %take square root since these were made from squared means
    B(dyad(7),dyad(7)) = .2*sqrt(temp77);
    B(dyad(6),dyad(7)) = .2*sqrt(temp67);
    B(dyad(7),dyad(6)) = .2*sqrt(temp76);
    B(dyad(5),dyad(6)) = .2*sqrt(temp56);
    B(dyad(6),dyad(6)) = .2*sqrt(temp66);
    B(dyad(6),dyad(5)) = .2*sqrt(temp65);

    % do inverse wavelet tranformation
    NewImage = W' * B * W;
```

```
    % add new image to the TotalImage matrix
    TotalImage = TotalImage + NewImage;

    % detect edges
    [E2, threshold] = edge(NewImage, 'canny');
    NewEdge = edge(NewImage, 'canny', 2.2*threshold);
    NewEdgeS = bwareaopen(NewEdge,25);

    % add new edge to the edge CI matrix
    EdgeCI = EdgeCI + NewEdgeS;


    % display one of the runs
    if k == 1
        figure(3);
        colormap gray
        imagesc(NewImage)
        EdgeEx = edge(NewImage, 'canny', 2.2*threshold);
        EdgeExS = bwareaopen(EdgeEx,25);
        figure(4);
        colormap gray
        imagesc(-EdgeExS)
    end

    k = k-1;

end

%find average of total image data; display it
ATotalImage = TotalImage/k2;
figure(5);
colormap gray
imagesc(ATotalImage)

%find average of total edge data; display it
AEdgeCI = EdgeCI/k2;
figure(6);
colormap gray
imagesc(-AEdgeCI)

% take the log of the average edge data; display it
LAEdgeCI(:,:)=log(AEdgeCI(:,:));
figure(11);
colormap gray
imagesc(-LAEdgeCI)
```

Bootstrapping Driver

```
%These are ratios (edge)/(edge + inside area) for cancer cases
cases = [0.6364
        0.3118
        0.4868
        0.1896
        0.2197
        0.2559
        0.2712
        0.4417
        0.4183
        0.4923
        0.4183
        0.4015
        0.4419
        0.5364
        0.5882
        0.3245];

B=100000; %number of bootstraps

%perform boostraps for cancer cases
mbs=[];
for i=1:B
    mbs=[mbs mean(bootsample(cases))];
end

[f1 x1]=ksdensity(mbs);
figure(1)
plot(x1, f1,'r-' , 'LineWidth' ,2)
hold on

%These are ratios (edge)/(edge + inside area) for benign controls
controls =[0.2274
        0.3533
        0.3003
        0.2466
        0.3871
        0.4396
        0.1149
        0.1290
        0.2000
        0.5000
        0.4800
```

```
        0.3378
        0.2678
        0.1834
        0.4375
        0.3620];

%perform boostrap for benign controls
mbsc=[];
for i=1:B
    mbsc=[mbsc mean(bootsample(controls))];
end
[f2 x2]=ksdensity(mbsc);
plot(x2, f2,'b-' , 'LineWidth' ,2)
legend('cases' , 'controls' ,2)

crit = 0.3555; %set threshold
plot([crit crit],[0   14],'g-' )

%find alpha and beta at threshold
beta = sum(mbs < crit)/B
alpha = sum(mbsc>crit)/B

%find F-measure, sens, and spec across a specified range
Fs=[]; Spc=[]; Ses=[];
range = 0.35:0.001:0.36;
for crit =range
   tp = sum( mbs > crit); %number of true positives
   fn = sum( mbs < crit); %number of false negatives
   fp = sum( mbsc > crit); %number of false positives
   tn = sum( mbsc < crit); %number of true negatives

   se = tp/B; %sensitivity
   sp = tn/B; %specificity
   PPV = tp/(tp + fp); %positive predictive value
   Fs = [Fs  harmmean([se PPV])]; %F-measure
   Spc = [Spc 1-sp]; %specificities across entire range
   Ses =[Ses se]; %sensitivities across entire range
end

%Plot F-measure for specified range
figure(2)
plot(range, Fs,'-')
xx=Spc;
yy=Ses;
xlabel('Threshold')
```

```
ylabel('F-Score')

%Plot ROC curve for specified range
figure(3)
plot(xx,yy,'o-')
hold on
plot([0.05 0.09],[0.9 0.94],'r:','Linewidth',2)
xlabel('1-Specificity')
ylabel('Sensitivity')
```

Bootstrap Function

```
function  vecout = bootsample(vecin)
% Bootstrapping from the array "vecin" by random selecting the rows
% Usage
%   vecout = bootsample(vecin)
% Input
%   vecin - nxp data matrix.
%   n - sample size
%   p - dimension of a single observation
% Output
%   vecout - a single bootstrap sample, size n x p.
% Example
%   bootsample([1 2; 2 3; 3 4; 4 5])
%   ans =
%      4     5
%      3     4
%      4     5
%      3     4

[n, p] = size(vecin);
selected_indices = floor(1+n.*(rand(1,n)));
vecout = vecin(selected_indices,:);
```

REFERENCES

[1] ADA Council on Access, P. and on Scientific Affairs, I. R. A. C., "Dental sealants," *Journal of the American Dental Association*, vol. 128, no. 4, pp. 485–488, 1997.

[2] Ahovuo-Saloranta, A., Hiiri, A., Nordblad, A., Worthington, H., and M., M., "Pit and fissure sealants for preventing dental decay in the permanent teeth of children and adolescents," *Cochrane Database of Systematic Reviews*, vol. 3, p. CD001830, 2008.

[3] Altekruse, S., Kosary, C., Krapcho, M., Neyman, N., Aminou, R., and Waldron, W., "Seer cancer statistics review, 1975-2007," 2010.

[4] Azarpazhooh, A. and Main, P., "Pit and fissure sealants in the prevention of dental caries in children and adolescents: A systematic review," *Journal of the Canadian Dental Association*, vol. 74, no. 2, p. 171177, 2008.

[5] Barker, L., Griffin, S., Jeon, S., Gray, S., and Vidakovic, B., "Evidence-based clinical recommendations for the use of pit-and-fissure sealants: A report of the american dental association council on scientific affairs," *Journal of the American Dental Association*, vol. 139, no. 3, p. 257268, 2008.

[6] Beauchamp, B., Caufield, P., Crall, J., Donly, K., Feigal, R., Gooch, B., Ismail, A., Kohn, W., Siegal, M., and Simonsen, R., "Evidence-based clinical recommendations for the use of pit-and-fissure sealants: A report of the american dental association council on scientific affairs," *Journal of the American Dental Association*, vol. 139, no. 3, p. 257268, 2008.

[7] Beiruti, N., Frencken, J., van t Hof, M., and van Palenstein Helderman, W., "Caries preventive effect of resin-based and glass ionomer sealants over time: a systematic review," *Community Dentistry and Oral Epidemiology*, vol. 34, pp. 403–409, 2006.

[8] Birkes, D. and Dodge, Y., *Alternative Methods of Regression*. New York, NY: Wiley, 1993.

[9] Daubechies, I., *Ten lectures on wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia, PA: Society for Industrial and Applied Mathematics, 1992.

[10] Diagnosis and of Dental Caries Throughout Life, M., "National institutes of health consensus statement 2001 march 26-28," *Journal of the American Dental Association*, vol. 132, no. 8, pp. 1153–1161, 2001.

[11] Eckley, I. A., Nason, G. P., and Treloar, R. L., "Locally stationary wavelet fields with application to the modelling and analysis of image texture," *Journal of the Royal Statistical Society Series C-Applied Statistics*, vol. 59, pp. 595–616, 2010.

[12] Efron, B., "Bootstrap methods: another look at the jackknife," *Annals of Statistics*, vol. 7, pp. 1–26, 1979.

[13] Efron, B. and Tibshirani, R. J., *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press, 1994.

[14] El-Naqa, I., Yang, Y., Wernick, M., Galatsanos, N., and Nishikawa, R., "A support vector machine approach for detection of microcalcifications," *IEEE Transactions on medical imaging*, vol. 21, no. 12, pp. 1552–1563, 2002.

[15] Flandrin, P., "Wavelet analysis and synthesis of fractional brownian motion," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 910–917, 1992.

[16] Ghosh, M., Chen, M.-h., Ghosh, A., and Agresti, A., "Hierarchical bayesian analysis of binary matched pairs data," *Statistica Sinica*, vol. 10, pp. 647–657, 2000.

[17] Griffin, S., Oong, E., Kohn, W., Vidakovic, B., and Gooch, B., "The effectiveness of sealants in managing caries lesions," *Journal of Dental Research*, vol. 87, no. 2, pp. 169–174, 2008.

[18] Heath, M., Bowyer, K., Kopans, D., Moore, R., and Kegelmeyer, P., *The Digital Database for Screening Mammography*. 5th International Workshop on Digital Mammography, Toronto, Canada, Madison, WI: Medical Physics Publishing, 2000.

[19] Hsueh, H., Liu, J., and Chen, J., "Unconditional exact tests for equivalence or noninferiority for paired binary endpoints," *Biometrics*, vol. 57, pp. 478–483, 2001.

[20] Jaeckel, L., "Estimating regression coefficients by minimizing the dispersion of residuals," *Annals of Mathematical Statistics*, vol. 43, pp. 1449–1458, 1972.

[21] Kahn, H. and Sempos, C., *Statistical Methods in Epidemiology*.

[22] Kestener, P., Lina, J., Saint-Jean, P., and Arneodo, A., "Wavelet-based multifractal formalism to assist in diagnosis in digitized mammograms," *Image analysis and stereology*, vol. 20, no. 3, pp. 169–175, 2001.

[23] Llodra, J., Bravo, M., Delgado-Rodriguez, M., Baca, P., and Galvez, R., "Factors influencing the effectiveness of sealants: a meta-analysis," *Community Dentistry and Oral Epidemiology*, vol. 21, no. 5, pp. 261–268, 1993.

[24] Lu, Y. and Bean, J., "On the sample size for one-sided equivalence of sensitivities based upon mcnemar's test," *Statistics in Medicine*, vol. 14, pp. 1831–1839, 1995.

[25] Mallat, S. G., *A wavelet tour of signal processing*. San Diego, CA: Academic Press, 1998.

[26] Martin, J., Moskowitz, M., and Milbrath, J., "Breast cancer missed by mammography," *American Journal of Roentgenology*, vol. 37, no. 2, pp. 142–162, 1979.

[27] United States Department of Health and Human Services, "Healthy people 2020," *http://healthypeople.gov/HP2020*, 2010.

[28] University of South Florida, "Digital database for screening mammography," *http://marathon.csee.usf.edu/Mammography/Database.html*.

[29] Mejare, I., Lingstrom, P., Petersson, L., Holm, A.-K., Twetman, S., Kallestal, C., Nordenram, G., Lagerlof, F., Soder, B., Norlund, A., Axelsson, S., and Dahlgren, H., "Caries-preventive effect of fissure sealants: a systematic review," *Acta Odontologica Scandinavica*, vol. 61, pp. 321–330, 2003.

[30] Mickenautsch, S. and Yengopal, V., "Caries-preventive effect of glass ionomer and resin-based fissure sealants on permanent teeth: An update of systematic review evidence," *BMC Research Notes*, vol. 4, pp. 22–39, 2011.

[31] Morikawa, T., Yanagawa, T., Endou, A., and Yoshimura, I., "Equivalence tests for pair-matched binary data," *Bulletin of Informatics and Cybernetics*, vol. 28, pp. 31–45, 1996.

[32] Moss, F., Ward, L., and Sannita, W., "Stochastic resonance and sensory information processing: a tutorial and review of application," *Clinical Neurophysiology*, vol. 115, no. 2, p. 267281, 2004.

[33] Nam, J., "Non-inferiority of new procedure to standard procedure in stratified matched-pair design," *Biometrical Journal*, vol. 48, pp. 966–977, 2006.

[34] Netsch, T. and Peitgen, H., "Scale-space signatures for the detection of clustered microcalcifications in digital mammograms," *IEEE Transactions on medical imaging*, vol. 18, no. 9, pp. 774–786, 1999.

[35] Nevitt, J. and Tam, H. P., "A comparison of robust and nonparametric estimators under the simple linear regression model," *Multiple Linear Regression Viewpoints*, vol. 25, pp. 54–69, 1998.

[36] Nicolis, O., Ramirez-Cobo, P., and Vidakovic, B., "2d wavelet-based spectra with applications," *Computational Statistics & Data Analysis*, vol. 55, no. 1, pp. 738–751, 2011.

[37] Normand, S.-L. T., "Tutorial in biostatistics meta-analysis: Formulating, evaluating, combining, and reporting," *Statistics in Medicine*.

[38] Ramirez-Cobo, P., Lee, K. S., Molini, A., Porporato, A., Katul, G., and Vidakovic, B., "A wavelet-based spectral method for extracting self-similarity measures in time-varying two-dimensional rainfall maps," *Journal of Time Series Analysis*, vol. 32, no. 4, pp. 351–363, 2011.

[39] Scholz, F., "Weighted median regression estimates," *Annals of Statistics*, vol. 6, pp. 603–609, 1978.

[40] sealants in the prevention of tooth decay, D., "National institutes of health consensus statement," *Journal of Dental Education*, vol. 48, no. 2 suppl, pp. 126–131, 1984.

[41] Sidik, K., "Exact unconditional tests for testing non-inferiority in matched-pairs design," *Statistics in Medicine*, vol. 22, pp. 265–278, 2003.

[42] Sievers, G., "Weighted rank statistics for simple linear regression," *Journal of the American Statistical Association*, vol. 73, pp. 628–631, 1978.

[43] Simonsen, R., "Glass ionomer as fissure sealant a critical review," *Journal of Public Health Dentistry*, vol. 56, pp. 146–149, 1996.

[44] Simonsen, R., "Pit and fissure sealant: review of the literature," *Pediatric Dentistry*, vol. 24, pp. 398–414, 2002.

[45] Tango, T., "Equivalence test and confidence interval for the difference in proportions for the paired-sample design," *Statistics in Medicine*, vol. 17, pp. 891–908, 1998.

[46] Theil, H., "A rank-invariant method of linear and polynomial regression analysis," *Indagationes Mathematicae*, vol. 12, pp. 85–91, 1950.

[47] Veitch, D. and Abry, P., "A wavelet-based joint estimator of the parameters of long-range dependence," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 878–897, 1999.

[48] Vidakovic, B., *Statistical Modeling by Wavelets*. New York: John Wiley & Sons, 1999.

[49] Wang, T. and Karayiannis, N., "Detection of microcalcifications in digital mammograms using wavelets," *IEEE Transactions on medical imaging*, vol. 17, no. 4, pp. 498–509, 1998.

[50] Wendt, H., Rouxl, S. G., Jaffard, S., and Abry, P., "Wavelet leaders and bootstrap for multifractal analysis of images," *Signal Processing*, vol. 89, no. 6, pp. 1100–1114, 2009.

[51] Yengopal, V., Mickenaustch, S., Bezerra, A., and Leal, S., "Caries-preventive effect of glass ionomer and resin-based fissure sealants on permanent teeth: a meta analysis," *Journal of Oral Science*, vol. 51, pp. 373–382, 2009.

VITA

Erin Kinzel Hamilton is a PhD student in Bioengineering, within the School of Biomedical Engineering. Erin began her PhD in 2006. She is currently under the co-advisement of Dr. Brani Vidakovic (since 2010) and Dr. Paul Griffin (since 2009) and is working with the CDC as an ORISE Research Fellow. Her current research interests include high frequency data analysis and classification, multiresolution image enhancement, and meta-analysis of clinical trials. Prior to this work, Erin was involved in research investigating cortical reorganization related to visual processes, as well as various usability studies in healthcare settings. She earned her B.S. in Industrial & Systems Engineering from Georgia Tech in 2005.