Complexity Reduction in Image-Based Breast Cancer Care

Markus Harz

DISSERTATION
zur
Erlangung des Grades eines
Doktors der Ingenieurswissenschaften
— Dr.-Ing. —

vorgelegt dem Fachbereich 3, Mathematik/Informatik
der Universität Bremen, Deutschland

Erstgutachter:	Prof. Dr. Heinz-Otto Peitgen
Zweitgutachter:	Prof. Dr. Horst K. Hahn
Externe Stellungnahme Gutachter:	Prof. Nico Karssemeijer
Datum des Kolloquiums:	16. September 2014

To my family.

Acknowledgements

The work I will present in my thesis would not have been possible without the many discussions with my supervisors Heinz-Otto Peitgen and Horst Hahn. Nico Karssemeijer, who acted as an additional examiner, was equally dedicated and provided directions and guidance. I am incredibly thankful for their inspirations and the necessary support to pursue ideas that emerged from the discussions. They answered my questions with patience and endurance and provided insights where I even didn't have a question yet. Their involvement with the topic of my thesis was a never-ending source of motivation.

Joachim Georgii helped me to understand the details of the finite element method and his implementation of a fast multigrid solver I have used. Tobias Böhler contributed his breast registration expertise in the biopsy preparation work. Lei Wang's segmentation algorithms provided essential ingredients to separate the breast tissue from adjacent tissues. Andrè Homeyer was a source of solid knowledge and ideas when I needed to discuss machine learning and validation approaches. Felix Ritter and I developed the ideas underlying the breast MRI reading prototype. We designed the interaction schemes and jointly worked with the radiologists to understand their image reviewing workflow. Hendrik Laue was always a discussion partner with a broad knowledge extending from cancer biology to contrast agent kinetic modeling, as was Fabian Zöhrer for several aspects in human-computer interface design and many other aspects. Ina Kompan, my office mate physically, and my soul mate in terms of thesis writing, shared thoughts and tea with me whenever we were in the office on the same day. In the end, we will have gone through the thesis writing phase almost synchronously. I am thankful to many others at Fraunhofer MEVIS, notably the 12:30pm Mensa group who were a spiritual support throughout the long writing phase. The encouragement created through the familial spirit in our working environment made it possible to combine the daily working routine with the pursuit of a dissertation.

Supervising students also taught me a lot. I had the pleasure to meet several gifted students. I developed the methods for lesion detection and classification in breast MRI together with Abhilash Srikantha and Jennifer Loose. Suzan Akbey pursued the biopsy preparation algorithm implementation and evaluation. Simon Benten programmed large parts of the iOS code for the breast MRI reading prototype and also contributed to the workstation server.

The work would not have been possible without many discussions with radiologists and clinical experts. Among them, Kathy Schilling takes a special place with her experience, explorative mind and not least her hospitality. During six month that I was allowed to visit her team in Boca Raton, I gained invaluable insights. Discussions with her and her husband Joseph Colletta, a breast surgeon, provided me with a solid working knowledge of optimal therapeutic breast cancer care. Ritse Mann, Nijmegen, also influenced my work with comments, encouragements, and most notably with a never ending stream of image data of all kinds. Roland Holland, a renowned breast pathologist also working in Nijmegen, was an inspiration for my concerns with ductal carcinoma in situ (DCIS). I am grateful for the opportunity to meet many more experts: Gillian Newstead, Robert Schmidt, Gregory Karczmar, and many others at the University of Chicago have seen some of the work evolve and shared their comments with me. Ulrich Bick, Charitè Berlin, made it possible for us to use a larger dataset of hand-segmented breast lesions.

Parts of the research that led to results in presented in this thesis European Community's Seventh Framework Programme(FP7/2007-2013) under Grants No. 224538 (HAMAM), and No. 601040 (VPH-PRISM).

I dedicate this thesis to my family: my wife Katia, who's love, originality and outspokenness formed the pillars required for an undertaking as this one, my children Phileas and Jerenike, who are the true masters in the art of unobstructed observations, of being captivated by detail, and of making sense of complex things; and my parents Gottfried and Ursula Wenzel, who fed

my curiosity since I was a child and had a steadfast belief in my capabilities to achieve my goals. They all can't be thanked enough for enduring me in the past time.

Abstract Breast cancer care is complex. The diversity of tumors and cancers of the breast creates demand for a plenitude of imaging modalities and appropriate evaluation techniques, requires the choice between many therapy options, and can make it a challenge to select among diagnostic tools and curative approaches. This thesis provides contributions in three areas, where particular challenges prevail, some more and some less complex.

First, in the clinical scenario of diagnostic image evaluation, the thesis focuses on algorithms to provide computer-aided detection and diagnosis of mass lesions and non-mass-like contrast uptake in breast magnetic resonance imaging (MRI). The application of novel spatio-temporal texture features is explored for the task of automated mass diagnosis. For enhancement patterns other than masses, a method for the detection and delineation of arbitrarily shaped regions is proposed that is based on the bilateral symmetry of the breast's uptake characteristics. From the segmented regions, a set of task-specific features helps to differentiate benign and malignant lesions with high specificity. The generality of the symmetry metric makes it applicable to a variety of other clinical tasks. Both contributions are fully automated and can be integrated into clinical decision support systems.

Next, turning to user interfaces and interaction in diagnostic image analysis, a novel breast MRI reading paradigm will be introduced, which is exemplified with a breast MRI reading workstation prototype dubbed the MRiPad, alluding to the iPad being used for MRI reading. Instead of mouse and keyboard, the MRiPad is operated using multi-touch gestures. The MRiPad implements a hanging protocol based workflow for the reading of breast MRI examinations complemented by a context sensitive selection of interactive, gesture-operated diagnostic tools. Further, a iPad-based prototype for mammography screening is proposed that attempts to combine the ad-hoc image perception available in light boxes and alternators for film screen mammography with the structured reading provided in digital mammography reading workstations. Quantitative and qualitative evaluations underscore the belief that gesture-based interaction on mobile devices may be a viable alternative to conventional user interface paradigms. Beyond the two prototype implementations, the mobile device is introduced as a personal key of hospital personnel, acting differently depending on location, user, and task. The scope of this general concept opens interesting applications in other areas within and outside hospitals.

Last, contributions to the physics-based modeling of breast tissue deformations will be introduced. Two clinical problems are presented to highlight the scope and applicability of the existing framework and its extensions proposed in this thesis: the prediction of the breast shape change from the prone to the supine positioning to support lesion location visualization for surgery planning and other tasks, and the prediction of the breast's shape change due to the compression in a MRI breast biopsy device. This may aid radiologists to target lesions with higher confidence. Contributions will be presented that extend a linear elastic finite element model based deformation simulation in a way that maintains the high performance of the simulation. Specifically, they allow for the approximation of non-linear elastic behavior, the emulation of sliding governed by friction, and the morphing of surfaces of elastic bodies.

Title image: A breast cancer cell imaged by a scanning electron microscope. A spiculated structure of the cell's surface characterizes the cell. Cancer cells are better characterized by their microbiological traits, but current research with scanning electron microscopes looks at cancer cell interactions with changing environments. Image of National Cancer Institute, U.S.A., released without restrictions.

List of Figures image: The first documented breast cancer case dates back to about 1600 BC, reported in Egyptian writings. The image shows ushabtiu, figurines which were in ancient Egypt thought to aid the deceased in their work in afterlife. (Image licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license.)

Zusammenfassung

Brustkrebs-Fürsorge ist komplex. Die Verschiedenartigkeit von Tumoren und Krebsarten der Brust erzeugt die Notwendigkeit, ebenso verschiedenartige Bildgebungstechniken für eine angemessene Befundung einzusetzen. Die beteiligten Ärzte müssen zwischen einer Vielzahl an Therapieoptionen wählen, und es kann eine Herausforderung sein, die optimale Wahl der eingesetzten Mittel und des kurativen Ansatzes zu treffen. Die vorliegende Arbeit trägt in drei herausfordernden Bereichen von verschiedener Komplexität zu Entscheidungsfindung der Ärzte bei.

Zunächst wird das klinische Szenario der diagnostischen Bildauswertung betrachtet, wo Algorithmen zur computer-unterstützten Detektion und Charakterisierung von verdächtigen Regionen in kontrastmittelgestützter Magnetresonanztomographie (MRT) im Zentrum stehen. Zunächst werden Textur-Merkmale zur Differenzierung massenartiger Funde vorgestellt und untersucht, die im Unterschied zu bekannten Textur-Merkmalen alle drei räumlichen und die zeitliche Dimension berücksichtigen. Für Anreicherungsmuster, die segmental oder diffus sind, wird anschliessend eine Methode entwickelt, die beliebig geformte Regionen aufgrund der bilateralen Asymmetrie segmentiert. Auf Basis des Segmentierungsergebnisses wird für jede gefundene Region ein Satz von Merkmalen errechnet, die abermals zur Differenzierung gutartiger von bösartigen Läsionen dienen. Der symmetriebasierte Detektionsansatz ist allgemein formuliert, so dass er für eine Zahl anderer Felder einsetzbar ist. Die vorgestellten Verfahren sind vollständig automatisiert und können somit in einem computerbasierten Entscheidungsunterstützungssystem eingesetzt werden.

Anschließend wendet sich die Arbeit der Interaktion in der bildbasierten Diagnostik zu. Ein neuartiges Paradigma zur Befundung von Brust-MRT wird eingeführt und ein Demonstrator vorgestellt, den wir MRiPad nennen, was darauf anspielt, dass ein iPad eingesetzt wird, das Maus und Tastatur vollständig ersetzt. Mehr-Finger-Gesten erlauben die Navigation durch ein Bildanzeige-Protokoll, das durch kontextsensitive Werkzeuge zur Detailanalyse ergänzt wird. Parallel dazu wird ein ebenfalls durch ein iPad gesteuerter Demonstrator vorgestellt, der sich der effizienten Arbeit im Mammographie-Screening-Szenario widmet und darauf abzielt, die intuitive Interaktion, die mit filmbasierten Mammogrammen am Lichtkasten möglich war, durch eine Gestensteuerung umzusetzen. Qualitative und quantitative Evaluierungen beider Demonstratoren unterstreichen die Annahme, dass gestenbasierte Interaktion auf Mobilgeräten eine Alternative zu den konventionellen Benutzerschnittstellen darstellen könnten. Über die beiden Demonstratoren hinausgehend wird ein Konzept skizziert, in dem das Mobilgerät als persönlicher Schlüssel für die Klinik-Informationsinfrastruktur dargestellt wird, das unterschiedliche personalisierte Informationen, aber auch Interaktionen ermöglicht, die an Benutzer, Ort und Aufgabe dynamisch angepasst werden.

Zuletzt werden Beiträge zur realitätsnahen Simulation von Brustgewebe-Deformationen präsentiert. Zwei klinische Probleme werden vorgestellt, die die Reichweite und Anwendbarkeit der vorgeschlagenen Erweiterungen eines bestehenden Frameworks ausloten. Das ist zum einen die Vorhersage der Veränderung der Brustform beim Übergang von der Bauchlage in die Rückenlage, was eine Anwendung in der Visualisierung zum Zwecke der chirurgischen Operationsplanung hat. Zum anderen wird die Formänderung, die durch die Kompression der Brust in einer MRT-Biopsiespule erfolgt, simuliert. Dies kann Radiologen helfen, ein Biospieziel mit größerer Sicherheit zu identifizieren. Die vorgeschlagenen Algorithmen erweitern ein Finite-Elemente-Modell, dem ein linear-elastisches Materialgesetz zugrunde liegt, so daß in begrenztem Umfang nicht-lineare Materialeigenschaften sowie zusätzlich friktionale Verschiebungen realistischer abgebildet werden können, ohne die hohe Performanz des Frameworks zu beeinträchtigen.

Contents

1 Complexity in Breast Cancer Care 1

1.1 Breast Cancer Care 1
1.1.1 Complexity 5
1.1.2 Cancer 7
1.1.3 Complexity in Breast Cancer Care 14
1.1.4 Ways to Approach Complexity 19

1.2 Imaging Breast Cancer 23
1.2.1 The Mammary Gland 23
1.2.2 Epidemiological Relevance 24
1.2.3 Imaging Methods 26

1.3 Complexity in Breast Cancer: Prospects 36

1.4 Thesis Overview 37

1.5 Data Used in the Thesis 37

2 CADe/CADx-Tools for Breast MRI 39

2.1 Machine Learning 41
2.1.1 Brief History and Terminology 41
2.1.2 Common Pitfalls 48
2.1.3 Recommended Approaches 61

2.2 DCIS from a Different Perspective 65
2.2.1 Biomarkers and Diagnostic Criteria 65
2.2.2 Non-mass CADx Requirements 69

2.3 Symmetry in Non-Mass CAD 71
2.3.1 Non-Mass Lesion Characterization 72
2.3.2 Automated Characterization 80
2.3.3 The Predictive Value Of Symmetry 83

2.4 4D Texture of Mass Lesions 85
2.4.1 Contribution 85
2.4.2 Texture Feature Classification 87

2.4.3 Comparison and Perspectives . . . 106
2.5 Future of CADx 107

3 Workflow Support for Breast Cancer Diagnosis . . . 109
3.1 Mobile Devices in Medical Environments 111
3.1.1 Division of Functionality . . . 114
3.1.2 Spread of Mobile Devices . . . 116
3.2 Clinical Breast MRI Workflow 118
3.2.1 Screening and Diagnostic Breast MRI . . . 119
3.2.2 Interventional Breast MRI . . . 122
3.3 Multi-Touch Based Breast MRI Workflow Support 124
3.3.1 Designing Interface and Interaction . . . 124
3.3.2 Techniques . . . 125
3.3.3 Results . . . 131
3.4 Multi-Touch Mammography Screening 137
3.4.1 Mammography Screening History . . . 137
3.4.2 Gesture Controlled Workstation . . . 140
3.4.3 Evaluation . . . 143
3.4.4 Conclusion . . . 146
3.5 Discussion 147

4 Intervention Support . . . 151
4.1 Biomechanical Simulations 153
4.1.1 General Approach . . . 153
4.1.2 Roots . . . 153
4.1.3 Finite Element Discretization . . . 159
4.2 Finite Element Method in Breast Care 161
4.2.1 Deformation Simulation Framework . . . 162
4.2.2 Interactivity . . . 164
4.2.3 Physical Plausibility . . . 166
4.3 Applications in Breast Cancer Care 176
4.3.1 Surgery . . . 176
4.3.2 Biopsy . . . 185
4.4 A Perspective on Applications 200

Conclusion . . . 203

A Complexity: Example . . . 205
A.1 Complex System, Simple Behavior 205

B iPad MRI Reading Workflow Prototype . . . 209
B.1 Questionnaire for Workflow Evaluation 209
B.2 Questionnaire for Qualitative Evaluation 211

C Implementations 215
C.1 Experiments in Feature Subset Selection Bias 215
C.2 Machine Lerning on DCIS Data Set 218
C.3 Histogram Operations 220
C.3.1 Histogram Stretching 220
C.3.2 Histogram Equalization 220

Bibliography 221
Books 221
Articles 224
Conference Proceedings and Others 244

Index 251

List of Figures

1.1 Breast shape variability . . . 2
1.2 Responding in systems from simple to chaotic . . . 5
1.3 Breast cancer cell . . . 7
1.4 Hallmarks of cancer . . . 10
1.5 Development of DCIS . . . 12
1.6 Screening mammography instrumentation . . . 16
1.7 Breast anatomy . . . 25
1.8 Mammogram Example . . . 28
1.9 Breast coil . . . 29
1.10 Contrast Enhancement Curves . . . 30
1.11 Motion Correction . . . 31
1.12 DCE-MRI Example . . . 32
1.13 HiSS example images . . . 33
1.14 Machine learning score histogram . . . 34

2.1 SVM terminology . . . 45
2.2 SVM decision boundaries . . . 47
2.3 Bias and variance . . . 50
2.4 Bias-variance trade-off . . . 51
2.5 Confusion matrix . . . 53
2.6 Example of a ROC curve . . . 56
2.7 Machine Learning tools . . . 60
2.8 Nomogram . . . 62
2.9 DCIS examples . . . 67
2.10 Mammary papilla detection . . . 74
2.11 Evaluation of nipple detection algorithm . . . 75
2.12 Performance of the nipple detection algorithm on example cases . . . 76
2.13 Projection distance . . . 79
2.14 DCIS detection/characterization results for two cases . . . 82
2.15 Directions of a co-occurrence matrix . . . 88
2.16 Co-occurrence matrix example . . . 89
2.17 Co-occurrence matrix in time . . . 90
2.18 Examples of benign and malignant lesions . . . 92
2.19 Results from predictions of all features on all lesions . . . 94

2.20 ROC curves of LRA-BE model . . . 95
2.21 Feature selection procedure. . . . 97
2.22 Classification results . . . 99
2.23 Texture classification ROC curve . . . 99
2.24 Bias of Random Forest on random data . . . 104
2.25 Unbiased classifier performance for 4D texture CAD . . . 105

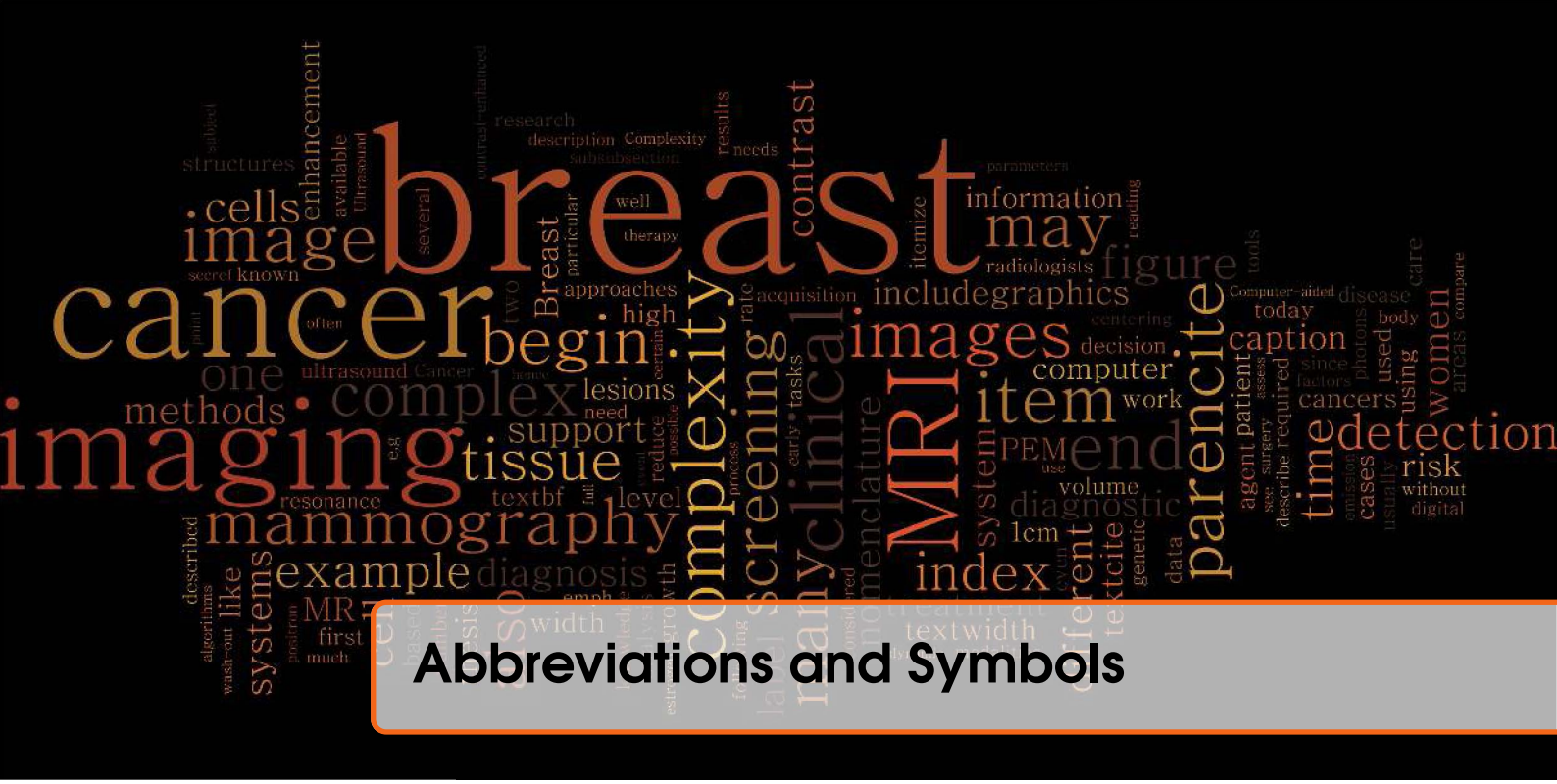
3.1 Interaction paradigm . . . 112
3.2 Information confluence . . . 113
3.3 Mobile devices . . . 114
3.4 iPad screens used for breast MRI . . . 126
3.5 Workstation and setup, and available tools . . . 127
3.6 iPad gestures controlling the workstation . . . 127
3.7 The iPad tool menu . . . 129
3.8 Measuring a distance with a gesture . . . 129
3.9 Device collaboration . . . 131
3.10 iPad MRI workflow prototype during interaction demonstration. . . . 133
3.11 iPad patient overview screen . . . 134
3.12 Qualitative assessment of gesture usability . . . 135
3.13 Quantitative evaluation results . . . 136
3.14 Alternator used in pre-digital mammography screen reading. . . . 139
3.15 Special keypad for mammography screening . . . 141
3.16 Schematic of prototype hanging protocol. . . . 142
3.17 Gestures: Advancing, roaming, zooming, history. . . . 143
3.18 Example of a visual sign in the study . . . 145

4.1 Facsimiles of Hooke and Newton treatises . . . 154
4.2 Energies and Deformation . . . 155
4.3 Poisson ratio . . . 157
4.4 Deformation simulation in MeVisLab: Overview . . . 163
4.5 Sensor controls deformation . . . 165
4.6 Result of shape-based update . . . 168
4.7 Influence of different stiffness update mechanisms . . . 170
4.8 Stress Metrics Comparison . . . 172
4.9 Breast model deformation by gravity . . . 173
4.10 Deformation for different parameter settings . . . 179
4.11 Sliding simulation on volunteer data . . . 181
4.12 Morphing a volume mesh to a target surface . . . 182
4.13 Morphing the volunteer volume mesh . . . 182
4.14 Outline of the simulation steps arranged in a feedback loop . . . 183
4.15 Illustration of vacuum-assisted core needle biopsy . . . 186
4.16 Two systems for MRI-guided biopsy . . . 187
4.17 Examples of peri-interventional scans . . . 188
4.18 Deformation fields for the two non-linear registrations . . . 192
4.19 Correspondence quantification . . . 194
4.20 Landmark correspondence evaluation . . . 196
4.21 Biopsy deformation results (projections) . . . 197

A.1 Cellular Automaton, wire and electron. . . . 205
A.2 Cellular Automaton, close-up . . . 206
A.3 CA prime number generator. . . . 207

B.1 Evaluation forms . . . 213

C.1 Bias of Naïve Bayes on random data . . . 217



Abbreviations

ABVS	Automated breast volume sonography, page 35
ADH	Atypical ductal hyperplasia, page 69
AIC	Akaike information criterion, page 95
ANN	Artificial neural network, page 43
ANOVA	Analysis of variance, page 138
BCT	Breast conserving therapy, page 124
CA	Cellular Automaton, page 207
CA	Contrast agent, page 29
CAD	Computer-aided detection/diagnosis, page 22
CADe	Computer-aided detection, page 17
CADx	Computer-aided diagnosis, page 17
CC	Cranio-caudal, page 27
CNB	Core needle biopsy, page 188
COSY	COrrelation SpectroscopY, page 29
CV	Cross validation, page 58
DCE	Dynamic contrast enhanced, page 4
DCE-MRI	Dynamic contrast-enhanced magnetic resonance imaging, page 17
DCIS	Ductal carcinoma in situ, page 15
DICOM	Digital Imaging and Communications in Medicine, page 115
DNA	Deoxyribonucleic acid, page 8
DSS	Decision Support System, page 21
FDG	Fluorodeoxyglucose, page 35
FEM	Finite element method, page 155
FFDM	Full-field digital mammography, page 22
FID	Free induction decay, page 33
FSM	Film-screen mammography, page 139
Gd-DTPA	Gadolinium-Diethylene-Triamine-PentAcetate, page 29
GLCM	Gray-level co-occurrence matrix, page 87
GPU	Graphics processing unit, page 165

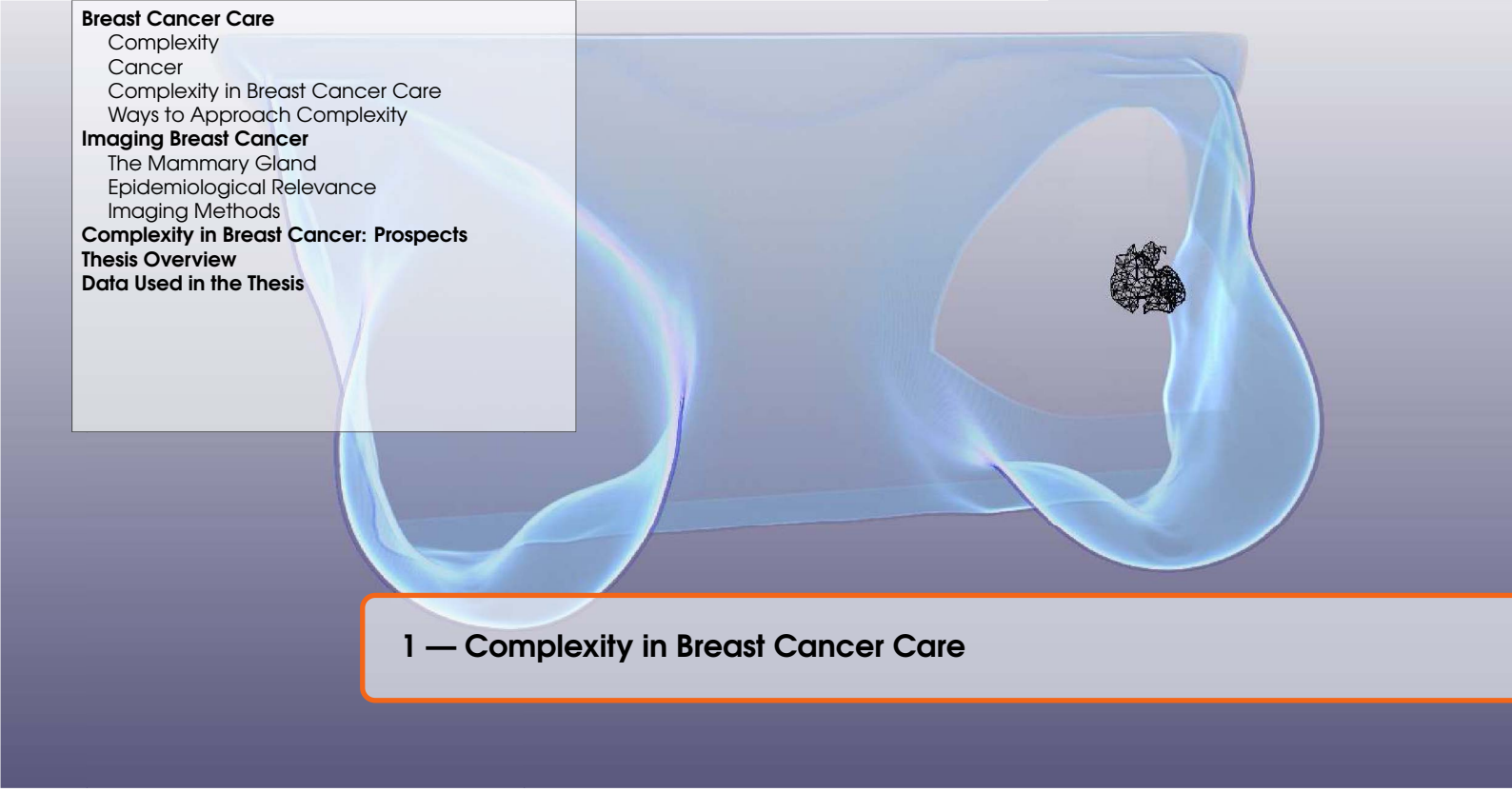
HiSS High spectral spatial resolution MRI, page 33
IDC Invasive ductal carcinoma, page 70
IP Internet Protocol, page 133
JPEG Joint Photographic Experts Group, page 133
LBP Local binary patterns, page 107
LCIS Lobular carcinoma in situ, page 39
LOO Leave-one-out (in cross validation), page 95
LRA-BE Logistic regression analysis with backward elimination, page 95
LTP Local ternary patterns, page 107
MARS Multivariate adaptive regression splines, page 102
MCC Matthews correlation coefficient, page 55
MIP Maximum intensity projection, page 31
ML Machine learning, page 41
MLO Medio-lateral oblique, page 27
MRI Magnetic resonance imaging, page 3
MRSI Magnetic resonance spectroscopic imaging, page 29
OBIA Object-based image analysis, page 72
PACS Picture archiving and communications system, page 115
PDE Partial differential equation, page 161
PE Peak enhancement, page 87
PEM Positron emission mammography, page 35
PET Positron emission tomography, page 35
PPV Positive predictive value, page 96
PR Pattern recognition, page 41
QR code Quick response code, page 127
RF Radio frequency, page 27
ROC Receiver operating characteristic, page 55
SGLD Spatial gray level dependence matrix, page 87
SVM Support vector machine, page 44
TCP Transmission Control Protocol, page 133
TDLU Terminal duct lobular units, page 65
TTP Time to peak, page 87
US-SWI Ultrasound shear wave imaging, page 34
WLAN Wireless local area network, page 117, 118

Symbols

$\mathcal{D}$ Context dictionary, page 78
$\mathcal{M}$ Suspicion map, page 78
$\mathbf{c}$ Stiffness tensor, page 158
χ Displacement function, page 160
$\boldsymbol{\varepsilon}$ (Cauchy) strain tensor, page 158
E Young's modulus, page 158
F Force (Hooke's law), page 158
I Image in $\mathbb{N}^4$, page 76
$\mathbf{K}$ Global stiffness matrix, page 163
$\mathbf{K}^e$ Element stiffness matrix, page 162
k Spring constant (Hooke's law), page 158

$\mathbf{M}$ Material matrix, page 159
ν Poisson ratio, page 158
N Shape function, page 162
$\boldsymbol{\sigma}$ Cauchy stress tensor, page 157
$\mathbf{u}$ Displacement field, page 160

Breast Cancer Care
Complexity
Cancer
Complexity in Breast Cancer Care
Ways to Approach Complexity
Imaging Breast Cancer
The Mammary Gland
Epidemiological Relevance
Imaging Methods
Complexity in Breast Cancer: Prospects
Thesis Overview
Data Used in the Thesis



1 — Complexity in Breast Cancer Care

Breast cancer care is arguably a complex matter. The clinical goal in breast cancer care is definitive treatment of the patient, meaning that after treatment, the patient doesn't experience a relapse. This puts the emphasis on those tasks directly influencing treatment decision making, since optimizing treatment use is supposed to benefit patients most (Mandelblatt et al. 2013).

The work presented in this thesis will approach the complex nature of breast cancer care from different perspectives, and will propose approaches in parts of this medical field that help to increase the efficiency of work and are expected to benefit treatment decision making. This comprises tasks in particular in the areas of image-based detection and diagnosis of breast lesions, and intervention planning based on radiological images.

The introductory chapter summarizes the clinical tasks involved in breast cancer care with an emphasis on image-based decisions and procedures. It outlines and defines the scope of clinical breast cancer care as it will be understood in this thesis, providing insights into cancer biology, breast cancer imaging, and image-based decision making. Special attention is paid to the complex nature of all involved tasks, and it will be seen how not only computational results may contribute to a reduction of task complexity, but how changes in the human-computer interface design may help to improve work efficiency.

1.1 Breast Cancer Care

Breast cancer care, in general, involves many clinical disciplines: Clinical oncologists assess the genetic predisposition and family history, asking for the risk of a woman to develop breast cancer, and are the experts to device chemotherapy regimen tailored to the individual women and breast cancer case. Radiologists and their technical assistants image the breasts with various means and aims from screening to image-guided interventions (biopsies). Radiotherapists and surgeons may specialize in the treatment of breast cancer. In some countries, gynaecologists perform the same or some of the interventional procedures that specialized breast surgeons perform. In addition, pathologists need to provide diagnoses based on the particular cellular alterations seen in the many types of breast cancers, contributing an important part to treatment decision making. Treatment decisions are nowadays often discussed in regular multi-disciplinary tumor board meetings, where all the above specialists congregate.

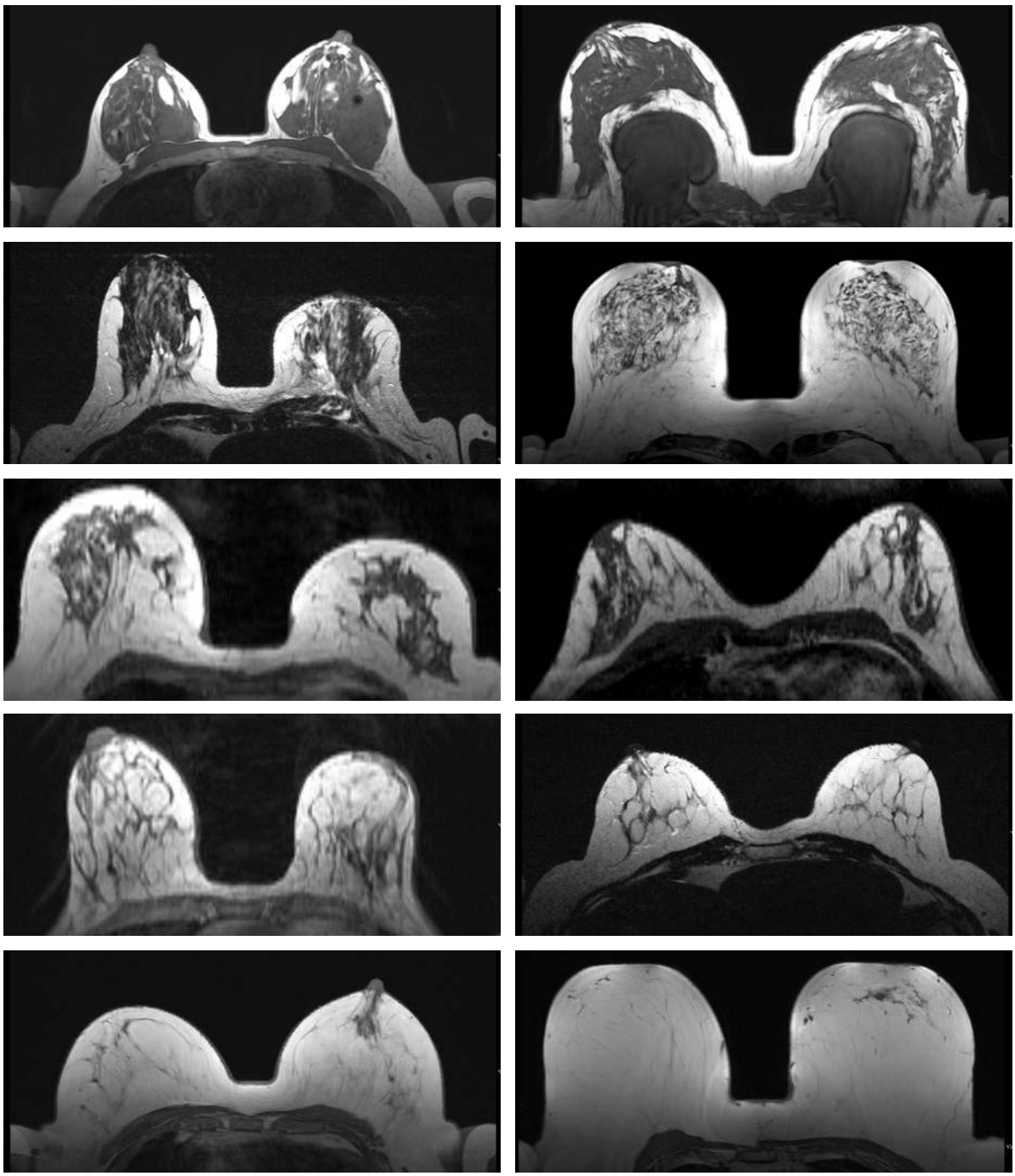


Figure 1.1: Breast shape variability. The outer shape as well as the internal structure of breasts and their deformability is highly variable, which has prevented the establishment of successful statistical appearance models and also limits the generation of many other models like for example biomechanical and functional ones. The figure shows central slices of breast MRI volumes. Bright voxels in the breast correspond to fatty tissue, dark voxels are parenchyma. The top right image depicts implants. Note the variability of shape and texture between examples, and within one example. The volumes are loosely sorted by breast density, dense breasts on top, fatty replaced breasts to the bottom of the figure.

Breast cancer care seen from the perspective of the healthy women looks different, and it begins with images. At some age in her life, she will be invited to participate in a breast cancer screening program, aiming at early detection of small tumors. She will be imaged in regular intervals, and in case a change is detected, she will be recalled to have diagnostic images taken to substantiate the suspicion. In case imaging and a biopsy confirm a cancerous lesion, she will meet the clinical oncologist and, depending on her eligibility, physicians of the therapeutic disciplines. Her treatment begins after the treatment decision has been taken, and will be monitored by imaging.

Breast cancer, seen from the perspective of the information processing sciences, is characterized by a high grade of inter-subject variability in visible growth patterns and aggressiveness, which is likely due to the high degree of variability of the host organ, and which reflects in the information that can be obtained by imaging. In this regard, the breast is different compared to many other organs of the body. From subject to subject, breasts not only show very different sizes, but more importantly, a very variable internal tissue composition, which in addition changes per subject with aging and the menstrual cycle. It is noteworthy that some amount of the defining elements on the macroscopic, the molecular and the genetic level is not directly visible to radiological imaging methods, compare Sec. 1.2.1, p. 23. The composition variability hampers efforts to build statistical knowledge, which is a common approach for computer based image interpretation in other organs like for example the brain. Statistical models, also called atlases, allow to interpret unseen data with respect to the "general model", which offers efficient and elegant ways to tell normal cases from diseased ones, locate abnormalities, describe deviation from the norm, etc. Computer-aided detection and characterization can hence be built upon such atlases.

For the breast, however, few approaches have been described, and none have had remarkable influence, except to support the segmentation of fibroglandular tissue, like in S. Wu et al. (2012). The variability of the internal breast tissue composition also impedes attempts to model its biomechanical behavior which depends on knowledge of tissue properties like stiffness. This is a research direction fed from the desire to aid surgeons visualize target movement, or to help plan cosmetic outcomes after partial mastectomy with plastic surgery.

This thesis' scope will be limited to those aspects of clinical breast cancer care that involve images: the detection, diagnosis, and treatment of breast lesions. The main clinical discipline in this regard is radiology. Hence, in the remainder of this thesis, the tasks of radiologists will be treated, and links to other disciplines considered only where required.

Radiologists face growing numbers of images, and besides, they face a growing number of modalities, physical machines generating images, among which they have to choose the most suitable ones for a given patient and clinical question. There will be additional information on the patient from questionnaires, for example regarding her own and her family's history of breast cancer. While today, radiologist may still be capable to have all information on a patient in mind, and may know the appropriate choice among all available imaging and therapy options, options and information items both grow in number. New genetic tests become available and get cheaper and will in the future be another factor to include in decision taking. New imaging modalities like for example contrast-agent-less dynamic breast magnetic resonance imaging (MRI) may allow increased numbers of MRI procedures. Specialized imaging may reveal additional information in selected populations, or in certain indications, and so on. While one breast MRI series[1] may appear to be comparatively simple to understand and interpret by trained radiologists, there are many types of breast alterations that are either hard to distiguish, or cannot be distinguished at all from only one MRI contrast. Hence, multiple contrasts are

[1]In the context of radiological images, the *image* refers to one two-dimensional image. A number of such images that are regularly sampled along a path through a volume are called a *series* or a *volume image*. In this thesis we sometimes refer to series, even with more than three dimensions, also by the term "image", which is common practice in medical image analysis. Series are in a MRI protocol gathered into a *study*.

acquired in one study, and correlation with other modalities, such as mammography, needs to be made. However, all series in one study together bear vastly more information, and it has to be integrated spatially and with all additional information — to a large extent, this information integration today happens in the mind of the clinician.

Even single imaging modalities on their own reveal several challenges. Some modalities are utilized in clinical screening, like e.g. mammography, but also MRI is employed in certain risk groups. Images are acquired to follow the women over the years, but even if they are stored in a central system, the level of ability to correlate one patient's images over time may severely limit the information a radiologist may extract, and the task of temporal and spatial correlation is tedious, error-prone, and time-consuming. A time-spanning correlation, however, is required to assess changes from one visit to the next.

Considering the example of breast MRI, radiologists are facing a growing number of MRI pulse sequences they can choose from, and which display different tissue characteristics with different gray level contrasts, e.g. emphasizing the water or the fat component. Usually, a selection of such pulse sequences is collected into the so-called scanning protocol for breast examinations. Besides high-resolution anatomic images, the most crucial component in any breast MRI scanning protocol is today the dynamic contrast enhanced (DCE) imaging sequence (see Sec. 1.2.3). It is like all other parts of the clinical breast scanning protocol subject to active research. New variants with increased temporal or spatial resolution emerge frequently. Others propose novel ways to acquire multiple contrasts in one acquisition, etc. Keeping track of those developments, and judging their clinical benefit or field of potential application requires continuous education.

Today, imaging protocols are considered optimal if they allow to acquire a maximum of diagnostically relevant different contrasts in a minimal amount of time per patient. Optimizing this protocol by reducing scanning time is an important economical goal, because with a high patient throughput, breast MRI operates more cost-efficient. Beyond economic considerations, such optimized imaging protocols are also clinically valuable, since only a high patient throughput — beginning with short acquisition times and including patient-specific optimal choices of contrasts to gain the information offers the highest probability of detection success — may allow to offer breast MRI screening on a broader basis. Consequently, many research efforts strive to reduce the time required to measure a given contrast. This benefit in speed can then either be turned into a higher throughput of patients, or into a protocol comprising more contrasts per patient. Either way, the total number of MRI examinations that are performed increases.

Assuming an optimized breast MRI protocol that results in a set of imaging contrasts, the review of these acquired images needs to keep pace with the speed in which patients are imaged. Employing more radiologists to review the images is one option, which in many cases will be considered too expensive. The alternative is a reduction of reading time per MRI examination, which may for example be enabled through carefully designed computer support (see Sec. 3.3).

From the imaging data collected for one woman in one visit, additional information may be harvested if not only the exams of each imaging modality are considered individually (e.g. first assess the mammograms, then the MRI series). Instead, integrated and spatially correlated image analysis and presentation of several modalities could help to identify corresponding locations more reliably, and could help to derive novel features from combined data in computer-based analyses.

Finally, numerous difficulties arise from the interdisciplinary setting in breast cancer care. Many types of breast cancer are known, and it is also well established knowledge that there are delicate and subtle differences that need to be respected when treatment options are considered (compare Sec. 2.2). Surgery is but one option besides chemotherapy, radiotherapy, cryoablation, and others. In some cases, combinations of the above are in order, but regardless of the choice,

thoughtful diagnostic imaging and biopsy-proven pathology are required to define the therapeutic path to take. Therapy monitoring then needs to find reliable estimates of cancer burden over time, so that success or non-responder status are established early enough to adjust the treatment regime.

Many of these remarks not only apply to breast cancer care, but to image-based oncology in general. Breast cancer, however, is particular in some aspects: it affects many more women than any other cancer, and it is to today's knowledge largely independent of personal conductance of life apart from the general influences detrimental to health, and from the well-known cancer risk factors smoking, alcohol abuse, and obesity. This means that targeted approaches to breast cancer prevention that aim at life style or life circumstances are unavailable.

1.1.1 Complexity

Complex, complicated, simple, and chaotic systems have been conceptually distinguished by Snowden (2000). According to his taxonomy, the ways to act in such systems are inherently different, and are characterized by according principles (compare Fig. 1.2). Acting in simple systems is possible using best practices, responding to the system after a categorization of sensed information. In complicated systems, more thorough analysis of information is required, and in complex systems, the cause-effect relationship is not known until probed, sensed, and post hoc explained. Chaotic or disordered systems are not of interest in our considerations.

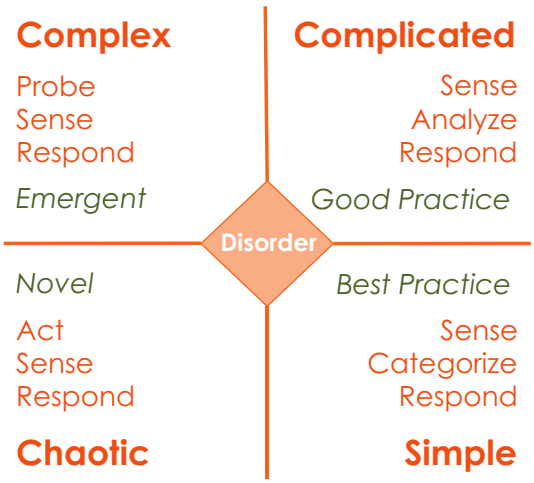


Figure 1.2: Responding in systems from simple to chaotic. After: Snowden (2000)

Systems that can be described, albeit in sometimes very involved fashion, in descriptions, algorithms, or workflows, are usually called "complicated". To follow the algorithms may require many experts and trained personnel, but a complicated system is typically independent from its environment (or the dependencies can be anticipated and charted in action-reaction prescriptions). Some examples for complicated, but not complex actions are to fly a rocket to the moon, or to conduct a surgical procedure. Complicated procedures can be automated to a certain degree, and usually they are.

Complexity is not to be mistaken for only a higher level of complicatedness. Complexity, as some authors state, is a multi-level effect. There have been many attempts to define complexity. There are also many models for real world phenomena that claim to be complex, and there is even a growing research direction that emerged in the aftermath of the chaos and fractals boom in the 1980s. The scientific community concerned with *complex systems analysis* is interdisciplinary and diverse, and its branches name themselves after the aspect they are considering.

To define complex systems, an often used way is to describe common characteristics of systems that are considered complex. Following Flake (1998), call a system *complex* if it is

1. parallel, containing multiple similar simple basic building blocks;
2. sequentially, iteratively, or recursively repeating actions and reacting on feedback of its environment;
3. adaptive, learning from the feedback for example by evolution.

This may be called a *structural* approach to complexity. Alternatively, a complex system can be described by the following questions, following Lloyd (2001). He categorized existing approaches that attempt to quantify the level of a system's complexity. He structured the metrics of complexity that he collected into three categories, and summarized them with one key question per category. A system is complex, if the following metrics yield high values.

1. Is it hard to describe?
   Repetitive patterns are easier to describe than more complex patterns. Random patterns, on the other hand, are also easy to describe, because they contain little to no information. Metrics like Entropy, Minimum Description Length, Algorithmic Information Content, or Fractal Dimension have been proposed to answer this question. Some are computable for small systems.
2. Is it hard to (re-)create?
   Imagine you are to enumerate the steps required to create the system: how long does it take? Most prominently, Algorithmic Complexity proposed by Kolmogorov (1963) defines the complexity of a system indirectly to be the length of a computer program producing the system. While historically important, it is empirically useless and doesn't exhibit desired properties of a complexity metric. Thermodynamic Depth, another metric, describes complexity as the minimum sequence of action to reach the system state from equilibrium, and derived from this, Statistical Complexity measures the entropy of an abstract automaton, where some examples of its computation exist (Shalizi 2006, 2012).
3. What is its level of organization?
   This expresses the notion that self-similar systems are less complex than multi-level systems that are different on each level and hence more difficult to describe. The amount and quality of exchange in the system hierarchy is another important quantity measuring the level of organization.

The two approaches to characterize complex systems contain one another: a system that is recursive and parallel, reacting on input and producing output, and changing in reaction on its environment, is typically neither easy to describe nor to create, and it requires a non-trivial degree of organization to stay alive. Conversely, systems that are simple to describe or create will likely not be highly parallel nor reactive, interactive, or adaptive.

Though many authors have defined metrics that quantify the complexity of a system, only few authors report their application to systems that practically matter (Lloyd 2001; Mitchell 2009), and even these examples are comparatively constrained. For reasons that are summarized for example in Landauer (1988) as well as (often whimsically) detailed in the references and summaries given by Shalizi (2012), attempting to formalize and quantify the complexity of practical, real-world tasks is practically unfeasible, either because most complexity metrics cannot be computed at all, or because they are only computable for toy problems — but ironically not for *complex* problems since these cannot be described. To measure the complexity of clinical tasks is consequently not an undertaking pursued in this thesis.

We will instead look at cancer from the perspective of microbiology and tumorigenesis to understand cancer's complex nature, before we propose approaches that anticipate the complexity of the clinical tasks underlying them. There is no theoretic reason to claim that there can never be simple solutions to deal with complex systems. The above Fig. 1.2 suggests the same: The

way to deal with the complex system is emergent and cannot be deduced from the system by analyzing it. Instead, the successful management of a complex system develops in interaction with the system through probing, sensing, and responding — generating a new hypothesis. It is this property of complex systems that underlies some of the approaches presented in this thesis. The Appendix A.1 expands more on the relationship between complexity of a system and comprehensibility of it. An example shows how a system can be constructed that conceals its purpose through the complexity required to achieve it, but which shows a behavior that can be described in a simple algorithm. However, the potentially simple solution to or description of a complex system is never simple to reveal, but instead requires deep research and understanding.

1.1.2 Cancer

From the practical perspective in this thesis, the fundamental building blocks of the system in question — breast cancer — are the cells. Cells, the genetic information they carry, and the phenotype emerging from the genes, are the lowest level of information that is considered in today's clinical decision-making, which justifies this choice. In the following, the path is laid out from normal cells to tumors and cancers.

Figure 1.3: Breast cancer cell as pictured by a scanning electron microscope. Source: National Cancer Institute (NCI).

A single cell is a complex system in and by itself. It interacts with its close vicinity, and is capable of sending and receiving messages from more distant locations. Cells catabolize and anabolize substrates for example to produce energy or to grow, and most importantly, they comprise a mechanism to reproduce themselves by copying the DNA in a process that involves a plenitude of cell-internal signals and correctly initialized internal states. This process results in mitosis, the division of the cell, emitting two genetically identical cells, mutation left aside.

A cell is, by our definition above, a complex system: It is parallel (although not in the strict sense of parallel similar building blocks); it is iterative in performing the same actions over and over again, reacting on feedback; and with mutation, it has the ability to evolve.

Organs usually exhibit a multitude of different cells fulfilling the function of the organ together. Human cell types are in the order of a hundred, and different organs comprise different subsets of the total. From the organ perspective, a higher level of complexity adds to the cell level. Cells become the basic building blocks, and between them, the parallelism of processes and the sequential or iterative nature of actions becomes obvious. Organs have a function, and they fulfill it by reacting on stimuli by their environment — on both the organ level, and the cell level. Cells receive messenger molecules that trigger cellular mechanisms, which in turn result in organ function.

Turning to the breast as an example, the epithelial cells react on hormonal changes either delivered by the endocrine system during the menstrual cycle, or administered in hormone replacement therapy after menopause, by building up lobular structures in a strictly controlled proliferative process. These cells are again torn down with the loss of external or endocrine hormones. For these pathways to function, cells expose receptors for signals they need to react on, like for example the estrogen receptors.

Next, the emergence of uncontrolled proliferation, tumorigenesis and carcinogenesis, can be viewed as complex processes. Tumors, as Hanahan and Weinberg (2011) put it, are "complex tissues composed of multiple distinct cell types that participate in heterotypic interactions with one another. [They] have increasingly been recognized as organs whose complexity approaches and may even exceed that of normal healthy tissues. [The] complex interactions between the neoplastic and stromal cells within a tumor and the dynamic extracellular matrix that they collectively erect and remodel" is not known today.

Carcinogenesis

Most cells in the body are programmed to divide, and most of them also to die after a fixed number of cell divisions. The cell life cycle follows a delicate balance of different kinds of inter- and intra-cellular signals, and researchers are only beginning to chart the signaling pathways. Each signal may cause a reaction elsewhere. Some of the most important signals are growth-control factors, inducing or prohibiting cell division. If certain defects in the signaling processes occur, growth may escape control. Cancer is the result of a number of such defects that collaboratively contribute to excessive growth. It is the remarkable and frightening property of cancer cells that they effect the required changes out of their own capabilities. Their excessive reproduction rate eventually leads to mutations that are perfectly adapted to the local environment — an ironic twist of the "survival of the fittest" theme.

Genetic mutations at DNA loci that, combined, potentially result in carcinogenesis occur many times all over the body and are a perfectly normal aspect of cell division in the body. The vast majority of defective alterations is cleaned up by protective responses like DNA repair processes and immune surveillance and defense. These mechanisms, however, may also fail, so that it is a matter of stochastics whether or not a cell with a certain accumulation of DNA defects evolves into a proliferating disease. The statistical chance is usually expressed as a "relative" or sometimes even an "absolute" risk of a person to acquire a genetic mutation that grows out of bounds — the "cancer risk". In this context, a body of literature looks at inheritable genetic factors, environmental influences, living conditions and their respective contributions to the individual risk, and epidemiology considers life circumstances and environmental factors as contributors (compare IOM (Institute of Medicine) (2012) and references therein).

It is crucial to note that not a single of the factors alone is a sufficient condition for cancer, nor is any of them a necessary factor. Not only the original defective cell and the cell cluster originating from it have to have specific carcinogenic genetic defects, but most importantly, they are supposedly only able to survive in tissue with characteristics that foster proliferation, the so-called extracellular matrix forming the basement membrane and separating the epithelial

cells from the supportive connective tissue, called the stroma. It is the stroma that has become the subject of a wide body of research over the last decade (compare e.g. Arendt et al. (2010), Basanta et al. (2011), Finak et al. (2008), Levental et al. (2009), Myhre et al. (2010), Ng et al. (2009), Polyak et al. (2010), Proia et al. (2005), Sharma et al. (2009), and Vargas et al. (2012) and others). Speaking of breast cancer, it is generally known that windows of susceptibility exist from *in utero* to the postmenopausal age, with endogenous and exogenous influences that promote DNA defects adding up to the gross (breast) cancer risk. The extracellular matrix itself is also regarded as a decisive ingredient, since it mediates the signaling between stroma and epithelium.

Cancer is a multifactorial disease, and a multistage process. Several hundred distinct types are known, and several dozens are clinically differentiated today. Clinically, they are assigned to groups with shared properties, like receptor expression, or the origin of their growth, caused by a great heterogeneity of genetic alterations (Bombonati et al. 2010; Perou et al. 2000). Though it is true that no two cancers are genetically equal, it is a clinically impractical extreme position. Still it shows the complexity of speaking about cancer, which also manifests in the history of attempts to cure and prevent cancer (Mukherjee 2011). The complexity of breast cancer progression is today being researched with genetic, epigenetic, and proteomic technologies, and, like many times before, researchers hope to finally find the tools to cure or even prevent breast cancer (Perou et al. 2000).

Hallmarks of Cancer

Elucidating tumorigenesis in more depth sharpens the appreciation of cancer as a complex system. In fact, cancer has been described as the perverted, but perfected version of our (complex) selves, an optimized counterpart exploiting the normal tissue around it, harvesting its resources and altering it to its own benefits (Mukherjee 2011).

Cancer is induced by mutations at oncogenes, genes that play vital roles in central cell signaling circuits and are hence key to cancer formation. While mutations naturally occur at a certain rate, the rate may increase in presence of several personal and environmental factors like for example radiation or smoking, but also genetic predisposition. Mutation needs to impede several central control mechanisms of the cell, and their statistically extremely unlikely concurrence forms the "hallmarks of cancer". Eight such mechanisms — only eight, remarkably, since hundreds of different cancers are distinguished — have been proposed by Hanahan and Weinberg (2000, 2011). They need to be accumulated by subsequent genetic defects, and together cause a cell cluster to transform into a tumor and eventually into a *cancerous* cell cluster. The contribution of the tumor microenvironment has also been highlighted by the same authors (Hanahan and Coussens 2012).

Resist cell death (apoptosis) Apoptosis has to be differentiated from necrosis, which is cell death "in a blast", often causing inflammation. Apoptosis happens silently. Cell-internal and external triggering pathways and guardian sensors can lead to apoptosis in reaction to defects or advert environmental conditions. Cancers are able to cut these pathways.

Sustain proliferative signaling Normal cells have mechanisms that prevent excessive growth in order to maintain viable tissue structures. Tumor cells need to undermine these mechanisms to proliferate, but also carefully controlled to prevent the transition into a non-proliferative senescent cell state.

Evade growth suppressors Strong redundant programs to prevent excessive proliferation and suppress tumor growth exist in the body, and they need to be circumvented by tumors.

Activate evasion and metastasis The crucial trait of cancer, setting it apart from tumors and non-malignant dysplasias, is its ability to spawn cells, and, more importantly, the ability of

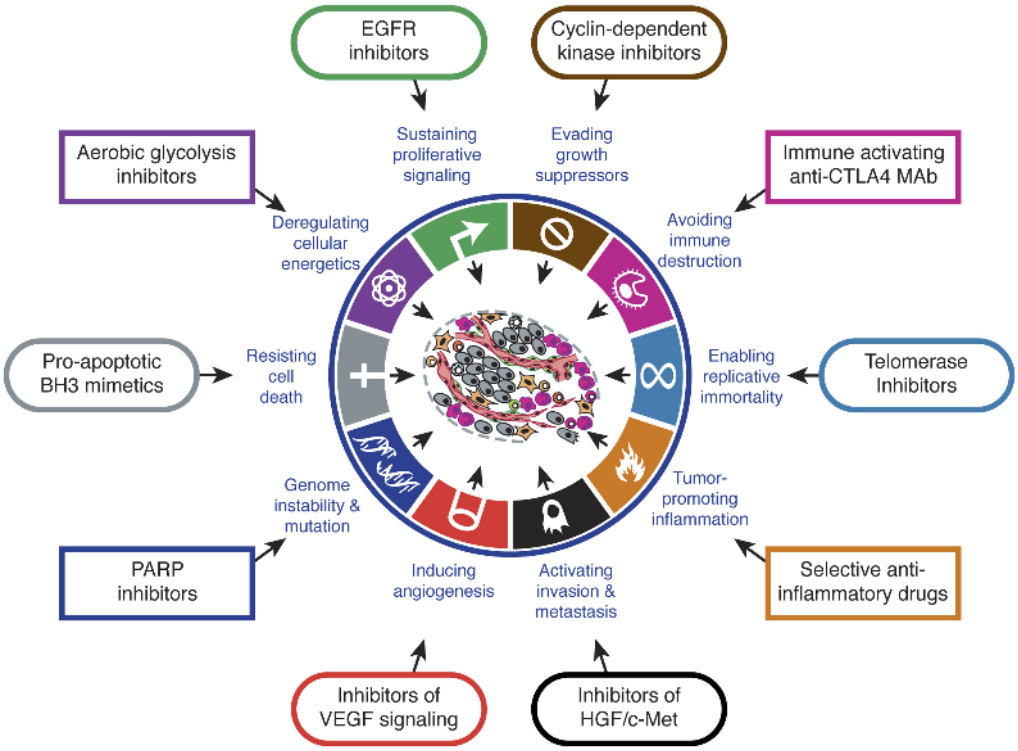


Figure 1.4: Hallmarks of cancer. From: Hanahan and Weinberg (2011).

such cells to root in distant organs and tissues. This ability is perhaps the most complex one when considered individually, and certainly the one where the greatest gaps in knowledge exists.

Enable replicative immortality A molecule bound to the end of the DNA is being used up in each cell division. Cancer cells need to rebuild these molecules or prohibit their consumption to become replicatively immortal.

Induce angiogenesis Tumors activate angiogenesis early in their development. To maintain ongoing neovascularization, tumors utilize inflammatory cells in their vicinity. The "angiogenic switch" is in fact not a single switch, but again a complex interplay of inhibitors and inducers of angiogenesis.

Deregulate cellular energetics Otto Warburg was the first to discover the reprogramming of cellular energy production in cancer cells, lending his name to the effect, henceforth known as the "Warburg effect" (compare Costello et al. 2005; Gatenby et al. 2004). Even in aerobic conditions, cancer cells use a far less efficient pathway to generate energy for cell division, making them independent of oxygen supply. Glucose consumption needs to be upregulated instead to sustain anaerobic energy supply.

Avoid immune destruction The immune defense of the body fights cancer cells. Cancer cells need to adapt to this and escape the immune system's control. Tumors are more likely to grow in the presence of immune system deficiencies.

The difference between tumor and cancer has been pointed out by Lazebnik (2010). Tumors are not necessarily malignant; they lack invasion and metastasis. Still, the words *cancer* (that possesses the capability to metastasize) and *tumor* are often used almost synonymously. Hanahan and Weinbergs works have consequently been criticized for contributing to this loss of clarity, most importantly since only one out of the hallmark capabilities is generic to cancers (namely, metastatic activity), while all others are shared by tumors alike.

Tumors and cancers are obviously complex. No mechanistic description has so far been possible, at best small parts of the signaling pathways are elucidated, understood, and can be simulated or used in breast cancer care. Optimal detection, diagnosis, and treatment of a highly complex disease requires to deal with this complexity in some way. Differentially diagnosing a breast cancer and robustly proposing a definitive treatment requires more knowledge than is available in most instances. While cancer is "predictable" in many aspects, and many cause-effect relations in treatment options are known, important details are still missing from the picture, which relates to mostly all of the hallmark capabilities above, including in particular the metastatic behavior. Hence, probing rather than sensing is the starting point of many of today's diagnostic image-based approaches, which is the characterizing decision making behavior for complex systems (cf. Fig. 1.2).

Biology and Development Pathways of DCIS

Regarding breast cancer diagnosis and treatment, DCIS is among the most challenging breast cancers. It is only one of the types of non-mass like segmentally growing lesion types, but by far the most prevalent one. A detailed description will yield an insight into how the complexity of a disease is sometimes poorly addressed by modern image analysis, but perhaps also by clinical cancer care. This thesis will pick up the detection and characterization of DCIS and non-mass like enhancement patterns depicted with DCE-MRI in Sec. 2.2.

It has been discussed vigorously by tumor biologists and pathologists what the nature of DCIS is, and various groups have found different, and sometimes even contradictory, explanations and characterizations. DCIS has been described a precursor lesion that inevitably progresses into aggressive, invasive forms of cancers (Lakhani 1999), but more recent research suggests that it may remain indolent or even regress. Also, the focality and multi-centricity of the disease has been discussed, which will be summarized later. In this light, DCIS is the subject to an intense clinical debate regarding the imaging features to look for, and treatment options. In the image computing community, DCIS has also received some attention, but its detection and automated characterization is considered an unsolved task apart from mammography-based techniques that have severe limitations. Pure DCIS accounts for 15–25% of "all breast cancers"[2]. Its incidence increased seven-fold with the advent of screening mammography, which is today the primary detection source. DCIS is more prevalent in white women than in black or other ethnicities and the risk for developing DCIS increases in the age 40–50 years (Dabbs 2012).

"Ductal carcinoma in-situ (DCIS) is a proliferation of abnormal epithelial cells, confined by the basement membrane of the mammary ductal system" (Patani et al. 2008). DCIS is a heterogeneous group of lesions with diverse characteristics.

Biologically, DCIS is not "just there", emerging out of nothing. It has been observed in pathology, that specimen containing DCIS also exhibit dysplasias of different degrees. Knowing from other cancers that dysplasia and hyperplasia are sequentially followed by invasive forms led researchers to the conclusion that this causative chain applies to breast cancer as well. These pathways on which hyperplasias develop into proliferative, but in-situ forms, and further on into invasive cancers of the ducts and lobular units has been a subject to studies for several years already. Lakhani (1999) summarized the evidence for this molecular transition model in 1999, stating that "a transition [to IDC] from normal epithelium to invasive carcinoma via non-atypical and atypical hyperplasia and in situ carcinoma" is the most likely hypothesis. Leonard et al. (2004) have later reviewed eight studies reporting the progression of untreated DCIS, and found a span of 14–75% of lesions to develop into invasive cancers. The wide span of lesions that are reported to become aggressive raises doubts on the prior causative assumption. The finding that a significant number of DCIS lesions might not become aggressive at all is also making the

[2]Quotation marks indicate that there is discussion as of if DCIS should be called a cancer after all.

clinical decision of how to treat an individual DCIS patient much more difficult. Despite, even recently some reports state that the favored hypothesis is "that invasive ductal carcinoma (IDC) evolves progressively through sequential stages, usually from ductal hyperplasia without cellular atypia to atypical ductal hyperplasia, to DCIS, and, eventually, to IDC" (F Sardanelli et al. 2008), although the subtle difference has to be noted that the author does not state that ductal hyperplasias is the root cause of IDC.

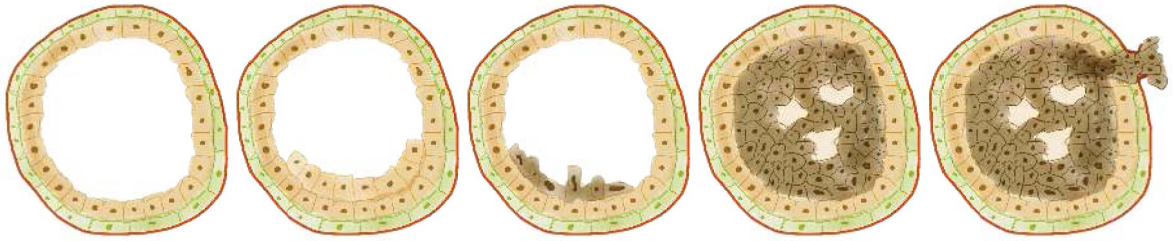


Figure 1.5: Schematic of the development of DCIS. *Cells*: The red line indicates the basement membrane. The thin (green) outmost cells are the myoepithelial cells, forming the basal layer. The orange layer are the epithelial cells that surround the lumen. Cells in darker brown are cancer cells. *From left*: Section of normal duct — Hyperplasia — Hyperplasia with atypia — Ductal carcinoma in situ — Ductal carcinoma breaking the basement membrane.

But the picture changed when the view widened. Clinical reports began to underline that not all DCIS develop into invasive cancer. It was assumed that hyperplasias are caused by some genetic alterations, and the next progression is caused by another genetic alteration that is acquired in the ADH stage. However, it was found that significant "global alterations [of genes] occur at ADH and are maintained in later stages of DCIS and invasive ductal carcinoma", suggesting that these genetic alterations occur early on, and are not the cause of the progression. Stated differently, the same factors lead to invasive as well as non-invasive types of cancer (Valenzuela et al. 2007), or yet in another light, there are genetic alterations that will be present in the earliest stage of hyperplasias that determine a lesions potential to become invasive. The same has been observed in a study that initially aimed to depict the transition from DCIS into invasive cancers in a transgenic mouse model. This work ultimately found that only nine out of the 21 initially detected lesions developed into invasive cancers while others remain indolent or even regress. So this study ended with "image-based evidence that DCIS may be a non-obligate precursor lesion with highly variable outcome" (Jansen, Conzen, Fan, E. J. Markiewicz, et al. 2009).

Active surveillance of clinically observed DCIS may hence be an option, but arguably any two-arm study that randomizes women with DCIS to either the arm where a tight imaging-based control is executed ("active surveillance"), or where surgical excision is conducted will face problems in ethical approval. Hence, no large-scale studies exist, but from selected women who refuse surgical treatment, at least some evidence has been collected by (Meyerson et al. 2011). In their work, they examined the development of DCIS in 14 women who opted out of definitive surgery. At the time of the report, six of these women remained under surveillance, with one to almost seven years (median 30 months) without indication of invasive cancer. Noting these numbers, a potential interdisciplinary research direction for tumor biologists and imaging physicists is to establish imaging biomarkers for the aggressiveness and growth prediction of DCIS, so that a predictive parameter can be assessed minimally invasive or non-invasively.

Today, pathologists and oncologists begin to see a more differentiated picture as they observe xenografts and more and more histopathological stainings of human developing cancers. Besides H&E staining[3], the standard staining used to assess sliced specimen, immunohistochemical stainings that show very particular tissue characteristics like hormone receptor status or HER2 status, are used (Yeh et al. 2008). Also, the advent of digital pathology begins to change the

[3]H&E: hematoxylin and eosin

picture of how the assessment of pathological images turns from qualitative and sometimes rough qualitative estimations into quantitative imaging biomarkers (Di Cataldo et al. 2012).

It is from such examinations reported, that precursor lesion develop into invasive forms on different time scales. "[It] appears that atypical cells or tumor cells find themselves at an early fork in the road and commit to either a well differentiated path or a poorly differentiated path" (Wiechmann et al. 2008). It is today accepted that DCIS "is a heterogeneous, unicentric precursor of invasive breast cancer" (Pinder 2010). The results obtained using a mouse model in which particular factors have been enabled or disabled to assess their influence on the paths of progression support the same paradigm shift. Knowing that DCIS shows numerous genetic alterations, making it genetically almost indistinguishable from invasive tumors, the contributions of the surrounding tissue are taken into consideration. In conclusion, "the progression of in situ to invasive breast carcinoma may not be due to the intrinsic properties of the tumor epithelial cells acquired during tumor evolution but determined by complex interactions among all the cell types that compose the tumor microenvironment" (Hu et al. 2008).

Until today the question is not settled — in the contrary: All current results only seem to indicate that the understanding of breast cancer progression (probably like cancer genesis and progression in general) is only in its infancy (Mukherjee 2011). Reeves et al. (2012) conclude in their review that out of all genetic and environmental factors considered, all but one correlate with the genesis of DCIS and IDC to the same degree, suggesting that DCIS is a precursor lesion to IDC. The one differing factor, the body mass index, may promote the progression from DCIS to invasive cancer. Consequently, molecular biology tries to explain findings that are generated by gene expression profiling of DCIS and invasive cancers, to reveal more and more pathways and messengers that contribute to or inhibit the progression (S. Lee et al. 2012; Scribner et al. 2012; Tamimi et al. 2008). Increasingly, interest is also directed to the stromal microenvironment and its character and influence as an agent that interchanges with the tumor, promoting or inhibiting its progression (Ma et al. 2009; Sharma et al. 2009; Vargas et al. 2012). From today's perspective, the "hallmarks of cancer" (see Sec. 1.1.2) have been revised by their originators to include the crucial interaction between stroma and tumor; included in this expanding field of view is the notion of the diversity of possible individual variations in tumor and stromal cell composition in terms of their carcinogenic mutations. A consequence of this variability is that "targeted" therapies, that bore the promise of eliminating the roots of cancer, no longer seem to be a feasible option. Frequent relapses after such therapies speak the same language (Hanahan and Coussens 2012).

One interesting new research direction links gene expression studies to imaging, notably breast MRI, relating enhancement patterns to molecular cancer profiles (Yamamoto et al. 2012). Not surprisingly, cancer genes correlate with imaging appearance, and this is good news for our efforts to derive prognostic markers from imaging studies.

In sum, the belief of a linear pathway of tumor progression has lost ground. Instead, the crucial role of the microenvironment has been described and offers a new paradigm for future research. Additionally, the interaction of the human (cancer) genome with the human microbiome has only recently entered the research stage, suggesting that in the future surprises can be expected (Khan et al. 2012).

With the recent success in the early detection of non-invasive cancers and the accompanying success of surgical treatment of such lesions, researchers became interested in a more thorough understanding of the progression from precursor lesions into aggressive, invasive cancers. The hope is that in these pathways, junctions can be identified that can be targeted by imaging, or by drugs, to the benefit of the patient. Hence, model systems of different kind have been set up, including computational models that describe how certain aspects of the molecular transitions influence progression, but more importantly also mouse models that can be observed with tools resembling human breast cancers, allowing for potentially easier transfer of knowledge.

Furthermore, since mice can be sacrificed during the studies to assess tissue samples in much greater detail, imaging and histopathology can be correlated, and imaging methods that are not available for in-vivo measurements can be employed.

Sacrificing a mouse diseased with a model of early DCIS after injection of contrast agent allows to examine the tissues in the vicinity of the ducts for accumulation of contrast agent using x-ray fluorescence microscopy. By this method, it has been confirmed that Gadolinium, the contrast agent molecule, "penetrates and collects inside neoplastic ducts" and that these ducts can be identified in-vivo using CE-MRI (Jansen, Paunesku, et al. 2009). Jansen, Conzen, Fan, E. Markiewicz, et al. (2011) conclude that transgenic mouse models can be used to demonstrate the utility of MRI in longitudinal studies of early stage mammary cancer disease progression and response to therapy.

! Transgenic or xenograft mouse models are mice injected with 0.5×10^6 to 5×10^6 tumor cells. The injection may be subcutaneous or orthotopic (at the original tumor site). Tumors grow reproducibly in immunocompromised mice. Xenograft mouse models for human cancer types are sometimes criticized because the receiving mice have to be immunodeficient to allow the injected cells to grow, which is an unrealistic microenvironment. So-called humanized mice can help in this case: for these models, human bone marrow is inserted into the mice to recreate the immune system. These models are known to mimic the human tumor microenvironment appropriately, but are expensive. Overall, xenograft mouse models are generally considered to be sufficiently close to human breast cancer to raise hopes that a transfer of results is possible. An alternative to xenografts are GEM (genetically engineered mice), where gene defects are artificially created to induce tumor growth. A competent immune system provides a realistic microenvironment (Richmond et al. 2008; Teicher 2009; Vargo-Gogola et al. 2007).

1.1.3 Complexity in Breast Cancer Care

Computer-based methods to support image-based clinical care of breast cancer are one way to deal with growing numbers of images, to optimize the usage of imaging modalities respecting their individual benefits and drawbacks, and to account for the limited number of radiologists with limited time. Some novel approaches that help with the detection, diagnosis, and treatment of breast cancer will be presented in this thesis. While some of the methods are generic and may as well be applied to other medical fields, others are specific to the disease breast cancer.

Starting early in the history of medicine, breast cancers have been identified, named, and operated on — without big success. Successively, the methods of diagnosis have been refined, as have the methods of treatment and patient care. Today's clinical breast cancer care standard comprises *screening* to detect cancers early, *diagnostic imaging* using sophisticated technologies to depict a suspicious area optimally, combined with clinical oncology to assess a woman's personal risk. *Treatment* is more and more local, and only in fewer and fewer cases radical resection of whole breasts, a mastectomy, is necessary. Even then, cosmetically convincing outcomes can be achieved with nipple- and skin-sparing surgical techniques followed by breast reconstruction.

This status has been achieved with the help of computers in all areas of cancer care. Computer based image analysis, image-guided simulations, and image based quantitative tissue characterizations are fast, reproducible, available at any time, and may even have predictable variance and bias. For these reasons they have become indispensable tools in image based clinical care in general, and for oncology in particular. In fact, imaging in oncology generates a huge demand for computer support, and both the amount of acquired images and the opportunities to apply digital image analysis have grown in the last decade, with indications that the growth will continue.

One significant change was the wide-spread introduction of digital mammography followed by the start of population-based mammography-based breast cancer screening. This necessitated

substantial developments in image review software support including the workflow that had to be made as fast as the reading using light boxes. Today, mammographic x-ray images turn 3D with the development and adoption of tomosynthesis, so that available software and trained review methodology will likely need to be revised once more. Likewise, pathological images are increasingly digitized, or histopathology slides are directly digitally scanned, offering new research directions, but also providing new challenges for the integration of disciplines. For diagnostic application, many imaging technologies have been developed to optimize breast cancer depiction for various scenarios, like ultrasound and breast MRI. Their differential contribution to the overall diagnosis requires radiologists to review them synchronized, which brings about the challenge to devise spatial correlation methods to reduce the workload of this task.

Shortcomings can still be identified in all areas, and motivate much of the work presented subsequently:

1. Machine learning approaches are becoming more popular due to a quickly growing number of quantitative imaging biomarkers, defined for multiple modalities. Large feature spaces comprising these biomarkers pose specific problems for automated analysis, particularly when only small sample sizes are available — a frequent situation in the medical area. The thesis will investigate these problems in theory and experiments, and apply the result on two practical examples, the classification of mass lesions of the breast, and the detection and characterization of non-mass enhancing lesions.

2. In diagnosis, a noteworthy unexplored area in the landscape of computer-aided diagnosis exists, which concerns the type of breast cancer lesions with the most promising potential of definitive treatment: ductal carcinoma in situ (DCIS). The thesis will examine the yield and potential of a method to detect and quantify non-mass enhancing breast lesions, among which DCIS is the most prominent one.

3. As described above, the number of screening and diagnostic images grows rapidly, imposing a high workload on the radiologist. This is a consequence of the widespread introduction of 3D modalities, such as MRI, but also breast tomosynthesis and automated volumetric breast ultrasound. The computer support for efficient reading of the increasing amounts of images has not kept pace with the growth of image databases.

4. Lastly, a mental gap between radiology and surgery exists in the hospital, which is reflected by computer science's inability to provide suitable tools supporting the communication between the disciplines. In the scope of this thesis, approaches will be presented that are in principle feasible to fill this gap by providing the technical framework for deformation simulation of the breast, a prerequisite to derive visualizations of the breast virtually placed into arbitrary positions.

This thesis will describe contributions to complexity reduction in breast cancer care in the three clinical areas screening, diagnosis, and therapy support, with the intention to make clinical routine processes more efficient and more robust. To this end, the notion of clinical *workflows* is used, by which we describe the operating procedures and algorithms by which routine processes are organized, regardless of their formalization. A workflow, in our use of the term, comprises the tools that are used, the sequence of actions and decisions, and the input and output. Input in most cases is data, in the context of our work mostly radiology data. Output of a workflow may again be data, or a clinical decision.

After a concise overview of the complexities in the above three areas, a general overview of approaches to complexity reduction is presented. It is the purpose of this chapter to summarize the current state of the art in image-based clinical breast cancer care from screening and diagnosis to therapy, highlighting past improvements in the clinical workflow as well as current tools and research efforts that reduce the complexity of the modern tasks. It is relevant to note an

apparent conflict: Workflows have above been introduced as operating procedures, and these are per definition a sign of simple systems (see above). Yet, the act to take decisions in the workflow is not simple, but can often be complicated, and sometimes complex. Hence, we will also describe approaches that reduce workflow complexity.

Screening

Breast cancer screening today is screening with film screen x-ray mammograms and increasingly with full-field digital mammography where the analog film has been replaced by a high-resolution digital detector plate. We point out the existence of highly optimized computer workstations that guarantee high-throughput screening, supported by hanging protocols guiding the radiologist through the images of a screening session, and special keypads for efficient interaction.

In European countries, screening mammography is predominantly offered to women aged 50–69 on a biannual basis. The invitation rate averages at about 80%, and the average participation rate across all programs is about 48% (data from 2005–2007; Giordano et al. (see 2011)). It should be noted that according to many reports, women aged 40 or older should be included in the invitations, since cancers in the age group 40–49 years tend to have a worse prognosis than cancers detected later in life, although on the other hand the impact of frequent exposition to x-ray radiation has to be weighted higher than in older age groups.

(a) (b)

Figure 1.6: Screening mammography. Left: Alternator, the machine used to read film mammograms in screening mammography prior to digitization. Right: modern full-field digital mammography screen reading workstation with special keypad.

From the German report on screening mammography, the characteristic numbers for the first screening round (women invited for the first time in their lives) show the workload associated with population-based screening: In a biannual screening program inviting women from the age of 50, 54.3% of invited women participated in mammography screening. The invitation rate was 91.6% of eligible women. The recall rate in Germany (the number of women that are asked to return for a diagnostic workup of a suspicious finding) was 6.12%, which is an average value in international comparison. The cancer detection rate was 0.819%, i.e. in about 8 of 1.000 screened women a malignancy was found (Malek et al. 2009). In comparison, in the U.S.A. women are offered screening mammograms annually in the age group 40–79 years. However, there is no population-based screening program. This may be the reason why the recall rate is comparatively high at 10%; the cancer detection rate, however, is only reported to be about 0.45% (Kerlikowske et al. 2011). These numbers only apply to the first screening round and are different in subsequent rounds. Also, the variations between federal states for example in Germany are large and depend on many factors.

Screening for breast cancer is likely to change in the future, owing to a better understanding of the influences of personal breast cancer risk, breast density, and genetic predisposition on the sensitivity and specificity of mammography. In some well-defined cases, mammography performs poorly, most prominently in women with breasts with a high fraction of mammographically dense glandular tissue. Particularly small cancers, which are likewise mammographically dense, may hide in the superposition of glandular tissue. The core issue is that dense tissue is most often found in young women, and that cancers in premenopausal women (and in particular in those with a hereditary BRCA1 gene mutation) are more frequently of the aggressive, fast growing basal (triple-negative) types which cannot be treated with a targeted therapy today (American Cancer Society 2013b).

For young women, women with a genetic predisposition, a family history of cancer, or risk factors like dense breasts, screening with alternative imaging methods has been and is being explored in large-scale retrospective analyses and prospective trials. The imaging technologies utilized are dynamic contrast-enhanced breast magnetic resonance imaging (DCE-MRI) in many flavors (see Sec. 1.2.3), hand-held two-dimensional and automated three-dimensional (volume) ultrasound imaging, and despite the higher radiation dose even in some cases positron emission mammography. Breast tomosynthesis may as well reduce the occlusion problem of the 2D projection imaging in mammography by a limited three-dimensional acquisition.

From an overview of detection rates, Lehman (2012) reports a sensitivity of less than about 40% for mammography and ultrasound in a high risk screening population, while MRI outperforms both at a sensitivity of an overall 84.6%. For high risk women, screening with MRI alone is advocated by parts of the scientific community (Francesco Sardanelli et al. 2012), which makes the clinical procedure selection easier, but is currently prohibitive in terms of procedure time, costliness, and reading time, if applied beyond very selected populations. Upon larger-scale adoption of MRI, the workload in radiology departments and reading rooms is bound to increase.

Automated volume ultrasound, a relatively new modality that is under consideration to become an adjunct screening modality, it can be observed how computer detection and diagnosis of lesions is being developed to a level where it might serve as a reading aid, perhaps even to detect normal cases and filter them out before a radiologist has even looked at them. According computer-aided algorithms for the detection and characterization of breast cancer have for example been presented by Fujita et al. (2008) and Tan, Platel, R. Mus, et al. (2013) and show very promising initial results, matching radiologists in sensitivity and specificity.

Since screening implies high throughput image reading, computers have been used as second opinions for many years. The most often implemented scheme is computer-aided detection (CADe), a paradigm where for example the human observer rates the images first, before the marks of the CADe system are displayed on the images, and the radiologist either changes the categorization or neglects the marks. This paradigm is today challenged by several proposals that are likely to improve diagnostic accuracy over the old approach (Hupse et al. 2013).

Summarizing the screening scenario, the following tasks are extracted:

Task 1.1 — Personalized Screening. Decide who needs which personalized imaging protocol in screening, pertaining to the type of imaging and the frequency. Personal risk needs to be an integral part in the determination of the screening strategy.

Task 1.2 — Include CADe/CADx. CADe is often over-sensitive, potentially leaving the radiologist with the task to sort out the unimportant (false positive) detections, but computer-aided diagnosis (CADx) on the other hand is known to increase the performance of inexperienced readers and the confidence in their diagnosis also for experienced readers. Usage of CAD is not always at the required level, hence a thoughtful clinical integration of such tools is required.

Task 1.3 — Handling the Workload. The sheer amount of images emerging from mammography screening today is almost impossible to handle, and expected to increase. MRI for selected screening applications is on the verge and will produce even more work. Information needs to be synchronously assessed from different images, and workflows and dependencies become more complex, and might require changes.

Diagnosis

Diagnostic imaging comprises images acquired after establishment of a suspicious finding from the primary detection modality. Many different scenarios exist that determine the sequence of imaging modalities employed for clinical workup of the finding. All available modalities, like mammography and tomosynthesis, hand-held and automated ultrasound as well as ultrasound shear wave elastography, contrast-enhanced breast MRI, breast-specific gamma imaging, and positron emission mammography have distinct places in differential diagnosis.

For many imaging modalities, automated or semi-automatic manual segmentation algorithms to delineate anatomical structures are available, for example for the pectoral muscle, the nipple, the parenchymal structures and the adipose tissue, for the breast outline and for suspicious findings. Likewise, the extraction of descriptive parameters based on these structures is well-researched. For example, breast density and breast volume can be assessed, and from the delineation of the anatomy, the reporting of finding locations can be eased by computer-generated measurements. Displaying lists or image markup showing computer-detected abnormalities in breast MRI has proven to be helpful particularly to inexperienced readers (Lehman et al. 2013).

Hence, computer support for many tasks in the common imaging methods is mature, but in particular in the diagnostic setting, the multi-modal assessment of a finding is clinically lacking and only explored in research projects (e.g. Giger et al. 2013; Yuan et al. 2010). Instead, quantitative imaging features are usually derived from the individual modalities and fused into the suggested action by the radiologist, and in case of a classification as probably malignant, a biopsy will follow to establish the diagnosis.

Many potentially helpful CADx systems have been proposed for mammography and breast MRI, but in particular the latter lacks dedicated support for the class of malignancies that are best seen and worst differentiated: segmental enhancements and non-mass-like lesions.

Also, as breast MRI is increasingly utilized as a cancer staging tool and decision aid, convincing concepts for high-throughput reading and annotation of breast MRI studies need to be devised. They might complement the existing successful diagnostic applications for breast MRI that are optimized for the comprehensive analysis of single cases. Some of the available diagnostic software for this purpose, however, does not conform with high-throughput requirements in diagnostic image reading anymore. User interfaces tailored at the extensive workup of a few breast MRI, for example, are too complicated and inflexible to handle larger amounts of patient cases efficiently.

For the scenario of differential diagnosis and diagnostic image assessment, the following complex tasks are extracted:

Task 1.4 — Software Support. Radiologists deal with complex user interfaces, hampering efficient diagnostic image reading. Differential diagnosis requires them to keep too much additional knowledge in mind — example cases, state of the art study results, patient history, etc. The provided CAD marks are not always usefully integrated into the application.

Task 1.5 — Interdisciplinary Diagnostics. The integration of other disciplines and their results, and dealing with the amount of possible modalities and protocols is poorly supported in clinical software support.

Treatment

Treatment, as much as it is image based or might be improved with computer support, encompasses options like chemotherapy, radiation therapy, and surgery. All are based on a differential diagnosis, which also brings about a prediction of the likelihood of each treatment option to succeed. Chemotherapy, for example, depends on the receptor status of the cancer; radiation therapy is possible if the lesion is not too close to the skin or other risk structures; surgery in many facets can be chosen for definitive treatment of smaller areas (lumpectomy) or disease with a larger extent (segmental or total mastectomy). The tumor burden, which can be understood as the affected tissue on biopsy, is decisive for the therapy choice — and notoriously hard to estimate from radiological images, particularly in the case of in situ cancers or in cases with distant foci or micro-metastases. They all have in common that very small cancer foci are potentially missed in imaging.

It is experience from prior cases and knowledge from research, condensed into guidelines and decision tree diagrams that help clinicians to select the most promising treatment option. Depending on the treatment, further planning steps are to be taken. The dependencies of cancer type, chemotherapy options, response predictions, adjunct surgical option, etc. are subject to a bi-annual expert consensus meeting, see Goldhirsch et al. (2013). While this consensus meeting gives little unanimous advice on how to treat which cancer, it can be observed from the documentation that with improved diagnostic tools including genetic testing and quantitative analysis of histopathology slides, a more targeted and more locally confined treatment may become feasible. Remarkably, though, only a minimal consensus has been reached by the 51 panel members of the 2013 meeting. Partially contradictory evidence from the reviewed studies may be one reason, the hypothetical nature of causes and effects in therapy another — both signs of the complex nature of treatment decision making due to the diverse genotypes and phenotypes of breast cancer. The results of the consensus meetings are unsatisfactory from the practical perspective. Therefore, a German expert working group congregated after prior meetings to translate the panel results into practically useful guidelines (Beckmann et al. 2009). In any case, the treatment of breast cancer is as much affected by the incompleteness of knowledge like the other two areas, screening and diagnostics.

1.1.4 Ways to Approach Complexity

When considering approaches that claim to reduce the complexity of one or more tasks involved in clinical breast cancer care, different aspects and types of complexity are addressed. In all, the approaches may be divided up into three categories: Firstly, *qualitative categorization* of imaging findings may aid in a less subjective description of breast cancer. When considering computer aid to reduce complexity, two other approaches are more interesting: *direct decision making support* in scenarios where human judgment often errs, and support of a more fundamental sort, perhaps best described as *ensuring task fulfillment preconditions*. While the first kind of support is usually given in the form of posters or printed or online atlases, the second type is potentially implemented in a computer algorithm. The third approach requires only limited computer-based support, instead the major developments are being made on a conceptual level.

Support by Qualitative Categorization

In clinical practice, efforts have been made to distinguish breast cancer types to ease communication and decision making. In a positive interpretation, this kind of simplification may make breast cancer more accessible despite its complexity, easing decision making because fewer variables have to be considered. However, simplification inevitably ignores some amount of knowledge on breast cancer, and hence ignores complexity, leading to wrong decisions, or to outcomes that are not conforming with the implicit assumption made by the simplification that lead to the decision.

One example for this is decision making in chemotherapy selection, here described according to an interview with an oncologist in a large university hospital. A very limited number of distinct therapy regimes is today accepted for breast cancer; a typical number of regimes to choose from is four, each available with slight modifications in actual medication choice, dose, and cycle number. Hence, all breast cancer types need to be described such that their eligibility for at least one of these regimes can be deduced. This is done based on a histopathologic assessment of a tissue sample, for which typically four parameters are described: the hormonal status (ER/PR), the HER2 receptor status, and the growth rate. Each of them is qualitatively described to fall into the low, intermediate, or high level of occurrence, yielding a total of at most $3^4 = 81$ categories. Each category is mapped to one of the four available chemotherapy regimes. Inhomogeneities in cancer composition (reflecting the complexity of how cancer grows) is largely neglected, instead, the most aggressive component determines the therapy regime. The consequence is that an optimal treatment of the actual cancer in its entirety is not guaranteed. Response rates of chemotherapy are thus mediocre, and research into the root cause of non-responders is hampered since no quantitative data describes the tumor as a whole, using the available image information.

Looking at the description of breast cancer seen in radiological images, there has been a long-standing desire to homogenize the verbal description made in reports. As one example, the ACR BI-RADS® atlas details since 2003 how to report on the appearance of breast cancer imaged using mammography, and since 2013 also for magnetic resonance imaging (MRI), and ultrasound. The patterns of image appearance are captured by assigning descriptors to them, and for example for MRI the atlas defines the terminology to describe any abnormality seen in breast images. Breast cancer is here divided into two major categories, masses and non-mass enhancement, and both are then subsequently characterized according to three (masses) and two (non-mass enhancement) criteria with predefined descriptors in each.

Perhaps striving for completeness, W A Kaiser (2008) lists 147 "signs in MR mammography" with their interpretation towards different types and gradings of breast tumors, normal alterations of the breast, and benign lesions. These signs describe morphology as well as functional aspects reflected in the uptake curves of contrast enhanced breast magnetic resonance imaging (compare Sec. 1.2.3). Some are readily eligible to be implemented into computer-based quantitative evaluations of MR images (e.g., the average relative enhancement in the lesion), others are very qualitative and subjective in their assessment and probably even for trained human observers hard to differentiate (e.g. "multiple unilateral regional enhancement" and "diffuse unilateral enhancement"), and it is questionable if such a differentiation into discrete categories instead of a quantitative description of the characteristic property is clinically helpful.

These approaches hence have limitations generated by the approach they take to simplify the complexity of breast cancer.

Support in Direct Decision Making

CADe (Computer-aided detection). Detection of abnormalities in images (or, more generally, of a search target in a source domain) usually employs characteristic features of the target that can be quantified from the image. CADe hence aims to distinguish lesion locations from non-lesion locations. Depending on the application, the aim of CADe can be high sensitivity, for example in screening where there should be no missed cancer. Depending again on the way how detection results are displayed to the user, a high specificity can be mandatory to reduce the number of false positive marks. Another clinical goal for CADe may be to signal that a case is "not normal".

CADx (Computer-aided diagnosis). Sometimes overlapping with the detection task, computer-aided diagnosis (CADx) characterizes the lesion locations further, for example by scoring or grading them. The score may or may not be visualized, and several options exist for the

visualization (color, metric, symbolic, etc.). CADx typically uses more sophisticated features, and they might be used in a classification mechanism that results in a binary or continuous scoring of each detection result. CADx may hence be considered as an extension of CADe. Some variants include interactive CAD (Hupse et al. 2013), but also the color overlays generated from dynamic contrast-enhanced MRI are sometimes considered to implement CADx. Such maps signal areas of highest suspicion by opaque signal colors. Computerized assessments reduce the complexity of decision making if the per-modality CAD results are fused by the computer, also excluding biases of human observers.

Spatio-temporal correlation and multi-modal analysis. Between different images (mammography, MRI, ...) acquired in one visit, and within one modality over the years, correlation of spatial locations can be eased with computer aid, particularly when the breasts may have changed due to aging, hormonal status, surgery, growing disease. This enables multi-modal assessment of findings, taking away much of the ambiguities when reviewing modalities individually.

CBR/DSS (Case-based reasoning and decision support systems). A natural extension of CADe/-CADx are decision support systems that use the CAD score or the decisive, most influential features contributing to the score to retrieve similar cases from a database. These templates can be presented to the decision maker, together with an annotation of their respective (a priori) decision. Instead of memorizing and remembering template cases themselves, radiologists can be supported in this task by the computer.

Volumetry. Reliable and reproducible measurements in images without observer variability, no deception that human observers may show (compare optical illusions), and no systematic misjudgments in quantitative results qualify the computer for such tasks (compare Jan Hendrik Moltz 2013).

Modeling like e.g. pharmacokinetic modeling, cancer growth modeling, therapy response prediction, and other mathematical models attempt to capture aspects most relevant to a certain decision task. Some permit to compute a prediction of outcomes with varying input parameters (e.g. cancer growth models reacting on drugs), other models take empiric measurements as an input to the model to compute physiologic parameters from them (pharmacokinetic modeling).

Normals detection. A recent trend is to not try to identify the diseased cases out of the bulk, but to reduce the number of cases the clinician has to pay more attention to by eliminating the certain normal cases.

Motion compensation. Measures to reduce artifacts from motion, particularly in time series like DCE images that need to be analyzed regarding the pharmacokinetic properties of tissues, are indispensable tools to create a standardized data appearance for human or computer-based assessments.

Support on the Conceptual Level

Besides these approaches which present themselves as computing and information processing problems and are hence solved using computer programs, there are further complexities and challenges that require information processing only as an adjunct to support a transformation of the working environment.

Workflow. Thorough research is required to assess, understand, and finally model and modify existing workflows in radiological image reading. One result of such considerations are the digital mammography hanging protocols for mammography screening workstations that have been modeled after the best practices of film-screen mammography, where the mammogram films were hung on a backlit screen on a machine that is called an Alternator. Today, an emerging challenge is the usability of "random access toolbox" style computer programs that offer all their functionality and all available images without user guidance. Getting from this presentation mode to an intuitive user interface is the topic of workflow research.

Hardware. Special keypads and devices can allow easier navigation, or support a particular way to work. Again using the example of screening mammography, the alternator that was utilized in film screen mammography has been replaced by specialized high-resolution monitors, and the alternator's physical user interface has been remodeled with a special keypad after transition to full-field digital mammography (FFDM).

Context awareness. Information extraction from images may improve the intuitiveness of computer software by providing tools only where they are required and sensible. This may be understood as a generalization of the principle of interactive CAD to other image analysis tools, like segmentation, measurement, annotations, and others. Only applicable choices are offered at any point in space and time.

Navigation. In the surgical setting, navigation aids are becoming more frequently utilized in several ways. Starting with surgery planning (aid in determining margins to risk structures, aid to envision surgical options), preparation, and surgery (image guidance by planning visualization, planning information projected onto patient surface), and outcome control by comparing planning and post-surgery images.

Interdisciplinary integration. One of the greatest challenges of today is the interdisciplinary integration. Diagnostically relevant information is not only provided by the images the radiologist looks at, but from many other sources. These include patient records, risk factors, additional imaging of other disciplines (most notably, pathology). It can be suspected that computer visualizations and adequate information presentation may bridge many clinical gaps, like e.g. between radiology and pathology, or radiology and surgeons. Diagnoses and therapy decisions may both be positively affected by such integration.

1.2 Imaging Breast Cancer

When images are the subject of analysis, two central aspects have to be considered upfront: the first concerns the relationship of the entity that is imaged and the imaging method, and the second concerns image analysis and presentation of analysis results with the aim of diagnostic interpretation. These questions are the central topic of the methods described in this thesis and will be asked in subsequent chapters. Before, aiming at a better understanding of the potential contribution that computer support can provide for clinical tasks that require the interpretation of images, the object being imaged and the imaging processes will be examined closer in this introductory chapter.

The breast as a developing organ will be described to sensitize the reader for the time frame and spatial scale on which early development of cancer and its precursors emerge. The most influential external and internal factors causing breast cancer provide a general idea of the mechanisms in cancer development. The different imaging tools are introduced with descriptions of their respective areas of optimal applicability. Current developments drawing from the knowledge gained about the molecular pathways of carcinogenesis will briefly be described, but the focus will be on the areas breast cancer screening, diagnostic imaging, and treatment decision making tools.

1.2.1 The Mammary Gland

The female breast is an organ that develops and differentiates in distinguished stages during the lifetime of a woman. The developmental stages can be described in terms of the morphological appearance of the milk producing lobules (Russo et al. 2004). The developments through these stages manifests in the differentiation of the lobules from poorly differentiated ones in the virginal breast of a nulliparous woman to fully differentiated lobules in the mammary gland of lactating women. One major stratifying feature of women towards their susceptibility to breast cancer growth is consequently their parity, because many endocrine processes involved in the full differentiation of the mammary gland positively affect the risk. The risk of malignant breast diseases is additionally linked to the duration of the individual developmental phases. Most notably, an early menarche and a late menopause increase a woman's lifetime risk of breast cancer. Oppositely, childbearing and breast feeding are known to lower the risk — but only if the woman is not too old upon first child birth. Women with a late first pregnancy (after the age of 35) are known to bear an increased lifetime risk to acquire breast cancer than nulliparous women.

The molecular pathways behind these effects are predominantly linked to the hormone estrogen, which is produced in the ovaries, but to a large amount also in the breast itself through two dominant pathways that synthesize estrogen from precursor substrates (ibidem, page 92). Estrogen promotes the proliferation of healthy breast tissue as well as neoplasms. In this sense, it is essential in the maturation of the breast. It should be noted, though, that the estrogen level of the body is of little importance in breast cancers given their capability of de novo synthesis of estrogen from circulating precursors in the breast stroma and epithelium. In fact, the tumor estrogen level is the same in pre-menopausal and post-menopausal women, although the circulating estrogen decreases substantially after menopause. In addition to estrogen inducing cell proliferation, estrogen is also a chemical carcinogen and an agent inducing mutations.

In addition to estrogen, the cyclical differentiation and regress of the breast's parenchymal structures during the menstrual cycle in itself contributes to the cancer risk. Each cell division during differentiation of the milk-producing units called the terminal duct lobular units bears the risk of DNA defects (IOM (Institute of Medicine) 2012).

Looking at Fig. 1.7, the terminal duct lobular units and lobules are seen colored in purple. Each individual small lobule is connected to a milk duct. Between menarche and menopause, in

each menstrual cycle a proportion of these lobules is built, and most of them destroyed again if no pregnancy prevents this through elevated levels of prolactin, which is expressed in the enlarged pituitary gland, and activating the transition from ductal to alveolar composition of the mammary gland. The relevant change with respect to the susceptibility for breast cancer is the full differentiation. It has been hypothesized that cells of the ductal units that have not undergone this process of full differentiation are more capable of developing into cancer cells than those that underwent the process (Russo et al. 2004).

1.2.2 Epidemiological Relevance

One in ten women will be afflicted with breast cancer in her life[4]. Every minute, one more woman is diagnosed with breast cancer worldwide. Every fourth minute, one woman dies from it. In a global perspective, cancer incidence is expected to rise due to a surge of detections in developing countries which is an effect owing to the adaption of western lifestyle including, most importantly, the habit of smoking. Today, breast cancer occurs among all cancers reported with a higher proportion in high income countries (12.5% of all cancers by organ) as opposed to low income countries, where only 7.7% of all reported cancers are breast cancers. However, it is expected that the proportion of breast cancers among all cancers will increase almost twofold in the developing world opposed to the western countries (Economist Intelligence Unit 2009). Breast cancer is also fatal: globally, every third woman diseased with breast cancer will die from it. Again, in low income countries, the fatality is higher than in high income regions (56.3% vs. 23.9%, respectively, Economist Intelligence Unit (ibidem)).

The scientific community looks back at many years full of successful projects to improve early detection, achieve a more robust diagnosis with non-invasive or minimally invasive methods, and to treat breast cancers without disfiguring surgery. The effect of screening on the detection rate is clearly visible, as well as the improved methods to detect early cancers. The reported incidence of invasive cancers has increased in the 1980s, and remained largely constant, while the number of reported smaller and more diffuse in-situ cancers continues to grow. It remains obscured to the scientist how many of the added number of detected small cancers are due to better detection methods, and how many stem from a higher (yet always unknown) baseline number of such cancers in the population. Assuming that reports on death causes are reliable, the absolute number of women dying of breast cancer remains constant.

Today's standard to report mortality figures is to provide either age-adjusted numbers of individuals saved, or numbers of life years saved. In the first case, the crucial statistical cornerstone is that for mortality figures, a "reference population" is chosen from all that are in the analysis, and is stratified into a number of age groups, and the cancer deaths for each age group are expressed as proportions from the whole. For each other population to be compared in the study, the same stratified proportional numbers are calculated, and the cancer deaths are then adjusted to match the population age profile of the reference population (J C Bailar et al. 1986; John C Bailar et al. 1997).

In contrast, in the second approach, the emphasis is on the number of life years saved, by assuming average longevity figures. A woman cured from breast cancer at age 65 with a life expectancy of 70 years would add 5 years to the total, a woman saved from breast cancer at age 40 would add 30 years. (Breslow et al. 1988).

Clearly, there is a normative aspect in the choice of the calculation. The first choice (age-adjusted cancer deaths) implicitly treats each life equally "worthy", counting the deaths prevented independent of expected years alive after cure. The calculation according to the second choice in contrast will result in higher performance figures for any program that saves more younger

[4]Male breast cancer is less common. Of all breast cancers, 99% are detected in female breasts, and 1% in male. This thesis only considers breast cancer detection, diagnosis, and treatment in the female breast.

Figure 1.7: Breast anatomy. Fatty tissue (orange in the picture) encloses the parenchymal structures, consisting of the terminal ductal lobular units (TDLUs, brown-purple), from where the ducts extend to the nipple. To the left, the rib cage with the pectoral muscle delimits the breast tissues. Connective tissue, the Cooper's Ligaments, extend from cranial beneath the skin surface and the subcutaneous adipose tissue towards the nipple, and from caudal as well. Additional connective tissue stabilized the breast tissue, extending from the rib cage to the skin. Lastly, usually between 8 and 12 compartments consisting of ducts and TDLUs are wrapped in connective tissue. These structure-giving elements are not depicted in the figure. Creative Commons Attribution 3.0, Patrick J. Lynch, medical illustrator; C. Carl Jaffe, MD, cardiologist.

individuals than one that works better for the higher age groups in the population. It could be stated that the first calculation favors health systems that focus on treatment, while the second is geared towards early detection and prevention.

Age-adjusted numbers of lives saved are for example reported annually by the American Cancer Society (American Cancer Society 2013a), and they are the more common reference standard. Unless otherwise indicated, mortality figures henceforth are expressed in this sense.

1.2.3 Imaging Methods

It has sometimes been coined that the only clinicians that really *see* breast cancer are pathologists. They have the tools to look at individual cells, and with different methods to stain pathological specimen, they can even assess some functional aspects of individual cells. Still, what pathologist look at is a temporal freeze-frame of a *process* that is cancer. It is impossible today to characterize the complete genetic status of cancer cells in a clinical setting, though efforts are being made to categorize breast cancer by their genetic signatures. In this light, radiological images will remain clinically important also in diagnostic decision-making and therapy preparation and monitoring.

Having described the structures of the breast that will change naturally or through carcinogenesis, we will now turn to the in vivo imaging of these structures. Several methods have been established in clinical imaging of the breast that have particular advantages and disadvantages. Not all of them are relevant to the methods and approaches described in this thesis. Therefore, this section does not aim to provide an overview based on the clinical relevance or cancer detection rate. Instead, the imaging methods central to this thesis will be described at a greater level of detail and put into clinical perspective.

The different imaging methods exploit different physical properties of the breast tissue to generate image contrast. Mammography uses the interaction of low-energy photons with the electron shell of atoms, magnetic resonance imaging builds upon the excitation of spins to generate magnetization effects, ultrasound imaging measures the reflextive and transmissive properties of materials when excited by ultrasonic waves, and nuclear imaging methods predict the location of radioactive decay events of a radiotracer injected into the body. More details on imaging physics of the most prevalent methods and how breast cancer may be detected by them follow subsequently.

The physical resolution and the signal to noise ratio of a given imaging system determines which of the breast structures and pathological features of lesions will be visible. Ductal carcinoma in situ (compare Sec. 2.2.2) may serve as an example. Digital mammography excels in the depiction of microcalcifications, small solid crystals. Digital mammography is even capable of showing the shape of the crystals, and thus to differentiate between calcium oxalate (an indicator of benign lesions), and carbonated calcium hydroxyapatite, more likely associated also with malignant type proliferations (Baker et al. 2010; Tabár et al. 2007, 2008). Despite the high detail, in many cases the true extent of the disease is not well appreciated in mammography alone, hence other imaging needs to be employed. Ultrasound can as well show the microcalcifications, but it is impossible to differentiate the shape. On the other hand, tissue changes obscure to mammography may be visualized for example using ultrasound shear wave elastography, an imaging technique sensitive to tissue stiffness which is known to increase around a tumor. Lastly, the neovascularization of the affected tissue is best depicted using contrast enhanced breast MRI, where calcifications are never seen, but the gross volume of the lesion is often much better appreciated.

The following more detailed description of imaging methods does neither in ordering nor in extent reflect the frequency or importance of their clinical use, but orients itself on the further subject of thesis contributions. Sorted by clinical relevance, mammography would be followed

by ultrasound and MRI, and all other imaging methods are restricted in their application to very specific clinical needs.

X-Ray Mammography

X-rays (low-energy photons) interact with the electron shell of atoms. They may be scattered or absorbed, which leads to noise, or to attenuation in the ray of the absorbed photon. Passing and attenuated photons are detected behind the imaged object, which is placed in the ray direction of the x-ray radiation source. This results in an image contrast depending on the thickness and permeability of the tissues imaged. For mammography, low-energy x-rays are required so that soft tissues with small density differences can be distinguished.

X-ray mammography is the current standard of clinical breast cancer screening, and to a large extent also for diagnostic imaging. Notably, this is not the result of an informed decision in the light of the complexity of the disease and a hypothesis about why and how x-rays should be employed to image a particular trait of breast cancer, but rather the result of chance discoveries and a tradition that has formed out of it (compare Sec. 3.4.1 for a short discussion, as well as Mukherjee (2011)). The recommended equipment is the full-field digital mammography system (FFDM), but sometimes, "digital film" (a digital detector plate sensitive to x-rays) may be used in conventional film mammography equipment, replacing the analog film. Analog film screen mammography is today mostly applied in private practices and is no longer considered a standard of care, and can for example not be employed in certified mammography screening centers in Germany.

The image acquisition, however, hasn't changed in terms of patient positioning: The breast is compressed between the detector plate and a compression paddle, and this arrangement can be tilted to acquire two angulated views of the same breast. Clinical routine today is to image once in cranio-caudal (CC) compression (horizontal), and once in oblique medio-lateral (MLO) direction (45° rotated, extending the field of view into the axillary tissue). Based on the two views, a finding that is seen in both projections can be located up to a certain accuracy in the breast tissue volume.

Breast MRI

The MRI Signal. The signal generation in magnetic resonance imaging is based on the excitation of the spin-1/2 of protons (effectively the nucleus of the hydrogen atom), which are available in the tissue in abundance. Quantum-physically, a spin-1/2 proton is magnetic, hence it reacts on radio-frequency (RF) pulses of the proper frequency by tilting its magnetization vector. Those radio-frequency (RF) pulses are generated by the transmit coils of a magnetic resonance (MR) scanner. Manipulating a sufficient number of protons using RF pulses can be employed to achieve a macroscopic rotating magnetization effect that can be measured with receiver coils that are tuned to the expected frequencies. The signal received is called the free induction decay (FID), a mixture of exponentially dampened sinusoids, where the amplitude of each sinusoid corresponds to the amount of nuclei resonating on that frequency. Generally, water and lipids are the dominant contributors to the mix, and unless a specific metabolite with a particular resonance frequency needs to be assessed, by far the largest part of the detectable FID (free induction decay) is not recorded in imaging applications to keep the readout time as short as possible.

For actual image generation, a much more sophisticated orchestration of RF pulses is required to achieve excitation of protons in a spatially resolved fashion. Also, for modern, fast image acquisition pulse sequences, the reconstruction of the received raw signals imposes additional algorithmic and computational demands. Furthermore, in contemporary MRI scanners, multiple transmit and receiver coils need to be operated and fused to speed up the process of image acquisition.

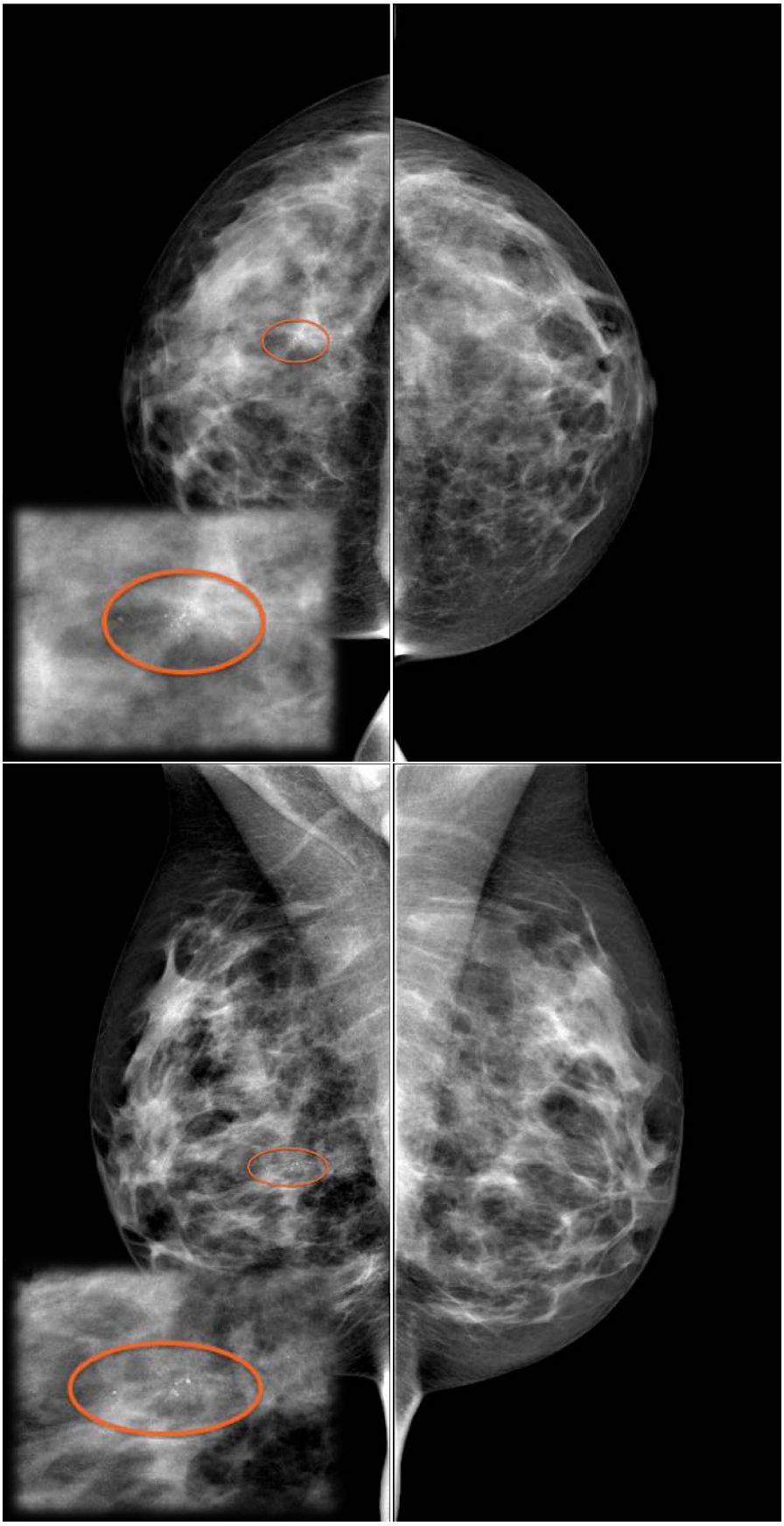


Figure 1.8: Left and right CC and MLO mammograms of a patient with extensive disease in the right breast. The only visible trace is a cluster of microcalcifications presenting as bright white spots.

In MR spectroscopy, the FID is typically recorded for between 0.5 and 2 sec. From the Fourier-transformed FID, the concentration of metabolites can be quantified. With a healthy reference, pathologically elevated or diminished concentrations of metabolites can be identified in a spectroscopic multi-voxel image. The underlying mechanism by which metabolites appear as different peaks in the spectrum is the slight change of the proton resonance frequency in their local magnetic environment (shielded or exposed) in the metabolite. The diagnostic utility stems from the knowledge that for example Choline, a metabolite involved in the construction of cell walls, is usually elevated in proliferating cells.

Breast MR spectroscopy is intensely researched in only a few places worldwide, with leading contributions of Mountford et al. (2012) and Ramadan et al. (2012), who go beyond typical 1D spectra, acquiring 2D single-voxel COSY spectra[5] of less than 1 cm^3. Also, a more robust application of 3D-^{1}H-MRSI for breast applications is researched in which after the acquisition, the reconstruction grid can be shifted to reduce partial volume effects (Baltzer et al. 2013; Gruber et al. 2011).

Figure 1.9: 18-channel breast coil. Images courtesy of Siemens Healthcare, Erlangen, Germany.

For optimal breast imaging, a dedicated breast coil is indispensable. Since breast imaging protocols usually comprise a plenitude of different MRI contrasts, like contrast-enhanced dynamic acquisitions with a T_1 weighting, anatomical images using a T_2 weighting, diffusion images, etc., time is a critical factor, and modern developments address this with coils that are suited for parallel imaging, which speeds up the image acquisition by a factor of up to four. The breast coil is placed on the MRI table, such that the woman can lay down onto it, facing towards the floor (so-called prone positioning). The breasts are fixated in two recesses of the coil using soft paddings. A typical image characteristic when using a breast coil is excellent contrast in the area of the breasts, but a strong reduction of signal intensity along with decreased signal-to-noise ratio towards the spine. For use in biopsies, breast coils usually also offer replaceable parts of the recesses, such that biopsy devices can be attached for image-guided minimally invasive procedures.

Contrast-Enhanced MRI. While in the future, MR based methods for breast imaging may become available that reduce or substitute the administration of contrast agent for lesion visualization, today's clinical method of choice in diagnostic breast MR imaging is contrast-enhanced MRI. A bolus of MR-visible contrast agent, Gd-DTPA, is administered intravenously in a typical concentration of 0.1 mmol/kg body weight. The contrast agent (CA) reaches the breast approximately one minute post injection. MR images are acquired before ("baseline scan") and after the contrast agent (CA) arrives in the breast. CA is confined to the vascularity since the molecule is too large to permeate through the vessel walls. In the case of a cancer, CA will still accumulate in the tumor-surrounding tissue, because growing tumors satisfy their nutritional and oxygen demands by expressing vascular growth factors in abundance. Many new

[5] COSY = COrrelation SpectroscopY; used to assess spin coupling in molecules using a proton spectroscopic experiment. Keeler (2006) gives a nice introduction of this and other MR spectroscopic experiments.

vessels will grow in the vicinity of the neoplasm, and since the speed of growth prevents mature vascularization, the vessel walls will be leaky. This again allows the CA molecules to transfer into the intercellular matrix, the interstitium. Depending on the leakiness, the contrast agent either continuously accumulates during the MR image acquisitions, or it comes to a halt, or even washes out again if the permeability is high enough. In this order, the wash-out behavior is indicative of benign, borderline, and malignant lesions. Usually, the wash-out is considered together with the amount of CA accumulating in the region before the wash-out.

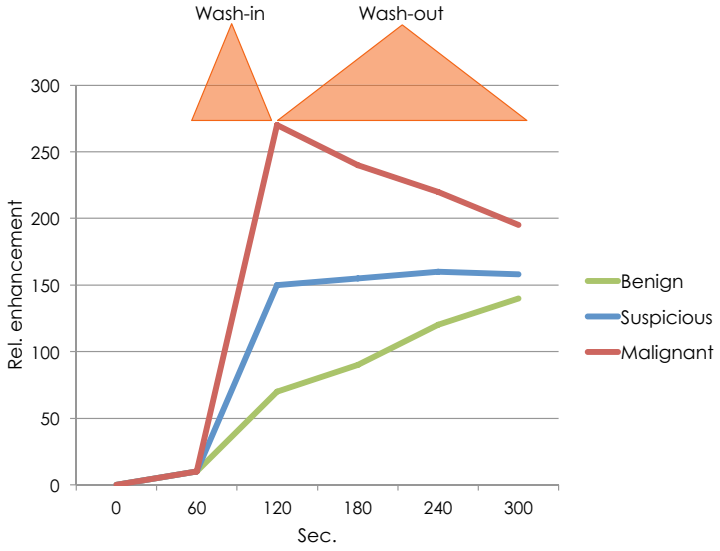


Figure 1.10: Typical curves of contrast enhancement accumulation. Units of the abscissa are seconds, counted from the time of contrast agent administration. Ordinate is in relative enhancement units (percent over baseline enhancement). Benign lesions usually exhibit low enhancement that progresses. Malignant lesions show a strong initial enhancement and a marked loss of enhancement ("wash-out") after the peak. Suspicious lesions show intermediate maximum enhancement associated with a plateau in the late phase.

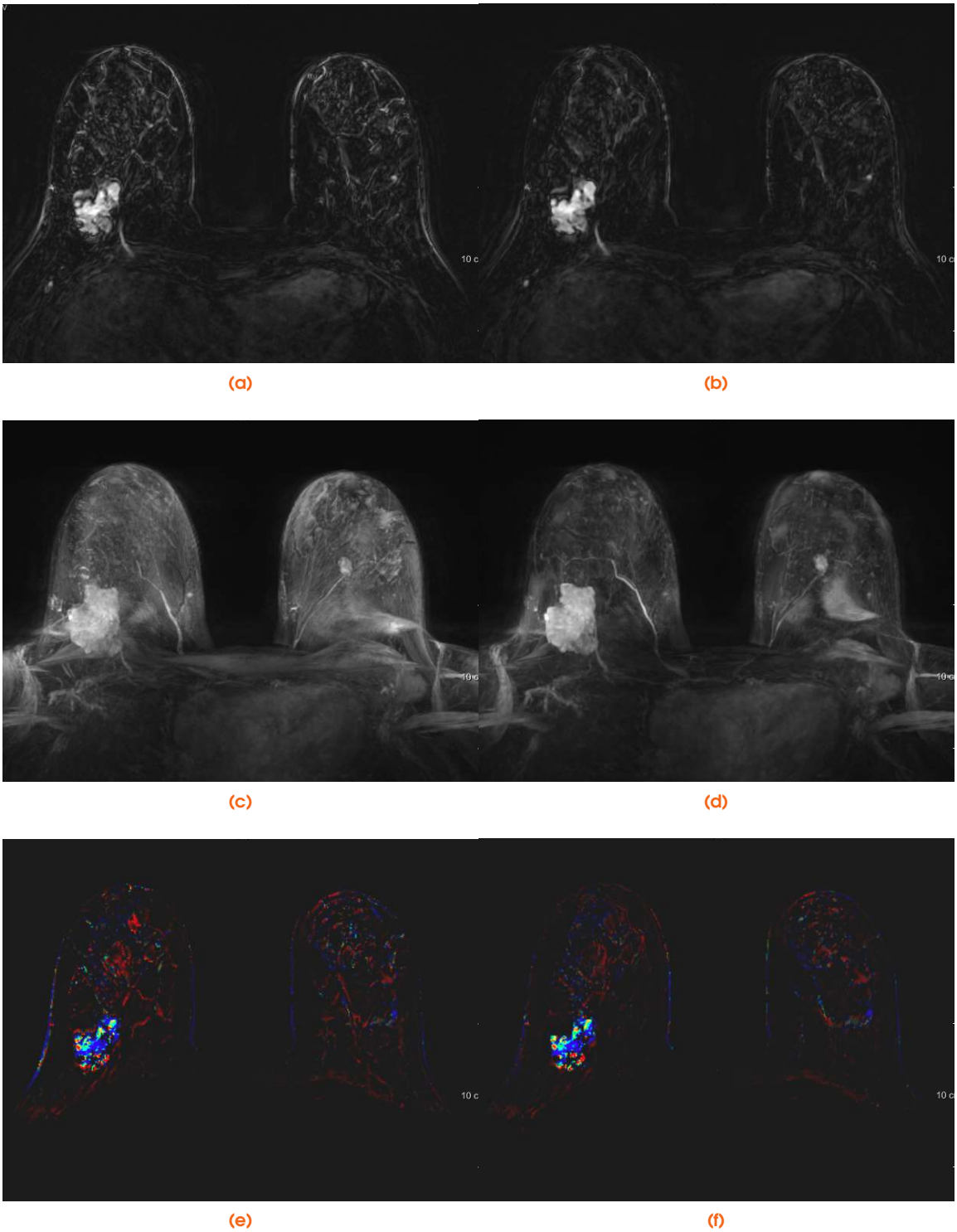
Gd-DTPA is one of the most prominent contrast agents for MR angiography and cancer imaging. The long name is Gadolinium-Diethylene-Triamine-PentAcetate with the structural formula $A_2[Gd(DTPA)(H_2O)]$. Introduced in 1987 as the first MR contrast agent, the strong paramagnetic properties of the molecule reduce the T_1 relaxation time yielding hyperintense regions in T_1-weighted MRI where it accumulates. The paramagnetic property is caused by the seven unpaired electrons with parallel spins on the outermost shell of the Gd atom.

Other Gadolinium based contrast agents are Gd-DOTA or Gd-BOPTA, which is more stable and therefore safer to use, since Gadolinium is toxic to the human body if not bound, and may cause nephrogenic systemic fibrosis in patients with weak kidneys. Studies have shown the superiority of Gd-DOTA over Gd-DTPA for cancer detection (Martincich et al. 2011), but Gd-DTPA remains to be a very frequently used MR contrast agent.

Gadolinium complexes are hydrophilic, implying that they cannot pass intact cell membranes and will accumulate only in the interstitium (and intravascularly). Also, it is excreted quickly and without interaction in the body through the renal system.

Typical Image Analysis for DCE-MRI. Clinical evaluation of contrast-enhanced breast MRI can be performed in different manners. Without dedicated computer support, a rapid automatic, endless succession of the five to ten time points before, during, and after administration of contrast agent can be replayed in the so-called *cine mode*, while the radiologist navigates through the data. Enhancing areas will "flash" from dark to bright.

Before applying any computer-based automated analysis, preprocessing steps are normally executed to improve image quality. This can include motion correction, which has been an area of intense research in the past years. A summary of approaches can be found in Boehler, Zoehrer, et al. (2011), and Boehlers PhD thesis is a good starting point for readers with



(a) (b)

(c) (d)

(e) (f)

Figure 1.11: Demonstration of the influence of motion correction on typical image processing algorithms. This is a case with subtle motion. Left: before, Right: after motion correction. Rows: (1) Difference images of early enhancement. (2) Maximum Intensity Projection (MIP) images of the before. (3) Color maps calculated from early relative enhancement magnitude and wash-out behavior. Opacity encodes strength of wash-in, color encodes type of wash-out. MR images courtesy Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands.

a deeper interest in the topic. Motion correction shifts and deforms all time series with respect to a reference time point. Usually, the first post-contrast time point is selected as a reference. Depending on the motion correction algorithm, the parameters of rigid or non-rigid transformations are optimized to minimize a distance metric between target and reference image. There is again a variety of choices available. The result is a time series where anatomically similar regions are at the same spatial location throughout the series. This removes artifacts in subtraction images, but more importantly also errors in the quantitative or descriptive analysis of the data, as detailed below.

Using minimal computations, a difference image (defined as the voxel-wise absolute value of the subtraction) of any post-contrast time point and the native, unenhanced time point reveals areas with uptake as bright spots, while all other areas are close to zero. A typical setup is to calculate not only the absolute difference image for the wash-in phase, but also a subtraction image between maximum enhancement and late enhancement, quantifying the washout. Here, it is required not to take the absolute value, but units of relative enhancement to differentiate between more benign lesions that show continuous uptake after the initial phase, and malignant-type enhancement patterns that are characterized by a negative slope after the initial enhancement. Commercial breast MRI reading software today usually employs schemes of color-coding of the enhancement characteristics such calculated (Dorrius et al. 2011). An example for a color map generated by assigning the transparency channel to the relative peak enhancement, and the color (green, blue, red) to the wash-out strength, is seen in Fig. 1.12.

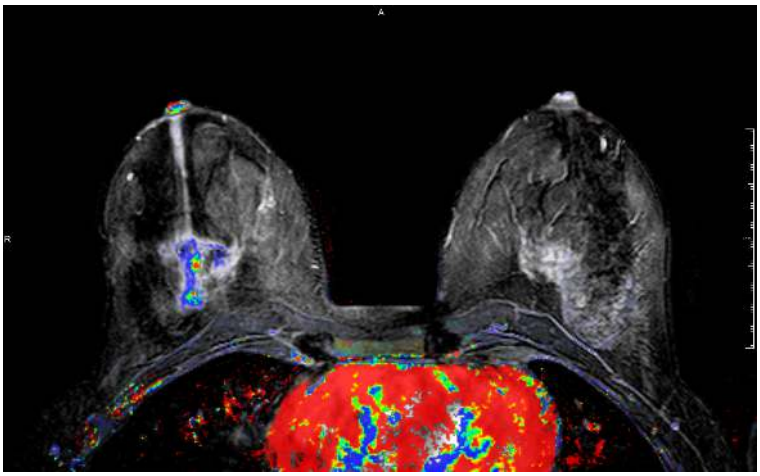


Figure 1.12: DCE-MRI of the same patient shown in Fig. 1.8. The contrast agent wash-in and wash-out characteristics are overlaid on a subtraction image that emphasizes the areas of maximum vessel leakiness.

Still, even such maps leave much of the image interpretation to the reader; in particular, objective and reproducible diagnostic results will be limited without quantitative descriptions of the contrast agent wash-in and wash-out behavior. Descriptive curve parameters are a remedy: the slope of enhancement, the time to the maximum enhancement and its magnitude, and the slope of the wash-out are most prominent parameters assessed in enhancement curves. Their value ranges, however, depend on many parameters that vary from clinical site to site, and even within one site between patients and examinations. Real quantitative evaluations require a contrast agent reference which is sometimes taken from the aorta enhancement.

More physiologically motivated are pharmacokinetic models that assess the time curves using a model of exchange between tissue compartments. Prominent examples are the models by Tofts et al. (1991), by Brix et al. (1991), and by Larsson et al. (1990). They yield k_{trans}, a transfer constant describing the leakiness of the vasculature, and ν_e, the extracellular leaky volume fraction which is permeated by the vasculature.

It has been seen that for DCE-MRI, the timing of the MRI sequence is of crucial importance both for the descriptive parameters and for the pharmacokinetic models (Laue et al. 2010). Therefore, hard-to-control variability is introduced through the missing knowledge of the exact timing of contrast agent administration together with the image acquisitions.

New Developments. While generally the marked wash-out behavior has proven to be a specific predictor of malignancy, cases with plateau or progressive enhancement are indeterminate. Recent research hence begins to focus on the wash-in phase in itself. It has recently been shown that the time it takes the contrast agent to travel from the administration site to the cancer location is a strong predictor of malignancy (R. D. M. Mus et al. 2012). To gain these insights, MRI pulse sequences had to be developed that do not read out the entire signal for each image, but share some of the data intelligently from acquisition to acquisition. By this, one of the acquisitions of a typical T_1-weighted dynamic acquisition that usually takes 60–110 sec, one such acquisition may be replaced by the fast sequence, fitting up to 20 frames of the about 4 sec long acquisitions in. Besides the cited novel descriptive parameters, the accuracy of the pharmacokinetic modeling can be increased (Laue et al. 2010), with the additional promise of a much shortened overall acquisition time if the conventional scans may be removed.

MRI pulse sequences that don't rely on contrast agent are increasingly being proposed as alternatives particularly aiming at serial MRI for screening purposes in risk populations. In principle related to MR spectroscopy, one noteworthy approach to breast imaging utilizes a longer acquisition of the excitation response, the FID — just long enough to quantify parameters of the water and lipid resonances. This so-called High Spectral-Spatial (HiSS) imaging is still in clinical evaluation, but in comparison to the dominant breast MRI technique, contrast-enhanced MRI, it appears to be a promising future alternative (compare Fig. 1.13).

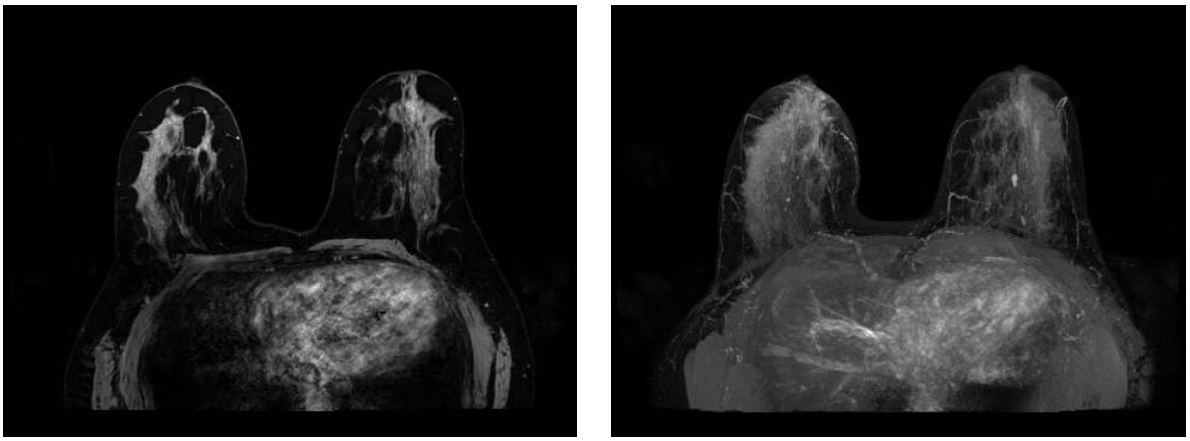


Figure 1.13: HiSS example images. Left: one slice of the volume data, in-plane resolution 480×480 voxels, imaged on a 3.0 T scanner. Right: MIP of the same volume. The contrast in the images is derived from the peak height of the water resonance reconstructed for each image voxel from the temporally Fourier-transformed FID. Images courtesy University of Chicago Department of Radiology, Chicago, IL, U.S.A.

With HiSS, features of the water and lipid resonances can be converted into high-resolution volumetric images. Voxel sizes in 1.5T MRI HiSS is in the order of $1 \times 1 \times 2$ mm and took about 1 minute at highest parallelization (M. M. Medved et al. 2010); it has also been demonstrated that higher in-plane resolutions and multi-slice acquisitions are feasible with this technique (M. Medved, Karczmar, et al. 2012; M. Medved, Newstead, et al. 2010).

Such developments move MRI-based screening of a high-risk population into reach. On the other hand, specific computer support tools to sort and read the increasing volume of breast MRI become indispensable. It is offered in the form of computer aided detection and diagnosis systems (CAD; cf. Sec. 2). One current direction of research is to detect findings automatically, and also automatically assess them based on pattern search and classification algorithms. For

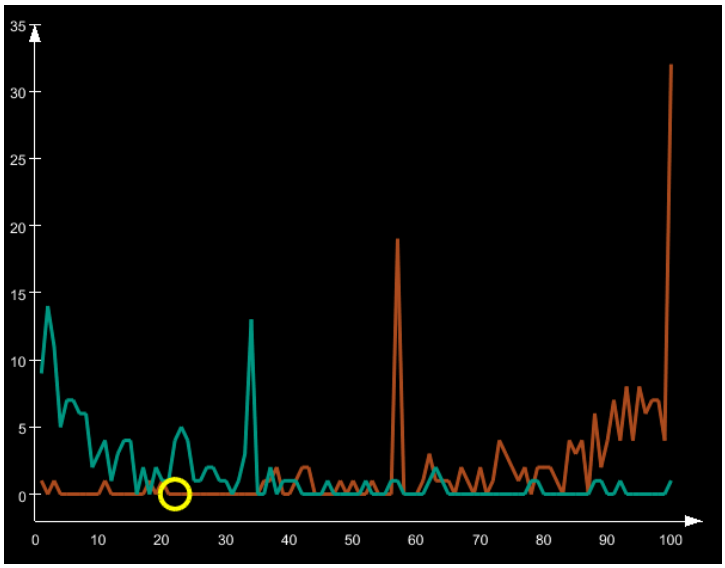


Figure 1.14: A machine learning score histogram from an early approach in decision support using a Support Vector Machine trained on descriptive kinetic features of DCE-MRI. The green and red curves visualize the number of benign and malignant training instances that received a certain score (score is on x axis). The circle shows the score of the currently investigated lesion.

example, the assessment result can be condensed into a score that corresponds to the malignancy, the risk, or any other clinically relevant information used in decision making. Such a score could either be provided to the radiologist numerically, or it may be encoded into a color map, which is perhaps a suitable choice to direct radiologist attention to suspect locations in MR images. If the score numerically rates a finding, it may be employed to sort the list of all findings, enabling the radiologist to navigate from finding to finding by relevance. It is here, where machine learning helps to generate the scores.

Together with a visualization of the decision basis of the score's origin, e.g. in the form of a score histogram (Fig. 1.14), decision aid can be provided in an interactive way that offers even more information to the clinician than only a score. Such a system may also provide most similar benign and malignant examples from the learning database — lesions corresponding to the two classes with predictions close to the current one (Giger et al. 2013).

More detail on several aspects of clinical breast MRI will be provided in the following chapters tailored to the specific needs. Readers with a deeper interest in the physical foundation of MRI and pulse sequence development in general and for breast MRI in particular may refer to descriptions that are accessible for novice readers (Elster et al. 2001; Levitt 2001; McRobbie et al. 2003), as well as to standard textbooks, which require more dedication to appreciate their wealth (Callaghan 1993; Haacke et al. 1999). There are also comprehensive textbooks focusing on the clinical utilization of different MRI techniques (Reimer et al. 2003). For specific information on suitable MR-based (and other) imaging methods in breast imaging, W A Berg (2006) can serve as a concise reference.

Other Approved Methods

Besides x-ray mammography and breast MRI, methods currently in clinical use encompass

Ultrasound Hand-held ultrasound imaging is used as a reliable and readily available imaging method. Ultrasound probes for different applications exist, and yield high-resolution images. Several advanced ultrasound imaging techniques have been developed from the base technology: *Doppler ultrasound* imaging allows to visualize the flow, for example to assess tissue vascularization. *Ultrasound shear wave imaging* (US-SWI) is capable of

imaging tissue stiffness, which appears to be correlated with malignancy. *Automated breast volume sonography* (ABVS) offers the benefit of separated image acquisition, for example by technologists, and image analysis by the radiologists. This technology is also beneficial for computerized analyses.

Tomosynthesis Multiple x-ray mammographies are acquired with slightly tilted angles. A limited three-dimensional reconstruction into a semi-volumetric data set can be calculated from the raw images. It is expected that morphology is better depicted and that tomosynthesis improves on the greatest limitation of mammography, the superposition of different densities, eventually obscuring lesions behind dense parenchymal tissue. From the tomosynthesis images, the typical four-view set of mammographic images can be deducted computationally.

Advanced Breast MRI Several MR sequences including fast contrast enhanced imaging of the CA uptake phase (e.g. TWIST), functional imaging using diffusion-weighted imaging (DWI) sequences, high spectral and spatial resolution imaging (HiSS), and metabolic imaging using MR spectroscopic imaging (MRSI) are under investigation in trials, or explored in routine protocols.

Positron Emission Mammography The physical principle of positron emission mammography (PEM) is the same as for positron emission tomography (PET), only a different detector geometry is used. The process for clinical PEM/PET imaging relies on F^{18}-FDG (fluorodeoxyglucose or 2-Deoxy-2-[^{18}F]fluoroglucose), radioactively labeled glucose. F^{18} is taken up in cells with glucose metabolism, but due to the F^{18} occupying the place of a OH component in glucose, the molecule cannot be further metabolized in the cell and remains there until it radioactively decays. Some PEM protocols today achieve reliable detection and differentiation at dose levels in the order of a conventional 4-view screening mammography examination (Thakur et al. 2012). With proper reconstruction algorithms, the commercialized PEM scanner allows for the detection of breast lesions approximately of the physical size of individual detector elements, which is in the order of about 2 mm diameter. PEM has been reported to operate at 91% sensitivity and 93% specificity, and to reveal approximately the same number of additional disease as for example MRI (PEM: in 14% of patients; MRI: 15%). PEM reportedly showed improved specificity compared with MRI, but ongoing studies have to substantiate this claim (Specht et al. 2012).

1.3 Complexity in Breast Cancer: Prospects

With many imaging methods under development, and with genetic tests being researched intensively, an increase of complexity when dealing with breast cancer care is very likely (ASCO – American Society of Clinical Oncology 2014). The increase of complexity is not limited to the direct actioning of the involved clinical personnel when they perform their clinical routine. They will need to oversee a widening range of possibilities in diagnosis and treatment as new imaging methods and new drugs become available. With new tests being invented and clinically tested, the actors will have to include more knowledge and more information sources.

Beyond evidence-based decision making, early results indicate that patient-doctor communication produces psychological effects in personalized medicine that will not be neglegible anymore, because they influence the decisions patients (and doctors) make. "The most personal part of personalized medicine is often the interaction that takes place between the physician and patient", is what Kelly et al. concluded, finding that the patients who were informed of their poor prognosis by a specific genetic test chose the *less* aggressive treatment compared to the group who did not undergo the genetic testing, who chose chemotherapy more frequently. Hence, a relevant research direction in personalized cancer care is the study of human-human interaction considering the level and type of information.

From all evidence collected above, it will be evident that "ignoring the complexity" is no promising option today, and will become even less so in the future, when a widened range of information sources becomes available and produces effects that are perhaps unpredictable from today's perspective. Instead, a thorough study of all aspects of the complex system is suggested.

1.4 Thesis Overview

In this thesis, we will present methods and algorithms that support clinical tasks from screening to intervention, covering topics in automated and robust image analysis and predictive modeling of breast shape changes with and without a shape target.

We will consider the different levels of clinical image based breast cancer care, starting in Chapter 2 with the computer-based detection and characterization of masses and non-mass enhancing lesions in dynamic contrast-enhanced breast MRI. Methods will be shown to assess lesions using the dynamic change of their texture. While this work depends on the lesion to be segmented prior to the texture analysis, the following section will present an approach to detect and outline suspicious regions including segmental enhancing lesions that don't present as a mass. This novel hybrid computer-aided detection and diagnosis algorithm utilizes the bilateral symmetry of contrast enhancement together with several kinetic and morphological features that are specifically tailored to non-mass-like enhancing patterns.

Thereafter, Chapter 3 focuses on the subject of workflow improvements in screening and diagnosis, proposing intuitive and efficient mobile gesture-based systems. The iPad as a multi-touch capable portable device is employed in a novel general paradigm in which the quality of high-resolution image display on primary monitors is combined with efficient and intuitive gesture control on a personal device, supported by a context awareness mechanism reacting to both the application and the user's environment and location. This paradigm will be applied to improve breast MRI workflow, and to enhance the capabilities of a mammography workstation prototype.

Finally, Chapter 4 brings topics from the area of therapy into the center of attention. Support systems for image-based interventions will be presented that improve the workflow of minimally-invasive imaging-guided biopsies by fusing the diagnostic, high-resolution MRI with the interventional scan. Second, to support the planning of open surgery, a biomechanics-based simulation of breast deformation may help surgeons in the future to visualize the position of the index lesion more naturally.

1.5 Data Used in the Thesis

For this thesis, several disjunct data sets have been used. This is owed to the fact that the thesis has been developed from publications submitted between 2008 and 2013. The publications regarded a diversity of image analysis and modeling problems, reflected in the chapters of this thesis. It is the nature of the problems, that no one data set was suitable to be used throughout the thesis. Also, some of the developments have been partially supported by projects and research grants, like for example the HAMAM project (EC FP7) and the MARIUS project (Fraunhofer Society). Other data has been provided for limited and specific use only.

The following table lists the data sets and data sources together with their application in the thesis.

Table 1.1: Data sets used in the thesis. RadboudUMC (Nijmegen, NL); Charitè (Berlin, DE); BRRH (Boca Raton Community Hospital, Boca Raton, FL, U.S.A.); Own data (Bremen, DE)

Type	Origin	Purpose
DCE-MRI non-mass lesions	RadboudUMC	Non-Mass CADe/CADx (Sec. 2.3)
DCE-MRI mass lesions	Charitè	4D Texture CADx (Sec. 2.4)
DCE-MRI biopsy	RadboudUMC and BRRH	Deformation simulation model
MRI prone/supine	Own data	Biomechanical model development (Sec. 4.3.1)

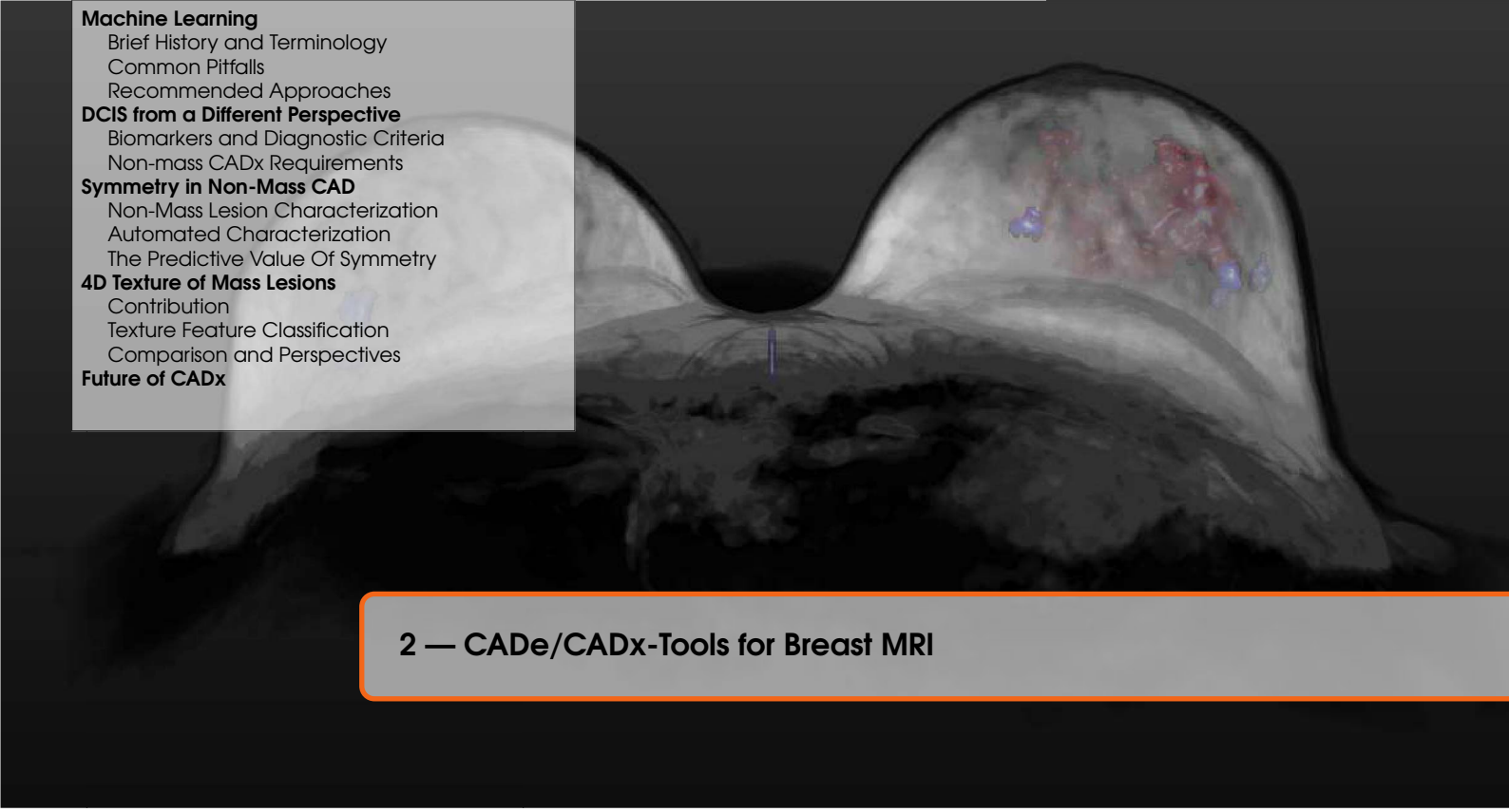
With the non-mass lesion data set, an evaluation of the mass lesion classification would have been possible under the premise of additionally obtained voxel-accurate segmentations of all

mass lesions in the data set. Likewise, the mass lesion data might have been used as control cases in the development of the non-mass lesion detection and characterization algorithm. For both data sets to be joined, however, the prevailing problem of data normalization would have been to be solved, which was not central to the two algorithms. Since both algorithms utilize descriptive parameters to capture the contrast agent uptake behavior in the tumor to derive features for machine learning, a sufficient degree of uniformity in the MRI acquisition protocol is mandatory. This level of uniformity can by experience never be assumed in data originating from different sites. Further, the mass lesion data has priorly been used for classification experiments based on 2D Haralick texture features, so that these older results can directly be compared to ours.

The volunteer data acquired for the biophysical modeling presented in Chapter 4 was particular in that it served the purpose of understanding the patterns of tissue deformation and movement between prone and supine acquisitions of the breast, once suspended in a breast coil (prone), and once freely moving around the chest either with the arm inside or outside the MRI field of view. No clinical data suited this purpose, and since volunteer data in general does not show any pathologies, nor is it acquired after administration of contrast agent, it serves the sole purpose of the study it was acquired for.

For the purpose of model building for deformation simulation between the diagnostic and interventional imaging involved in MRI-guided biopsies of the breast, no preconditions had to be met, hence data from different sites has been used for model development and validation.

Machine Learning
Brief History and Terminology
Common Pitfalls
Recommended Approaches
DCIS from a Different Perspective
Biomarkers and Diagnostic Criteria
Non-mass CADx Requirements
Symmetry in Non-Mass CAD
Non-Mass Lesion Characterization
Automated Characterization
The Predictive Value Of Symmetry
4D Texture of Mass Lesions
Contribution
Texture Feature Classification
Comparison and Perspectives
Future of CADx



2 — CADe/CADx-Tools for Breast MRI

COMPUTER-aided detection and computer-aided diagnosis (abbreviated CADe and CADx) with unsupervised methods have been researched over many years. "Unsupervised" in this case does not refer to the training process of the method (as in supervised versus unsupervised machine learning), but to the fact that no human input is required during execution of the CADe/CADx algorithm. Such algorithms are usually trained for their objective in a supervised fashion.

Where a distinction is required, methods that only find and mark abnormalities in an image will be called CADe methods. In contrast, CADx methods characterize abnormalities. In today's speech, the two terms are not always used in clearly separated meanings. In both cases, the result of the algorithms may be a continuous or discrete value that expresses the confidence that the detected abnormality is relevant, or a rank, or a label. Since the usage of the two terms is not always clear, and since practical implementations cannot always be assigned into one group, many references refer to CADe and CADx by only one common abbreviation, CAD.

CADe/CADx has a record of proven clinical utility for certain groups of cancers and some of the imaging modalities, and acknowledged difficulties for others. We will consider the computerized detection and diagnosis of a particularly challenging lesion type seen in DCE-MRI. This malignant, but non-invasive precursor lesion of invasive breast cancers develops in the milk ducts and towards the lobules. This cancer, ductal carcinoma in situ (DCIS), has not penetrated the basement membrane, but already begun to invoke changes of the surrounding tissue and vascular growth that make it visible in DCE-MRI. DCIS is by its imaging appearance and growth characteristic usually counted among the segmental or non-mass-like lesions.

Other cancers that look similar are in-situ cancers of the terminal duct lobular units (TDLUs), named lobular cancer in situ (LCIS). Also, benign breast tissue alterations have a similar imaging appearance in breast MRI, namely fibrocystic changes, fibroadenoma, and similar lesions. They can be distinguished only relative to DCIS/LCIS acquired under similar conditions, making it very hard to differentiate between benign and malignant non-mass-like enhancement patterns in a robust automated computer system. Sec. 2.2 summarizes the clinical importance, and the obstacles to a computerized differentiation of DCIS from other types. Since among the

The chapter image shows a volume rendering of an extensive ductal carcinoma in situ. The coloring corresponds to the wash-in and wash-out behavior of the lesion, where opaque colors encode high wash-in rates. The wash-out determines the hue, with red being the highest wash-out rate.

malignant-type non-mass enhancing lesions, DCIS is eight to ten times more prevalent than LCIS, we will in the following not particularly address LCIS, although epidemiological figures suggest a more likely development of invasive cancers from LCIS. In situ cancers account for about 10%–12% of screen-detected cancers (Barchielli et al. 2005; C. I. Li et al. 2005), while DCIS alone accounts for 20% of all cancers, according to Kell et al. (2005).

In screening, the detection of signs of these lesion types is a challenging task, since the lesions need to be detected when they are small. This requires the detection of signs of the disease in screening mammography based on microcalcifications and regional densities. Independent double reading by two radiologists or computer-aided detection is indispensable for this task. Upon a suspicious detection, a diagnostic workup either with mammography or adjunct modalities is required. To take therapeutic decisions, MRI is an often employed modality to assess extent and multi-focality of the disease. DCIS diagnosis on DCE-MRI is known to be hampered by ambiguous enhancement characteristics of DCIS and LCIS, often mimicking that of benign lesions, and particularly in the presence of fibrocystic changes or fibroadenomas the clear distinction of DCIS/LCIS from the benign findings is tedious and error-prone.

While non-mass like lesions account for a substantial, but minor part of all cancers in which treatment after successful delineation is promising, the larger fraction falls into the category of mass enhancing lesions, which are, compared to non-mass like enhancement, easier to detect and segment automatically. Their morphology is much less diverse, but there are many typical enhancement patterns that are known to distinguish benign from malignant lesion types. Mass-like lesions are the focus of the second methodological development that will be described in this chapter, complementing the DCIS detection and characterization method.

This chapter is structured as follows. We will review common pitfalls and best practices of machine learning in biomedical applications in Sec. 2.1. Building on the description of the pathology of DCIS given in the previous chapter, an outline of the complex task of its clinical differentiation in Sec. 2.2 prepares the presentation of a novel approach to the challenging task of detection and differentiation of non-mass like lesions in Sec. 2.3. The proposed approach makes use of features and insights derived in this chapter. Finally, we look at the knowledge gained from the analysis of mass lesion texture features utilizing a novel extension of texture features to time-resolved dynamic contrast enhanced breast MRI data (Sec. 2.4). The aim here is to provide a starting point for a scanner- and protocol-independent robust characterization of mass lesions.

Parts of this chapter have been published in

- Jennifer Loose et al. (2009). "Assessment of texture analysis on DCE-MRI data for the differentiation of breast tumor lesions". In: volume 7260. DOI: 10.1117/12.812971. URL: http://dx.doi.org/10.1117/12.812971
- Abhilash Srikantha, Markus T Harz, et al. (2012). "Symmetry-based detection of ductal carcinoma in situ in breast MRI". in: *European Journal of Radiology*. Volume 81, pages 158–159
- Abhilash Srikantha, Markus Harz, et al. (2013). "Symmetry-Based Detection and Diagnosis of DCIS in Breast MRI". in: *Proc. SPIE Medical Imaging*. Volume 8670. DOI: 10.1117/12.2000061. URL: http://dx.doi.org/10.1117/12.2000061
- Lei Wang, Markus Harz, et al. (2014). "A robust and extendable framework towards fully automated diagnosis of nonmass lesions in breast DCE-MRI". in: *IEEE International Symposium on Biomedical Imaging*. Volume accepted

2.1 Machine Learning in Medical Applications

Generally, machine learning techniques have been applied to problems in image analysis for a long time. Within the biomedical sciences, it forms a research direction in its own right. While clinical usage of machine learning to provide computer-predicted diagnoses usually faces obstacles in the formal approval processes worldwide, it is beyond any doubt that computer-based evaluations of images can surpass the performance of human observers for certain applications. In particular in mammography screening, CADe already serves as a second reader to detect suspicious areas. In automated volumetric breast sonography (ABVS), the computer-based detection and diagnosis of lesions reaches an area under the receiver operating characteristic (ROC)[1] curve matching that of the best experienced radiologists (Tan, Platel, Twellmann, et al. 2013), and also in DCE-MRI, computer-generated color maps suggestively colored red for malignant areas and blue to green for less suspicious regions provide indispensable support in rapid image reading.

The motivation to review machine learning methods and tools for their evaluation in more detail is that despite (or because of) the vast number of text books and references, there does not seem to be a general guidance for those who want to apply, not develop, machine learning techniques in a way that prevents common pitfalls, and evaluate their results in a comparable and publishable fashion. Of the introductory publications, Domingos (2012) is closest to what I am aiming at, though Domingos focuses more on an overview of relevant terminology. To provide a safe selection of classification methods, and corresponding best practices for evaluation on a theoretically well-founded basis is my purpose in the first sections of this chapter.

2.1.1 A Brief History of Machine Learning in Medical Science

The terminology in the area of artificial intelligence (AI) has many ambiguous and synonymous terms, so that some explanatory remarks and clarifying definitions are required upfront.

Artificial intelligence (AI) is commonly understood to encompass the processes of perception, decision taking, and conduction of successful actions. It entails the full information processing chain including sensors that provide information to actuators with feedback that modify the environment.

Within the broad area of AI, two subcategories can be delineated that cover perception tasks and decision taking, respectively. The first comprises pattern recognition (PR) and machine learning (ML), the second knowledge discovery and data mining. Methods in the latter two categories attempt to extract groups, anomalies, and dependencies from data, while PR and ML are the area of our interest. Pattern recognition (PR) is usually conceived of as the process of assigning a potentially fuzzy label to an input value. This is for example implemented in variants of neural networks. Lastly, machine learning (ML) is the process by which learning from experience, examples, or observations is usually associated (inductive learning). From the data, a model (concept, rule, ...) is built. In some variants, improving the model with more experience (examples, observations) is conceptually integrated, or easy to add.

Classification can hence be understood as an implementation of PR and ML: Learning from data to assign a label to input values. Classification has to be differentiated from regression which as well serves the purpose of assigning a learned output to input, but while classification methods assign nominal class labels to the input value, regression learners output continuous values, also

[1] The name of this criterium dates back to World War II, where it was examined how reliably the operators of radar receivers would distinguish blobs on the radar image. Their performance was tracked in the receiver operator curve, and their ability to detect hostile movement was described in the receiver operating characteristics, ROC. It was only in the 1970s that the generality of the performance characterization approach was brought to attention and used in the medical field (Tape 2014).

named scores sometimes. Note that the purpose of machine learning and classification is not to learn an exact representation of the observed input data (the training instances), but to find a generalized model of the process generating the observed data or behavior.

An important aspect of classification and regression learning algorithms is that they need to be supervised: for the machine learning method to build the model, it is indispensable to provide the correct answer to it, together with a measurement standard to quantify the prediction performance. This practically means that the learning step is guided by knowledge of the true decision of each example case in the training data. Unsupervised methods, on the other hand, are called clustering methods. Clustering commonly utilizes calculations to quantify the difference between examples to partition the feature space into compartments that contain examples that are similar with respect to the difference calculation. One simple difference calculation method is the Euclidean distance when the examples are described by coordinates for example in $\mathbb{R}^n$, but many more approaches have been proposed. In the prediction phase, unseen examples are tested against the clusters to yield a decision on how similar they are with respect to each.

A short introduction into machine learning with particular attention to the most prevailing pitfalls one usually encounters in ML on medical images will be presented to equip the reader with a solid understanding of common terms and best practices in machine learning. This introduction refrains from rigorous definitions of the machine learning task or mathematical details of each algorithm, but subsequently summarize the relevant terminology used in the remainder of the section.

Definition 2.1 — Features and Feature Vectors. A feature is a specification of an *attribute* describing an instance and its value. The possible values are from the *domain* of the attribute.

We distinguish categorial and continuous attributes, and among the categorial attributes nominal (no ordering between the values) and ordinal ones (having an ordering). For practical purposes, attributes are usually composed into an ordered set, so that features can be represented by their values in the same ordering. *Feature vectors* are then vectorial representations of the feature values for all attributes in the set, often denoted by the boldface letter $\mathbf{x}$. The terms "feature" and "attribute" are often used as if they were synonyms, which is not strictly true, but since in the machine learning literature, several expressions are commonplace, we adhere to them rather than to a rigorous usage of nomenclature.

Definition 2.2 — Predictor, Classifier, Regressor. An algorithm that predicts a class or a score of an unknown example by utilizing a model learned from known examples and their true class or score, is called a *predictor*.

We speak of a *classifier*, if the method returns discrete decisions, i.e. assigns each test example to one of the learned classes.

We speak of a *regressor*, if a continuous score is returned.

Classifiers are trained to approximate an unknown, potentially non-linear function f that connects the feature vector $\mathbf{x}$ to a classification y:

$$y = f(\mathbf{x})$$

When we refer to more than one instance, for example to describe a training or test set in machine learning, the (row) feature vectors of the instances are collected into an ordered set of feature vectors $\mathbf{X}$, and a (column) vector of corresponding classifications, $\mathbf{Y}$ captures the ground truth. The machine learning task is then to estimate the unknown f in $\mathbf{Y} = f(\mathbf{X})$.[2]

[2] In some machine learning textbooks, the classification task is formulated using $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \{\pm 1\}$

Today, the separating lines between pure discrete classifiers and regressors with continuous output are no longer drawn very sharply. Instead, for many regressors it is commonplace to map their potentially unbound output score to some interval and threshold it to derive a categorial decision, and on the other hand to employ classifier-internal error estimations to express its certainty in the decision.

The machine learning methods employed to learn patterns in the data changed over time, driven by computer capabilities and theoretical developments in the field. Giger et al. (2013) summarize the current status of breast image analysis for the detection and diagnosis of breast cancer. Their review shows that two methods are predominantly employed in machine learning tasks in breast cancer: support vector machines and neural networks. We will briefly review both before we introduce alternative methods that present robust and easy to use, yet very time-efficient alternatives to support vector machines and neural networks.

Artificial Neural Networks While types of computing schemes that today are subsumed into the category *artificial neural networks* (ANNs) can be traced back to before the emergence of computers[3], they gained a huge interest only with the increasing capabilities of computers and the crucial developments leading to reinforcement learning (Werbos 1974), by which categorial problems in machine learning pointed out by Minsky et al. (1969) could be solved. There is no commonly shared definition of a neural network, but in practice, those statistical models that (1) have a set of "weights" that are (2) adapted in a learning algorithm to (3) approximate non-linear functions of their input are called "neural".

Many types of neural network architectures exhibiting these traits have since been developed, with sometimes bold claims respective to their closeness to learning mechanisms of the human brain. Reinforcement learning has been employed to strengthen the connections in neural networks, bearing a resemblance to the human brain not only in terminology. However, training of ANNs and in particular their re-training to include the evidence of new samples in the model remains a crucial process that can be computationally intensive.

We cannot review the vast literature on neural networks and their types and topologies here. Instead, we briefly describe the type that prevails in applications of supervised learning, which is the feed-forward multi-layer perceptron. Such a neural network comprises input neurons, hidden layer neurons in one or more layers, and output neurons. The modeler of the topology needs to choose the number of hidden layers, the number of neurons per hidden layer, and the wiring of neurons. Input layer neurons are as many as there are features to observe, and the number of output neurons is determined by the number of classes to distinguish. Each neuron basically implements a function that takes all its input values (along the connections coming in from other neurons), and calculates the output value that is emitted to all connected neurons.

The training process then in principle needs to find function parameters for all neurons that are consistent with the training data, i.e. when any of the training data is fed into the input layer, the output layer neuron corresponding to the true class should display the highest value.

The topology of the classical types of ANNs — the number of input nodes, hidden layers, and their connectedness — determined their areas of application and their performance. A suitable topology was not always known a priory, and changing the topology required tedious re-training. Integrated approaches were hence described to change the topology at runtime (e.g. ART and ART2). In other approaches, the network topology was optimized by for example using genetic

where $\mathcal{X}$ is the domain from which cases x_i are taken, and the y_i are labels. Note that in this case, no assumptions are made on $\mathcal{X}$; for example, the x_i could be words, images, numbers, ... The machine learning task is again to infer $f : \mathcal{X} \to \{\pm 1\}$ from the input data. Compare (Schölkopf and Smola 2001).

[3] McCulloch et al. (1943) proposed a mathematical model of a neural network that lead to the development of the Hebbian learning, an attempt to describe the plasticity of information storage in the brain (Hebb 1949) and later the perceptron (Rosenblatt 1958).

algorithms (which were, in turn, a second hype in machine learning that we will not consider further), trying to find solutions for static problems.

Another fundamental difficulty is that the number of required neurons grows with the number of parameters in the model. Growing numbers of features will slow down the training process. Also, the number of training instances needs to exceed the number of parameters by orders of magnitude. Because of these restrictions, ANNs are today very much less commonly employed than before, and have been replaced by faster, more flexible and robust, and easier to interpret approaches. Applications of ANNs have a long history of success for example in the scoring of mammographic densities and microcalcification clusters (see for example Brake et al. 2000; Papadopoulos et al. 2002; Sahiner et al. 1996; Varela et al. 2006; Y. Wu et al. 1993), which may be caused by the fact that the "neural network hype" fostered their application in about the same years that digital mammography became widely employed. Today, other machine learning techniques are also explored in abundance.

So far, only the most prominent type of classical ANNs, the multi-layer perceptron, was considered, which for example Amato et al. (2013) focused on exclusively in a recent review paper on the applications of artificial neural networks in medical diagnosis.[4] Many other neural network types of classical topology have been described, like the more complex architectures implementing short-term memory, with forgetting, or with online training capabilities.

Those networks became larger, slower, and were hard to train. In 2006, however, the developments in neural network based research have been reignited by results in training methodology. Hinton et al. (2006) proposed a learning algorithm that allowed to train a hybrid network with multiple hidden layers efficiently, outperforming competing machine learning algorithms in hand-written digit classification. Bengio et al. (2007) and Ranzato et al. (2006) contributed to the learning methods, extending it to other networks and ultimately leading to an applicable learning algorithm that is today considered the state of the art in pattern recognition and unsupervised extraction of meaningful information from the most complex data sets (Razavian et al. 2014; Sermanet et al. 2013). Two of the pioneers in this research, Geoffrey Hinton and Yann LeCun are today employed by Google and Facebook, respectively, underscoring the acceptance and applicability their methods gained.

Support Vector Machines Support Vector Machines (SVM, proposed by Cortes et al. (1995), see also (Vapnik 1998)), are another hallmark in machine learning, and like ANNs, they are today widely used and extremely well researched.[5] They are rooted in statistical learning theory and are thus a theoretically well-understood method. For many applications in practical situations, however, they have been superseded by faster and more robust, easier to parameterize methods. Introductory mathematical treatments of the basic SVM algorithms as well as a motivation from statistical learning theory can be found in Burges (1998) and Schölkopf and Smola (2001). I largely follow their notations in the following description.

To understand the principle of SVMs, we start by showing how the most simple form of SVMs (linear SVM for linearly separable data) are applied once the model for the data is obtained. We begin by assuming training instances $\mathbf{x}_i$ from the data domain, $\mathbf{X} = \mathbb{R}^d$. The decision function of the linear SVM takes as input the unknown $\mathbf{x}_0$ and calculates on which side of the separating hyperplane the instance is. Let $(\boldsymbol{w}, b)$ be the canonical representation of the separating hyperplane, with $\boldsymbol{w}$ being normal to the hyperplane and pointing into the direction of the positive class, and b the distance from the origin. Then, the decision function is

$$f(\mathbf{x}_0) = (\langle \boldsymbol{w}, \mathbf{x}_0 \rangle) + b) \tag{2.1}$$

[4] The most well-known neural network simulator is the Stuttgart Neural Network Simulator, SNNS, which runs on Linux and Windows systems and is available for free (`http://www.ra.cs.uni-tuebingen.de/SNNS/`). Besides the perceptron, virtually any other neural network topology can be created and trained, notably with the exception of deep learning networks.

[5] Cortes et al. (1995) have been cited over a 10,000 times.

It is the task of SVM training to estimate the class separating hyperplane. The learning algorithm solves a constrained optimization problem consisting of an objective function τ and inequality constraints. Given m training examples, the objective function is to minimize the length of $\boldsymbol{w}$:

$$\min_{\boldsymbol{w},b} \tau(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{2.2}$$

subject to the inequality constraints

$$y_i \cdot (\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geqslant 1 \quad \forall i = 1, \ldots, m. \tag{2.3}$$

This method leads to a maximum margin hyperplane, i.e. that particular hyperplane that maximizes the distance between the closest positive and negative instance. The margin can be calculated from the constraints; it turns out to be $2/\|\boldsymbol{w}\|$, which is why by minimizing $\|\boldsymbol{w}\|$ the maximm margin classifier is obtained.

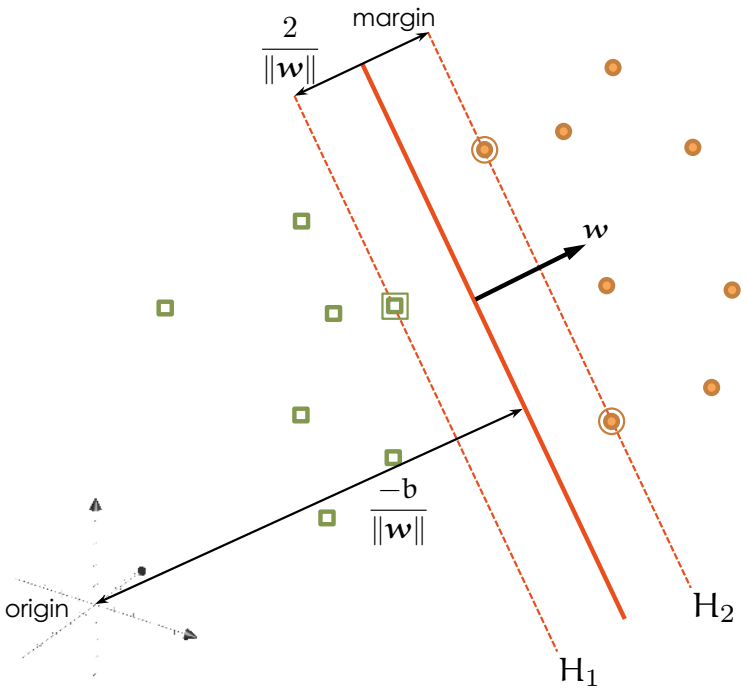


Figure 2.1: SVM terminology. Instances of a two-class learning problem are squares and disks. Support vectors are marked with a double line. They characterize the SVM fully and lie on the two hyperplanes H_1 and H_2.

Rewriting the optimization problem in a Lagrangian formulation, two advantages are brought about: (1) The constraints will become easier to deal with, and (2) the data — the $\boldsymbol{x}_i$ and labels y_i — will only appear in inner products between vectors, enabling the generalization of the method to the nonlinear case by using the famous "kernel trick". The Lagrangian is formed by introducing positive Lagrange multipliers to multiply the inequality constraints with and subtract them from the objective function, while adding the Lagrange multipliers for equality constraints. There are as many Lagrange multipliers as there are constraints, hence there is one for each $(\boldsymbol{x}_i, y_i)$. Adhering to the literature, we denote them α_i, $i = 1, \ldots, m$.

This leads to the Lagrangian

$$L \equiv \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{w} \rangle + b) + \sum_{i=1}^{m} \alpha_i, \tag{2.4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two instances (not necessarily a vector dot product since we didn't make assumptions on the domain of $\boldsymbol{x}$)).

L needs to be minimized with respect to $\boldsymbol{w}$ and b while the derivatives of L with respect to the α_i vanish. Convexity of objective function and of the point set fulfilling the constraints allows one to maximize the equivalent dual formulation, requiring to maximize L, this time requiring that the derivatives of L with respect to $\boldsymbol{w}$ and b vanish (Boyd et al. 2004). From this, it follows that $\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i$, and, substituting this into the Lagrangian, we arrive at the dual formulation

$$\hat{L} = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \tag{2.5}$$

Maximizing $\hat{L}$ with respect to α_i gives the solution $\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i$ in the linear separable case. Of the m Lagrangian multipliers, only those corresponding to *support vectors* will be greater than zero in the solution and lie on either H_1 or H_2 (compare Fig. 2.1). All others have $\alpha_i = 0$ and lie either on one of H_1, H_2 or on the side of the according plane. The support vectors hence describe the decision boundary, giving the name to the method (Wernick et al. 2010).

The Karush-Kuhn-Tucker conditions are satisfied for the SVM problem as stated above (see Burges 1998), and together with the convexity, it can be shown that solving the SVM problem is equivalent to finding a solution to the Karush-Kuhn-Tucker conditions. This allows to find a b which is not directly given by the outlined training procedure.

Still, so far only the separable case was considered. Through the introduction of a cost function, the constraints from Eqn. (2.3) on the solution can be relaxed where required. The cost function is introduced through slack variables ξ_i in the constraints, allowing some training examples to be on the "wrong side" of the separating hyperplane. Some function of the slack variables needs then to be added to the objective function that is minimized. Selecting the penalty term to be $C \sum_i \xi_i$ makes it disappear in the dual Lagrangian, and only adds an upper bound on the Lagrange multipliers as $0 \leqslant \alpha_i \leqslant C$.

Further, we also wish to learn separating hyperplanes without the linearity assumption. A trick described by Aizerman et al. (1964) helps to do this without many changes to the approach. We have seen how by writing the problem in the dual Lagrangian, the training data $\boldsymbol{x}_i$ only appear in inner products defined on the data domain, and we didn't need to make any assumptions on either the data domain or the inner product definition. The trick — the "kernel trick" — consists of defining a mapping from the data domain into an Euclidean space $\mathcal{H}$ where an inner product is defined. The mapping $\boldsymbol{\Phi}$ is

$$\begin{aligned} &\boldsymbol{\Phi} : \boldsymbol{X} \to \mathcal{H} \\ &\quad \boldsymbol{x} \to \hat{\boldsymbol{x}} := \boldsymbol{\Phi}(\boldsymbol{x}). \end{aligned} \tag{2.6}$$

Mappings can be defined that convert the instances into a space $\mathcal{H}$ of a different dimensionality. The inner product then needs to be calculated in $\mathcal{H}$ as well, but fortunately, $\langle \hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j \rangle = \langle \boldsymbol{\Phi}(\boldsymbol{x}_i), \boldsymbol{\Phi}(\boldsymbol{x}_j) \rangle$ can be written as $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for some kernel k that corresponds to the inner product in $\mathcal{H}$ via the mapping $\boldsymbol{\Phi}$ (Schölkopf and Smola 2001, p. 25).

Mapping into $\mathcal{H}$ has the advantage that no assumptions need to be made on the domain of the instances, except it being a set. In particular, no canonical inner product (vector dot product) needs to be defined on it. By mapping the instances into $\mathcal{H}$, the so-called *feature space*, which is Euclidean and hence has a dot product defined on it, one gains the freedom to choose a suitable dot product by choosing the appropriate mapping. Interestingly, there are certain cases

that allow to compute the inner product in $\mathcal{H}$ without explicitly computing the mapping $\mathbf{\Phi}$. The art is to come up with the required kernel k implementing a desired mapping. The kernel has to fulfil certain conditions to be admissible, which makes it difficult in the general case. The other way is also possible: choose the kernel first, and prove that it is admissible. Thankfully, some very useful kernels have been proposed early in SVM research, namely the polynomial kernel, a Gaussian radial basis function kernel, and others. If the data domain X is a subset of $\mathbb{R}^d$, the Gaussian kernel has the particular property to map the data into an infinite-dimensional $\mathcal{H}$. By choosing a suitable dot product — one that corresponds to a non-linear mapping from the domain of the original data into the feature space —, classes can become separable that are not in their data domain (Aizerman et al. 1964; Schölkopf 2001; Schölkopf and Smola 2001).

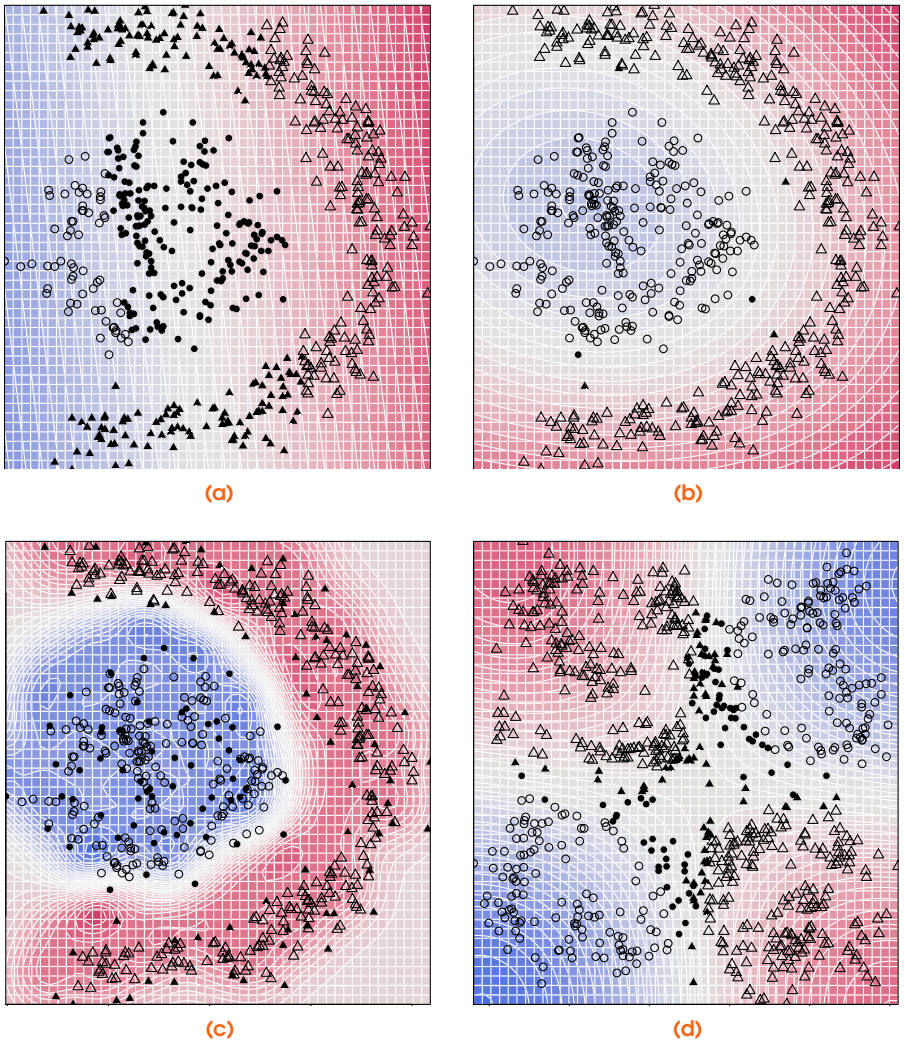


Figure 2.2: Different SVMs trained on different data sets. (a) A linear SVM is incapable to separate the instances in this data set. (b) A RBF SVM does a much better job. The small number of support vectors (filled symbols) indicate a predictor that generalizes well. (c) A RBF SVM overfitting the same data: note the high number of support vectors. (d) The XOR problem, which cannot be solved by linear classifiers. The RBF SVM shown in the figure again captures the class separation.

Already for linear decision boundaries (i.e. without mapping the data into a feature space), SVMs are much more robust than conventional linear classifier training algorithms like e.g. Linear Discriminant Analysis (LDA). LDA relies on a classical criterium to determine the model parameters, maximum likelihood estimation of the class covariances (Fisher 1936). SVMs, other than these methods, don't take into account sample points distant to the separating plane, but are designed to focus on the points that are harder to separate.

The frequently used SVM with a Gaussian radial basis kernel function takes as parameters the size of the Gaussian and the cost parameter C that has to be tuned to a application-specific compromise value. By experience, the generalization ability of SVMs (the bias-variance tradeoff) is very sensitive to the specific parameter selection. Hence, research has proposed parameter search procedures to yield (locally) optimal parameters (for example by Chapelle et al. (2002)), or genetic algorithms to optimize features and parameterization (Frohlich et al. 2003; Huang et al. 2006). Still no such automated local search guarantees to prevent the SVM from overfitting (compare Fig. 2.2). With sufficient expertise in the real-world domain from which the data emerges, and a thorough insight into the inner workings of the SVM, however, state-of-the-art classification results can be obtained.

Employing SVMs is in practice also often impeded by their requirement of a much larger number of training examples than are features in the model. Particularly in medical imaging, where the input may for example consists of hard-to-collect medical images together with corresponding pathologically obtained gradings, this is not always easy to meet. Feature selection procedures therefore are a frequent choice to keep the number of parameters low enough to ensure stable learning and to yield a model that abstracts from the training data and does not overfit. With this in mind, the high performance potential of SVMs has been shown in many studies where they are compared against other predictors both for classification and regression tasks (Meyer et al. 2003).

2.1.2 Common Pitfalls

Many of these problems, like overfitting, feature subset selection, and model search, have been tackled by more recent learning methods, most prominently by boosting (Freund et al. 1996; Viola et al. 2001) and bagging (bootstrap aggregation, Breiman (1996)) techniques. The Random Forest classifier is a general purpose machine learning tool of this category proposed by Breiman (2001). But also a simple statistical learning function, the Naïve Bayes method, provides a very robust and extremely fast method to deal with a high-dimensional feature space with almost no negative side effects in the presence of inconsistent, correlated, or useless features.

More detailed descriptions of both methods will be provided after a short introduction into difficulties often impeding machine learning in high-dimensional data. Particular to the situation in the medical imaging domain, study data sets are often small, and the desire to compare quantitative imaging biomarkers for their predictive power in decision making tasks leads to the *curse of dimensionality*. Machine learning in such conditions easily brings about *biased* classifiers if no attention to the evaluation methods is paid. Confronted with too many potential features and too few examples, *feature subset selection* is one way to countervail the curse of dimensionality, but at the cost of an additional source of bias. When all before has successfully been dealt with, the evaluation of the trained classifier requires *appropriate criteria*. These cornerstones of machine learning will subsequently be detailed and examined with a focus on practical advise.

Curse of Dimensionality

If many attributes are derived from the data, a high-dimensional feature space is implied, in which the machine learning method has to find and combine the most useful features. The classifier training task in many algorithms is to find decision boundaries between instances of different classes, compare Fig. 2.2.

To understand the problem that arises due to high feature space dimensionality better, some considerations will be presented. We will be looking at the situation where the number of features p exceeds the number of training examples, n, often denoted as the $p \gg n$ problem.

Machine learning may be thought of as the task of finding densities in feature space. It is this density estimation that is unsuccessful in high-dimensional spaces if the number of examples is too low. Density estimation can for a thought experiment be considered a histogram binning process. Suppose 100 samples are binned into 10 bins, then for the one-dimensional case, on average 10 samples are assigned to each bin, which may be acceptable. In two dimensions, only one sample is assigned to each bin, and in three dimensions, only every tenth bin has one sample, which is no longer a meaningful histogram. This has been called the "curse of dimensionality" (Bellman 1961).

In other words, the examples (instances, samples) don't occupy the feature space; they are sparsely distributed. When for example k-nearest-neighbor classification methods are applied, in commonly available implementations a distance function is employed under which all samples will look as if they are at equal distance to each other, differing only by quantities that cannot be distinguished from numerical errors due to the limits of a computer's representation of numbers. The distance calculation is then useless (Aggarwal et al. 2001).

A second, more mathematical thought experiment shows how intuition fails us when talking about high-dimensional feature spaces. If one considers a circle inscribed into a square, most area (about 75%) is occupied by the circle. We typically tend to extend this intuition into higher-dimensional spaces, but this is wrong.

! Let d be the dimensionality of the feature space. The volume of a d-dimensional cube is then described by $(2r)^d$ with r being the half edge length. A hypersphere with radius r in the same d-dimensional space has volume $\frac{2r^d\pi^{d/2}}{d\Gamma(d/2+1)}$; the Γ function is $\Gamma(n) = \int_0^\infty e^{-x}x^{n-1}dx$. Now consider the ratio of the two volumes while d increases:

$$\lim_{d\to\infty} \frac{1}{(2r)^d} \frac{2r^d\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2}+1)} = \frac{\pi^{\frac{d}{2}}}{2^d\Gamma(\frac{d}{2}+1)}$$

For $d = 1$, the ratio is one. For $d = 3$, it drops to only about $\frac{1}{2}$, and for $d = 6$, it is already smaller than $\frac{1}{12}$. The spheroid volume very rapidly vanishes against the cube volume, hence against all intuition most of the space is in the corners. For samples in this space it means that the probability of instances being "close" gets extremely low in high dimensional spaces.

Even worse for our trust into intuition, it can in the same thought model be shown that for an arbitrarily thin outer shell of the hypersphere, almost all volume of the hypersphere concentrates in this outer shell.

Another common intuition is likewise wrong: A Gaussian distributed variable in one or two dimensions has its probability mass in the center. Again, intuition tells us that this ought to be the case in higher dimensions as well, but the contrary is true: It can be shown that if a hypersphere in the center of the Gaussian is considered, with higher dimensionality, the probability mass is in the tails of the distribution. Note from above that the volume of a hypersphere approaches 0 when the number of dimensions increases. This is equivalent to saying that the major part of the probability mass is outside any hypersphere centered at the center of the Gaussian. A Gaussian in high dimensions can hence be considered a heavy-tailed distribution, while still the highest probability density is in the center. This observation impacts methods that are based on finding densities in the data, like for example maximum likelihood estimation.

The problems arising in and ways to deal with high dimensional data are today an extremely well researched topic, and once researchers are aware of the problem, a plenitude of methods is available that allow to deal with high-dimensional data and few examples efficiently[6], and

[6] See for example Kriegel et al. (2009) and the references therein.

we will see how high-dimensional feature spaces derived from lesions segmented in contrast enhanced breast MRI can be reduced and subdued to fast machine learning methods that are robust and successful.

Bias and Variance

In noisy, real-world data, a classification model usually cannot both fit the training data optimally, and simultaneously generalize well to unseen data (there is of course no theoretic argument why such a case should not exist). Usually, assumptions are made in the modeling process, e.g. regarding the number of degrees of freedom, or the general problem characteristics (linear or nonlinear? Which type of nonlinearity?). Regarding the number of degrees of freedom, there are two ways how a model can fail to generalize. Regard as a simple example the fitting of a polynomial to an observation of measurements. If the polynomial model has too few degrees of freedom, its predictions will be poor because of a lack of flexibility, and likewise, when it has too many degrees, it will overfit the noise in the training data and be a bad predictor on new data as well.

This behavior is expressed more formally in the statistical bias and variance of a predictor which can be obtained by splitting the prediction error into two summands. It is not the aim of this paragraph to reproduce the knowledge available in many excellent text books, for example by Bishop (1995). Rather, a good intuition shall be provided that helps to understand and critically review machine learning based biomedical research papers. For a simple understanding, bias is the systematic aberration of the average of multiple measurements from the true value, and variance characterizes the between-measurement distances.

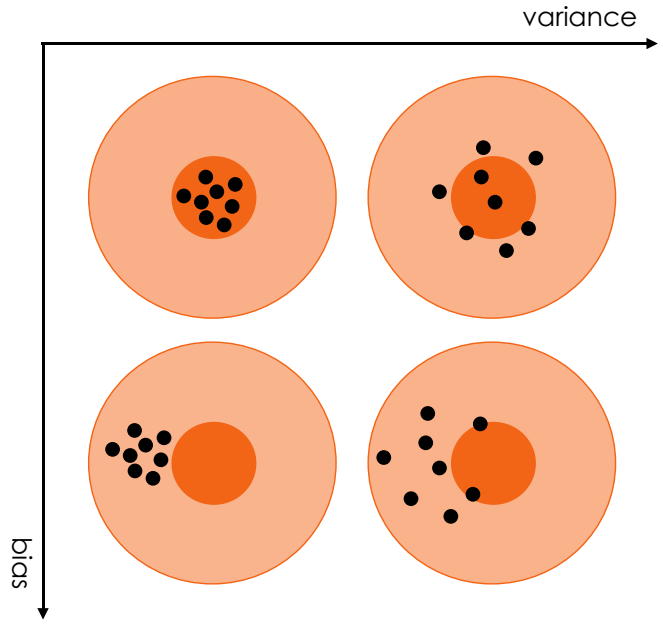
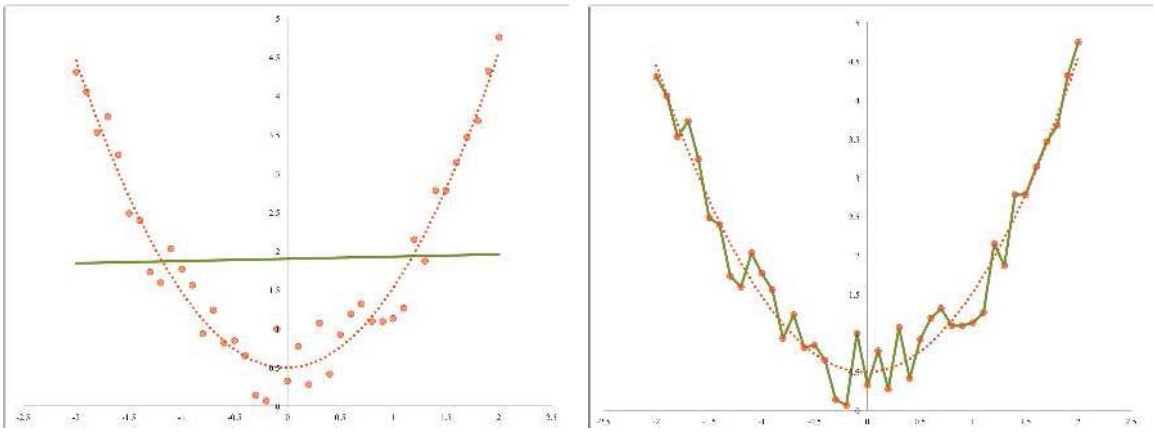


Figure 2.3: Bias and variance. Top left: low bias and variance. Lower left: high bias, low variance. Top right: low bias, high variance. Lower right: high bias, high variance. Bias and variance are in ISO 5725 termed trueness and precision.

Again considering a polynomial function as the generator to be modeled, the bias of an estimator of this function is the difference between the estimator's average value (on all data sets) and the true value of the generating function. The bias of a classifier is introduced in model building and describes how much the model estimates deviate from the true estimator on the training samples. Many features in the model help to minimize the deviation, leading to a low bias and a high-complexity model. Selecting only some features from a list of potential features reduces the model complexity, which is desired to generalize from the test set, but in

the extreme may not allow a good data representation because the real data characteristics may no longer be fitted, because a feature has been removed that performed poor on the data seen in training, but is important for the model transfer to independent test data.

The variance of the estimator quantifies the data dependence of the prediction error (between all data sets), or how the classifier generalizes from the training set. Allowing many free parameters may fit the training data very well (low bias), but the average error of predictions will increase. The classifier is in the extreme overfitted, and its performance depends on the test set: it has a high performance variance. The performance of a low-variance classifier, on the other hand, is the same on the training and the test set, but usually not very high on either.



(a) (b)

Figure 2.4: The two graphics show the extremes in model building exemplified on the estimation of a second degree polynomial $h(x)$ (dotted orange) that has been disturbed by white noise γ yielding measurements (orange dots). **Left**: A low-variance model. The green line is some arbitrarily chosen function $g(x)$ which clearly is equally wrong for any realization of $h(x) + \gamma$, which is expressed by a low variance. **Right**: A zero-bias model. The green line here is the linear interpolation between the measurements. Any new realization of $h(x) + \gamma$ will yield a different error, resulting in a high data-dependent variance (in fact, the variance equals the variance of the noise), while the bias is zero since all data points are matched exactly.

Feature Subset Selection Bias

It is common practice in machine learning to reduce the number of features to consider for the machine learning task. This is meant to prevent the predictor from overfitting the data, and sometimes also to reduce the runtime of the training algorithm by removing highly correlated features from the feature set. This can be done either in an unsupervised, or a supervised manner. The unsupervised process will rate the yield of a feature withholding the classification of the examples from the selection process, for example by assessing the type and characteristics of the distribution the feature follows, or by searching for clsuters in the feature domains. A further important approach is to assess the similarity of features for example expressed in the correlation coefficient and merge or remove similar ones. Note that similarity measures are prone to the "curse of dimensionality"-effect as discussed before, hence in some problems unsupervised feature selection may not be straight-forward. Supervised feature selection methods, on the other hand, know the features and the class assignment and can hence directly assess the potential benefit of a feature for classification. A higher bias can be expected from the supervised feature selection process opposed to the unsupervised.

An appropriate "wrapper" approach to supervised feature subset selection has been described by Kohavi and G. H. John (1997), who advised to hold back a testing sample to estimate the performance of the trained classifier on. However, when a data set X containing n examples is available for the machine learning task, and n is comparatively small against the number of features p, many researchers use the full X to select the most appropriate $p\prime$ features, and again use the full X, but with the reduced set of features, to conduct the predictor training and

evaluation in a cross-validated way. Sometimes, the feature subset selection and the classifier learning are encapsulated in the same cross-validation wrapper.

The problem in this approach is that it implicitly assumes that the value distribution in each attribute in the data represents the distribution of this attribute in the data the predictor is later applied to — an assumption that cannot easily be made. In fact, any data that is available practically can only be an unknown sample of the full data basis — unknown in the extent of its representativeness of the data basis. Attributes that are discarded from such a data sample during the attribute selection process may contain important information to predict unknown instances from the data basis. In practice, the result is thus a predictor with a performance on the training/testing data that is different from the performance on unknown instances from the data basis. Note that a similar over- or underestimation (bias) of the predictor is also possible if no feature selection is applied, but feature selection increases the bias.

Consequently, the approach to select features and train the predictor on the same data X received much theoretical criticism and critical comments also in the area of medical image analysis (Kupinski et al. 1999), and it can be proven that there is a substantial selection bias particularly if the number of training samples is small compared to the number of attributes (Jensen et al. 2000).

It has been examined in much detail how the selection bias in regression affects model building, and also how to avoid much of the bias (A. Miller 2002). The most important result here is that when selecting features by composing feature subsets and training a regressor on the subset, the bias gets stronger the more subsets are evaluated. In particular, in an exhaustive search, where one model is built for each conceivable subset from the feature set, the bias will be overwhelming. Much less of a bias is introduced if for example a forward selection procedure is used.

The situation is less severe for the task of classification compared to regression. Despite the same potential source of bias introduced by performing both feature selection and classification on the identical X, in practice only a negligible influence on the performance of the classifier is noticeable, due to the fact that the decision boundary is in many cases not affected by the selection bias. Singhi et al. (2006) derive this result for Bayesian inference, where it is easily visualized using artificial test data from normally distributed features. In fact, a bias in classification can be noticed in imbalanced training sets, but it still is very small even for a ratio of 1:4 in favor for one class.

Despite its reassuring conclusions, the study by Singhi et al. has limitations of practical importance: The model used for their analysis is a Naïve Bayes classifier with the assumption of normally distributed data in all attributes. Since the bias incurred by the feature selection process selecting from features with normal distributions is symmetric, in their model no bias occurs for classes with even weights, and even in imbalanced data, the bias is small. However, when considering real-world data, skewed distributions and attributes that are differently distributed are a very reasonable assumption. The bias will in these settings behave much less predictable, and might be much higher.

It is hence good advise to be cautious whenever any feature subset selection is required for the learning task. A remedy is brought about by applying feature selection and predictor training in a cross validated fashion. In this approach, only part of the data is used for feature selection and to train the predictor, while the remainder is used to validate the model such obtained. This is repeated multiple times, and the final predictor performance estimate is calculated by averaging over the multiple validation runs. In this methodology, the cross validation covers both feature selection and training. This is only feasible, however, if a sufficient number of instances is available. In any other case, explicit feature subset selection before classification model building is not possible without accepting a biased performance estimate.

Critera for Prediction Performance Evaluation

The starting point for any analysis of performance is a set of true results, often expressed as $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ of length n if there are n test examples. The y_k are from a fixed set of possible results, $\mathcal{Y}$, like for example nominal descriptions in the case of classifier algorithms (e.g. $\mathcal{Y} = \{$"malignant", "benign"$\}$) or in the case of regression algorithms, $y_k \in \mathcal{Y} = \mathbb{R}$. The prediction algorithm gives one prediction per test example, with the test examples represented in a second set $\mathbf{x}$ with the same number of entries and in the same ordering as $\mathbf{y}$. Intuitively, one y_k is the true result for the corresponding x_k.

Several criteria have been proposed and are in common practice to compare these two vectors and derive a scalar performance score for the prediction algorithm in question. Some of them, however, are subject to intense criticism, either because their value is doubted, or because they are being misused, or both. We will examine in particular the frequently used, but criticizable scores, and recommend those that find agreement in the literature.

In general, the criteria need to depend on the domain of $\mathbf{y}$, which can be binary, nominal multi-class, or numeric and continuous. Likewise, the predictor can independently yield binary, multi-class, or continuous numeric decisions. A common way to depict the relations between truth and predictions is the so-called confusion matrix[7] exemplified in Fig. 2.5.

		Prediction			
		benign	**malignant**	Sum	
Truth	**benign**	**tn = 403**	**fp = 197**	600	Specificity = tn/(tn+fp) = 403/600 = 67%
	malignant	**fn = 125**	**tp = 825**	950	Sensitivity = tp/(tp+fn) = 825/950= 87%
	Sum	528	1022	1550	
		NPV = tn/(tn+fn) = /528 = 88%	PPV = tp/(fp+tp) = 825/1022 = 81%		

Figure 2.5: The confusion matrix of a lesion malignancy likelihood predictor. Each entry in the central 2×2 matrix is a number of cases. The aim is to have as little cases as possible in the malignant/benign and benign/malignant combinations (shaded red).

From the confusion matrix, further relevant performance descriptors can be derived, most notably the true and false positive classifications and true and false negative classifications (tp, fp, tn, fn, respectively), which correspond to the matrix entries. Associating a positive prediction with prediction of malignancy ("the diagnosis has been made"), for example the true positive predictions are the lower right entry (825 cancer cases have correctly been classified as malignant). The sums can be used to calculate sensitivity and specificity which express the capability of the predictor to pick up the malignant cases (sensitivity) while also classifying benign cases correctly (specificity). In medical image analysis, high specificity is often weighed very high since with false positive cases additional costs are connected to confirm the diagnosis or reject it. Positive and negative predictive values express how good the predictor is at confirming the diagnosis or reassuring the patient of her health.

[7] The name might indicate that it is visible from the matrix if the classes are "confused" by the algorithm. It is also known as the contingency table, a term that is easier to trace back to its origins in the works of Karl Pearson.

Scalar Performance Criteria

Coming back to the scalar scores used to assess predictor performance, the most prominent example for the last category above (a doubtful criterion often misused) is perhaps the kappa coefficient, also written with the Greek letter κ. Originally designed to answer the question *if* two independent raters of cases with two possible outcomes agree by chance or not, kappa coefficients are often used not only to assess the agreement of more than two raters on cases with more than two outcomes, but in addition the assumption of rater independence is most often violated, and the coefficient (ranging from 0 to 1) is misinterpreted to indicate the level of rater agreement over chance. As Uebersax (2010) puts it, "one should be concerned about using a statistic that is the source of so much controversy".

The kappa coefficient is defined for two raters and two possible outcomes as follows:

Definition 2.3 — Kappa statistic.

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \tag{2.7}$$

with p_0 the observed probability of agreement between two raters and p_c the probability of chance agreement based on the assumption of complete independence and calculated from the marginal probabilities (ibidem).

If both raters judge the same number of cases, p_0 is calculated from the number of identical judgments divided by the number of all judgments. p_c is the part of the definition that gives rise to the criticisms, since it is calculated as if the two raters randomly choose from the two options. For both raters, their respective prior probability to choose the first and second option are independently calculated from how often they choose the first over the second, and combined into the probabilities that they by chance choose the same option. The probability of chance agreement, p_c, is then the product of the probabilities of chance agreement on the two options. The equation is not well-defined for some cases, for example if both raters rate all cases the same. In this case, p_c is 1, and no κ value can be calculated because the denominator vanishes.

Calculating the kappa value like stated in Eqn. (2.7) is done with the aim to adjust the observed agreement with the chance agreement to arrive at a more meaningful rater agreement characterization. However, calculating p_c from marginal probabilities assumes raters that guess completely on every case, and it assumes they guess with probabilities matching the marginal proportions of the observed rating. Both are assumptions that are very unlikely met.

Thankfully, Kraemer et al. (2002) looked at the possible valid applications of kappa statistics. It turns out that many practical judgments can still be made, as long as care is taken to not violate the assumptions of the κ calculation. Unfortunately, particularly in graphical machine learning tools like Orange and Weka (see above), the implementation of κ is not directly revealed, so that a careless use is too easily possible.

Another often used and misinterpreted measure of prediction performance is the accuracy.

Definition 2.4 — Accuracy. The accuracy of a predictor is

$$\alpha = \frac{tp + tn}{tp + tn + fp + fn} \tag{2.8}$$

It measures the proportion of correct predictions (true positives and true negatives) in the population.

Accuracy alone is prone to misjudge prediction performance for imbalanced data, which can be a relevant problem in medical applications, where it is often difficult to achieve well-balanced training and test data. If there are far more instances of one class, the uninformative classifier that always predict the class with higher prevalence can have a higher accuracy than a more useful one that also predicts the other class, at the cost of more false classifications (Zhu et al. 2007). Hence, the ISO 5725 norm has been issued to redefine accuracy to be the combination of two values: trueness and precision. The latter is identical to the positive predictive value, PPV, while trueness equivalent to the bias (compare in Fig. 2.3 on page 50). Trueness is a term introduced in the ISO norm to avoid misinterpretations of the term bias outside the technical domain.

In publications involving machine learning a somewhat more rarely used criterion is the Brier score that measures the accuracy of probabilistic predictions of exclusive outcomes. The Brier score measures the distance between prediction and true value, both being between 0 and 1. Hence, low values are indicative for a good predictor.

Definition 2.5 — Brier score. The Brier score is defined as the averaged sum of squared differences between forecast probability f_i and true probability o_i of a binary event, N times forecast:

$$\beta = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2 \tag{2.9}$$

The Matthews Correlation Coefficient (MCC) is one alternative to the above criteria. The MCC characterizes a classification model in terms of true and false positive and negative predictions (i.e. the confusion matrix) and is known to be a good criterion of classifier performance when certain preconditions are asserted. The MCC is derived from the Pearson's Correlation Coefficient and like it takes on values between -1 (worst) and 1 (best performance) (Baldi et al. 2000):

Definition 2.6 — Matthews Correlation Coefficient (MCC). The Matthews Correlation Coefficient (MCC) defined in terms of the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classifications is

$$\mathrm{MCC} = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}}.$$

Constellations in which the correlation coefficient will be relatively high can be constructed, but they require predictions with equally very few FP and very few TP cases, which is not a very helpful classifier anyways.

Receiver Operating Characteristics

We move on to receiver operating curves (ROC) and their analysis, which has become the probably most common tool to express observer performance and compare observers in the medical field (Metz 1978). Many good introductions and publications cover this subject well, and the method is widely accepted as a robust and comprehensive description of a predictor's performance. Hence we only emphasize some central aspects without going into detail.

For a ROC analysis, the predictor scores are thresholded over their full range, at each point calculating the numbers of true and false positive and negative decisions based on the current threshold. The corresponding sensitivities and specificities are plotted against each other, yielding the ROC curve. Note that in the case of a predictor that only outputs binary decisions, the ROC curve has only one operating point.

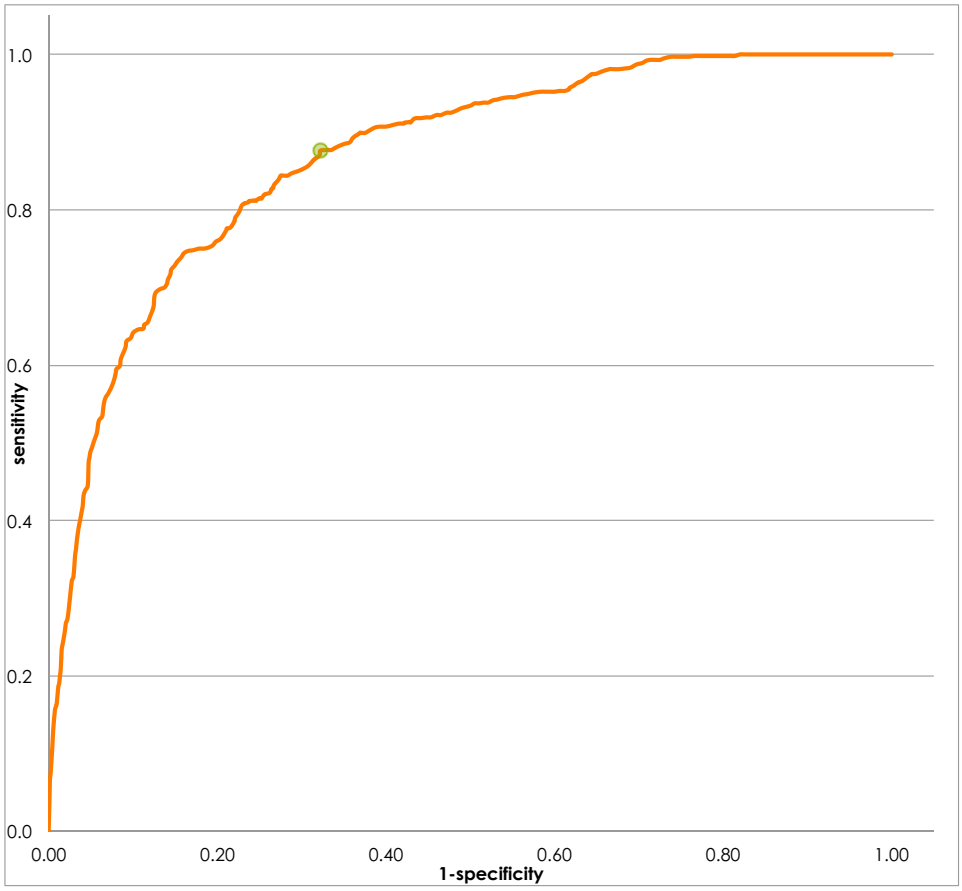


Figure 2.6: Example of a ROC curve. The ROC curve of prediction performance on the 4D textures data set (see Sec. 2.4). The AUC is 0.88, with one possible operating point at a sensitivity of 0.88 and a specificity of 0.68 for the prediction of malignant cases (marked with a green circle). The two axes are sometimes also labeled "true positive rate" (sensitivity) and "false positive rate" (1-specificity).

The most prominent scalar score derived from the ROC curve is the area under it (area under the ROC curve, AUC), ranging from 0.5 to 1, with one being the perfect predictor and 0.5 being chance predictions.[8] Calculating the AUC can safely be done with the simplest of all methods, the trapezoidal summation of partial areas from point to point (Lasko et al. 2005; Zweig et al. 1993). The nonparametric AUC is known to be downwards biased on small sample sizes, but on the other hand it is widely applicable since no assumptions on the layout of the data are made. An example of a ROC curve is shown in Fig. 2.6. Parametric estimates of the AUC assume certain data characteristics which are not always easily estimated or known. Since they do not offer substantial benefit over the nonparametric estimation of the area under the ROC curve, they are not further discussed.

Another (quite well hidden) pitfall in AUC estimation can in some cases be the influence of the way the scores are generated. n-fold cross validation or leave-one-out cross validation are often employed to produce robust estimates of AUC by averaging over folds. They are biased depending on the data balance, which can be circumvented by balancing the data. Airola et al. (2010) show in their works that n-fold cross validation and leave-one-out cross validation in heavily unbalanced data sets with only few examples and features can underestimate the AUC by as much as 0.1.

In cases where the data is heavily unbalanced, and the data set is very small, the computationally demanding but nearly unbiased leave-pair-out (LPO) cross validation can be employed, where in each fold one of every possible pair of a positive and a negative example are held out

[8] If an AUC of $a < 0.5$ is achieved, all decisions have to be negated, and the new AUC is $a' = 1 - a$.

from the data set, and they are tested against each other example in the data set, recording if the classifier gives the higher score to the "positive" class in each pair-wise test. An AUC estimate can in this case calculated according to the Wilcoxon-Mann-Whitney statistic,

$$\hat{A}(X, f) = \frac{1}{|X_+||X_-|} \sum_{x_i \in X_+} \sum_{x_j \in X_-} \mathcal{H}(f(x_i) - f(x_j)), \tag{2.10}$$

where $\mathcal{H}$ is the Heaviside function, X_+ and X_- are the positive and negative held out examples from the data set X, and f is a real-valued function learned from the training data. f is the predictor; it takes an example and returns a prediction between 0 and 1. Lastly, $\mathcal{H}$ is defined as

$$\mathcal{H}(a) = \begin{cases} 1 & \text{if } a > 0 \\ 0.5 & \text{if } a = 0 \\ 0 & \text{if } a < 0 \end{cases} \tag{2.11}$$

In all other cases, balanced n-fold cross validation is a good estimator of classification performance and can safely be used. In application scenarios it might also be a downside of LPO cross validation that no ROC curve is obtained, but only an estimate of the AUC, hence, no operating points may be selected, and the practically important balance between sensitivity and specificity of the predictor is not revealed.

Optimal operating points (combinations of specificity and sensitivity values on the curve) can be selected on ROC curves, taking into account the class proportions and expected cost of false positive and false negative decisions. From the ratio of the costs and the prevalences a slope can be calculated, and the optimal point on the ROC curve is the one where the curve has this slope (Zweig et al. 1993). The same authors also advise on the calculation of confidence intervals and required sample sizes to rate predictors against each other on the basis of their ROC curves, but both topics are beyond the required understanding for this thesis.

As a last remark pertaining to the field of medical decision making, it should be noted that the full area under the ROC curve may be misleading for the judgment of a medical test. If there is an upper limited of accepted false negative decisions (missed cancers, for example), one is probably more interested in the performance of the predictor in the interval below the limit. In this case, a partial AUC is often calculated, particularly when comparing prediction methods. Mind though, that since now an arbitrary threshold on accepted sensitivities or specificities is set, the partial AUCs are no longer easily comparable across studies.

Validating the Predictor Performance

Generally, a large number of training samples and an equally large number of test samples is the optimal precondition to train a predictor with good generalization ability. Certainly, a large number alone is not sufficient, but the training set needs to represent all target classes, and the examples need to cover the feature space of their respective classes. Also, a balanced training set (or a priori knowledge of the true proportions and their costs) is required to avoid unwanted bias. Under these conditions, optimal predictors may be obtained with inductive learning methods. Statements about a predictor's average error will on the other hand be more accurate with increasing test set sizes. Training and testing examples need to come from the same population.

Hence, the optimal condition for predictor training is a large balanced training and test example basis from one population. It is mandatory to train the predictor on the training set alone, and not optimize it towards the test set.

If the available data doesn't allow for a split, which is a very likely case in many practical problems in machine learning for medical data, alternative methods need to be employed.

Repeated random subsampling validation iteratively splits the available data into training (e.g. 70%) and test set (30%). This method will deliver different results on the individual runs, but with a large enough number of runs, the performance estimates will vary less. The advantage is that the number of runs is independent of the split, such that a low-variance error estimate can be achieved with more runs.

This is different in alternative validation strategies that split the data, namely k-fold cross validation, where a typical choice is $k = 10$. The full set is split into k equal parts, and each in turn is hold back from training and later used as a test set. The extreme version is $k = N$ with N the number of available examples in the full set. This setup is called leave-one-out cross validation (CV). Particularly in the medical context, it is sometimes varied into a leave-one-patient-out cross validation to make sure that no error underestimation occurs because data from one patient is included in both training and test set. Leave-one-(patient)-out CV performance predictions are known to be very little biased, but affected by high variance. A suitable compromise has been proposed with the averaging of 30 repeated 10-fold cross validation runs (Ounpraseuth et al. 2012).

Varma et al. (2006) compare 10-fold CV and leave-one-out CV, concluding that both have a positive bias, overestimating the true error. Their main contribution, however, concerns cases where feature selection precedes predictor training in a cross validation setup. The feature selection process must not use the full data set before the split; instead, the feature selection process (particularly if it includes cross validation) needs to be carried out on each training set selected from the total. The same applies to any parameter tuning of the predictors.

Facing small original example data sets, ways have been explored to generate artificial data from the available data. We do not follow this path for long, but point the interested reader to Q. Li (2007) who used kernel estimation methods to estimate the probability density function of the original data and generate samples from this distribution. Their analysis shows that the generalization performance of a predictor increases while the variance of the generalization performance is reduced when trained on virtual data compared to the originally small data set. Small in his experiments is 50 examples; at twice the number, the gain is much smaller. Another cross validation method proposed by Fu et al. (2005) is Bootstrap CV, where the examples of the training set are drawn from the population with replacement.[9]

Ounpraseuth et al. (2012) compare bootstrap CV and k-fold CV and conclude that the strong negative bias of bootstrap CV is "too high a price to pay for its reduced variance". In the same direction, Varma et al. (2006) characterize these approaches as low variance methods at the cost of with sometimes high bias.

Optimizing Predictor Performance

We have discussed several performance criteria in the preceding sections. It has become clear that the selection of the criterion has to be taken carefully, and we have seen how the criteria may depend on the data and application. In a more abstract consideration, we briefly turn to the common practice to optimize results achieved in machine learning tasks by using several feature selection methods each followed by a choice of machine learning methods, potentially in several cross validation setups. The crucial point here is not this process as such, but the way to report on it. Whenever a paper only reports the results of only one methodology, without comparing it to previous or own results with other methods on the same data, the new results have to be taken with some caution. They may have been obtained by optimizing the performance through predictor selection.

[9] Further reading on these and related methods is for example provided by Efron et al. (1983) and Kohavi (1995).

Permutation of class labels can be used to model the null hypothesis that none of the features is correlated with the true class. This approach has been proposed by Boulesteix et al. (2008) and can be employed to obtain a non-informative data set preserving the realistic structure. A data set such constructed may serve to quantify the potential bias induced by the process of optimal classifier selection. Boulesteix et al. (ibidem) have shown that the median minimal error rate can be as low as 31% based on the permuted non-informative data if optimized classifiers are used. Note again that we are talking about the error rates on random data. As a remedy to the situation, the authors advise to report on all evaluated classifiers to provide a more unbiased view on the expected classification accuracy of the classifier.

The Meaningful Predictor Evaluation

We have learned about the intrinsic statistical bias and variance of predictors, introduced by the modeling choices. Bias of another kind can involuntarily be introduced during the model *selection* process, this time affecting the results regardless of the actual chosen method. The usual research procedure is to train a classifier, observe its performance, alter its parameters and train again, until a satisfactory performance level is obtained.

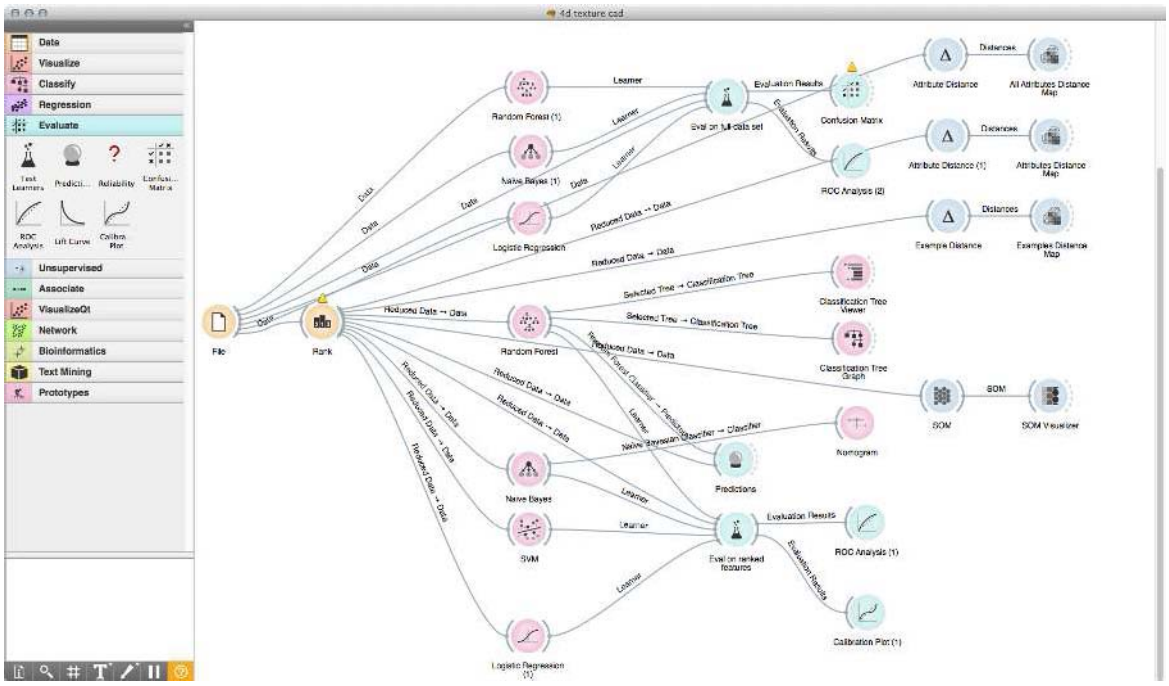
Eventually, other classifiers are also tested on the same data, which is being made very easy by the many freely available machine learning tools like Weka (Witten et al. 2005) or Orange (Demšar, Zupan, et al. 2004) that readily accept all types of data and let users explore all types of classification methods at the press of a button, not to mention the countless software packages implementing machine learning algorithms that can be loaded into numerical computation software packages like Matlab and Octave, S and R, Scilab, or IDL, to name only the most prominent ones. Classifiers can be evaluated in n-fold cross validation schemes, leave-one-out cross validation, or n-fold data sampling of a training set and test set.

This ease of operation has in the recent years led to a landslide of research papers that apply machine learning to problems in biomedical sciences. Many of the modern problems in machine learning in biomedical science have to deal with huge amounts of features, most prominently DNA microarray analysis, where easily tens of thousands of features are evaluated for a limited number n of samples, leading to the $p \gg n$ problem — the curse of dimensionality. But also medical image analysis can be affected by this, with one example being the automated analysis of high-resolution ex vivo MR spectra of tissue samples in attempts to replace histopathology by rapid classification of spectra, and we have explored several methods to reduce the feature space dimensionality without imposing modeling constraints that hamper the detection of structures in the data (M. Harz, Diehl, et al. 2009; Wenzel, Merkel, Althaus, Nölte, et al. 2006; Wenzel, Merkel, Althaus, and Peitgen 2006). Such high-resolution MR spectra often have tens of thousands of highly correlated features, while typically only a few dozen examples are available.

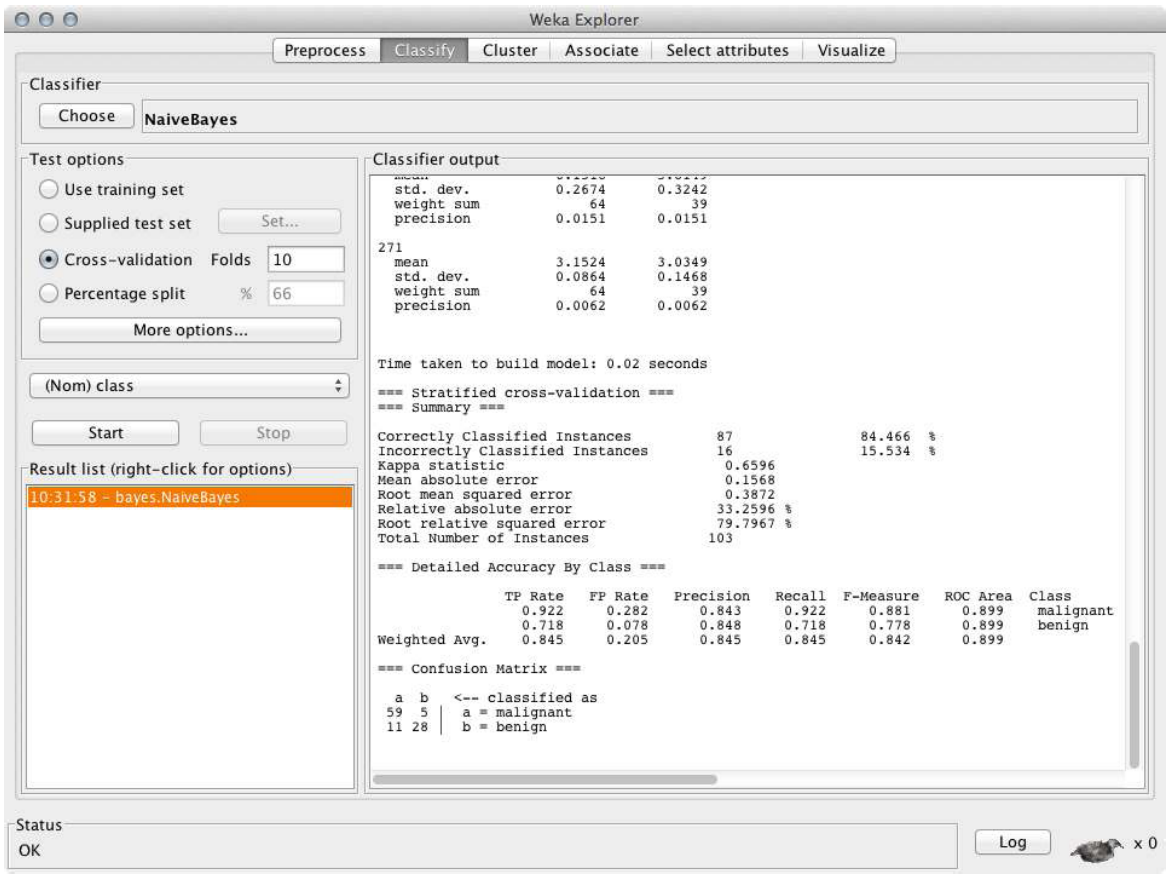
Particularly the errors in earlier studies in DNA microarray analysis provide lessons to learn for machine learning in high-dimensional feature spaces. Meta-research has looked at the quality of such research (Michiels et al. 2005). In a comment on this review, Ioannidis (2005) sarcastically remarks that the majority of studies have a predictive value comparable to flipping a coin, and the remainder of studies "barely beat horoscopes". He remarks that "without highly specified a-priori hypotheses, there are hundreds of ways to analyse the dullest dataset", ensuring that there will always be a positive result to report.[10]

Boulesteix et al. (2008) have looked at this problem in great detail, revealing that the way how models are selected and how results are reported will bias the study and determine

[10]This situation resonates in the emergence of journals like the *Journal of Negative Results in BioMedicine* that explicitly encourages researchers not to tune data and analysis until positive results are obtained, but instead to report failure and unsuccessful methods as well. The journal, by the way, has had only very few contributions up to today.



(a)



(b)

Figure 2.7: The graphical user interfaces of two prominent graphical machine learning and knowledge discovery tools that are freely available, (a) Orange (DEMŠAR, CURK, et al. 2013), and (b) Weka (WITTEN et al. 2005). Data loading, feature selection, classification, clustering, and many more tools are available in both, and both offer graphical tools to plug machine learning algorithms into a pipeline. Both are also extensible with own algorithms, where Orange provides a python interface, Weka builds upon Java.

the reported outcome. VARMA et al. (2006) has in addition looked at the influence of feature selection cross validation in connection with classifier training, revealing common methodological pitfalls that introduce another bias. If the feature selection algorithm is not independently optimized from the classifier cross validation scheme, a strong bias will be introduced. Both topics address the problem of valid model evaluation procedures.

2.1.3 Recommended Approaches

Naïve Bayes

The so-called "Naïve Bayes" may have a name that sounds slightly derogatory, and that it has been dubbed "Idiot's Bayes" by some might cause prejudice that the very well-established methodology certainly doesn't deserve. It is true, a Naïve Bayes classifier is very simple to implement and apply, but still performs very well when used on the correct data.[11] The attribute "naive" stems from the assumption that is made to simplify the equations. The Naïve Bayes assumption is that between the features there is no conditional dependence, which can be wrong even for very simple examples. We will discuss the impact of this assumption later.

As long as most features are conditionally independent, the Naïve Bayes classifier is not as affected by the curse of dimensionality as the classifiers discussed above since the learning mechanism inherently prefers models of medium complexity over others with too low or too high complexity (BISHOP 1995, page 386). A very convenient feature is that no assumptions to feature distributions need to be made, although the implementation with the assumption of Gaussian distributions represented by mean and variance makes the calculation of the model particularly easy.

We start with the well-known Bayesian formula expressing the posterior probability of an event given some observations, or, in our case, that a test example is of class C given features $F_1, \ldots, F_n$. This posterior probability $p(C|F_1, \ldots, F_n)$ equals the classes prior probability times the likelihood to observe the features for the class, divided by the evidence for the feature observation, or

$$\text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Evidence}}$$

.

Mathematically, this is

$$p(C|F_1, \ldots, F_n) = \frac{p(C)p(F_1, \ldots, F_n|C)}{p(F_1, \ldots, F_n)}. \tag{2.12}$$

The trick to arrive at of the formulation of the Naïve Bayes classification approach is to employ the chain rule. The chain rule states that

$$p(F_1, F_2|C) = p(F_1|C) \cdot p(F_2|C, F_1),$$

allowing to rewrite the "Likelihood" term into a product of $p(F_i|C, F_1, \ldots, F_{i-1})$. Finally, with the conditional independence assumption, we can write $p(F_i|C, F_j) = p(F_i|C)$, as long as $i \neq j$. It is here where the independence assumption manifests mathematically.

It is a beneficial property of the Naïve Bayes approach that meaningless features don't hamper the training and classification potential; they are automatically discarded in the calculation of the "Evidence" through the law of total probability. This is also referred to as *marginalization*, and is an inherent feature of the Naïve Bayes method. In addition, the "Evidence" term in the

[11] SCHÖLKOPF and SMOLA (2001) shows how the Bayes classifier is a special form of a very simple linear SVM.

denominator will be constant, such that up to a factor, the posterior probability $p(C|F_1, \ldots, F_n)$ can be simplified to

$$p(C|F_1, \ldots, F_n) \propto p(C) \cdot \prod_{i-1} p(F_i|C). \tag{2.13}$$

The classification of a new instance based on its feature vector is then done by calculating the above for all C, and selecting the C that maximizes the posterior probability.

Training a Naïve Bayes classifier is efficient. Firstly, the prior probability of each class $p(C)$ is calculated from the examples. The examples have then to be split into the classes, and the means and variances of all features need to be calculated per class, if a normal distribution of each feature is assumed. Note that this is not strictly necessary, since any parametric and even non-parametric distributions may be plugged into the framework.

There are, of course, limitations and sources of potential error in the application of the Naïve Bayes classifier, of which the most subtle ones require to be mentioned. Firstly, on skewed data (with much more training examples in one class C_i than any other) the most prominent class has a higher prior probability $p(C_i)$, effectively biasing all estimations towards this class.

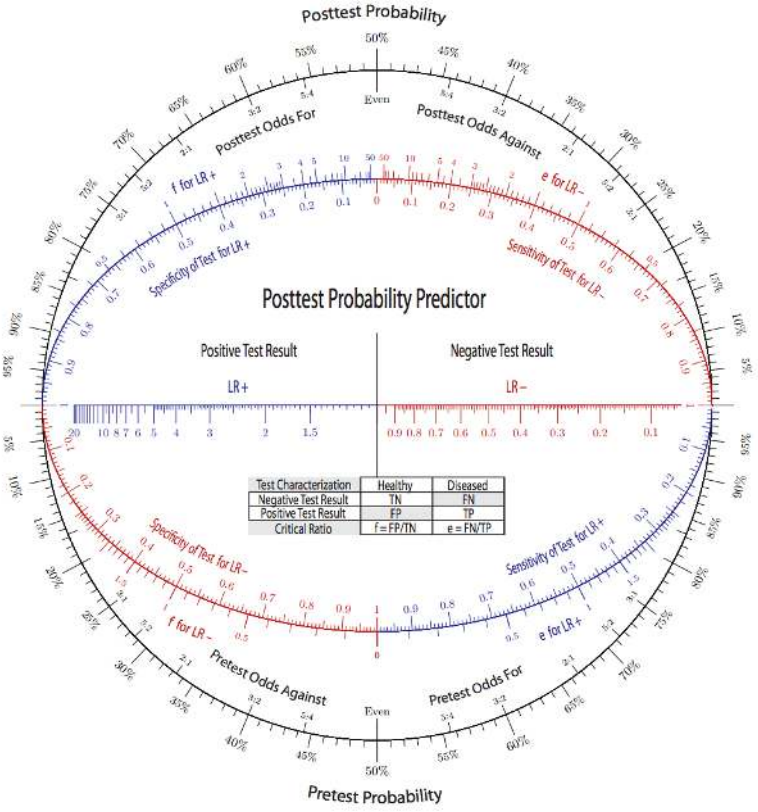


Figure 2.8: A nomogram implementing Bayes' theorem. It can for example be used to estimate the post-test probability of a disease from the pre-test probability and the test's likelihood ratio, which can help to decide if the test should be taken. By example: Assume a test with a positive likelihood ratio of 2.8 and a pre-test probability of 30%. By drawing a line from the bottom black scale Pretest Probability through the blue horizontal scale in the middle at 2.8 towards the top of the circle reveals a post-test probability of about 55% — the test should be taken. After: MARASCO et al. (2011).

Also, if one class C_i violates the feature independence assumption more than, say, C_j, it will be in favor. Suppose there is one feature F_k counting in favor of C_i, and there are for example two conditionally dependent features F_m and F_n contributing in the product as $p(F_m|C_j)$ and $p(F_n|C_j)$. Supposing that the features F_k and F_m contribute equally towards C_i and C_j,

respectively, then for one count towards C_i, there will be a double counting for C_j in Eqn. (2.13) (Rennie et al. 2003). A toy problem illustrating this is the classification of text snippets by counting word mentions. Thus, when classifying texts into ones talking about "Orlando" and ones about "San Diego", each appearance of "Orlando", "San", and "Diego" would be counted independently. Assuming Orlando is mentioned five times and San Diego three times in one text snippet, then the evidence for "Orlando" would be five, while for "San Diego" it would be three plus three — clearly a bias.

A more commonly known pitfall not specific to Naïve Bayes alone, but affecting many classifiers regards the topology of the classification problem, which under certain circumstances is problematic for Naïve Bayes: Any XOR-like decision problem cannot be modeled by Naïve Bayes if the features are all binary, i.e. if features can only take on two discrete values (Rish 2001). This is no relevant problem in the classification problems we will look at, where features are continuous or multi-valued.

We conclude our account of the Naïve Bayes classifier pointing the reader to the forgotten, but beautiful art of nomograms (compare Fig. 2.8). Nomograms are in their basic form a way to graphically calculate linear and simple nonlinear relationships between few variables. They can, however, be employed to visualize and interactively access a Naïve Bayes' prediction rules. The nomogram in this case visualizes the importance and influence of all features. In interactive implementations, feature values can be altered and their influence towards the prediction be understood.(Doerfler 2009; Marasco et al. 2011). Nomograms are for example implemented as a widget in the Orange framework (Demšar, Curk, et al. 2013), where they can be attached to a trained Naïve Bayes classifier. Nomograms are today employed in clinical decision making, e.g. to compare and predict treatment outcomes given associated risk factors (Albert et al. 2012), or the likelihood of recurrence of DCIS given 10 personal and disease-related factors (Yi et al. 2012). Since the parameter space is high, those nomograms are usually implemented in the form of online calculators.

Random Forests

Originally proposed by Breiman (2001), Random Forests quickly evolved to become a frequently used machine learning method in many scientific areas. Its principle is to construct a large number of decision trees (an *ensemble*) on partial data and only a subset of features, and compose ("bag") their results into the decision. The following description assumes basic knowledge on the construction of decision trees which is readily available for example in Wikipedia (2013a).

The algorithm proposed by Breiman can be summarized in only a few steps. Assume N to be the number of training examples and K the number of features characterizing an example. Set n to the number of examples to train one tree on, and k to the number of features used in each node of each tree. Breiman proposes $k = \log K$ as a choice for k. Construct t trees, each following the steps below:

1. Select with replacement n training examples for this tree. (Remainder serves as a test set for this tree.)
2. Grow unpruned tree. Per tree node, do:
   (a) Randomly select k features
   (b) Determine best split.

Note that for each tree, only a subset (usually about one third) of the complete data set are used to train the tree. When the tree is grown, the remaining examples can be run through the tree, and through the random forest built so far, and provide an estimate of the prediction error. These errors are known as the *out-of-bag* errors. Breiman even states that cross validation of

random forests is not required to estimate the total prediction error, since the procedure to build it yields unbiased error estimates on the way (Breiman and Cutler 2013).

Generally, random forests are very well-behaved regarding all pitfalls discussed above. Large feature numbers can effectively and efficiently be dealt with by a large enough training database, and the requirements regarding the proportions are far more favorable than for example SVM training. Even on small data bases with many features, the prediction error doesn't grow overly (it does, though), due to the Random Forest's inherent mechanism to neglect unimportant features automatically. Random forests are known to have an inherently low bias, and the variance can be reduced by growing larger numbers of random trees. Of particular interest for research purposes, where frequent re-training on newly available data is the routine, the high speed in training and prediction is of practical value. Also, the number of training examples has only little influence on the speed. Summarizing all said, we recommend Random Forests for general purpose tasks, particularly if aspects of the data and the distributions of the features is unknown.

2.2 DCIS from a Different Perspective

After the introductory description of the DCIS that focused on the microbiological level and the development pathways to and from DCIS, in this section the relevant groundwork will be provided to motivate a novel direction of image analysis. A comprehensive understanding of the diversity of DCIS's potential morphologic and kinetic appearance in radiological images is the required basis to derive the approach presented in Sec. 2.3.

2.2.1 Imaging Biomarkers and Diagnostic Criteria

The typical imaging appearance of DCIS has been studied for a long time already, starting with the seminal work of Holland et al. (1985). Dissecting mastectomy specimen and correlating them with the mammographic imaging appearance led to the conclusion that DCIS is frequently underestimated in size by imaging. This fact holds true until today and is the limitation to all imaging-based DCIS detection methods that can be devised. In particular MRI has only with recent developments begun to depict the fine detail of DCIS. Milk ducts are only in the peri-areolar area wider than 1–2 mm, and hence DCIS in its early stage, when it develops close to the terminal duct lobular units (TDLUs), will be seen only faintly if at all. Ducts average at 1 mm in Diameter, whereas TDLUs are of a diameter of far less than 0.5 mm upwards.

This, however, is the spatial in-plane voxel size that MRI sequences are beginning to resolve in clinical routine. This makes it more and more possible that early DCIS will be revealed with high-resolution techniques, also because with the growing knowledge of the microstructure and microbiological behavior of DCIS, imaging may in the future be tailored to the purpose, either by assessing the images in the vicinity (the stroma) of suspicious lesions more closely (Nabavizadeh et al. 2011) or by designing specialized MR protocols that quantify for example the vascularization (Hyodo et al. 2009).

The American College of Radiology breast MRI image reporting lexicon (ACR BI-RADS® atlas, E. A. Morris et al. (2013)) requires radiologists to describe the imaging appearance of non-mass enhancing abnormalities regarding many aspects. The reporting according to BI-RADS has become the standard. Among the modifiers applicable to the distribution of non-mass enhancements are focal, segmental, and regional, to cite those associated with the extent of the disease. A focal distribution is by this distinction characterized by a low involvement of breast tissue in only one quadrant. Segmental distributions should encompass enhancements in more than one duct, and regional enhancement is any that isn't described by the above (compare Macura et al. (2006)). There are many further modifiers (clumped, ductal, linear, etc.) that are to be reported as well, which potentially makes the task of verbally describing non-mass enhancments subjective and ambiguous.

! All DCIS imaging using contrast-enhanced MRI relies on the neovascularization of the tissue with quickly growing vessels that have defective vessel walls allowing blood to inundate the interstitial fluid. Contrast-enhanced MRI will pick up the signal from the contrast agent in this compartment. Other MRI-based imaging mechanisms exist, but are far less sensitive (DWI, ASL). Apart from this, vascularity can be displayed in Doppler ultrasound or with specialized ultrasound contrast agents, and tumor growth also can be imaged with contrast agents showing the glucose metabolism of cell growth. These techniques are, however, not in the focus of this part of the thesis. For a more in-depth treatment of the general imaging principle of breast MRI, see Sec. 1.2.3. For examples of DCIS as seen in DCE-MRI, compare Fig. 2.9 on page 67. In this figure, a standard coloring method has been applied to the data that encodes the wash-out behavior of the curve in any voxel by colors, with red corresponding to a malignant curve type exhibiting strong wash-out, blue encoding plateau curves, and green encoding persistent wash-in. In addition, voxels with a stronger relative enhancement in the first two minutes are less transparent. Usually, a threshold is applied to the initial relative enhancement, because it is believed that only abnormalities that enhance beyond for example 200% are to be further

considered. This color/opacity encoding has become a standard in many existing breast MRI workstations, though it arguably neglects the specific characteristics of the human visual system for example by using the color blue, and by using continuous transparency values, which are hard to differentiate.

It has to be noted that the research on DCIS is fundamentally complicated by the difficulties to define the disease appropriately. There are many more non-mass-like enhancement patterns to judge than only pure DCIS, and even pure DCIS is not one single disease, but is classified according to several factors[12] into low, medium and high grade, and are pathologically described based on their growth pattern for example as papillary (finger-like patterns in the ducts), cribriform (gaps in the cancer cell tissue), comedo (-necrotic; dead tumor cells), or solid. The heterogeneity of the phenotype of DCIS also leads to ambiguities in distinguishing it from atypical hyperplasia of the ductal epithelium, and even DCIS that has already broken through the basement membrane of the ducts is sometimes called DCIS "with microinvasion" if the invasive component is smaller than 1 mm (Pinder 2010).

In a pictorial report on typical DCIS appearance, Facius et al. (2007) only found 74 patients with pure DCIS lesions among 3583 routine MR examinations performed within 6 years, which amounts to only 2%. Other studies report that 20% of a total of more than 600 malignant lesions seen are DCIS, accounting for every third out of four seen non-mass lesions (both benign and malignant) (Jansen, Shimauchi, et al. 2011).

Holland et al. (1985) described DCIS appearance based on stained, thin-sliced mastectomy specimen. DCIS is considered a unicentric disease (Pinder 2010), but multiple centers evolving independently seem also to be possible: While Holland et al. reported only one multicentric lesion in over 100 cases of DCIS, later studies found five DCIS out of 14 to be multicentric, and additionally two DCIS that were considered multifocal at pathology (F Sardanelli et al. 2008). The same study cites further results confirming in particular the higher yield of multicentric DCIS. Other authors describe DCIS as multifocal, but also state that foci are rarely more than 10 mm apart and most often connected, so that they are pathologically counted as the same disease process (Mossa-Basha et al. 2010; Thomson et al. 2001). This gave rise to the hypothesis that DCIS grows within only one lactiferous system, in particular not extending from one segment of the breast into another. Ths is in opposition to the frequently proposed partitioning of the breast into four quadrants, reflected also in the surgical procedure called quadrantectomy. Instead, the areas belonging to one milk duct extending into the nipple might be a basis to reason about multicentricity and multifocality. The supply areas of one milk duct are unlikely to conform with the quartering of the breast. Rather, the 10–14 milk ducts can have quite arbitrary sizes, shapes, and locations. If DCIS truely grows along the ductal system, it is unlikely that it grows from the supply area of one duct into that of a different one, because this would require its growth into the nipple and from there back into the second duct, or a breach of the duct's basement membrane — invasion. DCIS, however, is not invasive by definition. Multicentricity and multifocality are commonly also defined with respect to quadrants: multifocal disease shows cancer foci with more than 40 mm distance in one quadrant, multicentric disease in addition crosses quadrant borders (Hayat 2008). This definition might appear poorly motivated, taking into account the morphological considerations made before. Similarly, DCIS should not be expected to be a segmental (involving more than one duct) or regional disease.

For the purpose of the following DCE-MRI based DCIS detection and diagnosis method, a presentation of the agreed typical appearance of DCIS in MRI will be given. Generally, it is obvious from the studies of DCIS that for a comprehensive and specific characterization, morphology plays an equally important role as contrast kinetics.

[12]Most commonly assessed are nuclear grade, mitotic rate, Ki-67, p53, estrogen and progesterone receptor expression, and HER2 status (Adler et al. 2012; Hieken et al. 2001).

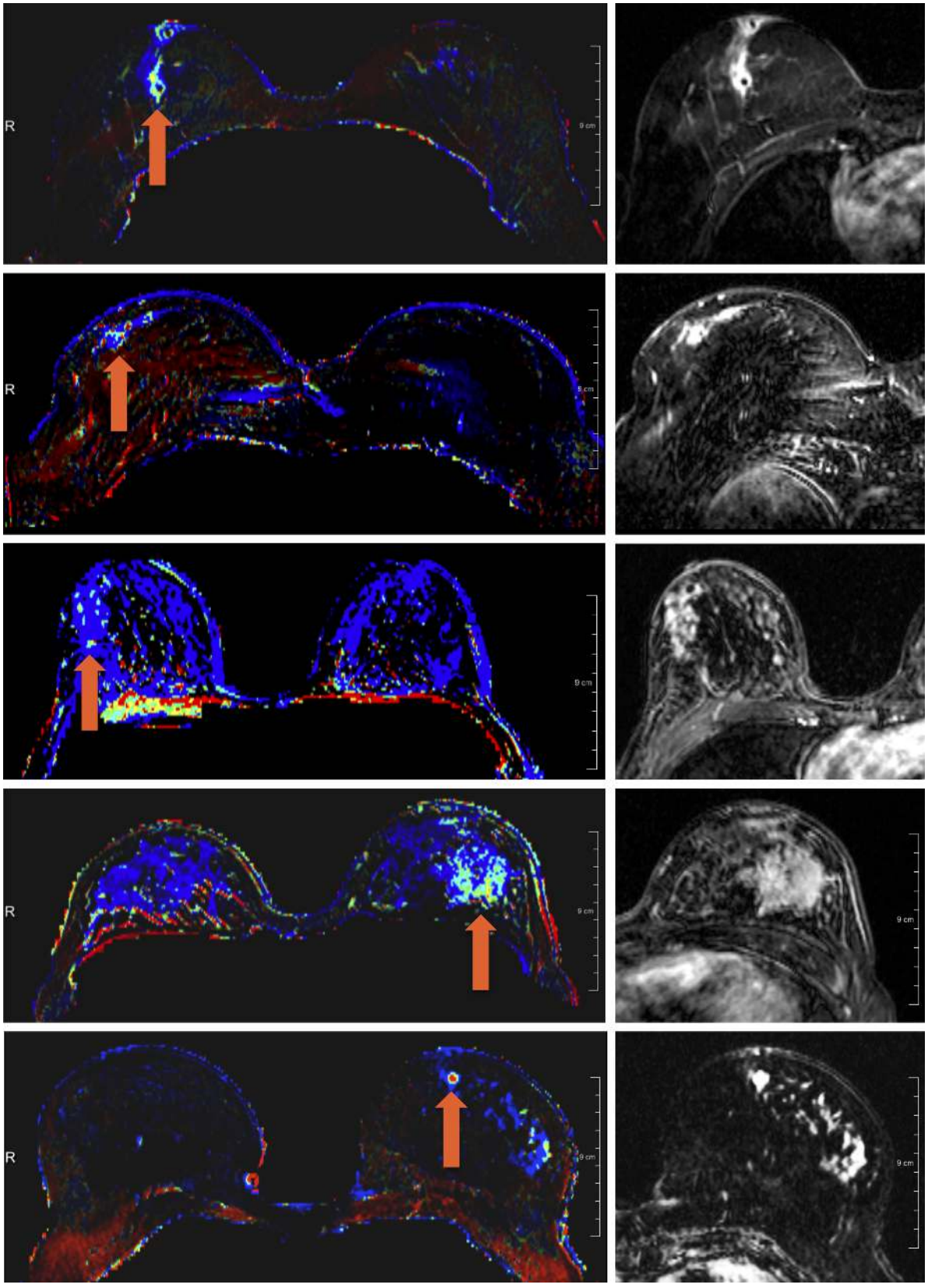


Figure 2.9: Differently looking manifestations of biopsy proven DCIS cases in contrast enhanced breast MRI, and the associated coloring using an encoding of the strength of wash-in by opacity, and the strength of wash-out by colors. Note that no regions show marked wash-out behavior. It can be seen that DCIS varies in shape. From linear structures extending towards the nipple, over segmental diffuse enhancing patterns, to focal and mixed appearances, all variations are possible. (MR images courtesy Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands.)

Over the recent years, the lack of reporting standards has lead to the development of the BI-RADS[13] criteria that describe imaging findings in terms of their appearance and kinetic properties. A malignant (BI-RADS class 5) DCIS will display early asymmetric (unilateral) enhancement in a segmental distribution, while suspicious lesions will show a less pronounced early enhancement and a morphology that is not clearly confined to the ducts (C. K. Kuhl et al. 2007; Raza et al. 2008).

While DCIS is generally being described as "non-mass like enhancement", this is only true for the majority of pure DCIS cases, but notable exceptions exist even in this group. Overall, 10%-20% may present as masses, while the remainder is described as clumped, ductal, segmental enhancements following the ducts. Segmental enhancement in this disease is defined to be directed towards the nipple, where the shape of the enhancing region may resemble a cone, consistent with growth along the ductal system (Raza et al. 2008). Interestingly, nuclear grade and mammographic appearance do not vary significantly with the MR morphology of pure DCIS (Jansen, Newstead, et al. 2007). The nuclear grade characterizes the variation in cancer cell nuclei and cell growth rate found in the tissue specimen, and is graded from low (very similarly shaped nuclei, and a low growth rate), to medium, to high grade (very dissimilar nuclei and a high growth rate), compare e.g. Leonard et al. (2004).

Detecting DCIS in breast MRI is usually done with reference to a number of parameters, divided into morphology and kinetics of enhancement patterns, with the assumption of a sufficient spatial resolution of the images (F Sardanelli et al. 2008). For descriptive kinetic parameters[14], the most salient choices are

- the initial enhancement percentage, i.e. the relative enhancement reached after a certain time, e.g. 2 min
- the time to the enhancement peak (TTP), which discriminates areas with washout (short TTP, e.g. less than 2 min) from those with continuous enhancement (long TTP),
- and the signal enhancement ratio, characterizing highly suspicious areas by their high relative enhancement, e.g. over 250%.

Clinically seen DCIS lesions most often display moderate wash-in characteristics, followed by an enhancement plateau curve type, rather than the marked wash-out characteristics like it is observed in invasive cancers (Newstead 2010). A reason for this is that the neo-vascularization of DCIS is more mature than in invasive lesions, hence less pronounced wash-out characteristics will be seen (Boetes et al. 2007). Among the pure DCIS lesions, however, other studies have found a mixture of all three wash-out categories: wash-out, plateau, and persistent curves. In these lesions, different from the morphological point of view, a variation of kinetics with mammographic appearance was observed, but again none with nuclear grade (Jansen, Newstead, et al. 2007).

Concluding, kinetic characteristics alone have to be considered insufficient and indecisive for an accurate analysis of DCIS.

Far more promising are the morphological features, but an important "[clinical] limitation of morphological descriptors is that they are qualitative and subjective" (Jansen 2011; Newell et al. 2009). In another line of investigation, Esserman et al. (2006) have attempted to established correlations of imaging biomarkers for DCIS derived from MRI with immunohistochemistry (IHC) variables and appearance on pathology. Their findings show how limited the depiction of imaging biomarkers that differentiate types of DCIS biology using MRI still is today, obtaining statistically significant ($P < .05$) correlation mostly for tumor size and density, but show

[13] Breast Imaging Reporting and Data System

[14] Descriptive parameters capture the shape of the enhancement curve. The other approach is a model based estimation of pharmacokinetic parameters, like e.g. using the Tofts or Briggs model.

very limited correlations only between MRI and four evaluated IHC variables[15]. Only the size, distribution, and density as seen on MRI correlated significantly with CD68 staining; in particular in certain high-grade DCIS this appeared to separate one subclass of DCIS. Besides, the authors found that other ER-positive DCIS were prone to two-fold size overestimation on MRI. The study, however, only looked at features from the BI-RADS catalog applying to DCIS.

Therefore, a better (automated) description of lesion morphology and a robust estimation of size and distribution may contribute to a more reliable diagnostic description of DCIS.

Jansen (2011) and others hence recommend to combine morphology and kinetics, and also to combine them in an intelligent fashion that weighs kinetic features based on lesion morphology. Also, from the results of mouse models, the hope arises that in the future image-based features can be derived that not only distinguish DCIS with invasive foci from those without, but also DCIS that will remain indolent from DCIS that will develop into invasive cancers. This, however, may require to take factors into account that are not captured with imaging alone, e.g. genetics, or risk factors (Jansen 2011; Jansen, Conzen, Fan, E. Markiewicz, et al. 2011).

Looking at the DCIS detection performance of other imaging methods shows that DCIS can be detected comparatively well in MRI (Boetes et al. 2007; C. K. Kuhl et al. 2007). It has been objected that one distinguished feature of DCIS, the formation of microcalcifications, cannot be detected in MRI, while it is an indicator seen in x-ray mammography, particularly on digital mammography. Calcifications in DCIS occur when a lumen filled with a sufficiently large quantity of liquid containing a sufficient concentration of calcium is closed by growing cancer cells. Over a certain amount of time, calcifications will form. Since in DCIS the lumen is the duct, and the growth is along the duct, microcalcifications will line up along the duct, eventually displaying elongated configurations with branching patterns on mammography. This prototypical imaging phenotype of DCIS calcification is extremely indicative for DCIS. On the other hand, calcifications are in general a very indeterminate finding; the definitely benign and definitely malignant appearances only account for a small fraction of all seen microcalcifications. Also, many of the studies that quote detection figures of mammography versus other imaging techniques are often biased since the definite diagnosis by pathology was given after detection on mammography. In fact, many additional DCIS are found contralateral to the index lesion on MRI (Boetes et al. 2007). It has also to be taken into account that a potential bias in many studies overrates the mammographic sensitivity against competing technologies (see next section).

2.2.2 Radiologist's Performance on Non-mass Lesions and Requirements for CADe/x

As initially indicated, invasive cancers are today easily visualized by MRI, with close to 100% sensitivity (Orel et al. 2001). Sensitivities for the MRI image-based visualization of DCIS, however, are reported between 77–96% (Raza et al. 2008), and for segmental enhancement patterns that are identified in the images, the differentiation is complicated, so that a high inter-observer variation in the classification and diagnosis of DCIS and atypical ductal hyperplasia (ADH) have been reported (Elston et al. 2000). The sensitivity number drops further when a more thorough analysis is performed. F Sardanelli et al. (2008) proposed to use the whole breast, sliced and examined by pathology, as a reference standard and then calculates sensitivities for both mammography and MRI that don't surpass 40%, and only in combination yield 46%. He concludes that the high sensitivities reported for MG/MRI reflect an incomplete study setup that overestimates the numbers because they are calculated from retrospective studies rather than a complete workup of the breast pathology. In his report, he notes that lesions that have been missed by imaging, but found by pathology, are typically smaller than 6 mm in diameter.

[15] ER (estrogen receptor status), CD34 (assessing angiogenesis), Ki67 (a proliferation marker), CD68 (assesses tumor-associated macrophages that indicate inflammation)

For lesions that are detected, it is important to diagnose them correctly. In a large data set, it has been attempted to find classifying features that differentiate normal controls from invasive and in-situ cancers of the ducts. Several kinetic, morphological, and texture features have been examined regarding their salience for classification, based on an automated segmentation of the enhancements using Fuzzy c-means clustering on the enhancement curves (Bhooshan et al. 2010). On their data containing 253 lesions (71 DCIS), the authors report AUC values of 0.83, 0.85, and 0.79 for the differentiation of DCIS from IDC, IDC and benign, and DCIS and benign lesions. Notably, in this study it was not a purpose to characterize the capability of the CAD to reduce false positive findings, but instead the differentiation of observer-provided abnormalities was assessed.

Recently, Yang et al. (2013) have independently proposed a method to differentiate benign and malignant cases based on the bilateral symmetry of three features estimated from the enhancement patterns in DCE-MRI. Their simple idea is to compare the overall average relative enhancement of left and right breast for early and late enhancement. These simple features result in an AUC of 0.78 using a Bayesian Belief Network classifier on their dataset comprising 130 retrospectively selected benign and malignant cases with unilateral findings with biopsy-proven outcomes. In conclusion, DCIS differentiation against benign findings in the breast is the most challenging task.

Novel attempts in contrast enhanced breast MRI imaging promise the potential for increased specificity. It has been proposed before to look at contrast agent uptake behavior during the perfusion phase (in the early contrast inflow phase) rather than diffusion part, requiring new MRI sequences that allow to image the first pass in higher temporal resolutions (Raza et al. 2008). This, however, is a subject that today only begins to be tackled in retrospective studies on clinical data, and sufficiently large data sets to draw initial conclusions are just emerging in hospitals employing fast uptake imaging sequences in their imaging protocol.

In this light, we will in the following propose a method for the automated detection and delineation of areas in conventional contrast-enhanced breast MRI that aims to pick up candidate regions that can further be examined using specifically tailored morphological features. With these methods, we contribute missing pieces to the art of computer-aided detection, characterization, and diagnosis of non-mass enhancement patterns. The methods are general enough to be adapted to emerging types of breast MRI sequences.

2.3 Symmetry in Non-Mass Lesion Detection and Delineation

Regarding computer-aided detection and diagnosis of DCIS, the above review of the state of the art shows that novel approaches are required (see Newell et al. (2009)). One reason for this is that the accurate delineation of enhancing areas is hampered by the diverse phenotype of DCIS, spanning from segmental enhancement to accumulations of small, but unconnected lumps. It is hence complicated to delineate DCIS with model-based or model-supported approaches which may be employed in the detection and segmentation of masses. On the other hand, the result of the detection and segmentation step determines the performance of features that describe the enhancement patterns and morphology of the area, which are ultimately submitted to train or query the predictor.

Learning from mass CADx approaches, many early attempts in DCIS characterization have looked at kinetic features, but they are proven to be insufficient to characterize non-mass like lesions. Even quantitative pharmacokinetic parameters appear to be diagnostically more useful in mass lesions than in non-mass lesions, because too often the enhancement characteristics of malignant non-mass-like lesions mimic those of benign mass lesions (Jansen, Shimauchi, et al. 2011).

Several approaches have tried to devise computer aid for the detection and delineation of non-mass-like findings in breast MRI. Some are based on automated segmentation of enhancing areas, for example using fuzzy clustering on the time-resolved data (ibidem) or by using Independent Component Analysis (ICA) on the time curves (Goebl et al. 2013). Various methods to derive quantitative features of the segmented areas have been employed, using either descriptive parameters for the wash-in/wash-out curves, or empiric or more biologically motivated pharmacokinetic models that result in quantitative parameters. In terms of sensitivity and specificity, however, none achieve a performance matching CADe/CADx approaches for the detection and classification of solid masses (Goebl et al. 2013; Jansen, Shimauchi, et al. 2011; C. Kuhl 2007; Orel et al. 2001). A recent description of a SVM-based machine learning approach classifying 84 lesions based on kinetic and morphological features including different kinds of moments (Zernike, Krawtchouk) has been described by Hoffmann et al. (2013), but again, only two of the six feature classes exceed the AUC = 0.7 level slightly, and only for linear SVM kernels. No fused classifier has been described in the publication, and it provides no insight into the classifier training, and no sensitivity/specificity figures allow a more detailed discussion of the results.

When looking at lesions that express vascular growth factors and hence trigger angiogenesis, one obvious thought is to assess the asymmetry of the vascular system in the bilateral breasts. There are reports where the symmetry of the breast vascular tree has for this purpose been estimated manually on maximum intensity projections or 3D reconstructions of contrast enhanced breast MRI. It has been established from these studies that the vascularity correlates with the presence of invasive breast cancers, but it has also been seen that this is less prominently related to the presence of DCIS (Orguc et al. 2012; Verardi et al. 2012). Also in the work of Newell et al. (2009), the performance of both texture and kinetic features doesn't appear to be sufficient to differentiate non-mass enhancements from benign cases. The authors reported an AUC of 0.78 on the data they used for method development, which is likely to be a positively biased estimate of the true performance. For all lesions, they report an AUC of 0.86, but in their data set, the proportion of mass lesions is much higher than in a data set we will look at in the subsequent method development, which is one of the factors making comparisons to prior work hard if it was not executed on the exact same data.

The human body is intrinsically symmetric; many parts of the body exist twice, like the limbs, eyes, and ears. The same is true only for fewer internal organs like the kidneys, but of the other internal organs, some are symmetric to a center plane, like the lungs and the brain. It is

hence intriguing to look at symmetry in a location-based manner, examining the corresponding segments jointly, to find indicators for disease in these organs. For the brain, this has been pursued successfully, and it is conceivable that the co-registration based approach of THIRION et al. (2000) can be applied for several such problems.

The female breast, however, is different. Not only the internal variation and variability are large, but also the differences in shape and size of left and right breast can be considerable. Therefore, we are convinced that a symmetry quantification on a voxel-wise basis, for example after applying non-rigid registration between left and right breast, is no viable approach. Rather, the two breasts need to be assessed in a manner that quantitatively compares the whole breasts.

Symmetry has been considered in medical image analysis for other organs than the breast, perhaps starting with the work of THIRION et al. (ibidem), who assess the level of asymmetry by evaluating the displacement fields computed in a non-linear co-registration of the images. While this approach is interesting and helpful in organs that are naturally very symmetric, we believe that for the breast, a direct evaluation of the displacement field is not helpful since the natural asymmetry of contralateral breasts in parenchymal structure is already high, but will still be dominated by the asymmetry introduced by differing sizes and bilateral variations caused by positioning in the imaging device. THIRION et al.'s work has consequently been recognized mostly in the area of brain image analysis and atlas building.

We hence propose an integrated CADe/CADx computation scheme building upon a novel symmetry criterion on so-called suspicion maps, which considers gray-value based symmetry without requiring co-registration, and even not assuming more than local consistency of gray value ranges. We derive a feature we dub the *contextual symmetry feature* that emulates the way DCIS is usually detected by radiologists. We further show how kinetic and morphological features help to achieve a voxel-level segmentation and a lesion candidate classification. This approach yields a performance approaching that of computer-aided methods used for masses — currently without including a plenitude of kinetic or morphological or textural features.

The approach described here unifies non-mass enhancement detection and diagnosis, though it will be described separately. Clinically, our developments may aid radiologists as a second reader, but it also has applications in automated reporting of DCIS findings, since from the segmentation result, characteristic features according to the BI-RADS standard can be derived with computer aid.

2.3.1 A Framework for Non-Mass Enhancement Characterization

Our contributions have been designed for easy extensibility or adjustments of all processing parts. The computational framework we propose consists of three major components. In the following, we describe algorithms to implement each of the components. The implementations already provide several configurations that open them to further research, but they may easily be extended to contain further functionality or substituted by alternative implementations.

Candidate Region Extraction We propose a novel scheme to extract candidate regions from contrast-enhance breast MRI. It is based on bilateral symmetry, assessed on subtraction images characterizing early uptake. As alternatives to subtraction images, descriptive dynamic parameter maps have been implemented, as well as texture feature maps. Both will not be treated here.

Feature Extraction We extract features from these regions utilizing modular implementations of the morphology and contrast agent kinetics features. As an alternative, an object-based image analysis (OBIA) toolkit (HOMEYER et al. 2010) has been fused into our framework so that its full capabilities may be harvested. We discuss the first in more detail, but also show preliminary results of the OBIA-based feature extraction.

Characterization Classifying the candidate regions may be done within the same OBIA framework using an efficient implementation of Random Forests; for research purposes, however, we exported all examples to assess the prediction performance utilizing different classification methods in greater detail.

Except the machine learning tools employed in the external validation of prediction models, all components are implemented in the MeVisLab environment, such that all general methods potentially also contribute to applications other than the specific task we developed them for.

Data

Our data set comprises a total of 49 patient cases acquired on 1.5 T scanners (all manufactured by Siemens, Erlangen)[16]. A dedicated breast coil (CP Breast Array, Siemens, Erlangen) was used in prone patient placement. Details of the specific settings of the DCE-MRI can be seen from Tab. 2.1. All patients were confirmed by histopathology based on specimen from vacuum assisted core needle biopsy or open (excision) biopsy, or after surgical removal of the index lesion.

Table 2.1: Details of the scanner settings used for the acquisition of the data used in this section. Res.: In-plane resolution. TR: repetition time. TE: excitation time.

Scanner	Field (T)	Res. (mm)	Spacing (mm)	TR (ms)	TE (ms)	Timing (sec)
Avanto	1.5	.66-.72	1.3	7.8	4	226 – 111 – 111 – 111
Sonata	1.5	.62	1.3	7.5	4	234 – 110 – 112 – 110
Symphony	1.5	.66-.70	1.3-1.35	7.8	4	201 – 90 – 90 – 90

In the images, experts marked the histologically proven areas of DCIS in 16 cases. In addition, 6 cases with benign and 23 with malignant tumors are in our data set to support an in-depth evaluation. The malignant lesions comprised invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), lobular carcinoma in situ (LCIS) and metastases from other organs. Benign findings at histology included fibrocystic changes (FCC), adenosis and hyperplasia. Two cases are without any remarkable enhancement (normal controls). In two more cases, the lesion was not annotated, hence they only contributed normal tissue samples.

All annotations have been performed by experienced readers, either providing a ellipsoid ROI covering the lesion and some surrounding tissue, or using the semi-automatic segmentation algorithm described in J H Moltz et al. (2009).

Preprocessing

Masking and Motion Correction For many of the subsequent processing steps, we require knowledge of the breast tissue region, and also the chest wall, which is in our application defined to be the boundary between breast tissue and the pectoralis major muscle. We segment the breast tissue area from the MRI images using a method based on the sheetness of the boundary layer. This method is tailored to the delineation of the breast tissue based on a non-fat-suppressed T_1-weighted image series (L. Wang, Filippatos, et al. 2011)

Also, we apply motion correction to all DCE-MRI images to remove artifacts emerging from patient motion (Boehler, Wirtz, et al. 2007). It is known that patient motion may mimic enhancements, so that they potentially introduce asymmetry, since motion frequently occurs unilaterally: During the MRI scanning process women relax and sink into the breast coils more and more deeply. Particularly the pectoral muscle relaxes and gives way for the breast tissue to expand into the breast coil. A visible trace of this motion can oftentimes be observed in

[16]Data courtesy of Radboud University Medical Center, Nijmegen, Netherlands

subtraction images, where the outline of the breast appears to be emphasized by a bright rim; compare Fig. 1.11 on page 31.

Nipple Detection Lastly, we estimate the nipple position, which will be used for enhancement shape analysis. We propose a straightforward method to detect the nipple position from contrast-enhanced breast MRI. We assume that the nipple position is towards the anterior part of the images, and we know that the nipple normally enhances bilaterally in DCE-MRI. The method multiplies the late enhancement image with a gray value ramp that decreases in value from the anterior image border to the posterior boundary, thereby penalizing posterior image parts and favoring enhancing anterior portions. The result is masked with the breast shape and cropped to include only the breast skin area of the most anterior part, and after an iterative refinement to a cluster small enough to represent a nipple, the algorithm finishes.

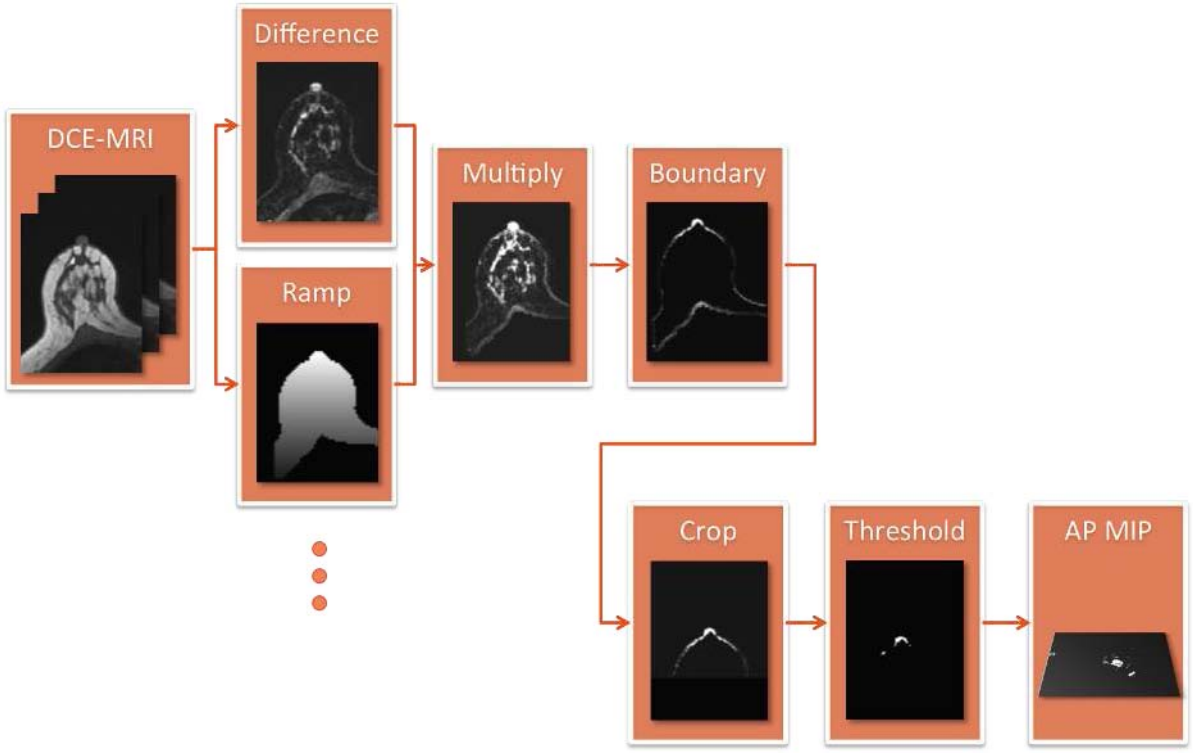


Figure 2.10: The mammary papilla (nipple) is detected in a sequential optimization pipeline. From top left, first the subtraction image of a dynamic series is multiplied with a gray value ramp from anterior to posterior. Of the result, only a boundary layer of a few voxels in the anterior half is retained and thresholded to 25% relative gray value. A coronal (AP) projection of the result is then submitted to an optimization process that leaves only the largest round component per breast, which is assumed to represent the nipple. The position is then converted back into the coordinate frame of the 3D image stack.

Note that at maximum 15 iterations are performed. The number of iterations required are a good indicator of success, hence the algorithm has inbuilt self-control which can be exploited to enhance the results in cases of failure: In such cases, the preselection of viable candidates could further be enhanced by including for example local curvature similar to the gray value ramp, but since this slows the algorithm down by an order of magnitude, it has not been included in the fast setup proposed here. The process is depicted in Fig. 2.10.

The algorithm can easily be extended to include new features that indicate the nipple position, for example local curvature. Any feature represented in a probability or score that is computed per image voxel can be combined arithmetically with the nipple position probability image created in the first step of the algorithm (indicated by the vertical dots in the figure). Currently, the probability is modeled firstly by assuming that the nipple is in the anterior portion of the MRI (towards the top of the image as it is displayed in the figure), and that the nipple enhances, which is why the difference image is taken into the computation.

The algorithm has been independently evaluated. For this purpose, the nipple positions of both breasts in 71 cases with and without disease have been annotated by the author, and the identified nipple position has been marked with one (x, y, z) position. The nipple is not always easily identified based on the outer shape, particularly in cases where the women were allowed to wear a shirt during the examination. This breast MRI acquisition technique is sometimes used to diminish motion artifacts since the shirt in the coil induces a slight fixation of the breast. In such cases, the nipple is often depressed, and sometimes only visible if the parenchymal structures and milk ducts leading towards the nipple are being followed. It helps in most cases to look at the enhancement to define the nipple position more accurately.

Three cases with about 100 mm deviation between ground truth and automatic marker may be considered outliers. They are due to a too small field of view in the projection step of the pipeline, which is easily adjusted dynamically. Without these cases, the average distance between manual and automatic annotation is 13 mm (±13.7 mm), the median distance is about 7.7 mm. Including the outliers, the mean distance is 15.2 mm (±19.6 mm), and the median distance increases to 8.0 mm. The two plots in Fig. 2.11 show the distributions of distances. Some typical results are presented in Fig. 2.12, where in particular the largest outliers are shown.

The accuracy of the nipple detection can be assumed sufficient for use to indicate the direction for the morphological parameters introduced later in this section. Simple modifications may improve the algorithm further, making it more reliable, but for the purpose of the following application, no further efforts are required. One interesting finding in the data remains unexplained: A consistent, though not significant difference in detection accuracy between left and right breast is observed, which is the reason why the two sides have been evaluated separately for the box plots.

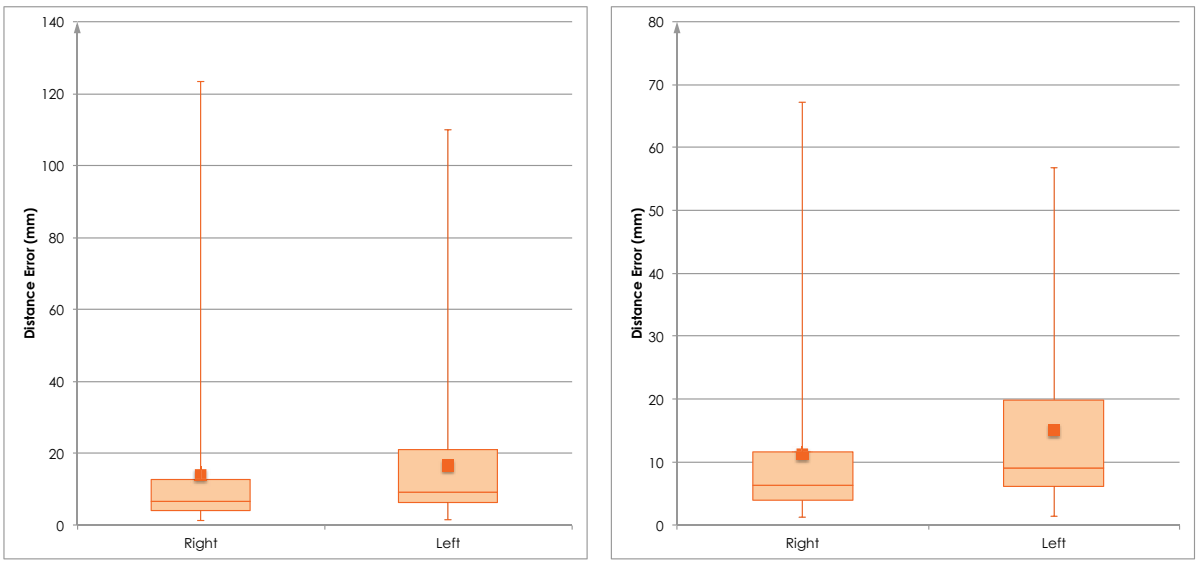


Figure 2.11: Evaluation of nipple detection algorithm, left with three outlier cases, right three outlier cases removed. Evaluation was conducted on 71 cases with ground truth nipple position annotation done by one experienced reader (Graphs on different scales).

Symmetry

Next, features from three feature classes are calculated from the data for the purpose of DCIS delineation. Subsequently, operations on images will be described. Clinical images are usually stored in integer format, i.e. the scalar gray value information is in $\mathbb{N}$. Hence, we define an image as follows:

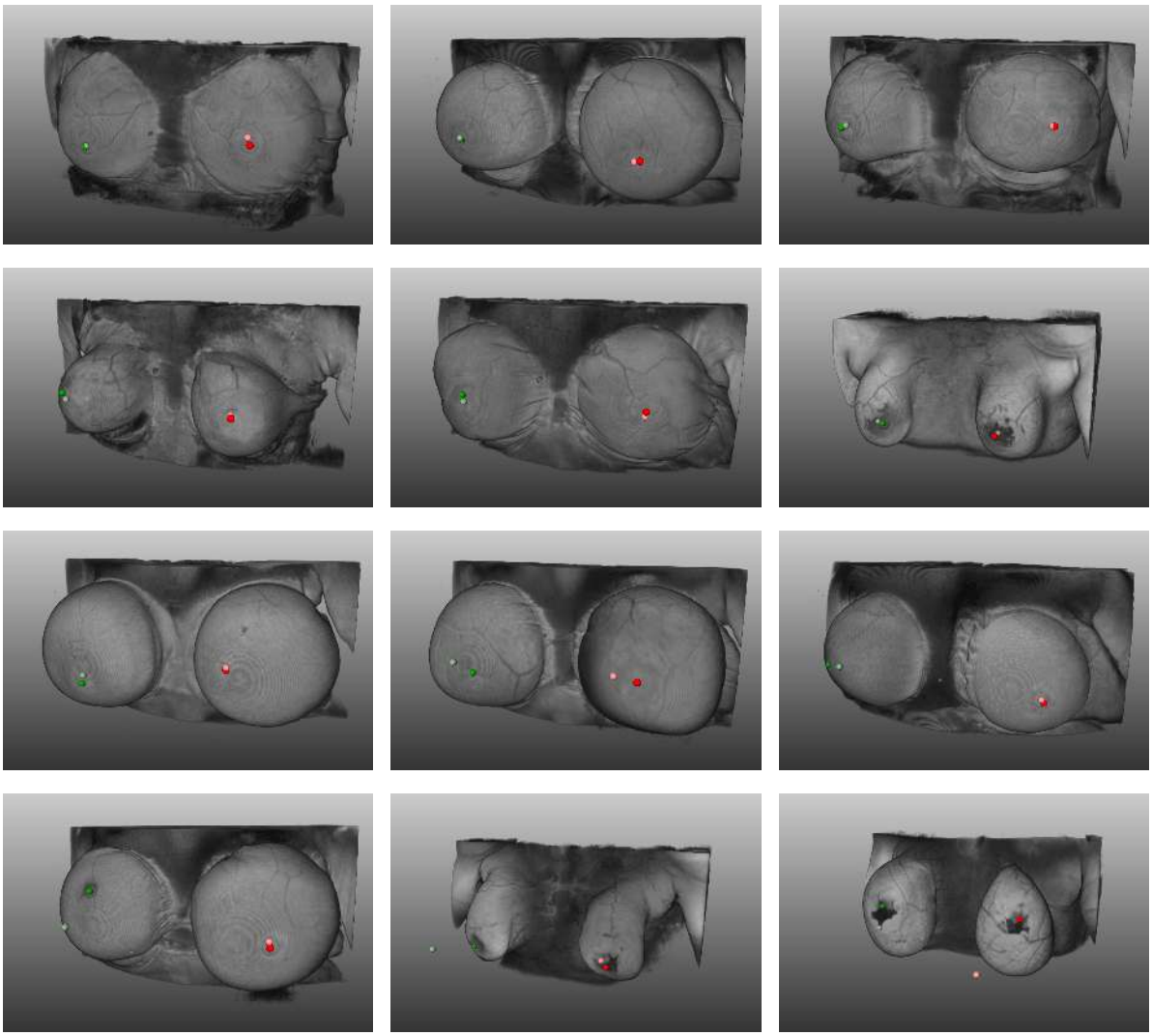


Figure 2.12: Performance of the nipple detection algorithm on example cases. The top two rows show cases with distances of under 10 mm, followed by one row of cases with more than one centimeter distance, and the last row, where the automated detection failed in one of the breasts. The manual markers appear in saturated colors, while the automatic markers are slightly de-saturated. Note that the proportion of outliers in the figure exaggerates the true proportion.

Definition 2.7 — Image. A discrete 4-dimensional image I is a set $I \in \mathbb{N}^4$, with atomic voxels $I(x, y, z, t) \in \mathbb{N}$. Intuitively, this describes a volumetric image in three-dimensional Euclidean space where each image voxel additionally has a number of time points.

The idea behind the symmetry criterion that we propose below is based on some basic assumptions and considerations. Firstly, we do not want to preprocess breast MRI so that scanner influence, protocol parameters, or imaging artifacts are being accounted for, rather, we want our method to be robust against those influences. We have to assume, however, that they are bilaterally comparable. This can safely be assumed for imaging and coil artifacts, and certainly for all protocol- and scanner-dependent parameters like contrast, voxel sizes, etc.

We further do not want to make assumptions on equal breast sizes or positioning, and in particular, we want to avoid the requirement of voxel-accurate left-to-right image registration in our processing. Hence, all criteria have to be independent of locations, but should capture local contexts, and should as well be independent of absolute intensities, but should quantify relative intensity changes. Moreover, a method that allows for experimentation with regard to the local

context to be considered might turn out to be advantagous.

Our proposed method fulfills all the above requirements. It measures the probabilities of observed instantiations of a certain local *context descriptor*. This descriptor captures local context information in the form of a spatial intensity variation histogram of the source image. A part of the breast MRI containing only one breast is analyzed by applying this context descriptor, and from the collected information we create what we dub a *suspicion map*. The suspicion map quantifies for the context descriptor how frequently and where its instantiations are observed in the target image (breast MRI part of the other breast), by looking up the probability of the context descriptor in the map. Depending on the actual implementation of the context descriptor, the suspicion map may have low entries for findings, or high ones.

The context descriptor can in our framework be defined arbitrarily. A simple and obvious choice is a tuple with the gray value of the current and a neighboring voxel, and in fact this descriptor practically shows robust results at moderate computational costs, as will be seen further on.

The descriptor probability map is for this simple descriptor equivalent to a non-symmetric horizontal gray level co-occurrence matrix (GLCM; compare Sec. 2.4.2), which also collects the number of occurrences of each possible gray level pair and can be normalized to interpret the entries as probabilities. But different from the usual application of GLCMs in image processing, this time we do not proceed by deriving secondary features from the GLCM (e.g. Haralick texture features), but interpret the target image directly using the encoded probabilities. For this particular context descriptor, we are interested in the areas of the map that correspond to *missing* instantiations of the descriptor, i.e. it was found in one side, but not in the other. This means, the suspicion that an abnormality has been detected is highest where the entries in the map are lowest. In other words, the suspicion map that the GLCM produces is practically an inverse suspicion map.

The proposed approach has the great advantage that it is conceptually trivial to extend to higher dimensions or more complex descriptors, but a simple consideration reveals that to follow this temptation would likely be punished by deteriorated performance in our application: We have to expect images of sizes that will fill only a few entries in the probability map, and might have very few voxels contributing to the entries. For example, one central image slice containing the area of interest of the breast can be assumed to provide around 20.000 voxels. Assuming only 256 gray values to fill the GLCM, giving tuples, it has $256^2 = 65536$ entries, already three times more than voxels to evaluate. In consequence, there will be only "infrequent" descriptor occurrences in the target image, which is interpreted as a equally high likelihood of abnormality throughout the target image. To overcome this, the number of distinct gray values needs to be reduced to yield fewer possible combinations of gray values and consequently a denser descriptor. This will be discussed in a later chapter in more detail (see Sec. 2.4.2 on page 87).

We learn from this thought experiment that for our purposes, the number of possible GLCM entries has to be kept small compared to the voxels in the image. Because we are still interested in an evaluation that takes the three-dimensionality of the images into account, we apply the processing sequentially in all three axis-aligned projections (transversal, sagittal, coronal). The results will then be fused into one three-dimensional suspicion image.

For the implementation, it is advantageous that we are not interested in the calculation of texture features from the GLCM, hence we can optimize both speed and space complexity by using a dictionary to store the number of gray level neighbor relationship instantiations. In the following, we consequently introduce a formalism of context descriptor probabilities to show the emphasis on probabilities and the extensibility of the framework more clearly. The result of the context quantification step is a map of the same size as the original image, containing probabilities in each voxel that this voxel is situated in a context that is not prominent contralaterally.

Capturing Local Context In the following, we look at single slices of images (i.e. at two-dimensional images) with only one time point and omit reference to z and t silently. Note that, since we are interested in the bilateral symmetry in breast MRI, we will look at left and right breast separately. When defining and applying the context dictionary, learning it is performed in one breast, for example the left, which we call the "source" breast. The learned dictionary is then applied to the other breast, the "target" breast.

Capturing the local context into the context dictionary is done using two functions, the first one mapping from a voxel position to a gray value combination, the second mapping from gray value combinations to probabilities.

Let us assume a n-dimensional image with gray values $g \in \mathbb{N}$. The first mapping is what we call the context descriptor. It maps from image positions in an n-dimensional image to a tuple of m gray values, or

$$\mathcal{C} : \mathbb{N}^n \mapsto \mathbb{N}^m. \tag{2.14}$$

The second function implements a look-up table. It takes one of the gray value combinations obtained by the above mapping, and returns a probability. This can be expressed by

$$\mathcal{L} : \mathbb{N}^m \mapsto \mathbb{R}. \tag{2.15}$$

The full mapping from an image voxel to the probability can then be perceived as the context dictionary $\mathcal{D} : \mathbb{N}^n \mapsto \mathbb{R}$. In a practical implementation, this can be thought of as a n-dimensional look-up table that is indexed by the gray values retrieved from the image using the context descriptor $\mathcal{C}$ and returns probabilities.

For our implementation, we define the context similar to the one used in a gray value co-occurrence matrix for two-dimensional images, using a direction and distance from a given image voxel (x, y). This can be expressed by a tuple $(\Delta x, \Delta y)$ for the two-dimensional case; the full context descriptor is then defined to map a voxel (x, y) to a tuple of gray values (a, b) with $a = I(x, y)$ being the gray value of the image at (x, y), and $b = I(x + \Delta x, y + \Delta y)$. Hence, we can write $\mathcal{C}(x, y) = (I(x, y), I(x + \Delta x, y + \Delta y))$.

The probabilities required to build the look-up table $\mathcal{L}$ are obtained by counting all occurrences of (a, b) in the image, and dividing the result by the number of voxels considered.

From the look-up table, the suspicion map corresponding to the target image can be obtained by considering the corresponding target breast I^T. The suspicion map $\mathcal{M}$ is of the same size as the target image, and its voxels have values obtained from the context dictionary by looking up the entry $\mathcal{L}(\mathcal{C}(x_0, y_0))$ for all (x_0, y_0) in the target image. Taking into account that not all $\mathcal{C}(x_0, y_0)$ need to exist in the context dictionary, the suspicion map is obtained as

$$\mathcal{M}(x, y) = \begin{cases} \mathcal{L}(\mathcal{C}(x_0, y_0)) & \text{if } \mathcal{C}(x_0, y_0) \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases} \tag{2.16}$$

Quantifying Asymmetry from Suspicion Map This results in an asymmetry score between enhancements of contralateral breasts, defined as a sum of the per-slice intra-breast differences in areas of suspicious regions: the scalar asymmetry score for one slice is the absolute difference of the entropies of the suspicion maps of left and right breast.

We are using the Shannon entropy definition, by which the entropy of a discrete random variable X with possible values $x_1, \ldots, x_n$ and a probability mass function $P(X)$ for a finite sample can be written as

$$H(X) = -\sum_{i=1}^{i=n} P(x_i) \log P(x_i). \tag{2.17}$$

The binary case, where we set $p = \Pr(X = 1)$, simplifies to $H(X) = -p \log p - (1-p) \log(1-p)$, using the logarithm with base 2.

The entropies for all slices are summed up, weighted with the number of breast mask voxels in this slice to rate asymmetry in central slices higher than asymmetry in auxiliary slices of the MRI that potentially exhibit widely differing breast coverage. In addition, the individual per-slice asymmetry values were averaged into one per-image total. These two features were computed from the maps generated in axial, sagittal, and coronal views, and averaged from the three projection directions. In addition, the features derived from only the axial views were added to the feature set to compare the performance of limited three-dimensional assessment with one direction alone.

Note that this step yields a decision about the entire case. For the further features, however, voxel-wise processing is required, hence we keep from this step a collection of 6-neighbor 3D connected components derived from the binarized suspicion maps for further consideration in the next two processing steps. These are further called the regions of interest, or for short, ROIs. Note that given by our implementation of the context descriptor, low entries in the suspicion map yield the ROIs. In our implementation where the input images are binned into 32 gray levels, a thresholding to keep only the zero probability regions in the suspicion map is appropriate.

Morphology

The second class analyzes morphological features on each candidate region (per slice) and consists of the principal directions of the mass distribution of the selected component and their ratio. Additionally, we propose a novel feature which we denominate projection distance variance. Essentially, it quantifies the directionality of enhancement towards the nipple, by calculating for every voxel in the DCIS candidate region its distance to a line connecting the component's center of mass with the nipple position. The variance of these projection distances to the generated line is calculated (see illustration in Fig. 2.13). Low values are indicators for enhancements that point towards the nipple. This feature alone already shows some separation between the normal and DCIS enhancements.

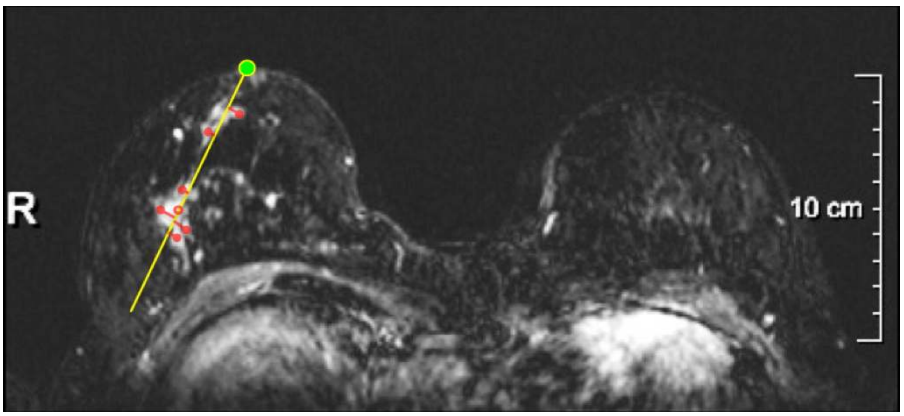


Figure 2.13: The projection distance feature captures the variance of distances of voxels in the lesion candidate to the line towards the nipple.

Kinetics

The third class analyzes kinetic features. We consider only descriptive parameters derived from the pre-contrast, peak (200 sec post contrast) and late (>300 sec post contrast) acquisitions. Our set of features has been carefully chosen from those that are known from literature to be most salient to characterize DCIS, and encompasses wash-in slope, wash-out slope, integral

under curve, relative enhancement at the first time point, time to peak and signal enhancement ratio (aka relative enhancement). Since these features are well documented in the literature, no detailed description is given here (Bhooshan et al. 2010; Jansen, Shimauchi, et al. 2011).

2.3.2 Automated Characterization

Classification is conducted on a per-ROI basis with the aim to show how much the classifier is capable of reducing the number of false positive findings of the detection algorithm. We chose to evaluate the performance by leave-one-patient-out cross validation, since this conforms best with the clinical scenario, where per-case variability in contrast agent uptake, motion artifacts, and other influencing factors will vary for all detected lesion candidates alike. In all experiments, a Random Forest (RF) classifier (Breiman 2001; Demšar, Curk, et al. 2013) and a Naïve Bayes (NB) classifier have been trained and evaluated.

Joining Symmetry, Morphology, and Kinetics

ROIs For the 49 cases, our region extraction algorithm on average finds about 19 ROIs per case (936 regions in total). We have considered ROIs to be representative of a lesion when there was a marked overlap with a lesion expert annotation. Since not always all diseased tissue (but only the later biopsied site with a definitive result from pathology) has been annotated, we have manually checked the expert annotations against all detected ROIs and deleted all ambiguous ROIs, i.e. those with clearly malignant uptake characteristics consisting of high and fast initial enhancement and pronounced washout behavior, when the ROI was close or bordering the annotated area. Highly suspicious ROI detections in the breast contralateral to the index lesion were only excluded in two cases: in one, where the expert annotation was clearly in the wrong breast, and in one where the pathology and target lesion morphology suggested presence of an extensive disease. Note that the manual preparation of the learning and testing data set was required since the detection results are voxel-accurate, while the expert annotations consisted of one or many spherical annotations that naturally overlap the lesion and some healthy tissue.

When matching those regions detected by the algorithm to those annotated by the experts, 10 ROIs have been detected in the cases containing benign lesions. 52 ROIs were considered to be examples of malignant lesions except DCIS, and 43 further ROIs represented DCIS lesions, summing up to a total of 95 regions of malignant types. The purpose in including cases with malignant and benign lesions besides the DCIS cases in our data set was to allow for more differential tests of the capabilities of the methods we propose.

The remaining 841 ROIs are situated in healthy tissue, identified in other parts of the breasts of diseased and normal cases alike. Frequently, the heart was detected, and also strong motion artifacts along the breast-air boundary often lead to false positive detections. Likewise, skin folding and inhomogeneity of the magnetic field of the MRI scanner account for a fraction of the false detections.

Evaluation Method and Results For performance evaluation, the 10 ROIs of benign lesions have not been individually classified, but taken into the class of normal ROIs, denominated the "benign" class. Likewise, the DCIS and malignant ROIs were together treated as the "malignant" class. Thus, 851 benign and 95 malignant ROIs formed the highly unbalanced data set for machine learning. Since it is known that Random Forests do suffer from class imbalance, we set up our training procedure in a wrapped procedure, where inside the leave-one-patient-out cross-validation, a n-fold instance subset selection is applied to the ROIs from the benign class. n is per case chosen such that roughly a 1:1 ratio is obtained between benign and malignant ROIs. On each such training set, the two classifiers (RF and NB) are trained and subsequently evaluated on the held-out case. All resulting AUCs, sensitivities, and specificities are averaged.

Joining all features (symmetry, morphology, and kinetics) results in a Random Forest predictor with an AUC of 0.91 on the two-class problem (benign vs. malignant). The operating point on the ROC curve is at a sensitivity of 76% and 82% specificity. The Naïve Bayes classifier reaches an AUC of 0.92 with a sensitivity of 84% and a specificity of 76%.

In Fig. 2.14, results are visualized for two exemplary data sets containing malignant lesions. In the figure, colors indicate true positive marks (red), true negative marks (blue), false positive marks (yellow), and false negative marks (orange). The lower of each image pair shows the expert annotation as a green blob overlaid onto the volume.

Internal Validation According to the introductory remarks on model evaluation, experiments were conducted to validate the results against randomized settings, with the aim to underpin that all achieved results are not due to errors in the experimental setup.

An experiment with permuted class labels (for the two-class problem benign vs. malignant) was set up according to Boulesteix et al. (2008). The labels were permuted randomly into five new orderings, and a Random Forest classifier with specifications as before was trained on each of them. The ROC curve and AUC value were inspected and recorded. No ROC curve deviated markedly from the diagonal in the diagram, and no Random Forest reached a performance above an AUC of 0.500, both corresponding to chance decisions. A statistical significance test to prove the difference to the above results was not conducted.

Additionally, the influence of different cross-validation method has been explored. In general, the predicted performance obtained using the leave-one-patient-out cross-validation as described above is greatly inferior to that obtained with leave-one-ROI-out cross-validation. This is expected, since the potential difference between ROIs of different patients due to patient-individual variation in the breast MRI acquisition is likely to be an influential factor, hence if by chance the patient's individual image characteristics are known during classifier training, this might positively bias the performance estimate.

Summary

Two relevant developments have been described so far, namely

1. the implementation of a symmetry criterion that is free of most assumptions and reasonably robust against influences in common clinical scenarios (bias field, changes of scanners or protocol, voxel sizes and isotropy, etc.); and
2. a simple and efficient way to include the morphology of ductal enhancements with the feature dubbed projected variance.

We found the symmetry-based detection of ROIs to be robust and reliable as long as the preconditions of the algorithm are met. The performance of the symmetry criterion alone is insufficient in the used data set to tell benign from malignant cases with the clinically desired accuracy, but it adds a novel feature to the classification task. Many avenues of research are open to improve the performance of both the symmetry criterion itself, as well as the lesion candidate region extraction the algorithm performs while calculating the symmetry score. The proposed morphology feature has not fulfilled the expectations, while still showing promising results that encourage further research in this direction.

In our general approach we acknowledge the conceptually interwoven nature of detection and characterization (CADe and CADx). The combined detection and characterization of non-mass-like enhancement patterns in breast MRI as presented above exploits the strengths of several approaches, and in particular included one decisive addition to achieve reasonable specificity without sacrificing sensitivity, bilateral asymmetry.

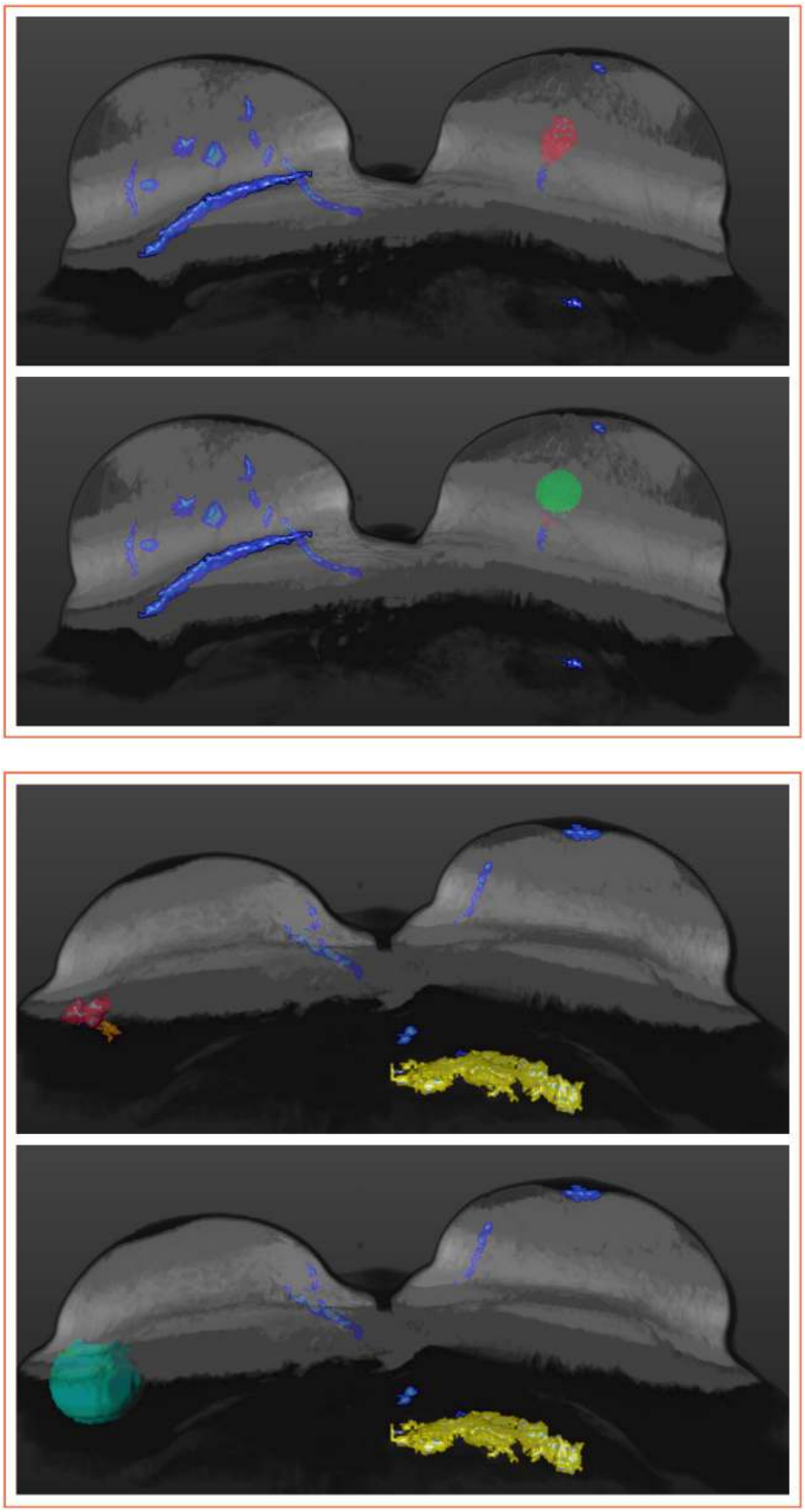


Figure 2.14: DCIS detection/characterization results for two cases. The top of each two images shows the detection results, colored according to the label received from the classifier. Red detection results are true positive detections, orange are false negative results, blue encodes true negative characterizations, and yellow areas are false positive detections. In second of each image pair, the expert annotation has been added in a green color.

2.3.3 The Predictive Value Of Symmetry

The symmetry based suspicious area detection shows potential as a preprocessing step to derive candidate lesions for further examination in machine learning settings. This brings motivation to think about extensions of the design and evaluation of the symmetry feature beyond the level that has been achieved. One principal limitation of the presented approach should be mentioned, however, which is that if the asymmetry assumption is broken, the algorithm will not provide useful results. This is most prominently the case when a woman presents with bilateral disease, and of course in cases of follow-up examinations after unilateral mastectomy without breast reconstruction or when an implant replaces the excised breast tissue. Large differences in positioning of left and right breasts in the coils may also hamper the results.

Of the potential approaches to detect bilateral asymmetry in the breast, we have proposed one that relies on local gray value neighborhoods alone. We acknowledge that alternatives exist that we haven't pursued thoroughly. We therefore aim to proceed in two general directions:

- Regarding the symmetry quantification, we propose to explore improvements of the symmetry criterion applied in this work. Non-rigid co-registration of the left and right breast might increase the accuracy of the symmetry criterion, because spurious detections caused by large differences in positioning might get reduced, and with them the false positive detection rate may drop. Criteria derived from the deformation field mainly describing the shape difference between left and right breast, considered together with the detection result, might serve as a novel feature as well.

  Moreover, discarding the intermediate results obtained during the calculation of the symmetry quantification might be avoided. Instead, a feature vector obtained from the quantification might allow more insights.

- Explore improved morphology features. Most importantly, we aim to describe the spatial configuration of detected lesion candidates more thoroughly to be able to classify enhancement patterns according to the ACR BI-RADS® categories for non-mass-like enhancing lesions. This requires to develop features that capture for example multi-centricity and multi-focality, two descriptors which are characterized by lesion candidates above a certain grade of suspicion, being situated close to each other (multi-focal case) or further apart. Also, the internal inhomogeneity of the enhancing area in terms of the enhancement curve types is of interest to distiguish regional, clumped, and segmental enhancements. To characterize linear enhancements and differentiate them from ductal enhancements (which can be cone-shaped), further morphology describing features ought to be developed.

We feel that the sensitivity and specificity can be improved upon. To increase both, several steps may in the future be taken. A vessel detection, segmentation, and masking step may be introduced to help to improve the rejection of negative cases. Bias field correction, used as a preprocessing method, may reduce the effects of segmental enhancements as well as skin brightening, which is observed in some images and is not always symmetric.

The correct detection of early DCIS is also still not trivial, and not convincingly covered by our approach, since these early and small enhancing structures neither introduce strong asymmetry, nor do they exhibit marked enhancement patterns. In particular, they are almost indistinguishable from nodular enhancements along the ducts and lobules which naturally occur during the menstrual cycle or are caused by hormone replacement therapy. We hope to gain insights from future data collections in high-risk populations that are followed over years.

In addition, we hope that the developments in imaging technology, most importantly MRI sequences focusing on the wash-in phase with high temporal resolution acquisitions (e.g., TWIST and HiT), and attempts to create new tissue contrasts, for example high spectral-spatial resolution imaging (HiSS) or DIXON imaging may provide means to derive decisive features.

Lastly, it will need to be examined in a more fundamental sense how detection and diagnosis — two sides of a coin — can be brought closer together, resulting in an approach that builds a non-mass-lesion likelihood map from all features derived from the image.

2.4 Dynamic Texture Features for Mass Lesion Characterization

In the following, older works regarding machine learning algorithms applied to breast MRI images of mass lesions will be described. From today's perspective, many of the approaches chosen appear well justified, in particular the development of textural kinetic features. The machine learning setup that was applied, however, was not the central aspect. It is hence subject to further investigations in this thesis. It is the aim to understand the limits of machine learning better, exemplified on the data set treated in the publication underlying this section. It is this thorough investigation of the impact of methodological choices like cross validation, feature selection, and different classification methods that makes it interesting to describe the older approach before discussing it.

Parts of the work in this section have been published as Jennifer Loose et al. (2009). "Assessment of texture analysis on DCE-MRI data for the differentiation of breast tumor lesions". In: volume 7260. DOI: 10.1117/12.812971. URL: http://dx.doi.org/10.1117/12.812971. The results of and the comparison with nonlinear classification methods is first reported in this thesis. Also, the original discussion has been altered and relates the presented results to publications published after our work, presenting sometimes similar approaches.

We propose a method for the fully-automatic differentiation of mass-like findings (invasive cancers that are circumscribed) in contrast-enhanced breast MRI. Our goal in this work is to increase the specificity, noting that mass detection by computer-aided detection (CADe) systems and automated segmentation is usually tuned to be over-sensitive to avoid missing small lesion candidates. The high number of false positive findings from computer detection algorithms, on the other hand, is impeding the diagnostic workflow, as it requires the radiologist to spend time on each finding. This may even lead to an increased number of biopsy procedures out of the doubts the CADe algorithm raises.

An approach is presented to characterize the texture of dynamic contrast enhanced MRI data in three spatial and one temporal dimension simultaneously. The features are evaluated for their individual and combined contribution to a benign-vs.-malignant decision. The choice of texture analysis was made on the assumption that texture features are independent of absolute gray values, and may hence contribute to a differentiated, reproducible, and observer-independent quantification of finding characteristics. Prior work already showed the general feasibility of texture-based analysis(W. Chen et al. 2007; Gibbs et al. 2003), but does either not use all information dimensions available, or does not analyze all information in a comprehensive way. We set out to assess a plenitude of potential approaches to include the temporal information of DCE-MRI into texture-based criteria, which will lead us to large numbers of features. Instead of using machine learning tools that are robust against such large feature spaces, we declared it our goal to contribute to the understanding of the important traits of mass lesions in DCE-MRI that are accessible to automated analysis.

Hence, the contributions in this work are: We first propose a procedure that, in a tree-like process, ranks good features *in combination* to build a simple logistic regression classifier on. We compare this approach with alternatives that select best features from each individual feature's contribution, and also with modern feature selection methods. Second, we propose several approaches to include the temporal development of enhancement patterns into the texture features. Lastly, we assess the performance of the proposed feature selection approach and subsequent model building against classification models computed by state-of-the-art methods that, however, don't reveal each feature's importance right away.

2.4.1 Contributions to DCE-MRI Texture Analysis of Mass Lesions

Texture analysis has previously been applied to breast MR images. Gibbs et al. (2003) applied spatial gray level dependence method to post contrast MR breast images. They showed that

benign and malignant lesions differ in terms of the spatial variations in voxel intensities. Woods et al. (2007) investigated the use of 4D co-occurrence based texture analysis to distinguish between nonmalignant and malignant tissues in DCE MR breast images employing a neural network based classifier. S C Agner et al. (2008) introduced *kinetic texture features*, also called *textural kinetics*, which are derived from first- and second-order statistical features for each voxel in a lesion over time, yielding a kinetic texture curve. While Woods et al. were able to demonstrate promising accuracy values on their data set by calculating 4D textures per voxel using a sliding window approach, Shannon C Agner et al. (2011) in her more recent publications provide evidence that their textural feature curves are competetive if compared with signal intensity curves and morphological features on one data set of 41 benign and malignant lesions.

Our work follows similar lines of thought, but is different in the crucial aspect that both Shannon C Agner et al. and Woods et al. calculate voxel-wise texture features for each time point while the work presented here integrates the spatial and temporal variation into one quantitative parameter per segmented lesion. Hence, while the cited approaches can provide color maps that indicate the composition of the breast (or only the lesion) in terms of texture measures, the approach proposed here yields a single scalar score between zero and one perhaps contributing to estimate the probability of malignancy using machine learning techniques as before.

We also anticipate the results of Newell et al. (2009) who demonstrate that a combination of enhancement kinetics and morphology (though in their work only represented by Haralick texture features derived from co-occurrence matrices of 2D image slices) is the most comprehensive and effective approach to the automated diagnosis of masses that include texture.

Our approach includes 4D regions consisting of lesion tissue and surrounding parenchyma, imaged during contrast agent wash-in and wash-out, in the calculations — instead of only assessing "most suspect curves" calculated from manually defined ROIs inside the lesion. Input to the algorithm were automatically segmented lesions based on user-defined seed points. Of course, the method can also be used to assess input ROIs given from a automatic lesion detection algorithm. The resulting predictor may be integrated into a computer-aided diagnosis (CADx) approach that is aimed at classification of lesions and may easily be extended also to detect them. The goal of the work is therefore to propose a preliminary yet effective CADx scheme based on a subset of texture parameters chosen with respect to their suitability for the task, based on a set of annotated data with known biopsy results.

The following work has three aims.

1. We wish to propose an integrated computation scheme that normalizes the data for the particular task of integrated 4D texture analysis. To this end, we differentially characterize several ways to include contrast kinetics into texture feature computation.
2. We will present a feature selection approach that is suited to assess the large number of texture features for their predictive power. This helps us to gain insights into the feasibility of the different ways to include contrast kinetics.
3. We will finally characterize the predictive power of the full feature set comparing a linear regression model based on only very few extracted features with nonlinear predictors built from the full feature set.

2.4.2 Linear and Nonlinear Classification of 4D Haralick Texture Features

Patient data

The patient data for this study was acquired on a 1.5 T whole-body MRI system (Siemens Vision, Siemens, Erlangen, Germany) using a T1-weighted 3D spoiled gradient echo sequence.[17] After the acquisition of the pre-contrast series and administration of Gd-DTPA contrast agent (CA), five post-contrast volumes were acquired in 69 sec intervals, yielding a series of images I_0 to I_5 (see (W. Chen et al. 2007) for details). A total of 96 patients were included in this study. Per patient, only the index lesion was used for analysis. Histological workup partitioned the data into 60 malignant and 36 benign lesions of various types and was used as the ground truth for classifier training. Three dimensional segmentation masks manually established by expert radiologists were available for all lesions.

Texture Features

Preprocessing All images were resampled to isotropic voxels. Histogram equalization and histogram stretching were independently applied to normalize the textural characteristics in two different ways and assess their respective benefit (compare Appendix C.3 for details). Additionally the gray value range was reduced to $N = 32$ distinct levels to keep co-occurrence matrices within computationally feasible sizes. Note that the order of these operation matters, since typical histograms tended to be heavily skewed, so that rebinning from the original 12bit data (4096 gray levels) to 6bit (32 gray levels) sacrifices much of the desired gray value changes by binning it into one target gray level. The histogram operations and rebinning were done for subtraction images I_s and 4D images I_{4D}, where I_s is defined as the voxel-wise gray value subtraction, $I_1 - I_0$. I_{4D} consist of I_0 and I_1.

Next, for any given voxel, characteristics of the intensity time curves were quantitatively described by eleven parameters, for example two curve slopes describing the wash-in and the wash-out effect; peak enhancement (PE), and time to peak (TTP). Thus, eleven parameter maps I^p_{param}, $p \in [\text{Slope1}, \text{Slope2}, \text{PE}, \ldots, \text{TTP}]$, are derived for each image.

Next, non-directional co-occurrence matrices of size $N \times N$ were computed by summing up the directional co-occurrence matrices, being four in the 2D case, 13 for 3D, and 39 for 4D images.

Co-occurrence matrix calculation The co-occurrence matrix $C_{(\Delta x, \Delta y)}(i, j)$ is defined on a two-dimensional image and describes the distribution of co-occurring values for a given distance and direction. The elements of C are the numbers of times that gray levels i and j occur in two voxels of the image that are apart from each other by the given distance in the given direction, which is defined by the offsets $(\Delta x, \Delta y)$. While one direction and one distance (or one tuple $(\Delta x, \Delta y)$) define a co-occurrence matrix, it is the usual habit to produce a more comprehensive local description of gray value co-occurrences by summing up the co-occurrence matrices of multiple directions combined with one or more distances. The summation is an element-wise matrix sum of all As an example, Fig. 2.15 depicts how four directions in distance 2 are sampled for the co-occurrence matrix. The resulting matrix C is also referred to as Spatial Grey Level Dependence matrix (SGLD) or Grey Level Co-occurrence Matrix (GLCM). C can become sparse if the gray value range is large and the image small such that only a few of the possible gray value combinations are found in the image.

The following definition captures the above more formally:

[17] Data courtesy of Prof. Dr. med. Ulrich Bick, Charité Universitätsmedizin, Berlin, Germany.

Definition 2.8 — Co-occurrence matrix. The co-occurrence matrix of a 2-dimensional image is defined as

$$C_{(\Delta x,\Delta y)}(i,j) = \sum_{x=1}^{m}\sum_{y=1}^{n} \begin{cases} 1, & \text{if } I(x,y) = i \text{ and} \\ & \quad I(x+\Delta x, y+\Delta y) = j \\ 0, & \text{otherwise} \end{cases}$$

where the pair $(\Delta x, \Delta y)$ is the characteristic element the co-occurrence matrix depends on.

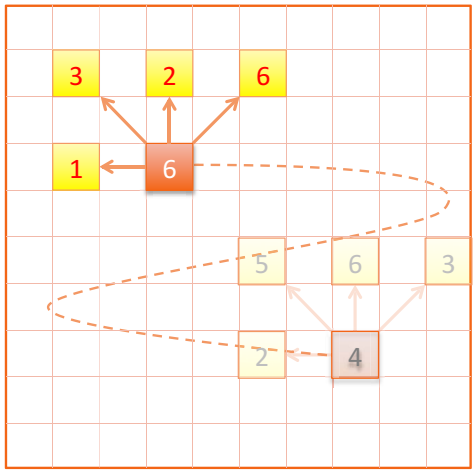


Figure 2.15: Directions of a co-occurrence matrix. The Figure shows the voxel of interest (orange) and four possible directions in a 2D matrix. For each direction a distance d is defined. In the Figure, $d = 2$ is indicated by the arrows. From the situation, four GLCMs will be built and accumulated, corresponding to the four pairs $(\Delta x, \Delta y) \in (-2,0), (-2,-2), (0,-2), (2,-2)$. The yellow voxels each contribute one entry in the GLCM together with the center voxel. Filling the GLCM can be conceived of as a filter operation into a 2D histogram.

Adhering to the definition, a GLCM is practically built by considering each source image voxel and its specified neighbor. The source image contains gray values $g \in \mathcal{N}$ in a value range $[a, b]$ with $a, b \in \mathcal{N}$. The GLCM will then be of size $(b-a) \times (b-a)$. Fig. 2.16 shows an example of a co-occurrence matrix of a 6×5 image with a value range of $[0, 3]$, resulting in a 4×4 GLCM. It should be noted that although the operation involved in the construction of the GLCM resembles a kernel operation on an image, GLCMs are in fact more comparable to a histogram.

Also note that several options have been described to build the co-occurrence matrix. Symmetric GLCM are built by increasing the $(i,j)^{\text{th}}$ entry and the $(j,i)^{\text{th}}$ entry simultaneously. This is equivalent to a symmetric setup of characteristic elements, e.g. $(\Delta x, \Delta y) \in (-2,0), (+2,0)$, but computationally slightly beneficial since neighboring gray values need to be looked up. This is how we built the GLCMs in the work described here. To obtain asymmetric co-occurrence matrices, the elements of C are updated strictly according to the characteristic elements. Asymmetric co-occurrence matrices cannot be used for the calculation of Haralick texture features.

Directional GLCMs are those obtained from only one (two) characteristic elements in the asymmetric (symmetric) case, for example the horizontal neighbors, the vertical neighbors, etc. If GLCMs corresponding to all four directions are averaged, the resulting matrix may be called direction-invariant. This averaging can in particular be obtained by normalizing the matrix.

If normalized such that the sum of all entries of a co-occurrence matrix equals one, an entry in a co-occurrence matrix describes the probability of the corresponding gray values to appear in distance d in the image. C is of size $N \times N$, with N being the difference of minimum

and maximum gray levels[18]. Since the number of distinct gray values for medical images is often high within the 12 bit per voxel limit (e.g. on average more than 500 gray values for the masked lesions under consideration), the number of gray values needs to be reduced in order to calculate co-occurrence matrices within reasonable times (Castleman 1996). It has been shown previously that the textural features do not depend significantly on the number of gray levels the image is reduced to, as long as the number does not fall below sixteen (W. Chen et al. 2007). Therefore, we rescaled the lesion images to 32 distinct levels.

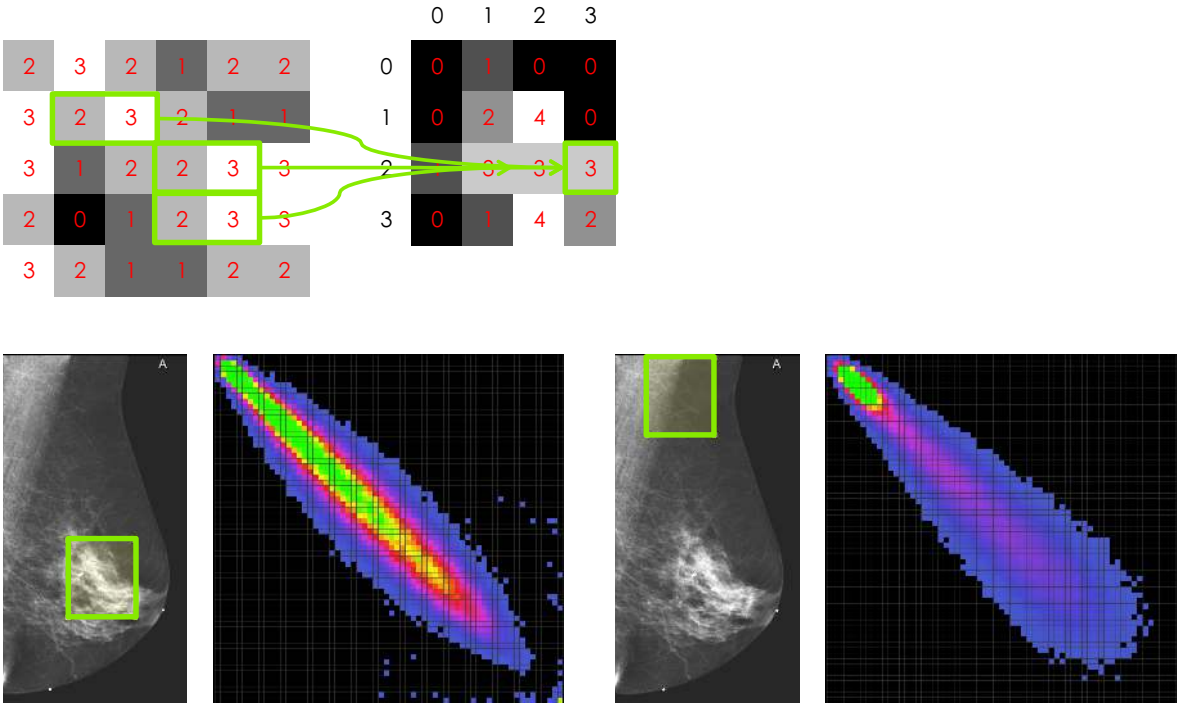


Figure 2.16: Top row: An example of a co-occurrence matrix of a 2D image. Left: The original image, the numbers indicate the gray values. Right: The resulting co-occurrence matrix $C_{(\Delta x, \Delta y)}$ for $\Delta x = 1, \Delta y = 0$, where the axes numbering corresponds to the four gray levels in the image. The characteristic element is again defined by the direction "right" and the distance 1, or the tuple $(+1, 0)$. With this characteristic element, the entry "3" in row 2, column 3 in the co-occurrence matrix indicates that the tuple $(2, 3)$ exists three times in the image. **Bottom row**: Two examples of image regions with the corresponding co-occurrence matrix, colorized to emphasize their different structures.

In our work we want to take advantage in the enriched information available in contrast enhanced 4D breast MRI volumes. Looking at the visual changes in enhancing lesions, several diagnostically relevant criteria may be observed, of which we cite only obvious ones. One important feature is called rim enhancement. In this dynamic pattern, the central region of a lesion enhances only after the lesion border. This is a strong indicator of malignancy. Internal dark septations becoming visible during enhancement of the finding indicate fibroadenoma, a benign finding. On the other hand, enhancing internal septations are signs of malignancy. More such patterns have been described (Glassman et al. 2013; Werner A Kaiser 2008). Consequently, one possible approach to harvest the information of the time-resolved series is by using texture analysis if the change of the texture can be quantified as it changes over time. While a calculation of 3D texture features for each time point is one alternative, we opt in this contribution for the calculation of co-occurrence matrices for full 4D image volumes to capture the information of wash-in and wash-out behavior implicitly in the texture feature. We wish to capture the direct neighborhood in all directions originating from the voxel of consideration in an asymmetric co-occurrence matrix. With a total of 8 direct neighbors in the 2D case, 26 in

[18]In the formulation of the symmetry-based DCIS detection, we have used a sparse representation of the co-occurrence matrix where the size did not depend on the maximum gray level, but the number of distinct gray level co-occurrences alone.

3D, and 78 in 4D, this requires 39 neighborhood directions to be considered for all inner voxels, i.e. those having neighbors in all directions.

A visual justification for the approach to treat the time dimension equal to the three spatial dimensions is given in Fig. 2.17 for three fictive template voxels representing enhancement curves of the malignant, suspicious, and benign type. It can be seen in the figure, that the curve types manifest in co-occurrence matrix entries in clearly separated areas of the matrix. In particular the off-diagonal elements corresponding to suspicious and malignant curve types are separated, as well as the elements close to or on the main diagonal of the benign and malignant types. Also note that this picture does not change regardless of whether symmetric or asymmetric co-occurrence matrices are calculated.

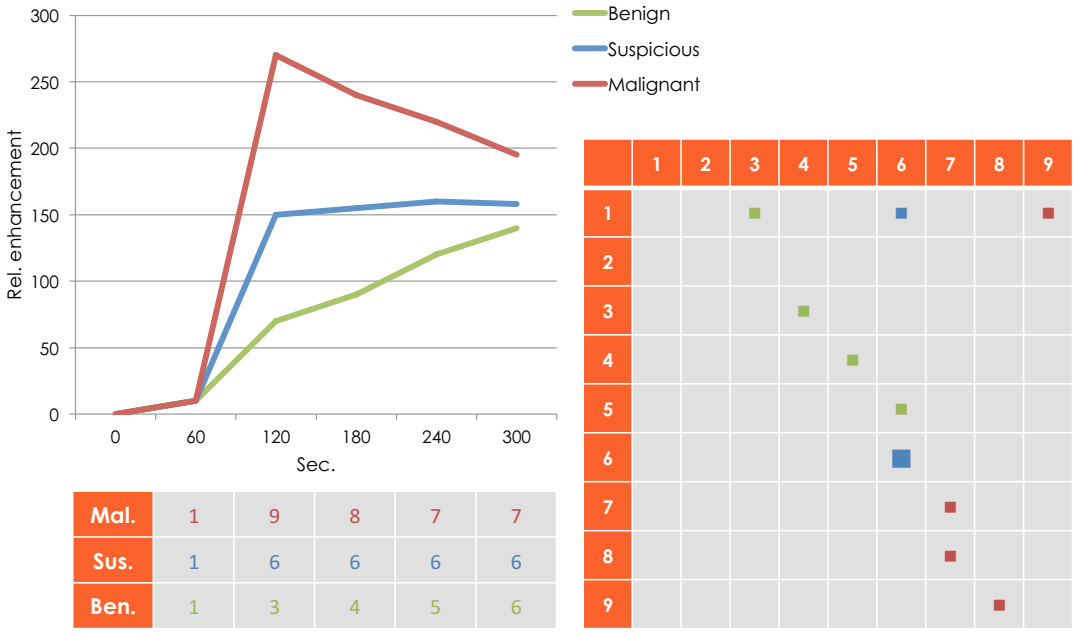


Figure 2.17: Thought experiment to justify the treatment of the time equal to the spatial dimensions. To the left, template curves for the three curve types are shown, together with a table of fictive curve values of all three. The table to the right shows the resulting co-occurrence matrix with entries colored according to the corresponding curve type, and square sizes corresponding to the number of occurrences of the gray value combination for the $(\Delta x, \Delta y) = (+1, 0)$ characteristic element.

This approach is complemented by experiments that extract the time course of contrast enhancement explicitly and calculate 3D texture features on it. On the full data set, comparisons between different approaches become possible, and insights into the salience of the different approaches.

Haralick Texture Features The goal of texture analysis is to obtain useful information by examining local variations in image brightness by quantifying the patterns of variations in image brightness within a region of interest (ROI) (Tourassi 1999). The spatial gray level dependence method proposed by Haralick et al. (1973) is frequently used because of its ability to capture the second order statistics of spatially or temporally collocated voxels. This information is contained in the co-occurrence matrix, as described above. The textural features are subsequently calculated from these co-occurrence matrices and describe the image in terms of higher-order statistics, e.g. homogeneity, linear structure, contrast, number and nature of boundaries present, or the complexity of the image.

If N is the number of distinct gray values; $C(i,j)$ is the (i,j)th entry of the co-occurrence matrix; $C_x(j) = \sum_j C(i,j)$ is the ith entry in the marginal-probability matrix obtained by summing the rows of $C(i,j)$.

$$C_{x-y}(n) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ |i-j|=n}}^{N} C(i,j), \text{ with } n=0,1, \ldots, (N-1)$$

$$C_{x+y}(n) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ |i+j|=n}}^{N} C(i,j), \text{ with } n = 2,3,\ldots,2\cdot N$$

Table 2.2: Haralick texture features.

Texture	Implementation
Angular Second Moment	$f_1 = \sum_i \sum_j C(i,j)^2$
Contrast	$f_2 = \sum_{n=0}^{N_g-1} n^2 C_{x-y}(n)$
Correlation[1]	$f_3 = \frac{\sum_i \sum_j (ij) C(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
Variance	$f_4 = \sum_i (i-\mu_x)^2 C_x(i)$
Inverse Difference Moment	$f_5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} C(i,j)$
Sum Average	$f_6 = \sum_{n=2}^{2N} n C_{x+y}(n)$
Sum Variance	$f_7 = \sum_{n=2}^{2N} (n-f_6)^2 C_{x+y}(n)$
Sum Entropy	$f_8 = -\sum_{n=2}^{2N} C_{x+y}(n) \log(C_{x+y}(n))$
Entropy	$f_9 = \sum_i \sum_j C(i,j) \log(C(i,j))$
Difference Variance[2]	$f_{10} = \sum_{n=0}^{N-1} (n-\mu_{x-y})^2 C_{x-y}(n)$
Difference Entropy	$f_{11} = -\sum_{n=0}^{N-1} C_{x-y}(n) \log(C_{x-y}(n))$
Information Measure of Correlation 1	$f_{12} = \frac{HXY-HXY1}{\max(HX,HY)}$
Information Measure of Correlation 2[3]	$f_{13} = (1-\exp(-2(HXY2-HXY)))^{\frac{1}{2}}$

[1] μ_x, μ_y, σ_x, and σ_y are the means and standard deviations of $C_x = \sum_j C(i,j)$ and $C_y = \sum_i C(i,j)$.
[2] μ_{x-y} is the mean of C_{x-y}.
[3] $HXY = -\sum_i \sum_j C(i,j) \log(C(i,j))$ where HX and HY are entropies of p_x and p_y and $HXY1 = -\sum_i \sum_j C(i,j) \log(C_x(i) C_y(j))$ and $HXY2 = -\sum_i \sum_j C_x(i) C_y(j) \log(C_x C_y)$, respectively.

Haralick texture features indirectly describe structures of the gray value image. The texture of the image will be reflected in the filling pattern of the co-occurrence matrix, which can be expressed by scalar-valued Haralick texture features. They can be distinguished from each other based on the structure of the corresponding co-occurrence matrix. Some of them can be explained with respect to the images via their representation in co-occurrence matrices.

Contrast, *Inverse Difference Moment*, and *Variance* describe how frequent co-occurrence matrix elements distant from the main diagonal are. The value for Contrast decreases and the value for Inverse Difference Moment increases when the entries of the co-occurrence matrix are located on the main diagonal, and vice versa. Variance describes the degree of variation of the gray values. Noting that high values on the main diagonal correspond to neighboring image voxels with very similar gray values, a dominant main diagonal corresponds to images with large homogeneous areas.

Entropy and *Angular Second Moment* describe the existence of dominant centers. Entropy is maximized and Angular Second Moment minimized when the gray values of the image are equally distributed, which will lead to a homogeneously filled co-occurrence matrix.

Correlation and the *Information Measures of Correlation 1* and *2* describe the occurrence of straight lines in the co-occurrence matrix, corresponding to image voxel neighbors that vary linearly across the image. Correlation is 1 for values on the main diagonal and -1 one for values on the main skew diagonal. Information Measure of Correlation 1 assesses the parallelism to the main diagonal, Information Measure of Correlation 2 increases with the presence of horizontal and vertical straight lines that indicate frequent voxels of one gray level next to many other gray levels.

In our work, the 13 of the 14 texture features defined by HARALICK et al. (1973) are calculated based on each of the images resulting from the preprocessing: the 4D co-occurence matrices derived from the images of baseline and early postcontrast phase, after both histogram operations (denoted $I_{0+2,eq/str}$ for histogram equalization/stretching), the the 4D co-occurence matrices derived from images of late phase and early phase (denoted $I_{2+5,eq/str}$), the 3D co-occurrence matrix directly derived from the subtraction image ($I_{Sub,eq/str}$), and the 3D co-occurrence matrices derived from all 11 descriptive parameters ($I_{Param,eq/str}$), in sum 375 features. The last feature is generally considered very unstable, which was also visible in our data, making us remove it from the data set.

Some texture features are independent of the quantity of contrast agent administered, since under monotonic gray value transforms, the co-occurrence matrix entries are only shifted, but not altered in their structure. Therefore, we follow other research proposing that Haralick texture features help to provide a viable tool for assessing lesion characteristics from MRI data in a differentiate, reproducible, and observer independent fashion. Texture analysis is applied to images that are preprocessed in different ways (W. CHEN et al. 2007; HARALICK et al. 1973).

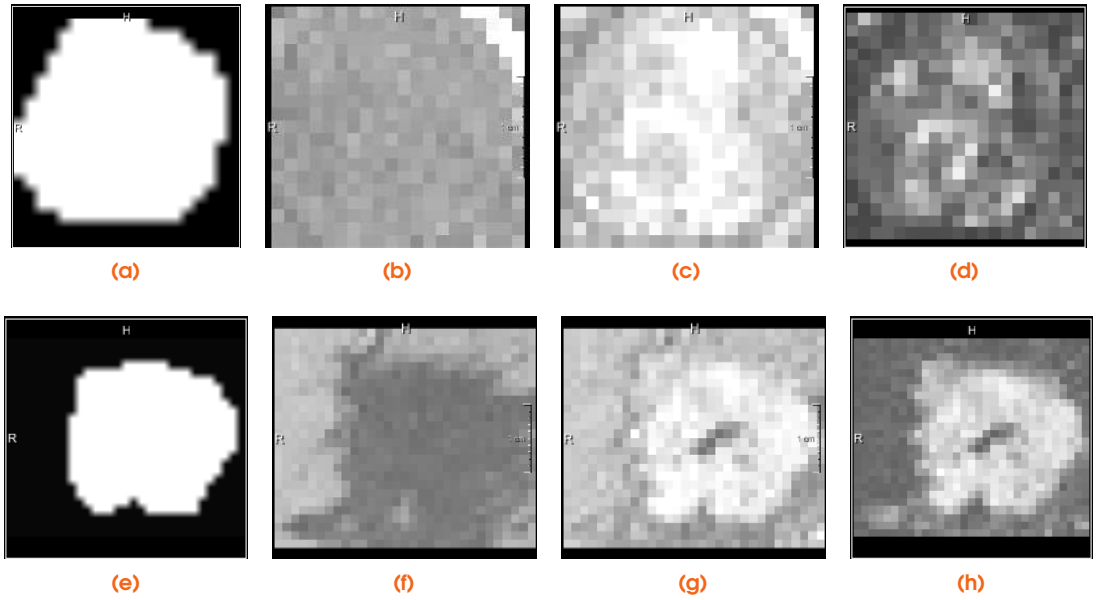


Figure 2.18: Example of benign and malignant lesion. Top row: fibroadenoma. (a) mask of the lesion; (b) baseline image at t =0 sec; (c) after t =158 sec; (d) subtraction image of the two time points. Images (e) to (h) show the same for a malignant lesion (invasive ductal carcinoma 30 mm of diameter).

Subtraction images A subtraction image is calculated by subtracting the pre-contrast image (reference image at $t = 0$ sec) from a post-contrast image (taken e.g. 158 sec after the reference image). Fig. 2.18d and Fig. 2.18h show a subtraction image of a benign and a malignant lesion from our data set. Two subtraction images are calculated for each lesion, one based on images manipulated with the histogram stretching method $I_{Sub,str}$ and the other one based on images manipulated with histogram equalization $I_{Sub,eq}$. 13 texture features are calculated based on each data set.

The dynamic component of DCE-MR images is only indirectly captured in textural features that are based on subtraction images. The loss of information introduced in this methodology is assessed by an alternative that is implemented to specifically capture the dynamic component of the DCE-MR images.

4D DCE-MR images Characteristics of the intensity time curve like wash-in and wash-out effect are strong indicators for malignancy and benignity. The gray values of the DCE-MR images reflect the enhancement of the contrast agent over time. The analysis of textural features based on 4D DCE-MR images tries to capitalize on these circumstances. The aim of textural analysis is to capture the spatial pattern in 2D or 3D. In this approach a temporal resolution is added and textural analysis is performed on 4D DCE MR images to capture the temporal distribution of the gray values and obtain additional information.

Degani et al. (1997) proposed to use one pre-contrast and two post-contrast time points (three-time-point method) to estimate wash-in and wash-out behavior, and in fact many decisive characteristics of contrast kinetic can already be calculated from these points if there is agreement on the absolute times of image acquisition and the overall acquisition protocol. We selected three time points out of the number of available time points according their absolute times of acquisition to increase robustness if acquisition parameters change (e.g. more time points are available, or their spacing changes). Co-occurrence matrices based on the three different time points are then calculated (I_{0+2} and I_{2+5}). The first is based on images acquired after 0 sec (pre-contrast image) and after 158 sec. This time point combination reflects peak enhancement. A second co-occurrence matrix captures the wash out phase based on images acquired after 158 sec and after 365 sec. Based on the two co-occurrence matrices two sets of texture features are calculated to describe the wash-in and the wash-out phase. Texture analysis is then applied to the obtained images.

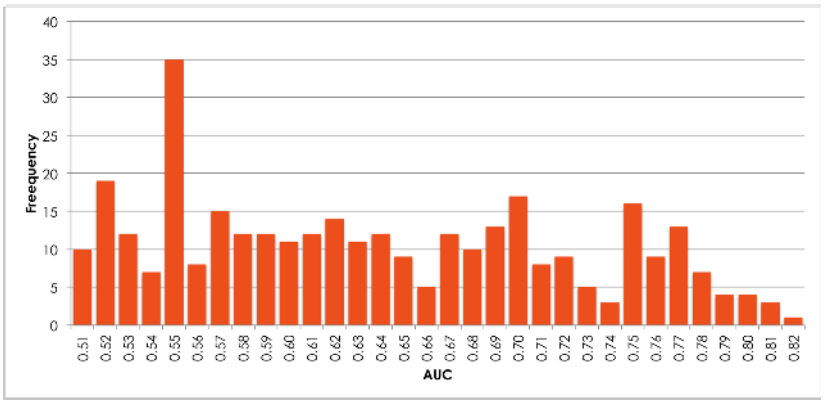
Descriptive Dynamic Parameter Maps Eleven dynamic parameter maps are calculated for each lesion, using the mean across the lesion voxels of each parameter map as an additional feature for classification. The parameter maps include the slope of wash-in and wash-out, the maximum enhancement overall and after the wash-in duration, the enhancement drop from peak to end of curve (wash-out), the area under the kinetic curve, the baseline enhancement, the time to the peak enhancement, to name the most useful ones. The classification rate of the dynamic features evaluated independently on our data using a logistic regression approach is promising (AUC $= 0.847$), but these features depend on many factors like tissue relaxation and the quantity of contrast agent administered (Behrens et al. 2007). Hence, we derive texture features from the 3D kinetic parameter maps. This is motivated by our expectation that texture features reduce the protocol dependencies, and are hence more suited to characterize contrast kinetics, provided they show similar discrimination power.

Assessment of Individual Features

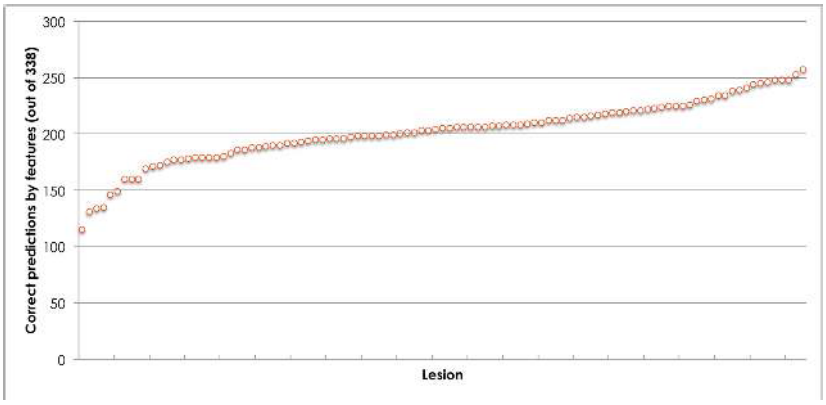
Before applying feature selection and classification methods to the data, we aim to gain a basic understanding of the performance level that can be expected from the features. To this end, nonparametric estimates of the AUCs of each feature were calculated, and while doing so, for each feature the classification results were recorded when thresholding it to the optimal value with respect to the prior class probability ratio.

From these recorded classifications, we yield a $n \times m$ matrix $\mathcal{F}$ with entries for all n features versus all m lesions, with $\mathcal{F}_{ij}$ indicating whether the ith feature is able to correctly predict the jth lesion.

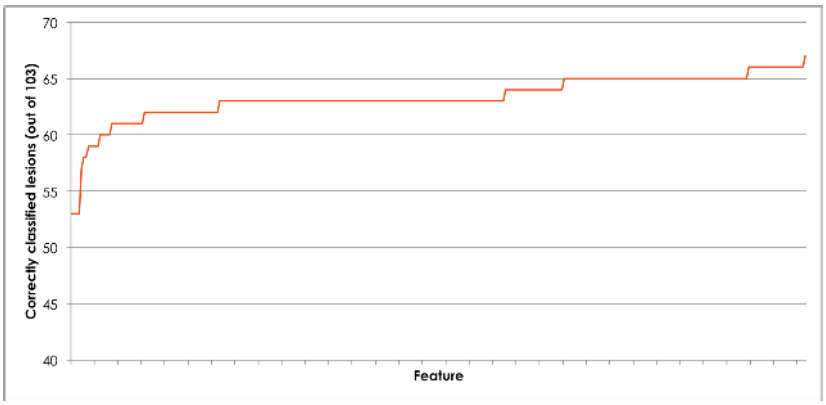
Fig. 2.19 shows summaries of the results gained from these experiments.



(a)



(b)



(c)

Figure 2.19: Results from predictions of all individual features on all individual lesions. ROC curves were obtained for all individual features to estimate their AUC towards the benign/malignant classification. (a) The histogram shows how many of the features obtain the AUC in the bins. Many features are barely above chance. (b) For each lesion, it has been recorded how many features classify it correctly, hinting at how simple it is to judge this lesion correctly. On the x axis, the lesions are sorted by increasing numbers of successful features. (c) The last plot shows for each of the 338 features (on the x axis), how many out of the total of 96 lesions it correctly predicts, indicating the discriminative power of the feature on the data set.

Linear Classification

To achieve optimal linear classification results that are still easily interpreted regarding the influence of each feature, we use step-wise logistic regression analysis with backward elimination (LRA-BE) to select the most salient features and train a linear regression model on them. Step-wise regression always includes a feature selection method. Essentially, regression model after regression model is being built, each with a changed set of features. Two central choices have to be made: the way how features are added or removed, and the criterion that is calculated to describe the model improvement. In our case, we employ backward elimination, which means the first model is built on the full set of features, and in each step the feature is removed that improves the model most. The salience of single features is expressed in terms of their Akaike Information Criterion (AIC) (Akaike 1974), which is defined as $a = 2k - \ln L$ with k the number of model parameters and L the model's maximum likelihood. To avoid the most common critique of the step-wise regression approach, we assess the model performance independent of the internal feature selection criterion. Instead, we wrap every model building step into leave-one-out (LOO) frameworks.

In a first step the performance of classification models for I_s, I_{4D}, and I^p_{param} are computed separately, with the aim to assess their individual predictive power. For I_s and I_{4D} LRA-BE was performed in a leave-one-out (LOO) framework using an initial variable pool of all 13 textural features. The resulting ROC curves and AUC values are shown in Fig. 2.20a to Fig. 2.20c. For the eleven parameter maps I^p_{param}, Tab. 2.3 shows the features included in the model, selected from the pool of 143 ($= 11 \cdot 13$) by their AUC value.

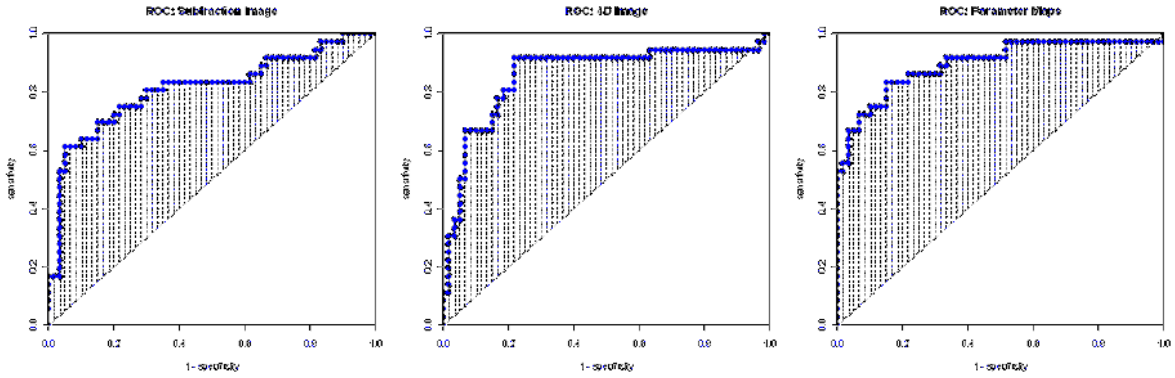


Figure 2.20: ROC curves based on the results of step-wise logistic regression in a LOO framework. From left: I_s: AUC $= 0.811$ — I_{4D}: AUC $= 0.858$ — I^p_{param}: AUC $= 0.896$

For each of the eleven dynamic parameter maps 13 texture features are computed, in sum 143 features per lesion. More features than samples are derived and some of these features are highly correlated, therefore it is not possible to apply step-wise LRA-BE to the whole feature set. While for the models so far, features have been selected either from small pools only, or a priori by their individual contribution, we will now turn to model building from all features at once. A new feature pre-selection procedure will be proposed to find an optimized model with respect to model complexity and predictive power.

The proposed pre-selection approach follows a fundamentally different line of thought. To avoid bias, not only features with an AUC above some manually set threshold should be used, but instead we assume that also features with smaller individual discrimination power are potentially helpful if their discrimination ability complements a feature of higher AUC in its blind spot, that is, in lesions where it fails. Obviously, many low-performance features with AUC values close to 0.5 on the full data set are available. These features may classify only a few

Table 2.3: Features from I^{p}_{param}; selected by AUC.

Dynamic Feature	Texture Feature	AUC
Slope 2	Sum Variance	0.843
Time to Peak	Information Measure of Correlation 2	0.832
Slope 2	Variance	0.827
Time to Peak	Angular Second Moment	0.825
Maximum Slope	Sum Average	0.824
Maximum Slope	Sum Entropy	0.822
Maximum Slope	Difference Entropy	0.820
Slope 2	Sum Entropy	0.816
Maximum Slope	Difference Variance	0.814
Maximum Slope	Contrast	0.814
Slope 1	Sum Entropy	0.812
Slope 1	Sum Average	0.811
Maximum Slope	Sum Variance	0.801
Slope 1	Difference Entropy	0.801

of the examples correctly that are misclassified by the high-AUC feature. It cannot be ruled out that it might be necessary to use both, high- and low-performance features, to reach an optimal classification result, leading to the requirement to develop a feature aggregation process that is transparent in all steps, and that lets us define the number of features desired for the final regression model. The step-wise regression algorithm with backward elimination is in principle able to find both features and reach high classification accuracy, but it is not possible to apply it to the whole feature set at once since it is extremely prone to overfitting, particularly when presented with a data set having many more features than examples.

Given a pool of candidate features to assess for their helpfulness in a model, our basic approach is to build many models, each using LRA-BE, since we aim to exploit the feature selection during model building that is inherent to LRA-BE. Instead of doing this once and on the full feature pool (which again potentially yields spurious results through overfitting), we do it in a leave-one-out fashion on subsets of the data, each time recording the features of the model and the model performance. The features to keep are this time selected by their significance. This procedure gives each feature equal chances to be chosen for the final regression model that is built on the melted-down feature set. The principal outline of the process is depicted in Fig. 2.21.

We expect better performance from a model based on the texture features of all input images (I_s, I_{4D}, and I^{p}_{param}). To combine them, we perform a variable selection process in multiple steps, stacking LRA-BE runs in LOO frameworks that are used to rank features by appearance in models built during LOO. After each LRA-BE run, only highly significant features ($p < 0.05$) are retained. This step is meant to reduce overfitting of LRA-BE that we observed. For the final aggregation step, only eight features from the last LOO process form the initial variable pool to train the final regression model. These eight features were selected as before; in particular, they were used in models of the final aggregation step more than 70 (out of 96) times, such that we consider them highly significant for the classification task.

Two of the features (Stretching.Integral.DifferenceEntropy and Stretching.Slope1.Entropy) were not selected in the individual selection procedures. They have individual AUCs of .60 and .64, respectively, hence perform poorly alone. Still, their relevance in the final model is confirmed by omitting them: if e.g Stretching.Slope1.Entropy is omitted, the AUC of the final model reduces to 0.824 (95% CI = [0.72, 0.90]). These features exemplify the problem specified above.

The final regression model is computed with a logistic regression based on the selected features from Tab. 2.4. Fig. 2.23 shows the ROC curve (AUC=0.907) of the final model. With the described procedure we achieve a positive predictive value (PPV) of 85% which is comparable

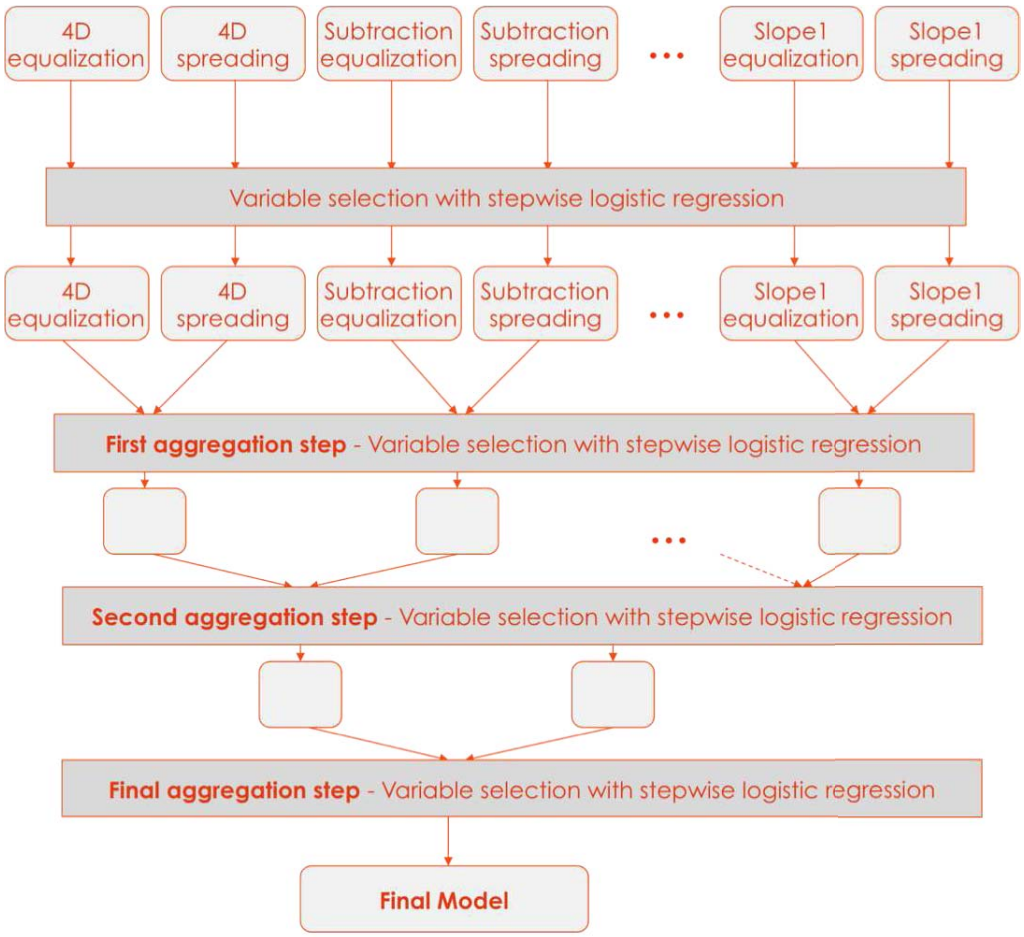


Figure 2.21: Feature selection procedure.

to Shannon C Agner et al. (2011), who reported PPV values between 72% and 86% for the features they applied in a similar approach, but on a different and smaller data set (compare Sec. 2.4.1).

The regression model that is finally obtained in our feature selection and training procedure calculates a score Γ from the features extracted for a given lesion according to the following equation:

$$\Gamma = -93.7 + [\alpha_1 \ \cdots \ \alpha_8]^{\mathsf{T}} \cdot [242.7 \ 73.0 \ 18.4 \ -18.2 \ 12.9 \ -11.1 \ 7.2 \ 5.1]. \tag{2.18}$$

In this linear model, the features are

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \\ \alpha_8 \end{bmatrix} = \begin{bmatrix} \text{Stretching.Slope2.AngularSecondMoment} \\ \text{Equalization.TTP.InverseDifferenceMoment} \\ \text{Equalization.TTP.Contrast} \\ \text{Equalization.TTP.DifferenceVariance} \\ \text{Equalization.TTP.DifferenceEntropy} \\ \text{Stretching.Integral.DifferenceEntropy} \\ \text{Stretching.Slope2.SumEntropy} \\ \text{Stretching.Slope1.Entropy} \end{bmatrix}$$

Table 2.4: Overview of the initial variable pool for the combined model. Features chosen after first variable selection step are in bold face.

Dyn. Feature/Image Type	Texture Feature	AUC
Slope 2	Sum Variance	0.843
Time to Peak	Info. Measure of Correlation 2	0.832
Slope 2	**Variance**	**0.827**
Time to Peak	**Angular Second Moment**	**0.825**
Maximum Slope	Sum Average	0.824
Maximum Slope	Sum Entropy	0.822
Subtraction	Sum Variance	0.822
Subtraction	**Sum Entropy**	**0.820**
Maximum Slope	**Difference Entropy**	**0.820**
Subtraction	Difference Entropy	0.818
Slope 2	Sum Entropy	0.816
Subtraction	Inverse Difference Moment	0.816
Subtraction	**Contrast**	**0.816**
Maximum Slope	Difference Variance	0.814
Maximum Slope	Contrast	0.814
Subtraction	**Entropy**	**0.814**
Subtraction	Variance	0.813
Slope 1	Sum Entropy	0.812
Slope 1	Sum Average	0.811
4D	**Contrast**	**0.797**
4D	**Difference Variance**	**0.797**
4D	Difference Entropy	0.794

Table 2.5: Confusion matrix of final model at a decision threshold of .56, which has been chosen in the above Fig. 2.22. **Rows**: true class. **Columns**: predicted class. A sensitivity of 95% can be reached at a specificity of 87%, with positive and negative predictive values of 92% each.

	benign	malignant
benign	33	5
malignant	3	59

Fig. 2.22 illustrates the separation ability of the final model. Lesions are scored 0 to 1, with 0 being indicative for benign, 1 for malignant lesions. Note, that the threshold level needs not be 0.5; instead, a threshold of 0.56 yields a slightly better result on the data set by classifying two additional benign findings correctly. This is reflected in the confusion matrix in Tab. 2.5.

To our surprise, the 4D features were largely outperformed by texture features that are computed based on either subtraction images or on kinetic parameter maps. With the feature selection process we have described above, 4D texture features are not contained in the final model. There are different possible explanations. Firstly, subtraction images and parameters maps also capture information on the time course of contrast agent delivery — perhaps in a more robust fashion than 4D texture features. Variability tests, where simulated data is systematically deteriorated, might reveal such influences. Moreover, it might be detremential to condense all information that is contained in the 4D structure into a single scalar value. This, however, would in our expectation affect conventional kinetic features at least as much as our integrative 4D texture approach that was expected to capture the complex information more holistically. Finally, it is possible that the data set we used in our analysis together with the lesion segmentation masks is more suited to the direct classification using kinetic features. Supporting this hypothesis is the fact that by visual inspection, the lesion masks tend to approach the lesion perimeter from the inside, yielding potentially more homogeneous tissues. Textural features, in contrary, might be capable to include information from the vincinity of the lesion. This, however, remains unexplored as of today.

In the light of the promising performance estimate on our relatively small data set, this leaves many further research opportunities that might result in a valuable addition to a CAD system. When the described system has been implemented in 2008/2009, it has been left for

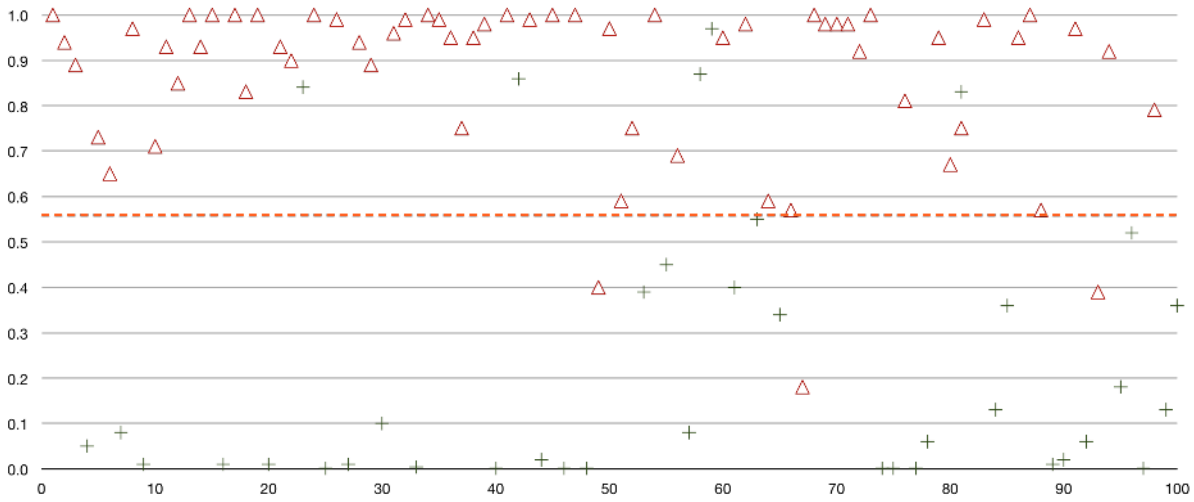


Figure 2.22: Classification results. Plus marks indicate benign, triangles malignant lesions. The score predicted by the linear model is on the vertical axis. The stippled horizontal line indicates one possible classification threshold (also compare Tab. 2.5).

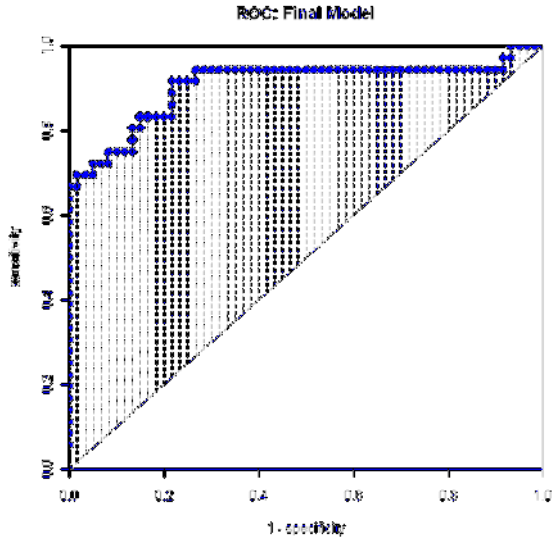


Figure 2.23: ROC curve of the final model with a corresponding AUC value of .91.

future work to look at non-linear classifiers, at other feature selection approaches and their correct application, and the potential positive bias introduced in the chosen approach. Those are aspects that make it interesting to look at the data set again, this time from the perspective of best practice methods we have described above. It will be of particular concern to quantify the magnitude of bias that might have been introduced by the combined feature selection/model learning procedure described above.

Linear Classification: Results and a Critical Perspective

In the work described so far, we started out with the aim to learn and understand the importance and influence of several options: how to include the time information in texture features, how to select and rate the best performing features, and how to build a linear model that shows feature dependencies in an unobstructed way. This decision brought about that the features have only been examined with respect to their individual yield or their contribution to a linear model alone. However, features neglected in the linear model (due to their individual poor performance) may still be valuable in context of other features. An example for a situation

where only a non-linear classification model is able to distinguish between two classes has been given in Fig. 2.2 on page 47.

To explore this more deeply, we will turn to non-linear classification models in the next section, after summarizing the lessons learned from the research so far and looking at more recent developments. With many lessons learned, looking back at the procedures that have been composed into the feature aggregation process, some remarks are in order.

Developments in 4D Texture Our approach to capture contrast agent wash-in and wash-out characteristics in texture has been comparatively successful. There were few comparable publications at the time of writing, and until today, no other groups have shown more convincing approaches to the problem, the work of Shannon C Agner et al. (2011) being the notable exception. They, however, estimate voxel-wise features, and follow them through time. Our classification method that is based on the preprocessed data in contrast incorporates all available information. In particular, neither of textural parameters or descriptive parameters alone were comparable to the result obtained with a model integrating both. Also, our results indicate that the informational content of time curves describing the CM distribution (wash-in and wash-out) alone is inferior to that of both textural parameters and textures of descriptive parameters. Our results show a performance that is in the range of reported accuracies of human observers as reported on different data sets in literature (sensitivities 79–94% (Heywang-Köbrunner et al. 2008), specificities 20–100% (Behrens et al. 2007); our scheme: 97% sensitivity at 70% specificity), and surpass results obtained on the same data before (W. Chen et al. 2007). Haralick features as calculated in our work in addition provide limited invariance to rotation since not only one direction is captured, but the entire neighborhood.

From the results we have obtained, it can summarizingly be concluded that texture metrics add a viable set of features to any classification approach for mass lesions seen in breast MRI. It needs to be further examined what the best way to exploit them is, however.

We proposed a classification scheme using textural features derived from both spatial and temporal neighborhood information in DCE breast MRI images. A central aim was to find features of breast MRI mass lesions that are robust or invariant to changes of MR acquisition protocols, contrast agent concentration, and other imaging parameters. Changes for example in the quantity of contrast agent administered, the number of post-contrast images, the temporal resolution and field strength, and certainly others more, all may require models to be retrained. Consequently, CADx of breast MRI lesions is still considered a challenge (Behrens et al. 2007). In fact, we assessed the model on an independent test set from another site with completely different acquisition parameters. Since the data set was heavily unbalanced, the results were not reported, because we considered them insignificant. With very few benign lesions, there was no way to assess discrimination power of the model on this data.

One particular characteristic, again in contrast with S C Agner et al. (2008) and Shannon C Agner et al. (2011), is that our texture features each are scalars for the entire 4D ROI, while Shannon C Agner et al. compute voxel-wise features for sliding windows of different sizes. This enables them to map and colorize the features for visualization, a feature we lack; we score the lesion as a whole, assuming that texture is a trait that is inherently connected to larger regions rather than small windows, and that therefore any texture feature benefits from information.

Feature Selection We didn't intend to develop a novel method to select features from a large feature pool. In fact, much of this effort was introduced in iterations, when we experienced instabilities that were caused by overfitting of the simple model. Hence, we addressed this problem by reducing the feature pool of each run of LRA-BE to a feasible size. This quickly led to the hierarchical layout of model generating layers that has been displayed in Fig. 2.21. The main obvious flaw is that there is no way the process may find a useful combination of

features with mediocre individual performance to be combined beneficially if they are in different branches of the hierarchy. A less obvious problem is hidden in the feature subset selection process itself, that despite the design in a leave-one-out cross-validation wrapper per layer, the entire procedure is not wrapped itself, essentially meaning that the source of positive bias discussed above is present in the presented learning procedure.

As we have learned, machine learning methods are well capable to deal with large feature spaces, and the "curse of dimensionality" can be dealt with by new developments in feature selection methods. Also, particularly a Naïve Bayes classifier is able to give insights into the importance of features and their positive or negative influence in the prediction model, robust against marginally important features and little effected by large feature spaces.

This made us propose to evaluate the texture features in nonlinear predictors, which was out of scope in our publication (Loose et al. 2009), but will be described in the following.

Nonlinear Classification

We have seen from the analysis of the features that some show nonlinear relationships that cannot be modeled by logistic regression. In addition, we note that the feature selection process is not beyond doubt in terms of a potentially positive bias of the learned model, since for feature selection and classification, the same basic mechanism is used — the predictive power of the attribute towards the class, assessed in a linear model like the final regressor.

To better validate the linear model built in the preceding section, a sound methodology employing classification on a feature-reduced data set needs to be developed, where feature subset selection is applied in an unbiased fashion. Hence, in a further set of experiments described in this section two aspects will be explored:

Feature subset selection bias We quantify the influence of feature selection when wrongly done with the knowledge of the full data set before splitting into training and test sets in a cross-validation scheme, using an artificial data set consisting of 300 features drawn from a uniform distribution and varied numbers of random examples.

Nonlinear classification We assess the potential for improvements using nonlinear classification methods, both with and without preceding feature selection, implemented in a cross-validation scheme that holds out a fraction of the data set for testing and develops the model (potentially including feature selection) with the dedicated training set. Evaluation is done on the hold-out sample. For comparison, a complete training with and without feature subset selection is done on the full data set.

Since feature selection effectively means that information is being thrown away, it is expected that the performance of the more forgiving classification methods deteriorates. Implementations for these crucial experiments are given in Appendix Sec. C.1.

General Setup We compare the performance of two classifiers. Each is parameterized with standard parameters as derived from the literature. We use a Random Forest with 100 trees and a Naïve Bayes classifier.

The Random Forest was parameterized to stop when nodes contain less than five examples. The number of features randomly selected for each split was determined automatically based on the rule of Breiman (2001) to be the square root of the available features. The maximum depth was not restricted.

During the experiments, data sets will be derived from the basis data set with varying numbers of top-scoring features. Throughout the experiments, a modern feature scoring method is used, based on the Multivariate Adaptive Regression Splines, MARS[19] (Friedman 1991).

[19]Free implementations, for example for R and python, are usually called EARTH, which is no acronym.

Multivariate Adaptive Regression Splines (MARS) is a non-parametric regression method that for the purpose of feature selection is used in a bagging setup, building many smaller models. From the bagged MARS regressors, a per-feature average contribution to the models is calculated, yielding a score for each feature. Note that MARS can also be applied directly to model the data in a regressor, but only on data sets that have fewer features than our case. It was originally designed to be used on data sets with 20 or fewer features.

Naïve Training In the first experiment, the often criticized practice of feature subset selection and classifier training on the same data set is executed to establish a base performance measure. Note, however, that we do not emulate the malpractice of optimizing the combination of feature subset selection and classification. Instead, we perform feature selection on the full data set and forward the reduced data set into the classifier training. In particular, we use a feature selection approach that is not included in the classifier itself, like for example a Random Forest based feature selector. For this reason, we expect only limited bias. A much greater bias would have to be predicted if the used features are estimated in an exhaustive search over possible feature subsets, or if the feature selection process is iteratively changed to optimize final classification performance.

Table 2.6: Results of MARS-based importance ranking. Four features of this list have been selected by our previously introduced feature aggregation algorithm in the final aggregation step, out of which two have been used in the final model.

Dyn. Feature/Image Type	Texture Feature	MARS importance
MTT	Sum Average	2.0
Maximum Slope	Entropy	1.7
4D	Difference Variance	0.8
Maximum Slope	Sum Entropy	0.8†
MTT	Inverse Difference Moment	0.8
CInt	Sum Entropy	0.7
MITR	Variance	0.7
4D	Correlation	0.6
Subtraction	Sum Entropy	0.6§
MITR	Difference Variance	0.5
Peak 2	Difference Variance	0.5
4D	Difference Entropy	0.5†
Maximum Slope	Inverse Difference Moment	0.4
Maximum Slope	Difference Entropy	0.4§
MITR	Correlation	0.4
MTT	Angular Second Moment	0.4
Peak 2	Angular Second Moment	0.4
Peak	Sum Average	0.3
Slope 1	Entropy	0.3
Slope 2	Difference Entropy	0.3
Peak 2	Information Measure of Correlation 2	0.3

† also selected into 22 features of final aggregation step
§ also selected into final 8 features of aggregation model

Comparing the features selected by MARS importance (Tab. 2.6) with the list of 22 features (and the eight features selected into the final aggregation model; Tab. 2.4) shows little detailed overlap. However, again the dominance of texture derived from descriptive dynamic parameter maps is obvious. Except three texture features directly derived from the full 4D image and one derived from the subtraction images, all others are derived from descriptive parameter maps. It is noteworthy, though, that other than in the feature selection process for the linear classification model, several 4D texture features have been included in the selection of most important features.

We have used not only the index lesions this time, but extended the data set with seven additional lesions found in the patients, resulting in 103 lesions total (39 benign and 64 malignant).To evaluate the predictors exhaustively, we performed multiple runs of evaluations with changed settings. Thus, we used once 50 and once 100 iterations of training and testing.

For the 50 iterations, 70% randomly sampled examples from the full set were drawn for training, keeping the remaining 30% for testing. This results in 1.550 classified examples. In the second run using 100 iterations, the split was 80%/20%, resulting in 2.100 classified examples. Since in our data, the proportion of malignant to benign cases is not extremely biased (approximately twice as much are malignant), we don't sample balanced training and test sets. We report classification accuracy, here defined only as the proportion of correctly classified examples, sensitivity and specificity, the area under the ROC curve (AUC), and the Brier score.

The following Tab. 2.7 summarizes the results of classification with the two methods capable to model nonlinear relationships between features. The comparison includes the use of the full feature set and a reduced feature set comprising the 20 most important features according to the MARS-based feature selection. In the table, sensitivity and specificity are reported with respect to the prediction of the malignant class.

Table 2.7: Mass classification performance with different classifiers. 100-fold random sampling of 80% training, 20% testing instances per classifier. Numbers in brackets indicate performance on full feature set; other numbers on data with 20 features preselected according to MARS-based importance ranking. (NB: Naïve Bayes. RF: Random Forest).

	CA	Sens	Spec	AUC	Brier
NB	(.798) .845	(.847) .890	(.719) .775	(.819) .904	(.402) .271
RF	(.799) .841	(.867) .910	(.688) .733	(.876) .916	(.284) .243

While a substantial performance gain of the Random Forest classifier is noted on the reduced feature space, it is tough to evaluate the Random Forest for its bias and variance, or overfitting, explicitly. Instead, the training process of Random Forests implicitly selects most salient features.

Validation Experiment with Random Data We have conducted an experiment on random data with a ratio of features to examples matching the given classification problem: 300 features in 100 examples were drawn at random from the uniform distribution. The two class labels were assigned to the examples in equal frequency, yielding 50 examples per class. While this deviates from the proportion in the 4D texture data set, the classification problem is not prone to bias due to a majority class preference in the classifier because a classifier cannot improve its performance by simply selecting the dominant class more frequently.

This data was then classified in the procedure outlined before, i.e. using feature selection and subsequent classification. Although the data set contained no information, classifiers were obtained that predicted the examples with consistently above chance accuracy, achieving AUC values of about 0.70 to 0.80, which is unexpected in random data. The explanation lies in the feature subset selection process. The higher-dimensional the feature space, the more likely it is that features can be selected that divide the sparsely distributed examples with some success. Note that it is due to chance alone if one of the 300 random features conforms with the true (random) trend in the data — it is however the implicit permit included in the feature selection step to remove the contradictory features from the model, which would harm classifier performance.

To gain more insight into the sensitivity of the feature subset selection process depending on the number of features selected, and the number of training examples available in the data set, we assess the variation of the classifier performance measured by its AUC with the number of training examples and number of selected features. The number of training examples was varied from 20 to 1,000, and the number of selected features from 10 to all 300. With more examples, it is expected that the AUC approximates 0.5, reflecting a chance decision. Fig. 2.24 shows a plot of the results for the Random Forest classifier. Indeed, both with a decreasing number of selected features from the total, and with a decreasing number of available examples for training, the classification performance reaches maximal values, while oppositely both high numbers of retained features make it impossible for the learning mechanism to find patterns in the data, and also high numbers of examples better occupy the feature space, such that the unpredictable

nature of the random data prevents the classifiers from separating it. For comparison with the results obtained using a Random Forest classifier, Appendix C.1 shows the results of the same experiment for a Naïve Bayes classifier.

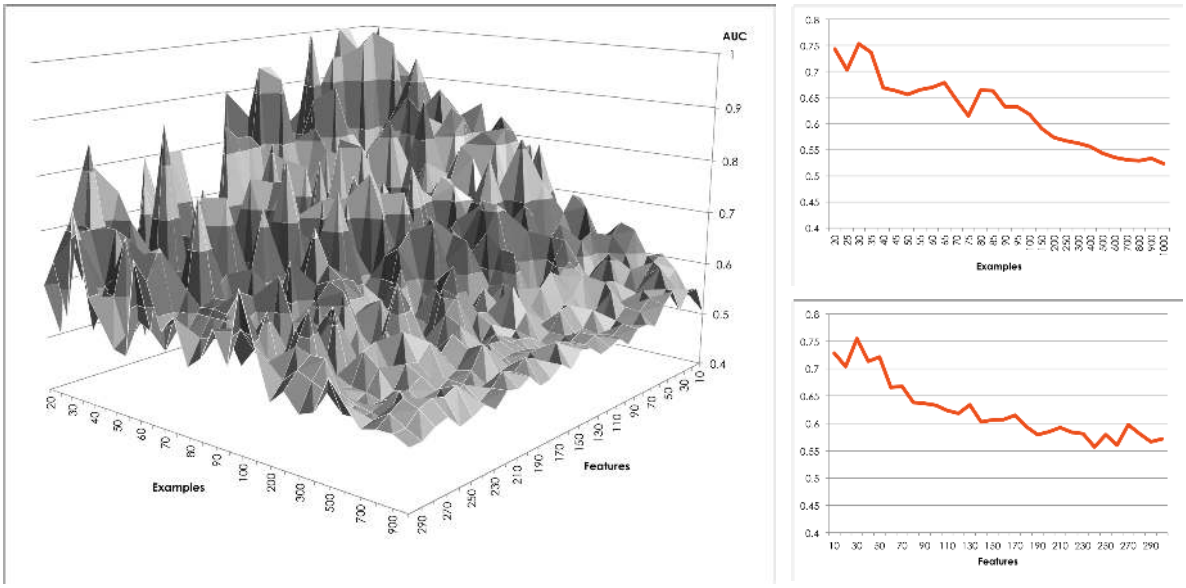


Figure 2.24: Illustration of Random Forest bias on random data. Left: Variation of AUC with numbers of examples and selected features. Right top: AUC of Random Forest averaged over all numbers of examples. Right bottom: AUC of Random Forest averaged over all numbers of selected features.

Validation Experiment on 4D Texture Data Set In the last experiment, knowledge on the expected performance of a classifier trained with an unbiased training scheme shall be gathered. A training procedure including a feature subset selection process was set up to obtain an optimized classifier on a reduced set of features. In this scheme, a cross validation wrapper encapsulates the sequence of feature selection and cross-validated training and evaluation on the reduced data. Of course both the validation wrapper and the internal training validation may be selected individually according to the data, to be either of a multiple random splitting of the data (used in this experiment in the wrapper), n-fold cross-validation (10-fold CV was used in this experiment), or leave-one-out cross-validation.

In this experiment, the resulting classifier performances of 50 runs of the following procedure were averaged for different numbers of selected features n_f:

1. Randomly select 75% of all examples into the training set, retain the remainder for evaluation.
2. Select n_f top-scoring features based on the training set, and filter the evaluation set to retain only the same n_f features.
3. Do a 10-fold cross-validated training/evaluation procedure on the training set, and keep all trained classifiers.
4. Predict all examples of the evaluation set using all kept classifiers, and average their results. This is the classifier performance of one run.

All 50 evaluation results for a Random Forest and a Naïve Bayes classifier have been collected according to the above training scheme to generate the statistical evaluation summarized in Fig. 2.25. The left boxplot reports the average, median, and maximum/minimum performances reached by the classifiers when selecting 20 top-ranking features. The range of AUCs of the trained classifiers on the test set is anywhere between 0.5 and 1.0, emphasizing the fragile and unpredictable nature of classifiers trained on problems with too many features, or too little data.

Averaged classifier performances across all trained classifiers for each number of selected features between 10 and 300 are shown in the plot to the right of the same figure. Here, it is noticed that feature selection does not benefit Random Forests (which perform an inbuilt feature selection)—in the contrary, it may prevent the Random Forest from optimizing its performance if it is presented too few features. Oppositely, the Naïve Bayes classifier requires some care when defining and selecting features to meet the criteria of feature independence and match its probabilistic framework to the type of feature distribution.

We are also able to quantify the bias that is brought about if the feature selection process is not wrapped in a cross-validation scheme from the comparison of classifier performances on training and testing sets. In Fig. 2.25, the difference of performance estimates on training and test set is clearly visible, as is the remaining difference even on the full feature set. Using either classifier, an AUC overestimation of about 15% is noted in the best case. It is interesting to note how the performance of the Naïve Bayes classifier drops after a peak at about 40 selected features, evaluated on the test set, while the Random Forest improves with available features. Again, the explanation might be the feature dependence. In the same Figure, note that the box plot corresponds to the statistics of only the one point on the "Features" axis of Fig. 2.25b at value 20.

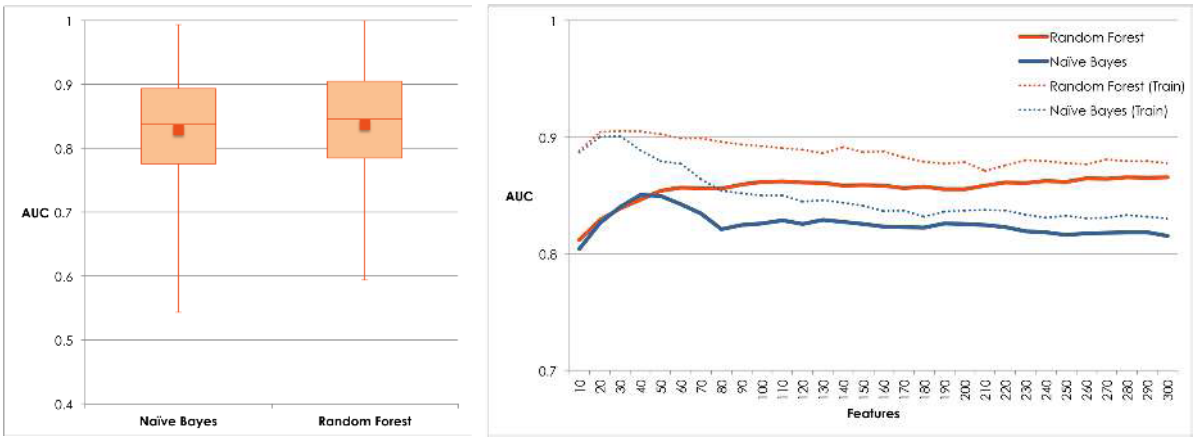


Figure 2.25: Unbiased classifier performance for 4D texture CAD. (a) Random Forest and Naïve Bayes classifiers show similar mean performance and variation in 100 runs of feature selection and training on 75% of the data, and testing on the hold out sample. 20 top-ranking features were selected. (b) The number of selected features differently affects the performance of Random Forest and Naïve Bayes. For comparison, the performance estimates of both classifiers on the training set are shown in stippled lines, while the estimates obtained on the held-back samples is shown in solid lines.

Nonlinear Classification: Lessons learned

With the feature selection and classification method described in Fig. 2.25, we have obtained average classification performance estimates with AUCs of about .86 for Random Forests. In a slightly different setup, feature selection, classifier training and classifier evaluation have been encapsulated in leave-one-out wrappers (instead of a 75%-25% split), so that it was likewise ensured that an independent test case has been retained before selecting the features. Then, the classifier performance for both the Random Forest and the Naïve Bayes exceed AUC values of 0.90.

We have continued to show that such a setup yields no predictions better than chance on a random data set of comparable size. We went on to quantify the bias when neglecting the proper learning/evaluation setup, and showed how a naïvely integrated feature subset selection and training method might yield a classifier with almost perfect class separation on purely random data.

The classifier performance estiamte on the mass lesion data set varies with the number of features selected. It stabilizes on a higher level for the Random Forest classifier than for the Naïve Bayes classifier, most likely because the Random Forest is capable to model the non-linear relationships in the features while the Naïve Bayes is not.

A consequence from these findings needs to be that when facing signs of the "curse of dimensionality", or if overfitting is suspected, experiments like the one conducted above should be undertaken to estimate the reliability of performance predictions of the trained classifier, alongside with employing a sound training procedure that does not add bias.

2.4.3 Comparison and Perspectives

A comparison of the results reported for the proposed linear regression model obtained with our proposed LRA-BE aggregation method, and classification with two robust classifiers suggests that the proposed scheme may be competitive. There are, however, indicators for a positive bias in the predicted performance based on the experiments performed in the section above.

All training and validation has been performed on one comparatively small data set of 103 lesions. Not all relevant shapes and morphological patterns of benign and malignant breast cancers can possibly be covered in such a small sample, even if limiting the study to lesions presenting as masses on MRI. Hence, a first extension of the study would be to take more examples of benign and malignant mass lesions into the data set. Here, data from different sites should be considered to assess the robustness of the features against varying acquisition parameters. Second, a thorough validation on an independent testing data set needs to be conducted to validate the generalization ability of all predictors.

The prime contribution of our work is therefore to provide a detailed insight into the topic of texture features calculated from differently preprocessed contrast-enhanced breast MRI, stressing the importance of descriptive dynamic parameter maps another time. We were able to train predictors with AUCs matching human observers. The features are easy to compute, since no modeling is involved.

A second key insight has been provided into the sensible topic of feature subset selection methodology. When a naive implementation of feature subset selection and subsequent training and evaluation on the same data over-estimates the achievable classification performance by as much as 15%, it has been learned that great care has to be taken in the process. As it finally turns out, the feature subset selection process designed to prevent optimistic bias yielded a Random Forest classifier with a performance similar to that on the full data set including all features.

Looking into the future, texture features like the ones computed in our contribution may be replaced by novel developments like local binary patterns (LBP) or local ternary patterns (LTP), though while such features excel in the characterization of patterns in large structured images (Liao 2010), their contribution to the classification of small areas has yet to be established. The only breast cancer related applications of LBP and LTP have been reported by Choi et al. (2012) and Nanni et al. (2012), using publicly available databases of digitized film screen and digital mammograms. In mammography, due to its higher spatial resolution, lesions occupy more voxels, which is probably favorable for LBP/LTP analysis. In breast MRI images, with a resolution orders of magnitude below that of mammography, no reported results are accessible.

2.5 Remarks on the Future of CADx in Breast MRI

The chapter described approaches to the detection and diagnosis of different types of breast lesions depicted by breast MRI — mass enhancements and non-mass enhancing lesions. Based on considerations about best practices regarding feature extraction and machine learning, lesions of these types have been treated with fundamentally different techniques. The performance for the classification of mass lesions is comparable to results achieved on the same data set, and acceptable both from a theoretic and clinical standpoint, but only effective and beneficial in a workflow based breast MRI reading scenario if combined with a lesion candidate detection scheme. Considering that the goal of this part of research was not to build a complete CAD system, but to explore a novel category of texture features derived from the time-resolved information of DCE-MRI, the contributions presented will potentially enhance existing CAD systems by increasing their specificity, and thereby provide a contribution in an integrated system.

The non-mass-like lesion detection and diagnosis method followed a completely different approach, not only in terms of extracted features, but also regarding the fused detection and diagnosis setup. The obtained classifier capabilities, in particular for the symmetry-related features, exceeds the performance of all reported systems so far, and competes with the diagnostic performance reported for human observers. A formal direct comparison is to be conducted on an independent and larger data set to substantiate such claims on a broader basis.

It is tempting to combine the two lesion classification systems. By design, the non-mass lesion CAD system is not specific for non-mass lesions in the detection stage, while both the symmetry features generated from the detection step, and the morphological features derived from the lesion candidates have been designed with this particular task in mind. Hence, candidate regions may as well be submitted to more than one specialized predictor, for example the texture feature based mass classification system and the non-mass lesion CADx.

With algorithms of this quality level at hand (many more are reported in the literature), the translation into clinically useful tools needs to be the prime objective of future efforts. There are challenges, however. Several potential impediments need to be considered:

- Market access regulations
- Radiologist acceptance
- Patient acceptance

Looking at the current clinical situation, machine learning based diagnostic systems for breast MRI are only available in disguise, for example termed “quantitative imaging biomarkers”, but most prominently in the form of colored overlays for the anatomic images. Such color maps translate diagnostically interesting characteristics of the contrast agent uptake curve into colors that are usually chosen to convey the machine-predicted malignancy likelihood to the human observer.

Market Access Regulations A reason for this covert result presentation is the strictly regulated market access for medical devices (which CADe/CADx is a part of) as well in the European Union as particularly in the United States, which is currently the largest market for medical devices worldwide. In Europe, any medical device (which includes computer systems and software) needs the CE[20] mark before it may be marketed. In the United States, the market clearance for class II medical devices (including diagnostic tests) is obtained in the so-called premarket notification (often also simply called 510(k)) process. This process requires the device to be similar to an already approved device. If this is not the case, a premarket approval process has to be initiated. The U.S. Food and Drug Administration (FDA) regulates CADe software,

[20] Conformité Européenne

but explicitly excludes CADx, suggesting to negotiate directly with the FDA prior to making any attempts to apply for clearance or approval (U.S. Food and Drug Administration 2012). The distinction between CADe and CADx is not made in Europe, suggesting easier market access for CADx algorithms.

Radiologist Acceptance CADx is well suited to reduce the complexity in many diagnostic decision making tasks. To this purpose, the user (the radiologist) needs to be convinced that a benefit for the patient will result from the usage of CADe/CADx — provided once the market clearance is possible. CADe usually produces more prompts than there are positive lesions, the false positive findings. Reviewing all of them is not deemed feasible by many radiologists. Studies have shown for mammography screening, that even in the presence of CADe marks, many marks that correctly indicate positive findings are ignored: in one study, as much as 71% of correct CADe marks were not regarded by the eight readers (Nishikawa et al. 2012). Several ways to overcome these problems can be discussed:

Interactive CAD does not show all CADe prompts at once, but only upon request of the radiologist, who can click any area of interest in the images, and bring up the computer-provided assessment of the area (Hupse et al. 2013).

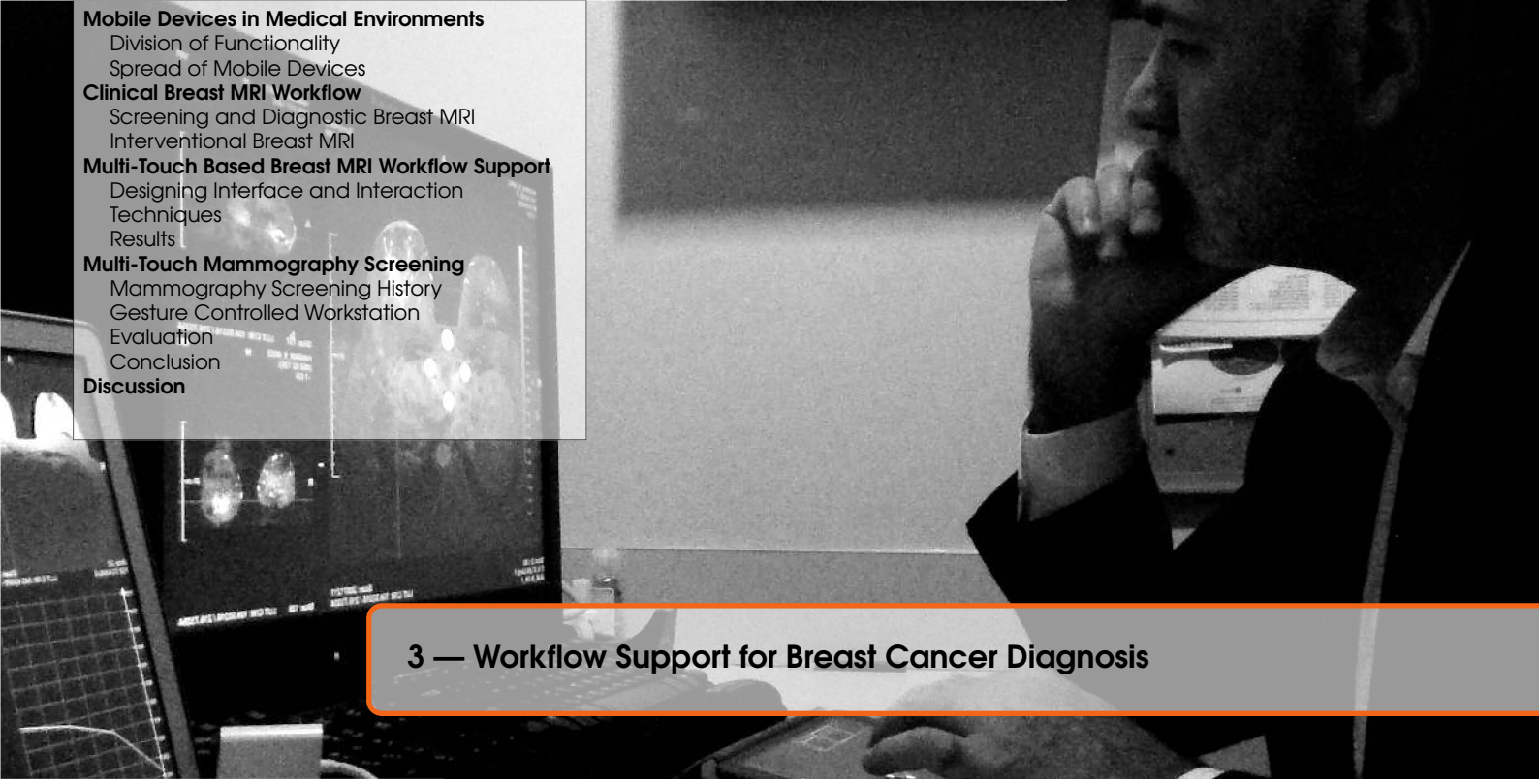
Continuous CAD does not show binary decisions that give the same emphasis to all marks the CADe generates, but shades them according to the confidence of the algorithm. This has been suggested among others by Nishikawa in Yeager (2012) and is somewhat similar to the generation of color maps from DCE-MRI data, where opacity indicates the relative enhancement during the wash-in phase, and color the wash-out slope.

Decision Support Systems based by case databases may offer similar benign and malignant findings based on the overall score, and based on the presence and value of the classification features. One such system was evaluated by Boroczky et al. (2013) and was shown to increase the confidence of some of the readers in the study.

Those suggestions aim at decision support that eases the radiologists' work without making them feel replaced or patronized by the computer.

Patient Acceptance Lastly, also the patients' trust in computer-generated diagnoses needs to be built up, by educating them about the potential of CADx for certain applications, and also about its possible ramifications. CADx performance in many cases compares favorably with human observers, which is why it is used as a second reader for example in mammography screening. Hence, a clear statement about the level of performance is required. Also, it should be put into the perspective of the clinical workload and cost-effective high-performance medicine which is on the one hand asked by patients, but not paid for by reimbursement policies in many countries. Computer support may help to maintain the diagnostic accuracy and the double blind reading standard of radiological images, when human second readers are replaced by computer-aided diagnostic systems.

Mobile Devices in Medical Environments
Division of Functionality
Spread of Mobile Devices
Clinical Breast MRI Workflow
Screening and Diagnostic Breast MRI
Interventional Breast MRI
Multi-Touch Based Breast MRI Workflow Support
Designing Interface and Interaction Techniques
Results
Multi-Touch Mammography Screening
Mammography Screening History
Gesture Controlled Workstation
Evaluation
Conclusion
Discussion



3 — Workflow Support for Breast Cancer Diagnosis

Mobile technology permeates hospitals. Annual reports show that by early 2012, almost two thirds of physicians already used an iPad — arguably the most prominent multi-touch tablet computer today — for professional purposes; half of them even at the point-of-care (Cooper 2013). The prevailing applications for mobile multi-touch devices used by radiologists and health care providers today seem to fall into the "peripheral brain" category (Robinson 2012): Those are applications that provide references or access to literature databases, they include sometimes very elaborate digital editions of anatomy atlases, or aid physicians with dedicated medical calculators. The most notable category of productivity applications, noted in several reports and subject to some studies already, is that of viewers and editors for the electronic health record (Teves et al. 2013).

The interaction with medical images displayed on the mobile touchscreen device has however not received the attention of researchers, examining the difficulties a professional environment with quality standards might impose on the utility of touch interaction.

In this chapter, a novel paradigm for clinical diagnostic software using a mobile multi-touch device for user interaction and dedicated monitors for image display is introduced. We show a demonstrator implementing a workflow-based breast MRI reading system tailored to multi-touch interaction. The demonstrator explores the feasibility of touch interaction for diagnostic reading of MRI patient cases. We show a patient-centric, workflow-oriented concept that is arranged around a multi-touch capable hybrid input-output device. Secondly, we transfer the concept to the challenging scenario of screening mammography reading and describe a second prototype implementing a flexible, yet standardized hanging mechanism for screening purposes.

Main contributors to this chapter have been Felix Ritter (user interface concepts) and the students Simon Benten (Breast MRI workflow system) and Merve Orhan (Multi-touch based mammography reading). Parts of this chapter have been published as:

- Markus Harz, Felix Ritter, et al. (2012). "A Novel Workflow-Centric Breast MRI Reading Prototype Utilizing Multitouch Gestures". In: *Breast Imaging – 11th International Workshop*. Edited by Andrew D A Maidment et al. Volume 7361. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pages 276–283

The chapter image shows a radiologist volunteering to evaluate the iPad-based breast MRI reading prototype described in this chapter.

- Felix Ritter et al. (2013). "Combining Mobile Devices and Medical Workstations for Diagnostic Reading of Medical Images". In: *i-com Zeitschrift für interaktive und kooperative Medien* 12.1, pages 2–9. DOI: 10.1524/icom.2013.0002
- Patent applications have been filed for the ideas and concepts in a wider scope: for Germany as DE 102011087150 A1, and for the United States of America as US 20130141366 A1.

3.1 Multi-Touch Mobile Devices in Medical Environments

Mobile devices with touch screens allow intuitive, gesture-based interaction with content and are today presenting themselves as the coming standard of interaction in both the private and the work environment for communication applications (email, mobile phone, e-book readers). The ease of use through the variety of natural gestures that emulate interaction with physical objects seems to appeal to users.

Research has tested this hypothesis since at least as early as 1988, when Wolf (1988) found that gesture control resulted in less motion and was faster to execute than mouse and keyboard control. A few years later, however, Sears et al. (1991) did not find a significant difference, but after some further years of maturation of gesture-based interfaces, the superiority of gestures over mouse interaction became more and more evident. Both Kin et al. (2009) and North et al. (2009) found significant speed advantages, and Yu et al. (2010) reported in his study that gestures were competitively fast while easier to learn, and the study participants also preferred this interface. Overall, there is a wide-held conviction that touch interfaces will become the dominant interaction paradigm for the near future.

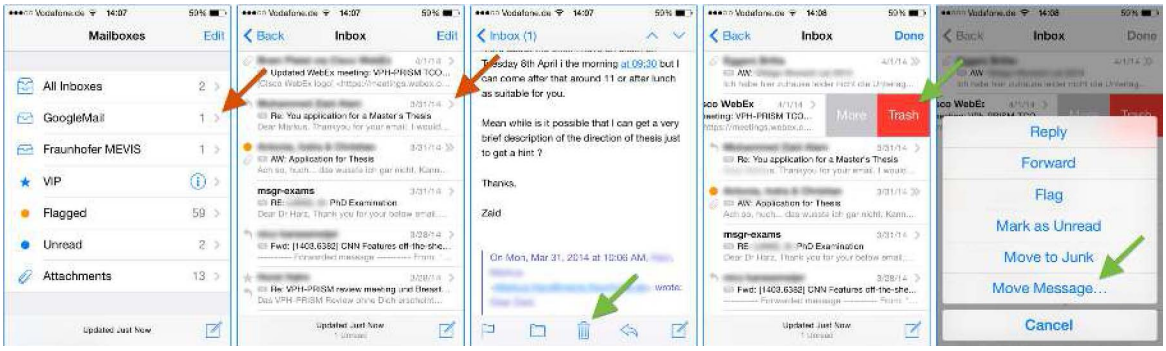
Comparing how applications that have been used on computers operated by mouse and keyboard are turned into applications on mobile devices with multi-touch capable screens reveals insights into the design limitations imposed on such mobile devices. Firstly, the screen real estate is severely limited. Instead of a monitor that provides 15–24 inches screen diagonals, smartphones like the iPhone or similar devices offered by other vendors offer screens of about 3.5–5 inches in diagonal. This limits the amount of information and interactive elements that can be arranged on the screen, making it necessary to redesign information display and interaction paradigm.

The radically restricted set of possible display and interaction methods fuelled inventions on different levels. As one consequence, user interface elements providing information display and interaction capabilities have been merged, like for example in lists in which items can be touched to offer detail information, swiped to show options, etc. This is in opposition to the way to implement interaction in a mouse and keyboard paradigm. Here, most often data is viewed in one area, and interaction is performed using dedicated buttons, entry fields, and so on (compare Fig. 3.1 for an illustration and example).

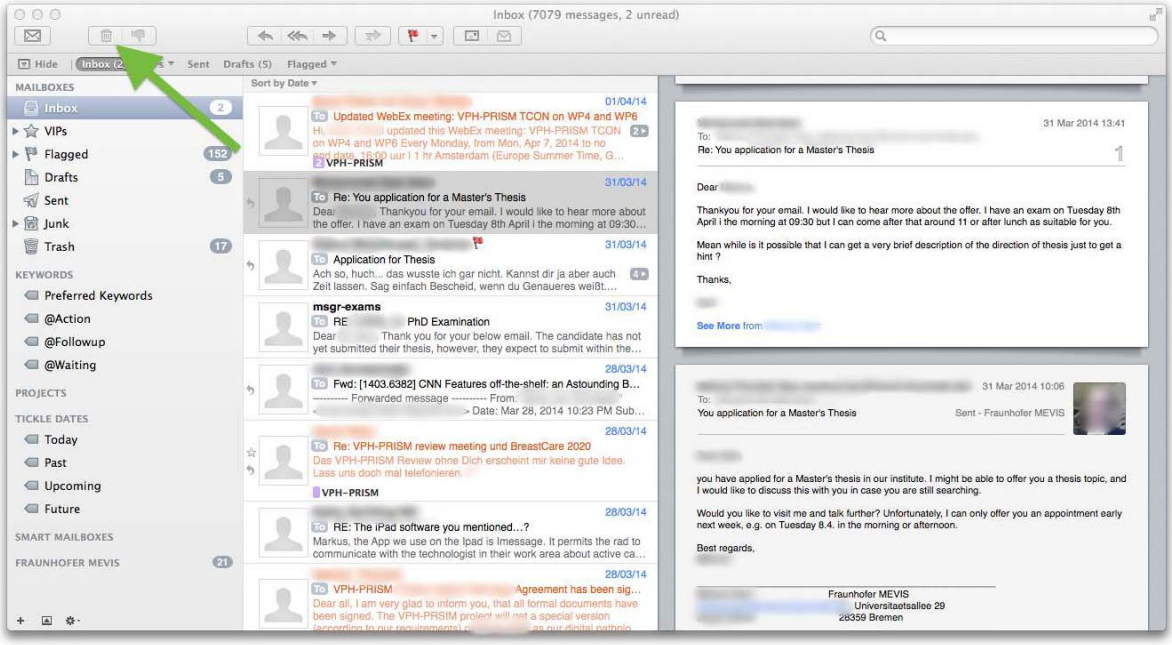
The approach to navigate in mobile applications is likewise affected. From overview information displays to details, from data to input, simple transitions need to be implemented that are animated such that the user receives a visual feedback about the connection between previous and current screen content. Importantly, there needs to be a cue about how to get back to the previous state. One often used navigation paradigm is a linear movement through the information, providing no shortcuts within the application that could confuse the user. Instead, from the overview, more detailed views of an aspect are available, and these might allow editing. Hence, applications need to become less complex, which can be achieved by either reducing their capabilities to the core task they need to fulfil, or by understanding the work they support better, so that they are intuitively understood by the user.

The new possibilities of interaction, namely the operation of applications through gestures and in a direct way has been exploited by many applications on mobile devices, like for example digital maps and navigation applications that can intuitively be operated using gestures to move, zoom, or select; task lists that are location-aware and integrate e-mail client and project accounting, etc.

Not only the information display and interaction methods have changed with the transition to mobile devices. Perhaps the mosts profound change emerged with the possibility to use data transfer to and from the device at any time. Mobile devices are “always on” since they are



(a)



(b)

Figure 3.1: Interaction with a mail reader on iOS and Mac OS X. In (a), typical actions when viewing a message are seen: moving from the mailbox overview (leftmost) to the inbox of one mailbox, and further to one message. Small indicators show navigation options (orange arrows). To delete a message, several options exist, either on the message level (middle screen), or after a horizontal swipe gesture on the message in the message list (second from right), compare green arrows. In (b), the user interface of a desktop mail application is seen. Again, deleting a message can be done in several ways: by a key stroke (backspace), by an icon (green arrow), with a trashcan icon that appears when hovering a specific location in the message display, or by selecting the according entry in the menu or the mail context menu.

derived from mobile phones, which require a permanent link to be able to receive calls. Mobile devices also offer multiple sensors, like cameras, motion sensors, light sensors, and more, which are exploited by some applications to improve the user experience. One example for this are map applications offering to shoot snapshots at locations, associate them with the place, and comment. Some map applications also offer augmented reality: point the device to the scene and see the camera image overlaid with background information on the building in sight. Of course, the map also may also know the details of contacts of the mobile device owner to offer directions, and for example calling a restaurant that one spots on the map is done by touching the phone number in the extended place mark information (compare Fig. 3.2).

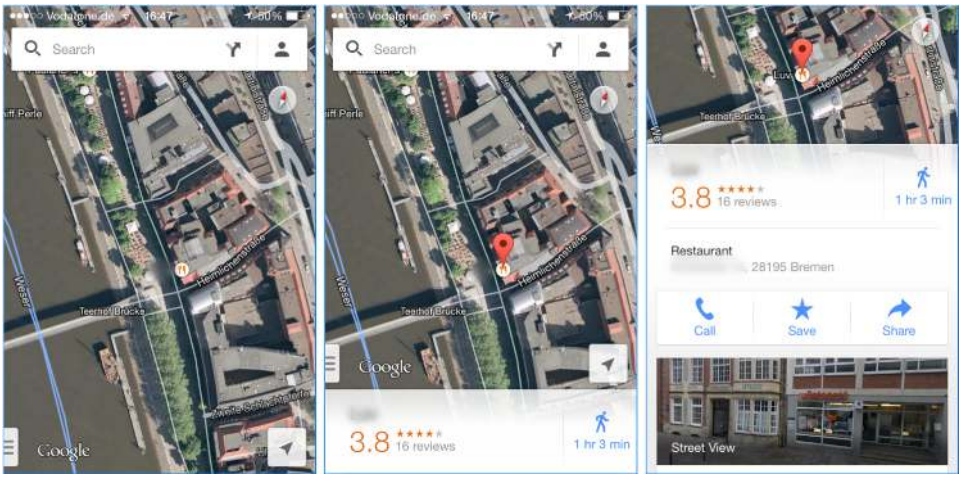


Figure 3.2: Choosing a dinner location in a map application. From left to right: Pan the map to the area where a dinner location is wanted. Tapping the map symbol indicating a restaurant brings up a short information. Tapping the short information brings up more information including images, and the option to initiate a phone call.

Many other domains profit from this confluence of information and connectivity in one device. This is not the case in the medical environment, and it can be argued that the mismatch between information richness and lacking support on mobile devices and with intuitive simple interfaces creates a demand that is today unfulfilled. Viewing or interacting, or even diagnosing medical images on mobile devices is reflected very critically. Reasons are most prominently the lack of FDA approved applications for this purpose, of which there is currently only one. With ongoing work to calibrate the displays of mobile devices, however, this might change in the future (De Paepe et al. 2012). Another often cited downside is that it is not favorable to occlude the image detail with fingers and hand while reviewing diagnostic images, and also, the screens often reflect the room lights, and get dirty from the fingers quickly. This might be one reason that the existing DICOM[1] compliant viewers that can connect to the hospital PACS[2] are not unambiguously welcomed. The limited screen space, missing features, unintuitive tools, unresponsiveness during interaction, or general instability of such viewers are some reasons not to employ mobile devices more frequently (Robinson 2012).

The usability and user experience advantages, however, continue to drive developments of applications worldwide: Mobile touch devices can be highly portable and lightweight, fitting into (larger) pockets while still providing sufficient screen space (compare Fig. 3.3). The screens offer extremely high pixel densities, resulting in crisp display even of small font sizes, and studies have shown that with an intuitive zoom functionality on the mobile device, also the diagnostic accuracy is on level with a full-featured PACS workstation (S. John et al. 2012; Mc Laughlin et al. 2012; McNulty et al. 2012).

[1]Digital Imaging and Communications in Medicine
[2]Picture archiving and communications system

Mobile multi-touch devices may as well give radiologists that are on call more flexibility choosing their workplace: with the iPad, so the promise, they can diagnose from the restaurant table, by connecting to the hospital PACS remotely. However, Robinson (2012) doubts that this alone will will make the case for the mobile devices, given that small notebook computers like for example the MacBook Air, are in a similar weight and battery endurance category, while offering a full keyboard and a larger screen. The most widely used radiology application on the iPad may be the OsiriX viewer. It operates smoothly, has connectivity to the hospital PACS, but offers only basic functionality. Still, there appears to be a wide-held belief that mobile devices will sooner or later become ubiquitous tools, spearheaded by tech-savvy clinicians who pioneer in using private devices and generating demand for solutions beyond the current range of applications (Marceglia et al. 2012).

Figure 3.3: Many types and sizes of mobile devices exist, most of them equipped with touch screens. Not all are equally well fit for the purpose of image review, but many may be put to good use if a suitable division of labor between device and environment can be found. Image by Brad Frost under the Creative Commons Attribution 2.0 Generic license (`https://creativecommons.org/licenses/by/2.0/`)

There is in fact much more than DICOM viewers or reference manuals a mobile multi-touch device may offer to clinical personnel. For digital pathology, Y. Wang et al. (2012) proposed a touch-controlled table to review slides. In the according study, touch interaction sped up the diagnostic process. In surgery, the iPad has pioneered as an augmented reality navigation aid (Schenk et al. 2013), after having been employed for the same purpose, but with a much less sophisticated setup years before[3]. Also, the iPad has been used to navigate lung surgeries and to augment ultrasound examinations of the heart by overlaying a volume rendering of the heart with the live image of the tracked ultrasound probe (Eguchi et al. 2012; Ford et al. 2012).

In our contribution, we propose a very different usage of the mobile device in settings where many images, or large images, need to be viewed for diagnostic or screening purposes.

3.1.1 Division of Functionality Between Devices

To emphasize the novel aspects of the application utilizing an iPad to support efficient workflows in breast radiology, an abstract categorization of interfaces involved in human-computer interaction helps. Out of many possible categorizations of human-computer interaction, in the following we will speak of its *technological* and its *functional* level. These levels can be thought of as layers, where the technology provides the basis for functionality, or, stated the other way around, certain functionality could not be provided without developments on the technological level.

As an example, this applies to functions of mobile phones in obvious ways: With the *technological basis* of ubiquitous connection to the internet, either using a mobile data network

[3] `http://www.youtube.com/watch?v=n5nbNIpqdAY`; accessed Nov. 2013

or a wireless local area network (WLAN) connection, a connection to one or more servers can be held active all the time. These servers may exchange several information items with the device, with the notable possibility of two-way data transfer. This allows to build supportive *functions* for the user like context-aware information either on demand or even in a "push" way, without requiring the user to take action to retrieve the information. Examples are location-based services, like reminding the user to buy an item on the shopping list when she is close to an appropriate shop. Push services are the standard for e-mail, so that the user is not required to retrieve new e-mail with a request. Undoubtedly, new technologies have changed the functions we expect to have available. The same will be true in the medical scene.

On the technological level, key components of human-computer interaction in a reading environment are image display devices and input devices. These items in the workplace interoperate with the human observer, or are manipulated by him or her. All other technical items can be abstracted from and are not subject to our considerations: a mechanism to retrieve images, which can for example be a wired or wireless network, or physical data storage devices; a computing device that takes care of the rendering, which can be a local workstation, or an image server with a remote client attached to the display device, or other setups. Input devices are manifold, starting from mouse and keyboard, specialized keypads, and voice dictation systems.

On the functional level, abstract tasks can be enumerated that are required for achieving clinically relevant goals. The images have to be selected and displayed, and some means to manipulate the images are normally also required for non-trivial clinical scenarios. Reporting a clinical decision is the final inherent functional requirement for any diagnostic reading system.

By example, a typical clinical scenario shall be described. Images are centrally stored in a PACS (Picture Archiving and Communication System), which can be a single or distributed server computer with a large enough data storage device attached to it to persist all images and secondary information. Whenever an image is taken with one of the attached PACS-enabled image capture devices (MRI, CT, PET, x-ray, ...), it is sent to the central PACS server. To read the images, they are either actively pushed to PACS clients or the PACS server is queried on demand (requested) by a PACS client. Data transfer is often physically implemented on a wired high-speed connection, but increasingly, cloud services are being used that emulate for the client a "local" data storage regardless of the client's physical location and network connection.

The images are then presented on the actual monitors of the PACS node. PACS nodes differ in their usage scenarios. While some are meant to bring up all kinds of images for a comprehensive review, others are used for diagnostic image reading. With the intended task of the node, the technical setup varies, most importantly regarding the display devices that can range from small (15") color monitors in landscape orientation to large, dedicated mammography monitors that are black-and-white and used in portrait orientation. It is crucial to note that in such a heterogeneous environment, there is no knowledge available in the PACS server about where, how, and why particular images are being shown. PACS clients are hence very often general-purpose tools, equipped with extensions that enable advanced functions for particular scenarios, like for example required in mammography screening. Most often, these advanced tools plug into the PACS client-server infrastructure by implementing the DICOM protocols to store and retrieve data. They thus don't require to run on the same computer as the PACS client, and will much more often be stand-alone workstations with a suitable hardware setup for the purpose they serve. Again, mammography screening is one prominent example; 3D volume rendering solutions, intervention planning tools etc. are others. All these examples require hardware that is not normally part of a PACS client, for example dedicated monitors (mammography), dedicated high-performance graphics capabilities, or hardware interfaces for sensors, tracking devices, or robotic devices.

3.1.2 Spread of Mobile Devices

There are only very few scientific publications exploring the subject of multi-touch-based interaction for image-based diagnostic reading. We speculate that on the one hand, mobile multi-touch devices are not available for long enough to be profoundly researched for this application, and on the other hand there is no clear paradigm visible how such devices should be employed to be useful and intuitive. Our contribution aims to help in both challenges. We want to provide a novel approach to integrate a mobile device in a clinical setting, and we want to present demonstrators that implement our paradigm on challenging clinical topics that can be explored further.

Prior work that we acknowledge has been presented by Lundstrom et al. (2011), who employ multi-touch gestures for interaction with medical images, and use a flexible tool selection menu. Their work, however, is aimed towards team interaction on a very large display table. A very similar approach has been presented for the application on digital pathology (Y. Wang et al. 2012) and for medical team meetings (Avila-Garcia, Trefethen, Brady, and Gleeson 2010; Avila-Garcia, Trefethen, Brady, Gleeson, and Goodman 2008). A work closer to the system proposed by us has been described in the US 2011/0113329 A1 patent publication. In this work, the author proposes a static setup on the mobile device, where touch wheels and buttons are depicted and usable with two hands. While this may be intuitive in the sense that the visual appearance of the user interface elements suggests the way how to interact with them, the approach is fundamentally flawed: a physical wheel is probably easier and more intuitive to operate than a graphical one, so a user interface requiring the user to turn a wheel should offer a physical wheel. BrainLab, on the other hand, employs a wall-mounted multi-touch display for use in brain surgical interventions, and has also foreseen the integration of a multi-touch mobile device, which is to be attached to the main system (Pat. EP 2 031 531 A2), thereby not extending the functionality of the wall-mounted monitor, and even limiting the capabilities of the mobile device. Lastly, IBM has proposed a system in which the image data is viewed on the mobile device (Pat. US 2010/0293500 A1), which is both obvious and in contradiction to the facts enumerated above.

Reviewing the cited body of scientific work, patents, and available tools, we have condensed the status quo of available approaches and concepts into the following categorization. It served as the outline for the design and implementation of the dedicated clinical support we want to contribute, extending the status quo.

Paradigm In many scenarios, the mobile device is foreseen to be used independently of the hospital information systems, though connected to it by WLAN or other wireless technologies. Either an App (Application, for example downloaded in Apples App Store or on the Android market), or a fully web-based user interface provides the functionality. Our approach will be fully integrated with the hospital information systems and provide unique benefits by this, augmented by device capabilities like location services.

Viewing In general, images are viewed on the mobile device, and interaction with the images is conducted on the screen. For interaction, a reduced set of tools is offered in toolbars, emulating the interface of the corresponding workstation software. For this work, we will pursue a contrasting approach in which images are not shown on the device, acknowledging however, that there are use cases prompting for a deviation from this paradigm.

Interaction Usually, the tools provided follow the workstation tools, and work like those. One notable exception in some cases is the zoom/pan functionality, where e.g. zooming is accomplished with a two-finger pinch gesture. We go beyond this, by suggesting gestures and interaction patterns that exploit more of the unique multi-touch device capabilities.

Workflow Workflow is generally not an issue addressed by the mobile apps, since they are mostly not intended to be used for diagnostic image reading. Hence, no structured review of images is implemented as of today, rather a random-access toolbox is provided. We oppose this paradigm by proposing adaptive tools that use prior knowledge on the images to pre-select most likely tools and actions.

Security To access data for example from the hospital PACS, the user in today's applications needs to log in to the hospital IT with his or her access rights. The device then acts as a remote viewing station. We will challenge this by conceiving of the mobile device as a personal key to the system (of course requiring security precautions elsewhere).

Implementation The choice today is often made between high-performance platform specific implementations (iOS, Android, and others) or platform-independent applications that run in web environments. Using the platform-specific implementation, all specific resources the platform offers can be leveraged, but development for multiple platforms gets more expensive. Using the web-based approach, sacrifices usually have to be made in terms of hardware access. Mixed approaches do exist, but they also compromise on features and capabilities. Our implementation concept tries to combine the benefits of both approaches, rather than their downsides.

Intended use Some image viewing mobile applications only seem to target the casual user who wants to have a quick look at a specific image while away from a workstation. For example, a radiologist on call might appreciate to see a case on his mobile device to decide if he needs to drive to hospital to see the emergency patient. Our concept rather envisions a central role in the daily routine of health care giving in a hospital environment.

In our contribution, we follow a different paradigm, propose a concept and setup that targets diagnostic reading, makes use of a different approach to security and data handling, and consequently has a different user group in mind: the radiologist doing diagnostic reading.

With our work, we want to challenge several aspects of conventional breast MRI workstations, and provide a clinically usable setup centered around a combination of stationary display devices and a mobile device that provides ubiquitous interaction with clinical data. Most importantly, we wish to move away from the static *random access toolbox* approach in current reading workstations, and introduce workflows into breast MRI reading.

We say that a workstation follows the *random access toolbox* (RAT) paradigm if it presents the available image series of a patient study on screen or in a selection list, and similarly offers image analysis tools in menu bars, buttons, pop-up context menus and so on. It is the user's task to decide which series to view, which tool to select, and where to apply it. Many workstations offer a plenitude of tools for zooming, like free zoom, fixed zoom levels, zoom-to-fit etc., several annotation tools, and different ways to measure or segment areas of interest.

We will present a prototype that follows a different paradigm we want to dub the *context-aware tools* (CAT) paradigm. We want to offer functionality and information at the time they are needed and for the images where they are needed. Also, we see the display devices with attached IT systems and hospital IT connectivity as one unit together with the mobile device, as opposed to the more conventional approaches of using mobile devices in which the mobile device emulates a mouse and keyboard to pose as a better remote, or replaces the diagnostic system altogether.

3.2 Clinical Breast MRI Workflow

To elucidate the properties a workflow based solution to efficient breast MRI reading needs to exhibit, the following considerations will quickly review the defining characteristics of *workflows* to proceed with an elaboration of the indications for breast MRI, from which two general classes of breast MRI image centered workflows will be derived and subsequently be described in more detail.

Workflow, by our definition, is an organized routine process. It has an objective, and creates an output from some input. In the breast MRI reading scenario, the primary input is the image data, and the output a clinical recommendation or decision derived from the data, e.g. proposing a follow-up examination after a certain time and with a certain recommended imaging technology, or the recommendation of a biopsy. Breast MRI is also employed to conduct minimally invasive tissue sampling (biopsy) procedures, and may be used to insert targeting aids (metal wire guides) into the breast tissue that are meant to help a surgeon to excise a lesion marked by the wire guides.

The workflow from input to output is usually executed by computers and radiologists, partially independently, partially cooperatively. Typical steps computer algorithms execute independently are those that prepare the data for reading. This may include motion correction to improve later quantitative tissue analyses, generation of color maps, calculation of subtraction images, detection of the breast tissue to mask air and other organs during display, and sometimes detection of lesion candidates. These calculations are done without interaction, and in a predefined order, triggered by the incoming data.

The part of the workflow involving the radiologist starts when the data is ready for review. The first question is how the radiologist learns about the availability of new data. While it is taken for granted that today an instant notification signals the arrival of a new email, the same mechanism is not commonplace in clinical environments (and perhaps not a helpful mechanism either). Rather, radiologists need to log in to the breast MRI reading workplace with their access rights, and query the database for the patient they wish to review. An alternative to this are work lists, where new patients are indicated to radiologists.

After loading the data into the reading workstation, the radiologist is presented with a default view of the data which can usually be tailored to the specific protocol in a given radiology department, and also to the presently available series. Breast MRI is optimally read on dedicated workstations rather than on viewers integrated into the PACS system, to provide a workplace with advanced viewing and analysis capabilities. In this way, the image presentation can for example automatically adapt to the diagnostic reading scenario or to intervention preparation (e.g. MRI-guided vacuum-assisted biopsy), and present results of the various image analysis algorithms or quality improvement methods discussed before (Sec. 1.2.3). Notably, dedicated workstations are not in all cases the first choice, particularly if a correlation to other imaging modalities needs to be performed for diagnostic purposes. Then, radiologists may be tempted to interrogate the images on the PACS client.

Indications for Breast MRI. The optimal set of tools, and the most helpful selection of information to display in a breast MRI workstation depends on the goal of the workflow. The goal again depends on the motivation for acquiring the MRI series of the woman. The following list comprises widely accepted indications to perform contrast enhanced breast MRI, under the condition that the equipment and expertise to conduct MRI-guided biopsy procedures are available on site (Destounis 2006):

1. Screening for high-risk patients (family history of breast cancer, own prior breast cancer, genetic predisposition).

2. Evaluation of tumor burden in cases with a recent diagnosis of breast cancer.

3. Evaluation of implants.
4. Search for occult primary cancer if metastases are detected.
5. Workup of cases with indecisive and indeterminate prior examinations (ultrasound, mammography, physical breast examination).
6. Surgery preparation and assessment of residual cancer in cases with close margins at histopathology.
7. Chemotherapy response monitoring.

Adding to these, MRI imaging is obviously performed when MRI-guided interventions are performed, like biopsies or placement of metal wire guides. Note that a MRI-guided biopsy is no indication for a breast MRI, but a consequence, which is why it does not add to the above list. All the indications can be summarized into two categories of workflows with shared aims: gaining diagnostic insight, or providing therapeutic or interventional guidance. This leads to the following two fundamental clinical goals involving breast imaging (references to the above list in parentheses):

Diagnostic insight Screening and diagnostic breast MRI including chemotherapy monitoring. The emphasis is on a comprehensive review of the examination in a workflow that guarantees that no suspect areas are missed (1 and 2–5; partially 7).

Interventional guidance Preparation, support, and monitoring of image-guided interventions like biopsies and metal wire guide placement. Efficient and reproducible targeting is the goal of a workflow targeting this group (6–7).

These general two categories will be further elaborated below to derive the requirements for approaches to the workflow-based presentation of breast MRI images and aids for goal-specific sets of tools.

3.2.1 Screening and Diagnostic Breast MRI

The American Cancer Society (ACS) recommends screening MRI in women with a relative breast cancer risk estimate of 20–25 % when estimated mainly from their family history (e.g. Gail model or Claus model). Genetic mutations and dense breasts also consolidate elevated risks suggesting screening by breast MRI in addition to breast self examinations (BSE) and mammography.

The German Interdisciplinary GoR level III Guidelines for the Diagnosis, Therapy, and Follow-up Care of Breast Cancer (Interdisziplinäre S3-Leitlinie für die Diagnostik, Therapie und Nachsorge des Mammakarzinoms) also propose to screen women at high risk, where a 30% risk by familial risk estimate, or a proven pathogenic BRCA1 or BRCA2 gene mutation make women eligible for MRI-based screening starting at age 20–25 years. This is to be complemented by ultrasound, breast self examination (BSE), and mammography, depending on breast density and age (C. H. Lee et al. 2010; Leitlinienprogramm Onkologie 2012).

Screening Breast MRI. Case numbers of breast MRI screening only begin to increase at specialized centers. Looking into the future, however, there will be a strong need to establish a breast MRI reading protocol for screening purposes, and it will very likely have more commonalities with the hanging-based approach used in mammography reading workstations than with breast MRI workstations as they are known today.

Correlation The prime task in breast MRI screening is the *robust and reliable spatial correspondence establishment* between previous and current MRI exams, and to some extent the correlation with other imaging results like mammography and ultrasound which are

recommended to be conducted in an interleaved fashion with breast MRI. Establishing correlations is today for example achieved by displaying and stacking through the images side by side, and manually. When images of different modalities are displayed in distinct software or even hardware systems, multiple workstations may have to be operated in parallel. For comparison of the most recent breast MRI images with prior examinations, the synchronization based on the current slice is the simplest conceivable support. In this case, even small positioning differences make the slice-based synchronization useless, and more typically, radiologists would browse through the current image series, and based on anatomical landmarks try to establish the correspondence to the prior series by browsing through it.

Annotation Once a suspect finding is seen, it needs to be *annotated and quantified*, which involves the description of location, enhancement characteristics, and size. The finding position needs to be described such that correlation with other imaging is possible, and in a way that enables later recovery of the position without ambiguity. Certainly, for this purpose a digital annotation stored with the data is preferable over a prose description in a radiology report.

Quantification Since follow-up assessment is most helpful if tiniest changes can be quantified in a reproducible manner, computer supported measurement and quantification of lesions and other imaging biomarkers throughout the history of images may be desired.

Overall, two supportive types of algorithms may clinically be appreciated: (1) those that allow to navigate from a finding in the prior MRI scan to the same finding in the current scan (or the other way around), and (2) those which allow to track lesion changes quantitatively. Both tasks have been subject to research in the past. For the spatial correlation between follow-up MRI exams, methods have previously been proposed, relying on deformable registration techniques (Boehler, Schilling, et al. 2010). Building upon these techniques, and combining them with lesion detection (Dorrius et al. 2011; Renz et al. 2012; Srikantha, M. Harz, et al. 2013; Takeda et al. 2012) and automated segmentation (J H Moltz et al. 2009), features extracted from the lesions can be traced over time, like for example those presented in Sec. 2.3 and Sec. 2.4.

Clinicians need to review the growing amount of information that is generated by high-volume producing imaging modalities like breast MRI when they are employed on an increasing number of women. This demands novel workflow support approaches. While the breast MRI screening scenario is not specifically addressed in the following contributions, it is worth considering in future research. It accounts for the largest portion of breast MRI exams, since 5–10% of women are at elevated risk by the criteria cited above, and may hence be offered regular breast MRI exams for extended screening, while only 0.1% of women are affected by breast cancer in their life, requiring a MRI exam for treatment decision taking. Approaches that will be described below are still general enough to be extended beyond their current purpose, and even beyond breast MRI.

Diagnostic, Curative, and Preoperative Staging Breast MRI. Unless screening MRI is conducted on larger patient cohorts, there is no marked difference between the workup of the MRI series for screening or diagnostic (curative) purposes. While there may well be a change in the future, there is probably no distinction today between screening breast MRI and diagnostic breast MRI reading, and dedicated tools are just beginning to emerge addressing specific needs and situations of them. In particular the two major tasks of correlation imaging with other modalities and the aim of quantitative description of findings is common to both indications for breast MRI. A difference may exist in the variety of images the breast MRI series need to be correlated with, since in a MRI-based screening scenario, fewer or no additional imaging will have been performed during the same visit.

On the other hand, the focus in diagnostic and curative imaging is less on the detection of suspect lesion than rather on their conclusive evaluation, necessitating CADx rather than CADe support. In preoperative staging, neither is required, but decision support systems might help to suggest therapeutic approaches (Séroussi et al. 2013) or display example cases with proven outcome that are similar to the current one, or even model patient-individual tumor response to different treatment options (Madabhushi et al. 2011; Takada et al. 2012) to facilitate and optimize clinical decision taking.

The diagnostic breast MRI scenario that will be the primary focus of this contribution restricts itself to the review of single MRI exams consisting of a number of MRI contrasts acquired in the breast imaging protocol. The clinical question that gave rise to the MRI examination is often one of the following:

- A women presenting for screening mammography is recalled because of a suspicious finding. The primary second look imaging modality is hand-held ultrasound, but not all breast cancers can be detected or conclusively diagnosed by ultrasound, so that in several cases a DCE-MRI is the imaging method of choice (Destounis 2006). This is in particularly helpful in women with dense breasts or high breast cancer risk. Overall, in the ACRIN 6666 trial, breast MRI has been reported to find additional cancers where mammography and ultrasound have failed to do so (Wendie A Berg et al. 2012), though the higher false positive rate attached to this improvement needs to be reminded of.
- A symptomatic women presents herself for diagnostic workup, and her case cannot conclusively be solved by diagnostic mammography or breast ultrasound. The aim is to determine the need of a biopsy to substantiate the suspicion.
- The most prevalent indication for DCE-MRI is preoperative staging of biopsy-proven breast cancers. In this case, a history of imaging already exists and has been reviewed, and the task is to correlate the MRI findings with the prior imaging and clinical data.

The imaging protocol may differ from radiology department to radiology department, but the centerpiece is today a contrast enhanced sequence of acquisitions. In addition, further morphological or functional sequences may be added to the protocol, with some depending on circumstances. The workflow during diagnostic workup of breast MRI will hence slightly deviate from site to site, but the shared goal needs the following actions to be taken:

- Assess the diagnostic quality of the images, for example by quantifying and describing motion artifacts, ghost images, and others that might hamper diagnostic reading of images. Consider personal details of the patient, like age, breast density, hormonal status, risk factors etc. which may impact the visual appearance of the images.
- Browse all series in an ordered fashion that ensures that all sequences have been reviewed, and that all areas have equally been examined, e.g. the left and the right breast and the axillary region.
- Correlate the current MRI series with prior MRI series, with mammograms from the same or a prior visit, and with ultrasound images from the same or a prior visit, to obtain a
- Generate measurements of findings that can be summarized in a report. Ideally, annotate the finding in a way that is persistent and makes it easy to recover the exact suspect region.
- Collect all diagnostic results (measurements, annotations) to suggest a BI-RADS grading, and persist the collected diagnostic results alongside with a recommendation (biopsy, short-term follow-up, etc.) in a report.

The ordering of the items may be a matter of personal habit, and they may be executed in parallel rather than one after another, e.g. artifacts, signal to noise, and patient motion

are assessed while reviewing the series, items and suspect findings to report are noticed while browsing the images, but in a second run annotated and taken into a report.

Workflow support for diagnostic breast MRI reading will need to support the tasks above. In this workflow, some actions can be supported by computerized analyses, for example the quantification of patient motion and the likelihood of motion correction to have succeeded (Boehler 2011), and some can be supported by proper display of information (risk, hormonal status, status of last breast MRI, etc.). Ensuring a workflow that makes navigation through the images and annotation of relevant findings efficient and intuitive, however, requires a goal-specific presentation of images and tools we will propose after the following brief account of breast MRI in the interventional setting.

3.2.2 Peri- and Intra-Interventional Breast MRI

When the decision has been taken to sample tissue from a suspect lesion, there are two major workflows depending on the clinical case.

Surgery preparation. One major aim in using breast MRI prior to intervention is to assess and quantify lesion extent, multicentricity and multifocality of the disease. The result determines the surgical options and depending on them also the way the lesion should be marked.

Disease affecting more than one quadrant of the breast is often considered for mastectomy, which can be skin-sparing, nipple-sparing, or radical. Image guidance is of limited help in the preparation of such interventions. Instead, it has its purpose in lesions restricted to one quadrant, where the total size of affected tissue suggests the image-guided placement of metal guide wires in or around the lesion. Here, breast conserving therapy (BCT) is a method that is proven to equal mastectomy in terms of years of disease-free survival, if accompanied by radiation therapy. A size limit of about 2 cm is considered to decide if only one metal wire is inserted peeking directly through the lesion, of if two or more wires demarcate the area of concern. Also, the lesion type is considered in this decision, where palpable mass lesions require less effort than non-palpable in-situ cancers. When two or more wires are inserted, the technique is usually called *bracketing*, as opposed to one wire, which is called *wire localization*.

Surgical excisions of the breast conserving type are also called “open biopsies” since they don’t aim at definitive treatment, but at tissue sampling in cases where minimally-invasive biopsy may be prone to fail. Open biopsy is often considered the reference standard biopsy technique. Insertion of the needle (wire) guides needs the confirmation of imaging, which can be achieved by any suitable imaging technique for the given lesion, including breast MRI, which often allows for a more accurate depiction of the extent of disease.

Based on the fraction and situs of the affected tissue, including involvement of the pectoral muscle, the nipple, and the skin, the surgical approach can be chosen. This comprises planning of immediate or delayed breast reconstruction, including the according procedures and techniques. Breast MRI is one modality to generate a patient-specific deformable model of the breast to assess the different reconstruction options; compare Sec. 4.3.1 and Georgii, Eder, et al. (2013). Such specialized tasks today require dedicated computer software (and sometimes also hardware due to high computational demands) and are not part of standard procedures.

Surgical techniques are manifold and vary with patient-individual anatomy and the specific aim. Some of the more commonly employed techniques are for example summarized in Kass et al. (2007). Particularly lumpectomies and their image-guided planning techniques are also explained in W A Berg (2006), including indications and contraindications setting them apart from mastectomies.

MRI-guided breast biopsy. MRI-guided breast biopsy will be the topic of Sec. 4.3.2, and will be discussed there including a description of the biopsy workflow. In general, many different types of targeting instrumentation have been devised, which differ regarding their physical layout. Similarly, different types of vacuum-assisted biopsy needles exist, which exhibit larger or smaller cavities, and either sharp or round tips to access lesions even under difficult circumstances (O'Flynn et al. 2010).

Dedicated software support is part of commercial breast MRI diagnostic workstations, and includes the automated detection of fiducials to obtain a spatial reference system for planning. Robotic needle placement is also explored in several research contributions (Chauhan et al. 2002; B. T. Larson et al. 2003; Mallapragada et al. 2011), but receives critical acclaim since practitioners consider the haptic feedback they receive from the stiff tissue around the tumor to be crucial in accurate punctures. Without feeling the lesion and fixating it by countering the pressure against the advancing needle, the lesion might escape the puncturing (Apesteguia et al. 2011).

Other interventional techniques. MRI guidance is also possible in other curative interventions for breast cancer, including radiotherapy planning and monitoring, and image-guided thermal ablation methods including radiofrequency ablation (RF ablation), focused ultrasound (FUS), and laser interstitial thermal therapy (LITT). Thermal ablation (including cryoablation) is a rare option for selected indications, and radiotherapy is predominantly performed using PET/CT both for planning and success monitoring.

3.3 Multi-Touch Based Breast MRI Workflow Support

In the following a framework for a workflow-centric breast MRI reading prototype will be described. We exemplify the use with a prototype tailored to diagnostic breast MRI reading, addressing one of the above three scenarios explicitly. However, since by the nature of the proposed workflow approach it is already foreseen to support any other application of breast MRI, adaption to other scenarios requires basically the definition of the underlying workflow.

The proposed system consists of a server side implementation that hosts the data and displays all images on screen. It is accompanied by a client side implementation that runs as a native program on the mobile device. The two parts are connected with a basic and efficient network protocol to exchange information and events using wireless LAN on the mobile device side, and any network connectivity to the same network on the server side.

The software demonstrator has been implemented based on MeVisLab. The device demonstrator has been implemented using the Apple inbuilt development environment, and is written for the iOS framework in ObjectiveC. The C++ implementation of the gesture framework has been accomplished based on the Qt framework and runs on the server side, and not on the mobile device. This eases the integration of mobile devices with different operating systems since it only requires the reimplementation of less critical components, and also allows a central control, extension, or change of the gesture system without device dependencies. With application-side interpretation and implementation of gestures, coherent interaction is guaranteed regardless of mobile device. Alternative technologies that use device capabilities while providing a device-independent programming environment usually have to choose the smallest common denominator of functionality and scope, while the restrictions in our approach are only those of the Qt gesture framework. Technically, for the workflow the device is only required to send one or multiple touch points, and all intended interaction is evaluated on the server computer.

3.3.1 Designing Interface and Interaction

Technically, in the typical clinical PACS environment described above, the system presented here is such a DICOM client to the PACS. It is equipped with functionality to communicate with the PACS server bidirectionally to retrieve patient data and images, and to store secondary captures and diagnostic information with the case. However, several aspects of the technical implementation anticipate developments in the health care environment (in this work, the hospital) that can be exploited to enable novel supportive functionality to ease care-giving.

To collect requirements for the design of a user interface that incorporates a mobile touchscreen device into the work routine, site visits were prepared. Firstly, a set of paper drawings suggesting different layouts for the patient overview screen on the touchscreen device were drafted. Second, a structured interview guide was designed to assess the mental representation of several technical aspects of DCE-MRI image formation, including MRI contrasts, image quality assessment, and image artifacts (compare Appendix B.1). Further, participants were asked to describe aspects of the computer-based image interpretation workflow without showing it on the computer. These descriptions were later compared to video recordings and observations of how the computer interface was actually used when reading breast MRI examinations. The application design and the included tools were then developed from the insights gained.

The train of thought leading to the actual implementation of our approach can conceptually be captured in the following five levels of user experience proposed by Garrett (2011): the Strategy, Scope, Structure, Skeleton, and the Surface level. In Garretts description, the five planes can be summarized by characteristic questions as follows:

Strategy The strategic level describes user needs and application objectives: What do users expect from the application, and what are the application provider's aims and interests?

In our application, we decided to support the reading of contrast enhanced breast MRI examinations, comprising several MRI contrasts. The research objective is to explore the minimally required set of supportive tools in the assessment of the data.

Scope The scope comprises functional specifications and content requirements: Determined by the strategic decision about which user needs should be fulfilled, find the specific requirements for content and functionality the application has to offer.

In our context, users expect to be able to review all contrasts, and to be able to visualize derived information, like for example color maps. In addition, closer inspection of target areas, and their quantitative characterization is clinically required.

Structure The structure describes how the provided features will work together, and how the user can reach all functionality so that tasks can be accomplished.

Our concept here is to employ a merely gesture-based concept in which all tools are local to the current view and position.

Skeleton The skeleton, in s nomenclature, outlines how user interface elements are arranged in principle, to support the structure of the work.

This is in our concept not a fully applicable design level, since the skeleton will need to change according to the application state and currently executed task. Overall, since our objective on the strategy level is to discard mostly all user interface elements and replace them by gestures, this level will need to be addressed differently.

Surface How user interface elements have to look like is likewise answered differently in our implementation, where the design of the gestures and their composition into a set of gestures that are sufficiently unique to not be mistakenly used, yet intuitive enough to be memorized easily.

3.3.2 Techniques

Paradigm

In our setup, the mobile device poses as a hybrid image display and interaction device, changing its role during the workflow. The fundamental principle is not to show images for clinical diagnosis on the device. The major reasons are (1) the limited screen real estate, where for example correlated viewing of orthogonal reconstructions in sufficient size is not feasible; and (2) the fact that during interaction the fingers will occlude a large portion of the small screen. Hence, in our opinion image display is better left with dedicated display devices. These can be tailored to their purpose, e.g. a defined contrast range for diagnostic reading, or large screen sizes in a operating theater. The mobile device also never stores data locally, which increases security, though the concept might also include secured storage of selected key images per patient for patient visits or interdisciplinary board meetings. In the general setting, the data is provided by a server that is typically connected to the hospital IT. The login procedure to access the data is accomplished by linking the mobile device with the display device.

Authentication and Location Awareness

In our setup, we think of the mobile device as a personal item belonging to the radiologist. He will log in to his device and authenticate towards it. To connect to a display device, different mechanisms are conceivable. In our current implementation, the internal camera of the mobile device reads a QR[4] code, a type of matrix barcode, that is displayed on the display device and that encodes the location and capabilities of the display device. By reading in the QR code, the

[4] Quick response

Figure 3.4: iPad screens. From left: (1) Reading the QR code. (2) The patient browser interface. (3) The iPad user interface during reading. MR images are displayed on the display devices (not shown).

mobile device learns about the display device, and configures itself such that the available tools are offered, and only the applicable patient data is shown to the radiologist. In practice, in a patient room only the data pertaining to the patients in this room are offered, and diagnostic tools, annotations, and reporting functionality are not provided; in a meeting room only the data of today's tumor board meeting might be shown with annotation functionality, while in a diagnostic reading room, all functionality will be provided for all patients assigned to the doctor.

Workflow

In contrast to existing workstations for breast MRI reading, we have removed all tool bars and menus from the application. Two reasons exist for that:

1. From an assessment of several experienced breast MR readers' usage of breast MR workstations, we observed that a very small subset of available tools was frequently used.
2. With a touch-based interaction paradigm, tool bars and menus are no longer a convincing means of interaction, because they necessitate a pointing device.

Workflow analysis was carried out in a community hospital breast care center, where an above-average number of MRI exams are being read (local high-risk population). We have been trained on the same workstation by independent experts prior to the workflow analyses, such that we knew all tools that can be employed during reading. With this background, we have observed four radiologists while reading MRI exams and video-taped their work. We have interrogated the radiologists on their tool usage, and on their reasons for using or not using them. One remarkable finding was that the only annotation tool that we have seen in use was an arrow pointing to a location of interest. Sizes have been determined with a digital ruler, and the results either stored using a screenshot or by dictation into a reporting system. Segmentation of findings for statistical kinetic characterization was never employed. We also noted that the tools were often selected via the menu bar and sub-menus of that, but less frequently using keyboard shortcuts or icons, of which many are available both in a central tool bar below the menu of the application, and in each single viewport. A notable problem reported by the radiologists was that the distance the mouse has to travel on screen is rather large, and that it is perceived difficult or impossible to remember which functions are available, and how or where to invoke them.

From these assessments of breast MR reading workflows, we designed a novel user interface both for the mobile device and the display devices. Main driving factors were to provide intuitive usage, and to minimize the necessity of large-scale movements to access functionality. The most complex interaction is patient selection and workup of patient history, which is consequently handled on the mobile device in an intuitive patient browser that shows the patient history with clinical events and risk data (cf. Fig. 3.4, middle). Annotations are correlated between sketch (left), image series (middle), and time line (right) using distinctive colors.

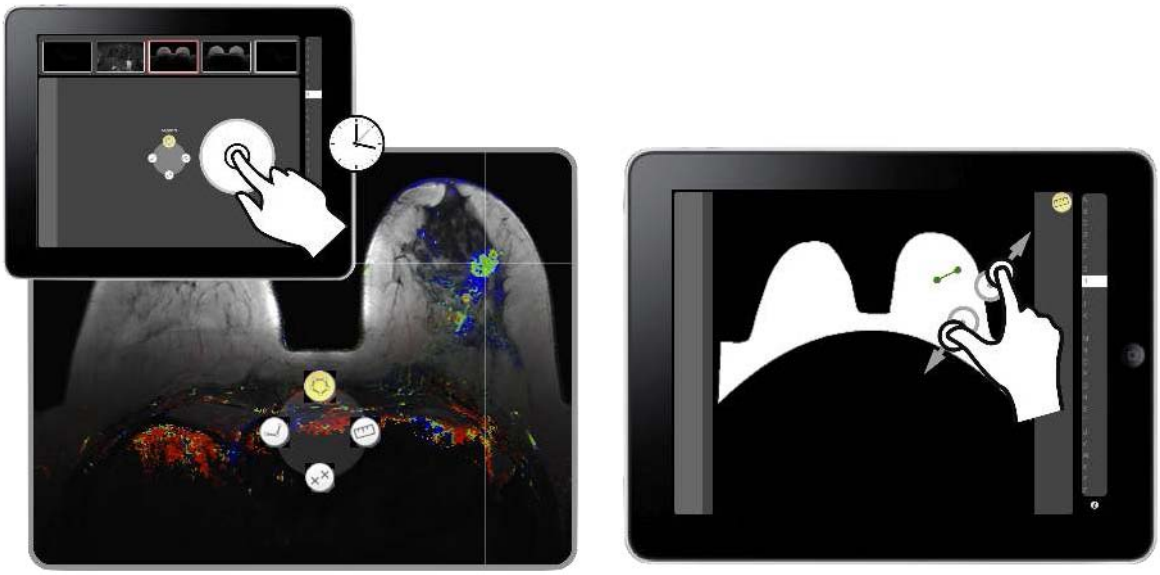


Figure 3.5: Application and tools. Left: a two-monitor setup is operated by the mobile device. Middle: Bringing up the context sensitive circular menu with a tap and hold gesture. Right: The iPad screen while a measurement is performed. Breast shape shown for orientation.

Image series thumbnails in the middle can be previewed in larger size on the diagnostic screens together with additional information. To select a series for diagnostic reading, the series thumbnail is double-tapped, which leads into the reading workflow. The screen on the mobile device changes to the appearance in Fig. 3.4 (right), where the predefined workflow steps (hangings) are indicated with icons on the top, while the rest of the space is left for gesture-based interaction. Preconfigured workflow steps are then executed swipe-by-swipe.

Navigation

Breast MRI diagnostic workstations usually offer viewports to show different aspects of the data, statically arranged on one or more monitors. The user may change the layout, or zoom viewports to fill the screen and interact. This is not feasible in our system. Instead, we defined a number of viewport arrangements (hangings), where always one viewport is shown in larger size and takes all input (the master viewport), and all others support the reading with additional information. These hangings are then executed in sequence.

Navigation and interaction is done using gestures, and all gestures always apply only to the master viewport. Other viewports providing additional image data, derived data, orthogonal projections of the master view etc. are continuously updated to match the position on the master viewport.

Figure 3.6: Gestures. From left: (1) One finger moves selection. (2) Two fingers swipe to stack through images, move through time points, dim color overlay. (3) One-finger tap-and-hold: bring up context tool menu. (4) Two-finger pinch: measurements.

Gestures are composed of a number of fingers used, and a pattern of how the fingers move. We have designed the gestures such that frequently used functionality is easier to access, and

that accidental triggering of harmful actions is avoided. Also, gestures are used similarly across contexts.

The following gestures are implemented currently (cf. Fig. 3.6):

One finger tap and move Moving one finger around on the iPad touch screen navigates on the current master viewport. Orthogonal projections follow. For a MIP, the slice of origin of the maximum intensity is selected, and the orthogonal projections shown for this.

One finger double-tap and move A double tap, where the second tap is not released, enters Window/Level control. Sliding the finger up and down increases or decreases brightness, left and right control contrast.

One finger tap and hold Placing one finger on the touch screen for more than 500 ms brings up the context-aware circular tools menu. The finger can then move into the direction of one of the tools, and when over it, the tool icon is colored to indicate successful selection upon release of the finger.

Two finger swipes Swiping with two fingers in parallel up-down moves through the image stack. In dynamic series, additionally the swiping of two parallel fingers in left-right is supported to navigate through the time points. In series with color overlays, the same left-right swipe controls transparency of the overlay.

Three finger swipes Parallel swiping of three fingers swaps between steps in the workflow (hanging protocol). Any hanging can directly be accessed by selecting it in the hangings bar above the touch area — either by pulling it into the highlighted center, or by double-tapping its icon.

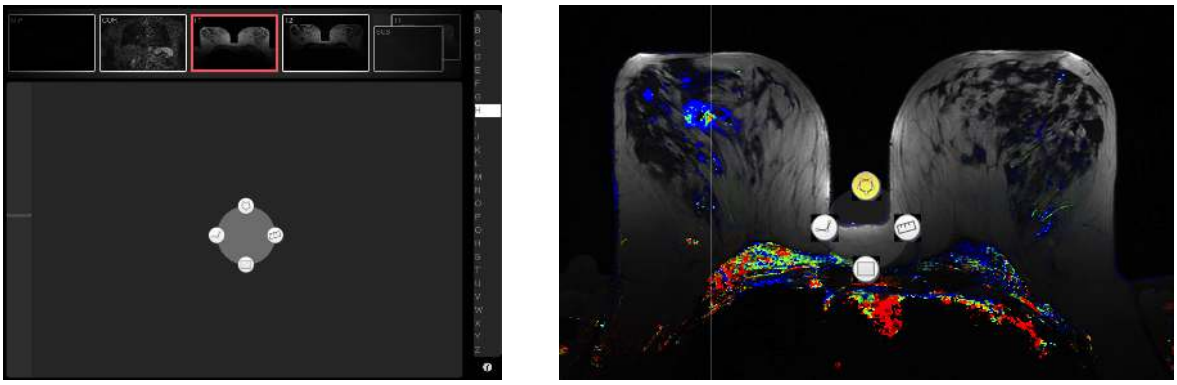
Five finger tap A simultaneous tap with five fingers saves all results when in a tools mode and leaves the tool, or finish the session when in default review mode. In this case, the application returns to the series selection screen.

This set of gestures can be adapted to further applications, e.g. mammography and tomosynthesis screening workflows. Regarding tool support during the workflow, all applications will share some common properties, as follows.

Tool Menu

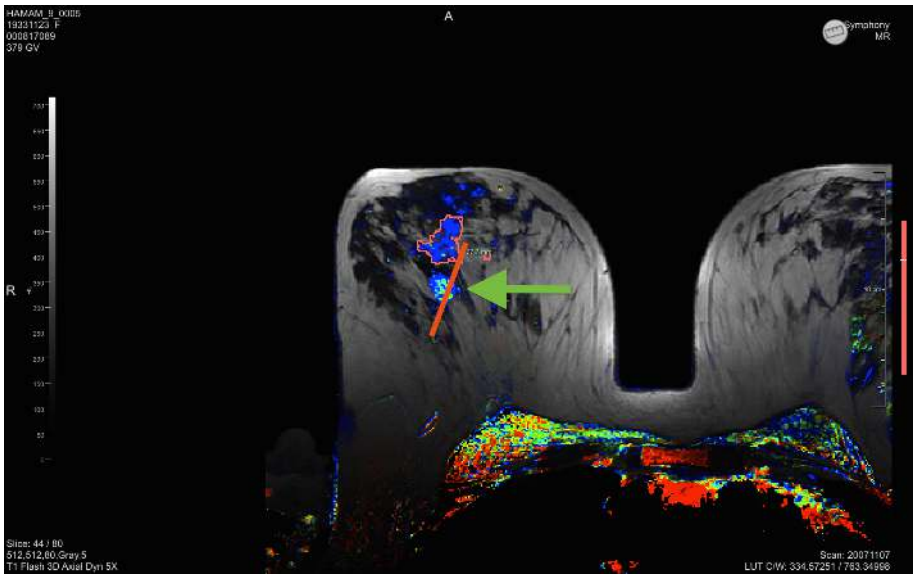
During each workflow step, tools are only offered when they apply to the reviewed series, and to the currently selected location. Our segmentation algorithm, for example, requires a subtraction image with sufficient contrast near the seed point. We thus propose to offer the segmentation tool only if the preconditions are met, i.e. if at the selected location in the series, the subtraction image shows gray levels above a preselected threshold. For other tools, similar context-aware aids can be provided by calculating image maps on which regions of potential tool availability are stored.

Bringing up the tool menu is done by touching and holding a point of interest. A circular menu around the finger on the touch display, and around the selected location on the display screens will show all options that apply at this location, e.g. measurement, annotation, and segmentation tools. Once a finding is segmented, it can be annotated. In any case it will be indicated with a color mark both on the screens and the touch display. Also, it will be stored in the case database, and displayed with a graphical icon in the patient overview screen. You can always navigate to the image slices corresponding to a finding by clicking the color bar. Of course, all views are immediately synchronized.

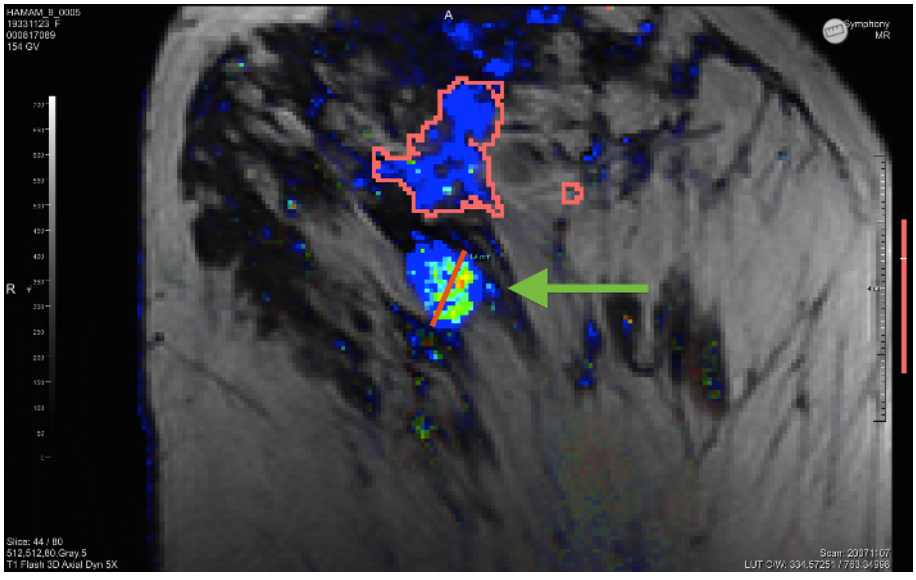


(a) (b)

Figure 3.7: The tool menu, shown on the iPad (left) and the monitor screen. On the monitor, the chosen tool is highlighted when the user slides the finger onto it.



(a)



(b)

Figure 3.8: Measuring the size of a target structure (green arrow) using the two-finger pinch gesture. The top image shows the state when the user started measuring. Zoom is then automatically increased when the measurement line length (orange line in images) shortens, like it is the case in the lower image. The red outline indicates the result of a priorly conducted automatic segmentation.

For size measurements, we implemented a two finger scale gesture that anticipates the desired size and zooms the images automatically to enable the precise measurement even of very small structures. When the size measurement mode is entered, two fingers on the iPad screen set the initial position of a ruler on the default viewport. After that, all movements and rotations the two fingers do on the iPad screen are converted into relative updates of position and rotation of the monitor ruler. By this, it is possible to refine a measurement iteratively, until it matches the desired start and end points. When small areas need to be measured, two fingers on the initial zoom setting cannot always be moved close enough together to perform the measurement. Our proposed remedy is to increase the zoom level depending on the length of the measurement line.

For reporting of finding locations, one finger interactively indicates the location of interest, and from precomputed locations of chest wall and nipples, the shortest distances are annotated and indicated. The segmentation of suspicious areas can be triggered from the same tool menu. An implementation of the smart opening algorithm (J H Moltz et al. 2009) with adaptions to the specific situation in dynamic contrast enhanced breast MRI takes the indicated position as the seed point for the region growing algorithm that is the first step in smart opening. After the segmentation, the result can interactively refined to fit the segmentation boundary optimally around the details of the area. Since only one parameter controls this, it is convenient to offer a one-finger gesture up and down on the iPad to control it.

Results of annotations are also visualized on the monitor next to the active viewport. A segmentation result is a 3D structure, hence its extent is visualized right to the image stack in a color that corresponds to the outline, and with a bar of a height that corresponds to the fraction and location within the 3D image stack (compare Fig. 3.8). Such an indicator has sometimes been called an elevator view, a denomination we also adhere to. Line measurements extend on one slice only, and are visualized on the same elevator view with a horizontal thick line. The current position on the stack is indicated with a white horizontal line.

Implementation Details

As briefly stated before, our implementation divides the responsibilities of iPad and server computer in a way that guarantees easy portability of the mobile device implementation while allowing to custom-tailor the specific gesture-based control only once, and centrally. We have also kept in mind the extension of the core prototype presented in this thesis both in terms of covered imaging modalities, and in terms of ubiquitous clinical pervasiveness regarding the coverage of locations.

Looking at the collaboration of devices, as depicted in Fig. 3.9, the intended technical setup will be elaborated. The Image and Authentication Server is a central instance storing the database of user access rights, and has access to the radiological (or any other) image database, for example the PACS. The four subunits in the figure symbolize working areas of radiologists, in which display computers connected to the central computer provide wireless connectivity and may store image data locally, for example the most recently acquired patient images in the radiology reading room, or the data of the patients of a patient room. In particular, the display computers are equipped with monitors or other display capabilities, like projectors.

Our scenario foresees for such display computers that they show a QR code the mobile device can read in using its camera. The QR code encodes the type of the display and the network location of the client computer, so that the mobile device can then negotiate the user access rights with the central authentication server via the display computer serving as a wireless access point towards the mobile device. Alternative configurations of this initial connection step are possible, for example the QR code may be placed besides the monitors on a wall, or Near Field Communication (NFC) tags may serve the same purpose. Displaying the QR code on screen, however, prevents fraudulent manipulations of the code in publicly accessible areas of the hospital.

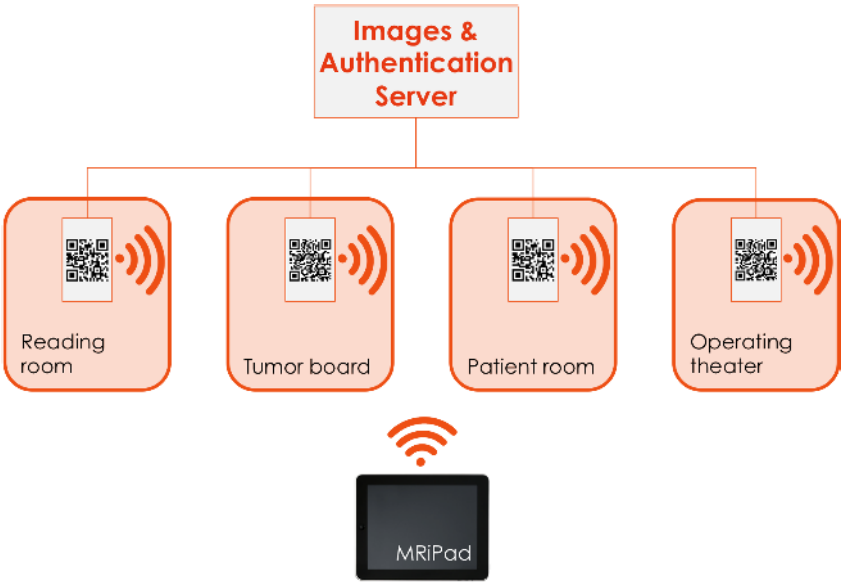


Figure 3.9: Device collaboration. Explanation see text.

During the initial negotiation, details about the user (the owner of the mobile device) will be sent to the authentication server, which will answer by providing a list of eligible functions tailored to the user's authentication level and access rights. At the same time, a list of patients depending on the location of the mobile device will be provided. The mobile device can then use the two server-provided pieces of information to configure the list of patients, and most importantly the functions the user can access using the user interface of the mobile device implementation.

Display computer and mobile device communicate based on a data stream over a TCP (Transmission Control Protocol) socket connection, defined by an IP (Internet Protocol) address and a port number. This network address is encoded in the QR code.

The messaging protocol between the mobile device and the local access point equipped with monitors has intentionally been kept as simple as possible. Using a web socket connection, both sides can send messages consisting of a message type identifier (ID) and a data part. This can be used to send a simple message, for example the mobile device may request the display computer to bring up a certain patient, or to advance one step in the workflow. The particular case where image data needs to be sent from the display computer to the mobile device is handled in a small protocol. Firstly, the mobile device sends a message to the display computer with an ID signalling the image request. The display computer, implementing the server side for this exchange, will in the second step send a message back containing the image ID and the image data in one message. For this, the image data can either be sent in raw format, or using JPEG compression.

3.3.3 Results

Our work has the aim to improve the workflow in breast MRI reading. Improvement can be made in various categories like speed, comfort, and usability, and in diagnostic image reading also in accuracy or reader confidence. The work we present addresses the first aspects most: usability and efficiency. Our belief is that increased intuitiveness complemented by a dedicated support for the typical steps in breast MRI reading will increase efficiency alongside.

However, evaluating a prototypical implementation of a reading workstation is difficult. The different levels of training with the old and the proposed environment may hamper a meaningful comparison of reading speed; likewise our prototype has not been performance-optimized. Also,

features of the proposed setup that improve convenience completely lack in other workstations, like for example the quick preview of all of the patient's imaging previews from mammograms, ultrasound or biopsy, together with their reports.

Therefore, two general evaluation paths are feasible:

1. Evaluate self-contained subsets of functionality that are available in both setups. Examples are tools like measurements and annotations. Those evaluations can be of a qualitative nature, assessing participant's preferences, or of a quantitative nature, rating implementation alternatives against each other in terms of speed, accuracy, or other metrics.
2. Evaluate the user responses to the proposed developments in a systematic fashion.

We have followed both directions. In a first qualitative clinical study based on a questionnaire, the aim was to assess the acceptance of the workflow based approach, and compare it with the workstation based routine examination of similar cases. Detailed experiments to assess accuracy and speed of the individual tools need to be conducted subsequently.

This answers the most critical remarks, which focused on the speed that can be achieved by using touch interaction instead of a mouse, demanding for experimental performance figures compared with special keypads, and of course mouse and keyboard. The second most critical remark concerned the accuracy, given that fingers always touch an area rather than a point. Both aspects have been addressed in an experiment focusing on the integrated zoom-and-measure gesture we will outline in the next section.

Qualitative Evaluation

For our qualitative evaluation study, four experienced breast radiologists in the breast center of a community hospital in Boca Raton, United States of America, have been asked to accustom themselves with the general concept and facilities the prototype offers. The system was set up on a MacBook Pro connected to one external monitor which provided the principal viewing space. An iPad Mini was used in the study. After an introducing demonstration, they were guided and supervised in the reading of one to two patients, before they were asked to review two cases on their own. Their actions were recorded with a digital video camera for later reference.

When they have completed their cases, we asked them to fill out an electronic questionnaire (see Appendix B.2). The questionnaire assessed their general preferences and opinions regarding mobile devices in private and work use. An extensive set of specific questions targeted the intuitiveness and possible inconveniences of the tools that have been proposed above.

All four radiologists equivocally stated that they own and use a touch screen mobile device very frequently since more than one year. Equally, they consent in their wish to be offered tools for these devices they could use professionally. Asked, however, which applications they know of or use in their professional environment, they either know none, or don't use those they do know. One participant mentioned the mobile version of the Mammography Reporting System (MRS), which was discarded quickly due to failures. Asked for their preferences in size, all felt the iPad Mini provided enough space to conduct the tasks they were given.

For the application, we first evaluated the patient overview screen (Fig. 3.11). The overall ratings spanned the range from medium to top score (avg. 4.0). Those who used the preview function that displays a blow-up of the thumbnail and the corresponding report tended to give higher scores, also for the preview function itself (scores of 4 and 5). The graphical findings indicator was judged by three participants averaging at 3.67. One participant remarked that a chance to start reviewing images of any modality from this overview screen would be desirable.

Turning to the MRI reading part of the prototype, we began with questions pertaining to the overall evaluation of the gestures. Like the more specific rating questions later, a scale from one

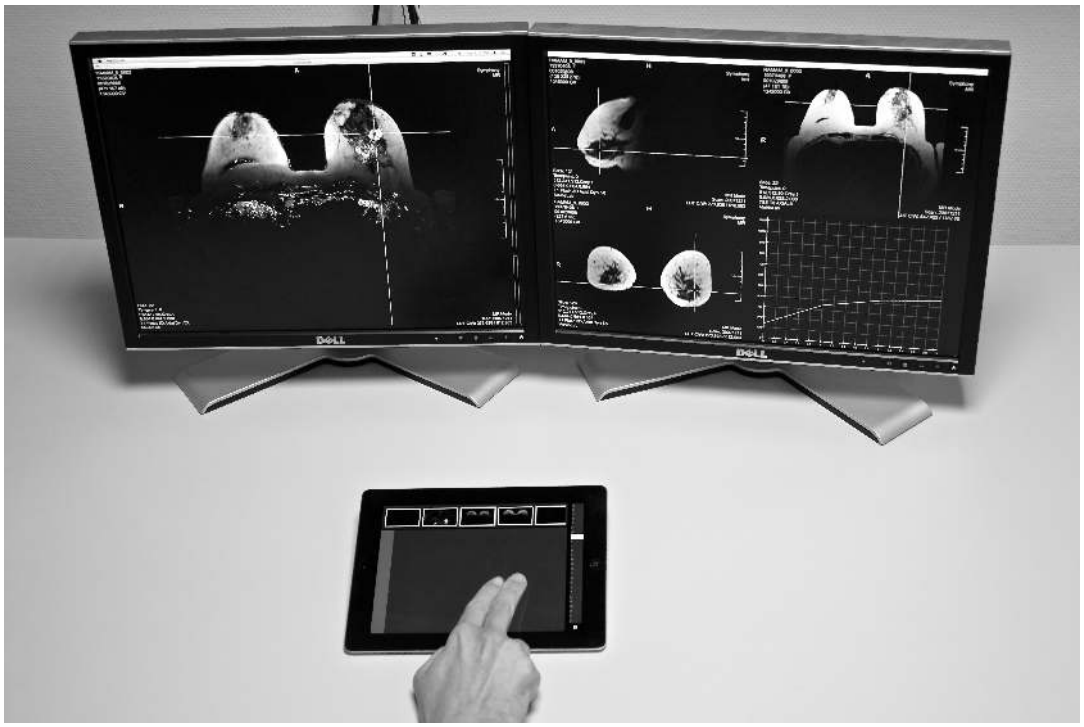


Figure 3.10: iPad MRI workflow prototype during interaction demonstration.

(worst) to five (best) was used to record satisfaction. All participants rated the workflow based approach with four or five out of five points, though only half of them prefer the workflow based approach to a more conventional workstation where tools are available all the time, everywhere, and a set of different MRI contrasts are visible to choose from.

Asked about the gestures in general, the participants found it moderate to easy (average scale value of 3.75) to remember the gestures after two training cases. No radiologist experienced difficulties to execute the gestures unambiguously, even those where they stated that they need training to execute them correctly. The most important comment was remarked by two participants who noted problems executing the two-finger stacking gesture. The reason was identified to be the setting of the maximum allowed distance between the fingers to be detected as a parallel swipe.

Next we reviewed opinions on the general workflow concept. Each facility to move through the hangings in the workflow was chosen by two of the four radiologists, using the three-finger swipe, or using the workflow step thumbnails. Except one participant, all deviated from the hanging protocol that was implemented in our workflow. Their reasons were either that the protocol didn't match their personal preference, or that the case required to switch to a different contrast after reviewing the images in a particular hanging. It was suggested to transport 3D positions between hangings, so that each new hanging continues where the old was left.

Lastly, the volunteering radiologists were asked to assess all tools they have used.

Segmentation. The segmentation tool has been used by two of the users. They rated the tool with four and five points, respectively, and indicated they were able to segment the target structure as they wanted. One participant emphasized the precision of the segmentation and the value for the description of morphology. The possibility to change the fitting parameter was appreciated. It was suggested to derive size and shape parameters from the segmentation automatically to reduce the manual work in reporting. Overall, there seems not to be much controversy about the desire to have a powerful one-click segmentation tool available.

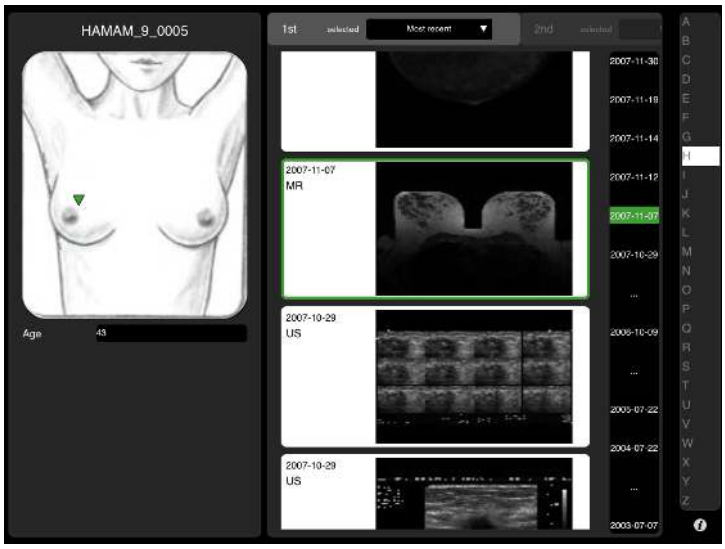


Figure 3.11: The patient overview screen features a graphical finding indicator, a list of thumbnails for all identified studies of different modalities, and a time line (left to right).

Distance from nipple. Three users rated the tool, all with four points out of five. The tool worked as expected, and delivered the desired information. The shared wish was to save the measurements with the lesion, again aiding the reporting. One reader commented that a sagittal plane for a second measurement would be necessary. Also, they wished for a screen capture function for documentation. Again there is little controversy about the utility and operability of a distance measurement tool that is supported by computer-detected chest wall and nipple position.

Size measurement. All four radiologists evaluated the size measurement tool. Two liked it very much, the other two gave four and three points. The three top raters subjectively found the gesture based measurement superior to mouse operation in terms of accuracy. In terms of speed, three raters believe the mouse is the faster option. We also asked for the ease of operation of the gesture versus the mouse, but here the opinions were indecisive, which is also the case for their personal preference. Still, we received very encouraging comments, even of those who prefer the mouse. Of all gestures and tools, this is the most controversially discussed.

The results are also reflected in the following table, where an overview is given on how intuitive and operable the gestures for various actions were rated to be.

The participants had room to leave us comments that were not covered by our questions. Besides the hope that tools like the presented might become available for general reading, in particular where many measurements are required, one comment indicated that a much larger touch screen, providing space enough to rest both hands comfortably on them, might improve the natural and intuitive operation, and provide more space for gestures that use both hands, or gestures that apply to more than one monitor.

Quantitative Evaluation

From the qualitative evaluation and many additional comments we received, it was obvious that a quantitative description and comparison of aspects of the prototype will be required to substantiate our claim that gestures can compete with mouse-based interaction.

It was as well obvious that a quantitative evaluation of the overall speed — leave alone diagnostic accuracy — of a gesture based system compared with a market-available breast MRI reading workstation is out of scope; there are too many free parameters that we would not be

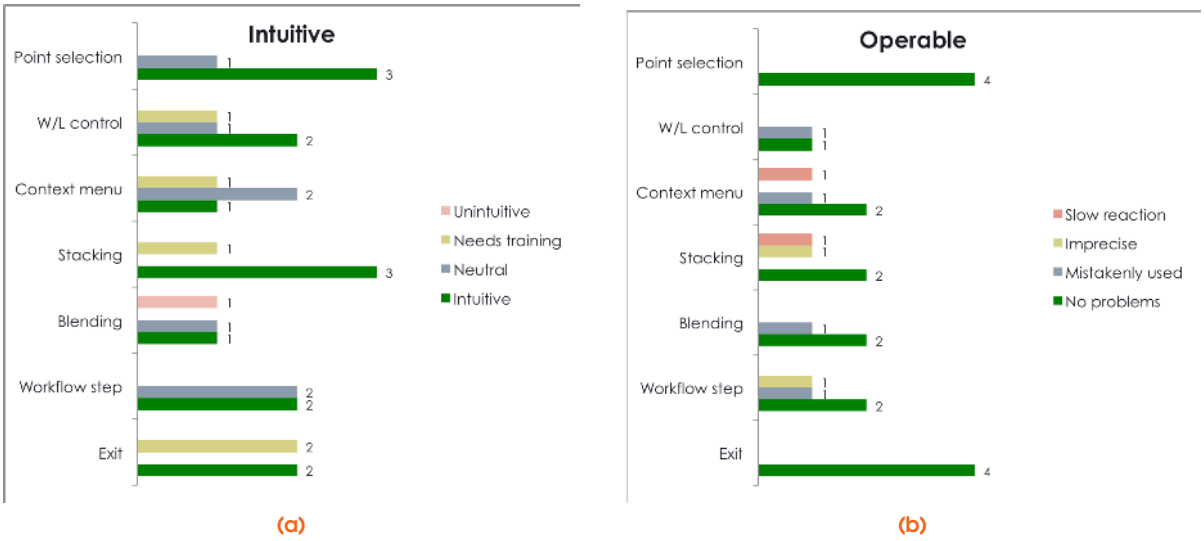


Figure 3.12: Summary of gesture assessments for intuitiveness (left) and operability.

able to control. We consider it a fundamental problem to evaluate prototypical developments like the one proposed here in a formal setup, and as a whole. Many of the most interesting aspects of it require sophisticated setups both from the perspective of study design, and from the technical standpoint.

- To compare the speed of the workflow-based reading approach as opposed to the "flat workstation layout" approach, a reading task needs to be defined that has a clear endpoint. One such endpoint may be to fill a standardized report. This, however, would quantify the suitability of tools that support report generation more than it assesses the workflow concept.
- To measure the intuitiveness of gestures instead of mouse/keyboard to perform the reading task would be possible if a novice user unfamiliar with both interaction techniques were to learn both, and perform a task with both. It is very unlikely, though, to find volunteers who have no prior experience with a mouse. Training with the gesture based system may reduce the bias towards mouse operation, but a carefully controlled study that also assesses the bias is required to obtain significant results.
- In clinical applications, another quantitative measure may be the diagnostic yield of the compared approaches. Since this requires experienced readers and a very well defined data set, this is no suitable option for our evaluation purposes in an early developmental stage.
- The most convincing studies will instead be those that test only very well-defined functions of the application, such as single tools, that can be isolated and can be applied to tasks that don't require domain knowledge.

In the light of these considerations, a quantitative study evaluating the size measurement gesture, which is combined with the zooming functionality, has been designed. A mouse-operated zoom and measure function has been implemented to compare it to the gesture-based measurement. To make the task independent of domain knowledge, all 45 target areas to measure were unambiguously visible in 2D binary images. Also, the 20 study participants had 30 training cases available to accustom themselves with the provided tools. The target structures had known diameters, and besides the deviation from these diameters, the distances of the measurement line endpoints from the structure borders were recorded and added into the overall precision error. In addition, the time to perform the measurements was taken. The recorded data was

analyzed using ANOVA[5] and Tukey's HSD[6] post-hoc test.

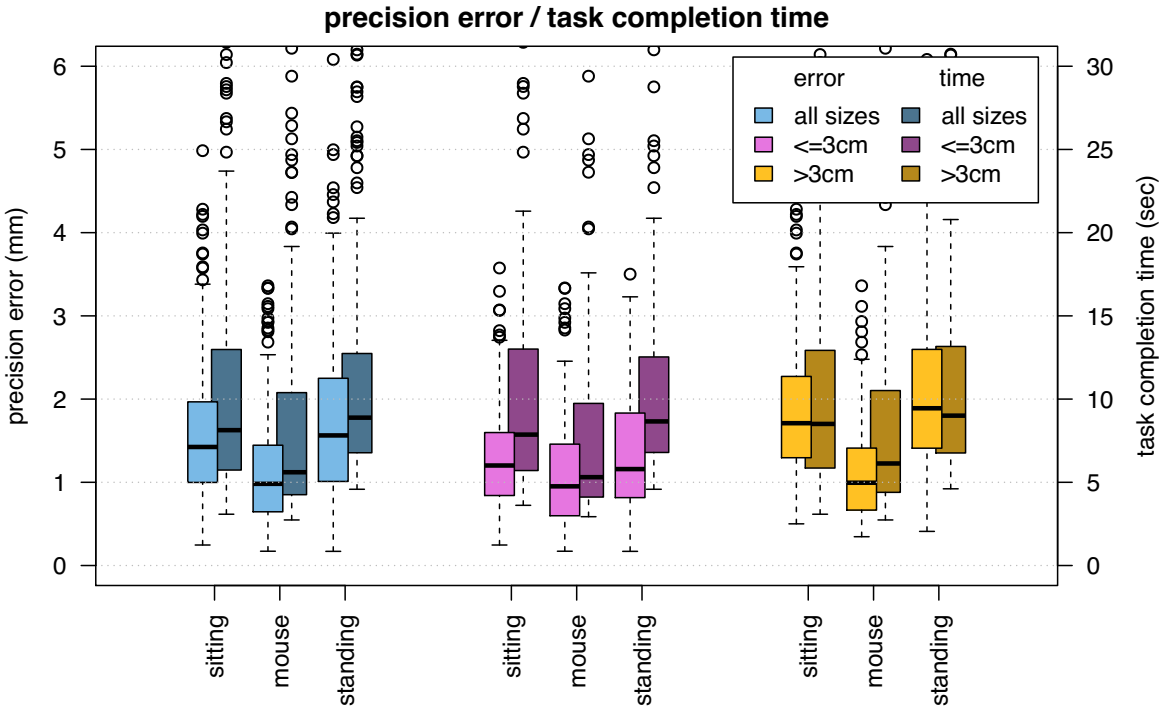


Figure 3.13: Error in size measurement (precision error) and task execution time for three sizes (colors), and three operator positions. From: RITTER et al. (2013)

Although the participants achieved a significantly higher precision using the mouse-based interaction than they did with the multi-touch gesture (see Fig. 3.13), the clinical relevance of the difference is not given. This insight can be gained by comparing the average precision errors of the multi-touch gesture in both sitting and standing operation of the touch device with the precision error of the mouse-based measurements, and by further setting the difference between these precision errors in relation to the typical in-plane voxel sizes of breast MR images, which are the objects in our study. While those images are typically of about a 1 mm by 1 mm in-plane resolution, the differences of averaged precision errors mouse vs. gesture are in the order of half millimeters or below. While not analyzed here, the differences in small objects is even less, since the combined zooming and measuring in the gesture-based tool apparently aid the volunteers. Hence, one common counter argument to gesture-based operation can be invalidated based on our data (RITTER et al. 2013).

The second aspect we were interested in was the performance of a gesture in terms of speed. Again, our volunteers were significantly faster to complete the measurements with the mouse, regardless of the size of the target structure. The difference of about 20% may be unbearable in applications were measurements are the main task. In breast MRI reading, however, the precision of the measurement is more decisive, and the amount of time spent on measurements is probably negligible against the total time it takes to review a MRI examination.

Also, in the study the choice and selection of the measurement tool was not included in the time measurements. The layout of typical breast MRI workstations may diminish the time performance of mouse based measurements easily to below the level of gestures by simply not providing a fast enough access to the tool.

[5] Analysis of variance
[6] Honestly significant difference

3.4 Flexible Mammography Screening Using the Multi-Touch Approach

The best proof for the claim of adequacy of a multi-touch based interaction setup will be the application to one of the most optimized scenarios in breast imaging, if not in clinical imaging at all: mammography screening based on full-field digital mammography (FFDM). In the following, a system design is presented that emerged from the study of mammography reading in the pre-digital era. While no rigorous comparative study of reading performance is conducted in terms of speed and detection rate, a study design for the evaluation of the prototypical implementation will be presented that allows to conduct limited performance quantification without the requirement to recruit expert mammographers, and without the need to prepare a careful case selection.

Since a combination of observations in the handling of modern digital mammography screening workstations and knowledge on the film-based screening workflow have led to the proposed design, we will in the following review not only the current state of the art in mammography screening, but also take a look back and put a spotlight on features of the film-based screening that were not transferred into the digital age due to limitations in the hardware.

3.4.1 Mammography Screening Past and Present

X-ray mammography has a long history, dating back to 1913 when Berlin surgeon Albert Salomon took x-ray images of huge numbers of breast mastectomy specimen, finding that they revealed details of the pathology he saw in histological images (Mukherjee 2011). It took decades before this led to the introduction of mammography as a tool to examine women's breasts in vitro with the aim of searching for the presence of suspicious findings. Film-based mammography, also called film-screen mammography (FSM), required to process the films before they could be assessed. Assessment was done using light boxes, large back-lit wall-mounted casings with a provision to attach the images to it.

In population studies, several projects aimed to find an optimal setup for the early detection of breast cancer using x-ray mammography. The cancer yield of the technology had to be weighted against the possible harms to a woman that the repeated x-ray may cause. This is a discussion that has not come to an end today, since the costs of population-based screening are increasingly debated in addition to the doubts regarding cancer yield raised by critics. Among the cancer screening advocates, it is generally agreed that annual bilateral two-view mammography screening of women aged 40–70 is the optimal setting that guaranties highest yields at lowest risk. It has to be noted, though, that mammography has an inherent blind spot spoiling its utility: it fails to show small cancers in high-risk populations, like young women with dense breast tissue. These cancers, however, are known to encompass the most aggressive breast cancer subtypes. For this reason, modern research tries to stratify women by their breast cancer risk factors, and to submit them to the optimal breast cancer screening program for their individual risk and morphology.

The positive effect of population-based mammography screening manifests itself as a reduction in mortality which is well-documented in countries with a long tradition of screening (Schoor et al. 2011). This improvement was brought about by a controlled increase in the referral rate, leading to an increase in cancer detections. The lower positive predictive value alongside rising costs associated with the more profound workup of the added number of detected suspicions has been acknowledged (Timmers et al. 2012). Extrapolating the performance figures of the screening program, where only breast cancers found in screened women are recorded, to the whole population based on Swedish numbers leads the Dutch National Evaluation Team for Breast Cancer Screening to assume a reduction of the risk of dying of breast cancer by 40% for screened women compared with women not participating (Heeten et al. 2009). In other countries, the same success is seen, for example in Germany, where screening has been implemented after the

model of the successful dutch screening program (Becker et al. 2008; Heywang-Köbrunner et al. 2008).

Also, a critical perspective has been taken on the effectiveness of mammography-based breast cancer screening to achieve a reduction of breast cancer induced mortality. Opponents of breast cancer screening often argue that the 0.05% absolute reduction in breast cancer mortality (corresponding to a relative decrease in mortality of 15%, Gøtzsche et al. (2013)) achieved by population screening for example in Norway require too many unnecessary biopsies and may result in the overtreatment of breast alterations like in-situ neoplasms that would without detection remain indolent or even regress spontaneously. These numbers are for several reasons misleading. Firstly, a relative reduction in breast cancer mortality of only 15% is widely surpassed in countries implementing a population based screening, like for example Sweden, Great Britain and the Netherlands, and also in Germany numbers of 25% and above are assumed (Becker et al. 2008). Secondly, converting the relative risk (the risk to die of breast cancer when it is detected) into the absolute risk is practically blurring the decisive facts for individual women who are much more likely interested in ways to improve their odds in case they are affected with breast cancer in their life.

Another example of misleading interpretations of breast cancer screening facts and figures is found in the studies of A. B. Miller et al. (2014). This group recently reported on a 25 year follow-up on the results of the Canadian breast cancer screening program, concluding that screening cannot be recommended over physical examination. There has been a harsh criticisms regarding the validity of the data decades ago, when the basis of this data was revealed (Kopans 1990). Also after the recent publication, it has instantly been criticized by the American College of Radiology as being "an incredibly misleading analysis based on the deeply flawed and widely discredited Canadian National Breast Screening Study (CNBSS)." (Farley et al. 2014).

Likewise, the 2009 recommendation of the US Preventive Services Task Force (USPSTF) created an outrage among breast care experts, when the USPSTF suggested that based on the available evidence (including the CNBSS study) women should be screened by mammography only from the age of 50 and only biannually (US Preventive Services Task Force 2009). This recommendation was vigorously debated for various ethical and technical reasons. Arguably, this is the most controversially discussed topic in image-based breast cancer detection..

Mammograms are usually acquired in two projections per breast, CC (cranio-caudal, or axial orientation), and ML/MLO (medio-lateral/oblique, or sagittal orientation, potentially tilted to visualize the axilla). This so-called 4-view screening mammography study is then compared with the prior exam (last year's study) to detect temporal change.

In the pre-digital scenario, all images were taken on film and processed. The processed images, for example 24×30 cm of size, would be hung up on a light box, similar to x-ray examinations of other organs. In population-based screening, however, with a manual image setup process like this, the amount of mammograms to assess per day, would have led to long wait times between two cases, which is impractical. This clinical requirement prompted the development of machines that provided capacities and mechanisms to store the physical mammogram films of a batch of screened women at once, and then move them case by case to the viewing area of the light box. While one study was being read, the next woman's images were already available within the machine, and upon the press of a button, the current case would be cleared from the screen, and the next case moved to the display. The machine was called an alternator (cf. Fig. 3.14).

In the late 1990s, the transition to digital mammograms became a more and more common topic. Many advantages were seen: digital images can be stored, transported, duplicated, manipulated, and organized in much easier ways than physical film mammographies. However, technical limitations of monitors and data processing mechanisms had to be overcome. The

Figure 3.14: Alternator used in pre-digital mammography screen reading.

image data for one single digital mammogram was enormous compared to contemporary computer's capabilities in terms of storage. Also, it required dedicated graphics hardware to drive the (likewise dedicated) high-resolution CRT (cathode ray tube) monitors. Next, reading 8 mammograms (four current mammograms and four priors) from disk or network required an extremely meticulous understanding of the data transfer protocols including network, hard drive, and the internal bus systems of the computer to achieve patient switching times that are feasible for screening routine, where sub-second times were expected for a "next patient" change.

Lastly, and most important for the work presented in this chapter, the workflow for the pre-digital light box scenario was such that it could not directly be transferred to the computer setup. In film-based mammography, all images, potentially even including several special views, were brought up at once — up to 12 images of sizes larger than A4. It was impossible to transfer this setup to digital displays directly, hence radiologists were observed to find out which images are examined in which sequence, and how long. A tool employed early in this process was eye tracking, and for example Beard et al. (1997) used such a setup to find out that most of the time spent in mammography reading is used in comparing contralateral images or ipsilateral images of prior and current visit, or the two projections of the ipsilateral side. Their study also examines several setups of workstation prototypes to display digital mammograms, but since they had no dedicated workstation available for timing comparison, their results remained to be validated. Their hypothesis was that the most crucial aspects are technical: when normal screening exams (ones without a finding) have to be loaded, displayed, read, and reported in about 15 sec, and suspicious cases have to be decided in less than 1 min, the throughput as well as the roaming and zooming tools have to be highly efficient.

More recent studies again asked if and why digital mammography reading takes longer than film screen reading, yielding inconclusive results. Pisano et al. (2002) found no significant

difference in reading speed at comparable detection performances, though their results indicated slightly faster readings of digital mammograms. They assign this unexpected speed-up to the comfortable tools available for roaming and zooming in the soft-copy reading workstation. However, Haygood et al. (2010) challenged this result, measuring significantly longer interpretation times on digital mammography compared to film screen mammography, also citing prior studies resulting in the same observation, for example the work of Berns et al. (2006), who in addition looked at the typical actions that led to prolonged reading times. Berns et al. observed that the image manipulation required on a digital workstation to emulate the hand-held magnifying glass when searching for microcalcifications is not optimally modeled. Since several randomized controlled trials have shown equivalence of digital and film-screen mammography, with evidence of a superiority of digital mammograms in dense breasts (C. H. Lee et al. 2010), the current recommendation is to screen with digital acquisition technique where available.

Concluding the comparison between film-screen and full-field digital mammography, there is apparently a potential to improve the speed and intuitiveness of digital mammography review by optimizing interfaces and interaction. When designing a novel interaction and interface, the objective is to achieve the fastest possible review of normal cases, and to ease the handling of those steps that are required in the assessment of abnormal mammograms.

Mammography screening today uses workstations with large high-resolution LCD monitors. The images are presented according to a hanging protocol, where each protocol step displays the images in a predetermined resolution and setup. Usually, an overview with the eight mammograms is provided first, before the current study is presented in contralateral setup of MLO followed by CC, and both in an additional 1:1 image pixel to monitor pixel resolution. Thereafter, comparisons of prior series with current series are provided, before the report screen is displayed. The important fact is that most of the workflow is fixed, which serves the purpose of a reproducible workflow in a hospital. Deviating from the norm is usually possible only using either the on-screen buttons and menus, which are often designed to be available at the mouse cursor location, for example with a right click.

To speed up the reading process in this setting, a special keypad is often part of the mammography screening workstation; one example is displayed in Fig. 3.15. It is equipped with dedicated buttons that are assigned the most frequently used functionalities, like "next/previous workflow step", "next/previous patient" etc. Also, the keypads sometimes offer additional buttons to which user-chosen functions can be assigned, like for example a preferred hanging of images. Such user function assignments need to be memorized and recalled when the function is desired during reading. In addition, workstations offer a set of supporting capabilities that roam the high-resolution mammograms automatically on the press of a button, so that all image voxels have been regarded in a 1:1 zoom ratio (one image voxel corresponding to one screen pixel). This is a requirement in some installations of population based screening programs. Any proposed change in the interface and workflow needs to assure this fundamental capability to ensure that all mammograms are fully covered. It is important to understand that recalling a patient is not done based on the first finding encountered, but for a re-evaluation of all suspicions. It is hence no option to speed up the reading by allowing to bypass some of the mammograms.

3.4.2 A Gesture Controlled Mammography Workstation

In line with the thoughts stated for breast MRI, we propose a changed workflow paradigm. While the inflexible way to review the images accounts for the most crucial drop in performance (Beard et al. 1997), in our proposed system the benefits of the hanging protocol based workflow are combined with direct access to predefined hangings. Further, the reading of priors for a current mammogram is facilitated by visually showing the two compared acquisition time points on a time line and supporting the rapid, yet comprehensible review of prior mammograms with

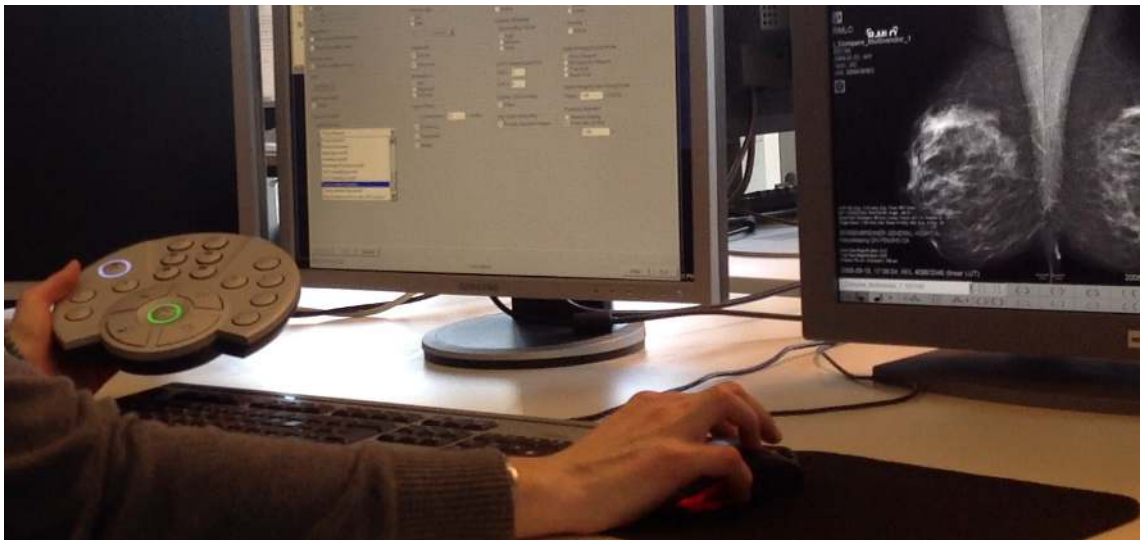


Figure 3.15: Special keypads for mammography screening can today be customized to suit an individual site's requirements.

a graphical representation of their transition. Given the constraints to the workflow depicted above, our system fuses the random-access capability, by which any predefined hanging can be pulled up, with a background tracking of seen mammogram areas.

Also, the inspection of mammograms by zooming and panning in the image will be shown with different approaches that allow for either a very natural action to magnify a mammogram, or a novel way of simultaneous zooming in two related mammograms, enabling a fast assessment of bilateral symmetry or current-prior comparison.

We show how a dedicated application that is based on the combination of a multi-touch mobile device with a prototypical mammography workstation enables these features.

Hanging Navigation

We assume that for patient selection and initialization of the screen reading a patient-centered browser is utilized, for example the one presented in the previous section. This contribution describes the deviating approach to the viewing of mammograms as opposed to the viewing of breast MRI. For diagnostic mammogram reading, it is of course possible to utilize all tools that have been described before, in particular the integrated measurement gesture to annotate the image.

To study user acceptance of a gesture-based approach to mammography reading, and to yield first performance estimates, a mammography workstation prototype with key features available in commercial screening workstations has been implemented. The features encompass a hanging-protocol-driven viewing workflow where each hanging presents a selection of the available images. It is possible to step through the hangings of the viewing protocol forward or reverse, and to jump to the next or previous patient. Importantly, the patient change time is much higher than feasible for mammography screening, hence a timed screening scenario cannot be studied with the prototype as it is configured today.

Employing multi-touch devices for mammographic diagnosis or screening is challenging because there is a long history of research into special keypads, which are frequently employed tools in this application field. Their shapes vary from vendor to vendor, and it appears that the one factor that makes radiologists like to use them more than mouse and keyboard, is their distinct layout with large buttons and controls assigned to unique tasks like "next patient" or "next hanging step". Some keypads have generic layouts that do not directly reflect the

mammography reading task consisting of steps through a hanging protocol based workflow, others are for example equipped with arrow-shaped buttons to indicate forward/backward steps.

Our hanging protocol is laid out based on a standard hanging protocol of a commercial software (Hologic SecurView®) and interviews and observations in mammography screening centers. It consists of an overview of all four current plus four prior mammograms, and continues with the back-to-back hanging of the current CC and MLO acquisitions. After that, comparison views of the current and prior CCs are presented in one hanging, and the same for MLO in another one. Next, comparisons of back-to-back priors of first CC, and then MLO are possible (compare the graphics in Fig. 3.16).

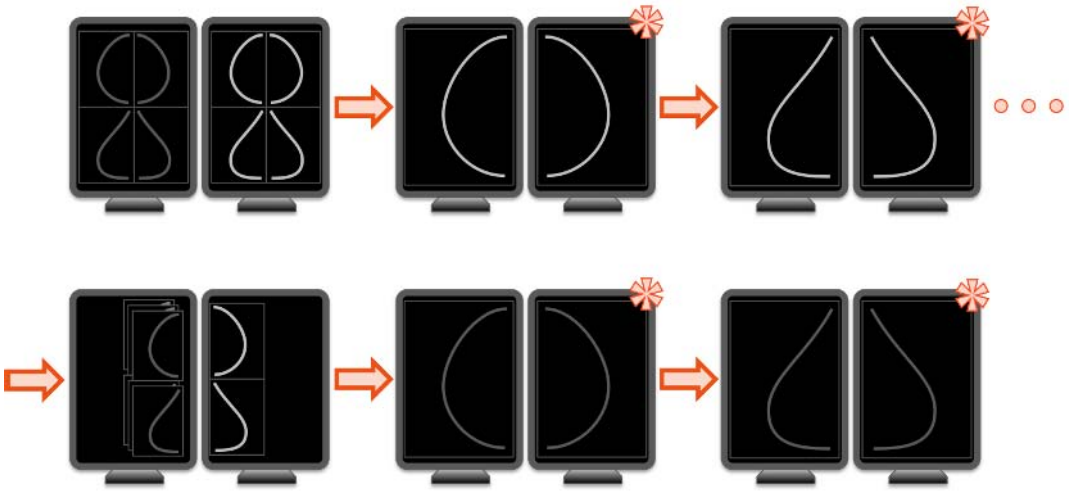


Figure 3.16: This schematic shows the steps of the hanging protocol that is implemented in the prototype. Darker breast shapes indicate prior mammograms, bright ones show the current mammograms. States marked with an asterisk allow for roaming mode.

Key Differences between User Interface Variants

Navigation and interaction with the mammography workstation prototype is implemented both for a mouse-and-keyboard paradigm, and for the multi-touch-based interaction using an iPad.

In the mouse-and-keyboard interaction pattern, key presses are assigned to the "next/previous step" actions for the hanging workflow. User interface buttons below the mammograms are assigned to the functions for "next/previous patient", and to advance in the roaming mode for a pair of images. The roaming mode has the purpose of showing the full mammogram (which is usually larger than what fits onto the physical resolution of the monitor in a 100% zoom level) in a 1:1 voxel:pixel ratio by dividing the mammogram into patches that fit onto the screen, and by allowing to iterate through these patches in sequence. Roaming is possible to be started from all back-to-back hangings, i.e. where exactly two mammograms are visible on the two screens.

Also, the application allows to magnify the mammograms interactively. A left mouse button click into the mammogram brings up a local magnifying glass in a 1:1 voxel:pixel ratio. Finally, four dedicated buttons in the user interface allow to access the four image pairs made of the current and prior acquisitions in the two projections. Finally, when prior acquisitions are available for a projection and laterality, they are visually indicated on the screen and can be selected with user interface buttons.

Using the iPad as the user interface, the interaction is mapped to gestures that are as easy as possible to learn and execute, with the aim to prevent the user from accidentally confusing gestures. A one-finger swipe is used for forward/backward stepping in the hanging protocol. Roaming is controlled with two-finger swipes. For magnifications, the combined zooming and panning which is well-known to touch device users from other applications is integrated.

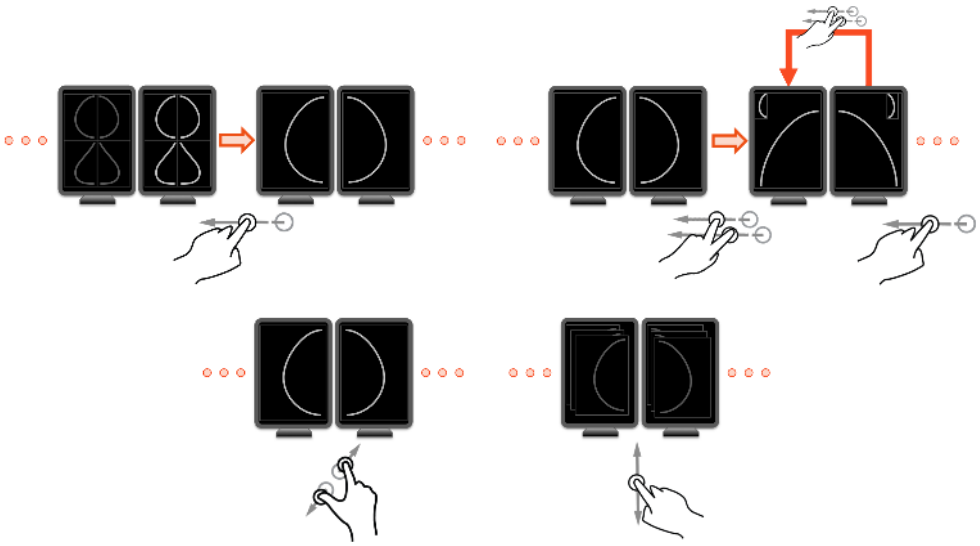


Figure 3.17: Gestures in the multi-touch based paradigm. Top row: Advancing in the hanging protocol using a one-finger horizontal swipe. Roaming in available hangings using a two-finger horizontal swipe. Bottom row: Zooming with the two-finger pinch gesture. Stacking through the history of prior mammograms using a one-finger vertical swipe.

"Direct access" to the four above mentioned image pairs is implemented by a tap gesture in one assigned quarter of the iPad. The quarters of the iPad correspond to the layout of the eight mammograms on the first hanging, i.e. the top left quarter corresponds to the prior CC mammograms (compare Fig. 3.16). Lastly, a five-finger tap or pinch gesture leaves zoom/pan and direct access modes.

While for the mouse-based magnification a conventional paradigm was chosen, the gesture-based magnification allows to magnify and pan two images at once with a mirrored pan direction in the back-to-back layout. This allows to inspect areas of interest in left versus right or current versus prior images. The extension to a correlated zoom and pan in two different projections requires to transform the panning amount with a function that for this purpose may be approximated depending on the angle between the two projection directions.

3.4.3 Evaluation

Very similar to the MRiPad prototype discussed before, the evaluation of the proposed approach is done quantitatively and qualitatively, for now with a stronger focus on the qualitative aspects. We have invited seven volunteers for the study described below. Out of these volunteers, one was a breast radiologist with experience in mammography screening, and one had prior experiences with the analysis of mammograms in computer algorithms. Most of the participants were experts in the field of computer science without mammography background, and all own a multi-touch capable mobile device.

While the population of volunteers is inhomogeneous and includes only one mammography expert, our experiments are still able to provide valuable results, since they are designed not to be a full reader study. Such a study would require carefully selected and well-documented clinical cases in large numbers, readers that participate in two temporally separated rounds of reading, and applications that are capable to provide all functionality the readers expect for a routine screening workup. In our case, we expect remarks about the design decisions with respect to the application domain, while the novice volunteers' performance can be assessed to gain insights into training time and the general feasibility of the implemented gestures. Their acquaintance with multi-touch devices and their average younger age makes their comments

valuable since it can be assumed that gestures will be rated with the background of existing standards.

A more evolved prototype incorporating the results of this evaluation will in the future allow a study to examine the iPad-based mammography screening approach in more depth. The prototype described here, however, is meant to demonstrate and test the general approach towards a more flexible and intuitive interaction and workflow design. Due to the fundamental novelty of the approach, it is required to assess the overall feasibility of design decisions, before a fine-tuned prototype is exposed to domain experts who should then rate it in comparison with a conventional mammography screening workstation. Quantitative evaluations, both regarding accuracy of detections and speed in operation of the two compared workstations are only then feasible.

For the purpose of semi-quantitative evaluations of task execution times in a population of non-expert volunteers, we propose a study design that is flexible enough to produce detection tasks which can be customized in their level of difficulty, so that any volunteer is able to fulfill the task. Our approach makes it easy to generate test images in arbitrary number.

Study Design

Each volunteer conducted two consecutive rounds of image analysis, in the first round using the mouse and keyboard, in the second using solely the iPad. Since we are less interested in the reader accuracy, we did not randomize the order of the two rounds among the participants. An iPad 3 has been used, providing a 9.7 inches display with a 2,048 $\times$ 1,536 px resolution (Retina display). The iPad was equipped with a stand that tilts its surface by approximately 15°.

In each round, eight sets of mammograms with varying numbers of prior images were presented to the volunteers. The task to find and report a sign we implanted into the digital mammograms had to be executed with the available interaction. We did not specifically ask the participants to read the images as fast as possible, but we measured their total task execution time for each of the two rounds.

The chosen task has been designed so that radiologists as well as lay people are able to fulfill it. In particular, no diagnostic analysis of mammograms is required. The sign to search for is a gray circle (compare Fig. 3.18) that has an opening into one of four possible directions, and upon finding the sign, volunteers were asked to describe the finding in a score form. The sign has been set in accordance with clinical scenarios:

- A new suspicion is discovered in the current mammogram, and it is not present in a prior mammogram. This corresponds to the screen-detected suspicion.
- A suspicion is seen in the current mammogram, and it is also visible, but potentially changed, in one or more prior mammogram. This mimics the task of watching a sub-threshold suspicion over the screening rounds.
- The current mammograms show no sign, but a sign is seen in a prior mammogram. This conforms for example with the situation when a benign breast tissue alteration resolves after some time.

Between the two rounds, the eight base images remained the same, but the implemented signs are located in other patients and other mammograms. All study participants were made aware of that, and in addition they were given a training case before the start for both interaction methods. They were asked to familiarize themselves with the workflow, user interface buttons, gestures etc. until they feel comfortable.

The time to analyze all cases was measured with a stop watch to gain insights into the performance. These timings are obviously limited in their significance since no participant

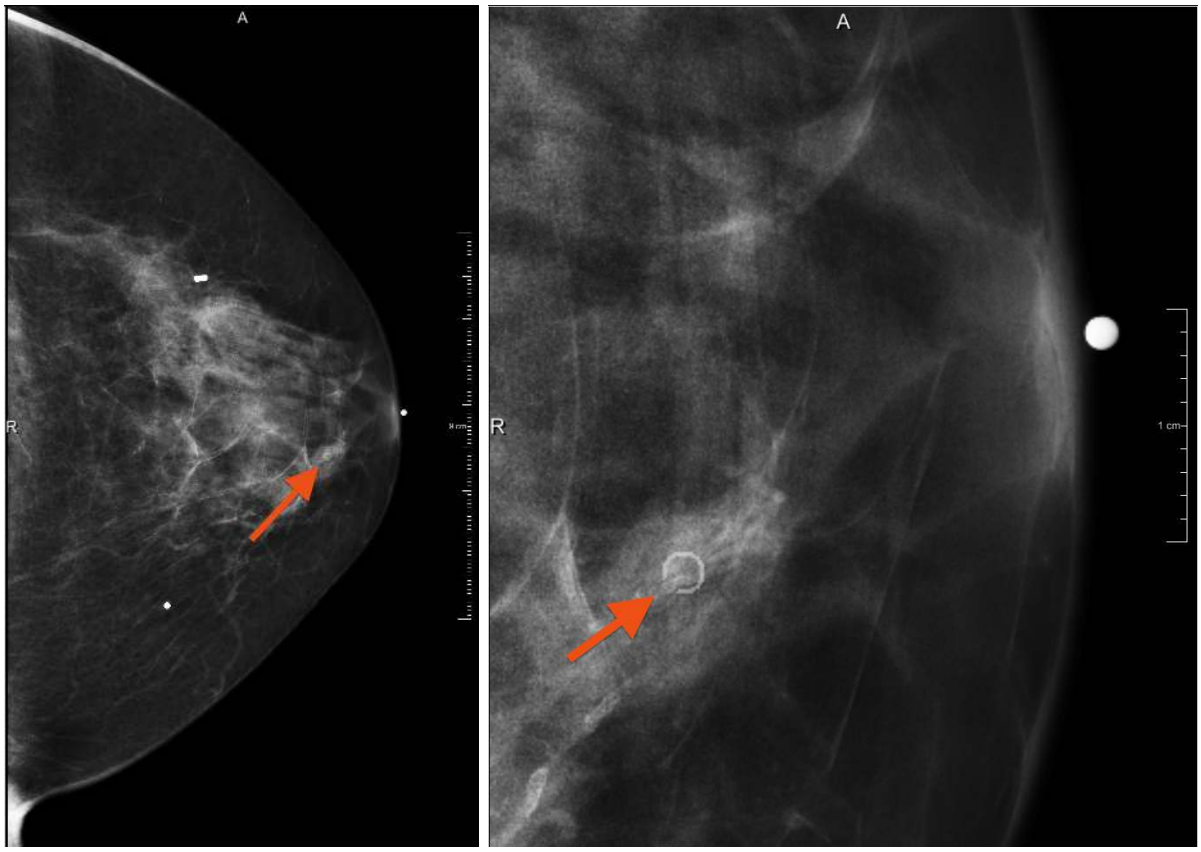


Figure 3.18: Example of a visual sign the volunteers in the study had to discover (left: in full mammogram, right: close-up). In the study, the contrast and location was chosen such that the sign was hard to discover and characterize in the overview screen to necessitate usage of the provided zoom tools.

performed the study twice (once mouse/keyboard first, once iPad first), but all participants saw the images using the mouse/keyboard user interface first. In addition, the screen layout of the buttons was not optimized, and no visual cues aided the participants. Instead, the function of each button was conveyed only by its textual label.

The basic results of the study were obtained using three online questionnaires designed with web-based forms and filled on a second computer available at the location of the study:

- A structured interview form evaluating prior experience with mobile and multi-touch devices. This form is filled after completion of both runs.
- One form each is filled alongside working through the cases, asking if a circle was found, and characterizing it: type, mode in which is was found (regular view, magnifying glass or pinch zoom, roaming mode).

All participants were able to complete the two rounds, hence no results had to be discarded.

Results

Quantitative Evaluation Our study participants had prior experience with mobile and multi-touch devices. All own a smart phone (a mobile phone with a touch screen and a screen size of above 3.5 inches diagonal), and three of them have more than three years of experience with multi-touch devices. Only two out of the seven participants use touch-screen devices professionally, but all stated they were open to use it for work.

From the timings taken during task completion, we note that the average completion time using the iPad-based gesture interface was about 20% below that of keyboard and mouse control

(19.4 min vs. 24.2 min), which has not been controlled for bias that can be caused by the study design, where mouse interaction was always performed before gesture interaction, perhaps leading to a training of the eye, and a better understanding of the task. Also note that particularly the button controls that had to be used in the keyboard/mouse setting were not optimized for speed and intuitiveness. The two most experienced participants had the fastest completion times, and the lowest difference between the two input methods.

Interestingly, the fraction of recovered signs was higher in the study part executed with gesture control, but in our study setup with only six marks in eight cases, this result lacks statistical power. However, in the gesture-controlled arm, participants appeared to prefer the pinch gesture for zooming over the quadrant roaming mode, while in mouse/keyboard interaction, the roaming mode was preferred over the magnifying glass tool, although it is more readily available by design.

Qualitative Evaluation As described above, we have paid attention to use gestures that have been utilized and validated before in consumer products, and that are as easy as possible to memorize. In fact, our volunteers found the gestures easy to remember with a mean score of 3.8 on a scale from 1 (hard) to 5 (easy). This applied to all gestures except the two-finger swipe gesture executing the roaming advance. Since quadrant roaming is a concept that is in its practical importance only known to mammography screening experienced radiologists, it can be speculated that the volunteers did not remember the functionality because they saw no meaning in the function triggered by it.

Asked for their preference between gesture and mouse/keyboard control, in none of the four categories “accuracy”, “speed”, “ease of operation”, and “personal preference”, the mouse/keyboard control obtained a higher score than the gesture control. Only the radiologist consistently rated mouse/keyboard superior, while the remaining participants were divided equally between the choice “gesture” and “indecisive” for all categories except “personal preference”. Here, the gesture control was preferred by five participants, and only one was indecisive.

All specific gestures have been evaluated using the same categories, plus “intuitiveness”. In general, the participants rated gestures to be easy to use, intuitive, and easy to remember more frequently than mouse/keyboard control, but also stated more often that training is required more than for mouse/keyboard. In terms of personal preference, many were indecisive, with a slight tendency to favor gestures over mouse and keyboard. This, however, can probably be attributed to the young average age of participants.

3.4.4 Conclusion

Overall, despite the small number of participants and a small number of cases each participant evaluated, valuable insights have been gained into trends in the preferences of users. Most importantly, our participant population suggested that gesture control may become a prerequisite of successful tools in the future. While individual design decisions in our prototype were rated ambiguously, the majority of participants still stated that they prefer gestures on a multi-touch capable device over mouse and keyboard. This inspires us to proceed in our work, and from our results so far, the main objectives for the next steps include some direct changes to design and implementation, and some more visionary concepts that might be suited to bring added value to the current setup.

The implementation details that hampered optimal workflow most were a lack of feedback when gestures have been detected, and a way to use defined zoom levels integrated with the pinch and pan gesture.

For the feedback, visual and audible cues are proposed that give feedback on the state of the application. Visual cues are already part of the iPad part of the prototype, for example

when using the direct access functionality to bring up a defined pair of mammograms with one tap. Further graphical hints on the iPad screen might support the intuitiveness of this gesture even more. Also, when the zoom/pan mode is activated by the pinch gestures, the iPad screen lights up to signal a mode change to the user. We have noticed in several participants that they attempted to return back to the previous state (i.e. unzoom and re-center) by swiping from left to right.

A further improvement might be brought about by audible cues that could help to signal the successful execution of a gesture to the user. This is a generally useful supportive feature since it removes uncertainty. A subtle sound could for example signal the end of roaming (last roaming image reached), and separate sounds for one-finger and two-finger swipes could immediately reflect the executed action to the user. Since the user interface is indirect, an optimal user experience requires an involvement of the user via such feedback mechanisms.

Regarding zooming, in commercial mammography reading workstations, a 1:1 pixel:voxel ratio zooming mode is usually offered to ensure that no resampling artifacts disturb image reading. In our seamless zoom gestures, it could be implemented that the zoom level snaps to that ratio when the user comes near it. Alternatively, a second zoom gesture could emulate a 1:1 magnifying glass. Also, proposing an efficient and easily memorable gesture for contrast and brightness adjustment would be desired.

Currently, the execution of the full hanging protocol in the defined ordering is mandatory, while shortcuts allow to jump to each image pair the reader is interested in. To fully harvest the flexibility of the direct access gesture combined with the seamless zoom tool, we suggest to implement two extensions. Firstly, we propose to remove the strict enforcement of the hanging protocol and to enable a shortcut to finalize the case at any time. To ensure that all parts of all current mammograms have been read, a second measure has to be taken: While the mammograms are read, the computer tracks the parts of the mammograms that have been viewed in sufficient resolution. When the user finalizes the case, he or she is presented with all remaining image tiles, accessible with the simple hanging protocol gesture. Alternatively, CADe algorithms could be integrated to restrict the tiles to those with suspicions.

It might appear that in the mammography setting, the iPad is employed solely in the function of a large touch pad. However, recalling the remarks about feedback through audio-visual cues reveals the importance of a device that provides both input and output functions. In addition, we have proposed in the context of the MRiPad prototype how in a more integrated and unified scenario, the mobile device might easily be extended in its functional scope to support many image analysis tasks like segmentation and annotation, and of course to collect case information directly on the iPad. One such addition relevant in mammography screening would be the reporting that could be executed on the device.

With a reworked prototype, a proper study including more radiologists and mammographers as volunteers, and more cases to study user performance in more detail is highly desirable. To this end, the performance of the server software in terms of load times of patient cases needs to be increased, and the workstation needs to be optimized for mouse/keyboard interaction to conduct tests with statistical relevance.

3.5 Discussion

We propose a novel system setup utilizing a multi-touch-capable mobile device to replace mouse and keyboard, where the image display is done on dedicated display devices, while auxiliary information and interaction capabilities are provided on the mobile device. This setup suggests that a less versatile device than an iPad or other mobile device might already be suitable to support the essential ingredients of our approach. We argue that this is hardly possible. In our

experiments, we found that the feedback the mobile device can give (by sound, by subtle changes of the interface that are noticable without directly looking at the screen) are a key factor to generate a user experience that is satisfying. In addition, the device can change its role and display additional information or provide more specific interaction and even allow data entry using an on-screen keyboard. It also stores data and is able to assist in hospital communication regarding the clinical cases.

Two application prototypes situated in the area of breast cancer screening and diagnosis, respectively, have been proposed. They tried to implement some of the benefits of fusing information display and interaction in a novel fashion. They have been described and evaluated in terms of performance and user experience.

Our goal for the future is to extend the prototype further to demonstrate the clinical feasibility and added value of the ubiquitous adaptive information processing mechanism of our approach, including collection of information (e.g. clinical assessments), processing of information (joining information about the user, his location, today's tasks, etc.), and delivering information in a highly specific and custom-tailored way. It is this capability of the mobile device that is the key to a fully integrated and pervasively simplyfied workflow. The limited screen space can easily be accounted for with a suitable setup as we have proposed it in this work.

Display devices exist in many places in the hospital, like e.g. in patient rooms, in meeting rooms, in dedicated diagnostic reading facilities, in the operating theater, and in recreational areas. These display devices may be of different nature: HD television in the patient room, projectors in meeting rooms, certified multi-monitor setups in the reading room and so on. The mobile device can then be utilized to provide seamless, location-specific interaction with images, always using the same general gesture-based approach, but providing task-specific tools or views on the patient data, depending on the current situation.

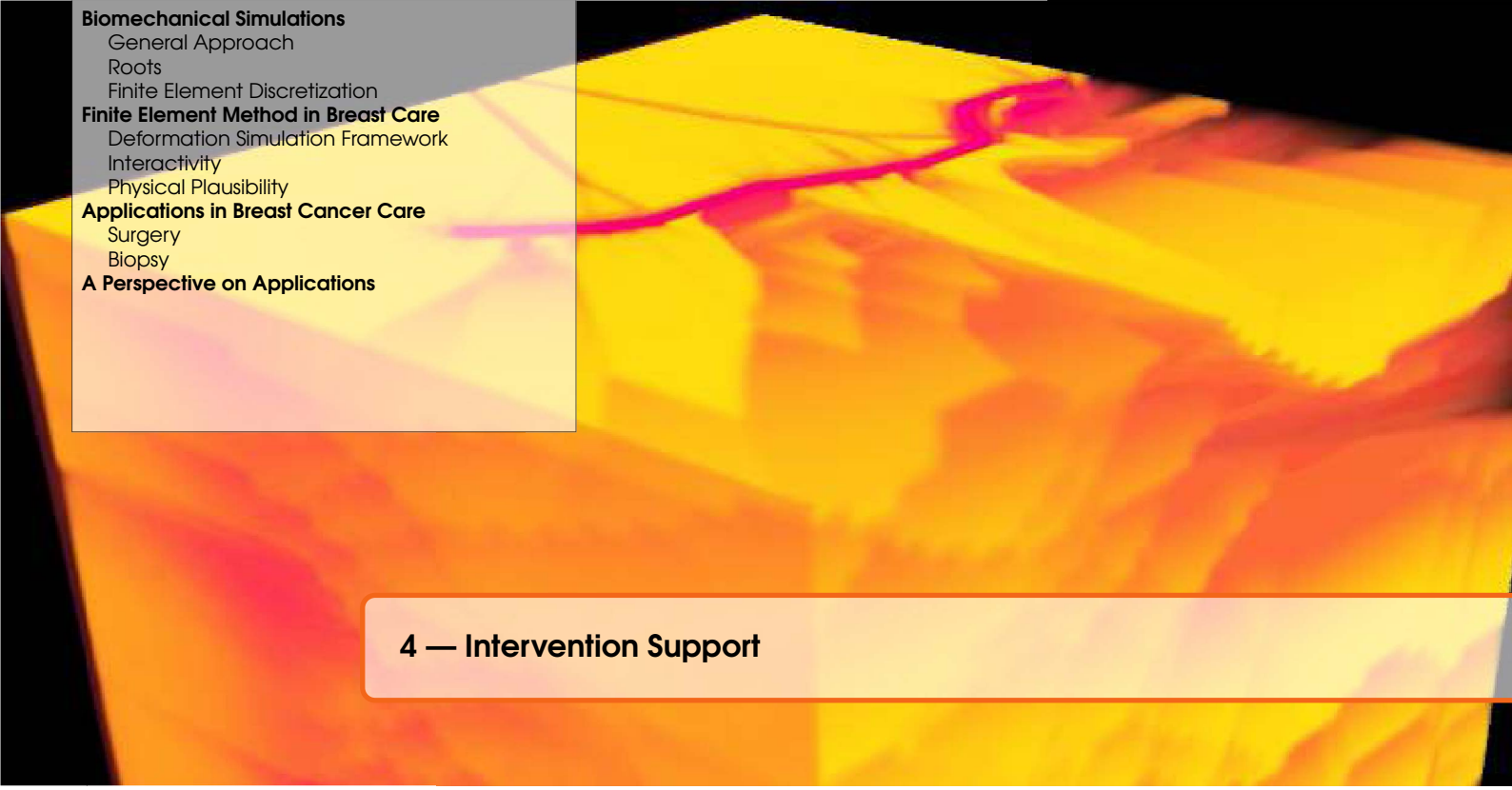
While we have presented a prototype that uses the iPad as a replacement and alternative to mouse and keyboard for two different clinical reading task, many of the proposed concepts are applicable on a more general level. In particular, to connect the mobile device to any viewing workstation by a simple login mechanism with subsequent exchange of mutually identifying information has many useful applications in hospital routine, but it may be understood as a basic concept as well, transcending the concept currently followed, where a dedicated control application has to be selected by the user, for example to control the home theater, or the music player, or smart home devices. In our proposal, one application unifies the different "remotes" into one intelligent application that adapts to the situation.

We consider this a substantial contribution to take much of the complexity out of clinical workflows, and out of today's hospital routine. Changing places frequently, interrupting the work and pick it up again, keeping track of information and needing to take information along are common requirements that usually require efforts in work organization. Mobile devices are particularly suitable in such environments. Their ability to connect to different information sources simultaneously or serially, combined with our proposed setup to control the image viewing in different environments with similar interaction concepts may help to make work more efficient by silently delivering task-specific information tailored to the role and needs of the user and the situation.

In the hospital environment, there are applications outside the radiology department, most importantly in emergency care. Not only can patient information be collected directly into an electronic system, but ultimately, emergency CT and x-rays may both be first operated, and then reviewed on the mobile device. The mobile device can help to select the appropriate imaging by providing a connection to a case database, and it may be used as a remote control for the scanner, before the images are visualized on the device and potentially even compared to similar cases to aid decision making. In addition, mobile devices are robust, and sterile wrappings exist that allow the device to be used even in the operating theater.

Given the generality of the underlying concept, a transfer of the location awareness concept and context sensitive mobile device integration to applications outside hospitals is conceivable. Assembly lines, where different machines, robots, and other systems need to be configured, controlled, and monitored, and all other complex setups in production and industry may benefit from workflow support that doesn't require a change of tool between work places and tasks, and that allows to unify information, control, and result assessment in one place.

Biomechanical Simulations
General Approach
Roots
Finite Element Discretization
Finite Element Method in Breast Care
Deformation Simulation Framework
Interactivity
Physical Plausibility
Applications in Breast Cancer Care
Surgery
Biopsy
A Perspective on Applications



4 — Intervention Support

Breasts are flexible, deformable organs with a high intra- and inter-individual variability in size, shape, and elasticity. The response to movement of the woman, to external forces, and to manipulation during medical procedures is likewise diverse and governed by many factors we will describe in Sec. 4.2. Approaches to model the deformation of soft tissue in a physical plausible way subject to those factors do exist, but often are limited in their applicability to the problems pertinent in breast care. Also, extensibility and flexibility of existing implementations is very limited, so that developments of a more general approach along the examples of two problems arising in breast care are warranted.

In breast surgery, the expected but unpredictable deformation makes procedures less robust and, from an efficiency perspective, longer than optimal. Training and experience are required for a surgeon to estimate the spatial shift of a lesion from imaging to surgery positioning, and to decide on the surgical procedure. Expected cosmetic outcome is one factor taken into account, and generally access to the index lesion is chosen so that both the cut to enter is minimized, as well as the path on which the surgeon advances towards the lesion is as short as possible. Visualizing the deformation before the procedure starts may help planning and procedure preparation.

In MR-guided breast biopsies, the target is visualized by a contrast-enhanced sequence. Though the success rate of this type of biopsy is high, and the technical support for the procedure established for years, improvements in the workflow are required. The most prevalent shortcomings are the lengthy and (from a patient's perspective) inconvenient dynamic contrast enhanced MRI (DCE-MRI) required for targeting, the requirement of matching the diagnostic scan to the interventional DCE-MRI, and the occasional target non-visualization that causes the abortion of 10% of the procedures (Brennan et al. 2011).

This chapter will outline the theoretical background of biophysical models of soft tissue deformation based on the finite element method. With an understanding of the principal limits of such models, both in terms of the achievable computational efficiency and achievable realism concerning the deformations that can be modeled, approaches with favorable properties will be presented. Implemented in an extensible framework, they allow to simulate large deformations of soft tissue with high computational efficiency and at a level of physical plausibility sufficient

The chapter image shows a colorized volume rendering of a distance field that is used for the definition of external force fields.

for the two example applications. Moreover, since the apporaches are of a general nature, they can be applied to problems outside image-based breast care.

The approaches described in this section have partially been published in

- Markus Harz et al. (2011a). "Real-time Breast Deformation using Non-linear Tissue Properties". In: *Lecture Notes in Informatics*. Edited by Hans-Ulrich Heiss et al. Volume P-192, page 442
- Markus Harz et al. (2011b). "Towards Navigated Breast Surgery Using Efficient Breast Deformation Simulation". In: *Medical Image Computing and Computer-Assisted Intervention, Workshop on Breast Image Analysis*. Edited by Christine Tanner et al.
- Markus Harz et al. (2012a). "Efficient Breast Deformation Simulation". In: *Workshop on Virtual Reality Interaction and Physical Simulation*. The Eurographics Association, pages 117–126

4.1 Finite Element Method-Based Biomechanical Simulations

4.1.1 General Approach

The Finite Element Method (FEM) is one way to make model problems computable that are expressed in continuous form in space and time. This is true for the model of material deformation, which is given by a partial differential equation of second order we will briefly introduce later.

The unknown in this equation is the deformation field. The FEM way to solve this equation is by considering points in discrete locations supporting continuous interpolation functions between them that both divide the domain into patches and simultaneously keep the full information of the continuous spatial domain accessible. Forces are then accumulated at these discrete points from the surrounding patches. In many practical cases, these points are placed on regular grids, which can aid in simplifying the calculations further.

To give the reader an idea about the fundamental concept of the simulation of elastic bodies, we review some of the origins of the underlying equations. Linear elasticity as expressed by Hooke's Law in short assumes a linear stress-strain relationship, which holds for real-world materials over a wide enough range of forces to be practically applicable.

Assuming three-dimensional space, a finite element can be thought of as a portion of material that has neighboring elements it shares some or all its faces with. This directly leads to the common abstraction and depiction of this concept as a volumetric mesh, consisting of vertices, and edges and faces between them, composed into the elements, which together span a volume, often called the domain and commonly denoted Ω.

The following brief account of the major components of a finite element method approach to the simulation of soft tissues follows the descriptions given by Boehler (2011), Braess (2013), Bro-Nielsen (1997), and Georgii (2007), consulting the works of M. G. Larson et al. (2013), Olver (2013), Sayaz (2008), and Zienkiewicz et al. (2005) for some aspects.

As initially stated, the elastic behavior of a body can be modeled by equations called partial differential equations. Partial differential equations consist of multi-variable functions and their partial derivatives, and only simple examples have analytic solutions.

The finite element method, in brief, is one way to numerically solve such partial differential equation without evaluating it only at discrete points (which is the approach of the finite difference method for the same problem statement), but by defining boundary values that allow for an analytic solution or approximation on the restricted domain (Olver 2013).

The following account of the finite-element-based simulation of body elasticity comprises two parts: it first motivates the mathematical problem arising from the composition of a material model describing how stress leads to strain, the motion equations describing how strain leads to displacements, and the boundary conditions that state what can and what cannot happen to the body. These three components will be composed into an equality of energies, known as the total potential energy functional, and having the displacement as the unknown (free) variable. Estimating a solution to this total potential energy equation may be done by different means; here, we will look at the principle of how a finite element discretization helps solving the problem.

4.1.2 Roots of the Total Potential Energy Functional

Robert Hooke and Isaac Newton were pivotal in the development of the laws that govern modern elasticity theory. Hooke's considerations about the elastic deformation of springs and solids in his 1678 *De Potentia Restitutiva* (compare Fig. 4.1a) included experiments with copper wires to

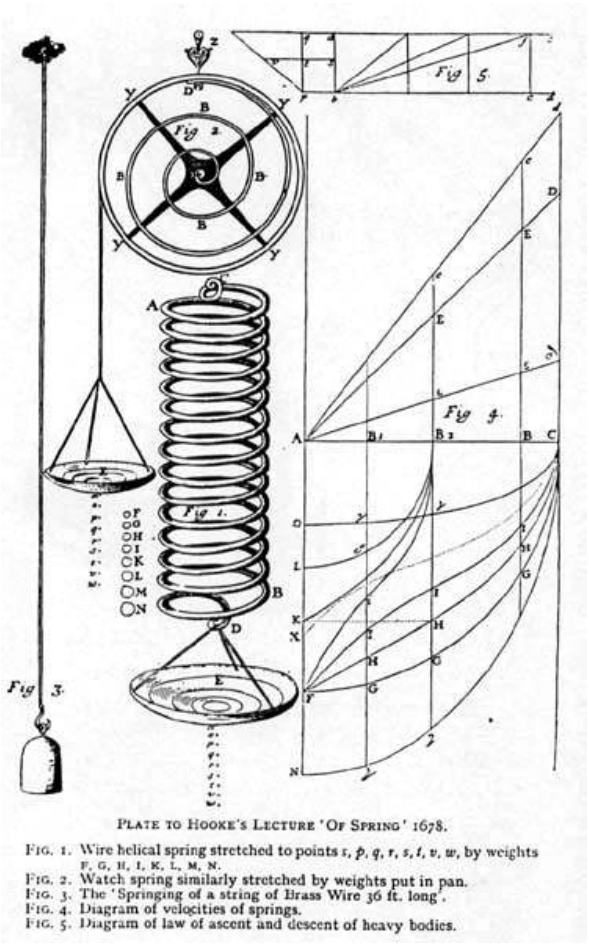
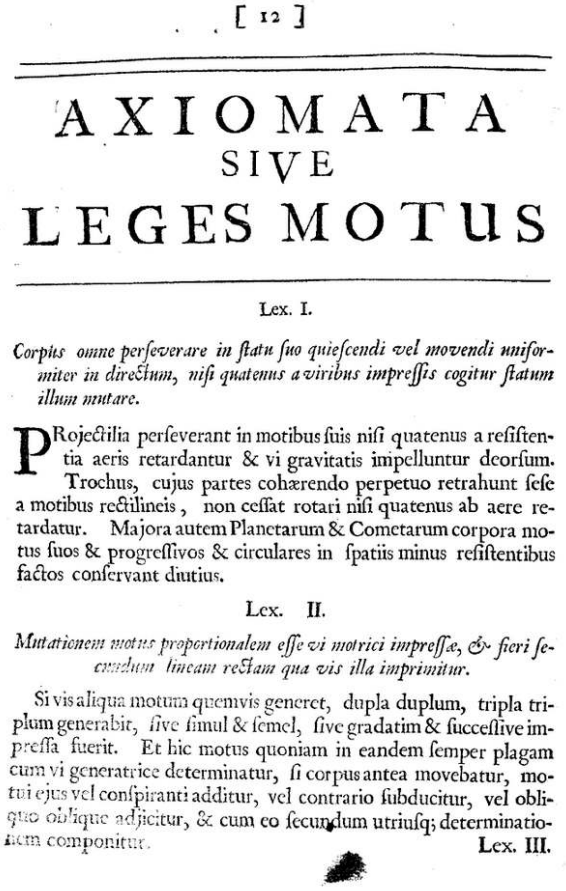


PLATE TO HOOKE'S LECTURE 'OF SPRING' 1678.

FIG. 1. Wire helical spring stretched to points *s, p, q, r, s, t, v, w*, by weights F, G, H, I, K, L, M, N.
FIG. 2. Watch spring similarly stretched by weights put in pan.
FIG. 3. The 'Springing of a string of Brass Wire 36 ft. long'.
FIG. 4. Diagram of velocities of springs.
FIG. 5. Diagram of law of ascent and descent of heavy bodies.

(a)



[12]

AXIOMATA
SIVE
LEGES MOTUS

Lex. I.

Corpus omne perſeverare in ſtatu ſuo quieſcendi vel movendi uniformiter in directum, niſi quatenus a viribus impreſſis cogitur ſtatum illum mutare.

PRojectilia perſeverant in motibus ſuis niſi quatenus a reſiſtentia aeris retardantur & vi gravitatis impelluntur deorſum. Trochus, cujus partes cohærendo perpetuo retrahunt ſeſe a motibus rectilineis, non ceſſat rotari niſi quatenus ab aere retardatur. Majora autem Planetarum & Cometarum corpora motus ſuos & progreſſivos & circulares in ſpatiis minus reſiſtentibus factos conſervant diutius.

Lex. II.

Mutationem motus proportionalem eſſe vi motrici impreſſæ, & fieri ſecundum lineam rectam qua vis illa imprimitur.

Si vis aliqua motum quemvis generet, dupla duplum, tripla triplum generabit, ſive ſimul & ſemel, ſive gradatim & ſucceſſive impreſſa fuerit. Et hic motus quoniam in eandem ſemper plagam cum vi generatrice determinatur, ſi corpus antea movebatur, motui ejus vel conſpiranti additur, vel contrario ſubducitur, vel obliquo obliquæ adjicitur, & cum eo ſecundum utriuſq; determinationem componitur.

Lex. III.

(b)

Figure 4.1: Facsimiles of Hooke's and Newtons treatises on the laws of elasticity and motion, respectively.

measure the elastic properties of solids. Almost contemporaneously, Isaac Newton published his *Philosophiæ Naturalis Principia Mathematica* in 1687, laying out the three laws of motion (compare Fig. 4.1b).

It took a number of thinkers and theorems to arrive at the generalized laws of motion (including energy equilibrium) that influence elasticity theory, and we will not follow this path. Suffice to say that from the kinetic energy of a system of particles, by taking partial derivatives in space and the total derivative in time, the generalized equation for motion can be derived. This, in turn, furthered the notion of the *action* of a body as the time integral of the sum of its total kinetic energy and its total potential energy, the so called *Lagrangian*. Note that action hence is a functional rather than a function; its value depends on the Lagrangian for all time it is integrated over (see HAND et al. (1998) and TORBY (1984) for details).

From Stress to Strain: The Material Law

We begin by clarifying the relation of stress and strain in a solid body, then see how strain leads to displacements, and how the material of the body influences this (compare Fig. 4.2).

Stress. Stress is a scalar value expressing force per oriented area and expressed in pascal (Pa) or, in SI base units, $1\frac{\mathrm{kg}}{\mathrm{m}\cdot\mathrm{s}^2} = 1\frac{\mathrm{N}}{\mathrm{m}^2}$. Stress is obviously measured in the same unit as pressure, which gives rise to the interpretation of stress as a body-internal reaction to external pressure exerted on the body. Stress is denoted (CHOU et al. 1967)

$$\sigma = \frac{F}{A}. \tag{4.1}$$

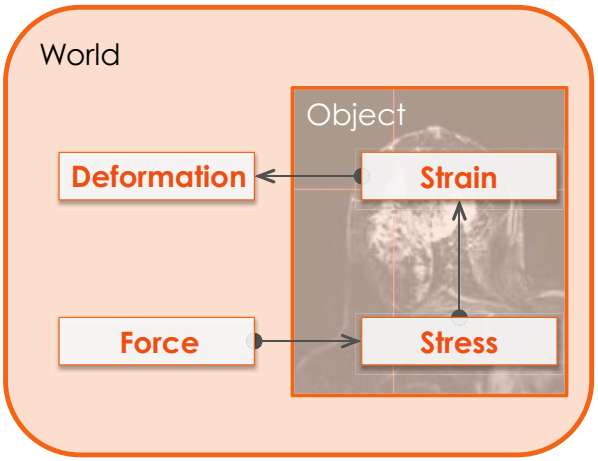


Figure 4.2: Relating internal and external energies with each other. After Bro-Nielsen (1997).

This assumes uniform distribution of stress over the area, and a direction of the force orthogonal to the area (uniaxial stress, engineering stress). Both can not generally be assumed in practice; instead external forces are vectorial quantities of arbitrary direction. Uniform distribution of stress, however, can be ensured by considering suitable areas.

In more complex situations, solid bodies experience combined stress, and there is no vector representation for this anymore. In this case, stress across a given area depends on the area's orientation in a more involved way. It was Cauchy[1] to note that the stress vector of any surface still always is a linear function of the surface normal; hence for three-dimensional Cartesian space, this function can be expressed in a 3×3 matrix, which is, owing to Cauchys denomination of stress as tension, in modern mathematics called a tensor: the Cauchy stress tensor.[2] Newtons laws of motion indirectly influenced Cauchys discovery, since they implied the law of the conservation of momentum. The Cauchy stress tensor is symmetric:

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_x & \tau_{xy} & \tau_{xz} \\ & \sigma_y & \tau_{yz} \\ sym. & & \sigma_z \end{bmatrix} \tag{4.2}$$

To capture the influence of the directionality on the reaction of the body, the stress is divided into orthogonal and in-plane components, denoted σ and τ, respectively. Common in engineering, the six-degrees-of-freedom tensor is denoted in Voigt notation as a column vector

$$\boldsymbol{\sigma} = [\sigma_x \sigma_y \sigma_z \tau_{yz} \tau_{xz} \tau_{xy}]^T$$

Common external forces causing stress to a solid body are gravity, which affects the body globally and homogeneously, and contact forces, occurring for example in collisions.

Strain. Strain is a body's reaction to stress, oftentimes observed in the form of the body's deformation. As with the stress, the strain can be written in a tensor, which is again symmetric under the infinitesimal strain assumption that needs the deformations to be infinitesimally smaller than any relevant dimension of the solid body. A derivation of the geometric linearization of the finite strain tensor (that is capable to capture large deformations) is out of scope of this work, but can be found elsewhere (Wikipedia 2013b).

[1] Augustine-Louis Cauchy; 21 August 1789 – 23 May 1857

[2] Any linear function connecting two physical vectors are today called tensors. The Cauchy stress tensor is by this classification a second order tensor.

The linearized (Cauchy) strain tensor again comprises normal strain ε and shear strain γ and is written

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_x & \gamma_{xy} & \gamma_{xz} \\ & \varepsilon_y & \gamma_{yz} \\ sym. & & \varepsilon_z \end{bmatrix} \tag{4.3}$$

Like the stress tensor, the strain tensor is often written in Voigt notation, $\boldsymbol{\sigma} = [\varepsilon_x \varepsilon_y \varepsilon_z 2\gamma_{yz} 2\gamma_{xz} 2\gamma_{xy}]^\mathsf{T}$. A geometric interpretation on the example of a rectangle deformed into a rhombus, the normal strain describes the elongation or compression along the coordinate axes, while shear strain prescribes the angular changes. Note that the strain tensor is sometimes also defined in terms of displacements, already anticipating the strain-displacement relationship we shall describe next. If $\boldsymbol{\varepsilon} = 0$, there is only rigid body movement like translation and rotation; hence it is sometimes defined that a displacement is a deformation where the strain tensor doesn't vanish everywhere (M. G. Larson et al. 2013).

Hooke's Law. Hooke's Law connects stress and strain—although Hooke expressed it differently at that time. Looking at springs, he found that for a certain range the extension of the spring under a given load is proportional to the load, or:

$$\mathsf{F} = \mathsf{k} \cdot \mathsf{x} \tag{4.4}$$

where F is the force to extend the spring by the amount x. The constant factor k is the spring constant, a material property.

In our context, a generalization of this law using the tensors for stress and strain is

$$\boldsymbol{\sigma} = \mathbf{c} \cdot \boldsymbol{\varepsilon}. \tag{4.5}$$

With the strain $\boldsymbol{\sigma}$ and the stress $\boldsymbol{\varepsilon}$ being second-order tensors, $\mathbf{c}$ is a $3 \times 3 \times 3 \times 3$ fourth-order tensor with 81 entries, called the *stiffness tensor*. While stress and strain are independent of the body, $\mathbf{c}$ is a material property that may vary with environmental conditions like temperature or pressure.

Exploiting symmetries of the three tensors, it turns out that $\mathbf{c}$ only possesses 21 degrees of freedom called elastic moduli, which are further reduced when only considering isotropic material, that is, material that reacts the same regardless of orientation (Braess 2013). In fact, for isotropic material, the linear relationship expressed in $\mathbf{c}$ is characterized by only two moduli: *Young's modulus* $\mathsf{E} \in \mathbb{R}$ and the *Poisson ratio* ν; compare Fig. 4.3.

Young's modulus E may be interpreted as a materials resistance against pressure, and like pressure, its unit is pascal ($\frac{\mathrm{N}}{\mathrm{m}^2}$). Material with a lower E modulus are softer than those with higher E. Values for E for materials at the far ends of the natural range are $10 - 100$ kPa for rubber, and $1,220,000$ kPa for diamonds.

Poisson's ratio ν, on the other hand, characterizes the behavior of the material under stress. ν determines how much volume a material may loose or gain under tension or pressure, or, in another interpretation, how the thickness of the body changes when the body is elongated or compressed. The value range of ν is from -1 to 0.5. Some materials with a negative Poisson's ratio are called auxetic materials and widen when elongated and narrow when compressed. Besides some rocks and minerals the most notable example of an auxetic material is paper which gets thicker when stretched in-plane (Blumenfeld et al. 2011; Steinberg et al. 2002). The majority of materials, however, have values between 0 (e.g. cork and foam) and 0.5 (e.g. rubber), and in particular tissues of the human body mostly tend to show ν close to 0.5, implying they are almost incompressible.

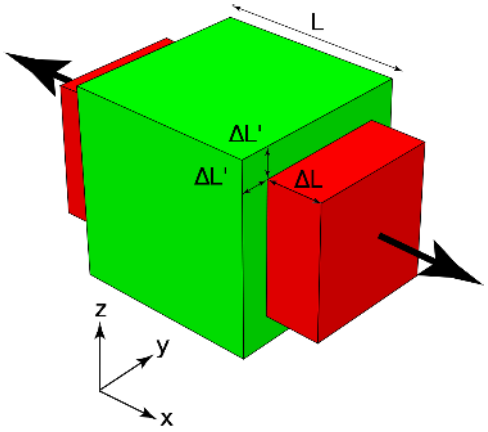


Figure 4.3: Assuming tension along x, a first-order approximation of the Poisson ratio ν of a linearly elastic material is $\nu = \frac{\Delta L'}{\Delta L}$. The definition, however, is in terms of strain: $\nu = -\frac{d\varepsilon_{trans}}{d\varepsilon_{axial}}$

Hooke's Law expressed in terms of Poisson's ratio and Young's modulus, after some arguments from tensor arithmetic and the insight that the strain tensor can be decomposed into a constant *volumetric strain tensor* and a varying *deviatoric strain tensor* component (Milton 2002, page 23), becomes

$$\boldsymbol{\varepsilon} = \frac{1}{E}(\boldsymbol{\sigma} - \nu[\mathrm{tr}(\boldsymbol{\sigma})\mathbf{I} - \boldsymbol{\sigma}]), \tag{4.6}$$

where $\mathbf{I}$ is the second-order identity tensor. From this relation, the matrix form of Hooke's law can be derived, noting that a fourth-order tensor may be written in Voigt notation as a 6×6 matrix. In matrix form, Hooke's law for isotropic materials is then written as

$$\begin{bmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \\ 2\gamma_{yz} \\ 2\gamma_{xz} \\ 2\gamma_{xy} \end{bmatrix} = \frac{1}{E} \begin{bmatrix} 1 & -\nu & -\nu & 0 & 0 & 0 \\ -\nu & 1 & -\nu & 0 & 0 & 0 \\ -\nu & -\nu & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2(1+\nu) & 0 & 0 \\ 0 & 0 & 0 & 0 & 2(1+\nu) & 0 \\ 0 & 0 & 0 & 0 & 0 & 2(1+\nu) \end{bmatrix} \begin{bmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ \tau_{yz} \\ \tau_{xz} \\ \tau_{xy} \end{bmatrix}, \tag{4.7}$$

where $2\gamma_{ij}$ is the *engineering shear strain*. Inversion of this relation yields

$$\begin{bmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ \tau_{yz} \\ \tau_{xz} \\ \tau_{xy} \end{bmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & \nu & 0 & 0 & 0 \\ \nu & 1-\nu & \nu & 0 & 0 & 0 \\ \nu & \nu & 1-\nu & 0 & 0 & 0 \\ 0 & 0 & 0 & (1-2\nu)/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & (1-2\nu)/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & (1-2\nu)/2 \end{bmatrix} \begin{bmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \\ 2\gamma_{yz} \\ 2\gamma_{xz} \\ 2\gamma_{xy} \end{bmatrix}, \tag{4.8}$$

which is the matrix notation of Eqn. (4.5). For isotropic materials, we shall express this form in short notation as

$$\boldsymbol{\sigma} = \mathbf{M}\boldsymbol{\varepsilon} \tag{4.9}$$

with the new symbol for the material matrix $\mathbf{M}$ to indicate the change from the general form to isotropic linear elasticity.

Hooke's law is not a law of nature, but a relationship under the small strain assumption. It is hence also called a constituent equation, deducted from reasoning and measurements (M. G. Larson et al. 2013).

From Strain to Displacement: The Motion Equations

Materials under stress may exhibit deformations, which are described in form of displacements. The equations governing this relationship are called motion or *kinematic* equations. The fundamental choice to make is how to describe such displacements: the first option is to describe it at material points, i.e. at points moving with the body; the second option is to describe it in fixed spatial coordinates, figuratively watching the body move by. Fortunately, in linearized elasticity there is no difference between the material and the spatial reference frame (Basaran 2008), dubbed the Lagrangian and the Eulerian frame of reference, respectively.

In the Lagrangian frame of reference, displacements are thought to be added to points of the body and moving with the body. We denote the displacement function $\chi : \Omega \to \mathbb{R}^3$, mapping points from the body into new positions. χ defines a displacement field that is defined anywhere on the domain Ω (the body). The usual notation for this is $\mathbf{x}' = \chi(\mathbf{x}) = \mathbf{x} + \mathbf{u}$. To describe for an infinitesimal element of the body how the displacement function affects this element is the purpose of the displacement or deformation gradient, $\nabla\chi$ which is defined as $\nabla\chi = \frac{d\mathbf{x}'}{d\mathbf{x}}$ with components $\nabla\chi_{ij} = \frac{\partial \mathbf{x}_i'}{\partial \mathbf{x}_j}$. In matrix form, we obtain

$$\nabla\chi = \begin{bmatrix} \frac{\partial x_1'}{\partial x_1} & \frac{\partial x_1'}{\partial x_2} & \frac{\partial x_1'}{\partial x_3} \\ \frac{\partial x_2'}{\partial x_1} & \frac{\partial x_2'}{\partial x_2} & \frac{\partial x_2'}{\partial x_3} \\ \frac{\partial x_3'}{\partial x_1} & \frac{\partial x_3'}{\partial x_2} & \frac{\partial x_3'}{\partial x_3} \end{bmatrix} \tag{4.10}$$

with $\mathbf{x}' = \chi(\mathbf{x})$.

Using the simplified, linear Cauchy strain tensor expressed in engineering strain (M. G. Larson et al. 2013, page 203), i.e.

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right),$$

we can rewrite the engineering strain vector to become

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \\ 2\gamma_{xy} \\ 2\gamma_{yz} \\ 2\gamma_{xz} \end{bmatrix} = \begin{bmatrix} \frac{\partial u_x}{\partial x} \\ \frac{\partial u_y}{\partial y} \\ \frac{\partial u_z}{\partial z} \\ \frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x} \\ \frac{\partial u_x}{\partial z} + \frac{\partial u_z}{\partial x} \\ \frac{\partial u_y}{\partial z} + \frac{\partial u_z}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x} & 0 & 0 \\ 0 & \frac{\partial}{\partial y} & 0 \\ 0 & 0 & \frac{\partial}{\partial z} \\ \frac{\partial}{\partial y} & \frac{\partial}{\partial x} & 0 \\ \frac{\partial}{\partial z} & 0 & \frac{\partial}{\partial x} \\ 0 & \frac{\partial}{\partial z} & \frac{\partial}{\partial y} \end{bmatrix} \cdot \begin{bmatrix} u_x \\ u_y \\ u_z \end{bmatrix} = \mathbf{B}\mathbf{u}. \tag{4.11}$$

We will meet the compact notation of the above, $\boldsymbol{\varepsilon} = \mathbf{B}\mathbf{u}$, when we compose the energy functional, using the deformation gradient and the material matrix from Eqn. (4.9). In this undertaking, we also need a further expression connecting the engineering strain and engineering stress into the *internal strain energy* (ibidem):

$$U = \frac{1}{2}\int_\Omega \boldsymbol{\varepsilon}^T \boldsymbol{\sigma}\, dx \tag{4.12}$$

Composing the Forces: The Total Energy Functional

The partial differential equation governing linear elasticity can be understood from the insight that the total force acting on any material volume must vanish ibidem. The forces acting on a volume $\Omega \in \mathbb{R}^3$ can be classified into two kinds: volume forces f penetrating the object, like for example gravity, and contact forces acting on the surface $\partial\Omega$, like for example pressure. Volume forces are described by force densities, contact forces by vector fields.

Contact forces, on the other hand, are expressed in the stress tensor $\boldsymbol{\sigma}$. The force on a small surface ds with unit normal $\mathbf{n}$ is hence $\boldsymbol{\sigma} \cdot \mathbf{n} ds$, and the total force F on the object is

$$F = \int_\Omega f \, dx + \int_{\partial\Omega} \boldsymbol{\sigma} \cdot \mathbf{n} \, ds \tag{4.13}$$

This can be reformulated exploiting the divergence theorem[3], yielding

$$F = \int_\Omega (\mathbf{f} + \nabla\boldsymbol{\sigma}) dx \tag{4.14}$$

Cauchy's equation of equilibrium follows by setting

$$(\nabla\sigma)_i = \sum_{j=1}^{3} \frac{\partial \sigma_{ij}}{\partial x_j}, \; i = 1, 2, 3,$$

and noting that in equilibrium, the forces must vanish everywhere:

$$\mathbf{f} + \nabla\boldsymbol{\sigma} = 0. \tag{4.15}$$

This gives us a vector-valued partial differential equation (PDE) that we next combine with a second partial differential equation: A corollary of the constitutive equation Eqn. (4.6) can be found by exploiting a different notion of the elastic moduli called the Lamé coefficients μ and λ (Braess 2013). After rearranging for $\boldsymbol{\sigma}$, the system of PDE becomes

$$-\nabla\boldsymbol{\sigma} = \mathbf{f} \tag{4.16}$$

$$\boldsymbol{\sigma} = 2\mu\varepsilon(\mathbf{u}) + \lambda(\nabla\mathbf{u})\mathbf{I}. \tag{4.17}$$

This can be combined into the Cauchy-Navier equation:

$$\mathbf{f} + \mu\Delta\mathbf{u} = (\lambda + \mu)\nabla(\nabla \cdot \mathbf{u}) = 0 \tag{4.18}$$

It is this equation for which we are looking for a solution $\mathbf{u}$, and the way to do this is to derive a variational formulation that can be solved with methods for numerical optimization. For a mathematical coverage of this, see (M. G. Larson et al. 2013). The derivation is out of scope of this introduction.

However, with the basic insight that all energies have to be in equilibrium, the internal and external forces can be composed into (Georgii 2007)

$$\Pi = \frac{1}{2}\int_\Omega \varepsilon^T \boldsymbol{\sigma} \, dx - \int_\Omega g^T \mathbf{u} \, dx - \int_{\partial\Omega} \mathbf{f}^T \mathbf{u} \, dx, \tag{4.19}$$

which are the internal strain energy from Eqn. (4.12), and the external volume and contact forces from Eqn. (4.13) expressed in terms of the displacement $\mathbf{u}$.

Variation with respect to the displacement yields the functional that is to be minimized using, in our case, the finite element discretization.

4.1.3 Finite Element Discretization

The key to the finite element method to solve the PDE from Eqn. (4.19) is to understand how piecewise analytic solutions on Ω together may approximate the solution of the complete PDE. This is effected by discretizing Ω into finite elements Ω^e, in our case, tetrahedrons consisting of four nodes P_q defined by their spatial location $\mathbf{x}_q = [x_q, y_q, z_q]^T$ with q the global node index. The deformation at node q is consequently $\mathbf{u}_q = [u_q, v_q, w_q]^T$ (Bro-Nielsen 1997).

[3]Divergence theorem: The net flow (of any quantity) through a surface is given by the difference of all sources and sinks.

Elements and Shape Functions

Looking at one tetrahedron Ω^e, its four nodes are $P_i^e, i = 1, 2, 3, 4$. Note that the local vertex index and the global node index above, denoted q, are not the same. Between the nodes of one element, the displacement $\mathbf{u}$ can be interpolated linearly using so-called *shape functions* $N_i : \mathbb{R}^3 \to \mathbb{R}$. These shape functions are associated with the nodes of the element and need to be selected such that their support is locally confined to the volume of the element, in other words, the shape function for a node is 0 at all other nodes.

The displacement field $\mathbf{u}$ can then be interpolated for any point $\mathbf{x}$ within element e (Bro-Nielsen 1997):

$$\mathbf{u}(\mathbf{x}) = \sum_{i=1}^{4} N_i^e(\mathbf{x})\mathbf{u}_i^e. \tag{4.20}$$

Note that the $\mathbf{u}_i^e$ are vectors. The definition for the $N_i^e(\mathbf{x})$ for linear interpolation in tetrahedral elements is given for example in (Bro-Nielsen 1997; Georgii 2007) and is not reproduced here. We note, however, that with Eqn. (4.20) now we have a locally continuous expression for $\mathbf{u}$ that we can partially differentiate with respect to x, y, z to arrive at a transformed and discretized expression of the strain energy (cf. Eqn. (4.12)) for one element and depending on $\mathbf{u}$ (Bro-Nielsen 1997):

$$U(\mathbf{u}) = \frac{1}{2} \sum_e \int_{\Omega^e} (\bar{\mathbf{u}}^e)^T (\mathbf{B}^e)^T \mathbf{M} \mathbf{B}^e \bar{\mathbf{u}}^e \, d\mathbf{x}. \tag{4.21}$$

For the tetrahedral element, $\mathbf{B}^e \in \mathbb{R}^{6\times 12}$ is the element-dependent Cauchy strain tensor from Eqn. (4.11) that can now be written in terms of the partial derivatives of the shape functions (see Boehler 2011; Georgii 2007). $\bar{\mathbf{u}}^e$ is the compound element displacement vector $\bar{\mathbf{u}}^e = [(\bar{\mathbf{u}}_1^e)^T, (\bar{\mathbf{u}}_2^e)^T, (\bar{\mathbf{u}}_3^e)^T, (\bar{\mathbf{u}}_4^e)^T]^T$, and $\mathbf{M}$ the material matrix (compare Eqn. (4.9)).

The Element Stiffness Matrix

The solution to Eqn. (4.21) is found when the total strain energy is minimal, which happens when the first variation of U vanishes: $\delta U(\mathbf{u}) = 0$. Splitting $\delta U(\mathbf{u})$ into element contributions and noting the independence of the orthogonal contributions, Eqn. (4.19) simplifies to the per-element equilibrium equation

$$0 = \int_{\Omega^e} (\mathbf{B}^e)^T \mathbf{M} \mathbf{B}^e \bar{\mathbf{u}}^e \, d\mathbf{x} - \bar{\mathbf{f}}^e, \tag{4.22}$$

where the compound 12×1 vector $\bar{\mathbf{f}}^e$ collects all element forces. Note that all contributions to the integral are constant, which is the crucial result enabling the short notation and efficient implementation thanks to the resulting linear matrix equation

$$\mathbf{K}^e \bar{\mathbf{u}}^e = \bar{\mathbf{f}}^e. \tag{4.23}$$

Comprising the material matrix and the Cauchy strain tensor, $\mathbf{K}^e = (\mathbf{B}^e)^T \mathbf{M} \mathbf{B}^e V^e$ is the *element stiffness matrix*[4] (Bro-Nielsen 1997).

[4] V^e is the element volume.

The Global Stiffness Matrix

The global problem, $\mathbf{K}\bar{\mathbf{u}} = \bar{\mathbf{f}}$, requires to compose the global stiffness matrix $\mathbf{K}$ from the element contributions, commonly referred to as the *global stiffness matrix assembly*. The pivotal notion is that of node re-indexing. In the formulations above, nodes were indexed local to the element. To fill the correct entries in the global stiffness matrix, however, the contributions always affect nodes expressed in global indices. For the force vector, $\bar{\mathbf{f}}$, a similar treatment applies. It is of particular importance for our later work to remark that the linearity of the $\mathbf{K}^e$ with Young's modulus allows a fast update of the local stiffness by simple scaling.

4.2 Finite Element Method in Breast Care

In this thesis, two clinical problem statements will be addressed:

Surgery preparation and support Diagnostic breast MRI examinations are conducted with the patient placed in prone position in the breast coil. The breasts are slightly fixated by soft paddings, often giving them an unnatural shape. Surgery is conducted in supine position, and the breasts deform into their natural shape. The lesion position is marked with metal wire guides, which serve the surgeon's access planning and interventional orientation. Optimally, surgeons could plan their intervention image-based, and one way to allow for this is by means of a simulation of the deformation from prone to supine positioning.

Spatial correlation for MR-guided biopsy MRI-guided breast biopsies are performed based on target selection in diagnostic DCE-MRI. For the biopsy targeting, a contrast-enhanced MRI is again taken with the breast placed in the specialized biopsy coil to visualize the lesion for targeting. The breast is this time compressed to keep the target in place when puncturing with the biopsy needle. Ideally, image information from the diagnostic DCE-MRI could be fused with the interventional images. One potential solution is again to simulate the deformation and shift between the imaging conditions to yield a deformation field suitable for spatial correspondence establishment.

More detailed problem statements will be found in the sections describing the two applications, Sec. 4.3.1 and Sec. 4.3.2. Before, we shall introduce the technical prerequisites contributing to the proposed solutions of the two scenarios, which are partially provided by the simulation framework, and by several extensions that are central to this thesis and will be introduced in detail. These extensions emerge from requirements, which themselves follow from the clinical application scenarios.

Speed is required at different levels in surgery planning and surgery support. Planning may be based on results of simulations that run for hours. It may benefit from interactive manipulation, though, which requires near real-time simulation performance. Interventional simulation, on the other hand, requires the highest level of performance and efficiency. Biopsy preparation requires a fast solution to the correspondence problem since it needs to be calculated between acquisition of the interventional planning scan and the needle placement. Any calculation time will be wait time.

Interactivity is the possibility to interact with the simulation as described before, and change parameters like the direction of gravity during computations, for example to explore models and their behavior in these changing environments. Secondly, interactively influencing the model shape (i.e. interacting with its surface or volume) opens applications generating visual feedback and potentially also force feedback.

Physical plausibility is a requirement for successful translation into clinical practice. Breast tissue is very soft, and due to its composition it behaves fluid-like, anisotropic, and non-linear in reaction to forces. A model that allows for non-linear, anisotropic material behavior is

thus required. Also, the breast tissue macroscopically shifts around the chest wall when changing position from prone to supine, an effect that needs to be modeled. During surgery in the operating room, the breast shape is influenced by manipulation. The shape change needs to be tracked and submitted into the deformation model.

Accuracy is required to a different extent for the two scenarios. For surgery preparation and to a certain extent also surgery support, a visualization of the overall change of breast shape will already be helpful. For biopsy targeting, accuracy in spatial correspondence needs to be of the order of the biopsy sample size, which is about 10 mm in diameter. This level of accuracy needs to be reached around the target, but not everywhere.

Generality assures that the solution can be applied to other scenarios, and can be applied if parameters of the scenario change, by including them in the model. Solutions without generality can not be extended easily, and require large efforts or further specialized additions to the model to transfer them to the new scenario.

Concerning breast tissue modeling, the tissue types and their patient-specific interconnectedness can hardly be accounted for in a realistic manner. Even if it were possible to image all tissue structures contributing to the shape and deformation of the breast, building a model on a fine enough scale to capture these structures would lead to unfeasible computational efforts, such that the more promising option is a modeling at a coarser scale.

Early efforts to model the breast deformation hence abstracted from tissue types and treated the breast volume as a homogeneous, isotropic material with linear elastic behavior (Azar et al. 2002). Later, some attempts to model heterogeneous tissue types were proposed, while others assume a homogeneous material, but introduce anisotropic or non-linear material laws, or combinations thereof. The works of Christine Tanner et al. (2011) and Whiteley et al. (2007) provide comparisons of selected approaches.

Pertaining to the task of prone-to-supine deformation simulation, researchers have tried to match the breast shapes or to derive one from the other by employing finite element analysis using non-linear material laws (Vijay Rajagopal et al. 2008; Vijayaraghavan Rajagopal et al. 2007). However, the patient-specific modeling effort and computational complexity of this approach are both high and thus unlikely to be employed in clinical routine. Currently available fast implementations of dynamic non-linear models (Han, J. H. Hipwell, Christine Tanner, et al. 2012; Han, J. Hipwell, et al. 2011) are based on explicit finite element approaches, which limit the magnitude of the largest possible time step as described before. Meanwhile, other research has looked at the feasibility of pseudo non-linear finite elasticity simulations and found them to be close, though not equal, to fully non-linear finite elasticity formulations (Whiteley et al. 2007). These pseudo non-linear formulations are closest to the work proposed in this contribution.

The interested reader is referred to publications aimed at researchers in engineering subjects (Bro-Nielsen 1997; Georgii 2007; Zienkiewicz et al. 2005) as well as more mathematically oriented introductions (Johnson 1987; M. G. Larson et al. 2013; Sayaz 2008) of the finite element method to obtain in-depth insights into many of the aspects touched upon in this introduction. The concise and comprehensive 25-pages overview of deformable models by Nealen et al. (2006) is recommended for a reader looking for the big picture.

4.2.1 Speed: The Efficient Corotated Cauchy Strain Based Framework

The implementations we propose are based on a multigrid finite element framework (Georgii 2007; Georgii and Westermann 2006), which efficiently simulates deformations of the breasts using the so-called co-rotated Cauchy strain formulation (Georgii and Westermann 2008; Rankin et al. 1986). One novel aspect of our contribution is to update the per-element elastic

modulus based on the shape change that the element experiences in a given simulation step. By this explicit per-element elasticity update, we effectively model a non-linear isotropic material law.

In our finite element framework, the element matrices are precomputed with a fixed elastic modulus E_0. Due to the linearity of the underlying material law, the element matrix of a particular element can then be obtained by scaling K^e by the stiffness value of this element relative to $E_0 \in \mathbb{R}$. Therefore, we can update the stiffness values within the assembling process of the corotational formulation at nearly no additional computational costs and thus achieve a fast update of stiffness values in the FE model analogously to previous approaches (Dick et al. 2008; Schiwietz et al. 2007). To efficiently update the data structures of the numerical multigrid solver, we make use of a fast approach to compute sparse-sparse matrix products (Georgii and Westermann 2010).

The solver has been wrapped into a module within the MeVisLab image processing platform to enable rapid prototyping of applications that need support of high-performance biomechanical simulations. Besides this fast solver, the simulation framework has been extended by two further solvers: one is an interface wrapper around the NiftySim explicit solver where several non-linear material laws are available (Z. A. Taylor et al. 2013). NiftySim has been ported to the graphics processing unit (GPU) to provide high update rates. The second addition is an implementation of the non-linear Green strain tensor following Georgii (2007), also in an explicit solver approach. Both have not been used in our studies reported here, but are available for comparisons in future work.

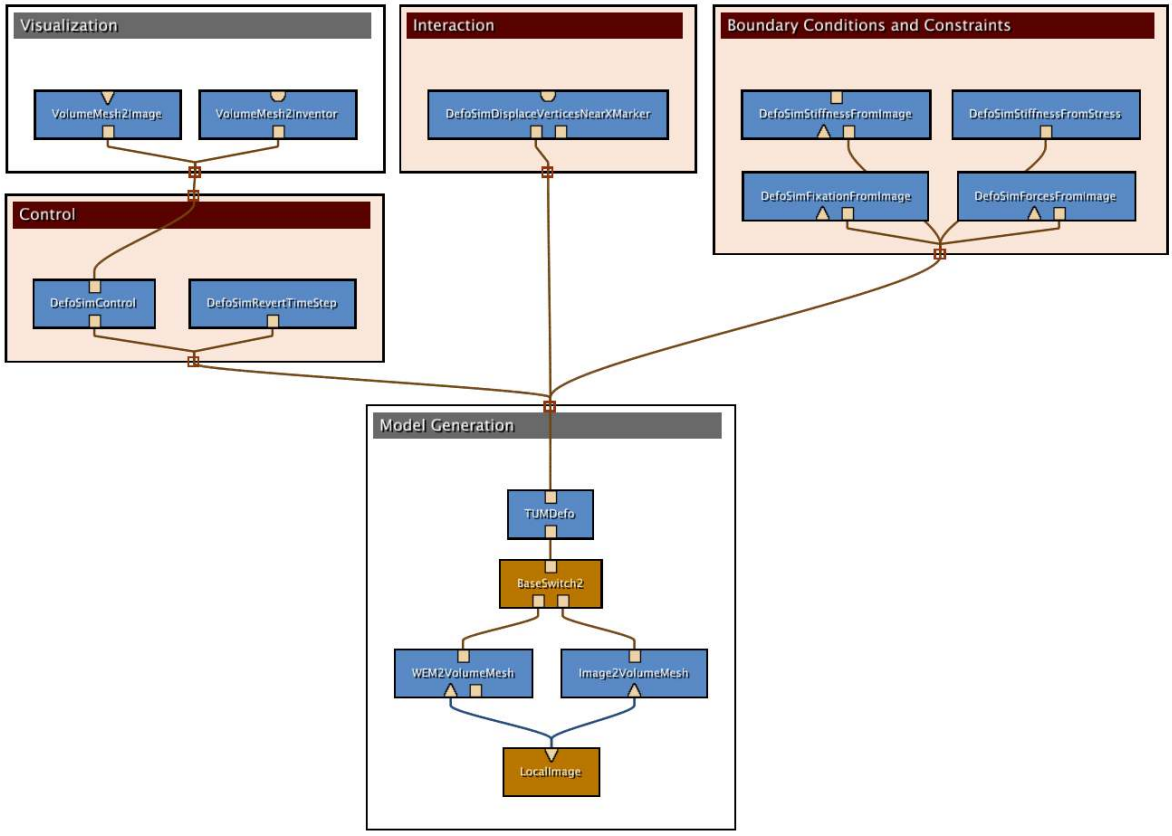


Figure 4.4: Overview of efficient deformation simulation framework in MeVisLab. The framework can be divided into five parts: (1) Model generation from a patient image; (2) control of boundary conditions like Dirichlet and Neumann boundary conditions and control of mesh state; (3) control of the simulation like material properties and time step size; (4) interaction with the model; and (5) visualization of the mesh and its internal state. Areas with contributions in this thesis are marked in color.

Internally, the framework communicates using shared data structures both for the volume

mesh and for the simulation parameters. Both are encapsulated into one object which is centrally provided by the simulation module `TUMDefo` (compare Fig. 4.4). Any extension requiring access to either the simulation or the volume mesh can share the object. It is left to the user of the framework to define the order of updates and calculations, which can be done in different convenient ways, thanks to the flexibility of MeVisLab and the framework. In particular, the framework has been set up such that any connected extension yields full control over the simulation object, including interface functions to trigger the simulation of one or more time steps.

A typical usage of the framework and its extensions might look as follows:

- During prototype development, the modules encapsulating the extensions are added to the network, and parameterized by hand. Their update behavior can be controlled by implementations of suitable triggers, which can be events or user interface elements.
- When prototype development progresses, Python scripts can overtake the task of parameterization, extension updating, and simulation triggering.
- Lastly, towards high-performance applications, it is advisable to convert the Python script based application logic into C++ code that directly accesses the simulation framework for read-out of input data and writing of results, and likewise directly accesses the simulation object to trigger the simulation of one or more time steps.

Looking at the functionality of the extensions more closely, the most powerful mechanisms are encapsulated into the extensions in the groups `Control` and `Boundary Conditions and Constraints` that are parts of the group addressing the physical plausibility. Haptic and physical interaction with the volume mesh is addressed in two extension suitable for different applications and described next.

4.2.2 Interactivity

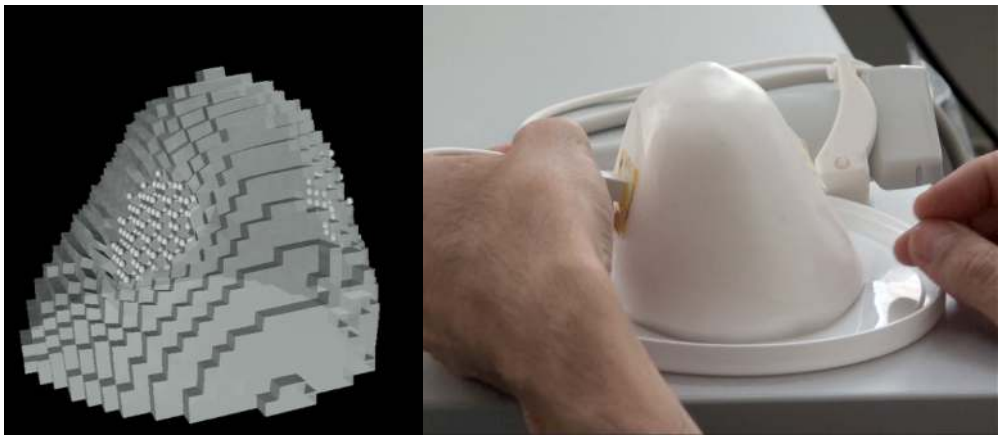
Applying Pressure Locally

Based on user input, the displacement of one or more vertices within a given radius around the user-defined location are updated to their new positions in one step. The simulation is highly robust against large external perturbation requests, and no special care must be taken to regularize the submitted displacement. Of course, the limits of the simulated material dictate certain restrictions.

The implementation allows to limit the influence of the user-provided input to surface vertices, or to allow surface and internal vertices to be affected. It is a nice property of the implementation that it is straight forward to change the boundary condition from a Dirichlet boundary condition (a vertex displacement) to a force input proportional to the displacement indicated by the user input. It is likewise possible to include feedback into the implementation. When a force (or displacement) has been exerted in one time step, the reaction on the input will be reflected by stress and internal body forces in the next step. These can be read out and fed back to the user. One potential application is a force feedback device that allows to "feel" the elastic properties of a virtual organ. Fig. 4.5 shows a demonstration application where a tracked sensor position placed on a soft breast phantom is transferred into displacement constraints of surface vertices of a model of the phantom. The phantom parameters have been set empirically so that the observed deformation matches the simulated one.

Applying Global Force Fields

If the target surface is known, for example by a surface scan acquired during the intervention, or if any other information is known about where displaced landmarks have moved, we propose



(a) (b)

Figure 4.5: A probe of an electromagnetic (EM) tracking device (a) is used to control the deformation simulation plugin. The marked vertices in (b) are displaced according to the probe displacement. By initializing the position of the EM tracker with a known position on the phantom, the sensation of interaction with the computational model is created.

morph the model to the target surface by applying forces at specific vertices of the finite element mesh. By means of this approach we can achieve a better matching of the model and the patient if we assume the surfaces to be sufficiently close in the beginning. However, as a side note let us mention that this step is not amenable to capture the whole prone-supine deformation in the interior. In particular, for a known target surface, we propose the following approach.

Firstly, we represent the target surface in a three-dimensional image containing a one voxel wide crust of the target volume. From this representation, we calculate the three-dimensional Euclidean distance transformation.[5] The distance field is calculated on both sides of the body—the inside and the outside. On the distance field, we apply the same three-dimensional derivative of Gaussians kernel filter introduced above to generate a vector field. At each point in this vector field, the vector points into the direction of the shortest path to the target surface. For our purposes, we interpret the vector directions as forces we wish to apply to surface nodes of the model. Their magnitude is obtained by scaling the vector field with the Euclidean distance. By this, we obtain larger force contributions for vertices farther from the target surface and smaller force updates close to it. In all, calculation of the distance field is done in 2–4 sec depending on image size, and can easily be sped up using GPU implementations of distance transform (Schneider et al. 2010) and Gaussian filtering to be feasible for interactive updates of the surface morphing.

Given the finite element mesh that will be altered, we determine the distance vector at the position of all vertices that shall be displaced. A force proportional to the vector's length (which is again proportional to the distance from the target surface) is applied in direction of the vector. Forces are accumulative: after every simulation step, the force is read at the new image position, and added to the force already stored at the vertex. Note that due to the dynamics of the simulation, overshooting of the displacements might occur, so that the nodes of the morphed body pass through the target surface. In such cases, due to the nature of the force field, the vectors added in the next step imply a force into the opposite direction, and the approach still converges as expected. Our implementation of the algorithm has options to affect only subsets of the vertices, e.g. only surface vertices, or only those that have no other constraints (fixation, displacement, etc.).

[5]The chapter image shows a colorized and illuminated volume rendering of such a distance field.

4.2.3 Physical Plausibility

Three methods are proposed that aim at the goal of a higher level of physical plausibility in the simulation of large deformations of the soft breast tissues. The dynamic adaption of material parameters per finite element allows to emulate a non-linear material law. During dynamical simulation, it is important to control the amount of changes to the element parameters carefully, in particular paying attention to a regularized adaption of material properties. Therefore, in the next two sections, not only approaches to update local material properties are presented, but also control mechanisms that ensure a stable dynamic simulation over time.

Dynamically Adapting Local Material Properties

Real-world materials exhibit non-linear reactions to stress: when a certain stress level is exceeded, they react with a non-linear change of stiffness. In contrast, linear material properties in isotropic materials following Hooke's law are not capable to reflect this property. Our goal is to implement a simplified isotropic non-linear material law by adapting the per-element elastic modulus in every simulation step. Specifically, we adapt this method to simulate plausible breast deformations, which can be mimicked by a relatively soft base elastic modulus in combination with a stiffening of the elements under load.

For a clinical setting, a compromise has to be found between fast calculations and sufficiently realistic material behavior. The novel aspect of our contribution is the update of the per-element stiffness based on the stress it experiences in a given simulation step. Updating the element stiffness requires two components, which will be defined in the following sections:

Metrics to quantify element stress Two general approaches will be shown: Direct derivation of a stress norm from the stress tensor, and shape-based metrics that quantify anisotropic deformations.

Stiffness update rule A mechanism that converts the result of the metric into a stiffness update is required, which can be either a single equation, or an algorithm. Both variants will be explored.

Implementing different stress tensor norms and evaluating their suitability for the task will help to decide upon the most robust measure of element stress. It is physically convincing to look at the eigenvalues of the stress tensor, since the direction of the principal stress may be employed to model anisotropic non-linear material laws. Secondly, we consider the trace of the stress tensor as a rotation-invariant per-element metric, and lastly, the von Mises stress tensor norm (Bathe 2002) is used, which is likewise a per-element non-directional metric. For the second category above, we will look at shape-based metrics, assessing the deformation of the element with respect to its undeformed state directly.

The stiffness is then adjusted per element and time step, according to the stress it encounters. Since no prior work has approached the modeling of tissue stiffness with an approach similar to the described one, numerous variants of the real-time stiffness update have been implemented and evaluated to compare their characteristics and performance. Options include linear and non-linear stiffness updates, damped and direct updates, and saturation effects. Finally, a method is presented that ensures the stability of the solution by allowing for a reversal of arbitrary numbers of simulated time steps, introducing an "undo-redo" mechanism for dynamic simulations that has not been described before.

For better comparability, a simple and regular mesh has been set up, consisting of a ground plate and a soft box on top of it that has an adjustable size and number of tetrahedral finite elements and can thus be used to study effects on single elements, or macroscopic effects on a larger scale. The box has material parameters that are used for the breast model as well, i.e. a Poisson ratio (the ratio of transverse to axial strain of a stretched/compressed material; cf.

Fig. 4.3) of 0.48 and a base stiffness of 2 kPa. In the computer simulations, gravity direction and strength were used to control the magnitude of deformations.

Custom visualization methods allow to view the deformation along with characteristic attributes like, e.g. the principal stress directions and the von Mises norm.

Stress Tensor-Based Metrics

As one alternative, we propose a stress norm that takes the relation of the first eigenvalue to the sum of second and third eigenvalues of the stress tensor into account, weighted by a factor α. This is based on the consideration that the stiffening of tissues should occur predominantly when the tissue is exposed to stress in one dominant direction. Since the eigenvectors of the stress tensor capture the principal stresses in the directions of the planes orthogonal to the largest stresses, the directionality of the forces can be evaluated. The corresponding eigenvalues give the stress magnitude in the respective directions. Let $|\lambda_0| > |\lambda_1| > |\lambda_2|$ be the sorted eigenvalues of the stress tensor. If

$$|\lambda_0| > \gamma \cdot (|\lambda_1| + |\lambda_2|), \tag{4.24}$$

i.e. if the largest eigenvector is larger than the sum of the other two times a constant γ, then the relative element stiffness s_r^e for element e is set proportional to the largest eigenvalue:

$$s_r^e = \alpha\lambda_0. \tag{4.25}$$

The constant γ determines the level of anisotropy the stress of the element should exhibit. For $\gamma < 1$, stress in the form of pressure or tension from all directions causes similarly high stiffness updates as for unidirectional pressure or tension, while increasing values of γ require stress from one direction. If the condition in Eqn. (4.24) is not met, no update occurs, which effectively means that only from a certain threshold of anisotropy on, the stiffness is proportionally increased.

Second, the trace of the stress tensor, $\mathrm{tr}(\sigma) = \lambda_0 + \lambda_1 + \lambda_2$ is an alternative scalar norm of the stress tensor that is easy to implement and fast to calculate, and also cited as a measure of gross stress (Dick et al. 2009). In fact, since the stress tensor is symmetric, it can be diagonalized by means of the eigenvector decomposition, and the eigenvalues are the diagonal elements of the diagonalized stress tensor. It is invariant under transformations, hence it is also independent of stress directionality.

Lastly, the von Mises stress tensor norm,

$$\|\sigma\|_{\text{Mises}} = \sqrt{3\sum_{k=4}^{6}\sigma_k^2 + \frac{3}{2}\sum_{k=1}^{3}(\sigma_k - \bar{\sigma})^2}, \tag{4.26}$$

takes into account all elements of the raw stress tensor to calculate a scalar metric of the stress. It has originally been implemented to serve as a reference standard, since it is a well-established stress norm, and to compare its performance with all other methods. In fact, it proved to be superior to other metrics for the given purpose.

Shape-Based Metrics

In another line of thought, two per-element shape-based metrics have been calculated for comparison. In both, the shape of the element is compared in undeformed and deformed state. Note that the five undeformed tetrahedrons comprising one voxel cube are dissimilar in shape, such that a more general implementation had to be found. The first implementation calculates

for each vertex of one element the largest relative change of distance to the three opposite edges, i.e. the maximum from twelve distances per element. Based on the largest relative displacement, the relative element stiffness is updated. Each of the distances is calculated from three vertices of the tetrahedrons according to

$$\mathrm{d} = \frac{|(\mathbf{x}_0 - \mathbf{x}_1) \times (\mathbf{x}_0 - \mathbf{x}_2)|}{|\mathbf{x}_2 - \mathbf{x}_1|}. \tag{4.27}$$

The second approach calculates for each vertex in one element the maximum relative change of distance of the vertex to the opposite face, which is the maximum of four distances per element. With the unit normal of the face given by

$$\mathbf{n} = \frac{(\mathbf{x}_1 - \mathbf{x}_0) \times (\mathbf{x}_2 - \mathbf{x}_0)}{|(\mathbf{x}_1 - \mathbf{x}_0) \times (\mathbf{x}_2 - \mathbf{x}_0)|},$$

the distance d becomes

$$\mathrm{d} = \mathbf{n} \cdot \mathbf{x}_0 + \mathrm{p} \tag{4.28}$$

Both measures have been implemented to calculate the distance orthogonal to edge or face, and from the barycenter. Calculations from the barycenters greatly decrease computational complexity, at the cost of reduced robustness in case of degenerate tetrahedrons. Fig. 4.6 shows a close-up of the result obtained with a shape-based update, and Fig. 4.7 further examples with other stress metrics and several update mechanism. A deformation of this magnitude has not been possible to achieve without the local stiffness update proposed here. The simulation would have resulted in physically implausible solutions that visually look as if the simulated body has permeated its own surface, which is due to a number of inverted elements.

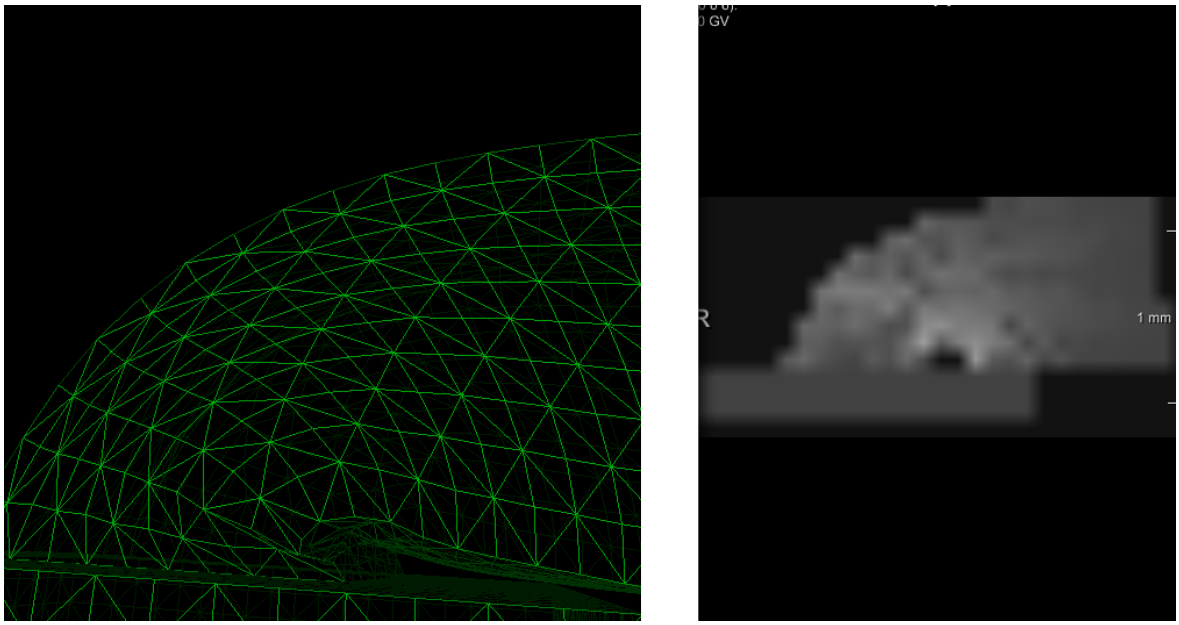


Figure 4.6: Result of shape-based update (here: absolute update based on vertex-to-face deformation change). Left: close-ups of the bending of the mesh under extreme gravity. Right: a central slice through the relative stiffness values observed inside the mesh.

Direct Stiffness Update

Regardless of the chosen stress tensor norm or shape based metric, a stiffness update has to be derived from the scalar result. A natural choice coming to the mind is a power law, reflecting

that real-world materials behave non-linear under compression, e.g., the stiffness increases exponentially when the stress grows. The initial implementation thus raises the stress tensor norm or shape based metric to a user defined power, which quickly turned out to be affected by strong oscillatory behavior during the dynamic simulation.

A linear dependency between stiffness and the metric proved to be more stable and is the basis for all further descriptions. The oscillations for the first time observed visually in the power law stiffness update model led to the development of the following stiffness update damping mechanism that effectively prevents oscillations in the dynamic simulation.

Initially, for all metrics an absolute update mode had been implemented. In the absolute update mode, the relative stiffness is set to $s_r = 1 + \alpha \cdot \|\sigma\|$ with α a user-defined factor, and $\|\sigma\|$ the chosen stress norm.[6] This might lead to oscillations, eventually only after considerable time, because the explicit updates of the per-element elastic modulus (stiffness) in the simulation steps increases the stored elastic energy in the linear elastic model and thus affects the stability of the approach. In other words, the internal forces in the body are mainly proportional to the elastic modulus, and thus updating this value while keeping the vertex displacements constant causes higher elastic energy. Therefore, we propose a dynamic simulation model exhibiting damping and inertia. In this model, the elastic modulus is updated in increasingly smaller steps, which is accomplished by an automatically adjusted stiffness damping factor.

Damped Stiffness Update

The first category of damping methods that have been explored are time-dependent damping mechanisms. A function $\omega(t)$ can be designed that decreases from 1 to 0 when t increases. Let $s_r(t)$ be the relative stiffness value calculated from the stress measure at time t, and $s_r(t-1)$ be the previous value that was set for the element. The new value is then not taken directly, but adapted to be

$$s_r(t) = (1 - \omega(t))s_r(t-1) + \omega(t)s_r(t). \tag{4.29}$$

By this setup, the weight of the relative stiffness the element had in the previous time step prevails with increasing t. For the time-dependent damping function $\omega(t)$, different functions have been tested. Exponential damping is often encountered in nature, thus we implemented exponential damping for the stiffness update:

$$\omega(t) = e^{-\alpha\pi t^2}. \tag{4.30}$$

Alternatively, a simplified damping that is easily scaled such that it assumes 0 for any value $t_0 > 0$ is given by

$$\omega(t) = \frac{\left(\cos\left(\frac{\pi t}{t_0}\right) + 1\right)}{2}. \tag{4.31}$$

In a second category, the damping mechanism only depends on the average per-element stress change. The advantage is that no maximum time t_0 has to be set at which the stiffness update stops, but rather the update magnitude is proportional to the average stiffness change in the

[6] In fact, for oriented stress tensor norms, like for example the shape based metrics, α can be set to different values depending on the direction to emulate anisotropic material.

elements. This method has proven to be superior to all others, as will be explained in the results section (Sec. 4.3.1). This damping method calculates the average absolute per-element relative stiffness change over all mesh elements in one time step and allows in the next time step a per-element stiffness update according to $s_r(t) = (1-\omega(t))s_r(t-1) + \omega(t)s_r(t)$ from Eqn. (4.29), where $\omega(t)$ this time is the average absolute per-element relative stiffness change in the last time step, cropped to 1.0 to avoid amplification of the update.

A test application was designed to examine the per-element reaction on stress over time. In this application, it is possible to set start and end values for the parameters controlling the simulation and the extension, and to increase/decrease these parameters during simulation. This is intended to avoid abrupt jumps in external forces or updates. The application then carries out the desired number of simulation steps. Since each finite element has a state variable, this can be recorded after each simulation step. Then, the parameters are varied according to the prescriptions, and the next iteration starts.

From one run, we thus obtain a time resolved recording of the state of all mesh elements during the simulated time. The most important parameter in the state variable that is being recorded is the relative per-element stiffness resulting from the stiffness update. It hence traces the series of stiffness updates. Further details and explanations will be provided in the results section below.

Phantom Experiments

In the phantom experiments, it has become obvious that the scalar von Mises stress norm is more robust than the directional eigenvalue based norm, in which the largest eigenvalue is related to the sum of the other two eigenvalues. The stress tensor trace figured out to be unsuitable for the task. Still, while the results with the von Mises stress norm and the eigenvalue-based stress measure looked promising visually, more or less pronounced oscillations were observed after longer simulation runs and have been addressed with the damping mechanisms described above. We have studied the onset of oscillations empirically by simulating forces acting on a bar.

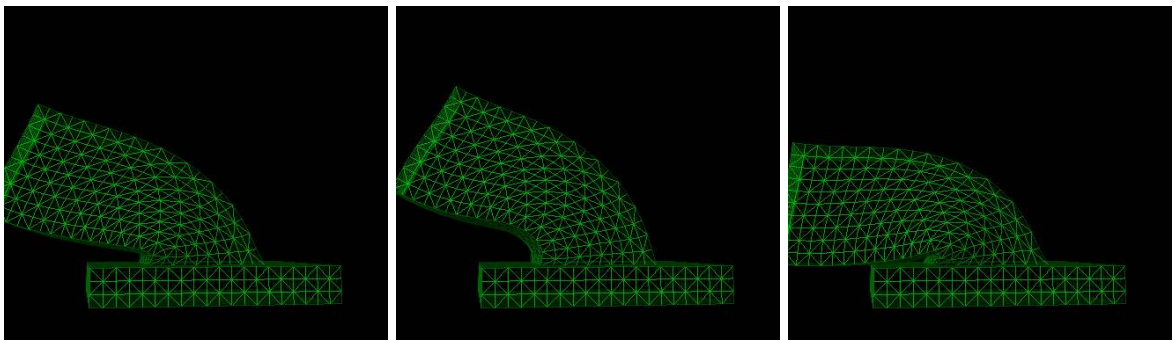


Figure 4.7: A comparison of the influence of different stiffness update mechanisms. Per column, a different update mechanism is applied. (a) absolute update; (b) time-damped update; (c) adaptive damping update.

The time-dependent update damping shown in Eqn. (4.31) appeared to be satisfactory at first with t_0 set to any duration that allowed for 20–40 simulation steps. Macroscopically the deformed mesh came to rest. Looking at the stress field revealed slightest changes to the stiffness, however, and continuing the simulation without damping eventually led to increasing oscillations and finally macroscopic instability of the simulation expressed in physically impossible solutions where elements invert and the mesh penetrates itself (compare Fig. 4.7 and Fig. 4.9 for examples).

The most robust damping method turned out to be the continuous update proportional to the average element stress change. Fig. 4.7 illustrates the outcomes of the simulation after

comparable simulation duration. While the destabilizing effects of insufficient damping might occur a long time after superficially the simulation appears to be stable, this damping scheme for much longer periods of time shows no sign of destabilization in any of the tests.

For a closer analysis, we tracked the update of relative stiffness, averaged over all elements, from simulation step to simulation step. We combined the update mechanisms (absolute, time-damped, adaptively damped) and stress norms (von Mises, eigenvalues, shape change). It is important to note that the resulting evolution of relative stiffness update magnitude is largely independent of the starting configuration and parameters, hence no normalization or thoughtful selection of the update factor, α, is required. We have in these experiments observed that only methods independent of the temporal evolution of the simulation are successful in stabilizing the dynamic simulation over time. This is against intuition, since it would be expected that for a macroscopically resting body, the internal forces are in equilibrium. Our experiments show, however, that this is not the case, although the relative stiffness updates are several orders of magnitude smaller in this state than during macroscopically observed deformation. In time-dependent damping, these minimal numerical errors quickly build up oscillations, eventually causing strong loads in some elements which again react with inversion. The adaptive damping, on the other hand, makes slight adjustments to the stiffness and is thereby capable to maintain a state of observed rest.

By resampling the relative stiffness of each element to a resolution that depicts the tetrahedrons of the volume mesh, a checkerboard-like pattern in the artificial phantom becomes visible (compare Fig. 4.8 for an illustration). The stress norm jumps in magnitude from one element to its neighbor, possibly causing an ill-posed system matrix that causes numerically inaccuracies. Consequently, oscillations and break-down of the mesh might be the result, conforming with the observations. The analysis tools employed in these experiments provided insights into the stability of the different stress norms. In particular, the shape-based metrics using the simplified barycenter-based distance calculations were discarded because they exhibited the strongest element-to-element relative stiffness jumps conforming to their inferior stability noted before. Because the shape-based metrics using orthogonal distances are computationally more expensive than the von Mises stress norm, further developments settled with the von Mises norm combined with the adaptive damping mechanism.

Patient Data

We have assessed the performance of the implementation on volunteer data sets sampled to $5 \times 5 \times 5$ mm resolution, yielding varying numbers of elements (cf. Table 4.1 on page 179). The resolution can be reduced to $10 \times 10 \times 10$ mm to increase speed by almost an order of magnitude. For the practical application, only the simulation of either the left or the right breast is required; a restriction of the field of view to this area would also double the speed.

The visual and qualitative results with and without the stiffness update are compared in Fig. 4.9. Especially, for one single voxel, we plot the von-Mises stress norm over the time. Note that the color map and graph scale are the same in the top and bottom row of the figure. While without stiffness update, elements begin to invert after a number of simulation steps, this is avoided when the per-element stiffness is updated proportional to its von-Mises stress. This calculation is fast: even in a non-optimized, per-element serial computation it only increases the simulation time by 25% (cf. Table 4.1 on page 179).

Most importantly, the goal of stabilizing the simulation during application of strong forces has been achieved. It is now desired to achieve the same robust behavior without having to determine suitable parameters for the above algorithms. The following mechanism addresses this requirement.

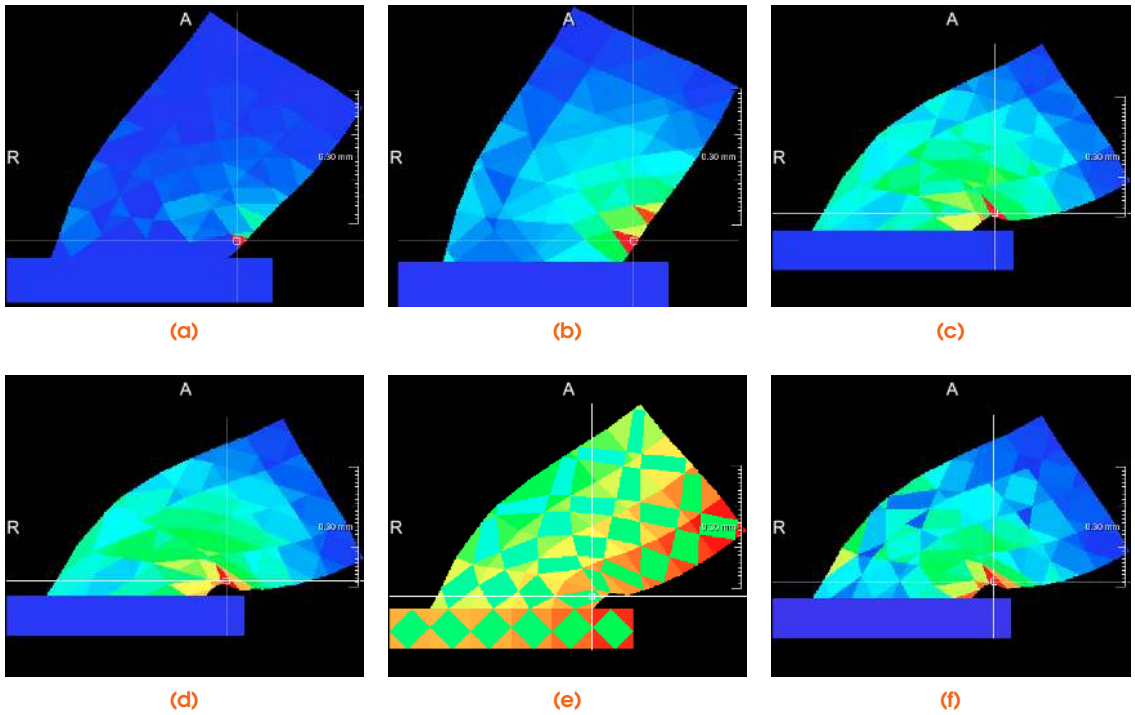


Figure 4.8: A qualitative comparison of the stress metrics. The images show two-dimensional cross-sections through a map of element stress values according to the following metrics: (a) eigenvalue; (b) von Mises; (c) vertex-edge barycentric; (d) vertex-edge orthogonal; (e) vertex-face barycentric; (f) vertex-face orthogonal.
The HSV color scale from blue to red encodes the element's relative stiffness, corresponding to its stress, always covering the full range. For all simulations, the same magnitude and direction of gravity has been assumed. Correlation with the stability of simulations, those metrics performed best where few neighboring tetrahedras developed large differences in relative stiffness over time. (e) and (f) stand out with the largest differences between neighbor tetrahedras. Considering stability and computational cost, the von Mises stress norm (b) is a good compromise, although neighboring elements show very different magnitudes of relative stiffness.

Improved Non-linear Material Modeling Using Time Step Reversal

Our approach to model non-linear material behavior is motivated by the observation that in a model coarse enough for real-time update speeds, a very low Elastic modulus is required to account for the dominant contribution of fatty (loosely coupled) tissue in breast deformations. However, this yields unrealistic behavior for large deformations, manifested in inverted elements at contact points or in folds. Locally stiffening the material at these points effectively prevents this behavior as described before and published in M. Harz et al. (2011a,b).

To model the non-linear material law aspect responsible for a macroscopic stiffening of material under stress, we use the von-Mises stress tensor norm (Bathe 2002) to quantify the element stress (see Eqn. (4.26) above).

The von-Mises norm provides a rotation-invariant scalar metric of the overall stress variation imposed on an element. We assume that breast tissue is composed of lumps of material that can move about with little friction. Macroscopically, this requires a small elastic modulus to observe realistic deformations. However, under compression real-world material exhibits a non-linear behavior, since now the stiff tissue parts determine the material behavior rather than the low-friction movements, i.e., macroscopically the stiffness increases with the stress.

As proposed before, the element relative elastic modulus can be modified per simulated time step according to $E_r = 1 + \alpha \cdot s_{VM}$ with $\alpha \in \mathbb{R}$ a user-defined scalar factor greater than zero and $s_{VM} \in \mathbb{R}$ the von Mises stress norm.

These explicit updates of the per-element elastic modulus have to be performed carefully to ensure stability of the approach, because internal forces in the body are mainly proportional

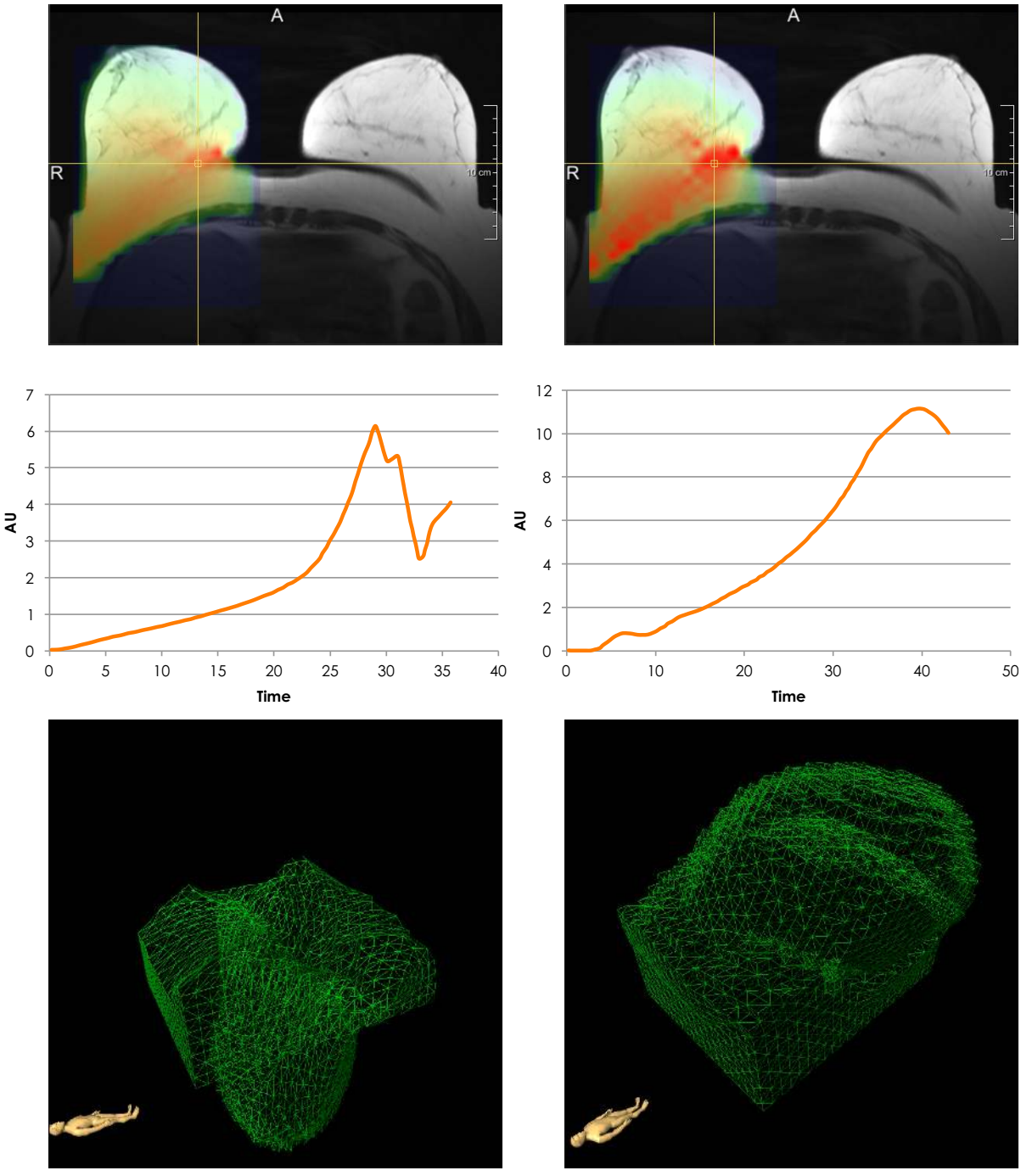


Figure 4.9: LEFT COLUMN. Top: Axial slice through the breast volume, with a color overlay indicating the stress (von Mises norm) that the elements are experiencing. No stiffness update is applied to the elements. Middle: The graph shows a plot of the von Mises stress over a number of simulation steps for the voxel indicated by the crosshair (in artificial units). Note the peak at the 29th simulation step. Bottom: Shortly after the peak, the deformed mesh begins to exhibit physically implausible behavior.
RIGHT COLUMN. Same model, this time with stiffness update. The stress that the elements can take with the stiffness adapted according to the von Mises stress norm saturates on an almost doubled level and at a later simulation step. A physically realistic and stable deformation is obtained when for example simulating the supine positioning of a patient during surgery using a FE model derived from an MRI acquired in prone positioning.

to the elastic modulus, and thus updating this value while the deformation is not modified increases the stored elastic energy. Instead of limiting the elasticity update by damping, we propose a novel mechanism that adapts the α such that instabilities are avoided at all. Since α is no longer fixed, two things are required: (1) a mechanism to try out a guess for α, and (2) a metric for the goodness of α.

Our metric is a binary decision on the integrity of the mesh, indicated by the lack of inverted finite elements. Inverted elements occur when the local force acting on an element is larger than what the element can bear. Then, physically implausible vertex positions are the only numerically valid solution to the FEM system of equations. Consequently, our approach to avoid inverted elements is to revert a simulation time step, increase α, and repeat the time step. This can be done iteratively until the first suitable α is found that stiffens the elements enough to withstand the forces in the simulation time step without inversion. A second option to overcome element inversion is to re-mesh the model into smaller elements, either locally or globally. Both are no feasible options in our performance-oriented framework since they both induce higher computational demands, either spent in re-meshing or in each simulation step to solve the larger system of equations.

The time step reversal is of course only possible with a time integration scheme that allows for a storage of all characteristic information of the mesh in each time step, such that it can be reverted. This is for example the case for dynamic Euler integration where all information is contained in any two consecutive displacement fields. Consequently, it is sufficient to store u_{t_0-1} and u_{t_0} to revert to the simulation state at time t_0. Our actual implementation even allows for an "undo" history, by using n time step reversal plugins that store a simulation state in a round-robin fashion. Then, the simulation can be reverted to any of the n stored states, and continued from there.

Once the simulation proceeds, α can be decreased again to enable a maximum movement of the breast tissue. Note that some careful control has to be added to avoid consecutively increase and decrease of α. However, our implementation does not dependent on user-provided parameters. Instead, from an initial α set by the user, the simulation starts, and only when an inverted element is detected, α is adjusted as described above. Now, when relaxation of α starts, we detect oscillations: as soon as α has been decreased, increased, and decreased again, we adjust the speed of α changes proportional to the magnitude of α. This converges to a stable α. For performance reasons, α should initially be set to a knowingly too high value, and the algorithm will decrease it as described.

Note, however, that a new α has to be found if any of the boundary conditions change, in particular, if the direction of gravity changes, which is relevant in our scenario. In our application the deformation simulation does not require to run at a constant update rate, and especially the computationally challenging deformation from prone to supine can be performed before the intervention.

Emulating Breast Tissue Sliding on Chest Wall. While in previous work Dirichlet boundary conditions were applied on the vertices on the breast-chest interface, we herein propose to replace the fixations by displacement constraints that are updated in each simulation time step similar to the approach of Georgii, Lagler, et al. (2010). This enables us to emulate a sliding of the tissue on the chest wall, governed by friction, with the possibility to include inertia effects.

This is implemented by computing the forces for all vertices that have a displacement constraint, and projecting the resulting force vector into the plane locally tangential to the chest wall. To find this tangential plane, prior to the simulation a vector field is calculated from the chest wall mask obtained in the segmentation step. The image I is convoluted with two smoothed Gaussian derivative filters in x and y direction,

$$G_\sigma^x = \frac{\partial G_\sigma}{\partial x} = x e^{-\frac{x^2+y^2}{2\sigma^2}}$$

and similarly for the y and z directions. From the derivatives, the gradient direction can be computed. Let $\frac{\partial I}{\partial x} = G_\sigma^x * I$ be the convolution of the image I (similar for y), then the gradient direction is

$$\phi = \mathrm{atan2}(\frac{\partial I}{\partial y}, \frac{\partial I}{\partial x}),$$

for the xy plane, and similarly for the other planes, and the magnitude is the square root of the sum of squared directional derivatives.

For our purpose, $\sigma = 2$ mm is chosen because this is in the order of twice the in-plane voxel resolution of current morphological MRI sequences, such that a mild averaging is obtained that reduces the staircase effect of the binary mask borders.

Since in practice not all tissue will slide equally freely, and since for the intended application, all sliding will have an upper limit of movement, limiting factors are implemented to tune the sliding to match the empirically observed behavior. Of course, this calculation can also be bypassed to allow arbitrary sliding governed by friction and shape alone.

1. **Exponential decay** imposed on the displacement, according to $d = d + D_f e^{-a\|d\|}$, where d is the displacement so far, $D_f = f dt^2/m_v$ is the displacement imposed by the current force f on the vertex with mass m_v, dt is the size of the time step, and a is a factor influencing the speed of decay.

2. **Linear decay** imposed on the displacement, according to $d = d + (a + 1)D_f$, where the slope a should be negative and governs the speed of decay.

3. **Gaussian shaping** of the permissible shift to model the observation that the breast will not slide near the sternum, nor near the axilla. This is controlled by a bimodal 2-dimensional Gaussian function

$$z = Ae^{-\left(\frac{(x-x_0)^2}{2\cdot\sigma_x\cdot\sigma_x} + \frac{(y-y_0)^2}{2\cdot\sigma_y\cdot\sigma_y}\right)} + Ae^{-\left(\frac{(x-x_1)^2}{2\cdot\sigma_x\cdot\sigma_x} + \frac{(y-y_1)^2}{2\cdot\sigma_y\cdot\sigma_y}\right)},$$

   and calculating the sliding displacement d as $d = d + zD_f$. The two modes are preferably centered at the breast centers in the xz-plane, and should have σ set such that the extend of the breasts is covered. To this end, the center of the breasts can be determined automatically based on the nipple positions known from our segmentation algorithm, while an appropriate value for σ is obtained by simply setting it to half of the diameter of one breast which again is roughly one quarter of the image's bounding box.

It is particularly noteworthy that this implementation does not require an analytic model of the surface, but works with any given geometry, in particular one derived from the patient images. We use the segmented pectoral muscle surface to derive the sliding surface, while in some publications, a coarse approximation to the chest wall is used, oriented on the air-filled volume of the lungs and approximated by a cylinder (Han, J. Hipwell, et al. 2011).

4.3 Applications in Breast Cancer Care

4.3.1 Surgery Preparation Using Efficient Breast Deformation

Breast cancers that are detected early, for example by biopsy and MRI imaging, and are determined to be locally invasive only, are eligible for breast conserving therapy, where only the cancerous tissue together with a safety margin is excavated. These therapy approaches, however, suffer recurrence rates from 8% to 14%, and specimen pathology reveals positive (cancerous) margins in between 17% to 59% of the cases (Coopey et al. 2011). Additionally, although the procedure is comparatively easy and routine, poor cosmetic outcomes are frequent complications, which is due to the way of indicating the target area with metal wire guides that are inserted under mammographic or MR image guidance, and that serve as a surgery path indicator to the surgeon. In effect, depending on lesion position and wire insertion direction, this can cause suboptimal access paths leading through unnecessary long parts of healthy tissue instead of shortest paths from skin incision to lesion. Additionally, the required safety margins in this setup are large, and still not always met when the diseased area is invisible intraoperatively.

Several attempts have been made to improve surgical outcomes, including incorporating intra-operative ultrasound imaging. Some authors propose robot aid for the placement of biopsy needles (Mallapragada et al. 2011), but require a fixated breast where the needle is placed. Approaches that help the surgeon to navigate more safely are more challenging and less developed. The study of Alderliesten et al. (2010) uses MRI-based navigation and tackles the problem of tracking the breast surface from a supine MRI scan to the surgery position, but neither support navigation in open breast surgery nor account for the much more challenging deformations from the prone positioning. Support of open surgery is attempted in the ultrasound-based approach presented by Sato et al. (1998). Their approach, however, requires a tracking equipment to be positioned in the operating room, and will display the superimposed target area only in a computer monitor.

Another approach is to simulate the deformation of the breast from the known shape seen in the diagnostic MRI scans (in prone positioning, meaning face-down) into the shape seen in the face-up, supine, positioning on the surgery table. This approach is commonly known as the prone-to-supine deformation and has been tackled in several publications. To understand the difficulty to simulate this deformation, a brief revisit of the breast anatomy with emphasis on the tissue composition and biomechanical behavior of breast tissues is required (please refer to Fig. 1.7 on page 25). The breast shape is predominantly governed by four tissue types: (1) the skin, which is nearly incompressible but capable of virtually unconstrained shear movements; (2) the fatty tissue, which can be conceptualized as small lumps that are not interconnected and hence behave like a viscous fluid; (3) the glandular tissue, namely milk ducts and lobules, which are connected to the nipple and partially to the fatty and stromal tissue surrounding them, and are not extensible; and (4) the Cooper ligaments, a connective tissue reaching from the subcutaneous tissues to the chest wall and accounting for the breast shape for the most part. They are not connected to any other tissue. In MRI, the fatty and glandular tissues can be seen and segmented, sometimes also the skin layer can be estimated. Cooper ligaments are not visible due to the limited image resolution.

Some research focuses on augmented reality approaches, displaying target structures on the skin of the patient, after calculating the transformation from prone to supine. The contributions of Carter et al. (2006) and Palomar et al. (2008) address many modeling aspects that we are dealing with as well. Carter et al. (2006), however, uses supine images of the patient to define the target of deformation of the diagnostic (prone) breast MRI image, and the work of Palomar et al. (2008) has a different (and perhaps less complicated) objective of simulating the deformation from supine to standing position.

Large Deformations of Breast Tissue: Overview

We use prone MRI scans to automatically create a patient-specific model of the breast to model the deformation from the prone position (using breast coils) to the supine position used for surgical procedures. We extend an approach that combines automatic segmentation with the generation of multi-resolution tetrahedral meshes (M. Harz et al. 2011b), which are an effective basis for fast, efficient deformation simulations.

Our goal is to improve the modeling of breast deformations incorporating isotropic non-linear material laws (ibidem). We will use the per-element elastic modulus update described above to mimic a non-linear material law. We keep with the assumption of breast tissue homogeneity, since, as described above, for a physically realistic simulation, the required level of detail is beyond computational feasibility for our application. Instead we will enhance the efficient deformation simulation approach with algorithms that improve the plausibility and accuracy of the results while keeping the computational efficiency of the approach. Therefore, the approach is suitable for extensions towards intra-surgery deformation tracking, which is one of the goals of our work.

1. We utilize the time step reversal mechanism that effectively avoids inverted elements when adjusting the per-element elastic modulus in a simulation step (see Sec. 4.2.3).
2. The simulation of frictional sliding between interfacing bodies governed by vertex forces is employed to increase the physical plausibility and along with this the matching of simulated and measured supine breast shape (see Sec. 4.2.3).
3. We utilize the framework extension to set vertex forces per element to deform the breast towards the measured supine target surface (see Sec. 4.2.2).

Our overall method maintains its computationally beneficial properties: It avoids complicated measurements to determine non-linear material parameters, and it can be integrated into existing corotated linear-elastic code. We approximate microscopic material tissue properties on a macroscopic scale by modifying the material under load to emulate natural behavior. Inverted finite elements are avoided by the time step reversal mechanism presented.

Using the Framework: Methods

Our proposed system starts with MR images taken in the prone position that are automatically segmented into deformable and fixed tissues using the methods proposed by L. Wang, Filippatos, et al. (2011). These methods have been adjusted to provide necessary information for the sliding approach, as it will be described in more detail in this section. The segmentation is used to setup the FE model. A highly efficient FEM-based breast deformation approach is then used to simulate the shape change from prone MRI to the desired patient position, even allowing for real-time simulations of breast/patient repositioning at moderate mesh resolutions, which is relevant for the surgery scenario depicted above. An overview of the interplay of the single algorithmic steps is given in Fig. 4.14.

Our overall approach is designed to work in clinical practice; hence we have devised tools to support automatic patient-specific model generation from image acquisition to initialization of simulation, including the segmentation into the relevant tissues. This goes beyond other work, where the meshing is automatic, but requires undisclosed efforts for segmentation (Ruiter et al. 2006). The algorithmic methods described in the following sections are novel developments that have been applied to volunteer data. For a thorough analysis, we additionally describe results based on computational phantoms that exhibit characteristics suitable to demonstrate the benefits of the proposed methods, and sketch the benefits on volunteer data by example.

Data. The software phantom used for most developments is a half-ellipsoidal shape that is modified to resemble prone and supine shapes of the breast, while preserving the volume of the phantom. These shapes are easy to generate in any size and resolution. Also, in such a computational phantom, the amount of deformable tissue located between the breast tissue and the rigid, non-deforming rib cage can be changed from a thin layer to a thicker layer.

In addition to the phantom data, volunteer data acquired on a 3.0 T Siemens Trio scanner using a 16-channel breast coil were available. Standard non-fat-suppressed T2-weighted breast MRI series were obtained from five volunteers with voxel sizes in the order of $1 \times 1 \times 5$ mm, once in prone (facing down) and once in supine (facing up) position.

Model building. All data sets were automatically segmented into rigid and deformable tissue, where the breast parenchyma and the adipose tissue were considered elastic, and the thorax was considered fixed. From the segmentation algorithm, a three-dimensional gradient field is also automatically estimated on the surface separating the chest and the breast and will be used to define the sliding behavior. A volumetric tetrahedral mesh was generated from a isotropically down-sampled version of the breast tissue mask (about $(2.5\text{ mm})^3$ voxel size), resulting in meshes consisting of between 50k and 300k elements, depending on the actual size of the breast and the field of view.

All preprocessing steps supporting the generation of patient-individual models have been unified in a convenient application, together with the possibility to define landmarks, actually perform the deformation simulation, and visually trace and digitally record various metrics during simulation. All performance figures that will be presented in the results have also been produced with this application.

Simulation. For the simulation, the three framework extensions described before have been employed. According to the von Mises stress tensor norm, the per-element stiffness is updated in every simulation step. This is controlled by the adaptive update magnitude constraint that makes the local updates depend on the overall change. Frictional sliding on the pectoral muscle boundary is unconstrained if the full volume including both breasts is simulated, but for simulations only acting on the target breast, the medial vertices are fixated (Dirichlet boundary conditions). After dynamic simulation until the average vertex displacement update is below a threshold (visually no more shape changes are noticeable), we apply forces to the surface nodes of the mesh that are directed towards the supine target shape. This last step is meant to emulate the scenario in which a target surface during surgery is provided by optical scanning, and to verify the suitability of the approach.

How Large is Large? How Fast is Fast? Results

One paramount goals is to maintain the performance of the underlying corotational finite element framework. Therefore, we have evaluated the performance on several data sets with 50k to 300k elements, and observed the expected linear scaling of execution time with the number of mesh elements for one time step. Table 4.1 lists timings for the simulation including the elastic modulus update and the sliding on the chest wall.

The timings in Table 4.1 were measured on an Intel Core i7 Quad 2.66 GHz mobile CPU. Note, that the timings are dominated by the time required by the simulation step. The additional time required to update the elastic modulus of each element and to compute the sliding on the chest wall is small as can be seen in the respective column (up./sl.). Note that the time step reversal increases the number of total simulation steps and is therefore not reflected in the Table. The simulations are not necessarily converged after 25 iterations, though in our experience the deformations imposed by gravity are typically settling after 0.5 to 1.0 sec wall time, which is after 15–30 iterations at time steps of 0.033 sec.

Table 4.1: Timing statistics. 25 simulation time steps (sim.) were performed, and the same number of per-element elastic modulus update and sliding operations (upd./sld.). Timings were measured on an Intel Core i7 Quad 2.66 GHz mobile CPU.

Elements	Sim. (ms)	Upd./Sld. (ms)	Total 25 steps (sec)
47,895	189	35	5.6
99,390	400	73	11.8
158,340	644	109	18.8
158,340	628	109	18.4
313,335	1432	239	41.7

In the following, we independently describe the benefits contributed by the three proposed extensions to provide easier assessment.

Elasticity Update with Time Step Reversal. The time step reversal based elasticity update has been developed and evaluated on an artificial data set prior to the application to breast MRI data. In these experiments, the proposed update criterion was able to produce plausible deformations of the model. Based on literature values, a Poisson ratio of 0.48 and a base elastic modulus of 2 kPa was chosen to characterize the breast tissue material (Han, J. Hipwell, et al. 2011). For the example given, the fraction of time reversal steps was about 20% of all simulation steps until convergence to a stable state. This can be assumed to be a worst case scenario: α was initially chosen one order of magnitude too high and the α decrease factor (and consequently also the increase factor for reverted time steps) was set to a high value to arrive at a convergence fast. Under these conditions, the deformation simulation cannot keep pace with the elasticity drop caused by quickly decreasing α, such that inversion occurs several simulation steps "too late" and have to be reverted while α is increased again.

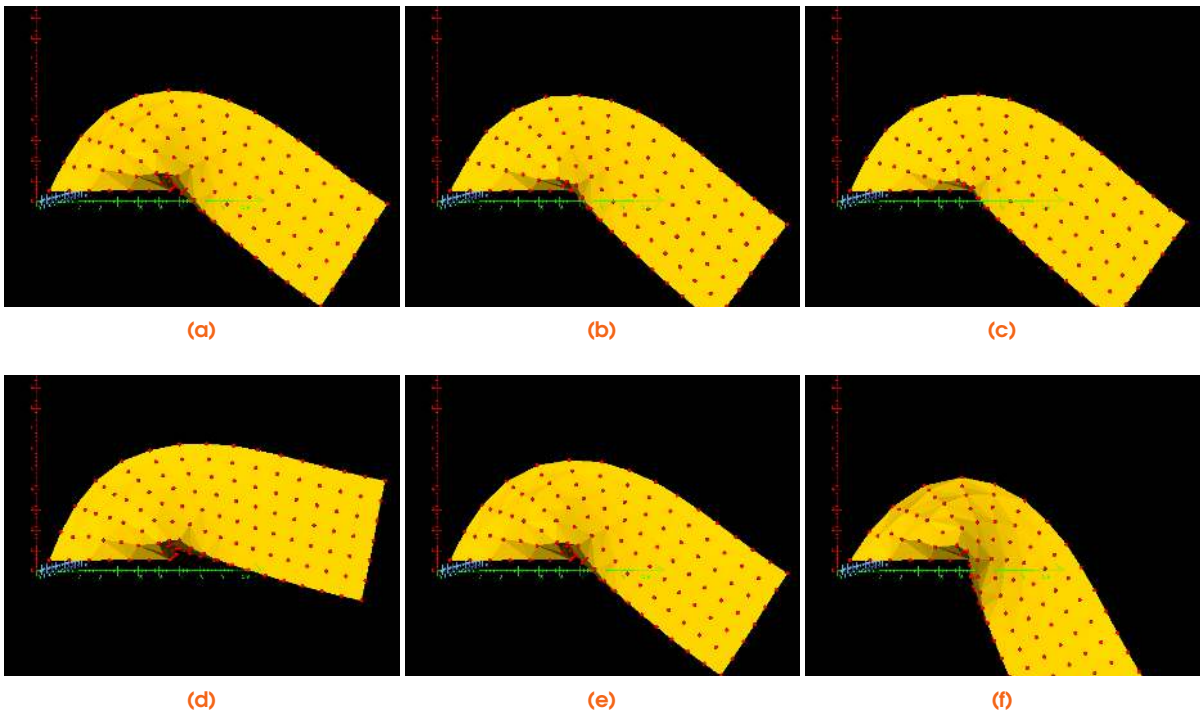


Figure 4.10: Maximum deformation with α automatically adapted for Poisson ratios of 0.48, 0.38, and 0.28 using a fixed elastic modulus of 2000 kPa (a-c) and decreasing elastic moduli of 3000 kPa, 2000 kPa, and 1000 kPa at a fixed Poisson ratio of 0.48 (d-f). The proposed algorithm is able to adapt α to all parameter settings automatically, and from a common start setting.

Our approach to find a suitable factor α by decreasing and increasing its value based on

observed inverted elements has shown a robust behavior. We have tested the method for different combinations of Poisson ratios and elastic modulus as shown in Fig. 4.10. In practice, the parameters can be set to a robust general preset and the user will not have to adjust them. With this mechanism, the element elastic modulus update supports continuing numerically stable dynamic simulations at minor additional computational expenses. In contrast, without a dynamic adaption of α, the depicted deformations would not have been possible without physically implausible solutions.

Sliding. Sliding has been tested on artificial and volunteer data. We have generated several computational phantoms ranging from simple proof-of-concept models to phantoms modeling two breasts on a curved chest wall to assess the behavior of the sliding implementation in detail.

In Fig. 4.11, images of a volunteer with a cup size of C to D are considered. The volunteer has been imaged in prone and supine position. We have chosen this volunteer for presentation since larger breasts are obviously more difficult to model than smaller or stiffer breasts. Hence, we expected that the robustness and physical plausibility can be assessed more rigorously in this data. We have generated orthogonal cross-sections of both breasts after deformation simulation including the sliding algorithm and have compared the result to the supine data at a matching cross section. Note how the slightly bent parenchymal structure visible in the transversal projection of the anatomically left breast (the breast to the right hand side in the top cross section in Fig. 4.11b) is deformed into the simulated supine position shown in Fig. 4.11d.

Surface Morphing. Surface morphing has been developed using a breast software phantom consisting of a half ellipsoid that has been generated analytically in two configurations: one resembling the prone breast shape in the breast coil, and one resembling a stretched-out configuration in supine positioning. This is not the intended scenario for the application of the surface morphing algorithm, but rather a setup that is intended to display the performance under extreme conditions.

The configurations were constrained to have similar volumes and are both attached to a planar surface, so that the critical edge observed in MRI scans with breast coils is also modeled. To make the model as realistic as possible, we have moved the undeformed "prone" model out of the center of the "supine" model, and also applied gravity in a direction oblique-angled (not orthogonal) to the plate. The back vertices of the planar surface have been fixated (Dirichlet boundary condition). The two structures can be seen in Fig. 4.12a, and consist of roughly 100k tetrahedral elements. Elasticity and Poisson ratio have been set to values similar to the volunteer breast models described above.

Gravity has been simulated before the force field was applied to the surface vertices. But by means of our surface morphing, the target shape is closely approximated without any parameter adjustment, and in addition the final breast phantom volume is close to the volume of the target shape. Note in particular how the sharp edge between ground plate and breast model is adapted to the target shape.

On the volunteer data seen in Fig. 4.11, we have continued the simulation using surface morphing, but with the sliding algorithm disabled. Fig. 4.13 illustrates this. Morphing in our volunteer data set leads to about 30% volume loss of the model to achieve a match of both surfaces, due to surface forces acting against Dirichlet boundary conditions of the nodes that are attached to the chest wall instead of again sliding. When looking at the prone and supine MR images of our volunteer data sets, it is seen that in the supine scan the breast tissue is distributed very differently, both in cranio-caudal and in lateral orientation.

In particular, for the diagnostic prone DCE-MRI, as much as possible axillary tissue is imaged as well to assess lymph nodes, and is part of our segmentation mask. In the supine position, this tissue is no longer situated behind the visible breast surface, which gives reason for the change of volume observed when matching the surfaces. This might indicate that the breast

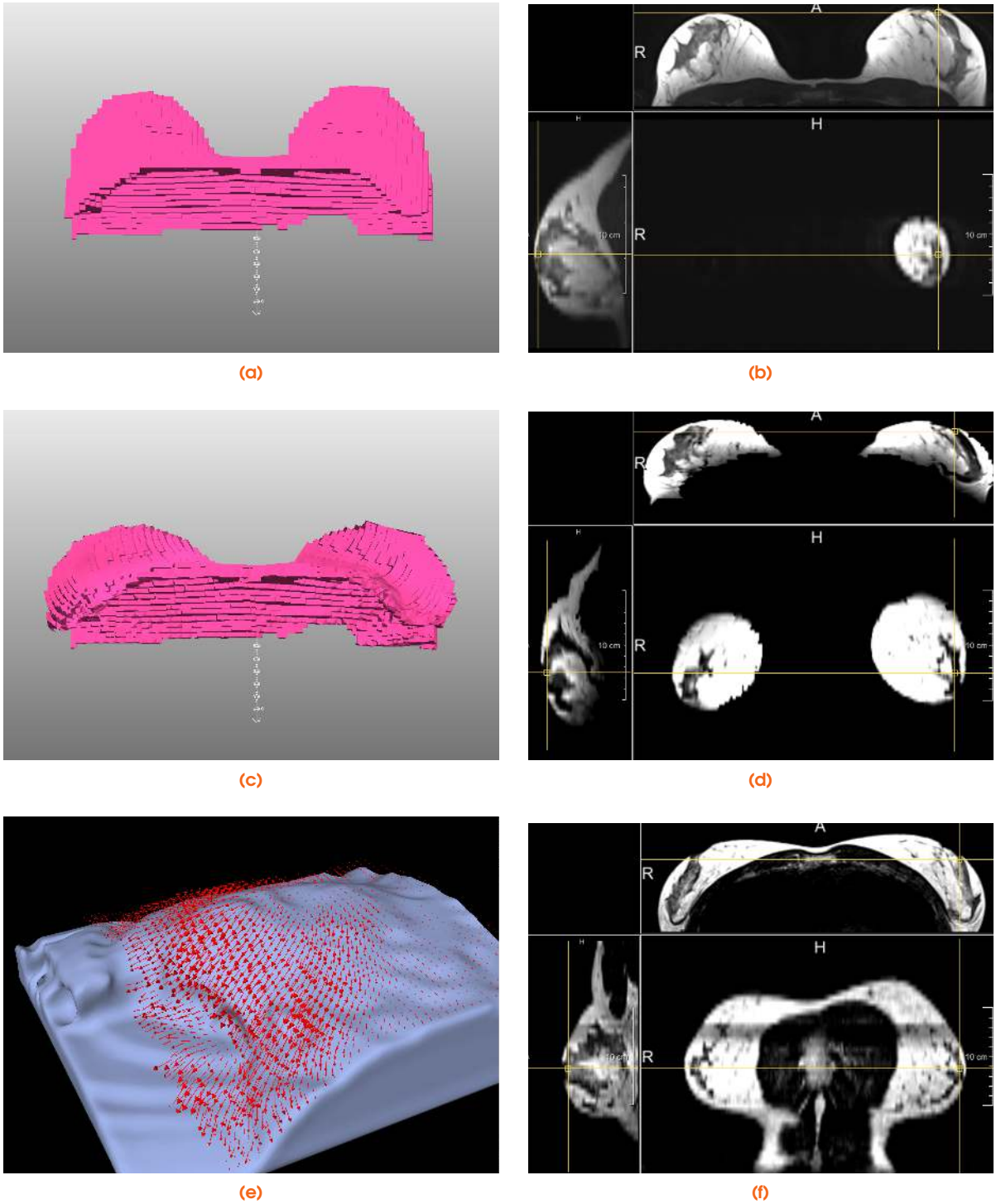
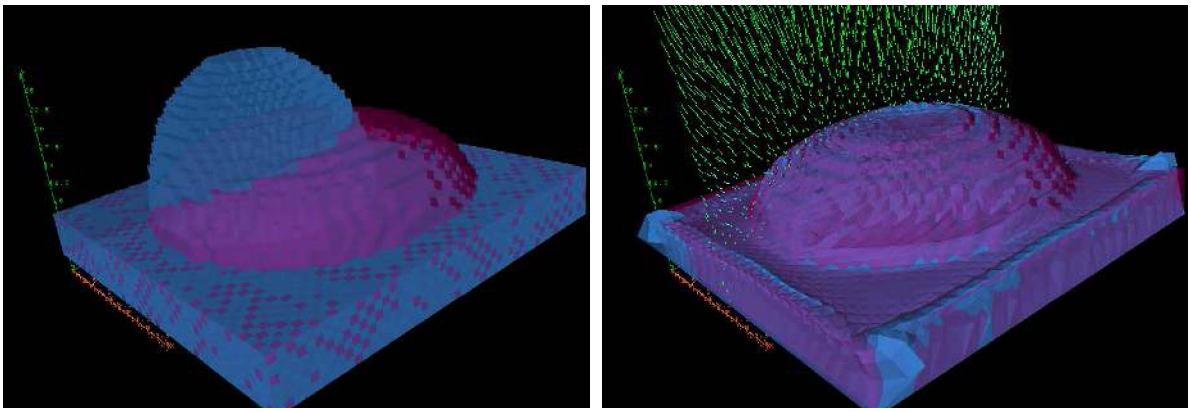


Figure 4.11: Sliding simulation on volunteer data. (a), (b) Surface rendering of the mesh before and after deformation simulation. (c) Vectors indicate the sliding on the approximated chest wall surface (blue). (d), (e) Sections through the prone and simulated supine MRI; (f) Section through a supine MRI scan of the same volunteer for comparison.

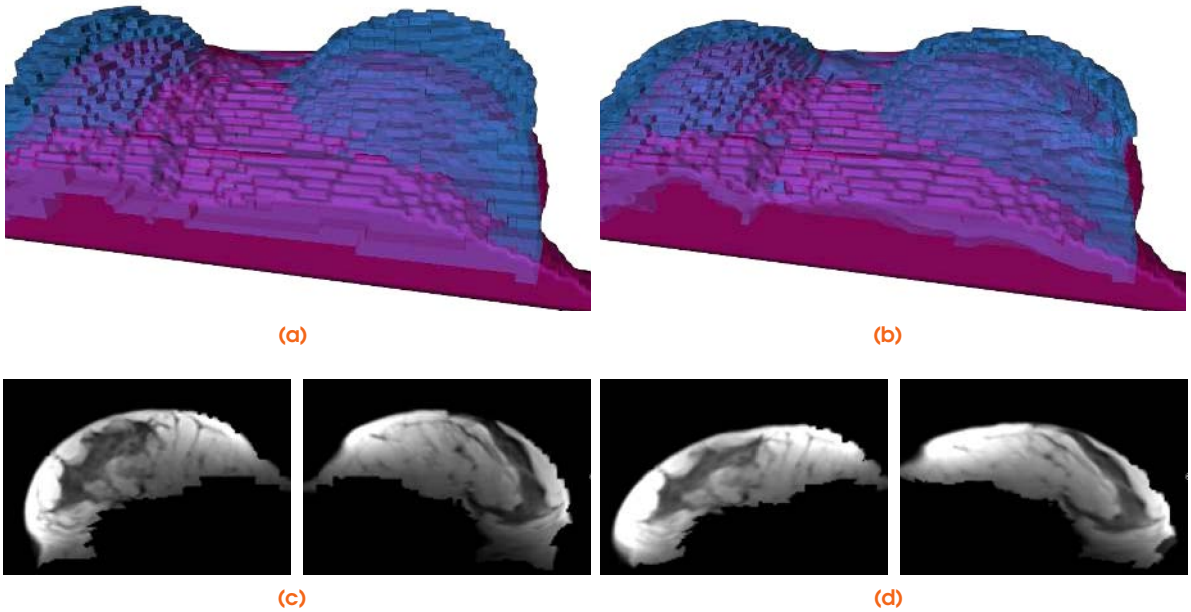


(a) (b)

Figure 4.12: Morphing a volume mesh to a target surface. (a) The model shapes before deformation simulation. (b) Final deformation using forces attached to surface. Forces are proportional to the green arrows (larger with increasing distance from target surface). Only the lower edge of the plate is fixed, while the plate itself models deformable adipose tissue.

tissue parameters in the simulation may need to be set to a much lower Poisson ratio to allow for controlled compressibility.

A current limitation of the approach is that the chest wall model obtained from the prone data is bounded by the arms. This causes the breast model during simulation to not slide as much as in reality, and, as a consequence, the surface morphing needs to account for a larger distance between the surfaces. We intend to solve this issue by improving the modeling of the chest wall so that a more realistic sliding around the chest is obtained.



(a) (b)

(c) (d)

Figure 4.13: Morphing the volunteer volume mesh generated from the sliding algorithm (cf. Fig. 4.11) to the target surface generated from the supine MRI data. (a) The model shapes obtained from the sliding simulation; (b) after seven iterations of surface morphing. (c), (d): Orthogonal reformatting of the states in (a) and (b), showing the same structures as in Fig. 4.11 for the right/left breast.

Combining the Approaches. All algorithms described before are modular additions to the basic simulation code. They can be parameterized and arbitrarily enabled and disabled. We propose a combined pipeline according to Fig. 4.14. The processing iteratively performs the following

steps:

1. Add vertex forces at current displaced positions (Surface morphing).
2. Update element elasticity based on von Mises stress and α.
3. Simulate: calculate new displacement field.
4. Check for inverted elements
   (a) If none, decrease α; back to 1.
   (b) If inverted elements, increase α and revert positions; back to 2.

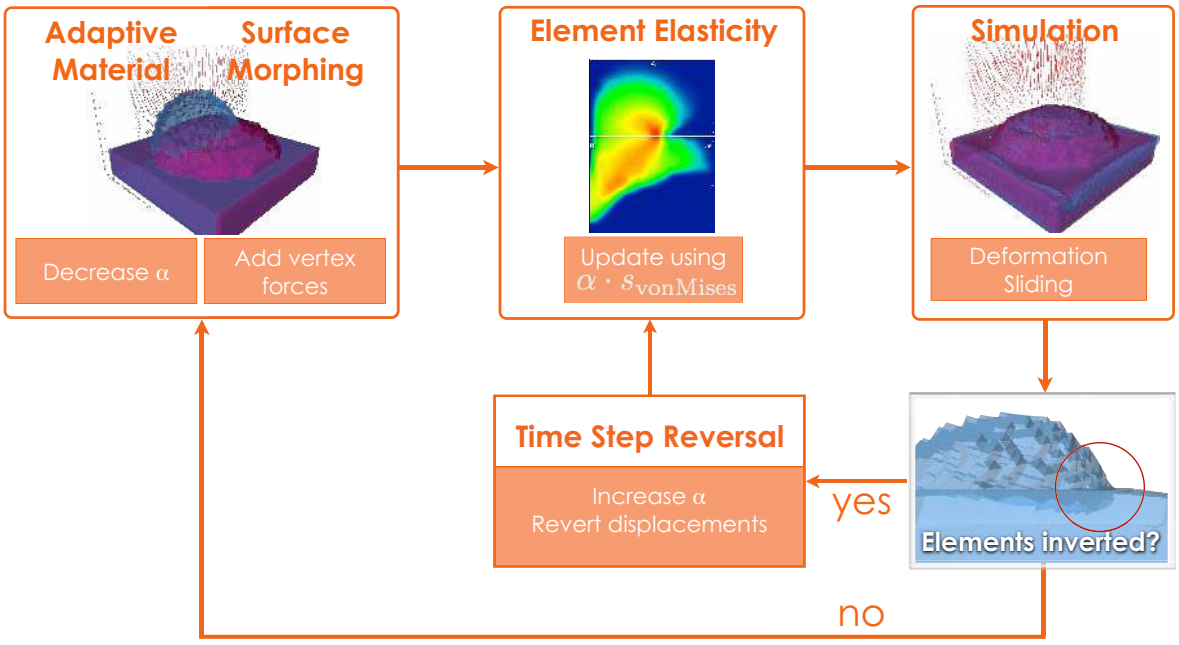


Figure 4.14: Outline of the simulation steps arranged in a feedback loop. The loop allows for undo operations of those simulation steps that result in inverted elements.

Note that the vertex forces do not need to be reset explicitly: In case inverted elements are detected in step 4., no forces at the current positions are added, because the algorithm iterates from 2. The vertex positions, however, are restored to the state before the current. Consequently, the vertices are again in the positions where they received the last vertex force update (which has not been reset), but elements will be stiffer now due to the increased α. The next iteration starts at step 2., with forces like in the previous iteration.

We have so far successfully applied the proposed pipeline of algorithms to the breast model depicted in Fig. 4.12 and can state that generally the combination of the algorithms increases the robustness and quality of the overall approach.

We have demonstrated a set of efficient algorithms that each individually improve simulation of breast deformations, and we have shown examples demonstrating the effectiveness of the approaches. By means of the automatic stiffness update in combination with the timestep reversal, we are able to mimic the behavior of breast tissue on a coarse scale, and thus we are able to achieve fast update rates of that approach.

The simulation of sliding boundary conditions on the chest wall is especially designed for the prone-supine deformation problem, where the breast tissue moves significantly along the chest wall; the benefit of our approach is that the sliding is determined automatically by inspecting

the acting forces at the Dirichlet boundary conditions. Moreover, we have shown that a surface morphing approach can be successfully employed to deform one surface into the shape of a given second surface, even when they are not very similar.

We have proposed a pipeline joining all individual developments into a robust setup and demonstrated as a proof-of-concept the application of individual and combined algorithms to volunteer data, and of the pipeline to a breast phantom. Several improvements have yet to be integrated, though, including the interoperability of the different modular framework extensions to increase the user-friendliness and robustness of the combined computation pipeline. Ways to achieve this will be discussed below.

Perspectives

Out of the many alternatives that exist to simulate soft tissues, we have chosen a combination of corotated Cauchy strain in an efficient multigrid framework, aiming at a physically realistic but still efficient simulation that is capable of interactive update rates. There are methods that may yield more accurate simulation results in particular for the prone-to-supine deformation simulation we wish to discuss.

Using specialized shape functions V Rajagopal et al. (2004) decided to use a mesh of hexahedral elements with tri-cubic Hermite basis functions instead of a trilinear interpolation, which may be justified because fewer elements may be able to capture the curved geometry using this shape function. In these meshes, a neo-Hookian (non-linear) material model is used. The computation time is not documented, but sine each node in the tri-cubic Hermite model has eight degrees of freedom, opposed to only one in trilinear interpolation, it is expected to exceed computation time from this perspective; in addition, and more importantly, no use can be made of the pre-calculated stiffness matrix which is only possible in linear material laws. The employed implementation used by V Rajagopal et al. is freely available at `http://www.cmiss.org`.

Using GPU implementations Explicit finite element solvers (those which calculate solutions per vertex, and not simultaneously) lend themselves to parallelization, for example on the multiple cores of any typical programmable GPU like in (Han, J. H. Hipwell, Z. Taylor, et al. 2010). The speed of computation then scales with the number of GPU cores. Any linear or non-linear material law can be simulated without prohibitive additional computational effort. Still, since for explicit solvers, the maximal time step is limited, there is always a considerable surplus of computation required. Also, the speedup by using the GPU instead of a multi-core CPU is not always as big; for example D. J. Morris et al. (2012) reports only a factor of two in their implementation of an explicit FEM solver using linear stress-strain relationships. The most severe drawback, however, is the inflexibility of GPU programming from a software engineering standpoint. While there is sophisticated support available to program parallel GPU code within standard C++ code, it is not as easily possible to design a modular extensible framework like the one we have used above.

Summarizing, the major advantages of our approach over the alternatives, is its high performance combined with a framework approach. This together makes it possible to support interactive scenarios while offering approximated non-linear material properties. Most importantly, however, no domain-specific assumptions are made. Instead, a general solution applicable beyond breast deformation simulation has been developed. We will discuss a further application of our framework to a different task in breast cancer care below (Sec. 4.3.2).

Our approach enables a number of future research directions which have the potential to further improve the specific clinical procedures. One idea is to extend the surface morphing approach by not only dragging the model to the target surface, but to adjust other boundary conditions (e.g. the gravity direction) such that a situation with as little as possible local stress

is obtained. Another way to achieve a more stable simulation is to replace direct displacement constraints by forces acting in the direction of the displacement, with magnitudes proportional to the desired displacement. In this fashion, the deformation simulation acts as a regularizer for the force field and is more likely to yield results with fewer element inversions that have to be suppressed by time step reversal.

The extension modules concerning the time-critical updates of boundary conditions are currently not implemented in the most efficient way. They are generally eligible for parallelization, which is most evident for the calculation of stiffness updates when only element-local parameters need to be known or calculated. This is also where the greatest gains are expected.

It may also be considered to relax the mesh estimated from the prone DCE-MRI into a gravity-free state before simulation of gravity. The breast shape in the breast coil is loaded by gravity and by unknown compressive forces by the fixation paddles in the coil. In our work so far, we use a simple approach to determine the gravity-free state, which does not account for non-linearities in the material behavior. We plan to refine the process of generating a gravity-free state similar to previous approaches, where the reference state was iteratively approximated from a loaded configuration (Vijayaraghavan Rajagopal et al. 2007). In addition, for a realistic relaxation we want to estimate the unknown compression introduced by the breast coil during relaxation. All this is at the expense of additional computational costs.

Some future effort has to be spent to automate the application and algorithms to process volunteer data fully automatically and robustly, since breast shapes and sizes, and their deformation behavior due to tissue composition vary significantly. We aim at testing the algorithms on a large number of data sets and derive robust parameters which might depend on patient-specific values such as breast volume and breast density, potentially also the structure of the parenchyma. Additionally, in our framework it is possible to compare the approximated non-linear material law with implementations using true anisotropic non-linear material laws.

Our methods are not specific to the task of surgery preparation or deformation tracking during surgery. It can equally well be employed to simulate the breast compression in other imaging procedures, like mammography or tomosynthesis (compression with one movable plate against one fixed plate; friction of breast surface on plates), or whole breast ultrasound (compression with one plate against the chest, no or minimal friction of breast surface on plate), or the matching of a pre-biopsy diagnostic MRI to the peri-interventional biopsy MRI where often the lesion position of interest is hard to recover. This application will be discussed more closely in the following section.

4.3.2 Biopsy Workflow Support Using Deformation Modelling

Targeting lesions for biopsy in MRI-guided breast biopsy procedures relies on the administration of contrast agent and a dynamic contrast enhanced MRI scan. The index lesion will then be punctured with the biopsy needle, and a control scan reveals correct needle placement with respect to the targeted position. In the work we describe here, we propose to simulate the breast deformation from the diagnostic MRI scan to the MR biopsy preparation MRI taken before contrast agent administration. We describe the technology and setup, and provide a proof-of-concept on retrospective example data where the lesion was visible in the biopsy scan. The proposed method may allow to biopsy some of the 10% of lesions that would no longer enhance in the biopsy device. In addition, the procedure is faster, and by not requiring contrast agent, cheaper. We present a concept study with qualitative and quantitative evaluation on data from a small-size pilot study.

Breast biopsies are the gold standard decision making tool to determine the diagnosis and suggest the treatment of a suspicious lesion. In principle, the biopsy will be conducted using

the imaging modality that displays the lesion first with a sufficient reliability, and accurately enough to determine the lesion extent. Among the most prominent biopsy techniques in breast cancer care are the following (Ames et al. 2011; Bruening et al. 2010; O'Flynn et al. 2010):

Stereotactic biopsies. With two oblique mammographic acquisitions of the same breast under unchanged compression, a lesion position can be targeted with highest accuracy. This technique can be used both with core needle biopsy devices or vacuum-assisted needles, the latter being the method of choice. Mainly, small clusters of microcalcifications, indicative of early DCIS, are the target lesion class that is biopsied in this modality.

Ultrasound-guided biopsies. Whenever a lesion is visible sonographically, hand-held ultrasound provides a preferred tool to visualize lesion and needle. The real-time control over needle position together with the monitoring of lesion displacement induced by the advancing needle tip are main factors for extremely low rates of failed procedures.

MRI-guided biopsies. There are cases where neither mammography nor ultrasound suffice to demonstrate the lesion in question. These cases arise when a suspicious finding from mammography and ultrasound remains indeterminate, and MRI is used as a decision making tool. Often, additional lesions display in MRI, or the extent of the finding is demonstrably larger than it appeared. Since in these cases, no other modality might guide the biopsy procedure, MRI is used for targeting.

In this section, we focus on the MRI-guided breast biopsy procedure and suggest a workflow improvement enabled by deformation simulation that has the potential to save time and money, contribute to more patient comfort, and help to biopsy those 10% of lesions that currently cannot be biopsied because they don't show up in the biopsy preparation MRI.

MRI-Guided Vacuum-Assisted Breast Biopsy

The indication for a MRI-guided breast biopsy is the same as for any other image-guided biopsy procedure: image-guided interventions for histopathological diagnosis and therapy planning in BIRADS-4/5 imaging findings should be performed guided by the imaging method the finding is best visualized with. In the case of MRI-guided breast biopsy, vacuum-assisted core needle biopsy (CNB) is internationally recommended because of its superior sensitivity (American Cancer Society 2013c; Leitlinienprogramm Onkologie 2012).

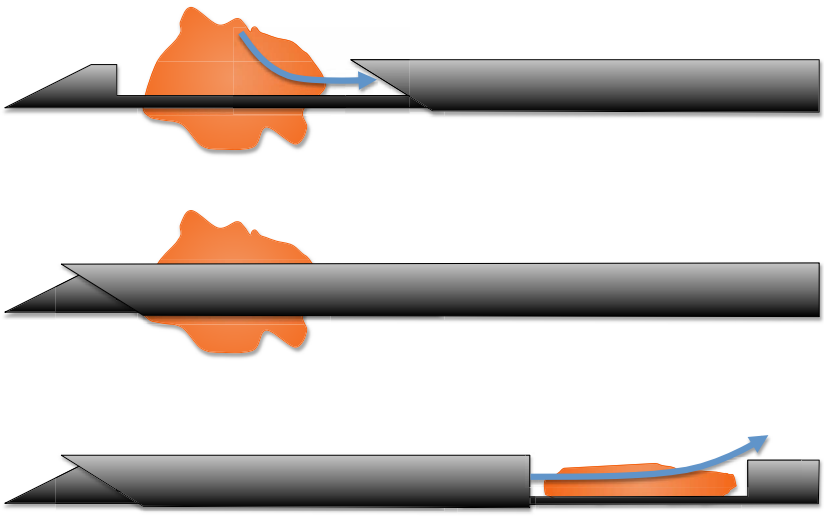


Figure 4.15: Illustration of vacuum-assisted core needle biopsy. *From top:* (1) The hollow needle has an opening into which tissue is sucked by a vacuum. (2) The outer part of the needle is advanced to cut the cylindrical tissue sample. (3) The vacuum transports the tissue sample along the needle into a container or an opening where it can be collected.

For the MRI-guided biopsy, the woman is placed in prone positioning into the breast coil, which is equipped with a biopsy grid on the lateral side. The breast is slightly compressed to avoid as much target movement as possible during the 30 min procedure.

The procedure starts with a fast DCE-MRI scan to depict the lesion. For this scan, and throughout the procedure, the breast of concern is slightly compressed within the biopsy coil. The compression is predominantly effected by the biopsy grid, which is in most biopsy-capable breast MRI coils inserted into the lateral casing. Biopsy coils also usually offer fiducials that aid the computerized targeting, for example small openings into which contrast agent-filled containers have to be inserted. These fiducials are automatically detected by the biopsy targeting software.

Using the MRI biopsy planning software, the operator places a target marker in the three orthogonal projections of the anatomical breast image of the preparation scan. From fiducials and manual mark, the targeting parameters can be determined. The radiologist reads off the grid position and the required needle insertion depth from the computer monitor and transfers them to the biopsy grid. One possible device arrangement, which is being used in the images in Fig. 4.17 is a grid into which a small cube can be inserted. The cube has a number of parallel holes so that for each grid position of the cube, multiple needle insertion positions exist (see Fig. 4.16). Alternate configurations exist for medial access, and even devices for cranial or caudal access have been designed. For complicated lesion positions close to risk structures, needles with round tips exists.

(a) (b)

Figure 4.16: Two systems for MRI-guided biopsy. (a) Grid system, where targeting is restrained to the grid discretization. (b) Post-and-pillar system that allows arbitrary positions to be targeted. Images were kindly provided by Royal Philips, Amsterdam, The Netherlands.

When the targeting has been completed and the insertion of the biopsy needle prepared, local anesthesia is administered to the woman, and a small incision of about 5m̃m allows the biopsy needle to be forwarded into the breast and guided to the suspect area. To confirm placement adjacent to the lesion, a second scan is performed. Some systems allow to insert a contrast agent-filled tube that is easily visualized in the MRI scan. When the correct needle placement has been confirmed, the biopsy needle is advanced through the outer sheath, and the biopsy begins.

While the needle cavity is open (compare Fig. 4.15, top), a vacuum sucks the tissue into it. The outer sheath rotates and advances, cutting a sample of the tissue, which is sucked through the hollow needle into a deposit where it can be extracted by the radiologist. Step by step, the cavity of the needle is rotated to sample tissue around the position, while the individual tissue

samples (the "cores") are taken. Twelve or more such cores are often sampled, summing up to about 1 cm^3 of tissue.

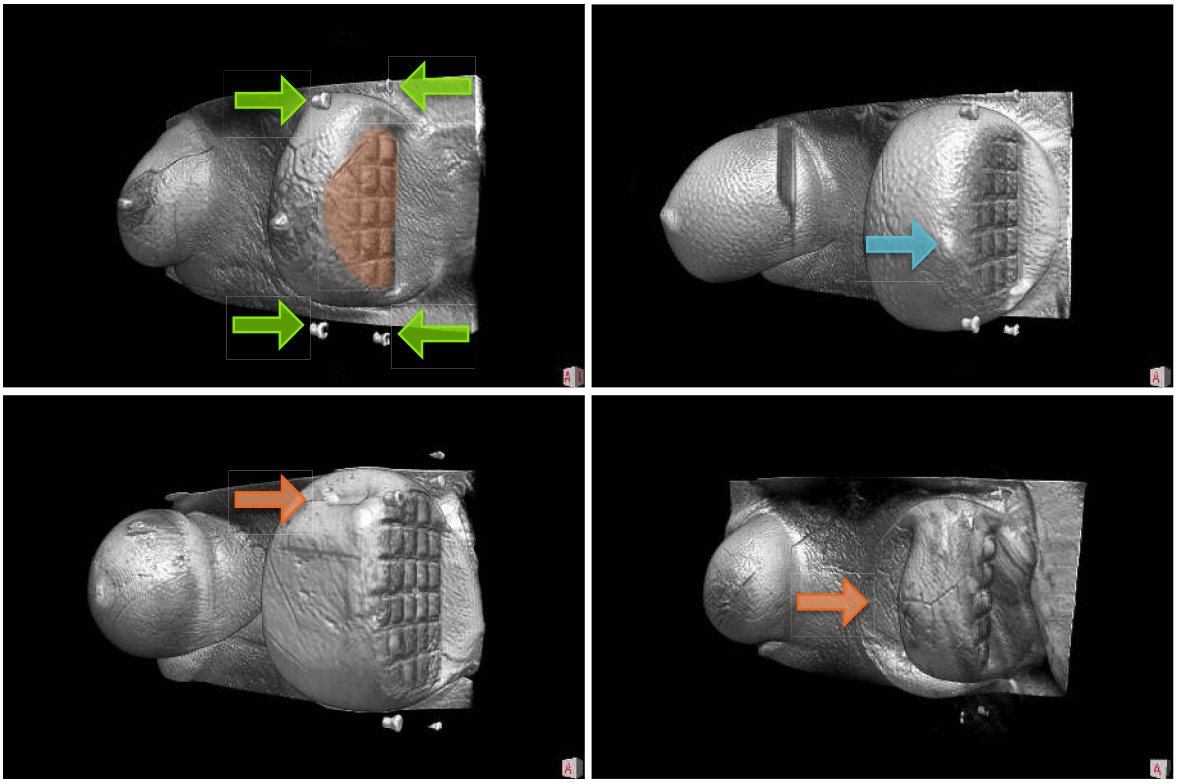


Figure 4.17: Examples of peri-interventional scans shown in illuminated volume renderings. In the top left example, the biopsy grid impressed into the breast tissue is indicated with a color shading. In addition, arrows mark the following: *Green arrows:* Artificial landmarks helping to set up the computational model of the biopsy grid. *Blue arrow:* Shifted nipple position. *Orange arrows:* Strong local or regional deformations.

After the biopsy, the concordant cavity position is easily confirmed using another MRI scan. The taken specimen are subsequently examined in the pathology department. Histopathology and imaging results of all employed modalities need to be correlated later to confirm the diagnosis and establish the definitive treatment of the suspect area.

Robust Biopsy Site Targeting: Contributions

The procedure leads to reportedly about 10% of targeted lesions that are no longer visible on pre-interventional MRI (Brennan et al. 2011; Viehweg et al. 2006). Although some reports state that those missed lesions are frequently benign and hormonal tissue alterations, they inevitably lead to a further short-term follow-up MR imaging, also inducing anxiety in the women. Also, particularly small lesions (<1 cm in diameter) are among those not recovered in pre-interventional MRI. On the other hand, such small cancers can be cured with almost certainty, so that a missed biopsy in these cases is particularly undesirable.

One way to improve lesion depiction particularly during subsequent confirmation scans is to administer the contrast agent not in one bolus which will be washed out long before the procedure came to its conclusion, but to give smaller amounts of contrast agent at several times during the procedure. This leads to a slightly increased amount of contrast agent necessary, but helps to visualize the lesion throughout the procedure if done properly (Stoeblen 2014).

Instead of an improvement like the one described, we chose to explore the feasibility of a changed workflow that may help to correlate the lesion not only between successive confirmation scans during the biopsy procedure, but also to the diagnostic MRI taken before. It also opens a way to speed up the procedure by substituting the contrast-enhanced series with computational

derivations of the target position altogether, although this is not the primary goal. We propose to utilize deformation modeling as described above to simulate the effect of the compression induced by the breast coil on the breast tissue. The outline of this approach is the following:

Prior Preparation Segment the breast tissue and the target lesion from the contrast enhanced diagnostic MRI scan and generate the patient specific breast model breast tissue mask.

Immediate Preparation Segment the breast tissue from a contrast-free fast acquisition in the biopsy MRI using the biopsy coil. The compression by the biopsy device should be applied in these images.

Matching Two alternative approaches are explored to match the diagnostic scan to the biopsy scan:

Simulation. Apply boundary conditions on the breast. Large displacement occurs from medial side towards lateral by the compression plate, and a light fixating displacement is induced by the biopsy grid from the lateral side towards medial. Afterwards, surface morphing accounts for the difference in surface shapes.

Registration. Automatic initial alignment using an affine registration algorithm; result refined by two steps of non-linear elastic registration.

Biopsy Transfer target lesion position to the deformed state and visualize for biopsy planning.

If successful, this novel procedure not only helps to biopsy even those lesions that cannot be visualized any more in the biopsy coil, but it potentially also saves time through sparing the contrast-enhanced acquisition. It may hence additionally save costs by substituting the contrast agent administration by a computational method, and lastly, it contributes to patient comfort since no injection is required in the biopsy situation. Even if these goals are today not reached, the increase in reliable target identification is apt to increase the radiologists confidence performing the procedure.

From Model to Prediction: Methods

The aim is to approximate the deformation of the breast from MRI in the prone position using breast coils to the compressed prone position used in MR-guided biopsy. When this is achieved, the current lesion position can be predicted from the diagnostic scan and the deformation model.

Data and Model Generation. Breast MR images of a total of 30 patients have been used in this study. The patients together underwent 31 MR-guided breast biopsy procedures. Out of this collection, we have used six biopsy cases for the initial method developments and the remainder for refinements and testing.

The breast MRI data used in this work consists of diagnostic and interventional data. The diagnostic data is acquired with optimized spatial resolution and with a longer coverage of contrast agent wash-out by several acquisitions. Our data, contributed by two sites, differed in this regard: Six patient data sets used for development were acquired in a standard five-time point DCE-MRI examination, with one baseline acquisition and four post-contrast time points in 60–90 sec delay. The testing data from another site has been acquired in a high temporal resolution protocol, where before and during the wash-in phase 20 time points are acquired with an approximately 4.5 sec between any two.

The interventional scans are likewise site dependent. The development data came from a site where unilateral scans in sagittal direction were used, and only one pre-contrast and one post-contrast time point is measured. The testing data from the second clinical site encompassed bilateral scans with a smaller field of view (FOV) in the anterior-posterior direction. Five time points are measured this time, one before and four after contrast agent injection.

We excluded two patients from the deformation simulation based approach since they were not correctly processed by the fully automated preprocessing and model generation pipeline. One patient had large implants. The automated segmentation assigned the implant to the class of rigid tissue types—in fact, though, implants are designed to deform similar to soft breast tissue. A second patient with very small breasts has been excluded because the automated setup of simulation boundary conditions fails to determine the fixation plates of the biopsy device. It is in our approach one important goal to build an automated processing mechanism and to assess its performance, which is why we did not manually correct the segmentation of the implant case, nor manually set the fixation plates of the second case. Of course both cases could be processed with manual interaction, but it is left for future work to design suitable extensions of the automated processing.

All data sets were segmented into rigid and deformable tissue. For reasons explained above, we do not distinguish between breast parenchyma and the adipose tissue, but consider them equally elastic, and only the thorax and adjacent pectoral muscle are considered rigid and fixed in position (Dirichlet boundary condition). Our proposed system starts with MR images taken in the prone position that are automatically segmented into deformable and fixed tissues using the methods proposed by L. Wang, Filippatos, et al. (2011). A volumetric tetrahedral mesh was generated from a downsampled version of this data, tuned to yield meshes consisting of at most 50,000 elements.

To model the breast compression that is induced by the biopsy device, we now look at the breast tissue outer contour, with the boundary to the chest wall excluded. To avoid later possible misses in the detection of the compressing devices (the biopsy grid on the lateral side and the breast coil padding on the medial side), we crop the contour image by looking at the voxel counts in sagittal slices. Thinking of the voxel count plotted against the sagittal slice number, the resulting curve is supposed to show two peaks where the breast contour is roughly parallel to the sagittal plane. These peaks are detected in a maximum search and used as a regional constraint for the search of the two planes. From this reduced mask image, we generate a point set that we thin out by iteratively removing every other point several times, until a suitable number of points remains.

Deformation Simulation. These points are submitted to an implementation of the RANSAC algorithm (Fischler et al. 1981) to find two parallel planes in the data that are roughly parallel to the sagittal viewing direction and have distance above a user-defined threshold (e.g. 6 cm). These planes are assumed to be infinite in the cranio-caudal direction, but limited posteriorly based on the detected size of the found plane.

For the intended clinical setting, a compromise has to be made between fast calculations and highly realistic material behavior. Fast computations are mandatory in this scenario since in the intervention situation, the image of the compressed breast is acquired for the first time, such that while the bulk work, which includes setting up the finite element model for the original breast MRI volume, may be computed in advance, the actual deformation simulation from the prone fixated diagnostic contrast-enhanced breast MRI to the interventional situation will be computed with the patient in the scanner. It hence has to be fast so that the benefits of the approach can be harvested. Hence, the above efficient FEM-based breast deformation framework is used to simulate the shape change from MRI to the positioning and compression in the current biopsy procedure.

An implementation of a compression simulation including sliding on the plane surface governed by friction has been adapted to consider planes with finite extents. In this implementation, a collision detection probes if surface vertices of the breast volume mesh are, after one movement step of the compressing plane, behind the plane. If so, the vertex is projected onto the plane by applying a displacement constraint. After that, forces are calculated for all vertices on the plane

to estimate their friction-governed sliding, which is again submitted to the vertex in the form of a displacement constraint (Dirichlet boundary condition).

To align the surface resulting from the compression simulation with the known surface of the compressed breast in the biopsy device, we utilize the surface morphing approach outlined. The resulting vertex displacements, describing a dense deformation field, are applied to the diagnostic MRI image for evaluation processes.

Registration-based approach

The registration based approach combines linear and non-linear image registration techniques. The linear registration step serves as an initialization step with the aim to diminish the distance between the images enough for the non-linear registration to succeed. All registration techniques we used are implemented in MERIT (Medical Image Registration Toolkit) (Boehler 2011).

Preprocessing. The tissue masks obtained for the FEM approach are utilized to restrict the calculations of the image registration techniques to the breast tissues. This intentionally neglects deformations in adjacent tissues to increase the flexibility of the algorithms within the region of interest. To reduce influences from soft tissues other than the target breast further, the breast mask of the biopsy image is cropped to the breast. In cases where the contralateral breast is partially included in the scan or wraps around in the phase encoding direction (a MRI imaging artifact common in unilateral scans), this has proven beneficial. On the other hand, the selection of the processing region has to ensure a region large enough to provide sufficient information to the registration algorithms. To this end, the two scans are initially chosen so that their intersection encompasses the target breast fully in both. By subsampling the input data, the influence of noise is reduced, while simultaneously the maximum amount of permissible deformation is increased by the larger voxel size.

Linear Registration. After rigidly shifting the image centers of gravity to align, the linear registration optimizes translation, scaling, and shearing for the three orthogonal axes simultaneously by maximizing the normalized cross-correlation (NCC) between the reference and template image. Here, reference image is the interventional scan, deformed template is the diagnostic scan. Rotation is intentionally prohibited since during plate compression the assumption is that the breast will not be rotated. In fact, this assumption may be violated in some cases; see the discussion for a reasons of this. The parameters of the affine transformation $\mathbf{T}$ are hence

$$\mathbf{T} = \begin{pmatrix} s_x & h_x & h_x & t_x \\ h_y & s_y & h_y & t_y \\ h_z & h_z & s_z & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{4.32}$$

where s is the scaling, h defines the shearing, and t the translational part of the transformation.

Note that for the registration, only the native, unenhanced images of both the diagnostic and the interventional scan are used; in particular, no knowledge about the location of the index lesion is assumed.

The choice of NCC as the distance metric is motivated by the fact that intensity values and image contrasts are often different between diagnostic and interventional scans. NCC is known to be invariant against linear intensity changes, and only mildly affected by global differences. Further, $\mathbf{T}$ is optimized with a multi-level Newton method on two levels. Transforming the template image is then a simple multiplication of the world matrix of the template image with $\mathbf{T}$:

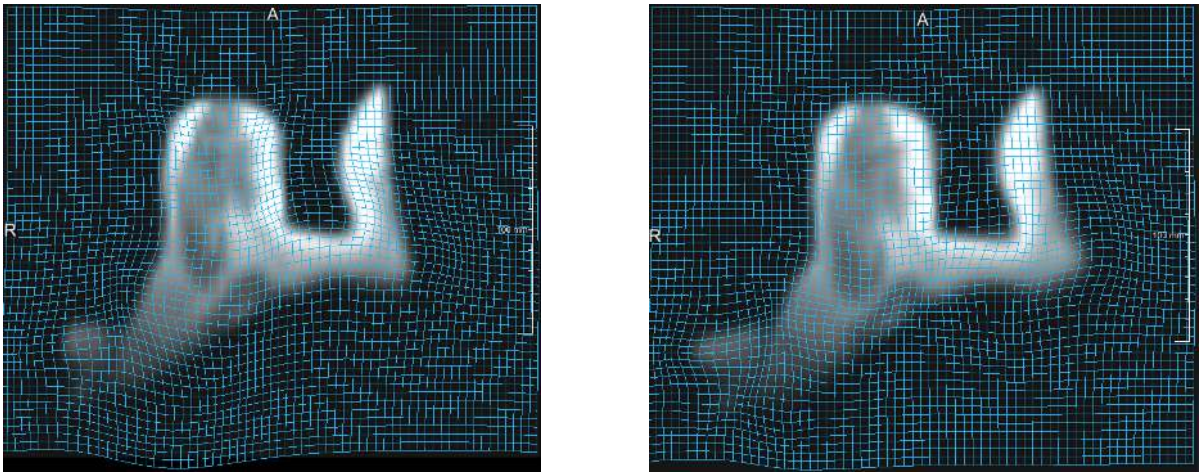
$$\mathbf{I}_T^{linear} = \mathbf{T}_{linear}(\mathbf{I}_T, \mathbf{q})$$

Non-linear Registration. Two identically parameterized runs of a non-linear image registration method follow the linear initialization. We have found one non-linear registration to be incapable of solving the registration problem under boundary conditions that prevent unstable results. Continuing with the interventional scan as the reference, the transformed diagnostic scan serves as the template image. The images are kept at the reduced resolution for the same reasons as before.

The space of permissible solutions is in the chosen diffusion based algorithm governed by the distance metric and the regularizer. To allow deformations orthogonal to the edges in the image, the normalized gradient field servers as the distance metric. The resulting deformation field is regularized with a Gaussian filter with 11 voxels diameter and $\sigma = 1.1$. This ensures deformation fields without foldings. Neumann boundary conditions allow deformations of the image borders, additionally contributing to the overall flexibility.

The deformation fields of the two non-linear deformations, $\mathbf{u}_1$ and $\mathbf{u}_2$ (compare for an example Fig. 4.18), are added to yield $\mathbf{u}(x) = x + \mathbf{u}_2(x + \mathbf{u}_1(x))$, and appended to the linear registration result, $\mathbf{T}$. For the template image, it then holds that

$$\mathbf{I}_T^{non-linear} = T_{non-linear}(\mathbf{I}_T^{linear}, \mathbf{u}).$$



(a) (b)

Figure 4.18: Deformation fields for the two non-linear registrations. Note the strength of deformation in the second iteration.

Evaluation

Landmark position differences are evaluated quantitatively. The landmarks are annotated by an experienced user. Landmarks have been set in the nipple position and at landmarks within the index lesion if it was uniquely defined in biopsy and diagnostic scan.

For the evaluation, the error between the estimated landmark position and the true landmark position can be assessed bidirectionally: from the given position in the diagnostic image, or from a given position in the interventional image. The average error for one landmark (pair) is then calculated from the errors in both directions.

For bidirectional landmark prediction, the non-linear part of the mapping needs to be inverted. There is no unique solution for non-linear deformation fields, hence the inversion operation introduces an error in landmark registration accuracy that has been determined to average below typical voxel sizes. The inversion algorithm employed in this work has been proposed by M. Chen et al. (2008) and utilizes a iterative method rooted in fixed-point theory that

converges to a pseudoinverse of the forward deformation field even if it has few local topology changes (wrapping artifacts).

Care has to be taken when setting up the inverse transformation. While the affine transformations can be inverted most easily (they apply to any location equally and have a uniquely defined inverse), the non-linear deformation or registration part is only defined for the image voxel coordinate ranges, and its inverse may be ambiguous. Additionally, the affine transformations are applied to the template image. They are hence a mapping from template to reference. Deformation fields, on the other hand, are usually given as target-to-source mappings, translating from the reference to the template image. When finally concatenating the transformations, it is important not only to invert the correct transformation, but also the order of their concatenation.

For a landmark prediction from template to reference, the affine transformation is applied, which shifts the reference frame into that of the reference image. Since the deformation field is given in voxels of the reference image, the world position needs to be converted to voxel coordinates, before the displacement given by the inverse deformation field can be added. Formally, this is expressed by

$$\widehat{\mathbf{x}}_\mathsf{T} = W_\mathsf{R}^{-1}\mathbf{T}W_\mathsf{T}\mathbf{x}_\mathsf{T} + \mathbf{u}^{-1}(W_\mathsf{R}^{-1}\mathbf{T}W_\mathsf{T}\mathbf{x}_\mathsf{T}), \tag{4.33}$$

with $\mathbf{T}$ indicating the affine transformation matrix, and $\mathbf{u}$ the displacement field. W are the voxel-to-world matrices of template and reference, and W^{-1} the world-to-voxel matrices. In the other direction, from diagnostic (template) to interventional (reference) image, the displacement field can directly be added to the voxel coordinates of the template image, before each world position is transformed with the inverse affine transformation. This is expressed in the following equation and illustrated by Fig. 4.19.

$$\widehat{\mathbf{x}}_\mathsf{R} = W_\mathsf{T}^{-1}\mathbf{T}^{-1}W_\mathsf{R}[\mathbf{x}_\mathsf{R} + \mathbf{u}(\mathbf{x}_\mathsf{R})] \tag{4.34}$$

A last subtlety is the potentially different voxel size in template and reference image, such that all measured landmark position differences $\mathtt{d}_\mathsf{R}, \mathtt{d}_\mathsf{T} \in \mathbb{R}$ need to be adjusted to the same reference frame. While measuring the distance in the reference image (interventional image) is simply

$$\mathtt{d}_\mathsf{R} = |\mathbf{x}_\mathsf{R} - \widehat{\mathbf{x}}_\mathsf{T}|,$$

for the distance $\mathtt{d}_\mathsf{T} \in \mathbb{R}$ in the template (diagnostic) image the coordinates have to be transformed into the reference image coordinate frame:

$$\mathtt{d}_\mathsf{T} = |\underline{\mathbf{x}}_\mathsf{R} - \underline{\widehat{\mathbf{x}}}_\mathsf{T}| = |W_\mathsf{R}^{-1}W_\mathsf{T}\mathbf{x}_\mathsf{T} - W_\mathsf{R}^{-1}W_\mathsf{T}\widehat{\mathbf{x}}_\mathsf{R}|$$

Robust and Accurate? Results

Between six and 14 landmarks have been identified in the 31 data sets. The nipple position has been marked in all image pairs, and the index lesion has been marked if it was present and uniquely defined in both images. Since some of the cases were aborted biopsies due to lesion non-visualization, and in others, either the lesion was not uniquely defined in the biopsy scan or had no sufficiently unique landmarks, the number of lesion markers is only 15.

For all cases, the simulation took about 2 min of preparation time (including automatic segmentation of the diagnostic MRI and setting up the volume mesh and matrices). These steps can be computed before the patient enters the MRI scanner for the biopsy procedure. Note that even the definition of the intended biopsy site could be done based on the diagnostic MRI already.

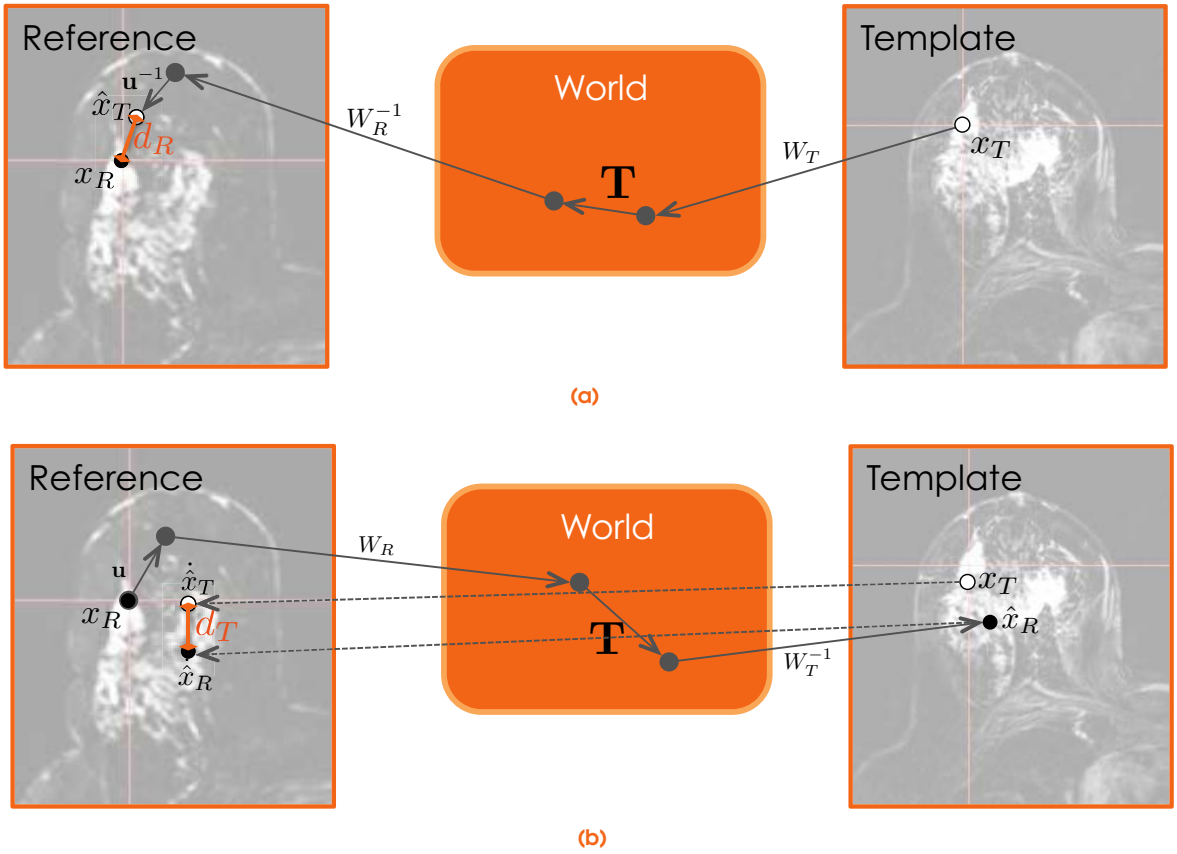


Figure 4.19: Bidirectional correspondence quantification for quantitative evaluation; (a) diagnostic→interventional; (b) interventional→diagnostic. T: template image space; R: reference image space; W: world coordinate frame. The red distances $\mathrm{d_R}$ and $\mathrm{d_T}$ are the landmark registration errors in reference frame and template frame, respectively. The dashed lines (lower figure) indicate the mapping back into the world reference frame to account for potentially different voxel sizes (see text).

When the preparation scan in the scanner has been acquired, the manual registration, and the simulation including finding the compression plates, simulating the moving plates, and morphing the surfaces, took about 1 min altogether. Fig. 4.21 gives a visual impression of the shape of the deformed diagnostic MRI with respect to the true interventional MRI (first row, left to right).

Deformation Simulation. Validation has been conducted for 30 out of the total of 31 data sets; one patient case was excluded since automatic segmentation failed for the interventional scan. The mean distance of landmarks over all landmarks of all data sets was 13.3 mm ± 5.4 mm, with a minimum of 6.5 mm and a maximum of 23.7 mm. More than half of the cases have landmark registration errors of 15 mm and below. Fig. 4.20a shows a plot of the per-patient landmark distances and their distributions. In the plot, the nipple position recovery error is indicated by the open circle; green triangle indicates the recovery error of the lesions that have been traced.

Note that out of nine outliers, six are the nipple landmarks, and that in further seven cases the extreme error reported corresponds to the nipple landmark. A reason for this and potential approaches to reduce the error will be discussed below. For the lesion landmarks, only one is an outlier, and two are extreme points of the reported distances.

Registration. The averaged per-patient distances range from 4.4 mm to 16.6 mm, excluding one data set with an average of 41.9 mm. This data set is considered an outlier, since the initial coarse rigid alignment step in the automated pipeline failed due to a missegmentation of the breast tissue. In the dataset, heavy wrap-over artefacts in the phase encoding direction (RL) may be the cause of this. Averaged over all data, including the outlier, the mean landmark registration error is 10.0 mm ± 5.6 mm. Excluding the outlier data set, this number drops to 9.0 mm ± 5.0 mm. About two third of the data set have average errors of below 10 mm between true and predicted landmark position; this holds for both directions. Details can be seen from Fig. 4.20b.

This time, out of four outliers, three are nipple landmarks. Still, nine nipple landmarks mark the extreme errors. No lesion landmark has been considered an outlier, and two are extreme errors of their data set. More than two third of the nipple markers are registered to within 15 mm distance, and eleven are within 10 mm error. Out of the 15 lesion landmarks, only four are recovered more than 10 mm from their marked position, while the remaining eleven are within a 10 mm range of their true location. More than one third of the 15 marked lesions are recovered with errors of less than 5 mm, and the maximum of lesion registration errors is 13.5 mm.

Quantifying the error of the inversion of the deformation field yields an average of 1.8 mm ± 1.6 mm (excluding the outlier data set reduces these number so 1.7 mm ± 1.4 mm). The differences between distance measurements in reference and template image are in the interval of 0.02 mm and 10.4 mm. The outlier has differences of up to 15.2 mm.

Perspectives

The clinical problem to correlate a biopsy target that was identified on diagnostic DCE-MRI with a location in the interventional biopsy planning DCE-MRI has been approached with two algorithmic setups. While the available data has been drawn from the clinically used contrast-enhanced series, both approaches have been implemented using the native, pre-contrast interventional scan alone to be in concordance with the long-term goal of potentially replacing the strict necessity of contrast agent during biopsy. The performance has been quantified, and sources of error been identified. Based on this, several improvements can be suggested.

Firstly, in some cases the segmentation method providing the basis to build the simulation model is not optimal for the data, in particular, implants are considered not to be breast tissue. While this is medically correct, in the context of the simulation they would need to be treated as

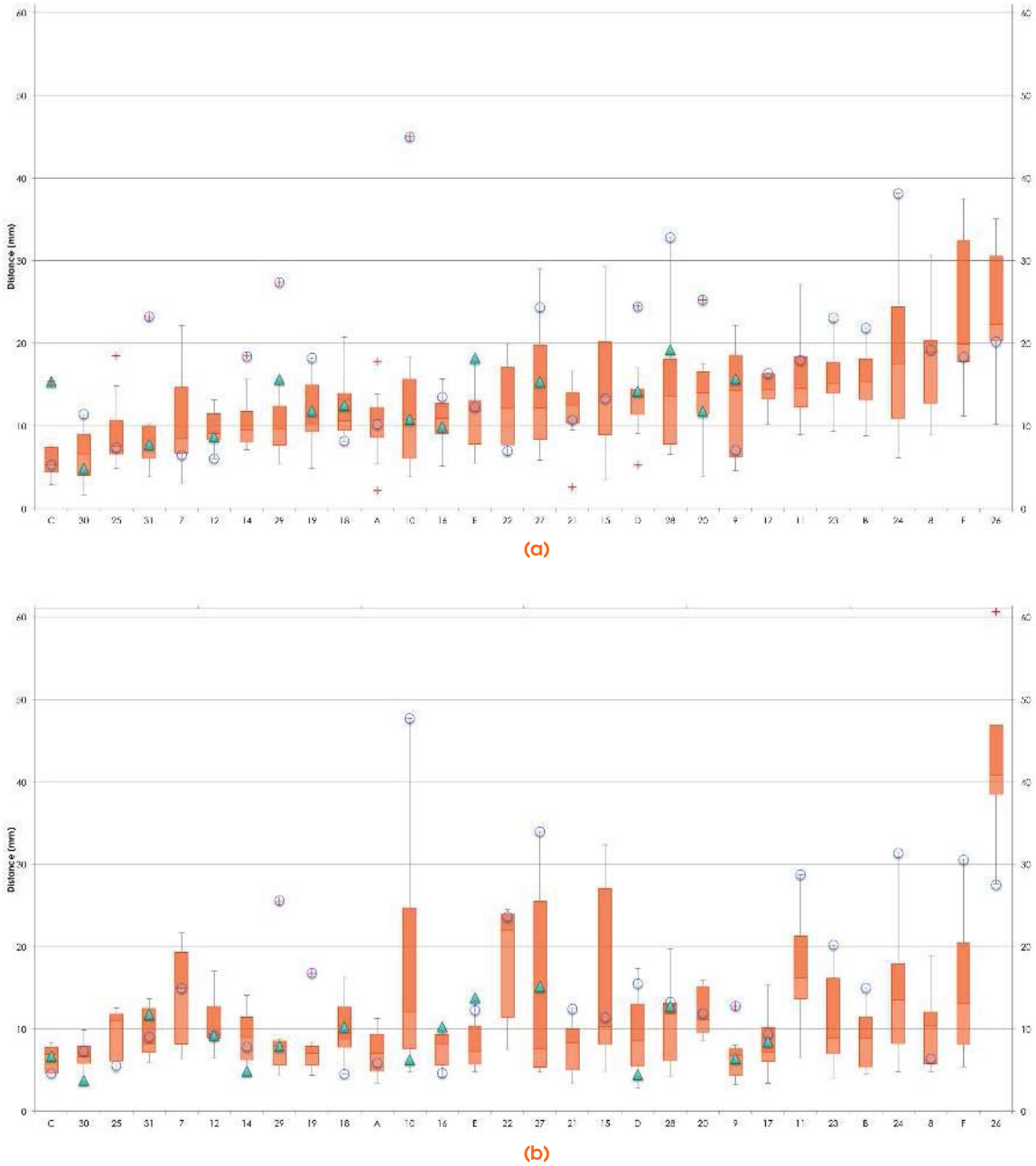


Figure 4.20: Boxplot of landmark registration errors for (a) deformation simulation and (b) registration approach. Cases are sorted by median error in deformation simulation based approach. Errors of prominent landmarks are visualized as follows: Circle: nipple distance. Green triangle: lesion. Red cross: outliers. Cases A-F are development data sets; numbered cases are from the test set.

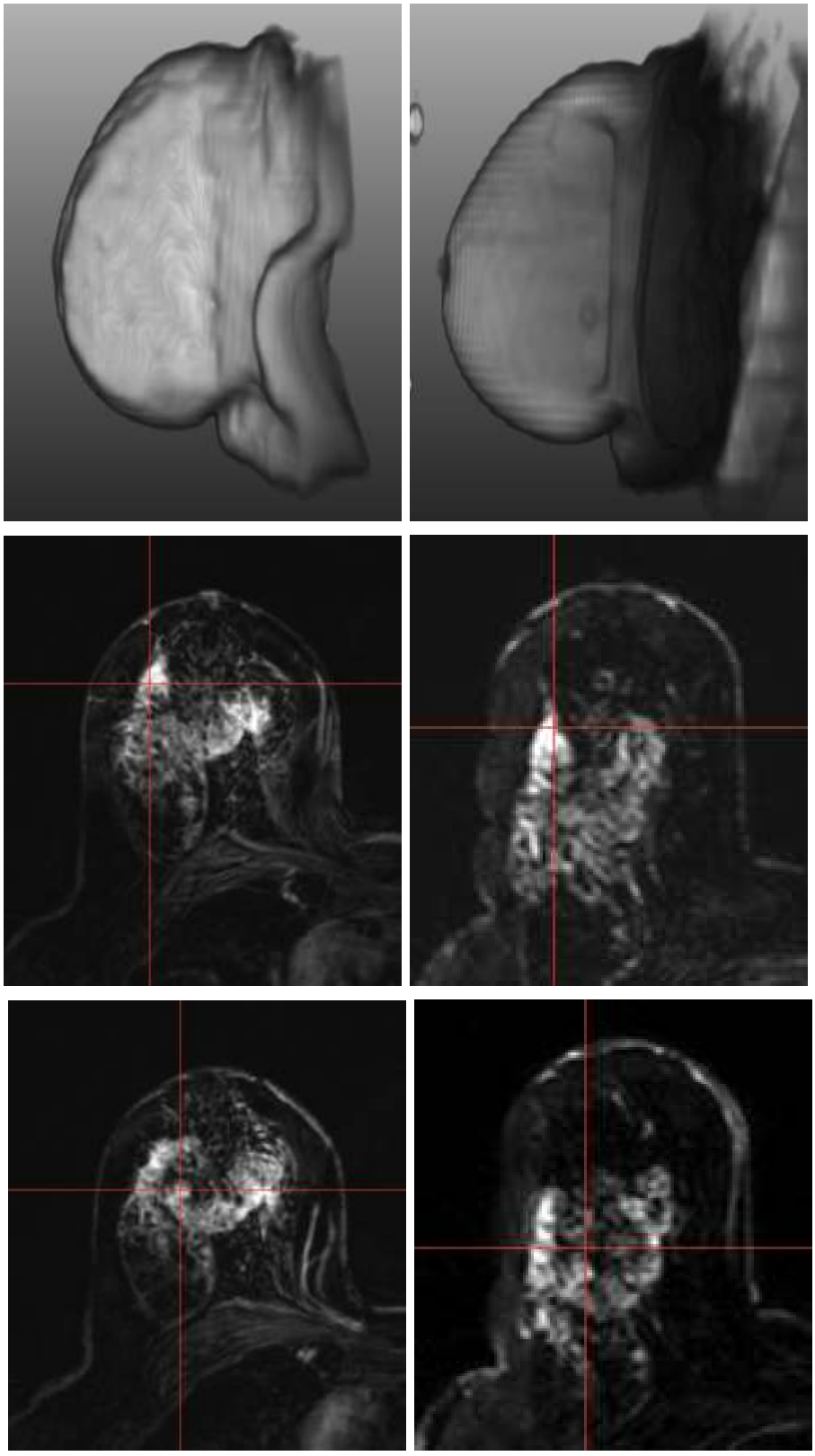


Figure 4.21: FIRST ROW: The 3d visualizations show the similarity in surface shapes after deformation simulation. Left: deformed diagnostic MRI; right: MRI acquired in the biopsy coil. MIDDLE/BOTTOM ROW: Two landmarks and their correspondences between diagnostic scan (left in each pair) and interventional scan are shown. The transfer of position from the diagnostic scan to the interventional scan has been achieved using the deformation field obtained from the registration.

soft tissue. The implants obviously are designed to deform roughly like breast tissue and should hence be included in the breast tissue mask, not the chest mask. Secondly, the deformation that is observed between diagnostic and pre-biopsy MRI indicated a more complex movement than what we currently model. In particular, it may be required to account for a two-way movement of the compressing device, first towards the chest, then medially.

A second observation is that the automatic segmentation needs to be adjusted to include the pectoral muscle in the "deformable tissue" mask. A close examination of the diagnostic and biopsy MRI reveals that with the compression applied to the breast, the muscle is also shifted and compressed. On the other hand, it appears not to be necessary to model the stiffness of the lesion. Particularly in the case with a larger lesions, the deformation was comparable to the surrounding tissue, while for small lesions, the lesion deformation will be small compared to the lesion shift, so that the first objective of future improvements is to overcome the deformation related problems.

On our limited data, we currently assess the main origins of the observed misalignment. Work we are carrying out includes to utilize different non-linear material laws in the simulation to assess the influence of the material model on the results.

In the second approach, we substituted the simulation by a hybrid linear-non-linear registration approach to establish the spatial correspondence on a fine level of detail. Here, we noticed that in particular the nipple position change between diagnostic and interventional MRI hamper a convincing registration for some cases. The fact that through the combination of linear and elastic registrations foldings of the deformation field can occur is another main drawback. In addition, the linear deformation step doesn't guarantee volume constancy, which is physically implausible.

From the biopsy cases of different sites we have learned how different the acquisition details are: the field of view depends on equipment and MRI sequence, different acquisition directions (sagittal and transversal) may reflect preferences of radiology departments, and different positioning in the biopsy coil is adapted from case to case for optimal lesion access, not for imaging reproducibility. As a visual consequence, the nipple position gravely changes from diagnostic scan to interventional scan and produces our cases with the larger errors. This has to be reflected in the algorithm by first accounting for the large and "arbitrary" deformations, which is more naturally integrated into the finite element simulation.

Overall, the biophysical modeling and simulation approach presents itself as a flexible and general method, applicable to diverse scenarios, once a domain-specific (i.e., organ specific) model has been defined. The registration based approach, on the other hand, achieves a slightly higher accuracy in the presented scenario of breast MRI guided biopsies if it is tuned to this specific application. Combining both approaches may be one further research direction, by for example adjusting the nipple position using the deformation simulation, and applying non-linear registration techniques to reach the final simulation. Initial experiments showed results that appeared to be inferior to both the individual approaches, but no thorough exploration has yet been carried out for this approach.

For the clinical scenario of MRI-guided biopsy, further support is easily provided with the approaches presented. While a forward link from the diagnostic DCE-MRI to the interventional scan is a potentially powerful technique to change the workflow from diagnosis to biopsy, the reverse mapping may help to interrogate the biopsy image for the correspondence of structures in the diagnostic scan. In addition, calculating a mapping from one peri-interventional scan to the next may help to assure successful targeting and needle placement once the contrast agent no longer visualizes the lesion. The rationale is that lesions and surrounding tissue are displaced in the order of several millimeters when the needle is advanced. This deformation is today hard to estimate from the MRI scans taken, which introduces a potential source of unsuccessful biopsies.

In addition, in breasts with many enhancing structures, sometimes the contrast agent wash-out behavior is the only differentiating and decisive factor, and is not seen in interventional DCE-MRI with two time points, looking only at the wash-in phase. This creates the necessity to identify the target among false positives with more reliability, and we believe our work provides an approach for this task as well.

4.4 A Perspective on Applications

The task to build a biomechanical breast deformation model remains challenging. The current research trends can be summarized into two directions: one is to improve the results in terms of accuracy, the other to improve the speed while maintaining a task-specific level of accuracy. The means to increase the predictive ability of biomechanical simulation of breast tissue include most importantly the application of non-linear material laws, including several kinds of isotropic and anisotropic formulations. The cost of this choice is obviously speed, and often also the user-friendliness in terms of model initialization that gets more complex with the complexity of the model.

This motivated many research efforts into simpler models, including geometric models to describe breast deformations (C Tanner et al. 2011) and statistical approaches that predict a learned target shape from a given breast image. Their potential applications are restricted either by the maximum amount of deformation, subject to the geometric model, or in the other case by the learned statistics. They offer the benefit of very fast computations, rendering them interesting either for selected problems or as a preprocessing step, followed by another method that refines the result. The proposed biomechanical simulation framework is more flexible, but can as well not claim to be physically realistic in all clinically relevant applications. It however maintains a high level of speed that allows it to be used interactively or again as a preprocessing step.

As deformations sometimes need to become larger than they can be simulated in simplifying models as well as in the proposed FEM framework, and since statistical parameter prediction approaches are inherently limited to the learned shapes, alternative ways to achieve a physically plausible and fast simulation are explored in the community. For fluid-like tissues, connected by fibers, an approach has been proposed by Costa (2012). The model conforms with some fibrous tissues in the body, and is flexible enough to model a variety of characteristics. Building the model from a surface triangulation of the body of interest, the authors propose three sources of internal body forces: a fluid-like interior filling that is incompressible, implying that, when one node looses volume, the difference needs to be assigned elsewhere; internal fibers that connect each node with each other node, and superficial fibers that connect neighboring surface nodes. Loadings and boundary conditions can also be modeled. The general idea proposed by these authors may well be extended to model the breast, although by principle the approach is limited to cases where the deformation doesn't result in folds. In this case, the convexity of the surface is violated, and the connective fibers will exert forces attracting the outsides of the shape to itself.

Many tasks in breast care have been and are being tackled by biomechanical simulation. Some examples are summarized in three general categories, but without claim of completeness or comprehensiveness:

Registration Biomechanical models can be utilized to establish correspondences between different imaging modalities, where the correspondence is hard to established based on gray values alone. Examples are mammography-to-MRI registration (Mertzanidou et al. 2012), tomosynthesis-to-MRI registration, ultrasound-to-MRI registration (C Tanner et al. 2011), and automated volume ultrasound (AVUS) to tomosynthesis (Georgii, Zöhrer, et al. 2013). The common approach is to model the deformation between the modalities by estimating the breast shape from the image information, and applying compression to convert from one shape to the other. A common property in this type of application is that biomechanical simulation mediates between two known shapes. High performance deformation estimation is imperative in some of these applications to optimize compression parameters towards minimizing the distance metric calculated between deformed template image and reference. The biopsy preparation scenario described above can be assigned into

this category, though as we have seen it can also be solved by relying on the gray values alone.

Simulation Opposed to this, scenarios exist in which the estimation or prediction of an unknown target shape is required, most prominently to visualize the deformation of breast tissue under varying gravity with the purpose of surgery preparation or support, or prediction of tumor location after deformation (Han, J. H. Hipwell, Christine Tanner, et al. 2012; M. Harz et al. 2012b; Palomar et al. 2008; Pathmanathan et al. 2008; Shih et al. 2010; Christine Tanner et al. 2011). Simulation of cosmetic outcome of implants is another practical application (Georgii, Eder, et al. 2013). More advanced topics include the simulation of surgical instruments penetrating the tissue. There is little research done in this direction specific to breast tissue. Simulation may as well be used to solve inverse problems that for example arise when external sensors yield empiric information on the displacement at sparse locations, and the forces causing them are to be estimated.

Fusion Simulation can be linked to registration. Biomechanical models may be employed as regularizers in gray-value based registration, smoothing dense deformation fields. In particular, if in an application the calculation of a dense deformation field is either not possible (for example due to only limited overlap between template and reference) or computationally prohibitive, simulation can aid to interpolate the available information. Biomechanical simulation can estimate a smooth deformation field conforming with the given information. This scenario of limited overlap of template and reference is for example met in hand-held ultrasound to MRI registration, where the US probe compresses the breast tissue. Both image information and surface deformation may be tracked, and a global deformation estimated by simulation. The proposed framework extensions allow to model any of these scenarios (compare Schiwietz et al. 2007).

There are many other applications of the framework extensions for efficient dynamic biomechanical simulations outside the area of breast care. Limiting the discussion to the medical field, the following is an incomplete list of topics where efficient simulation is being researched, and where the proposed framework with its extensions might contribute:

Other organs Deformation simulation is being asked for in a wide range of applications, and usually, the displacements and deformations are much smaller than that of breast tissue. Examples include the liver, where surgery may benefit from models that are able to predict deformations based on sparse external sensor information. Our developments have found application in this area in the context of master theses. In open brain surgery, the prediction of the brain shift once the neurocranium is opened is a persisting problem that involves deformation and movement of the brain governed by sliding with friction.

Training Haptic feedback in medical simulators is a prominent research topic. The aim is to provide realistic force feedback to the instruments (Dogan et al. 2011). For this application, the real-time speed is the most crucial requirement. Surgery simulators require additionally that the tissue can be cut or punctured, for which the required basis has been laid with the presented framework and its extensions.

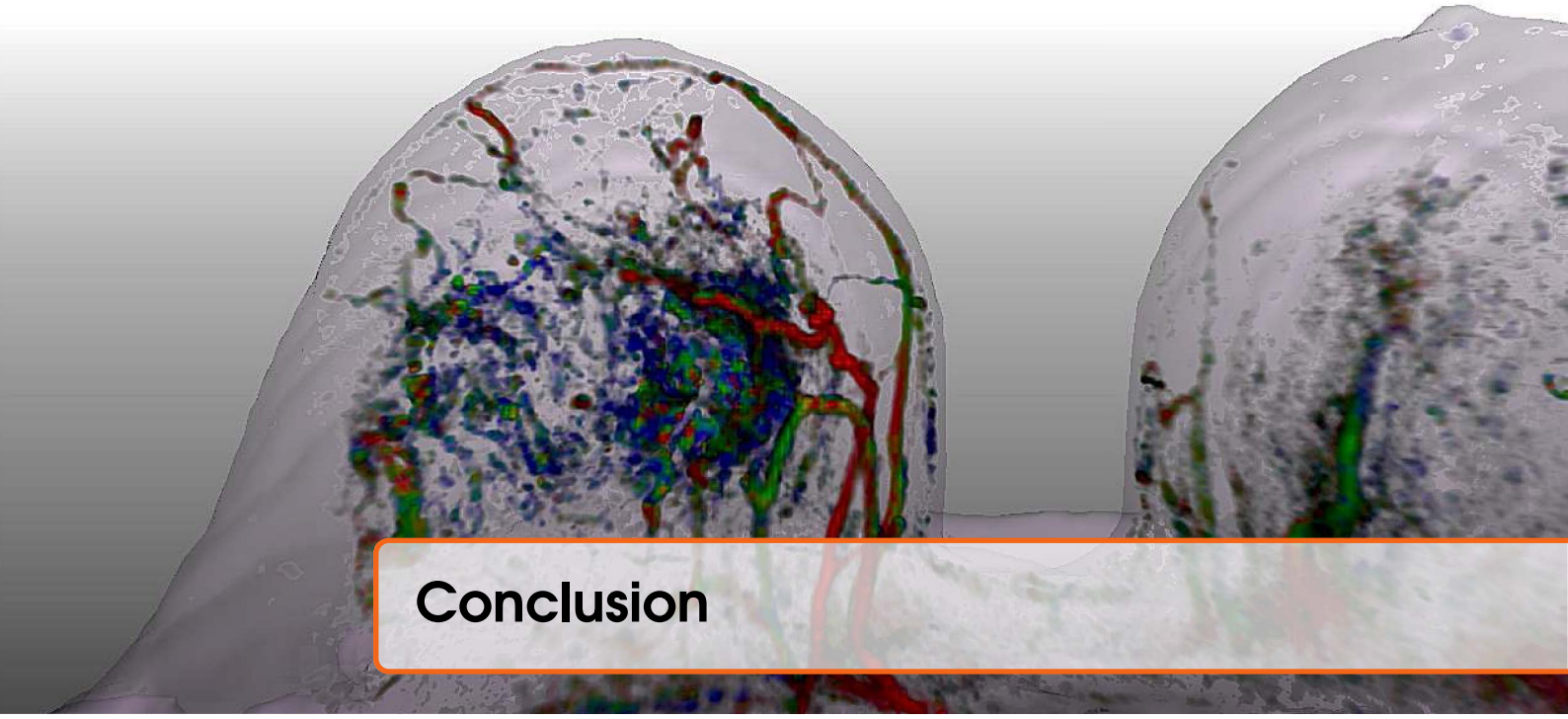
Preprocessing Some applications of fast biomechanical simulation emerge where accurate registration is the goal, but local deformations are too large to be modeled. One example is the tissue deformation when a tumor grows inside an organ (Zacharaki et al. 2008).

Validation FE based simulation can be used to create realistic deformations of organs and tissues in a reproducible, deterministic manner. This allows to create training or test data for the development or validation of other algorithms, like for example image registration algorithms (Boehler 2011) or statistical deformation models.

Some voices in the biophysical modeling community claim that from the engineering perspec-

tive, the problem of the simulation of any tissue is conceptually solved. While this may hold for a limited number of materials and applications even in the medical field, we do object the claim for the subject of breast deformation. As initially described, the shape and flexibility of the breast tissue is governed by the interplay of very differently deforming tissue types that together form a loosely coupled compound. Each component contributes to the overall appearance; their interplay, like sliding and friction, isotropic vs. anisotropic deformation, suitable non-linear material laws, is largely obscure to all imaging, and may only be determined on specimen. Even if all the tissue types and their parameters were known, it would by today's standards be practically impossible to create a biophysical simulation model that captures all the different behaviors and interactions that are known.

It is exactly this complexity that gives reason to approaches like the one presented in this chapter. The proposed efficient simulation doesn't claim to predict breast deformations with true-to-life accuracy. Rather, it helps to solve practical problems that require calculations in limited time, and with the possibility of interactive modification.



Conclusion

Contributions have been described in several parts of clinical breast cancer care, which share the common property to explore novel approaches to current clinical diagnostic or interdisciplinary challenges.

We started with an overview of the nature of breast cancer as a complex disease and introduced imaging methods for the detection and differential diagnosis of breast cancer. We enumerated the pressing clinical issues that arise from the changing preconditions in terms of work load and computer support, and derived possible approaches that help clinicians in these challenges. Along categories of clinical computer support, we approached computationally and conceptually simpler as well as more complex problem statements.

Starting with CADe and CADx, we focused on the reliable detection and diagnosis of one of the most challenging types of breast lesions, DCIS. After an introduction into the machine learning basics, motivating the use of such methods in our work, we provided an overview of the clinical importance of DCIS, and continued with the pathophysiological background to shed light on the reasons for the complicated diagnostic task of radiologists and computer algorithms alike in the diagnosis of DCIS from DCE-MRI. We have proposed a novel approach to the computer-based detection of non-mass lesions that yields candidate regions. Other than all approaches reported in literature, the bilateral symmetry of gray value neighborhoods underlies our detection method. From the resulting ROIs, we have derived shape and kinetic features to assess them in a predictor. We were able to report a predictor performance matching that of human observers. The algorithm has been evaluated on a data set comprising normal controls, DCIS cases, invasive mass-like lesions, and benign lesions of both mass like and non-mass like morphology. Second in this chapter, we presented our results on novel 4D texture features that have been explored for the differentiation of DCE-MRI detected mass lesions. Again, a machine learning approach led to results that compete with the sensitivity and specificity of human observers.

Anticipating the growing importance of breast MRI in the context of early detection of breast cancer in the screening of a high-risk subgroup of the population, we proposed the MRiPad, symbolizing a new paradigm in workflow-oriented breast image reading. Our breast MRI reading workstation prototype utilizing a multi-touch capable mobile device implements an exemplary hanging protocol for the reading of breast MRI examinations. We have in addition proposed to remove much of the menu and toolbar items to replace them by an adaptive set of functionality which we called the location and context aware toolset. This mechanism allowed us to present

the user with a minimal selection of suitable, interactive gesture-operated diagnostic tools. Underlying the prototype is a more general concept of location awareness for which we motivated applications beyond breast care and perhaps outside of hospitals.[7]

We turned to mammography screening, and using the knowledge gained in the work on breast MRI, a prototype has been proposed and investigated that shows how benefits of pre-digital screening techniques can be transported into a digital screening workstation by use of a gesture-based interface. The same paradigm as in the breast MRI scenario applies regarding the task of the mobile device. The tablet provides a gesture interaction interface, enhanced by subtle visual and audible cues to provide a user experience that is intuitive and unambiguous.

In our studies of both prototypes, we could establish not only a broad acceptance of the concepts, but also several indications that gesture-based interfaces are often qualitatively preferred by users, and don't impede accurate and fast execution of standard workflow components.

From suspicious lesion detection and screening workflow, and from image-based diagnosis and diagnostic workflow, we lastly turned our attention to interventional procedures, again based on DCE-MRI as the imaging basis. We have identified the requirement to support MRI-guided breast biopsies and breast surgery planning, and proposed to employ physics-based modeling of the breast tissue deformations for these tasks. The prediction of the large deformations the breast undergoes from the prone imaging position to the supine positioning during surgery can be the basis of both visualizations and planning tools for the interdisciplinary tumor board, supporting lesion location visualization, access planning, and risk estimation in an interactive view and perspective common to surgeons. During surgery, interactive updates can be based on the initially deformed supine shape. We have derived requirements for this scenario and provided implementations in an efficient finite element based simulation. Purpose-driven extensions of this framework proved to increase the physical realism of the simulation results.

With the same framework, we have we have contributed to the solution of the spatial correlation problem between pre-biopsy DCE-MRI images and the unilateral, compressed, low-resolution interventional images, reducing the ambiguity of lesion recovery during biopsy. Particularly in patients with many enhancing spots in the vicinity of the index lesion, this is apt to increase confidence in targeting. The method was validated on an extensive set of clinical cases, yielding target recovery errors of an order of magnitude sufficient for the application, though not sufficient to replace the interventional contrast enhanced acquisition by the simulation. We have compared the results obtained by application of the general biomechanical simulation framework with a specially adapted pipeline of image registration methods and showed that despite the careful tuning, the improvement in some cases is attended by decreased performance in others.

In summary, in this thesis contributions to topics in several areas of clinical breast cancer care have been presented, spanning from cancer detection to breast surgery, and including automation of otherwise error-prone manual tasks, giving decision aid and providing workflow improvements in DCE-MRI, and offering support for reliable targeting in MRI-guided biopsies. With these contributions to the clinical practice of image-based breast cancer care, the complexity gap, which was before bridged by subjective reasoning, is partially closed by quantitative, reproducible approaches.

[7]The concept is patent pending for medical applications.

A — Complexity: Example

A.1 Complex System, Simple Behavior

This appendix introduces an illustrative example from outside the medical world to show how a complex system displays a behavior that can be described in a very simple algorithm. The cellular automaton we will examine mimics some of the building blocks of a computing machine. It has been implemented using the Wireworld cellular automaton described by Brian Silverman in 1987 (see Wolfram 2002, p. 1117).

Cellular automata (CA) have been introduced by John von Neumann to prove that computer programs can reproduce themselves (they can; Wolfram (see ibidem, p. 876)), and have been described in many textbooks since, leading to sometimes exaggerated expectations about their explanatory power for real-world complex systems (Ilachinski 2001; Shalizi 2012). CA are based on extremely simple building blocks, often dubbed cells. These cells live on regular grids with a set of neighbors. They take on a state out of a limited set of predefined states. They follow deterministic rules that describe how they transition between states, potentially involving functions of time or external factors depending on the system that is modeled. Such cells, combined, exhibit astonishing properties when starting conditions and rule set are properly set. CA have been employed to model real-world problems, and they show patterns that are similar to several growth processes met in plants and animate life. Despite some successes, critics warn that though a CA can for example simulate the growth of spots on the leopard skin, it can provide no explanation why a leopard grows spots, while a polar bear does not (Shalizi 2012).

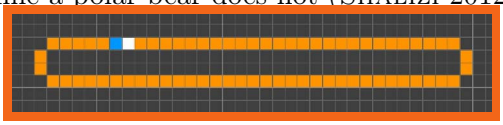


Figure A.1: The copper wire with one electron head (white) and tail (blue). Applying the rules in a dynamic simulation, the electron will move clockwise around the loop.

We consider in the following example cells that live on a regular grid and take on one of the following four states: they can be either copper, an electron head, or an electron tail, or they can be blank. A simple rule set defines the transition between the states, and semantically models how electrons (consisting of head and tail) will flow along a copper wire if the rule set is repeatedly applied to the starting configuration:

1. An electron head (white) will become a tail (blue) in the next generation.
2. An electron tail will become copper (orange).
3. If one or two neighbors of copper are electron heads, the copper becomes electron head.

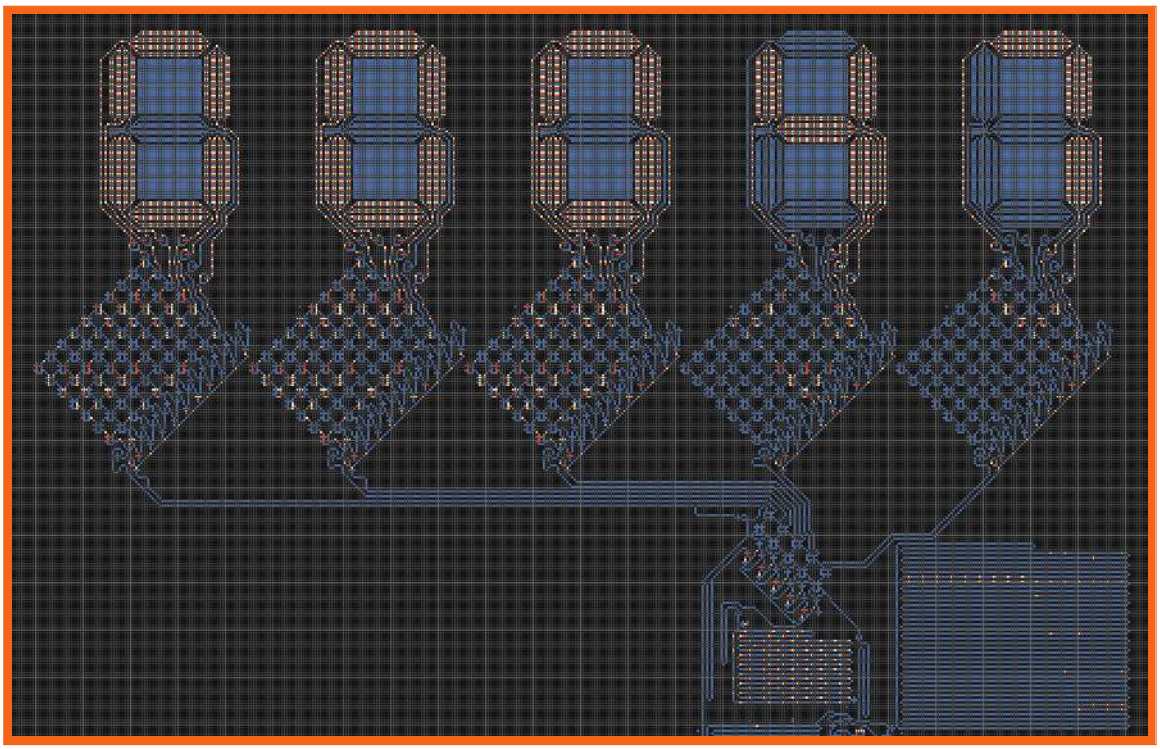


Figure A.2: A close-up of the cellular automaton calculating prime numbers. The copper wiring shows electron heads and tails in several locations.

The figure shows starting conditions that will let the one electron move along the copper wire in a clockwise direction. From this simple setup, wires and logical gates (OR, XOR, AND) can be constructed, and from them, higher-order functionality like diodes, on-off switches, circuits that do binary addition etc. Next, storage circuits can be defined that can be read and written. These circuits establish both the program memory as well as read-write registers. Clocks that emit electrons at regular intervals can be constructed from loops with exits, so that in a dynamic simulation a continuous trigger is established. These components ultimately integrate into a simple computer. The computer program can be implemented "hard wired" into the memory block, and the computer clock triggers the program.

The function of this computer design has been demonstrated by running a (brute force) prime number generator on it. Choosing a slow simulation speed allows to watch the electrons flow along the low-level circuit components. Zooming out and increasing the speed of simulation gradually lets one appreciate how a display (consisting of cells of the cellular automaton!) is eventually updated to count up prime numbers. This CA conveys a sense of how complex behavior emerges out of simple basic building blocks, and shows the multiple facets of complex systems. It takes thousands of iterations to produce the output of the first prime number, "00002", and increasingly more with every new prime number to follow, which is due to the runtime complexity of the implemented trial division algorithm (Owen 2013).

Considering this CA, the three defining characteristics of complex systems enumerated by Lloyd (2001) can be answered:

1. It is hard to describe on the cellular level, although simpler building blocks make up the whole design and are easier to describe.
2. It is not simple to recreate on the cellular level and without an understanding of the basic building blocks.
3. It is a multi-level system with a high level of organization.

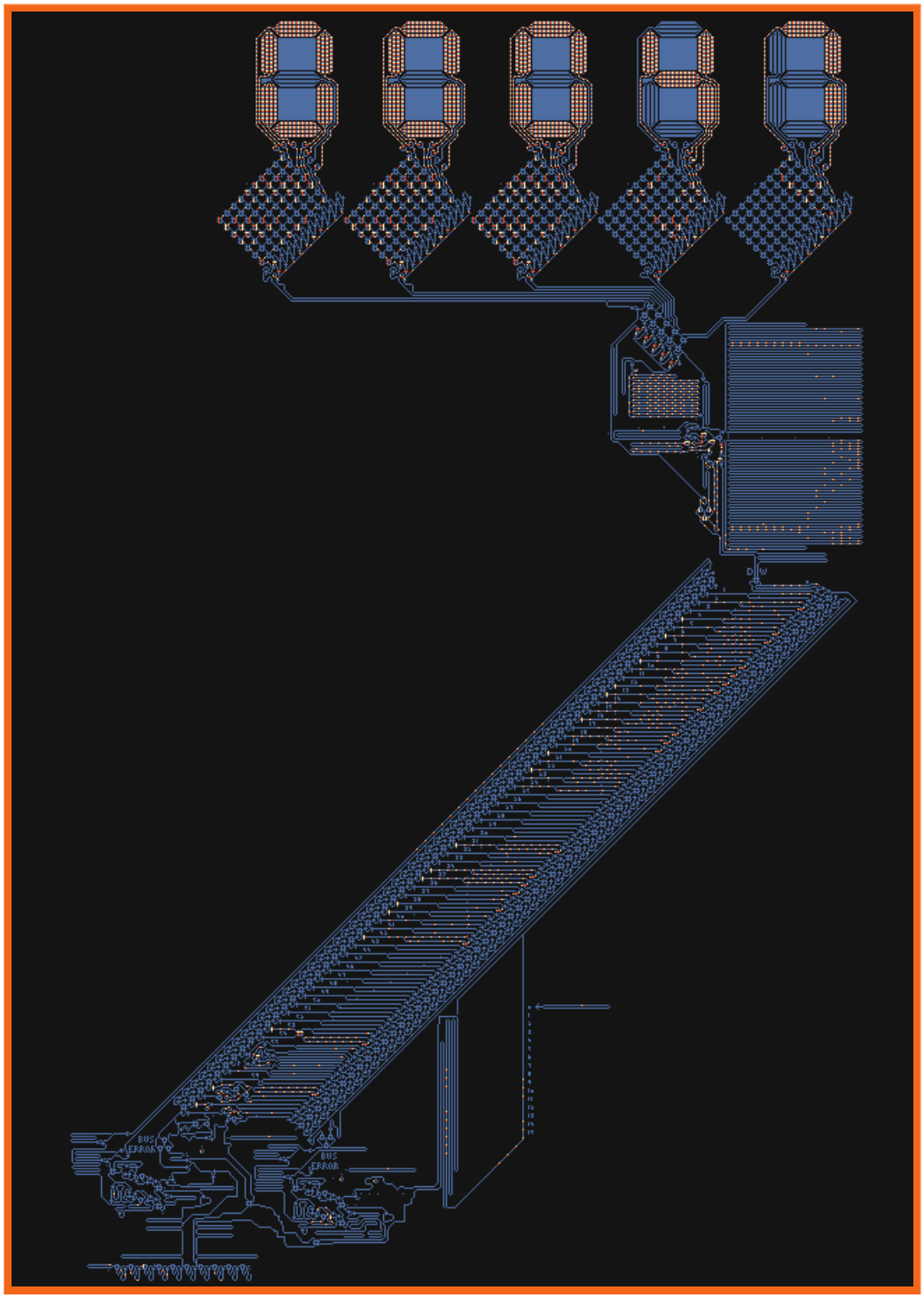


Figure A.3: The CA prime number generator has produced the number 47, using about 9.000.000 iterations. Components in the image are the "LCD display" (top), units to convert a decimal digit into a seven-segment display (below), and converters from binary to decimal. The diagnoal structures represent the registers in which the program is stored in machine code. The lower part of the diagonal structure contains the arithmetic-logic unit (ALU), and the clock and pulse generators for read and write access are shown underneath.

Also, on the scale of the structural criteria of Flake (1998), the system might be considered complex: it is parallel because it consists of multiple instances of basic building blocks that work concurrently, it is repetetive and may react on external input (although it practically doesn't), but it is not adaptive in that it cannot learn from feedback it receives. It is fully determined by its starting conditions and has no interaction with the outer world, although it is possible to think of extensions that allow for example the modification of the program.

Interestingly, the essential property or emergent behavior of the system can be described by the algorithm that it implements. The system as a whole is hence a fairly complicated, perhaps almost complex system, but one that can be predicted with a much simpler to understand and reproduce algorithm. It is very challenging to understand what the system is doing by looking at it statically. It is still an undertaking to understand it when watching it operate, and yet both can be achieved with enough time and dedication. It still figures as an example of how complexity emerges from a multi-level stacking of more basic building blocks: Consider an extension of the CA computer to store a much larger program and data, for example one that implements any advanced machine learning methods that takes input and provides output. Although theoretically, it will still be a fully deterministic system on the electron level, the prediction of output will be impossible to predict on the CA level.

There are other examples for complex systems which are found in nature (ant colonies, bees, and the human brain), in economics (stock markets), psychology and many others. Compared to the CA computer, they are harder to dissect. Such systems have been studied and described in many projects, but many of them can still not be predicted, nor can their state be fully sampled, and sometimes even the variables to sample to describe the system fully are unknown. Concluding, no complex systems can be identified that can be treated (diagnosed, understood) with a simple algorithm.

B — iPad MRI Reading Workflow Prototype

B.1 Questionnaire for Workflow Evaluation

Questionnaire Breast MRI i

Markus Harz

Fraunhofer MEVIS

06. April 2010

Questionnaire: The Radiologists View on Breast MRI

Exploring the way radiologists of different levels of experience conceive of image formation, image appearance, and image meaning will yield new insights into optimal support from the software side. We hypothesize that the mental image of computer scientists is inherently different from the image radiologists have of images, workflow, ...

I.	What do you know about image formation?
A.	T1
B.	T2
C.	T2*
D.	Dynamic contrast enhances series
II.	What do you know about the influence of contrast agents?
A.	Types of CAs
B.	Quantity to apply
C.	Time resolution of dynamic protocol

(a)

Questionnaire Breast MRI ii

III. How do you know when images can not be used diagnostically?
A. Motion artifacts?
B. FOV too small?
C. Contrast did not arrive?
D. Parenchyma enhancement?
IV. Breast MR reading workflow
A. Which tools do you use?
1. for segmentation
2. for measuring
B. Which facilities do you miss?
1. for segmentation
2. for measuring
3. for reporting
C. How do you assess sizes in more detail:
1. ... of lumps?
2. ... of regional enhancements?
D. How do you report the location of a finding?
1. Which reconstructions do you use?
2. How do you measure distances to fiducials?

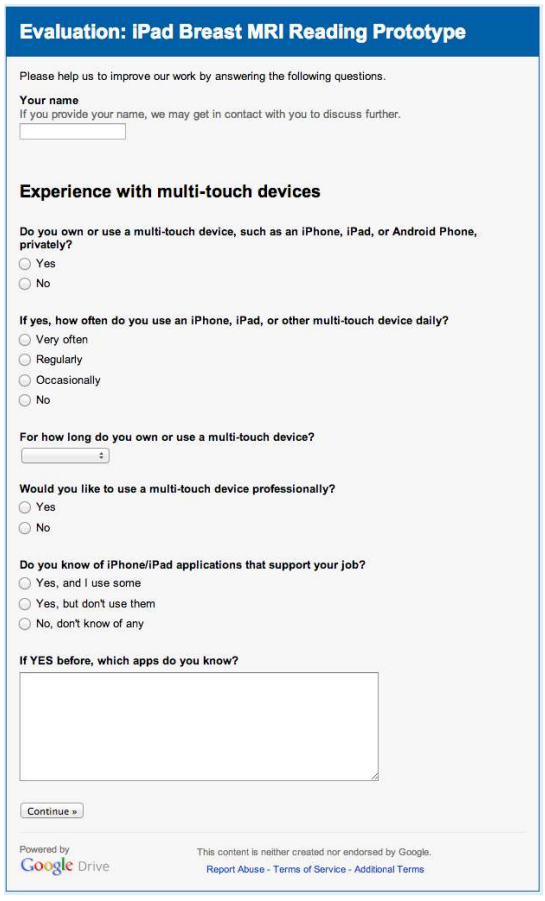
(b)

Questionnaire Breast MRI ii

III. How do you know when images can not be used diagnostically?
A. Motion artifacts?
B. FOV too small?
C. Contrast did not arrive?
D. Parenchyma enhancement?
IV. Breast MR reading workflow
A. Which tools do you use?
1. for segmentation
2. for measuring
B. Which facilities do you miss?
1. for segmentation
2. for measuring
3. for reporting
C. How do you assess sizes in more detail:
1. ... of lumps?
2. ... of regional enhancements?
D. How do you report the location of a finding?
1. Which reconstructions do you use?
2. How do you measure distances to fiducials?

(c)

B.2 Questionnaire for Qualitative Evaluation



Evaluation: iPad Breast MRI Reading Prototype

Please help us to improve our work by answering the following questions.

Your name
If you provide your name, we may get in contact with you to discuss further.

Experience with multi-touch devices

Do you own or use a multi-touch device, such as an iPhone, iPad, or Android Phone, privately?
- Yes
- No

If yes, how often do you use an iPhone, iPad, or other multi-touch device daily?
- Very often
- Regularly
- Occasionally
- No

For how long do you own or use a multi-touch device?

Would you like to use a multi-touch device professionally?
- Yes
- No

Do you know of iPhone/iPad applications that support your job?
- Yes, and I use some
- Yes, but don't use them
- No, don't know of any

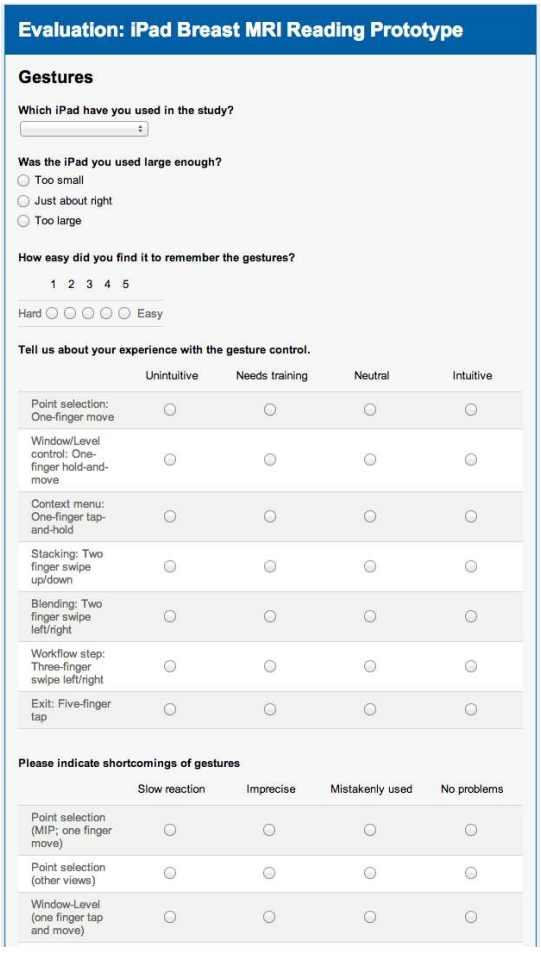
If YES before, which apps do you know?

Continue »

Powered by Google Drive

This content is neither created nor endorsed by Google.
Report Abuse - Terms of Service - Additional Terms

(d)



Evaluation: iPad Breast MRI Reading Prototype

Gestures

Which iPad have you used in the study?

Was the iPad you used large enough?
- Too small
- Just about right
- Too large

How easy did you find it to remember the gestures?

1 2 3 4 5

Hard ○ ○ ○ ○ ○ Easy

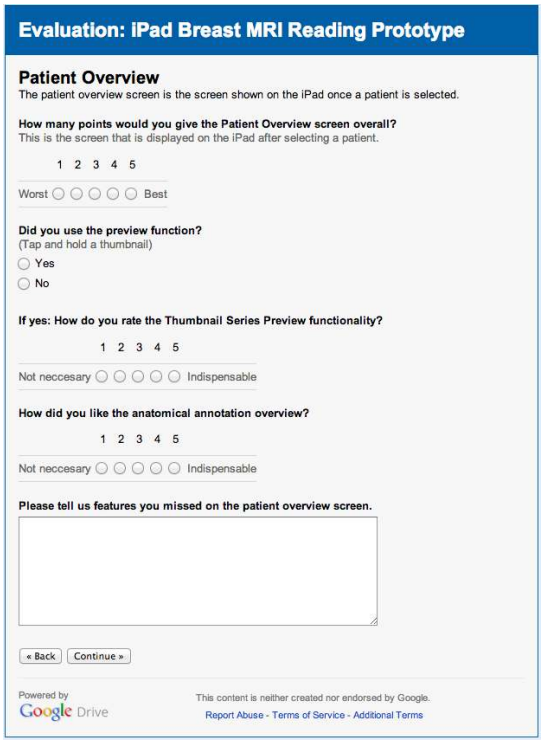
Tell us about your experience with the gesture control.

	Unintuitive	Needs training	Neutral	Intuitive
Point selection: One-finger move	○	○	○	○
Window/Level control: One-finger hold-and-move	○	○	○	○
Context menu: One-finger tap-and-hold	○	○	○	○
Stacking: Two finger swipe up/down	○	○	○	○
Blending: Two finger swipe left/right	○	○	○	○
Workflow step: Three-finger swipe left/right	○	○	○	○
Exit: Five-finger tap	○	○	○	○

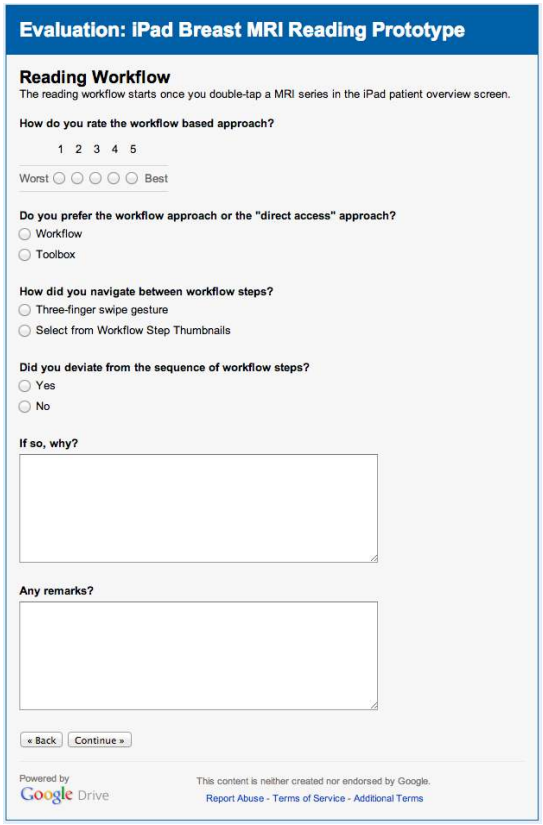
Please indicate shortcomings of gestures

	Slow reaction	Imprecise	Mistakenly used	No problems
Point selection (MIP; one finger move)	○	○	○	○
Point selection (other views)	○	○	○	○
Window-Level (one finger tap and move)	○	○	○	○

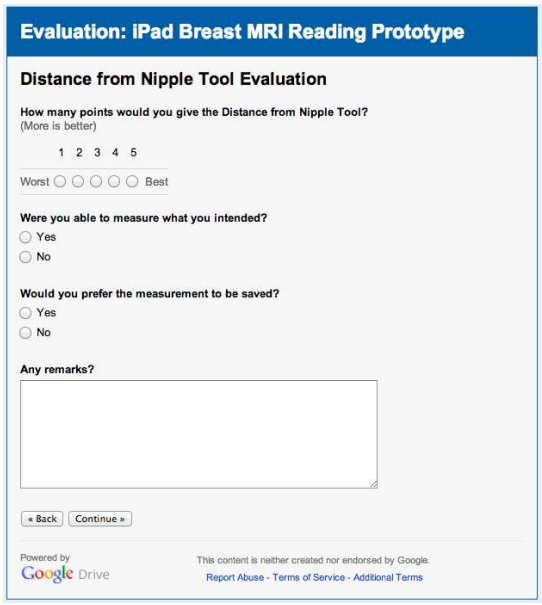
(e)


Evaluation: iPad Breast MRI Reading Prototype
Patient Overview
The patient overview screen is the screen shown on the iPad once a patient is selected.
How many points would you give the Patient Overview screen overall?
This is the screen that is displayed on the iPad after selecting a patient.
1 2 3 4 5
Worst Best
Did you use the preview function?
(Tap and hold a thumbnail)
Yes
No
If yes: How do you rate the Thumbnail Series Preview functionality?
1 2 3 4 5
Not neccesary Indispensable
How did you like the anatomical annotation overview?
1 2 3 4 5
Not neccesary Indispensable
Please tell us features you missed on the patient overview screen.
« Back
Continue »
Powered by Google Drive
This content is neither created nor endorsed by Google.
Report Abuse - Terms of Service - Additional Terms

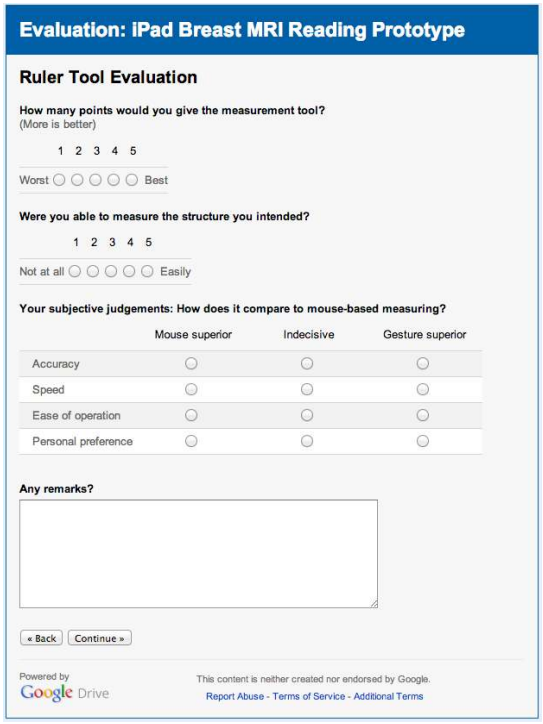
(f)


Evaluation: iPad Breast MRI Reading Prototype
Reading Workflow
The reading workflow starts once you double-tap a MRI series in the iPad patient overview screen.
How do you rate the workflow based approach?
1 2 3 4 5
Worst Best
Do you prefer the workflow approach or the "direct access" approach?
Workflow
Toolbox
How did you navigate between workflow steps?
Three-finger swipe gesture
Select from Workflow Step Thumbnails
Did you deviate from the sequence of workflow steps?
Yes
No
If so, why?
Any remarks?
« Back
Continue »
Powered by Google Drive
This content is neither created nor endorsed by Google.
Report Abuse - Terms of Service - Additional Terms

(g)


Evaluation: iPad Breast MRI Reading Prototype
Distance from Nipple Tool Evaluation
How many points would you give the Distance from Nipple Tool?
(More is better)
1 2 3 4 5
Worst Best
Were you able to measure what you intended?
Yes
No
Would you prefer the measurement to be saved?
Yes
No
Any remarks?
« Back
Continue »
Powered by Google Drive
This content is neither created nor endorsed by Google.
Report Abuse - Terms of Service - Additional Terms

(h)


Evaluation: iPad Breast MRI Reading Prototype
Ruler Tool Evaluation
How many points would you give the measurement tool?
(More is better)
1 2 3 4 5
Worst Best
Were you able to measure the structure you intended?
1 2 3 4 5
Not at all Easily
Your subjective judgements: How does it compare to mouse-based measuring?
Mouse superior
Indecisive
Gesture superior
Accuracy
Speed
Ease of operation
Personal preference
Any remarks?
« Back
Continue »
Powered by Google Drive
This content is neither created nor endorsed by Google.
Report Abuse - Terms of Service - Additional Terms

(i)

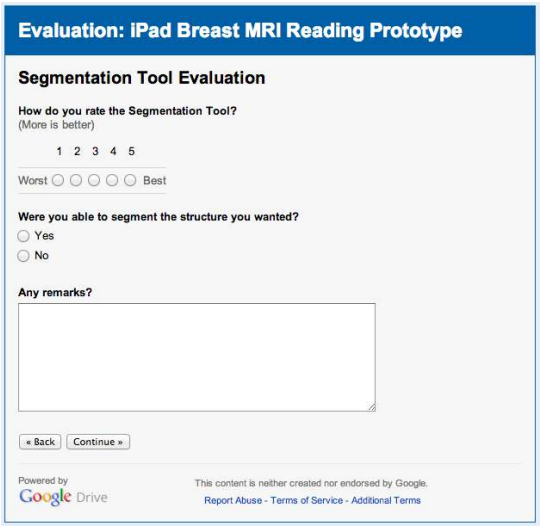
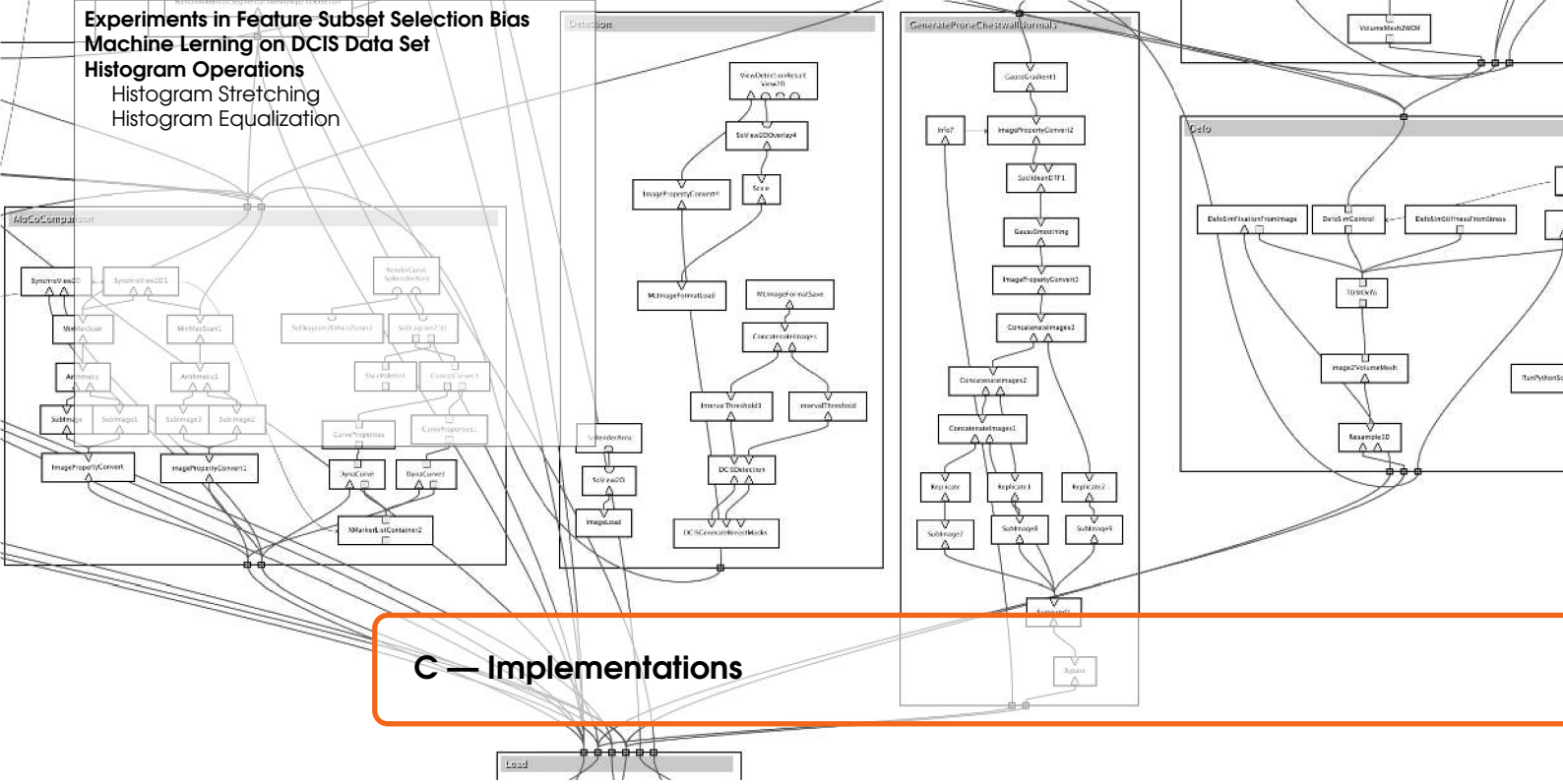


Figure B.1: This and previous pages: Evaluation forms for the Breast MRI iPad Reading Workflow Prototype.



C.1 Experiments in Feature Subset Selection Bias

The first listing shows how a dataset of 300 attributes, drawn from the uniform distribution, is constructed and classified using a Random Forest and a Naïve Bayes classifier. Different numbers of examples are generated.

```
import Orange
import orangecontrib.earth
import numpy

# Generate a domain with 300 features
features =[Orange.feature.Continuous(str(x)) for x in range(300)]
class_att = Orange.feature.Discrete("class", values=["0", "1"])
dom = Orange.data.Domain(features, class_att)

allScoresRF = []
allScoresNB = []
print "Examples, AUC (RF), AUC (NB)"

# Run experiments with 50...1000 examples, and record all AUCs
for numExpl in range(50,101,10):
        values = numpy.random.rand(numExpl,301)
        values[0:numExpl/2,300] = "0"
        values[numExpl/2:numExpl,300] = "1"

        td = Orange.data.Table(dom, values)
        # Feature subset selection using Earth Importance score on train data.
        filteredTrain = Orange.feature.selection.FilterBestN(data=td,
            measure=orangecontrib.earth.ScoreEarthImportance(), n=20)

        # Initialize the learner (need to initialize something iterable)
        learner = (Orange.ensemble.forest.RandomForestLearner(),
            Orange.classification.bayes.NaiveLearner())
        # Leave-One-Out training on the feature-reduced training set. This gives one
            classifier for each fold, i.e. as many as examples.
        loo = Orange.evaluation.testing.leave_one_out(learner, filteredTrain,
            store_classifiers=False)
```

```
        print str(numExpl) + ", " + str(Orange.evaluation.scoring.AUC(loo)[0]) + ", "
            + str(Orange.evaluation.scoring.AUC(loo)[1])
        allScoresRF.append(Orange.evaluation.scoring.AUC(loo)[0])
        allScoresNB.append(Orange.evaluation.scoring.AUC(loo)[1])
```

The following implementation was used to train a Random Forest and a Naive Bayes classifier on a data set after feature subset selection. To prevent selection bias, a fraction of the data is held out from the examples, and only on the remainder the classifier is trained and evaluated in a leave-one-out fashion. Each resulting classifier is retained and run on the test set, and all results are averaged.

```
import Orange
import orangecontrib.earth

def orangeMLScript():
        # Load data
        td = Orange.data.Table('/path/to/data.csv')

        numRuns = 50
        averagedRFScores = []
        averagedNBScores = []
        averagedRFScoresTrainSet = []
        averagedNBScoresTrainSet = []
        # Set up indexing scheme with a 75%-25% random split, stratified.
        indices2 = Orange.data.sample.SubsetIndices2(p0=0.75)
        indices2.stratified = indices2.Stratified

        for numFeatures in range(10,301,10):
                totalScoreSumRF = 0.0
                totalScoreSumNB = 0.0
                totalScoreSumTrainRF = 0.0
                totalScoreSumTrainNB = 0.0
                allScoresRF = []
                allScoresNB = []
                print "============================================================"
                print "Features: " + str(numFeatures)
                # Create splits to average classifier performance
                for fold in range(numRuns):
                        # Generate fold with new random seed
                        indices2.random_generator = Orange.misc.Random(fold)
                        ind = indices2(td)
                        tdTrain = td.select(ind, 0) # Training instances
                        tdTest = td.select(ind, 1) # Testing instances

                        # Feature subset selection using Earth Importance score on
                            train data.
                        filteredTrain =
                            Orange.feature.selection.FilterBestN(data=tdTrain,
                            measure=orangecontrib.earth.ScoreEarthImportance(),
                            n=numFeatures)
                        # Restrict the test data portion to the selected features.
                        filteredTest = Orange.data.Table(filteredTrain.domain, tdTest)

                        # Initialize the learner (need to initialize something iterable)
                        learner = (Orange.ensemble.forest.RandomForestLearner(),
                            Orange.classification.bayes.NaiveLearner())
```

```
                # Cross-validated training on the feature-reduced training set.
                    This gives one classifier for each fold, i.e. as many as
                    examples.
                cv = Orange.evaluation.testing.cross_validation(learner,
                    filteredTrain, store_classifiers=True)

                totalScoreSumTrainRF += Orange.evaluation.scoring.AUC(vc)[0]
                totalScoreSumTrainNB += Orange.evaluation.scoring.AUC(vc)[1]

                # Run each classifier on the test set, and average their AUCs
                scoreSumRF = 0.0
                scoreSumNB = 0.0
                for classifier in cv.classifiers:
                        resultOnTest =
                            Orange.evaluation.testing.test_on_data(classifier,
                            filteredTest)
                        scoreSumRF +=
                            Orange.evaluation.scoring.AUC(resultOnTest)[0]
                        scoreSumNB +=
                            Orange.evaluation.scoring.AUC(resultOnTest)[1]

                totalScoreSumRF += scoreSumRF / float(len(loo.classifiers))
                totalScoreSumNB +=scoreSumNB / float(len(loo.classifiers))

        averagedRFScores.append(totalScoreSumRF/numRuns)
        averagedRFScoresTrainSet.append(totalScoreSumTrainRF/numRuns)
        averagedNBScores.append(totalScoreSumNB/numRuns)
        averagedNBScoresTrainSet.append(totalScoreSumTrainNB/numRuns)

    return averagedRFScores, averagedRFScoresTrainSet, averagedNBScores,
        averagedNBScoresTrainSet
```

Complementary to the figure for Random Forests on random data, the following figure shows the AUC of the Naïve Bayes classifier on 10-1,000 examples with 10-300 features drawn from a uniform distribution.

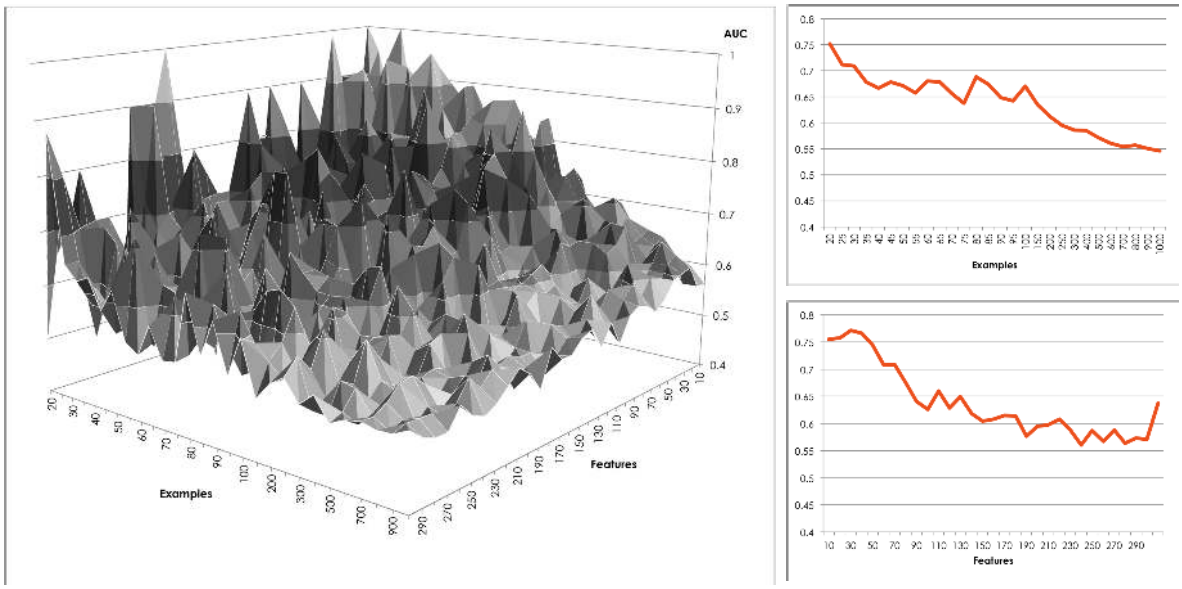


Figure C.1: Illustration of Naïve Bayes bias on random data. Left: Variation of AUC with numbers of examples and selected features. Right top: AUC of Naïve Bayes averaged over all numbers of examples. Right bottom: AUC of Naïve Bayes averaged over all numbers of selected features.

C.2 Machine Lerning on DCIS Data Set

The following script exemplifies how the python script interface to the Orange machine learning toolkit has been used to conduct the experiments on the DCIS data set. In a leave-one-patient-out cross validation, all intermediate results are printed to console. The script allows to either balance the data, or use it as it is (see comments).

Here, a two-class problem is solved. Orange as well allows to deal with multi-class classification problems, then returning performance figures for all pairs of classes. Two evaluation variants are visible: In the implementation as it is shown below, AUC is calculated using the Wilcoxon method; in addition, sensitivity and specificity are evaluated. This requires the test set to contain both classes in a binary classification task, and all classes in multi-class classification.

```
import Orange
import orangecontrib.earth
import numpy as np

def leaveOnePatientOutScript():
        # Load data
        td                              = Orange.data.Table('inputData.csv')
        domain                  = td.domain
        ml_domain               = Orange.data.Domain(["slope1", "slope2", "integral",
                                    "ttp", "relInitialEnh", "relPeakEnh",
                                        "washInWashOutRatio",
                                    "eVal1", "eVal2", "dotProd", "projVariance",
                                        "sym3Davg",
                                    "sym3Dtotal", "sym1Davg", "sym1Dtotal", "class"],
                                        domain)
        ml_data                 = Orange.data.Table(ml_domain, td)
        imageid_domain  = Orange.data.Domain(["image_id"], domain)
        image_ids               = Orange.data.Table(imageid_domain, td)
        image_ids.remove_duplicates()

        foldIndex = []
        idIndex = 0
        for i in range(len(td)):
                if td[i]["image_id"] == image_ids[idIndex][0]:
                        foldIndex.append(idIndex)
                else:
                        idIndex = idIndex + 1
                        foldIndex.append(idIndex)

        # Start Leave-Patient-Out CV
        for i in range(len(image_ids)):
                trainingSamples = ml_data.select_ref(foldIndex, i, negate = True)
                testSamples     = ml_data.select_ref(foldIndex, i, negate = False)
                # Separate out the "0" cases from the data
                class0Train     = trainingSamples.filter({"class":"0"})
                class1Train     = trainingSamples.filter({"class":"0"},negate=1)
                a1, c1, w1              = class1Train.toNumpy() # Convert both to numpy
                    arrays to join them.
                # Balance data:
                # nFolds                        = len(class0Train)/(len(class1Train))
                # Don't balance data:
                nFolds                  = 1

                #Create folds for the normal/benign cases.
```

```
class0TrainFolds = Orange.data.sample.SubsetIndicesCV(class0Train,
    folds=nFolds)
for fold in range(nFolds):
        class0Filtered  = class0Train.select_ref(class0TrainFolds, fold)
        a0, c0, w0           = class0Filtered.toNumpy() # Convert both
            to numpy arrays to join them.
        trainArrayFeat = np.concatenate((a0,a1), axis=0)
        trainArrayCls  = np.concatenate((c0,c1), axis=0)
        trainArrayClsRS =
            trainArrayCls.reshape(trainArrayCls.shape[0],1)
        finalTrainArray =
            np.concatenate((trainArrayFeat,trainArrayClsRS), axis=1)
        finalTrainData = Orange.data.Table(ml_domain, finalTrainArray)

        rfLearner   = Orange.ensemble.forest.RandomForestLearner()
        nbLearner   = Orange.classification.bayes.NaiveLearner()
        rfClassifier = rfLearner(finalTrainData) # Trains a classifier
            on the data. Undisclosed training...
        nbClassifier = nbLearner(finalTrainData)

        # Test the two classifiers
        testResult     =
            Orange.evaluation.testing.test_on_data([rfClassifier,
            nbClassifier], testSamples)

        # Evaluate the classifier with AUC, classification accuracy,
            and Brier Score.
        rfAUC = Orange.evaluation.scoring.AUC(testResult)[0]
        nbAUC = Orange.evaluation.scoring.AUC(testResult)[1]
        rfCA  = Orange.evaluation.scoring.CA(testResult)[0]
        nbCA  = Orange.evaluation.scoring.CA(testResult)[1]
        rfBrier = Orange.evaluation.scoring.Brier_score(testResult)[0]
        nbBrier = Orange.evaluation.scoring.Brier_score(testResult)[1]

        # Alternatively, evaluate data by sensitivity, specificity, and
            AUC calculated by the Wilcoxon method.
        # This only works if there are positive and negative instances
            in each fold.
        if len(testSamples.filter({"class":"0"},negate=1)) > 0 and
            len(testSamples.filter({"class":"0"},negate=0)) > 0:
                rfAUC =
                    Orange.evaluation.scoring.AUCWilcoxon(testResult,
                    class_index = 1)[0][0]
                nbAUC =
                    Orange.evaluation.scoring.AUCWilcoxon(testResult,
                    class_index = 1)[1][0]
                rfSensi =
                    Orange.evaluation.scoring.Sensitivity(testResult)[0]
                rfSpeci =
                    Orange.evaluation.scoring.Specificity(testResult)[0]
                nbSensi =
                    Orange.evaluation.scoring.Sensitivity(testResult)[1]
                nbSpeci =
                    Orange.evaluation.scoring.Specificity(testResult)[1]
                # Print CSV data that can easily read into Excel etc.
                print "%s, %d,     %5.3f, %5.3f,    %5.3f,  %5.3f,
                    %5.3f,    %5.3f" % (image_ids[i][0], fold, rfSensi,
                    rfSpeci, rfAUC, nbSensi, nbSpeci, nbAUC)
```

```
        pass
```

C.3 Histogram Operations

C.3.1 Histogram Stretching

Histogram stretching essentially distributes the occurrences of gray values across a given new value range. This is equal to a rescaling of the image gray values to a new range before computing the histogram.

The shape of the histogram (or, in general, the probability density function, PDF, of the distribution the histogram has been obtained from) qualitatively doesn't change with this operation, while the mean and standard deviation do change.

The operation by which each gray value g has to be transformed to spread a histogram that covers values in the range g_{min} to g_{max} to a new range given by h_{min} and h_{max} is described by

$$s(g) = h_{min} + \frac{h_{max} - h_{min}}{g_{max} - g_{min}}(g - g_{min}) \tag{C.1}$$

C.3.2 Histogram Equalization

Histogram equalization is a computation that makes the Cumulative Distribution Function (CDF) corresponding to a histogram linear. This operation amplifies gray values that are underrepresented in the image, and suppressing those that are dominant.

With an image I with gray values ranging from g_{min} to g_{max}, and the probability of a gray value g being $p(g)$, the CDF can be defined as

$$\mathrm{CDF}(g_0) = \sum_{x=0}^{g_0} p(x) \tag{C.2}$$

Transforming the gray level of each image voxel with the equalization function

$$e(g) = g_{max}\mathrm{CDF}(g) \tag{C.3}$$

will result in an image that has a histogram with a linearized CDF. Detail and a derivation can be found in Gonzales et al. (2007).

Bibliography

Books

American Cancer Society (2013b). *Breast Cancer Facts & Figures 2013-2014*. Edited by American Cancer Society. Atlanta: American Cancer Society (cited on page 17).

Bathe, Klaus-Jürgen (2002). *Finite Element Procedures*. Prentice Hall (cited on pages 166, 172).

Bellman, R (1961). *Adaptive Control Processes: A Guided Tour*. 'Rand Corporation. Research studies. Princeton University Press (cited on page 49).

Berg, W A (2006). *Diagnostic imaging: Breast*. Amirsys (cited on pages 34, 122).

Bishop, C M (1995). *Neural Networks for Pattern Recognition*. Clarendon Press (cited on pages 50, 61).

Boyd, Stephen and Lieven Vandenberghe (2004). *Convex Optimization*. New York, NY, USA: Cambridge University Press. isbn: 0521833787 (cited on page 46).

Braess, Dietrich (2013). *Finite Elemente. Theorie, Schnelle Löser und Anwendungen in der Elastizitätstheorie*. 5. Auflage. Springer (cited on pages 153, 156, 159).

Callaghan, Paul T (1993). *Principles of Nuclear Magnetic Resonance Microscopy*. Oxford science publications. Oxford University Press (cited on page 34).

Castleman, Kenneth R (1996). *Digital Image Processing*. Prentice Hall (cited on page 89).

Chou, Pei Chi and Nicholas J Pagano (1967). *Elasticity: Tensor, Dyadic, and Engineering Approaches*. D. Van Nostrand Company, Inc. (cited on page 154).

Dabbs, D J (2012). *Breast Pathology*. Saunders W.B. Elsevier/Saunders (cited on page 11).

The chapter image is licensed under the Creative Commons Attribution 2.0 Generic license.

Elster, A D and J H Burdette (2001). *Questions and Answers in Magnetic Resonance Imaging*. Mosby, a Harcourt Health Sciences Company (cited on page 34).

Flake, Gary William (1998). *The Computational Beauty of Nature*. The MIT Press (cited on pages 5, 208).

Garrett, Jesse James (2011). *The Elements of User Experience*. New Riders, 1249 Eigth Street (cited on page 124).

Gonzales, Rafael C and Richard E Woods (2007). *Digital Image Processing*. 3rd edition. Pearson Prentice Hall (cited on page 220).

Haacke, E M, R W Brown, M R Thompson, and R Venkatesan (1999). *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Wiley (cited on page 34).

Hand, L N and J D Finch (1998). *Analytical Mechanics*. Cambridge University Press (cited on page 154).

Hayat, M A, editor (2008). *Methods of Breast Cancer Diagnosis, Therapy, and Prognosis*. Springer (cited on page 66).

Hebb, Donald O (1949). *The Organization of Behavior*. New York: Wiley (cited on page 43).

Ilachinski, Andrew (2001). *Cellular Automata: A Discrete Universe*. Singapore: World Scientific. isbn: 9-810-24623-4 (cited on page 205).

Johnson, Claes (1987). *Numerical solution of partial differential equations by the finite element method*. Cambridge University Press (cited on page 162).

Kaiser, W A (2008). *Signs in MR-Mammography*. Springer-Verlag (cited on page 20).

Kaiser, Werner A (2008). *Signs in MR-Mammography*. 1st edition. Springer (cited on page 89).

Keeler, J (2006). *Understanding NMR Spectroscopy*. Wiley (cited on page 29).

Larson, Mats G and Frederik Bengzon (2013). *The Finite Element Method: Theory, Implementation, and Applications*. Springer (cited on pages 153, 156–159, 162).

Leitlinienprogramm Onkologie (2012). *Interdisziplinaäre S3-Leitlinie für die Diagnostik, Therapie und Nachsorge des Mammakarzinoms. Langversion 3.0, Aktualisierung 2012* (cited on pages 119, 186).

Levitt, M H (2001). *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. Wiley (cited on page 34).

Malek, D, P Rabe, and K Bock (2009). *Evaluationsbericht 2008–2009*. Ergebnisse des Mammographie-Screening-Programms in Deutschland. Kooperationsgemeinschaft Mammographie (cited on page 16).

McRobbie, Donald W, Elizabeth A Moore, Martin J Graves, and Martin R Prince (2003). *MRI from Picture to Proton*. Cambridge University Press. isbn: 0521523192 (cited on page 34).

Miller, A (2002). *Subset Selection in Regression*. Chapman {&} Hall/CRC Monographs on Statistics {&} Applied Probability. Taylor {&} Francis (cited on page 52).

Milton, Graeme W (2002). *The Theory of Composites*. Cambridge University Press (cited on page 157).

Minsky, Marvin and Seymour Papert (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA, USA: MIT Press (cited on page 43).

Mitchell, Melanie (2009). *Complexity. A Guided Tour*. Oxford University Press (cited on page 6).

Mukherjee, Siddhartha (2011). *The Emperor of All Maladies: A Biography of Cancer*. Fourth Estate (GB). isbn: 0007250916 (cited on pages 9, 13, 27, 137).

Olver, Peter (2013). *Introduction to Partial Differential Equations*. 2014th edition. Springer (cited on page 153).

Reimer, P, P M Parizel, and F Stichnoth (2003). *Klinische MR-Bildgebung: Eine praktische Anleitung*. Springer Verlag (cited on page 34).

Russo, J and I H Russo (2004). *Molecular Basis of Breast Cancer. Prevention and Treatment*. Springer (cited on pages 23, 24).

Schölkopf, Bernhard and Alexander J Smola (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press. isbn: 0262194759 (cited on pages 43, 44, 46, 47, 61).

Tabár, L, T Tot, and P B Dean (2007). *Casting Type Calcifications: Sign of a Subtype with Deceptive Features*. Breast Cancer: Early Detection With Mammography. Thieme. isbn: 9781588905802. url: `http://books.google.de/books?id=d0MglwCOoagC` (cited on page 26).

— (2008). *Crushed Stone-like Calcifications: The Most Frequent Malignant Type*. Breast Cancer - Early Detection with Mammography Series v. 2. Thieme. isbn: 9783131485311. url: `http://books.google.de/books?id=N4dJvJ6isi4C` (cited on page 26).

Torby, B J (1984). *Advanced dynamics for engineers*. HRW series in mechanical engineering. Holt, Rinehart, and Winston (cited on page 154).

Witten, Ian H and Eibe Frank (2005). *Data Mining: Practical machine learning tools and techniques*. San Francisco (cited on pages 59, 60).

Wolfram, Stephen (2002). *A New Kind of Science*. Wolfram Media, Inc. (cited on page 205).

Zhu, Xingquan and Ian Davidson (2007). *Knowledge Discovery and Data Mining: Challenges and Realities*. 1st edition. Gale Group (cited on page 55).

Zienkiewicz, O C, R L Taylor, and J Z Zhu (2005). *The Finite Element Method: Its Basis and Fundamentals*. Elsevier Butterworth-Heinemann (cited on pages 153, 162).

Articles

Adler, Esther H, Jaya L Sunkara, Arthur S Patchefsky, Leopold G Koss, and Maja H Oktay (2012). "Predictors of disease progression in ductal carcinoma in situ of the breast and vascular patterns." In: *Human pathology* 43.4, pages 550–556 (cited on page 66).

Agner, Shannon C, Salil Soman, Edward Libfeld, Margie McDonald, Kathleen Thomas, Sarah Englander, Mark A Rosen, Deanna Chin, John Nosher, and Anant Madabhushi (2011). "Textural kinetics: a novel dynamic contrast-enhanced (DCE)-MRI feature for breast lesion classification." In: *Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology* 24.3, pages 446–463 (cited on pages 86, 97, 100).

Aizerman, M, E Braverman, and L Rozonoer (1964). "Theoretical foundations of the potential function method in pattern recognition learning". In: *Automation and Remote Control* 25, pages 821–837 (cited on pages 46, 47).

Akaike, H (1974). "A new look at the statistical model identification". In: *Automatic Control, IEEE Transactions on* 19.6, pages 716–723 (cited on page 95).

Albert, J M, D D Liu, Y Shen, I W Pan, Y C T Shih, K E Hoffman, T A Buchholz, S H Giordano, and B D Smith (2012). "Nomogram to Predict the Benefit of Radiation for Older Patients With Breast Cancer Treated With Conservative Surgery". In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 30.23, pages 2837–2843 (cited on page 63).

Alderliesten, Tanja, Claudette Loo, Anita Paape, Sara Muller, Emiel Rutgers, Marie-Jeanne Vrancken Peeters, and Kenneth Gilhuijs (2010). "On the feasibility of MRI-guided navigation to demarcate breast cancer for breast-conserving surgery". In: *Medical physics* 37.6, pages 2617–2626 (cited on page 176).

Amato, Filippo, Alberto López, Eladia Mar´a Peña-Méndez, Petr Va v n hara, Ale v s Hampl, and Josef Havel (2013). "Artificial neural networks in medical diagnosis". In: *Journal of Applied Biomedicine* 11.2, pages 47–58 (cited on page 44).

Ames, Victoria and Peter D Britton (2011). "Stereotactically guided breast biopsy: a review." In: *Insights into imaging* 2.2, pages 171–176 (cited on page 186).

Apesteguia, Luis and Luis Javier Pina (2011). "Ultrasound-guided core-needle biopsy of breast lesions." In: *Insights into imaging* 2.4, pages 493–500 (cited on page 123).

Arendt, Lisa M, Jenny A Rudnick, Patricia J Keller, and Charlotte Kuperwasser (2010). "Stroma in breast development and disease." In: *Seminars in cell & developmental biology* 21.1, pages 11–18 (cited on page 9).

ASCO – American Society of Clinical Oncology (2014). "The state of cancer care in america, 2014: a report by the american society of clinical oncology." In: *Journal of Oncology Practice* 10.2, pages 119–142 (cited on page 36).

Azar, Fred S, Dimitris N Metaxas, and Mitchell D Schnall (2002). "Methods for modeling and predicting mechanical deformations of the breast under external perturbations." In: *Medical image analysis* 6.1, pages 1–27 (cited on page 162).

Bailar, J C and E M Smith (1986). "Progress against cancer?" In: *The New England journal of medicine* 314.19, pages 1226–1232 (cited on page 24).

Bailar, John C and Heather L Gornik (1997). "Cancer Undefeated". In: *The New England journal of medicine* 336.22, pages 1569–1574 (cited on page 24).

Baker, R, K D Rogers, N Shepherd, and N Stone (2010). "New relationships between breast microcalcifications and cancer." In: *British journal of cancer* 103.7, pages 1034–1039 (cited on page 26).

Baldi, P, S Brunak, Y Chauvin, C A Andersen, and H Nielsen (2000). "Assessing the accuracy of prediction algorithms for classification: an overview." In: *Bioinformatics (Oxford, England)* 16.5, pages 412–424 (cited on page 55).

Baltzer, Pascal A T and Matthias Dietzel (2013). "Breast Lesions: Diagnosis by Using Proton MR Spectroscopy at 1.5 and 3.0T – Systematic Review and Meta-Analysis." In: *Radiology* 267.3, pages 735–46 (cited on page 29).

Barchielli, Alessandro A, Massimo M Federico, Vincenzo V De Lisi, Lauro L Bucchi, Stefano S Ferretti, Eugenio E Paci, Antonio A Ponti, and Eva E Buiatti (2005). "In situ breast cancer: Incidence trend and organised screening programmes in Italy". In: *European journal of cancer (Oxford, England : 1990)* 41.7, pages 6–6 (cited on page 40).

Basanta, David, Jacob G Scott, Mayer N Fishman, Gustavo E Ayala, Simon W Hayward, and Alexander RA Anderson (2011). "Investigating prostate cancer tumour-stroma interactions - clinical and biological insights from an evolutionary game". In: *arXiv.org*. arXiv: `1108.0654v1 [q-bio.PE]` (cited on page 9).

Beard, David V, Peter Bream, Etta D Pisano, Pat Conroy, R Eugene Johnston, Patricia Breauning, Robert McLelland, and Richard Clark (1997). "A pilot study of eye movement during mammography interpretation: Eyetracker results and workstation design implications". In: *Journal of Digital Imaging* 10.1, page 14 (cited on pages 139, 140).

Becker, Nikolaus and Hans Junkermann (2008). "Nutzen und Risiko des Mammografie-Screenings". In: *Dtsch Arztebl* 105.8, pages 131–136 (cited on page 138).

Beckmann, Matthias W et al. (2009). "Zurich Consensus: German Expert Opinion on the St. Gallen Votes on 15 March 2009 (11th International Conference at St. Gallen: Primary Therapy of Early Breast Cancer)." In: *Breast care (Basel, Switzerland)* 4.2, pages 109–116 (cited on page 19).

Behrens, Sarah, Hendrik Laue, Matthias Althaus, Tobias Boehler, Bernd Kuemmerlen, Horst K Hahn, and Heinz-Otto Peitgen (2007). "Computer assistance for MR based diagnosis of breast cancer: present and future challenges". In: *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 31.4-5, pages 236–247 (cited on pages 93, 100).

Berg, Wendie A et al. (2012). "Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk." In: *JAMA : the journal of the American Medical Association* 307.13, pages 1394–1404 (cited on page 121).

Berns, Eric A, R Edward Hendrick, Mariana Solari, Lora Barke, Denise Reddy, Judith Wolfman, Lewis Segal, Patricia DeLeon, Stefanie Benjamin, and Laura Willis (2006). "Digital and screen-film mammography: comparison of image acquisition and interpretation times." In: *AJR American journal of roentgenology* 187.1, pages 38–41 (cited on page 140).

Bhooshan, Neha, Maryellen L Giger, Sanaz A Jansen, Hui Li, Li Lan, and Gillian M Newstead (2010). "Cancerous breast lesions on dynamic contrast-enhanced MR images: computerized characterization for image-based prognostic markers." In: *Radiology* 254.3, pages 680–690 (cited on pages 70, 80).

Blumenfeld, Raphael and Sam F Edwards (2011). "Theory of strains in auxetic materials". In: *Audio, Transactions of the IRE Professional Group on*, pages (cited on page 156).

Boehler, Tobias, Fabian Zoehrer, Markus Harz, and Horst Karl Hahn (2011). "Breast image registration and deformation modeling." In: *Critical Reviews in Biomedical Engineering* 40.3, pages 235–258 (cited on page 30).

Boetes, Carla and Ritse M Mann (2007). "Ductal carcinoma in situ and breast MRI". In: *Lancet* 370.9586, pages 459–460 (cited on pages 68, 69).

Bombonati, Alessandro and Dennis C Sgroi (2010). "The molecular pathology of breast cancer progression". In: *The Journal of pathology* 223.2, pages 308–318 (cited on page 9).

Boroczky, Lilla, Mark Simpson, Hiroyuki Abe, and Jeremy Drysdale (2013). "Observer study of a prototype clinical decision support system for breast cancer diagnosis using dynamic contrast-enhanced MRI." In: *AJR American journal of roentgenology* 200.2, pages 277–283 (cited on page 108).

Boulesteix, Anne-Laure and Carolin Strobl (2008). "Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction." In: *BMC Medical Research Methodology* 9, pages 85–85 (cited on pages 59, 81).

Brake, G M te, N Karssemeijer, and J H Hendriks (2000). "An automatic method to discriminate malignant masses from normal tissue in digital mammograms." In: *Physics in Medicine and Biology* 45.10, pages 2843–2857 (cited on page 44).

Breiman, Leo (1996). "Bagging Predictors". In: *Machine Learning* 24.2, pages 123–140 (cited on page 48).

— (2001). "Random Forests". In: *Machine Learning* 45, pages 5–32 (cited on pages 48, 63, 80, 101).

Brennan, S B, J S Sung, D D Dershaw, L Liberman, and E A Morris (2011). "Cancellation of MR Imaging-guided Breast Biopsy Due to Lesion Nonvisualization: Frequency and Follow-Up". In: *Radiology* 261.1, pages 92–99 (cited on pages 151, 188).

Breslow, L and W G Cumberland (1988). "Progress and objectives in cancer control." In: *JAMA : the journal of the American Medical Association* 259.11, pages 1690–1694 (cited on page 24).

Brix, G, W Semmler, R Port, L R Schad, G Layer Lorenz, and W J (1991). "Pharmacokinetic parameters in CNS Gd-DTPA enhanced MR imaging." In: *J Comput Assist Tomogr* 15.4, pages 621–628 (cited on page 32).

Bruening, Wendy, Joann Fontanarosa, Kelley Tipton, Jonathan R Treadwell, Jason Launders, and Karen Schoelles (2010). "Systematic review: comparative effectiveness of core-needle and open surgical biopsy to diagnose breast lesions." In: *Annals of internal medicine* 152.4, pages 238–246 (cited on page 186).

Burges, Christopher J C (1998). "A Tutorial on Support Vector Machines for Pattern Recognition - Springer". In: *Data Mining and Knowledge Discovery* 2.2, pages 121–167 (cited on pages 44, 46).

Carter, T, C Tanner, W Crum, N Beechey-Newman, and DJ Hawkes (2006). "A framework for image-guided breast surgery". In: *Lecture Notes in Computer Science* (cited on page 176).

Chapelle, Olivier, Vladimir Vapnik, Olivier Bousquet, and Sayan Mukherjee (2002). "Choosing Multiple Parameters for Support Vector Machines". In: *Machine Learning* 46.1-3 (cited on page 48).

Chen, Mingli, Weiguo Lu, Quan Chen, Kenneth J Ruchala, and Gustavo H Olivera (2008). "A simple fixed-point approach to invert a deformation field." In: *Medical Physics* 35.1, pages 81–88 (cited on page 192).

Chen, Weijie, Maryellen L Giger, Hui Li, Ulrich Bick, and Gillian M Newstead (2007). "Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images". In: *Magnetic Resonance in Medicine* 58.3, pages 562–571 (cited on pages 85, 87, 89, 92, 100).

Choi, Jae Young and Yong Man Ro (2012). "Multiresolution local binary pattern texture analysis combined with variable selection for application to false-positive reduction in computer-aided detection of breast masses on mammograms." In: *Physics in Medicine and Biology* 57.21, pages 7029–7052 (cited on page 106).

Coopey, Suzanne, Barbara L Smith, Stephanie Hanson, Julliette Buckley, Kevin S Hughes, Michele Gadd, and Michelle C Specht (2011). "The Safety of Multiple Re-excisions after Lumpectomy for Breast Cancer." In: *Annals of Surgical Oncology* (cited on page 176).

Cortes, Corinna and Vladimir Vapnik (1995). "Support-Vector Networks". In: *Machine Learning* 20.3 (cited on page 44).

Costa, Ivan F (2012). "A novel deformation method for fast simulation of biological tissue formed by fibers and fluid." In: *Medical Image Analysis* 16.5, pages 1038–1046 (cited on page 200).

Costello, Leslie C and Renty B Franklin (2005). "'Why do tumour cells glycolyse?': from glycolysis through citrate to lipogenesis". In: *Molecular and cellular biochemistry* 280.1–2, pages 1–8 (cited on page 10).

De Paepe, Lode, Peter De Bock, Olivier Vanovermeire, and Tom Kimpe (2012). "Performance evaluation of a visual display calibration algorithm for IPad". In: *Proc. SPIE* 8319 (cited on page 113).

Degani, H, V Gusis, D Weinstein, S Fields, and S Strano (1997). "Mapping pathophysiological features of breast tumors by MRI at high spatial resolution." In: *Nature medicine* 3.7, pages 780–782 (cited on page 93).

Demšar, Janez, Tomaž Curk, et al. (2013). "Orange: Data Mining Toolbox in Python". In: *Journal of Machine Learning Research* 14, pages 2349–2353. URL: http://jmlr.org/papers/v14/demsar13a.html (cited on pages 60, 63, 80).

Destounis, Stamatia (2006). "The Role of MRI and "Second-Look" Ultrasound for Evaluation of Breast Cancer". In: *Appl Radiol* 35.10, pages 10–20 (cited on pages 118, 121).

Di Cataldo, Santa, Elisa Ficarra, and Enrico Macii (2012). "Computer-aided techniques for chromogenic immunohistochemistry: Status and directions". In: *Computers in Biology and Medicine* 42.10 (cited on page 13).

Dick, Christian, Joachim Georgii, Rainer Burgkart, and Rüdiger Westermann (2009). "Stress Tensor Field Visualization for Implant Planning in Orthopedics". In: *IEEE Transactions on Visualization and Computer Graphics (Proceedings of IEEE Visualization 2009)* 15.6, pages 1399–1406 (cited on page 167).

Doerfler, Ron (2009). "On Jargon". In: *The UMAP Journal* 30.4, pages 457–493 (cited on page 63).

Dogan, Firat and M Serdar Celebi (2011). "Real-time deformation simulation of non-linear viscoelastic soft tissues". In: *Simulation* 87.3 (cited on page 201).

Domingos, Pedro (2012). "A few useful things to know about machine learning". In: *Communications of the ACM* 55.10 (cited on page 41).

Dorrius, Monique D, Marijke C Jansen-van Weide, Peter M A Ooijen, Ruud M Pijnappel, and Matthijs Oudkerk (2011). "Computer-aided detection in breast MRI: a systematic review and meta-analysis". In: *European Radiology* 21.8, pages 1600–1608 (cited on pages 32, 120).

Efron, Bradley and Gail Gong (1983). "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation". In: *The American Statistician* 37.1, pages 36–48. DOI: 10.1080/00031305.1983.10483087. eprint: http://www.tandfonline.com/doi/pdf/10.1080/00031305.1983.10483087. URL: http://www.tandfonline.com/doi/abs/10.1080/00031305.1983.10483087 (cited on page 58).

Eguchi, Takashi, Keiichiro Takasuna, Atsushi Kitazawa, Youhei Fukuzawa, Yasuo Sakaue, Kazuo Yoshida, and Makoto Matsubara (2012). "Three-dimensional imaging navigation during a lung segmentectomy using an iPad." In: *European Journal of Cardio-Thoracic Surgery* 41.4, pages 893–897 (cited on page 114).

Elston, C W et al. (2000). "Causes of inconsistency in diagnosing and classifying intraductal proliferations of the breast. European Commission Working Group on Breast Screening Pathology". In: *European Journal of Cancer* 36.14, pages 1769–1772 (cited on page 69).

Esserman, Laura J et al. (2006). "Magnetic resonance imaging captures the biology of ductal carcinoma in situ." In: *Journal of Clinical Oncology* 24.28, pages 4603–4610 (cited on page 68).

Facius, Mirjam, Diane M Renz, Henning Neubauer, Joachim Boettcher, Mieczyslaw Gajda, Oumar Camara, and Werner A Kaiser (2007). "Characteristics of ductal carcinoma in situ in magnetic resonance imaging". In: *Clinical Imaging* 31.6, pages 394–400 (cited on page 66).

Finak, Greg et al. (2008). "Stromal gene expression predicts clinical outcome in breast cancer." In: *Nature Medicine* 14.5, pages 518–527 (cited on page 9).

Fischler, Martin A and Robert C Bolles (1981). "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24.6 (cited on page 190).

Fisher, R A (1936). "The Use of Multiple Measurements in Taxonomic Problems". In: *Annals of Eugenics* 7.2, pages 179–188. issn: 2050-1439. doi: 10.1111/j.1469-1809.1936.tb02137.x. url: http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x (cited on page 47).

Friedman, Jerome H (1991). "Multivariate Adaptive Regression Splines". In: *The Annals of Statistics* 19.1, pages 1–67 (cited on page 101).

Frohlich, H, O Chapelle, and B Scholkopf (2003). "Feature selection for support vector machines by means of genetic algorithm". In: *IEEE International Conference on Tools with Artificial Intelligence. Proceedings*, pages 142–148 (cited on page 48).

Fu, Wenjiang J, Raymond J Carroll, and Suojin Wang (2005). "Estimating misclassification error with small samples via bootstrap cross-validation." In: *Bioinformatics (Oxford, England)* 21.9, pages 1979–1986 (cited on page 58).

Fujita, H et al. (2008). "Computer-aided diagnosis: The emerging of three CAD systems induced by Japanese health care needs". In: *Computer Methods and Programs in Biomedicine* (cited on page 17).

Gatenby, Robert A and Robert J Gillies (2004). "Why do cancers have high aerobic glycolysis?" In: *Nature Reviews Cancer* 4.11, pages 891–899 (cited on page 10).

Georgii, Joachim, Maximilian Eder, Kai Bürger, Sebastian Klotz, Florian Ferstl, Laszlo Kovacs, and Rüdiger Westermann (2013). "A Computational Tool for Preoperative Breast Augmentation Planning in Aesthetic Plastic Surgery". In: *IEEE J Biomed Health Inform* Epub ahead of print. (Cited on pages 122, 201).

Georgii, Joachim and Rüdiger Westermann (2006). "A Multigrid Framework for Real-Time Simulation of Deformable Bodies". In: *Computer & Graphics* 30, pages 408–415 (cited on page 162).

— (2010). "A Streaming Approach for Sparse Matrix Products and its Application in Galerkin Multigrid Methods". In: *Electronic Transactions on Numerical Analysis* 37, pages 263–275 (cited on page 163).

Gibbs, Peter and Lindsay W Turnbull (2003). "Textural analysis of contrast-enhanced MR images of the breast". In: *Magnetic Resonance in Medicine* 50.1, pages 92–98 (cited on page 85).

Giger, Maryellen L, Nico Karssemeijer, and Julia A Schnabel (2013). "Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer." In: *Annual review of biomedical engineering* 15, pages 327–357 (cited on pages 18, 34, 43).

Giordano, Livia L et al. (2011). "Mammographic screening programmes in Europe: organization, coverage and participation." In: *Journal of Medical Screening* 19 Suppl 1, pages 72–82 (cited on page 16).

Goldhirsch, A, E P Winer, A S Coates, R D Gelber, M Piccart-Gebhart, B Thürlimann, H-J Senn, and Panel members (2013). "Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013." In: *Annals of Oncology* 24.9, pages 2206–2223 (cited on page 19).

Gøtzsche, Peter C and Karsten Juhl Jørgensen (2013). "Screening for breast cancer with mammography." In: *The Cochrane Database of Systematic Reviews* 6, page CD001877 (cited on page 138).

Gruber, Stephan, Bogumil-Krystian Debski, Katja Pinker, Marek Chmelik, Guenther Grabner, Thomas Helbich, Siegfried Trattnig, and Wolfgang Bogner (2011). "Three-dimensional Proton MR Spectroscopic Imaging at 3 T for the Differentiation of Benign and Malignant Breast Lesions". In: *Radiology* 261.3, pages 752–761 (cited on page 29).

Han, Lianghao, John H Hipwell, Christine Tanner, Zeike Taylor, Thomy Mertzanidou, Jorge Cardoso, Sebastien Ourselin, and David J Hawkes (2012). "Development of patient-specific biomechanical models for predicting large breast deformation." In: *Physics in Medicine and Biology* 57.2, pages 455–472 (cited on pages 162, 201).

Han, Lianghao, John H Hipwell, Zeike Taylor, Christine Tanner, S Ourselin, and David J Hawkes (2010). "Fast Deformation Simulation of Breasts Using GPU-Based Dynamic Explicit Finite Element Method". In: *Proceedings of the International Workshop in Digital Mammography (IWDM 2010)* 6136, pages 728–735 (cited on page 184).

Han, Lianghao, John Hipwell, Thomy Mertzanidou, Tim Carter, Marc Modat, Sebastien Ourselin, and David Hawkes (2011). "A hybrid FEM-based method for aligning prone and supine images for image guided breast surgery". In: *ISBI 2011*, pages 1239–1242 (cited on pages 162, 175, 179).

Hanahan, Douglas and Lisa M Coussens (2012). "Accessories to the crime: functions of cells recruited to the tumor microenvironment." In: *Cancer cell* 21.3, pages 309–322 (cited on pages 9, 13).

Hanahan, Douglas and Robert A Weinberg (2000). "The Hallmarks of Cancer". In: *Cell* 100.1, pages 14–14 (cited on pages 9, 10).

— (2011). "Hallmarks of Cancer: The Next Generation". In: *Cell* 144.5, pages 29–29 (cited on pages 8–10).

Haralick, Robert M, K Shanmugam, and Its'hak Dinstein (1973). "Textural Features For Image Classification". In: *IEEE Transactions on Systems, Man, and Cybernetics* 3.6, pages 610–621 (cited on pages 90, 92).

Harz, Markus, Joachim Georgii, Lei Wang, Kathy Schilling, and Heinz-Otto Peitgen (2012b). "Efficient Breast Deformation Simulation". In: *Proceedings of the 9th Workshop On Virtual Reality Interaction and Physical Simulation*, pages 117–126 (cited on page 201).

Haygood, Tamara Miner, Jihong Wang, Deanna Lane, Eva Galvan, E Neely Atkinson, Tanya Stephens, and Gary J Whitman (2010). "Why Does It Take Longer to Read Digital Than Film-Screen Screening Mammograms? A Partial Explanation". In: *Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology* 23.2, page 170 (cited on page 140).

Heeten, Gerard J den and M J M Broeders (2009). "Nationwide breast cancer screening in the Netherlands". In: *MedicaMundi* 53.1, pages 35–39 (cited on page 137).

Heywang-Köbrunner, Sylvia H, Ingrid Schreer, and Walter Heindel (2008). "Bildgebung für die Brustkrebsfrüherkennung". In: *Dtsch Arztebl* 105.31–32, pages 541–547 (cited on pages 100, 138).

Hieken, T J, M Farolan, S D'Alessandro, and J M Velasco (2001). "Predicting the biologic behavior of ductal carcinoma in situ: An analysis of molecular markers". In: *Surgery* 130.4, pages 9–9 (cited on page 66).

Hinton, Geoffrey E, Simon Osindero, and Yee-Whye Teh (2006). "A fast learning algorithm for deep belief nets." In: *Neural Computation* 18.7, pages 1527–1554 (cited on page 44).

Hoffmann, Sebastian, Jamie D Shutler, Marc Lobbes, Bernhard Burgeth, and Anke Meyer-Bäse (2013). "Automated analysis of non-mass-enhancing lesions in breast MRI based on morphological, kinetic, and spatio-temporal moments and joint segmentation-motion compensation technique". In: *EURASIP Journal on Advances in Signal Processing* 2013.1, page 172 (cited on page 71).

Holland, R, S H J Veling, M Mravunac, and J H C L Hendriks (1985). "Histologic multifocality of tis, T1–2 breast carcinomas implications for clinical trials of breast-conserving surgery - Holland - 2006 - Cancer - Wiley Online Library". In: *Cancer* (cited on pages 65, 66).

Hu, Min et al. (2008). "Regulation of in situ to invasive breast carcinoma transition". In: *Cancer cell* 13.5, pages 394–406 (cited on page 13).

Huang, C L and C J Wang (2006). "A GA-based feature selection and parameters optimization for support vector machines". In: *Expert Systems with Applications* 31.2, pages 10–10 (cited on page 48).

Hupse, Rianne, Maurice Samulski, Marc B Lobbes, Ritse M Mann, Roel Mus, Gerard J den Heeten, David Beijerinck, Ruud M Pijnappel, Carla Boetes, and Nico Karssemeijer (2013). "Computer-aided detection of masses at mammography: interactive decision support versus prompts." In: *Radiology* 266.1, pages 123–129 (cited on pages 17, 21, 108).

Hyodo, Fuminori, Gadisetti V R Chandramouli, Shingo Matsumoto, Ken-Ichiro Matsumoto, James B Mitchell, Murali C Krishna, and Jeeva P Munasinghe (2009). "Estimation of tumor microvessel density by MRI using a blood pool contrast agent." In: *International Journal of Oncology* 35.4, pages 797–804 (cited on page 65).

Ioannidis, John P A (2005). "Microarrays and molecular research: noise discovery?" In: *Lancet* 365.9458, pages 454–455 (cited on page 59).

IOM (Institute of Medicine) (2012). "Breast Cancer and the Environment: A life course approach". In: (cited on pages 8, 23).

Jansen, Sanaz A (2011). "Ductal carcinoma in situ: detection, diagnosis, and characterization with magnetic resonance imaging." In: *Seminars in ultrasound, CT, and MR* 32.4, pages 306–318 (cited on pages 68, 69).

Jansen, Sanaz A, Suzanne D Conzen, Xiaobing Fan, Erica J Markiewicz, Gillian M Newstead, and Gregory S Karczmar (2009). "Magnetic resonance imaging of the natural history of in situ mammary neoplasia in transgenic mice: a pilot study." In: *Breast cancer research : BCR* 11.5, R65 (cited on page 12).

Jansen, Sanaz A, Suzanne D Conzen, Xiaobing Fan, Erica Markiewicz, Thomas Krausz, Gillian M Newstead, and Gregory S Karczmar (2011). "In vivo MRI of early stage mammary cancers and the normal mouse mammary gland." In: *NMR in biomedicine* 24.7, pages 880–887 (cited on pages 14, 69).

Jansen, Sanaz A, Gillian M Newstead, Hiroyuki Abe, Akiko Shimauchi, Robert A Schmidt, and Gregory S Karczmar (2007). "Pure ductal carcinoma in situ: kinetic and morphologic MR characteristics compared with mammographic appearance and nuclear grade." In: *Radiology* 245.3, pages 684–691 (cited on page 68).

Jansen, Sanaz A, Tatjana Paunesku, Xiaobing Fan, Gayle E Woloschak, Stefan Vogt, Suzanne D Conzen, Thomas Krausz, Gillian M Newstead, and Gregory S Karczmar (2009). "Ductal carcinoma in situ: X-ray fluorescence microscopy and dynamic contrast-enhanced MR imaging reveals gadolinium uptake within neoplastic mammary ducts in a murine model." In: *Radiology* 253.2, pages 399–406 (cited on page 14).

Jansen, Sanaz A, Akiko Shimauchi, Lindsay Zak, Xiaobing Fan, Gregory S Karczmar, and Gillian M Newstead (2011). "The diverse pathology and kinetics of mass, nonmass, and focus enhancement on MR imaging of the breast." In: *Journal of Magnetic Resonance Imaging* 33.6, pages 1382–1389 (cited on pages 66, 71, 80).

Jensen, David D and Paul R Cohen (2000). "Multiple Comparisons in Induction Algorithms". English. In: *Machine Learning* 38.3, pages 309–338. issn: 0885-6125. doi: 10.1023/A:1007631014630. url: http://dx.doi.org/10.1023/A:1007631014630 (cited on page 52).

John, Sindhu, Angeline C C Poh, Tchoyoson C C Lim, Elizabeth H Y Chan, and Le Roy Chong (2012). "The iPad Tablet Computer for Mobile On-Call Radiology Diagnosis? Auditing Discrepancy in CT and MRI Reporting". English. In: *Journal of Digital Imaging* 25.5, pages 628–634. issn: 0897-1889. doi: 10.1007/s10278-012-9485-3. url: http://dx.doi.org/10.1007/s10278-012-9485-3 (cited on page 113).

Kell, Malcolm R and Monica Morrow (2005). "An adequate margin of excision in ductal carcinoma in situ." In: *BMJ (Clinical research ed)* 331.7520, pages 789–790 (cited on page 40).

Kelly, C M and K I Pritchard (2012). "Personalized Medicine: What Exactly Is It and Can We Truly Measure It?" In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 30.18, pages 2173–2174 (cited on page 36).

Kerlikowske, Karla, Rebecca A Hubbard, Diana L Miglioretti, Berta M Geller, Bonnie C Yankaskas, Constance D Lehman, Stephen H Taplin, Edward A Sickles, and Edward A Sickles (2011). "Comparative effectiveness of digital versus film-screen mammography in community practice in the United States: a cohort study." In: *Annals of Internal Medicine* 155.8, pages 493–502 (cited on page 16).

Khan, Abdul Arif, Abhinav Shrivastava, and Mohsin Khurshid (2012). "Normal to cancer microbiome transformation and its implication in cancer diagnosis." In: *Biochimica et biophysica acta* 1826.2, pages 331–337 (cited on page 13).

Kohavi, Ron and George H John (1997). "Wrappers for feature subset selection". In: *Artificial Intelligence* 97.1-2 (cited on page 51).

Kolmogorov, Andrey (1963). "On Tables of Random Numbers". In: *Sankhya Ser. A.* 25, pages 369–375 (cited on page 6).

Kopans, D B (1990). "The Canadian Screening Program: A Different Perspective : American Journal of Roentgenology: Vol. 155, No. 4 (AJR)". In: *American Journal of Roentgenology* (cited on page 138).

Kraemer, Helena Chmura, Vyjeyanthi S Periyakoil, and Art Noda (2002). "Kappa coefficients in medical research." In: *Statistics in Medicine* 21.14, pages 2109–2129 (cited on page 54).

Kriegel, Hans-Peter, Peer Kröger, and Arthur Zimek (2009). "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering". In: *ACM Trans Knowl Discov Data* 3.1 (cited on page 49).

Kuhl, Christiane (2007). "The current status of breast MR imaging. Part I. Choice of technique, image interpretation, diagnostic accuracy, and transfer to clinical practice". In: *Radiology* 244.2, pages 356–378 (cited on page 71).

Kuhl, Christiane K, Simone Schrading, Heribert B Bieling, Eva Wardelmann, Claudia C Leutner, Roy Koenig, Walther Kuhn, and Hans H Schild (2007). "MRI for diagnosis of pure ductal carcinoma in situ: a prospective observational study". In: *Lancet* 370.9586, pages 485–492 (cited on pages 68, 69).

Kupinski, M A and M L Giger (1999). "Feature selection with limited datasets." In: *Medical Physics* 26.10, pages 2176–2182 (cited on page 52).

Lakhani, S R (1999). "The transition from hyperplasia to invasive carcinoma of the breast." In: *The Journal of Pathology* 187.3, pages 272–278 (cited on page 11).

Landauer, Rolf (1988). "A simple measure of complexity". In: *Nature* 336.6197, pages 306–307 (cited on page 6).

Larsson, H B, M Stubgaard, J L Frederiksen, M Jensen, O Henriksen, and O B Paulson (1990). "Quantitation of blood-brain barrier defect by magnetic resonance imaging and gadolinium-DTPA in patients with multiple sclerosis and brain tumors." In: *Magnetic Resonance in Medicine* 16.1, pages 117–131 (cited on page 32).

Lasko, Thomas A, Jui G Bhagwat, Kelly H Zou, and Lucila Ohno-Machado (2005). "The use of receiver operating characteristic curves in biomedical informatics". In: *Journal of Biomedical Informatics* 38.5 (cited on page 56).

Lazebnik, Yuri (2010). "What are the hallmarks of cancer?" In: *Nature Reviews Cancer* 10.4, pages 232–233 (cited on page 10).

Lee, Carol H et al. (2010). "Breast Cancer Screening With Imaging: Recommendations From the Society of Breast Imaging and the ACR on the Use of Mammography, Breast MRI, Breast Ultrasound, and Other Technologies for the Detection of Clinically Occult Breast Cancer". In: *Journal of the American College of Radiology* 7.1, pages 18–27 (cited on pages 119, 140).

Lee, Sangjun, Sheila Stewart, Iris Nagtegaal, Jingqin Luo, Yun Wu, Graham Colditz, Dan Medina, and D Craig Allred (2012). "Differentially Expressed Genes Regulating the Progression of Ductal Carcinoma In Situ to Invasive Breast Cancer." In: *Cancer Res* (cited on page 13).

Lehman, Constance D (2012). "Clinical indications: what is the evidence?" In: *European Journal of Radiology* 81S1, S82–S84 (cited on page 17).

Lehman, Constance D, Jeffrey D Blume, Wendy B Demartini, Nola M Hylton, Benjamin Herman, and Mitchell D Schnall (2013). "Accuracy and interpretation time of computer-aided detection among novice and experienced breast MRI readers." In: *AJR American journal of roentgenology* 200.6, W683–9. doi: 10.2214/AJR.11.8394 (cited on page 18).

Leonard, Gregory D and Sandra M Swain (2004). "Ductal carcinoma in situ, complexities and challenges." In: *Journal of the National Cancer Institute* 96.12, pages 906–920 (cited on pages 11, 68).

Levental, Kandice R et al. (2009). "Matrix crosslinking forces tumor progression by enhancing integrin signaling." In: *Cell* 139.5, pages 891–906 (cited on page 9).

Li, Christopher I, Janet R Daling, and Kathleen E Malone (2005). "Age-specific incidence rates of in situ breast carcinomas by histologic type, 1980 to 2001." In: *Cancer Epidemiology, Biomarkers & Prevention* 14.4, pages 1008–1011 (cited on page 40).

Li, Qiang (2007). "Improvement of bias and generalizability for computer-aided diagnostic schemes". In: *Computerized Medical Imaging and Graphics* 31.4-5, pages 338–345 (cited on page 58).

Lloyd, S (2001). "Measures of complexity: a nonexhaustive list". In: *Control Systems, IEEE* 21.4, pages 7–8 (cited on pages 6, 206).

Lundstrom, C., T. Rydell, C. Forsell, A. Persson, and A. Ynnerman (2011). "Multi-Touch Table System for Medical Visualization: Application to Orthopedic Surgery Planning". In: *Visualization and Computer Graphics, IEEE Transactions on* 17.12, pages 1775–1784 (cited on page 116).

Ma, Xiao-Jun, Sonika Dahiya, Elizabeth Richardson, Mark Erlander, and Dennis C Sgroi (2009). "Gene expression profiling of the tumor microenvironment during breast cancer progression". In: *Breast cancer research : BCR* 11.1, R7 (cited on page 13).

Macura, Katarzyna J, Ronald Ouwerkerk, Michael A Jacobs, and David A Bluemke (2006). "Patterns of enhancement on breast MR images: interpretation and imaging pitfalls". In: *Radiographics : a review publication of the Radiological Society of North America, Inc* 26.6, page 1719 (cited on page 65).

Madabhushi, Anant, Shannon Agner, Ajay Basavanhally, Scott Doyle, and George Lee (2011). "Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data." In: *Computerized Medical Imaging and Graphics* 35.7-8, pages 506–514 (cited on page 121).

Mallapragada, V, N Sarkar, and T. K Podder (2011). "Toward a Robot-Assisted Breast Intervention System". In: *Mechatronics, IEEE/ASME Transactions on* 16.6, pages 1011–1020 (cited on pages 123, 176).

Mandelblatt, Jeanne, Nicolien van Ravesteyn, Clyde Schechter, Yaojen Chang, An-Tsun Huang, Aimee M Near, Harry de Koning, and Ahmedin Jemal (2013). "Which strategies reduce breast cancer mortality most?: Collaborative Modeling of Optimal Screening, Treatment, and Obesity Prevention." In: *Cancer* 119.14, pages 2541–2548 (cited on page 1).

Marasco, J, R Doerfler, and L Roschier (2011). "Doc, what are my chances". In: *The UMAP Journal* 32.4, pages 279–298 (cited on pages 62, 63).

Marceglia, S, S Bonacina, V Zaccaria, C Pagliari, and F Pinciroli (2012). "How might the iPad change healthcare?" In: *JRSM* 105.6, pages 233–241 (cited on page 114).

Martincich, Laura et al. (2011). "Multicenter, double-blind, randomized, intraindividual crossover comparison of gadobenate dimeglumine and gadopentetate dimeglumine for Breast MR imaging (DETECT Trial)." In: *Radiology* 258.2, pages 396–408 (cited on page 30).

Mc Laughlin, Patrick, Siobhan O Neill, Noel Fanning, Anne Marie Mc Garrigle, Owen J O Connor, Gerry Wyse, and Michael M Maher (2012). "Emergency CT brain: preliminary interpretation with a tablet device: image quality and diagnostic performance of the Apple iPad". English. In: *Emergency Radiology* 19.2, pages 127–133. ISSN: 1070-3004. DOI: 10.1007/s10140-011-1011-2. URL: http://dx.doi.org/10.1007/s10140-011-1011-2 (cited on page 113).

McCulloch, Warren and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". In: *Bulletin of mathematical biology* 5.4, pages 115–133 (cited on page 43).

McNulty, Jonathan P, John T Ryan, Michael G Evanoff, and Louise A Rainford (2012). "Flexible image evaluation: iPad versus secondary-class monitors for review of MR spinal emergency cases, a comparative study." In: *Academic Radiology* 19.8, pages 1023–1028 (cited on page 113).

Medved, Milica M, Marko K MK Ivancevic, Olufunmilayo I OI Olopade, Gillian M GM Newstead, and Gregory S GS Karczmar (2010). "Echo-planar spectroscopic imaging (EPSI) of the water resonance structure in human breast using sensitivity encoding (SENSE)." In: *Magnetic Resonance in Medicine* 63.6, pages 1557–1563 (cited on page 33).

Medved, Milica, Gregory S Karczmar, and Gillian M Newstead (2012). "Do we really need contrast agents?" In: *European Journal of Radiology* 81 Suppl 1, S99–100 (cited on page 33).

Medved, Milica, Gillian M Newstead, Hiroyuki Abe, Olufunmilayo I Olopade, Akiko Shimauchi, Marta A Zamora, and Gregory S Karczmar (2010). "Clinical implementation of a multislice high spectral and spatial resolution-based MRI sequence to achieve unilateral full-breast coverage." In: *Journal of Magnetic Resonance Imaging* 28.1, pages 16–21 (cited on page 33).

Mertzanidou, Thomy, John Hipwell, M Jorge Cardoso, Xiying Zhang, Christine Tanner, Sebastien Ourselin, Ulrich Bick, Henkjan Huisman, Nico Karssemeijer, and David Hawkes (2012). "MRI to X-ray mammography registration using a volume-preserving affine transformation." In: *Medical Image Analysis* 16.5, pages 966–975 (cited on page 200).

Metz, Charles E (1978). "Basic Principles of ROC Analysis". In: *Seminars in Nuclear Medicine* 8.4, pages 283–298 (cited on page 55).

Meyer, David, Friedrich Leisch, and Kurt Hornik (2003). "The support vector machine under test". In: *Neurocomputing* 55.1-2, pages 18–18 (cited on page 48).

Meyerson, Anna F, Juan N Lessing, Kaoru Itakura, Nola M Hylton, Dulcy E Wolverton, Bonnie N Joe, Laura J Esserman, and E Shelley Hwang (2011). "Outcome of long term active surveillance for estrogen receptor-positive ductal carcinoma in situ." In: *Breast* 20.6, pages 529–533 (cited on page 12).

Michiels, Stefan, Serge Koscielny, and Catherine Hill (2005). "Prediction of cancer outcome with microarrays: a multiple random validation strategy". In: *Lancet* 365.9458, pages 5–5 (cited on page 59).

Miller, Anthony B, Claus Wall, Cornelia J Baines, Ping Sun, Teresa To, and Steven A Narod (2014). "Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial." In: *BMJ (Clinical research ed.)* 348, g366 (cited on page 138).

Moltz, J H et al. (2009). "Advanced Segmentation Techniques for Lung Nodules, Liver Metastases, and Enlarged Lymph Nodes in CT Scans". In: *Selected Topics in Signal Processing, IEEE Journal of* 3.1, pages 122–134 (cited on pages 73, 120, 130).

Mossa-Basha, Mahmud, Gina M Fundaro, Biren A Shah, Sharif Ali, and Milan V Pantelic (2010). "Ductal carcinoma in situ of the breast: MR imaging findings with histopathologic correlation." In: *Radiographics : a review publication of the Radiological Society of North America, Inc* 30.6, pages 1673–1687 (cited on page 66).

Mountford, Carolyn E, Christian Schuster, Pascal A T Baltzer, Peter Malycha, and Werner A Kaiser (2012). "MR spectroscopy in the breast clinic is improving." In: *European Journal of Radiology* 81 Suppl 1, S104–6 (cited on page 29).

Myhre, Simen, Hayat Mohammed, Trine Tramm, Jan Alsner, Greg Finak, Morag Park, Jens Overgaard, Anne-Lise Børresen-Dale, Arnoldo Frigessi, and Therese Sørlie (2010). "In silico ascription of gene expression differences to tumor and stromal cells in a model to study impact on breast cancer outcome." In: *PloS one* 5.11, e14002– (cited on page 9).

Nabavizadeh, Nima et al. (2011). "Topographic enhancement mapping of the cancer-associated breast stroma using breast MRI." In: *Audio and Electroacoustics Newsletter, IEEE* 3.4, pages 490–496 (cited on page 65).

Nanni, Loris, Sheryl Brahnam, and Alessandra Lumini (2012). "A simple method for improving local binary patterns by considering non-uniform patterns". In: *Pattern Recognition* 45.10 (cited on page 106).

Nealen, Andrew, Matthias Müller, Richard Keiser, Eddy Boxerman, and Mark Carlson (2006). "Physically Based Deformable Models in Computer Graphics". In: *Computer Graphics Forum* 25.4, pages 809–836 (cited on page 162).

Newell, Dustin, Ke Nie, Jeon Hor Chen, Chieh-Chih Hsu, Hon J Yu, Orhan Nalcioglu, and Min Ying Su (2009). "Selection of diagnostic features on breast MRI to differentiate between malignant and benign lesions using computer-aided diagnosis: differences in lesions presenting as mass and non-mass-like enhancement". In: *European Radiology* 20.4, pages 771–781 (cited on pages 68, 71, 86).

Newstead, Gillian M (2010). "MR imaging of ductal carcinoma in situ." In: *Magnetic Resonance Imaging Clinics of North America* 18.2, pages (cited on page 68).

Ng, Mei Rosa and Joan S Brugge (2009). "A stiff blow from the stroma: collagen crosslinking drives tumor progression." In: *Cancer Cell* 16.6, pages 455–457 (cited on page 9).

Nishikawa, Robert M, Robert A Schmidt, Michael N Linver, Alexandra V Edwards, John Papaioannou, and Margaret A Stull (2012). "Clinically missed cancer: how effectively can radiologists use computer-aided detection?" In: *AJR American Journal of Roentgenology* 198.3, pages 708–716 (cited on page 108).

O'Flynn, E A M, A R M Wilson, and M J Michell (2010). "Image-guided breast biopsy: state-of-the-art." In: *Clinical Radiology* 65.4, pages 259–270 (cited on pages 123, 186).

Orel, S G and M D Schnall (2001). "MR imaging of the breast for the detection, diagnosis, and staging of breast cancer". In: *Radiology* 220.1, pages 13–30 (cited on pages 69, 71).

Orguc, Sebnem, Isil Basara, Teoman Coskun, and Gokhan Pekindil (2012). "Three-dimensional vascular mapping of the breast by using contrast-enhanced MRI: association of unilateral increased vascularity with ipsilateral breast cancer." In: *Diagnostic and Interventional Radiology* 18.5, pages 454–9 (cited on page 71).

Ounpraseuth, Songthip, Shelly Y Lensing, Horace J Spencer, and Ralph L Kodell (2012). "Estimating misclassification error: a closer look at cross-validation based methods". In: *BMC Research Notes* 5.1, page 656 (cited on page 58).

Palomar, A Pérez del, B Calvo, J Herrero, J López, and M Doblaré (2008). "A finite element model to accurately predict real deformations of the breast." In: *Medical Engineering & Physics* 30.9, pages 1089–1097 (cited on pages 176, 201).

Papadopoulos, Athanasios, Dimitrios I. Fotiadis, and Aristidis Likas (2002). "An automatic microcalcification detection system based on a hybrid neural network classifier". In: *Artificial intelligence in Medicine* 25.2, pages 149–167 (cited on page 44).

Patani, Neill, Bruno Cutuli, and Kefah Mokbel (2008). "Current management of DCIS: a review". In: *Breast Cancer Research and Treatment* 111.1, pages 1–10 (cited on page 11).

Pathmanathan, Pras, David J Gavaghan, Jonathan P Whiteley, S Jonathan Chapman, and J Michael Brady (2008). "Predicting tumor location by modeling the deformation of the breast." In: *IEEE Transactions on Bio-Medical Engineering* 55.10, pages 2471–2480 (cited on page 201).

Perou, C M et al. (2000). "Molecular portraits of human breast tumours." In: *Nature* 406.6797, pages 747–752 (cited on page 9).

Pinder, Sarah E (2010). "Ductal carcinoma in situ (DCIS): pathological features, differential diagnosis, prognostic factors and specimen evaluation." In: *Modern Pathology* 23 Suppl 2, S8–13 (cited on pages 13, 66).

Pisano, Etta D et al. (2002). "Interpretation of digital mammograms: comparison of speed and accuracy of soft-copy versus printed-film display." In: *Radiology* 223.2, pages 483–488 (cited on page 139).

Polyak, Kornelia K and Raghu R Kalluri (2010). "The role of the microenvironment in mammary gland development and cancer." In: *Cold Spring Harbor Perspectives in Biology* 2.11, a003244–a003244 (cited on page 9).

Proia, D A and C Kuperwasser (2005). "Stroma - Tumor agonist or antagonist". In: *Cell Cycle* 4.8, pages 1022–1025 (cited on page 9).

Rajagopal, V, P M F Nielsen, and M P Nash (2004). "Development of a three-dimensional finite element model of breast mechanics". In: *Proc IEEE Engineering in Medicine and Biology* 7, pages 5080–5083 (cited on page 184).

Rajagopal, Vijayaraghavan, Jae-Hoon Chung, David Bullivant, Poul M F Nielsen, and Martyn P Nash (2007). "Determining the finite elasticity reference state from a loaded configuration". In: *International Journal for Numerical Methods in Engineering* 72.12, pages 1434–1451 (cited on pages 162, 185).

Ramadan, Saadallah, Pascal A T Baltzer, Alexander Lin, Peter Stanwell, Hayden Box, Werner A Kaiser, and Carolyn E Mountford (2012). "L-COSY of breast cancer at 3T." In: *European Journal of Radiology* 81 Suppl 1, S129–31 (cited on page 29).

Rankin, C C and F A Brogan (1986). "An element-independent co-rotational procedure for the treatment of large rotations". In: *ASME J. Pressure Vessel Tchn.* 108, pages 165–174 (cited on page 162).

Raza, Sughra, Monica Vallejo, Sona A Chikarmane, and Robyn L Birdwell (2008). "Pure ductal carcinoma in situ: a range of MRI features". In: *AJR American Journal of Roentgenology* 191.3, pages 689–699 (cited on pages 68–70).

Razavian, Ali Sharif, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson (2014). "CNN Features off-the-shelf: an Astounding Baseline for Recognition". In: *arXiv.org* (cited on page 44).

Reeves, Gillian K, Kirstin Pirie, Jane Green, Diana Bull, Valerie Beral, and Million Women Study Collaborators (2012). "Comparison of the effects of genetic

and environmental risk factors on in situ and invasive ductal breast cancer." In: *International Journal of Cancer* 131.4, pages 930–937 (cited on page 13).

Renz, Diane M et al. (2012). "Detection and classification of contrast-enhancing masses by a fully automatic computer-assisted diagnosis system for breast MRI." In: *Journal of Magnetic Resonance Imaging* 35.5, pages 1077–1088 (cited on page 120).

Richmond, Ann and Yingjun Su (2008). "Mouse xenograft models vs GEM models for human cancer therapeutics." In: *Disease Models & Mechanisms* 1.2-3, pages 78–82 (cited on page 14).

Ritter, Felix, Jumana Al Issawi, Markus Harz, Simon Benten, and Kathy Schilling (2013). "Combining Mobile Devices and Medical Workstations for Diagnostic Reading of Medical Images". In: *i-com Zeitschrift für interaktive und kooperative Medien* 12.1, pages 2–9. doi: 10.1524/icom.2013.0002 (cited on pages 110, 136).

Robinson, Jeffrey D (2012). "The skeptical technophile: iPad review." In: *Journal of Digital Imaging* 25.3, pages 365–368 (cited on pages 109, 113, 114).

Rosenblatt, Frank (1958). "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". In: *Psychological Review* 65.6, pages 386–408 (cited on page 43).

Ruiter, N V, R Stotzka, T O Muller, H Gemmeke, J R Reichenbach, and W A Kaiser (2006). "Model-based registration of X-ray mammograms and MR images of the female breast". In: *IEEE Transactions on Nuclear Science* 53.1, pages 204–211 (cited on page 177).

Sahiner, Berkman, Heang-Ping Chan, Nicholas Petrick, Datong Wei, Mark A Helvie, Dorit D Adler, and Mitchell M Goodsitt (1996). "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images". In: *IEEE Transactions on Medical Imaging* 15.5, pages 598–610 (cited on page 44).

Sardanelli, F, L Bacigalupo, L Carbonaro, A Esseridou, G Giuseppetti, P Panizza, V Lattanzio, and A Del Maschio (2008). "What is the sensitivity of mammography and dynamic MR imaging for DCIS if the whole-breast histopathology is used as a reference standard?" In: *La Radiologia medica* (cited on pages 12, 66, 68, 69).

Sardanelli, Francesco, Filippo Santoro, and Franca Podo (2012). "Screening high risk women with MRI alone?" In: *European Journal of Radiology* 81S1, S137–S138 (cited on page 17).

Sato, Y, M Nakamoto, Y Tamaki, T Sasama, I Sakita, Y Nakajima, M Monden, and S Tamura (1998). "Image guidance of breast cancer surgery using 3-D ultrasound images and augmented reality visualization". In: *IEEE Transactions on Medical Imaging* 17.5, pages 681–693 (cited on page 176).

Schoor, G van, S M Moss, J D M Otten, R Donders, E Paap, G J den Heeten, R Holland, M J M Broeders, and A L M Verbeek (2011). "Increasingly strong reduction in breast cancer mortality due to screening." In: *British Journal of Cancer* 104.6, pages 910–914 (cited on page 137).

Scribner, K C, F Behbod, and W W Porter (2012). "Regulation of DCIS to invasive breast cancer progression by Singleminded-2s (SIM2s)." In: *Oncogene* (cited on page 13).

Sears, Andrew and Ben Shneiderman (1991). "High precision touchscreens: design strategies and comparisons with a mouse". In: *International Journal of Man-Machine Studies* 34.4 (cited on page 111).

Sermanet, Pierre, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun (2013). "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks". In: *arXiv.org*. arXiv: `1312.6229v4 [cs.CV]` (cited on page 44).

Séroussi, B, C Laouénan, J Gligorov, S Uzan, F Mentré, and J Bouaud (2013). "Which breast cancer decisions remain non-compliant with guidelines despite the use of computerised decision support?" In: *British Journal of Cancer* 109.5, pages 1147–1156 (cited on page 121).

Sharma, M, A H Beck, J A Webster, I Espinosa, K Montgomery, S Varma, M Rijn, K C Jensen, and R B West (2009). "Analysis of stromal signatures in the tumor microenvironment of ductal carcinoma in situ". In: *Breast Cancer Research and Treatment* 123.2, pages 397–404 (cited on pages 9, 13).

Shih, Tzu-Ching, Jeon Hor Chen, Dongxu Liu, Ke Nie, Lizhi Sun, Muqing Lin, Daniel Chang, Orhan Nalcioglu, and Min Ying Su (2010). "Computational simulation of breast compression based on segmented breast and fibroglandular tissues on magnetic resonance images". In: *Physics in Medicine and Biology* 55.14, pages 4153–4168 (cited on page 201).

Specht, Jennifer M and David A Mankoff (2012). "Advances in molecular imaging for breast cancer detection and characterization." In: *Breast Cancer Research* 14.2, pages 206–206 (cited on page 35).

Steinberg, Niclas and Christer Fellers (2002). "Out-of-plane Poisson's ratios of paper and paperboard". In: *Nordic Pulp and Paper Research Journal* 17.2, pages 387–394 (cited on page 156).

Stoeblen, F (2014). "MR vacuum assisted biopsies and contrast injection optimisation". In: *Insights Into Imaging* 5.Suppl 1, page 372 (cited on page 188).

Takada, M et al. (2012). "Predictions of the pathological response to neoadjuvant chemotherapy in patients with primary breast cancer using a data mining technique." In: *Breast Cancer Research and Treatment* 134.2, pages 661–670 (cited on page 121).

Takeda, Kazuna, Shotaro Kanao, Tomohisa Okada, Masako Kataoka, Takayuki Ueno, Masakazu Toi, Hiroshi Ishiguro, Yoshiki Mikami, and Kaori Togashi (2012). "Assessment of CAD-generated tumor volumes measured using MRI in breast cancers before and after neoadjuvant chemotherapy." In: *European Journal of Radiology* 81.10, pages 2627–2631 (cited on page 120).

Tamimi, R et al. (2008). "Comparison of molecular phenotypes of ductal carcinoma in situ and invasive breast cancer". In: *Breast Cancer Research* 10.4, R67 (cited on page 13).

Tan, Tao, Bram Platel, Roel Mus, László Tabár, Ritse M Mann, and Nico Karssemeijer (2013). "Computer-Aided Detection of Cancer in Automated 3-D Breast Ultrasound". In: *IEEE Transactions on Medical Imaging* 32.9, pages 1698–1706 (cited on page 17).

Tan, Tao, Bram Platel, Thorsten Twellmann, Guido van Schie, Roel Mus, André Grivegnée, Ritse M Mann, and Nico Karssemeijer (2013). "Evaluation of the Effect of Computer-Aided Classification of Benign and Malignant Lesions on Reader Performance in

Automated Three-dimensional Breast Ultrasound." In: *Academic Radiology* 20.11, pages 1381–1388 (cited on page 41).

Tanner, C, N Karssemeijer, and G Szekely (2011). "Deformation models for registering MR and 3D ultrasound breast images". In: *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 582–585 (cited on page 200).

Tanner, Christine, Mark White, Salvatore Guarino, Margaret A Hall-Craggs, Michael Douek, and David J Hawkes (2011). "Large breast compressions: observations and evaluation of simulations." In: *Medical Physics* 38.2, pages 682–690 (cited on pages 162, 201).

Teicher, Beverly A (2009). "Human tumor xenografts and mouse models of human tumors: re-discovering the models." In: 4.12, pages 1295–1305 (cited on page 14).

Thakur, Mathew, Kaijun Zhang, Adam Berger, Barbara Cavanaugh, Sung Kim, Andrea Frangos, Colleen Dascenzo, Eric Wickstrom, and Charles Intenzo (2012). "Positron Emission Mammography (PEM): Beyond F-18-FDG". In: *J Nucl Med* 53 (Supplement 1), page 11 (cited on page 35).

Thirion, J P, S Prima, G Subsol, and N Roberts (2000). "Statistical analysis of normal and abnormal dissymmetry in volumetric medical images." In: *Medical Image Analysis* 4.2, pages 111–121 (cited on page 72).

Thomson, J Z, A J Evans, S E Pinder, H C Burrell, A R M Wilson, and I O Ellis (2001). "Growth pattern of ductal carcinoma in situ (DCIS): a retrospective analysis based on mammographic findings". In: *British Journal of Cancer* 85.2, pages 225–227 (cited on page 66).

Timmers, Johanna M, Gerard J den Heeten, Eddy M Adang, Johannes D Otten, André L Verbeek, and Mireille J Broeders (2012). "Dutch digital breast cancer screening: implications for breast cancer care." In: *European Journal of Public Health* 22.6, pages 925–929 (cited on page 137).

Tofts, P S and A G Kermode (1991). "Measurement of the blood-brain barrier permeability and leakage space using dynamic MR imaging. 1. Fundamental concepts." In: *Magnetic Resonance in Medicine* 17.2, pages 357–367 (cited on page 32).

Tourassi, G D (1999). "Journey toward computer-aided diagnosis: role of image texture analysis". In: *Radiology* 213.2, pages 317–320 (cited on page 90).

US Preventive Services Task Force (2009). "Annals of Internal Medicine | Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement". In: *Annals of internal medicine* (cited on page 138).

Valenzuela, Marcia and Thomas B Julian (2007). "Ductal carcinoma in situ: biology, diagnosis, and new therapies". In: *Clinical Breast Cancer* 7.9, pages 676–681 (cited on page 12).

Vapnik, Vladimir N (1998). "An overview of statistical learning theory." In: *IEEE Transactions on Neural Networks* 10.5, pages 988–999 (cited on page 44).

Varela, C, S Timp, and N Karssemeijer (2006). "Use of border information in the classification of mammographic masses". In: *Physics in Medicine and Biology* 51.2, pages 425–441 (cited on page 44).

Vargas, Ana Cristina et al. (2012). "Gene expression profiling of tumour epithelial and stromal compartments during breast cancer progression." In: *Breast Cancer Research and Treatment* 135.1, pages 153–165 (cited on pages 9, 13).

Vargo-Gogola, Tracy and Jeffrey M Rosen (2007). "Modelling breast cancer: one size does not fit all". In: *Nature Reviews Cancer* 7.9, pages 659–672 (cited on page 14).

Varma, Sudhir and Richard Simon (2006). "Bias in error estimation when using cross-validation for model selection." In: *BMC Bioinformatics* 7, pages 91–91 (cited on pages 58, 61).

Verardi, N, G Di Leo, L A Carbonaro, M P Fedeli, and F Sardanelli (2012). "Contrast-enhanced MR imaging of the breast: association between asymmetric increased breast vascularity and ipsilateral cancer in a consecutive series of 197 patients." In: *La Radiologia medica* (cited on page 71).

Viehweg, P, T Bernerth, M Kiechle, J Buchmann, A Heinig, H Koelbl, M Laniado, and S H Heywang-Kobrunner (2006). "MR-guided intervention in women with a family history of breast cancer." In: *European Journal of Radiology* 57.1, pages 81–89 (cited on page 188).

Viola, P and M Jones (2001). "Rapid object detection using a boosted cascade of simple features". In: *IEEE Conference on Computer Vision and Pattern Recognition. Proceedings* 1, pages I–1 (cited on page 48).

Wang, Yinhai, Kate E Williamson, Paul J Kelly, Jacqueline A James, and Peter W Hamilton (2012). "SurfaceSlide: a multitouch digital pathology platform." In: *PloS one* 7.1, e30783 (cited on pages 114, 116).

Wernick, M N, Yongyi Yang, J G Brankov, G Yourganov, and S C Strother (2010). "Machine Learning in Medical Imaging". In: *IEEE Signal Processing Magazine* 27.4, pages 25–38 (cited on page 46).

Whiteley, Jonathan P, David J Gavaghan, S Jonathan Chapman, and J Michael Brady (2007). "Non-linear modelling of breast tissue". In: *Mathematical Medicine and Biology* 24.3, pages 327–345 (cited on page 162).

Wiechmann, Lisa and Henry M Kuerer (2008). "The molecular journey from ductal carcinoma in situ to invasive breast cancer". In: *Cancer* 112.10, pages 2130–2142 (cited on page 13).

Wolf, Catherine G (1988). "A Comparative Study of Gestural and Keyboard Interfaces". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 32, pages 273–277 (cited on page 111).

Woods, Brent J, Bradley D Clymer, Tahsin Kurc, Johannes T Heverhagen, Robert Stevens, Adem Orsdemir, Orhan Bulan, and Michael V Knopp (2007). "Malignant-lesion segmentation using 4D co-occurrence texture analysis applied to dynamic contrast-

enhanced magnetic resonance breast image data." In: *Journal of Magnetic Resonance Imaging* 25.3, pages 495–501 (cited on page 86).

Wu, Yuzheng, ML Giger, K Doi, CJ Vyborny, BA Schmidt, and CE Metz (1993). "Artificial Neural Networks In Mammography: Application to Decision Making In the Diagnosis ofBreast Cancer"'. In: *Radiology* 187.1, pages 81–87 (cited on page 44).

Yamamoto, Shota, Daniel D Maki, Ronald L Korn, and Michael D Kuo (2012). "Radiogenomic Analysis of Breast Cancer Using MRI: A Preliminary Study to Define the Landscape." In: *AJR American Journal of Roentgenology* 199.3, pages 654–663 (cited on page 13).

Yang, Qian, Lihua Li, Juan Zhang, Guoliang Shao, Chengjie Zhang, and Bin Zheng (2013). "Computer-Aided Diagnosis of Breast DCE-MRI Images Using Bilateral Asymmetry of Contrast Enhancement Between Two Breasts." In: *Journal of Digital Imaging* (cited on page 70).

Yeager, David (2012). "In CAD We Trust? — Researcher Seeks Ways to Build Radiologists' Comfort With Computer-Aided Detection Software". In: *Radiology today Magazine* 13.2, page 6 (cited on page 108).

Yeh, I-Tien and Carolyn Mies (2008). "Application of immunohistochemistry to breast lesions." In: *Audio, Transactions of the IRE Professional Group on* 132.3, pages 349–358 (cited on page 12).

Yi, M, F Meric-Bernstam, H M Kuerer, E A Mittendorf, I Bedrosian, A Lucci, R F Hwang, J R Crow, S Luo, and K K Hunt (2012). "Evaluation of a Breast Cancer Nomogram for Predicting Risk of Ipsilateral Breast Tumor Recurrences in Patients With Ductal Carcinoma in Situ After Local Excision". In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 30.6, pages 600–607 (cited on page 63).

Yu, Lingyun, Pjotr Svetachov, Petra Isenberg, Maarten H Everts, and Tobias Isenberg (2010). "FI3D: direct-touch interaction for the exploration of 3D scientific visualization spaces." In: *IEEE Transactions on Visualization and Computer Graphics* 16.6, pages 1613–1622 (cited on page 111).

Yuan, Yading, Maryellen L Giger, Hui Li, Neha Bhooshan, and Charlene A Sennett (2010). "Multimodality Computer-Aided Breast Cancer Diagnosis with FFDM and DCE-MRI". In: *Academic Radiology* 17.9, pages 10–10. doi: 10.1016/j.acra.2010.04.015 (cited on page 18).

Zacharaki, Evangelia I, Cosmina S Hogea, George Biros, and Christos Davatzikos (2008). "A comparative study of biomechanical simulators in deformable registration of brain tumor images." In: *IEEE Transactions on Bio-Medical Engineering* 55.3, pages 1233–1236 (cited on page 201).

Zweig, M H and G Campbell (1993). "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." In: *Clinical Chemistry* 39.4, pages 561–577 (cited on pages 56, 57).

Conference Proceedings and Others

Aggarwal Charu C.and Hinneburg, Alexander and Daniel A. Keim (2001). "On the Surprising Behavior of Distance Metrics in High Dimensional Space". English. In: *Database Theory — ICDT 2001*. Edited by Jan Bussche and Victor Vianu. Volume 1973. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pages 420–434. isbn: 978-3-540-41456-8. doi: 10.1007/3-540-44503-X_27. url: http://dx.doi.org/10.1007/3-540-44503-X_27 (cited on page 49).

Agner, S C, S Soman, E Libfeld, and M McDonald (2008). "Novel kinetic texture features for breast lesion classification on dynamic contrast enhanced (DCE) MRI". In: *Proc. SPIE* (cited on pages 86, 100).

Airola, Antti, Tapio Pahikkala, Willem Waegeman, Bernard De Baets, and Tapio Salakoski (2010). "A comparison of AUC estimators in small-sample studies". In: *JMLR: Workshop and Conference Proceedings*. Edited by Saso Dzeroski, Pierre Geurts, and Juho Rousu, pages 3–13 (cited on page 56).

American Cancer Society (2013a). *Breast Cancer Facts and Figures 2011-2012*. url: http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-030975.pdf (cited on page 26).

— (2013c). *Exams and Test Descriptions*. url: http://www.cancer.org/treatment/understandingyourdiagnosis/examsandtestdescriptions/index (visited on 12/23/2013) (cited on page 186).

Avila-Garcia, M S, A E Trefethen, M Brady, and F Gleeson (2010). "Using interactive and multi-touch technology to support decision making in multidisciplinary team meetings". In: *Computer-Based Medical Systems (CBMS), 2010 IEEE 23rd International Symposium on*, pages 98–103 (cited on page 116).

Avila-Garcia, M S, A E Trefethen, M Brady, F Gleeson, and D Goodman (2008). "Lowering the Barriers to Cancer Imaging". In: *eScience, 2008. eScience '08. IEEE Fourth International Conference on*, pages 63–70 (cited on page 116).

Basaran, Salahi (2008). "Lagrangian and Eulerian Descriptions in Solid Mechanics and Their Numerical Solutions in hpk Framework". PhD. University of Kansas (cited on page 158).

Bengio, Yoshua, Pascal Lamblin, Dan Popovici, and Hugo Larochelle (2007). "Greedy Layer-Wise Training of Deep Networks". In: *Advances in Neural Information Processing Systems 19:Proceedings of the 2006 Conference*. Edited by B Schoelkopf, J Platt, and T Hofmann, pages 153–160 (cited on page 44).

Boehler, Tobias (2011). "Deformable Image Registration Methods for Clinical Applications of Magnetic Resonance Mammography". PhD thesis. Universität Bremen (cited on pages 30, 122, 153, 160, 191, 201).

Boehler, Tobias, Kathy Schilling, Ulrich Bick, and Horst K Hahn (2010). "Deformable Image Registration of Follow-up Breast MR Images". In: *Proceedings of WBIR*. Volume 2604. Lecture Notes in Computer Science. Lübeck: Springer, pages 13–24 (cited on page 120).

Boehler, Tobias, Stefan Wirtz, and Heinz-Otto Peitgen (2007). "A Combined Algorithm for Breast MRI Motion Correction". In: *Proc. SPIE Medical Imaging*. Volume 6514 (cited on page 73).

Breiman, Leo and Adele Cutler (2013). *Random Forests*. URL: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (visited on 10/31/2013) (cited on page 64).

Bro-Nielsen, Morten (1997). "Medical Image Registration and Surgery Simulation". PhD thesis. Technical University of Denmark, Lyngby, Denmark (cited on pages 153, 155, 159, 160, 162).

Chauhan, S, R Mishra, S Kumar, and M Y Teo (2002). "A robot for non-invasive breast cancer surgery". In: *Control, Automation, Robotics and Vision, 2002. ICARCV 2002. 7th International Conference on*, pages 425–428 (cited on page 123).

Cooper, Stephanie (2013). *New Study Reveals Physicians' Device and Digital Media Adoption Rapidly Evolving*. URL: http://manhattanresearch.com/News-and-Events/Press-Releases/physician-digital-media-adoption (visited on 11/11/2013) (cited on page 109).

Demšar, Janez, Blaž Zupan, Gregor Leban, and Tomaž Curk (2004). "Orange: From Experimental Machine Learning to Interactive Data Mining". In: *Knowledge Discovery in Databases: PKDD 2004*. Edited by Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi. Volume 3202. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pages 537–539. ISBN: 978-3-540-23108-0. DOI: 10.1007/978-3-540-30116-5_58. URL: http://dx.doi.org/10.1007/978-3-540-30116-5_58 (cited on page 59).

Dick, Christian, Joachim Georgii, Rainer Burgkart, and Rüdiger Westermann (2008). "Computational Steering for Patient-Specific Implant Planning in Orthopedics". In: *Proceedings of Visual Computing for Biomedicine 2008*, pages 83–92 (cited on page 163).

Economist Intelligence Unit (2009). *Breakaway: The global burden of cancer – challenges and opportunities*. URL: http://www.livestrong.org/pdfs/GlobalEconomicImpact (cited on page 24).

Farley, Shawn and Heather Williams (2014). *BMJ Article on Breast Cancer Screening Effectiveness Incredibly Flawed and Misleading*. American College of Radiology. URL: http://www.acr.org/News-Publications/News/News-Articles/2014/ACR/BMJ-Article-on-Breast-Cancer-Screening-Effectiveness-Incredibly-Flawed-and-Misleading (visited on 02/25/2014) (cited on page 138).

Ford, Steven T, Ivan Viola, Stefan Bruckner, Hans Torp, and Gabriel Kiss (2012). "HeartPad: real-time visual guidance for cardiac ultrasound". In: *WASA '12: Proceedings of the Workshop at SIGGRAPH Asia*. ACM Request Permissions (cited on page 114).

Freund, Yoav and Robert E Schapire (1996). "Experiments with a New Boosting Algorithm". In: *International Conference on Machine Learning*, pages 148–156 (cited on page 48).

Georgii, Joachim (2007). "Real-Time Simulation and Visualization of Deformable Objects". http://mediatum2.ub.tum.de/node?id=627732. PhD thesis. Technische Universität München (cited on pages 153, 159, 160, 162, 163).

Georgii, Joachim, Daniel Lagler, Christian Dick, and Rüdiger Westermann (2010). "Interactive Deformations with Multigrid Skeletal Constraints". In: *Proceedings of the 7th Workshop On Virtual Reality Interaction and Physical Simulation*, pages 39–47 (cited on page 174).

Georgii, Joachim and Rüdiger Westermann (2008). "Corotated Finite Elements Made Fast and Stable". In: *Proceedings of the 5th Workshop On Virtual Reality Interaction and Physical Simulation*, pages 11–19 (cited on page 162).

Georgii, Joachim, Fabian Zöhrer, and Horst K Hahn (2013). "Model-based position correlation between breast images". In: *Proc. SPIE Medical Imaging*. Volume 8670. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. DOI: 10.1117/12.2007472. URL: http://dx.doi.org/10.1117/12.2007472 (cited on page 200).

Glassman, Leonard and Marieke Hazewinkel (2013). *Radiology Assistant: Breast MRI*. URL: http://www.radiologyassistant.nl/en/p47a585a7401a9 (visited on 10/31/2013) (cited on page 89).

Goebl, Sebastian, Claudia Plant, Marc Lobbes, and Anke Meyer-Bäse (2013). "CAD-system based on kinetic analysis for non-mass-enhancing lesions in DCE-MRI". In: *Proc. SPIE Defense, Security, and Sensing*. Edited by Harold H Szu. SPIE (cited on page 71).

Harz, Markus, Volker Diehl, Bernd Merkel, B Terwey, and Heinz-Otto Peitgen (2009). "Fast unsupervised hot-spot detection in 1H-MR spectroscopy data using ICA". In: *Proc. SPIE Medical Imaging*, pages 7259–7268 (cited on page 59).

Harz, Markus, Joachim Georgii, Kathy Schilling, and Horst K Hahn (2011a). "Real-time Breast Deformation using Non-linear Tissue Properties". In: *Lecture Notes in Informatics*. Edited by Hans-Ulrich Heiss, Peter Pepper, Bernd-Holger Schlingloff, and Jörg Schneider. Volume P-192, page 442 (cited on pages 152, 172).

— (2011b). "Towards Navigated Breast Surgery Using Efficient Breast Deformation Simulation". In: *Medical Image Computing and Computer-Assisted Intervention, Workshop on Breast Image Analysis*. Edited by Christine Tanner, Julia Schnabel, Nico Karssemeijer, Mads Nielsen, Maryellen Giger, and David Hawkes (cited on pages 152, 172, 177).

Harz, Markus, Joachim Georgii, Lei Wang, Kathy Schilling, and Heinz-Otto Peitgen (2012a). "Efficient Breast Deformation Simulation". In: *Workshop on Virtual Reality Interaction and Physical Simulation*. The Eurographics Association, pages 117–126 (cited on page 152).

Harz, Markus, Felix Ritter, Simon Benten, Kathy Schilling, and Heinz-Otto Peitgen (2012). "A Novel Workflow-Centric Breast MRI Reading Prototype Utilizing Multitouch Gestures". In: *Breast Imaging – 11th International Workshop*. Edited by Andrew D A Maidment, Predrag R Bakic, and Sara Gavenonis. Volume 7361. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pages 276–283 (cited on page 109).

Homeyer, Andr'e, Michael Schwier, and Horst K Hahn (2010). "A Generic Concept for Object-Based Image Analysis". In: *Proceedings of the International Conference on Computer Vision Theory and Applications*. Volume 2, pages 530–533 (cited on page 72).

Kass, Rena B, D Scott Lind, and Wiley W Souba (2007). "Breast Procedures". In: ACS Surgery: Priciples & Practice (cited on page 122).

Kin, Kenrick, Maneesh Agrawala, and Tony DeRose (2009). "Determining the benefits of direct-touch, bimanual, and multifinger input on a multitouch workstation". In: *GI '09: Proceedings of Graphics Interface 2009*. Canadian Information Processing Society (cited on page 111).

Kohavi, Ron (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". In: *IJCAI*, pages 1137–1145. URL: citeseer.nj.nec.com/kohavi95study.html (cited on page 58).

Larson, B T, N V Tsekos, and A G Erdman (2003). "A robotic device for minimally invasive breast interventions with real-time MRI guidance". In: *Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on*, pages 190–197 (cited on page 123).

Laue, Hendrik, Anja Hennemuth, Volker Diehl, Markus T Harz, Horst K Hahn, and Heinz-Otto Peitgen (2010). "Influence of Contrast Arrival Time and Temporal Resolution in Diagnosis of Breast Cancer with DCE-MRI". In: *Proc ISMRM* (cited on page 33).

Liao, Wen-Hung (2010). "Region Description Using Extended Local Ternary Patterns". In: *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE Computer Society, pages 1003–1006 (cited on page 106).

Loose, Jennifer, Markus Harz, Hendrik Laue, Thorsten Twellmann, Ulrich Bick, Marga Rominger, Horst K Hahn, and Heinz-Otto Peitgen (2009). "Assessment of texture analysis on DCE-MRI data for the differentiation of breast tumor lesions". In: volume 7260. DOI: 10.1117/12.812971. URL: http://dx.doi.org/10.1117/12.812971 (cited on pages 40, 85, 101).

Moltz, Jan Hendrik (2013). "Lesion segmentation and tracking for CT-based chemotherapy monitoring". PhD thesis. Bremen: School of Engineering and Science, Jacobs University Bremen (cited on page 21).

Morris, Derek John, Eike Falk Anderson, and Christopher Peters (2012). "A modular framework for deformation and fracture using GPU shaders". In: *2012 18th International Conference on Virtual Systems and Multimedia (VSMM)*. IEEE, pages 267–274 (cited on page 184).

Morris, E A, C E Comstock, and C H Lee (2013). "ACR BI-RADS®Magnetic Resonance Imaging". In: *ACR BI-RADS®Atlas, Breast Imaging Reporting and Data System*. 5th edition. Reston, VA: American College of Radiology (cited on page 65).

Mus, Roel D M, Ritse M Mann, A Moyakine, Christian Geppert, Bram Platel, Nico Karssemeijer, and Jelle O Barentsz (2012). "MRI Screening of the Breast in Less than 2 Minutes: A Prelude to Extend MR Breast Screening Possibilities". In: *RSNA 2012 Scientific Assembly and Annual Meeting* (cited on page 33).

North, Chris, Tim Dwyer, Bongshin Lee, Danyel Fisher, Petra Isenberg, George Robertson, and Kori Inkpen (2009). "Understanding Multi-touch Manipulation for Surface Computing". In: *INTERACT '09: Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part II*. Springer-Verlag (cited on page 111).

Owen, Mark (2013). *The Wireworld Computer*. URL: http://www.quinapalus.com/wi-index.html (visited on 11/28/2013) (cited on page 206).

Rajagopal, Vijay, Martyn P Nash, Ralph P Highnam, and Poul M Nielsen (2008). "The Breast Biomechanics Reference State for Multi-modal Image Analysis". In: *IWDM '08: Proceedings of the 9th international workshop on Digital Mammography*. Springer-Verlag (cited on page 162).

Ranzato, Marc'Aurelio, Christopher Poultney, Yann LeCun, and Sumit Chopra (2006). "Efficient Learning of Sparse Representations with an Energy-Based Model". In: *Advances in Neural Information Processing Systems 19:Proceedings of the 2006 Conference*. Edited by B Schoelkopf, J Platt, and T Hofmann, pages 1137–1144 (cited on page 44).

Rennie, Jason D M, Lawrence Shih, Jaime Teevan, and David R Karger (2003). *Tackling the Poor Assumptions of Naive Bayes Text Classifiers*. Technical report. Artificial Intelligence Laboratory, Massachusetts Institute of Technology; Cambridge, MA 02139 (cited on page 63).

Rish, Irina (2001). *An empirical study of the naive Bayes classifier*. Technical report RC 22230 (W0111-014). New York: IBM Research Division (cited on page 63).

Sayaz, Francisco J (2008). *A Gentle Introduction to the Finite Element Method*. URL: `http://www.math.udel.edu/~fjsayas/anIntro2FEM.pdf` (visited on 10/27/2013) (cited on pages 153, 162).

Schenk, Andrea, Alexander Koehn, Ryusei Matsuyama, and Itaru Endo (2013). "Transfer of Liver Surgery Planning into the Operation Room: Initial Experience with the iPad". In: *10. Congress of the European-African Hepato Pancreato Biliary Association*, page 27.15 (cited on page 114).

Schiwietz, Thomas, Joachim Georgii, and Rüdiger Westermann (2007). "Freeform Image". In: *Proceedings of Pacific Graphics* (cited on pages 163, 201).

Schneider, Jens, Martin Kraus, and Rüdiger Westermann (2010). "GPU-Based Euclidean Distance Transforms and Their Application to Volume Rendering". In: *Selected papers of VISIGRAPP 2009, Communications in Computer and Information Science (CCIS) 68*. Springer-Verlag Berlin Heidelberg, pages 215–228 (cited on page 165).

Schölkopf, Bernhard (2001). "The Kernel Trick for Distances". In: Leen, T K, T G Dietterich, and V Tresp. *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*. Advances in neural information processing systems. MIT Press, * (cited on page 47).

Shalizi, Cosma Rohilla (2006). "Methods and Techniques of Complex Systems Science: An Overview". In: *Complex Systems Science in Biomedicine*. Edited by ThomasS Deisboeck and Kresh. Topics in Biomedical Engineering International Book Series. Boston, MA: Springer US. Chapter 2, pages 33–114. ISBN: 978-0-387-30241-6. DOI: `10.1007/978-0-387-33532-2\_2`. URL: `http://dx.doi.org/10.1007/978-0-387-33532-2%5C_2` (cited on page 6).

— (2012). *Review: A Rare Blend of Monster Raving Egomania and Utter Batshit Insanity*. URL: `http://vserver1.cscs.lsa.umich.edu/~crshalizi/reviews/wolfram/` (visited on 09/23/2013) (cited on pages 6, 205).

Singhi, Surendra K and Huan Liu (2006). "Feature subset selection bias for classification learning". In: *ICML '06: Proceedings of the 23rd international conference on Machine learning*. ACM (cited on page 52).

Snowden, David (2000). "Cynefin: a sense of time and space, the social ecology of knowledge management". In: *Knowledge Horizons : The Present and the Promise of Knowledge Management*. Edited by C. Despres and D. Chauvel. Oxford: Butterworth Heinemann. URL: http://books.google.com/books/?id=LR3amB6hHE0C (cited on page 5).

Srikantha, Abhilash, Markus T Harz, Lei Wang, Bram Platel, Ritse M Mann, Horst K Hahn, and Heinz-Otto Peitgen (2012). "Symmetry-based detection of ductal carcinoma in situ in breast MRI". In: *European Journal of Radiology*. Volume 81, pages 158–159 (cited on page 40).

Srikantha, Abhilash, Markus Harz, Gillian Newstead, Lei Wang, Bram Platel, Ritse M Mann, Horst K Hahn, and Heinz-Otto Peitgen (2013). "Symmetry-Based Detection and Diagnosis of DCIS in Breast MRI". In: *Proc. SPIE Medical Imaging*. Volume 8670. DOI: 10.1117/12.2000061. URL: http://dx.doi.org/10.1117/12.2000061 (cited on pages 40, 120).

Tape, Thomas G (2014). *Interpreting Diagnostic Tests*. URL: http://gim.unmc.edu/dxtests/ROC3.htm (visited on 03/28/2014) (cited on page 41).

Taylor, Zeike A, Stian Johnsen, and Sebastien Ourselin (2013). *NiftySim*. URL: http://sourceforge.net/projects/niftysim/ (cited on page 163).

Teves, J, B Chaparro, B Kennedy, R Chan, N Copic, R Riss, and J Simmons (2013). *Exploring iPad usage by Healthcare Professionals in a Pediatric Hospital*. URL: http://usabilitynews.org/exploring-ipad-usage-by-healthcare-professionals-in-a-pediatric-hospital/ (visited on 11/11/2013) (cited on page 109).

Uebersax, John (2010). *Kappa Coefficients: A Critical Appraisal*. URL: http://www.john-uebersax.com/stat/kappa.htm (visited on 09/28/2013) (cited on page 54).

U.S. Food and Drug Administration (2012). *Guidance for Industry and Food and Drug Administration Staff - Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions*. URL: http://www.fda.gov/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm187249.htm (visited on 11/25/2013) (cited on page 108).

Wang, Lei, Kostas Filippatos, Ola Friman, and Horst K Hahn (2011). "Fully automated segmentation of the pectoralis muscle boundary in breast MR images". In: *Proc. SPIE Medical Imaging*. Volume 7963. DOI: 10.1117/12.877645. URL: http://dx.doi.org/10.1117/12.877645 (cited on pages 73, 177, 190).

Wang, Lei, Markus Harz, Tobias Boehler, Bram Platel, Andre Homeyer, and Horst K Hahn (2014). "A robust and extendable framework towards fully automated diagnosis of nonmass lesions in breast DCE-MRI". In: *IEEE International Symposium on Biomedical Imaging*. Volume accepted (cited on page 40).

Wenzel, Markus, Bernd Merkel, Matthias Althaus, Matthias Nölte, and Heinz-Otto Peitgen (2006). "On dimensionality reduction for high-resolution ex-vivo NMR

spectra". In: *Proceeding ESMRMB*. Volume 19, Supplement 7. Magnetic Resonance Materials in Physics, Biology and Medicine, page 32 (cited on page 59).

Wenzel, Markus, Bernd Merkel, Matthias Althaus, and Heinz-Otto Peitgen (2006). "PCNSA for NMR Spectroscopy Breast Tissue Classification". In: *ISMRM DP Spect WS* (cited on page 59).

Werbos, Paul (1974). "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences". PhD thesis. Cambridge, MA: Harvard University (cited on page 43).

Wikipedia (2013a). *Decision Tree Learning*. url: `http://en.wikipedia.org/wiki/Decision_tree_learning` (visited on 10/31/2013) (cited on page 63).

— (2013b). *Infinitesimal strain theory*. Version 578774894. url: `http://en.wikipedia.org/wiki/Infinitesimal_strain_theory` (visited on 11/19/2013) (cited on page 155).

Wu, Shandong, Susan Weinstein, and Despina Kontos (2012). "Atlas-Based probabilistic fibroglandular tissue segmentation in breast MRI". In: *MICCAI'12: Proceedings of the 15th international conference on Medical Image Computing and Computer-Assisted Intervention*. Springer-Verlag (cited on page 3).

Index

A

Accuracy 54
Anatomy 25
Artificial neural network 43, **43**, 44
Attribute 42
AUC 56
Auxetic 158

B

BI-RADS 69, 83
Bias 50
Breast
  Anatomy 23
  Biopsy 187
  Cancer 7
  Coil 29
  Development 23
  Surgery 178
  Symmetry 71
Breast conserving therapy 124
Brier score 55

C

CADe . . . . . *see* Computer-aided detection
CADx . . . . . *see* Computer-aided diagnosis
Cauchy strain tensor 160
Cauchy stress tensor 157
Classification **41**
  Bias 50
  Curse of dimensionality 48, 59
  Evaluation 53, 59
  Performance criteria 54–56
  Validation 57
  Variance 50
Classifier **42**
Co-occurrence matrix 77, **87**
Computer-aided detection 17, **20**, **39**, 85
  Interactive CAD 109
  Mammography 41
Computer-aided diagnosis 17, **20**, **39**, 86
  Future 108
Confusion matrix **53**
Cooper's ligaments **25**
Cross validation 58
Curse of dimensionality 101

D

DCIS . . . . . *see* Ductal carcinoma in situ
Decision support system **21**, 109
Displacement field **160**, 162
Displacement function **160**
Domain 42
Ductal carcinoma in situ **11**, 39, **65**
  Example 67
  Microcalcification 69

E

EARTH . . . . . *see* MARS
Epidiomolgy (breast cancer) 24

Estrogen 23

F

F^{18}-FDG 35
Feature 42
Finite element method **155**
Full-field digital mammography 22, 27

G

GLCM . . . . . *see* Co-occurrence matrix

H

Haralick texture features **90**
Hooke's law **158**, 159
    Matrix notation 159

I

Imaging 26
in situ 39

K

$\mathbf{K}$ (Global stiffness matrix) 163
$\mathbf{K}^e$ (Element stiffness matrix) 162
κ (kappa) 54

L

Lamè coefficients 161
Local binary patterns 107
Local ternary patterns 107

M

Machine learning **41**
Magnetic resonance imaging **27**
    Dynamic contrast-enhanced 29
    Enhancement curve 30
    MRI guided biopsy 153
    Spectroscopy 29
Mammary gland 23
Mammography **27**
    Alternator 16
    Microcalcification 28, 40, 44, 69
    Screening 14, **16**, 139
    Workstation 16
MARS 102
Material matrix 159
MCC 55
MeVisLab 73, 126, 165
Microcalcification . . . . . *see* Mammography
Molecular pathways 23
Morphology
    Ductal carcinoma in situ 66, 68
    Features **79**
Motion correction 73
Motion equations **160**
MRSI 34

N

Naïve Bayes **61**
    Bias 62
    Chain rule 61
Nipple detection 74
Non-mass lesion 69
Nuclear grade 66, **68**

O

Oject-based image analysis 72

P

Pattern recognition 41
Poisson ratio **158**, 184
Positron emission mammography 35
Potential energy functional 155
Predictor **42**

R

Random forest **63**
Regressor **42**
ROC 55

S

Selection bias 51
Shape function 162

Stiffness matrix
Element 162
Global 163
Stiffness tensor 158
Strain 157
Stress 156
Stroma 8, **9**, 13, 23, 65
Support vector machine 34, **44**
SVM . . . . . *see* Support vector machine
Symmetry **71**, 75

T

Terminal duct lobular units 23, **25**, 65
Texture **85**, **87**
Classification peformance 97
Haralick features 90
of contrast kinetics 93
Total energy functional 160

U

Ultrasound 34
ABVS 34
Shear wave ultrasound 34
Volume ultrasound 34

V

Variance 50

W

Workflow **15**, 18, 21, 119

Y

Young's modulus **158**