Dissertation
submitted to the
Faculty of Physics
University of Bremen, Germany
for the degree of
Doctor of Natural Sciences (Dr. rer. nat.)

Computer-aided image quality assessment in automated 3D breast ultrasound images

Julia Schwaab

Referees:
Prof. Dr. Matthias Günther
Prof. Dr. Robert Martí
Date of colloquium: 22.01.2016

Abstract

Automated 3D breast ultrasound (ABUS) is a valuable, non-ionising adjunct to X-ray mammography for breast cancer screening and diagnosis for women with dense breasts. High image quality is an important prerequisite for diagnosis and has to be guaranteed at the time of acquisition. The high throughput of images in a screening scenario demands for automated solutions.

In this work, an automated image quality assessment system rating ABUS scans at the time of acquisition was designed and implemented.

Quality assessment of present diagnostic ultrasound images has rarely been performed demanding thorough analysis of potential image quality aspects in ABUS. Therefore, a reader study was initiated, making two clinicians rate the quality of clinical ABUS images. The frequency of specific quality aspects was evaluated revealing that incorrect positioning and insufficiently applied contact fluid caused the most relevant image quality issues.

The relative position of the nipple in the image, the acoustic shadow caused by the nipple as well as the shape of the breast contour reflect patient positioning and ultrasound transducer handling. Morphological and histogram-based features utilized for machine learning to reproduce the manual classification as provided by the clinicians. At 97 % specificity, the automatic classification achieved sensitivities of 59 %, 45 %, and 46 % for the three aforementioned aspects, respectively.

The nipple is an important landmark in breast imaging, which is generally—but not always correctly—pinpointed by the technicians. An existing nipple detection algorithm was extended by probabilistic atlases and exploited for automatic detection of incorrectly annotated nipple marks. The nipple detection rate was increased from 82 % to 85 % and the classification achieved 90 % sensitivity at 89 % specificity.

A lack of contact fluid between transducer and skin can induce reverberation patterns and acoustic shadows, which can possibly obscure lesions. Parameter maps were computed in order to localize these artefact regions and yielded a detection rate of 83 % at 2.6 false positives per image.

Parts of the presented work were integrated to clinical workflow making up a novel image quality assessment system that supported technicians in their daily routine by detecting images of insufficient quality and indicating potential improvements for a repeated scan while the patient was still in the examination room. First evaluations showed that the proposed method sensitises technicians for the radiologists' demands on diagnostically valuable images.

Contents

Preface

Since the dawn of history, people have suffered from and written about cancer. Especially breast cancer has been mentioned in nearly every period of history, since breast lumps, unlike other internal cancers, tend to manifest themselves as visible tumours. The oldest evidence of breast cancer was discovered in Egypt in 2015. The 4,200-year-old skeleton of an adult woman shows the typical destructive damage provoked by the extension of a breast cancer as a metastasis (Mourad & Stonestreet 2015). The unknown author of the Edwin Smith papyrus, which was written around the 17th century BC, describes "ball-like chest tumours" as "an ailment I will fight with", meaning that there is no cure.

In the 18th century, a local therapy seemed to be an option and the lack of anaesthesia did not prevent brave physicians from performing mastectomies (see figure 0.1), not always to the patient's best interest (Olson 2002).

Nowadays, breast cancer has not lost a bit of the terror it spreads, being the most common cancer in women worldwide. Although treatment options in more developed regions are manifold comprising surgery, chemo- and radiation therapy, it remains the second cause of cancer death in women after lung cancer (Ferlay et al. 2013).

Since early detection of breast cancer improves outcomes (Etzioni et al. 2003), screening programmes have been established. Multiple studies have shown that standard X-ray mammography screening reduces mortality from breast cancer (Tabár et al. 1985), but it is not equally effective in all women. Overall, the sensitivity of mammography for detecting breast cancer is around 80 %. However, in women with radiographically dense breast tissue, the sensitivity can get as low as 48 % (Kolb et al. 2002). Therefore, stratified screening programmes making use of other modalities as ultrasound (US) or magnetic resonance imaging (MRI), tailored to risk assessment based on family history, age, genetic profiles, and breast density, are proposed (Drukteinis et al. 2013). Specifically, automated (whole-) breast ultrasound (ABUS) has been shown to support the early detection of small invasive cancers that are occult on mammography in women with dense breasts (Drukteinis et al. 2013; Mandelson et al. 2000; Yaghjyan et al. 2011; Kelly et al. 2010a). Compared to hand-held US, ABUS systems acquire 3D volumes that can be stored on an image archiving system, thus enabling temporal comparison of exams with relevant priors. The acquisition can be performed by non-radiologists, e.g. technicians, consequently reducing the costs of the acquisition procedure. Although the acquisition is automated to a high degree, image quality still depends on the imaging procedure: inadequate device parameter settings or incorrect positioning of the transducer can cause diverse image artefacts, which may impair the diagnostic evaluation of the ABUS exam substantially.

An automated image quality assessment tool that detects artefacts directly after image acquisition has the potential of enhancing overall image quality

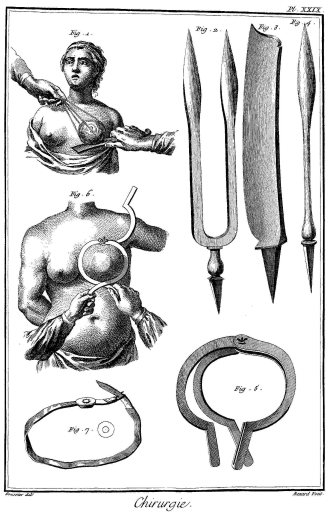


Figure 0.1: Sketch by Louis-Jacques Goussier showing some tools for mastectomy. Printed in "Encyclopédie Didérot", Planche XXIX, written by Louis de Jaucourt in 1752.

in ABUS screening images and, thus, facilitating diagnosis. Due to the high amount and the complexity of volumetric images that have to be analysed in an ABUS screening set up, there is a trend towards computer aided detection (CADe) systems (Tan et al. 2015; Moon et al. 2012). High-quality images are an essential pre-requisite particularly for these tools that are supposed to support radiologists in their daily routine. The work presented in this thesis was driven by the idea of a comprehensive fully automatic image quality assessment system that is able to process acquired images in real-time and to provide immediate feed-back to the medical technicians or nurses. Detailed information on potential artefacts in an ABUS image would allow immediate repetition of a scan with corrected parameters at low cost since an additional scan only takes several minutes and US causes no radiation burden to the patient.

Objectives

The present work has been embedded in the European Commission's FP 7 project ASSURE targeting personalised breast cancer screening programmes. The objective of this thesis is the development of an automated image quality assessment system for automated 3D breast ultrasound (ABUS), which has shown to be a valuable adjunct to X-ray mammography. The scope of ASSURE ranges from improved risk estimation over stratified screening examinations to sophisticated computer aided detection. A stratification of the screening population as suggested by preliminary results of the project will entail an increased amount of ABUS examinations. Good image quality is of highest importance for reliable diagnostics as well as for further image processing. An automated image quality assessment tool as envisaged in this thesis will process the images right after acquisition and alert the technicians if unwanted artefacts are detected such that the scan can easily be repeated with the patient still present in the examination room. To achieve this goal, the following steps have to be undertaken:

(i) Potential ABUS image quality aspects and artefacts have to be defined and discussed together with experienced clinical researchers. The relevance of specific quality aspects has to be evaluated in a reader study based on a data set of original images acquired in routine clinical care.

(ii) Characteristic physical and visual properties as well as the origin of the most relevant image quality aspects have to be examined and translated to quantitative metrics on different scales.

(iii) Since the nipple is an important landmark in breast imaging, automated assessment of nipple visibility has to be designed adjoining empiric atlases to existing methods based on probability maps.

(iv) Machine learning (classifier training) has to be deployed correlating features based on physical properties of ultrasound to ground truth annotations provided by clinicians.

(v) Integrated to an existing software framework for data management and workflow design, the developed algorithms will make up a first prototype for automated image quality assessment in ABUS, which will have to be evaluated in clinical routine with respect to usability and utility.

1 Basics

The female breast has a very complex structure that keeps on changing over a woman's lifetime. As described, e.g., in Berg & Yang (2014), there are different relevant landmarks in breast development. Starting with the menarche, the breast prepares for pregnancy every menstrual cycle by increasing the amount of stromal (connective) and ductal tissue, which is dismantled again if no fertilisation took place. In case of pregnancy, the ductal and lobular tissue proliferates even more to prepare the lactation. During menopause, fatty replacement of epithelium and stroma takes place. These versatile changes open up various chances for the genesis of cancer cells.

Figure 1.1 shows the basic anatomy of a female breast. The lobules making up the glandular tissue produce milk, which is transported to the nipple by the ducts. Fibrous tissue (ligaments) and fat are the main factors determining breast size and shape and hold the other tissues in place. If there is a high amount of fibrous or glandular (fibroglandular) tissue, a breast is considered as dense in contrast to a mainly fatty breast. Density plays an important role for the choice of suitable breast imaging modalities and is considered as risk factor for breast cancer development (McCormack & dos Santos Silva 2006) as will be discussed in the next sections.

1.1 Breast Cancer Screening

Breast cancer is the most common cancer that affects women, with 494,000 new cases diagnosed in the EU in 2012 and 143,000 women dying from the disease (Ferlay et al. 2013). While causes remain largely unknown, incidence is still increasing. Currently, approximately one in eight women develops breast cancer during her lifetime. It is generally accepted that early detection of breast cancer improves therapy outcomes (Etzioni et al. 2003), and population-based breast cancer screening programmes using X-ray mammography have been shown to reduce mortality: Schopper & Wolf (2009) reported breast cancer mortality reductions between 24 % and 48 % in ten countries among women having attended at least one screening session. Early detection of tumours allows for more effective and at the same time less radical treatment options maintaining the quality of life of these women.

Despite these clear benefits for women attending screening programs, still a substantial number of women die from breast cancer, even though they perfectly complied with screening protocols. Several studies concluded that approximately 30 % of breast cancers were detected in-between screenings (interval cancers) (Törnberg et al. 2010; Bennett et al. 2011). Furthermore, Otten et al. (2005) found that 25 % to 30 % of screen-detected cancers were retrospectively detectable on previous mammograms, and thus could have been detected earlier. The sensitivity of mammography is highly variable, ranging from 98 % for

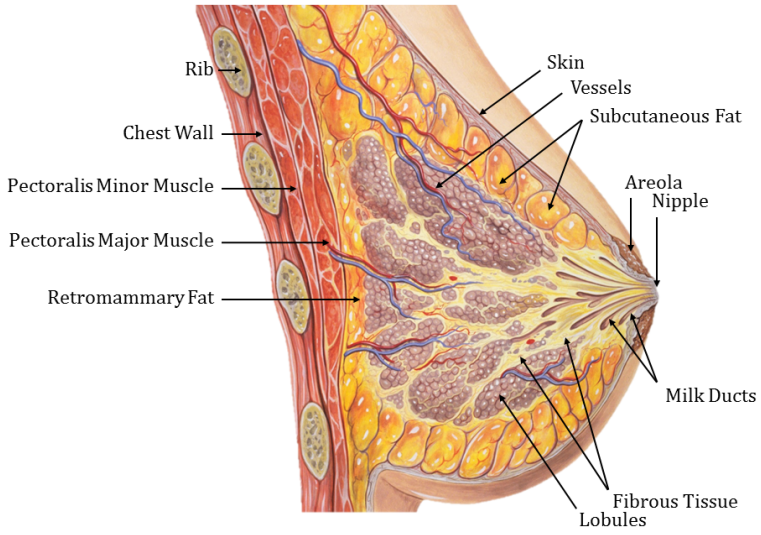


Figure 1.1: Anatomy of the female breast (Graphic by Patrick J. Lynch (illustrator) and C. Carl Jaffe (MD, cardiologist))

women with low amount of fibroglandular tissue to 36 % for women with dense breast parenchyma (Kolb et al. 2002). This reduced sensitivity is due to the fact that dense tissue has the same X-ray attenuation properties as tumours and thus both show equally bright on mammographic images. This causes tumours to remain masked for radiologists and thus breast cancer to remain undetected. As approximately 35 % to 40 % of the screening population have dense breasts, the huge impact of this limitation of the currently applied one-size-fits-all approach is evident. Furthermore, breast density is after age and the rare BRCA (BReast CAncer) gene mutation the third strongest breast cancer risk factor known to date: Women with dense breasts, i.e. with more than 50 % fibroglandular tissue, have been shown to have a three- to six-fold increased risk of developing breast cancer compared to those with little or no dense tissue (McCormack & dos Santos Silva 2006).

Therefore, stratified screening programmes tailored to risk assessment based on family history, age, genetic profiles, and breast density, are proposed (Drukteinis et al. 2013). The main parameters of personalised screening approaches are the different modalities that are available for breast imaging as well as the screening intervals. The imaging methods have to be employed wisely in order to increase sensitivity and specificity while minimizing cost and radiation exposure. Three important breast imaging methods are described in the following sections.

1.2 Breast Imaging

Breast cancer screening programmes do not serve as prevention—a misleading term that is often used in this context—but as early detection of lesions or micro-calcifications in the breast as a predictor of a tumour. Generally, these indications can be detected on standard X-ray mammograms. However, women at cumulative lifetime breast cancer risk of more than 20 % to 25 %, mostly due to BRCA gene mutation, are recommended to get a breast MRI examination. This encompassed roughly 1 % of women. For women with dense breast tissue, ultrasound is under consideration as adjunct to X-ray mammography, a combination of techniques that has recently yielded promising results in a Japanese randomized trial (Ohuchi et al. 2015).

1.2.1 X-Ray Mammography

In 1895, Wilhelm Conrad Röntgen discovered that the radiation he produced in a cathode tube was able to pass through matter and cast object specific shadows on a film. Only one year later, X-radiation, as he called the previously unknown "invisible" light, was used clinically to examine bone fractures or gunshot wounds.

According to the Bohr model, X-rays are generated if an electron from a higher atomic shell jumps over into a free position of an inner shell. The discrete difference of energy is released as a photon and characteristic for the element. To produce free positions in the inner shells, a target material, i.e. molybdenum for breast imaging, is shot with accelerated electrons that interact with the nuclei and the shells of the target atoms, producing bremsstrahlung (retardation radiation) as well as characteristic radiation, respectively. Technically, X-rays are produced in an X-ray tube where a tungsten filament, i.e. the cathode, is heated at a voltage of 10 V and a current of 10 A such that free electrons are emitted (thermionic emission). They are accelerated in vacuum towards the anode at a voltage of 30 kV to 250 kV and a current of a few 100 mA. Arriving at the anode, the electrons are decelerated by interactions with 1) K-shell electrons producing characteristic X-rays of energies around 17 keV for breast imaging, 2) nuclei causing bremsstrahlung, which makes up for the main part of produced radiation, and 3) outer shell electrons generating a line spectrum. Only 1 % of the electron energy is converted into X-ray production, whereas the rest is lost in heat by electron-electron collisions.

When X-radiation passes through tissue, the photons mainly interact with the electrons resulting in scattering and absorption. The exponential decrease of their radiation intensity I is described by the material-specific attenuation coefficient μ and depends on the thickness d of the imaged material. In a first approximation, it is $I = I_0 \exp(-\mu d)$. μ incorporates the different interactions that can appear: For photons with energies above 1022 keV, electron-positron-pair production can appear. Scattering is described by the Compton Effect, which lowers the signal to noise ratio since the scattered photon usually travels in a changed direction. Photoelectric absorption is the most likely reaction of low energy (some keV) photons and depends on the atomic number Z of the passed material. Thus, the differences in electron density of the traversed tissue are the basis of a contrasted image. As a side effect of the above described

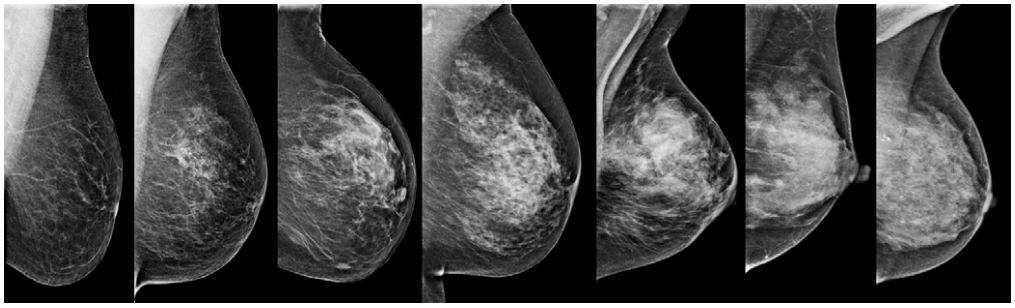


Figure 1.2: Examples of breast density patterns in X-ray mammograms (medio-lateral view), with overall density increasing from left to right

interactions secondary electrons are released, i.e. X-rays are ionizing radiation causing damage to living cells, which is a shortcoming of this technique. For mammography, the photon energies are particularly low yielding a low transmission, which results in a good contrast but also in relatively high skin doses.

X-ray imaging produces summation images integrating all attenuation coefficients along the path of the photons, such that the order of overlapping objects cannot be determined. In standard screening mammography examinations, two X-ray images per breast are therefore acquired from different views, generally a cranio-caudal (CC) and a medio-lateral (MLO) view. The various appearances of breasts with different density can be seen in figure 1.2.

1.2.2 Magnetic Resonance Imaging of the Breast

After Isidor Rabi had already shown the ability of nuclei to absorb high frequency electro-magnetic pulses using a molecular beam in a vacuum in 1938, it would not be until late 1945 that Felix Bloch and Edward Purcell would demonstrate nuclear magnetic resonance in condensed matter. Only several decades later, in 1973, Paul Lauterbur acquired a magnetic resonance image differentiating between normal and heavy water.

All nuclei that have an uneven number of nucleons and some nuclei that have uneven amounts of protons and neutrons possess a nuclear spin, which in turn is associated to a magnetic moment $\vec{\mu}$. Due to the nuclei's thermal energy at normal temperatures, the magnetic moments in a sample are distributed isotropically, but in a static magnetic field $\vec{B_0}$, the spins align parallel (spin-up) or anti-parallel (spin-down) to $\vec{B_0}$. The spin-down energy level is higher than that of spin-up and the population of these levels is described by the Boltzmann distribution. This results in a small surplus of spins that are oriented parallel to the static magnetic field and yields a measurable longitudinal net magnetisation $\vec{M}$.

In magnetic resonance imaging (MRI), $\vec{M}$ is excited by a radio frequency (RF) pulse of the Larmor frequency $\omega_L = \gamma B(x, y, z)$ being proportional to the local static magnetic field $B(x, y, z)$ and the gyromagnetic ratio γ, which is characteristic for each isotope (e.g. $\gamma(^1\mathrm{H}) = 42.6\,\mathrm{MHz/T}$). $B(x, y, z)$ is composed of $\vec{B_0}$ and three additional (orthogonal) gradient magnetic fields that are used for spatial encoding. The RF pulse forces the net magnetisation $\vec{M}$ to flip away from its original orientation and precess around $\vec{B_0}$, causing a transversal mag-

netisation that induces a measurable alternating voltage in a receive coil. The measured frequency depends on the local magnetic field whereas the amplitude encodes the strength of the transverse magnetisation, which in turn depends on the proton density and tissue-specific parameters.

The system is restoring its equilibrium state by exponential relaxation, i.e. the free induction decay (FID) of the MR signal. During the relaxation process, two independent effects superimpose: the spin-lattice interaction with relaxation time T_1 encodes the recurrence of the longitudinal magnetisation, and the spin-spin interaction with relaxation time T_2 describes the de-phasing of the spins. Since the human body consists of about 70 % water, the focus of MRI is on the hydrogen atoms that are bound in water molecules. The measured relaxation times are specific for the molecular structure and environment of the bound hydrogen atoms, allowing to differentiate between tissue types.

In dynamic contrast enhanced MRI (DCE-MRI), a paramagnetic contrast agent (CA), e.g. the Gadolinium-based Gd-DTPA, is injected intravenously before a time series of MR images is acquired to visualise the enhancement characteristics of the imaged tissues. Gadolinium causes the relaxation time to decrease resulting in images of higher contrast. At the first pass of the CA through the blood circulation, which is typically 45 s to 60 s after injection, it is predominantly intra-vascular allowing evaluation of perfusion, i.e. blood flow per unit volume. During the subsequent 2 to 10 minutes, the diffusion-based passage of CA into the extra-vascular (and extra-cellular) space is increased, and imaging during this delayed phase enables measurement of vascular permeability. For tumour tissue, the CA enhancement curve is changed compared to healthy tissue. Common DCE-MRI sequences for the breast focus on high spatial resolution allowing for detailed morphologic evaluation of lesions. One reference image is acquired before CA administration, followed by up to four images showing the maximum enhancement as well as the late behaviour of CA uptake, i.e. increasing enhancement, a plateau, or wash-out. New ultra-fast view-sharing MRI protocols allow imaging at high temporal resolution while retaining a high spatial resolution (Laub & Kroeker 2006). Using these sequences, a volumetric image of the breast can be acquired within 5 s enabling an accurate description of CA kinetics while at the same time maintaining detailed information of lesion morphology (Platel et al. 2014). First attempts to adjust the spatial and temporal resolution dynamically depending on the current CA behaviour have been taken by Kompan (2015).

Sample T_1 weighted breast MR images are described in figure 1.3. The maximum intensity projection (MIP) images are based on the subtraction of the pre-contrast image from (in this case) the first post-contrast image.

DCE-MRI of the breast is the most sensitive alternative for the detection of breast cancer (Mann et al. 2007; Lehman et al. 2005). Unlike X-ray mammography, MRI is unaffected by breast density and does not use ionizing radiation. However, the application of MRI for breast cancer screening in the general population is not practical because of its high costs, limited availability, use of contrast agents and variable specificity. Contrast agent-less MRI sequences such as ASL (Arterial Spin Labelling) (Buchbender et al. 2013) or HiSS (High Spectral and Spatial resolution) (Medved et al. 2011) are currently under development for breast cancer imaging. First results especially for other entities, e.g. the

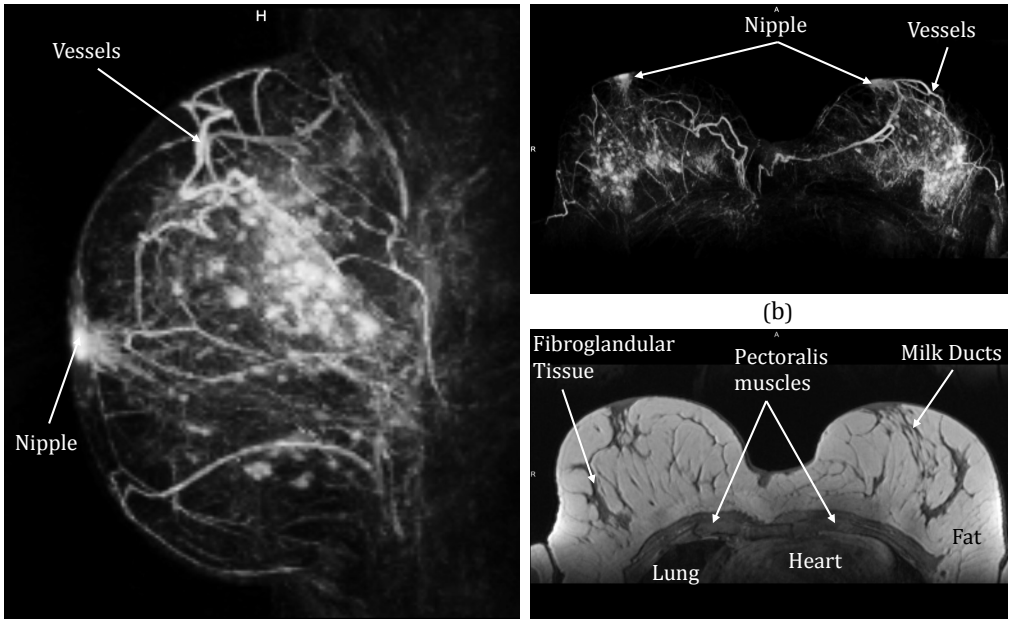


Figure 1.3: Sample breast DCE-MR images of the breast: (a) shows the sagittal maximum intensity projection (MIP) of the difference image of one breast, (b) is the transversal MIP, and (c) is a transversal slice of the T1 weighted image of the same breast.

head, are promising (Günther 2014), but those sequences are not yet ready to replace DCE-MRI for the breast.

1.2.3 Breast Ultrasound

An affordable and safe alternative to MRI and X-ray mammography is the use of ultrasound for breast examination, since no radiation or intravenous contrast agent is required. Large screening studies have shown that (classical) hand-held ultrasound has a high sensitivity for breast cancer, particularly for aggressive forms of cancer (Berg 2008). Apart from that, ultrasound image quality is unaffected by breast density and it has been reported to double the detection rates in dense breasts when used in combination with mammography (Kelly et al. 2010b). The disadvantage of time-consuming examination that has to be performed by a highly qualified radiologist has recently been technically resolved with the introduction of automated 3D breast ultrasound (ABUS) (Brem et al. 2015; Drukteinis et al. 2013), which can be acquired by trained medical technicians and analysed later by radiologists. Furthermore, this technique provides standardized volumetric images that can be compared to relevant prior examinations.

Physical Basics of Ultrasound

Medical image acquisition with ultrasound is based on the reflection and backscattering of insonated acoustic waves. In soft tissue, ultrasound waves can be considered as longitudinal waves, i.e. a propagation of compression and decompression in a medium manifested as a particle vibration along the propagation direction.

At transitions between different matters, e.g. muscle and fat, ultrasound waves are partly reflected and partly transmitted (refracted if the surface is not hit perpendicularly). Thus, the echo runtime indicates the distance between the transducer and the tissue border whereas the signal intensity contains information on the material properties. The relevant property of matter (material constant) is described by the acoustic impedance $Z = c_s \cdot \rho$ (for harmonic waves), which combines the density ρ and the speed of sound c_s in the particular material. If an acoustic wave passes the interface of two materials (see figure 1.4) with Z_1 and Z_2 with entrance angle θ_1, Snell's law holds true and yields the exit (refraction) angle θ_2 (relative to the normal of the interface)

$$\sin\theta_2 = \sin\theta_1 \frac{c_2}{c_1} \tag{1.1}$$

The reflection r describing the ratio between the reflected amplitude A_R and the incident amplitude A_0 is then given by

$$r(\theta_1, \theta_2) = \frac{A_R}{A_0} = \frac{Z_2 \cos\theta_1 - Z_1 \cos\theta_2}{Z_2 \cos\theta_1 + Z_1 \cos\theta_2} \tag{1.2}$$

If the incidence of the sound wave is perpendicular, this turns into

$$r(0,0) = \frac{Z_2 - Z_1}{Z_2 + Z_1}. \tag{1.3}$$

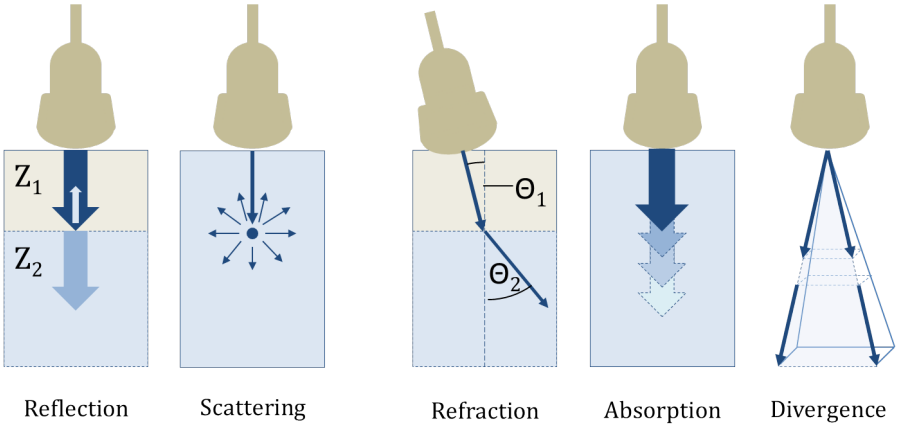


Figure 1.4: Different aspects of acoustic attenuation when an ultrasound wave traverses tissue. The ultrasound image is reconstructed from reflected and scattered signals. (Figure similar to a figure in Delorme & Debus (1998).)

Table 1.1: Speed of sound in different tissues. The acoustic impedance $Z = c_s \cdot \rho$ is computed from density ρ and speed of sound c_s. All numbers refer to body temperature of 37 °C (Numbers from Deserno (2011)).

Material	c_s *in* $\mathrm{m\,s^{-1}}$	ρ *in* $10^3\,\mathrm{kg\,m^{-3}}$	Z *in* $10^6\,\mathrm{kg\,m^{-2}\,s^{-1}}$
Bone	3600	1.70	6.12
Marrow	1700	0.97	1.65
Blood	1570	1.02	1.61
Muscle	1568	1.04	1.63
Water	1540	0.99	1.53
Fat	1400	0.97	1.36
Air	340	1.20×10^{-3}	4.08×10^{-4}

If $Z_2 \approx Z_1$, the reflection r becomes 0, i.e. the acoustic wave passes the tissue interface without being reflected. The cases $Z_2 \ll Z_1$ and $Z_2 \gg Z_1$ result in $r = -1$ and $r = 1$, respectively, which describe a total reflection with or without phase shift of 180°.

The typical values of acoustic impedance listed in table 1.1 yield a very strong reflection for interfaces from air to soft tissue as well as from soft tissue to bone, whereas the reflection ratio is small between different soft tissues. Therefore, "impedance-matched" (water-based) contact gel must be used for air-free coupling of ultrasound waves to the human body. Furthermore, it is almost impossible to acquire sonographic views behind bony structures or air filled organs.

Whereas the above described specular reflection visualizes flat smooth interfaces as the diaphragm or walls of major vessels, scattering appears on small objects, e.g. cells, and provides information on the inner structure of tissues. Scattering on a point source generates a spherical wave, which depends on the diameter a of the scatterer relative to the wave length λ: If $a \gg \lambda$, the scattering is geometric leading to a strong (diffuse) reflection as for example on small arteries and bile ducts of the liver, which appear brighter (hyper-echoic). If $a \approx \lambda$,

the scattering is stochastic and directional, as it occurs, e.g., in liver tissue and makes up for about 20 % of the total acoustic attenuation. Rayleigh scattering is relatively weak and occurs if $a \ll \lambda$, as for example in blood. Consequently, the interior of vessels is normally dark (anechoic).

Due to the coherent character of ultrasound waves, the reflected signals interfere with each other at the transducer aperture. Depending on the relative phase of the scattered waveforms, they can add constructively or destructively producing the "speckle" pattern, which is inherent to ultrasound images and potentially decreases image quality. Although the pattern is random, it is more or less constant over time and, thus, can be exploited, e.g., for motion tracking purposes (Notomi et al. 2005). Compound scanning, i.e. averaging over several scans acquired by differently steered ultrasound waves (Hoskins et al. 2010, p. 38), can decrease the speckle pattern and acoustic noise.

Apart from reflection on interfaces of different impedance Z, refraction at positions of changing speed of sound c_s, and scattering, absorption and divergence also contribute to the attenuation of an ultrasound wave (see figure 1.4). Absorption is the transition of ultrasound energy into heat and results in an exponential decay of the ultrasound pressure $p(x) = p_0 \exp(-\alpha x)$. The absorption coefficient α depends on the frequency as $\alpha = \alpha_0 f^m$ where α_0 and m (≈ 1.1 to 1.3) have to be determined experimentally for a particular material. Divergence is the attenuation of ultrasound intensity I due to the spread of the beam and can be expressed by the inverse square law: $I \propto 1/r^2$ with r being the distance to the source.

As mentioned above, the attenuation increases with increasing ultrasound frequency f, but as well does the spatial resolution. The capability to resolve (or to differentiate) two objects that are close to each other depends on the wavelength λ that is coupled to the frequency as $\lambda \cdot f = c_s$. Whereas the axial resolution (along the propagation direction of the sound waves) is approximately twice the wavelength λ, the lateral resolution (orthogonal to the propagation) is only four to five times the wavelength (Delorme & Debus 1998). Typical diagnostic ultrasound devices operate in the frequency range of 1.5 MHz to 20 MHz trading-off image depth and spatial resolution as shown in table 1.2.

Imaging Techniques

Technical implementation of ultrasound imaging is enabled by piezoelectricity that was discovered in 1880 by the French physicists Jacques and Pierre Curie. The piezoelectric effect describes the behaviour of certain solid materials that respond to mechanical stress with separation of charge resulting in a voltage, and vice versa. For ultrasound imaging, an array of piezoelectric crystals ("sub-aperture") is used to convert electrical signals into mechanical deformation that is coupled into the body (see figure 1.5). The reflections of the ultrasound pressure wave are echoed back to this (or another) sub-aperture transforming now the deformation into an electrical signal that can be measured. The runtime of the signal defines the spatial origin of the echo. Therefore, a constant speed of sound of 1540 m/s is assumed, which is a sufficient approximation for soft tissue imaging (see table 1.1), but still leads to image artefacts (see section 1.3).

There are different ways of processing and interpreting the electrical signal of the pulse echo. In A-mode (Amplitude) imaging, the amplitude (envelope) of the

Table 1.2: Penetration depth, spatial resolution of different ultrasound frequencies as well as typical organs that are imaged at (approximately) these frequencies. (Table similar to a table in Postema & Attenborough (2011, p. 157))

Frequency in MHz	*Wavelength in* mm	*Penetration depth in* cm	*Lateral resolution in* mm	*Axial resolution in* mm	*Clinical application*
2	0.78	25	3.0	0.80	Liver, Fetus, Heart
3.5	0.44	14	1.7	0.50	Kidney
5.0	0.31	10	1.2	0.35	Brain
7.5	0.21	6.7	0.8	0.25	Thyroid, Superficial Vessels
10.0	0.16	5.0	0.6	0.20	Prostate, Breast
15.0	0.10	3.3	0.4	0.15	Breast
21.0	0.09	1.1	0.36	0.13	Eye, Skin

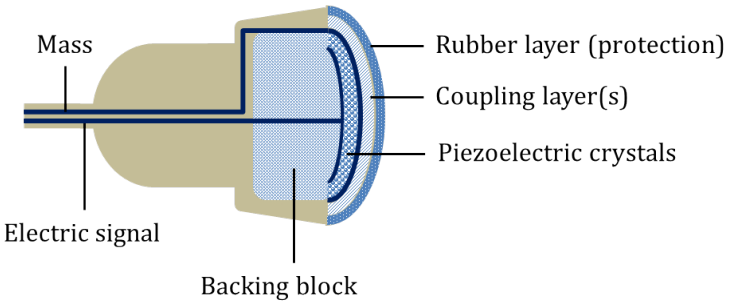


Figure 1.5: Simplified schematic construction of an ultrasound probe. The electrodes apply an alternating potential difference to make the piezo crystals contract, and receive the electrical signal if the crystals are distorted by the incoming echo. The backing block decreases ringing of the piezo elements, i.e. shortens the pulse length and increases the axial resolution. The coupling layers reduce the bridge the difference in acoustic impedance between crystals and contact fluid that is applied to the skin.

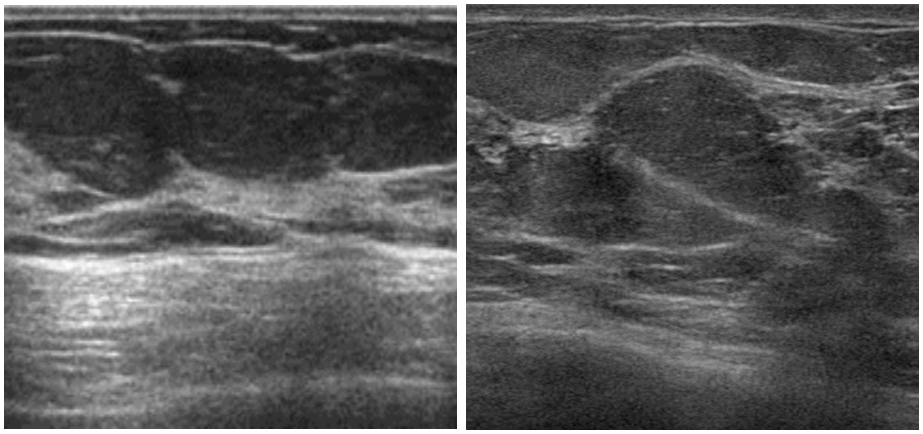
returning signal is displayed as a function of depth for a single line originating from one transducer element (piezo crystal). In B-mode (Brightness) imaging, the amplitude of the echo is coded in grey scales, such that the combination of several adjacent lines results in a 2D image. In M-mode (Motion) imaging, a single B-mode line is recorded over time by emitting and evaluating pulses in quick succession. This technique allows to see real-time motion, e.g., of a cardiac valve. Doppler ultrasound imaging can be used to visualize blood flow by employing the Doppler Effect that describes the change of frequency if sound waves are reflected from moving objects.

As mentioned above, a 2D image is made up of several adjacent 1D lines that encode the returning ultrasound signal intensity and runtime in grey scales. Depending on the intended application, different arrangements of piezo crystals are used clinically. The simplest configuration is a linear parallel array of traditionally 64 to 256 transducer elements out of which a sub-aperture is excited to transmit and receive a focused bundle of ultrasound waves. The subsequent excitation of adjacent groups of piezo crystals produces a rectangular B-mode image. Due to the finite speed of sound, the chosen image depth and the number of lines that are acquired for one image limit the maximum frame rate. It has to be noted that modern transducers can have much more single elements and that it is even possible to excite and read out all elements at once, increasing the frame rate significantly and enabling real-time 3D ultrasound (Bercoff 2011). Where tiny transducers are needed, e.g. for endo-rectal or gynaecological imaging, mechanical scanners that rotate or "wobble" a single transducer element are applied. These so-called sector scanners have a small aperture and produce a fan-shaped acoustic window that can also exploit the gap between two ribs for abdominal (inter-costal) imaging. A convex arrangement of piezo crystals can be found in curved arrays combining the advantages of both above described scanner types. It provides a good (wide) image in the near-field that even gets wider with increasing depth. However, it also gets coarser due to the divergence of the single lines.

Various effects cause a continuous attenuation of the ultrasound beam passing through tissue. It is however desirable that the same tissues are represented by the same pixel intensities on the resulting ultrasound image. Therefore, a time gain compensation (TGC) is introduced to amplify the returning echoes: The later a signal arrives at the transducer, the deeper—or rather longer—it has travelled into the body before being reflected, and the more it needs to be intensified. The TGC function can be set manually on most ultrasound scanners since the ideal setting depends on the imaged tissue structure.

Tissue Harmonic Imaging

Since ultrasound often suffers from noise and artefacts due to reverberation or aberration, modern devices use techniques that are more sophisticated than basic B-mode imaging. One example is Tissue Harmonic Imaging (THI), which deploys non-linear effects of sound propagation and reflection. As a sound wave travels through tissue, it produces regions of higher and lower pressure, which in turn influence the speed of sound c_s and consequently distort the original sine-like pulse. Since tissue is not linearly elastic, it contracts less than it expands. At higher pressure, c_s increases such that the peaks get pulled forward



(a) **(b)**

Figure 1.6: Two breast ultrasound images acquired (a) without and (b) with Tissue Harmonic Imaging. The image in (a) is clearly more affected by noise than the image in (b). Note that the images are not showing the same patient.

as the sound wave passes through tissue. The opposite effect applies to regions of lower pressure causing the trough of the signal to travel slower. This non-linear propagation results in an asymmetrical, saw tooth shaped wave, which is physically equivalent to a signal containing not only the bandwidth of fundamental frequencies of the original ultrasound pulse but also higher harmonic frequency components. The harmonics have shorter wavelengths as well as a narrower beam profile since they are mainly produced along the strong central ultrasound beam and not by weak components of the initial signal as scattered echoes or the edge of the transmit beam. This significantly improves the grey scale contrast resolution and reduces artefacts. Furthermore, the harmonic signals are less noisy, i.e. distorted or scattered, because they are produced in the body and, thus, only have to pass through the body wall (or skin fat layer) once. A general haze that overlies the top centimetres of an image in conventional ultrasound images due to reverberations between the transducer and the body wall layers is also eliminated by THI. The amplitude of the harmonic signal is very low and suffers from strong absorption. Several methods to detect the harmonics and to eliminate the unwanted fundamental echoes are available. High pass filters can remove the fundamental frequencies from the received signal. However, this technique requires a transducer that transmits a very narrow bandwidth of frequencies in order to avoid substantial overlap between (the highest) fundamental frequencies and (the lowest) harmonics. In single line pulse inversion, the fundamental and harmonic echoes of one line are recorded before an inverted pulse is applied to the same line. The resulting signal is subtracted from the first one, cancelling the fundamental echoes and sparing the harmonic information. Sample breast ultrasound images acquired with and without THI are shown for comparison in figure 1.6. It has been stated that THI can improve the tumour delineation and tissue differentiation when compared to standard ultrasound imaging (Clevert et al. 2007).

Automated 3D Breast Ultrasound

In automated 3D breast ultrasound (ABUS), a series of 2D ultrasound images covering the breast is acquired automatically by a transducer translating across the breast. Volumetric breast images are then generated by stacking the single slices together. This automation enables technicians to acquire these images, as opposed to hand-held ultrasound images that are attributed to radiologists. Although the real-time feedback inherent to hand-held ultrasound is lost in ABUS imaging, the 3D images promise to include more information at comparable image quality (An et al. 2015).

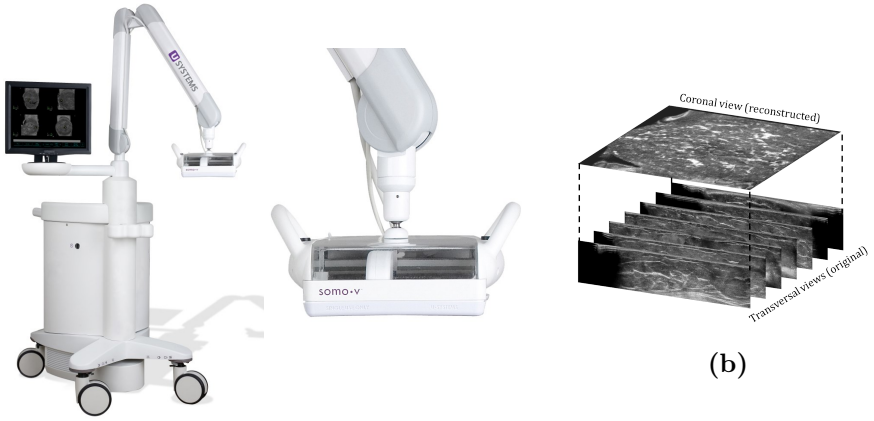
The most common commercially available systems are the ACUSON S2000 ABVS system by Siemens and the somo-v by U-Systems[1] (see figure 1.7a), which both have FDA approval for clinical use.

With these two systems, ABUS images are acquired by a wide linear array ultrasound transducer with more than 700 elements sliding continuously over one breast, which is gently compressed by a dedicated membrane while the patient lies in a supine position. During the sliding motion of the transducer, the ultrasound scanner acquires more than 300 transversal images covering a large segment of the breast. These single slices are stacked to a 3D ultrasound image that can be examined in multi-planar reconstructions (van Zelst et al. 2015) as shown in figure 1.7b. Depending on the size of the breast, up to five views of each breast are acquired. The positioning and compression of the breast are standardized to some extent and include anterior-posterior (AP), lateral (LAT), medial (MED), superior(SUP) or inferior views, the breast being gently pushed in these directions, respectively.

The ACUSON system produces images of a maximum size of $154\,\text{mm} \times 168\,\text{mm} \times 60\,\text{mm}$ as well as a minimum voxel size of $0.21\,\text{mm} \times 0.52\,\text{mm} \times 0.07\,\text{mm}$ in cranio-caudal (head-to-toe), medio-lateral (left-right) and antero-posterior (front-to-back) direction, respectively. Spatial resolution in antero-posterior direction depends on the chosen scanning depth, which is generally adapted to the breast size. The somo-v scanner acquires images with a maximum size of $146\,\text{mm} \times 168\,\text{mm} \times 49\,\text{mm}$ at similar resolution. However, the systems differ in the ultrasound frequencies they provide. Whereas the Siemens transducer can operate at frequencies between 5.0 MHz and 14.0 MHz (adjustable to the breast size), the U-systems device allows choosing between 8.0 MHz and 10.0 MHz. Some anatomical structures are referenced in the ABUS image shown in figure 1.8.

A single ABUS scan takes approximately one minute plus patient positioning and transducer set up. For a complete ABUS acquisition, the typical imaging time per patient is 10 to 20 minutes with additional preparation times of 5 to 10 minutes. Interpretation and reporting time for an experienced radiologist is approximately 7 to 10 minutes per examination (Kelly et al. 2010b; Skaane et al. 2015). Computer aided detection systems, which aim at improving and accelerating the image read, are therefore being developed. Images of consistently high quality are an essential pre-requisite for these tools.

[1] Now marketed as Invenia ABUS by GE Healthcare



(a) **(b)**

Figure 1.7: (a) The somo-v Scanner by U-Systems with the typical linear ABUS scan head sliding over the breast within its frame. (b) The acquired transversal slices are stacked together in multi-planar reconstruction.

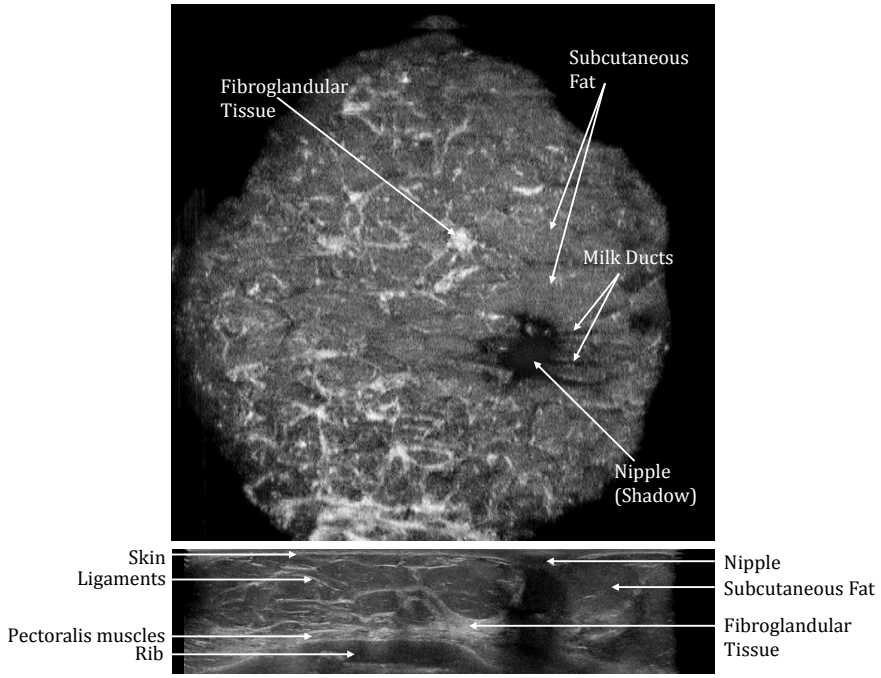


Figure 1.8: Sample coronal (*top*) and transversal (*bottom*) slices of an ABUS image.

1.3 Ultrasound Image Quality

Typical Ultrasound Artefacts

The above mentioned physical aspects of sonography can lead to imaging artefacts: structures that appear in the image for technical reasons but do not reflect anatomical reality. Ultrasound artefacts originate in idealised assumptions of sound expansion as explained below for some examples. It is important to note that ultrasound artefacts indeed may disturb diagnosis, but on the other hand, sometimes help to confirm findings.

The assumption that every pulse echo returning to the transducer has only been reflected once can lead to repetition or mirror artefacts. If the ultrasound waves are reflected back and forth between two successive reflectors, the increasing runtime makes the incoming signal mock a regular stripe pattern resulting from the repeated depiction of the reflectors. These reverberation lines can also appear if the ultrasound signal does not leave the transducer but is reflected within the coupling layers (see description of air artefacts in the next section). Nevertheless, repetition artefacts can be used diagnostically, e.g. to identify metallic OP clips. If there is one strong reflector in the field of view, it is possible that it reflects sound waves to the back side of an object that lies in front of it. This simulates a copy of the object behind the reflector and produces a mirrored version on the image.

Whereas a constant ultrasound velocity of $1540\,\mathrm{m\,s^{-1}}$ is assumed for image generation, c_s can indeed vary considerably in diverse tissue types (see table 1.1). This can produce runtime artefacts showing up as deformations especially evident if tissues with extremely different c_s are next to each other, as it is the case when imaging, e.g., the liver through the ribs (trans-costal).

Another idealised assumption is the equal attenuation of ultrasound in all media, which can lead to erroneous signal enhancements as well as to acoustic shadows. The latter one appears if a structure does not conduct the ultrasound to deeper layers in tissue, i.e. the sound is reflected totally at abrupt changes of acoustic impedance or it is absorbed completely. Bones or kidney stones are typical examples for such strong reflectors as can be seen in breast ultrasound images where the ribs cause severe acoustic shadows (see figure 1.9a). Posterior enhancement is caused by disproportionate time gain compensation and describes the fact that tissue behind structures with very low ultrasound attenuation appears brighter. Due to the reduced attenuation of, e.g., a cyst, the pixel intensity behind the cyst is overestimated and appears higher than in the surrounding areas. This behaviour is an important criteria for detecting cysts (see figure 1.9b).

Finally, it is assumed that the ultrasound beam is focused sharply, but actually the beam has a distinct width causing blurring or partial volume artefacts. Due to the finite extension of the beam, pulse echoes that originate in a volume are projected onto one pixel in the image. If the imaged structures are of similar size as the beam width, a mixture of these different tissues will be displayed.

Most of the above described artefacts are caused by the physical properties of ultrasound interacting with different media and cannot be altered or avoided, e.g., by a change of scanning parameters. Further image quality aspects that play a role especially in ABUS imaging are introduced in the following section.

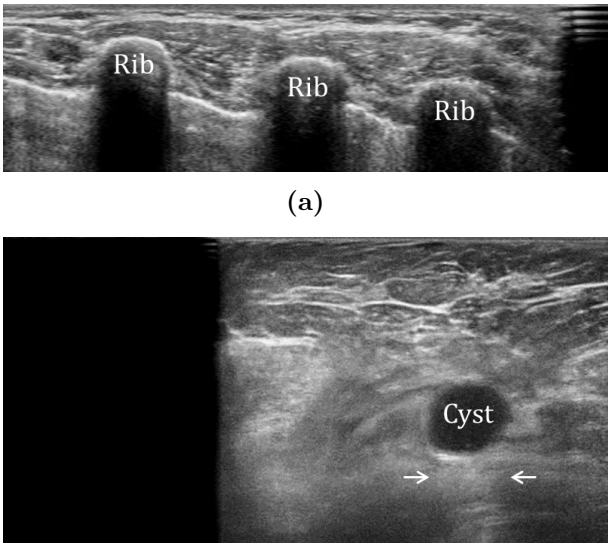


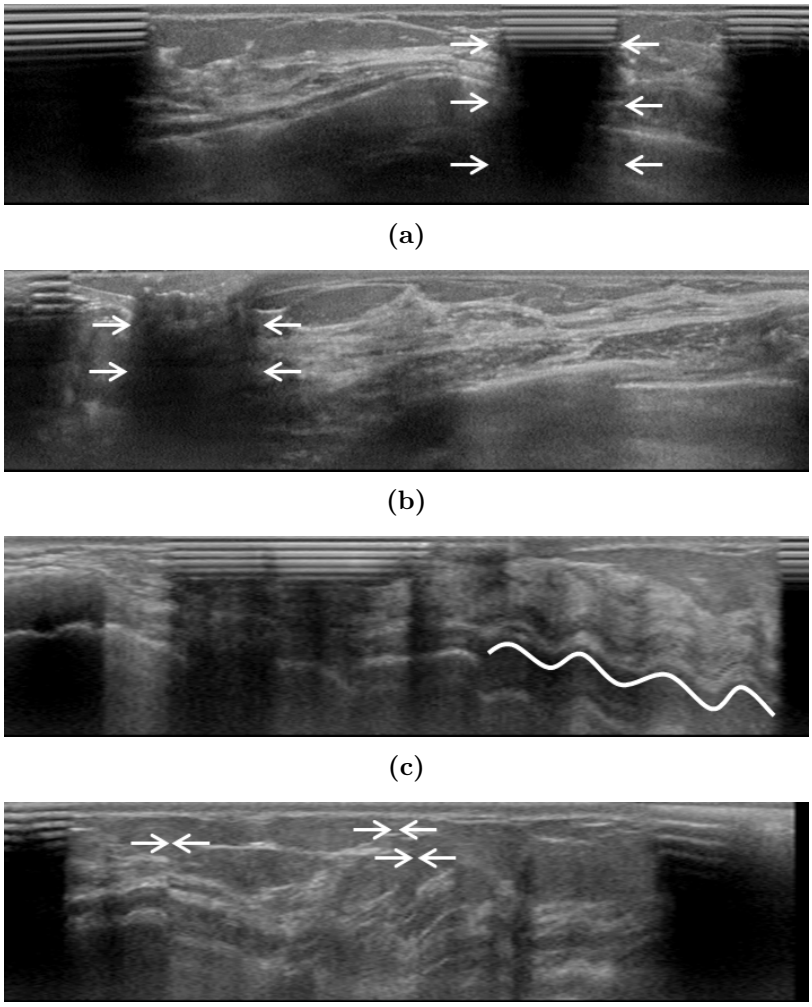
Figure 1.9: Artefacts in (AB)US imaging. The sagittal reconstruction in (a) has three strong acoustic shadows caused by the ribs. In (b), an anechoic cyst causes posterior signal enhancement (arrows) due to a very low attenuation within the cyst and consequently incorrect TGC.

Particular ABUS Image Quality Aspects

An important factor and unique feature of hand-held ultrasound imaging is the user interaction with the real-time image. Many artefacts can be resolved or explained by a slight movement or tilting of the ultrasound probe. In ABUS imaging, this interaction and real-time feedback is completely lost. Therefore, a standardized acquisition protocol ensuring high quality reporting, complete coverage of breasts and accurate temporal comparison of prior to current ABUS images is of highest importance in ABUS imaging. Although the image acquisition process is automated to a high degree, it still depends on the experience and training of the operating technician. The standard ultrasound parameters as scanning depth, contrast, and level must be adjusted. The transducer needs to be placed using the correct pressure and there must be a sufficient amount of contact fluid covering the whole breast in order to avoid imaging artefacts, which can either mimic or obscure pathology causing patient recalls. The most relevant image quality aspects of ABUS are described in the following.

A common artefact in ultrasound imaging are shadows (air artefacts) caused by a lack of contact gel, which is needed to couple the sound waves properly into the body. If this coupling is not provided, the high impedance change prevents the signal from leaving the transducer. Instead, the sound waves are reflected back and forth within the transducer layers causing a characteristic stripe pattern (reverberations) and a deep acoustic shadow on the image as shown in figure 1.10a.

Major concerns in ultrasound breast images are the nipple and the ducts. On the one hand, the nipple is an important landmark which helps, e.g., describing lesion locations. On the other hand, the nipple can cause severe shadows in an



(a)

(b)

(c)

(d)

Figure 1.10: Artefacts in ABUS imaging. Air artefact (a) and nipple shadow (b) in (orignal) transversal view. Wavy pattern (c) and discontinuities (d) in (reconstructed) sagittal view.

ultrasound image due to entrapped air and lack of contact between transducer and skin as shown in figure 1.10b. The ducts may also be filled with air and, thus, can produce elongated shadow regions along their axes. These shadows may cover important structures in the breast image and hinder a solid diagnosis.

Eventually, the success of the imaging process also depends on the cooperation of the patient. The acquisition of a single ABUS volume consisting of 318 2D transversal slices takes approximately 1 min. If the woman is breathing strongly, talking or coughing during the examination, the volumetric image will show a wavy pattern in the reconstructed image plane (see figure 1.10c).

As the volumetric image data set is reconstructed from several 2D ultrasound image slices, which are collected one after another by the transducer scanning over the breast, it is crucial that the transducer moves smoothly at constant velocity through its frame. If, however, the pressure of the transducer on the skin is too high or if there are hard tissue structures in the breast, the transducer motion may be hampered. In consequence, there will be discontinuities between the lines of the reconstructed images (see figure 1.10d).

Furthermore, the position of the nipple relative to the rest of the breast in the image is a quality aspect that should be considered. The nipple being too close to the edge of the (laterally compressed) breast might constrain the view of the radiologist on important areas and induce uncertainty about the true contour of the breast.

The shapes and sizes of female breasts vary strongly among different women, which results in slightly different shapes and contours in a breast image. Nevertheless, on an ABUS image all breasts should show a smooth and roundish contour line and more or less fill the image volume. An irregular breast shape on an ABUS image might indicate improper patient positioning causing skin folds, air cavities or even completely omitted regions of the breast.

1.4 Image Processing

Due to the increasing use of digitization of medical images, digital image processing is gaining importance in health care, such that the entire spectrum of digital image processing is now applicable to medicine (Deserno 2011). Generally, digital image processing covers four major areas: 1) image formation, 2) image visualization, 3) image analysis, and 4) image management. Whereas the first two aspects deal with the steps from acquisition to display of an optimized output of the image, the latter one refers to data storage, communication and transmission. The focus of this work is on the third one—image analysis—which includes quantitative measurements as well as abstract interpretations of biomedical images. An important factor for this analysis is a priori knowledge on the content and nature of the images, which must be implemented to the algorithms on a high level of abstraction. Due to the complexity of biomedical images, formulation of medical a priori knowledge is a challenging task. The discrepancy between the cognitive interpretation of a medical image by a physician and the simple representation of this image that is used by computer programs is called the *semantic gap* (Smeulders et al. 2000). The heterogeneity of medical images, the unknown delineation of objects, and the required robustness of algorithms are the three main aspects of medical imaging that hinder bridging this gap. Nevertheless, medical image analysis is a wide field of research that has already translated many ideas to clinical applications, e.g. computer aided detection (CAD) algorithms supporting radiologists in the interpretation of mammograms (first product approved by the U.S. Food and Drug Administration in 1998). Some basic image analysis tools that were used throughout this work are described in the following.

1.4.1 Otsu's Threshold

A very common and important task in image processing is separation of image foreground from background. In an ideal case, the histogram of an image has a deep and sharp valley between two peaks that represent object and background, respectively. Then, this valley can easily be chosen as threshold value. In real images however, this is seldom the case, e.g. due to noise roughening the whole histogram. Otsu (1975) proposed a threshold selection method that maximized the variance between foreground and background. Assuming that a threshold value k for the pixel intensity dichotomizes the image into two classes C_0 and C_1 (background and object) he computed the between-class variance as

$$\sigma_B^2 = \omega_0\omega_1(\mu_1 - \mu_0)^2 \tag{1.4}$$

where ω_0 and ω_1 are the probabilities of class occurrence and μ_0 and μ_1 the class mean intensity levels, respectively. Given a picture of N pixels and L grey-levels with n_i pixels at level $i \in 1, ..., L$, normalisation yields a probability distribution:

$$p_i = n_i/N, \quad p_i \geq 0, \quad \sum_{i=1}^{L} p_i = 1. \tag{1.5}$$

Then, the elements of equation 1.4 depend on k as follows:

$$\omega_0 = \sum_{i=1}^{k} p_i = \omega(k) \tag{1.6}$$

$$\omega_1 = \sum_{i=k+1}^{N} p_i = 1 - \omega(k) \tag{1.7}$$

$$\mu_0 = \sum_{i=1}^{k} i p_i / \omega_0 = \mu(k)/\omega(k) \tag{1.8}$$

$$\mu_1 = \sum_{i=k}^{N} i p_i / \omega_1 = \frac{\mu(L) - \mu(k)}{1 - \omega(k)}, \tag{1.9}$$

where $\omega(k)$ and $\mu(k)$ are the zeroth and first order cumulative moments of the histogram up to the kth level, respectively, and $\mu(L)$ is the total mean level of the original picture. Thus, the problem to be maximized over k is

$$\sigma_B^2(k) = \frac{[\mu_T \omega(k) - \mu(k)]^2}{\omega(k)\,[1 - \omega(k)]}. \tag{1.10}$$

The criterion measure takes a minimum value of zero if all pixels are either C_0 or C_1, and otherwise takes a positive and bounded value, i.e. the maximum always exists. It is noticeable that Otsu's method only uses zeroth and first order statistics making the computation relatively simple. Furthermore, an extension to multi-thresholding exists.

As can be seen in figure 1.11, this single threshold is not always sufficient to get a meaningful separation between foreground and background. But for ABUS images, Otsu's method yields a binary image of the breast that is clearly separated from the background. Small holes in the foreground can be eliminated by morphological operations as described in the next section.

1.4.2 Basic Morphological Operations

Mathematical morphology comprises many techniques for the analysis and processing of geometrical structures, which are commonly applied to digital images. The basic morphological operators are erosion, dilation, opening and closing, all of which are shift-invariant and can be applied not only to binary images but also to grey-scale images. In binary morphology, an image is considered as a subset of a Euclidean space $\mathbb{R}^d$ or an integer grid $\mathbb{Z}^d$, for some dimension d. This image is probed, i.e. convolved, by a simple, pre-defined structuring element, which is by itself a binary subset of the space or grid. The structuring element can have the shape, e.g., of a disk of radius r, a $n \times n$ square, or a cross. The four basic operations that are often used in medical image processing are depicted in figure 1.12 for a sample binary image and a circular structuring element. If E is a Euclidean space or an integer grid, A is a binary image in E, and B is the structuring element, dilation is defined as

$$A \oplus B = \bigcup_{b \in B} A_b. \tag{1.11}$$

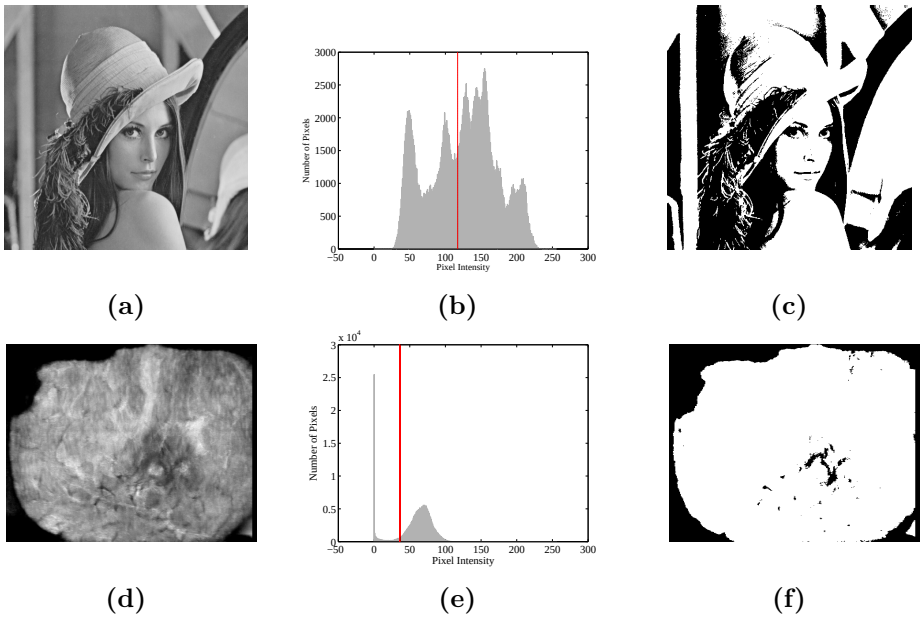


Figure 1.11: Otsu's threshold image filter applied to a sample image (top) and the mean projection of 50 coronal ABUS image slices. (a) and (d) are the original grey scale images, (b) and (e) show the corresponding histogram and Otsu's threshold level (red line), and (c) and (f) are the resulting binary images.

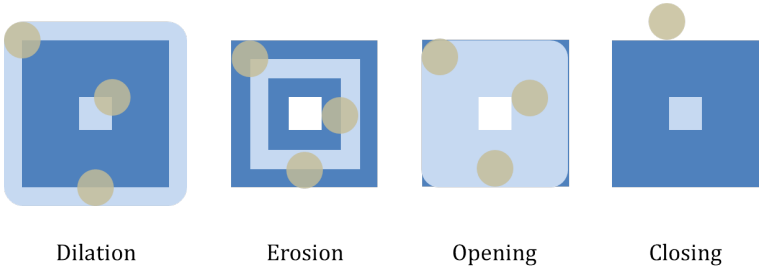


Figure 1.12: Basic morphological operations. The dark blue open square is the image that is dilated, eroded, opened and closed by the circular structuring element yielding the light blue shape.

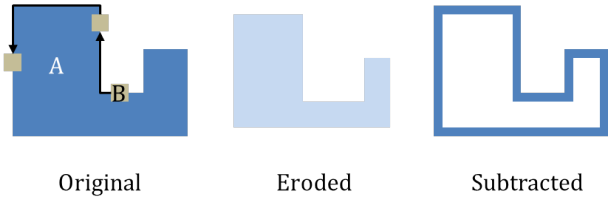


Figure 1.13: Principle of contour extraction algorithm that uses erosion and subtraction. The size of the structuring element B defines the thickness of the boundary (*right*).

Whereas dilation works like a low pass filter enlarging foreground structures, smoothing contours, or even filling small holes, erosion is the contrary operation that makes foreground regions shrink and cancels details that are smaller than the structuring element. Erosion can be written as

$$A \ominus B = \bigcap_{b \in B} A_{-b}. \tag{1.12}$$

Opening is the combination of erosion and dilation,

$$A \circ B = (A \ominus B) \oplus B, \tag{1.13}$$

which keeps all parts of the foreground that the probe image fits in. Opening keeps all points of the object that are covered by the translation of the structuring element along the inner border, i.e. it eliminates protrusions and bridges. Closing is the inverse operation—dilation followed by erosion—

$$A \bullet B = (A \oplus B) \ominus B \tag{1.14}$$

filling gulfs and holes. Closing adds all points from the background to the foreground that cannot inclose completely the structuring element translating along the outer boundaries of the object, i.e. holes within the foreground object are closed if they are smaller than the structuring element.

1.4.3 Boundary Extraction

The extraction of the boundary (or contour) of an object in a binary image is a very useful application of the above described morphological operators. Technically, the boundary of an object A can be obtained by eroding A by a suitable structuring element B and then subtracting the eroded set from A as shown in figure 1.13. The size of the structuring element defines the thickness of the extracted contour. If, for example, B is a square of 3×3 pixels, the contour line will be one pixel thick.

1.4.4 Hole Filling

A hole in a binary image may be defined as a background region surrounded by a connected border of foreground pixels. Whereas a closing operator might already do the job for small holes, this is not an option to erase bigger enclosed background regions since a sufficiently large closing operator would distort the

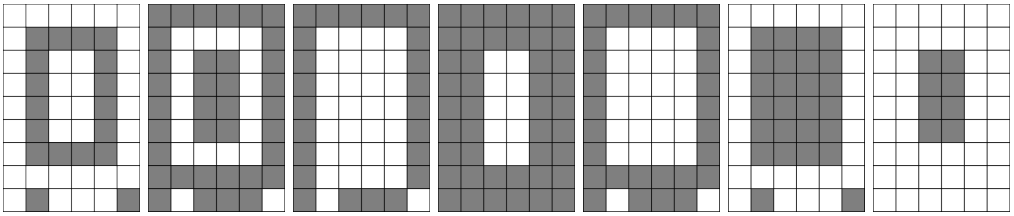


Figure 1.14: Principle of the hole filling algorithm based on morphological reconstruction. The original image I is dilated by the structuring element B being a 3×3 matrix of ones and masked by I^c. The rightmost image shows the isolated filled hole. Dark pixels represent the foreground. (Figure similar to a figure in Gonzalez & Woods (2008).)

rest of the image, i.e. it would not only fill holes but also abrade potentially relevant structures of the object contour. A more robust, iterative approach is employing *morphological reconstruction*. As described in detail by Gonzalez & Woods (2008), morphological reconstruction uses two images and a structuring element instead of only one image and a structuring element. Generally, the so-called marker image F defines the starting points for a morphological operation whereas the mask image G, e.g. the original image, constrains the operations to specific regions.

For hole filling of a binary image, the mask G is the complement I^c of the original image I. The marker F is a matrix of the same size as I with zeros everywhere except for the borders where it is $1 - I$ (see figure 1.14). Hole filling can be accomplished by iteratively dilating F by the structuring element B, e.g. a 3×3 matrix of ones, and computing the set intersection with I^c until the resulting image does not change any more. The essential operation is called *geodesic dilation* and is defined iteratively as

$$D_G^{(n)}(F) = D_G^{(1)}\left(D_G^{(n-1)}(F)\right) \tag{1.15}$$

with

$$D_G^{(1)}(F) = (F \oplus B) \cap G \text{ and } D_G^{(0)}(F) = F. \tag{1.16}$$

As shown in figure 1.14, the dilation of F with B starts at the borders and proceeds inward. The set intersection with I^c protects the original foreground pixels from changing during the iterations. In this simple example, one iteration is already sufficient to fill the hole as indicated in the final result H.

1.5 Machine Learning

The aim of the presented work is to measure the quality of ABUS images automatically. Due to the increasing throughput of medical images especially in a screening scenario, machine learning suggests itself to be included to the solution of this task, i.e. the computerized emulation of an image quality rating that had been provided originally by a clinician. The basic principles of machine learning as it will be used in this work are described in the following.

"The ease with which we recognize a face, understand spoken words, read handwritten characters, identify our car keys in our pocket by feel, and decide whether an apple is ripe by its smell belies the astoundingly complex processes that underlie these acts of pattern recognition" (Duda et al. 2001). What seems to be a very easy task for us with our highly sophisticated neural and cognitive capabilities can be a tough challenge for a machine. Nevertheless, we want machines to support us by recognizing patterns reliably and accurately as for example in object recognition and image classification, and thus, make them learn.

Whereas unsupervised learning approaches focus on the detection of patterns and clusters in data sets of unknown classes and categories, *supervised learning* provides an annotated set of training *instances* to derive concepts that can then be applied to an unseen instance in order to predict its class. Each instance is characterised by the values of *attributes* (or *features*) that measure different aspects of the instance (Witten & Frank 2005). The latter approach will be employed throughout this thesis. Features will be derived from the images using image processing on different scales, i.e. describing the image as a whole or on the level of smaller patches. They can be categorical or numerical, describing geometrical measures or histogram attributes, and can be based on the original images or on corresponding parameter maps.

In conclusion, the three main issues of supervised machine learning are the annotation of a sufficiently large training data set, the definition and computation of meaningful attributes (feature extraction), and the selection and instantiation of a suitable classifier algorithm. In order to evaluate whether a trained classifier will be able to predict the class of an unseen instance reliably, it is generally applied to a test data set of annotated instances that are disjunct from the training instances.

1.5.1 Feature Ranking

The design of a meaningful set of attributes from scratch is a challenging task. Prior knowledge, e.g. the particular view of an image, has to be translated into nominal attributes and visual characteristics have to be simplified until they can be computed automatically. As soon as a potential set of attributes has been defined and determined for a training data set, often the importance of those attributes to discriminate between the classes, i.e. their discriminant power, is computed using feature ranking (or attribute selection). Various methods are available for this purpose, out of which two are described below.

Pearson's Correlation

A straight forward approach to measuring the correlation between particular features and the class in an annotated training data set is to compute Pearson's correlation coefficient. It expresses the linear correlation between two quadratically integrable random variables X and Y and takes a value of 1 (-1) if there is a perfectly positive (negative) linear correlation. It is defined as

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)} \tag{1.17}$$

with σ being the true standard deviation and $\mathrm{Cov}(X,Y)$ describing the true covariance of the distribution. The empirical correlation coefficient ρ_e for paired values of a measurement (sample) is defined as

$$\rho_e = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{1.18}$$

where $\bar{x}$ and $\bar{y}$ are the empirical mean values of the measurement.

Since ρ_e is intended to uncover linear correlation between two variables that are on an interval scale, it is intuitive but not suitable to analyse the relevance of features for a separation into only two classes (class labels are not on an interval scale). Furthermore, if $\rho_e = 0$, there might still be a non-linear correlation between a feature and the class.

Information Gain Ratio

Another method to evaluate the relevance of single attributes for classification is to compute the *information gain ratio* with respect to the class. The *information* is defined in Witten & Frank (2005) as a measure of purity in a subset of training instances with respect to the amount of positive and negative class instances in this subset, i.e. the *entropy* H. Using the picture of a decision tree, the measure represents the expected amount of additional information that is needed to decide on the class of a new instance when it has arrived at a specific node of the tree.

If a training data set contains p positive and q negative instances with $p+q = n$, the contained information is computed as

$$\mathrm{info}([p,q]) \coloneqq H(p/n, q/n) = -(p/n)\log(p/n) - (q/n)\log(q/n). \tag{1.19}$$

If the data set is split into two subsets s_1 and s_2 based on two values a_1 and a_2 that a specific attribute a can take, $s1$ and $s2$ will contain p_1 and p_2 positive as well as q_1 and q_2 negative instances with $p_1 + q_1 = n_1$ and $p_2 + q_2 = n_2$, respectively. The *information gain* by the split on attribute a will then be

$$\begin{aligned} \mathrm{gain}(a_1, a_2) &= \mathrm{info}([p,q]) - \mathrm{info}([p_1,q_1],[p_2,q_2]) && (1.20)\\ &= H(p/n, q/n) - H(p_1/n_1, q_1/n_1) - H(p_2/n_2, q_2/n_2). && (1.21) \end{aligned}$$

The entropy is chosen as it is the only function that satisfies all of the following three necessary properties.

- If a data subset is pure, i.e. contains only instances of one class, the measure is zero.
- If the impurity (or randomness) is maximal, i.e. all classes are equally likely, the measure is maximal.
- The measure obeys the *multi-stage property*, meaning that it can reflect data splits into more than two classes. Such a split can be performed by splitting data into two subsets in a first stage, and then splitting up these subsets again in a second stage.

The information measure is however biased towards attributes with a large number of values since subsets of data are more likely to be pure if there are many different attribute values, each of which theoretically could serve as split criteria. These so-called highly branching attributes, thus, can cause over-fitting or fragmentation, meaning that the data is split into (too) many small sets. Therefore, the *intrinsic information* of a split is computed as the entropy of distribution of instances into branches. It reflects the amount of information that is needed to tell which subset an instance belongs to. Hence, attributes with higher intrinsic information are less useful. For the split described above, the intrinsic info is

$$\text{info}_{\text{int}}(a_1, a_2) = H(n_1/n, n_2/n). \tag{1.22}$$

Finally, these two metrics are combined to the information gain ratio as

$$\text{GR}(a_1, a_2) = \frac{\text{info}(a_1, a_2)}{\text{info}_{\text{int}}(a_1, a_2)} \tag{1.23}$$

which can be calculated for each feature to indicate its relevance for the automated classification.

1.5.2 Random Forests

A popular approach to machine learning is a *decision tree*, which is grown using a training data set $X = x_1, ..., x_n$ with ground truth class annotations $Y = y_1, ..., y_n$. A decision tree is made up of nodes, where the data splits into branches, and ends up in leaves if no further splitting is possible. A sample instance x' will follow a particular branch of the tree until it ends up in a leaf corresponding to a specific class, which x' is then assigned to. Tree growing starts from the whole training data set. Out of all features the one with the largest variance is used to introduce a linear split of data such that the intra-class variance of the resulting subsets is minimal. The procedure is repeated at each node using the corresponding data subset until either the class of all instances in the node is the same or another stopping criterion is satisfied, e.g. specified depth of the tree or number of instances in the node. Advantages of the decision tree method are the invariance under scaling and various other transformations of feature values, the robustness to inclusion of irrelevant features, and the descriptive, presentable models it produces. However, decision trees tend to learn highly irregular patterns and over-fit their training sets if they are grown too deep.

Random forests were proposed by Breiman (2001) to average multiple deep decision trees that are trained on different subsets of the available data, which will in turn reduce the variance, i.e. avoid over-fitting. The output class of a new instance is then the mode of the classes each tree voted for. For every decision tree in the forest the following steps are applied:

(i) Out of N instances in the training data set, $n < N$ objects are randomly selected with replacement ("bagging").

(ii) Out of M attributes, $m < M$ attributes are randomly chosen at each node and used to perform the data split, e.g. by minimizing the entropy.

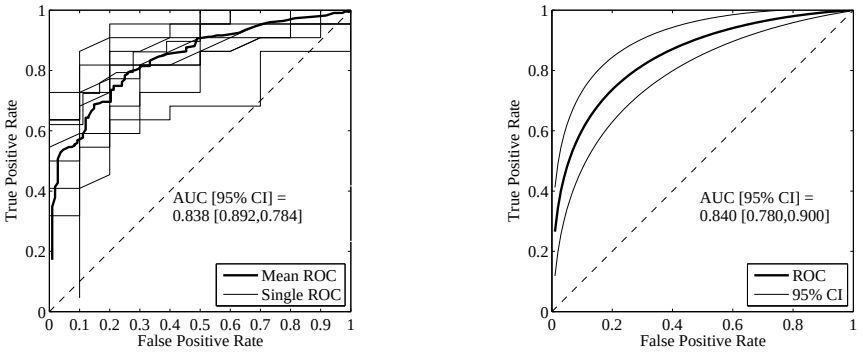
(iii) The tree is fully grown (until a defined stopping criterion is satisfied), but not pruned.

Several parameters need to be defined before growing a random forest. The number of trees can range from some hundreds to several thousands and is an important factor for computational speed. The maximum depth $d_{\max}$ of a tree can be limited, e.g. to the number M of available features. The minimum number $n_{\min}$ of samples (instances) required at a leaf node for it to be split can be restricted, e.g. to a low percentage of the available instances in the training set. Finally, the number m of randomly selected features at each node has to be defined. Breiman (2001) proposed to set it to $\log_2(M) + 1$.

1.5.3 Receiver Operating Characteristic

Measuring the performance of a classifier is normally achieved by comparing the class label output for each instance of a test data set to a corresponding ground truth annotation. In a two class system, talking of a "positive" and a "negative" class is common. If an actually positive instance was assigned the correct class label by the classifier, it is considered a "true positive" (TP), otherwise it is counted as "false negative" (FN). A factually negative instance that is classified correctly is called "true negative" (TN), whereas it is a "false positive" (FP) if it is assigned to the positive class. The true positive rate (TPR), also called sensitivity, is the amount of TPs divided by all actually positive instances (TP+FN). The false positive rate (FPR), on the other hand, is computed as FP/(FP + TN). The specificity is defined as 1 − TPR. These measures can also be used to compare the ratings of two readers with each other.

A more detailed analysis of the classifier can be retrieved from the receiver operating characteristic (ROC) curve, that displays the TPR as a function of the FPR for all possible decision thresholds (operating points) of the classifier. Whereas the diagonal through the origin in this plot represents a random classification, curves that have a steep slope and come close to the upper left corner of the plot, which stands for high sensitivity at high specificity, describe a good classification. The area under this curve (AUC) is therefore a standard measure in machine learning applications. In figure 1.15 two different types of ROC curves originating from the same cross-validation experiment are shown. The ROC plots in figure 1.15a are directly retrieved from the empirical data, whereas the curve in figure 1.15b represents the bivariate normal fit of the data.



(a) **(b)**

Figure 1.15: Sample ROC curves. In (a), empiric ROC curves resulting from a 10-fold cross-validation experiment are plotted. The thin lines describe the single folds, whereas the bold line is the merged ROC curve combining all folds. The given AUC value is derived from this merged curve. In (b), the same data has been fitted to generate smooth ROC curve of the merged folds. The 95 % confidence interval has been computed from the fitted curves of the single folds (not shown).

If the available training data is not large enough, it is common practice to perform cross-validation instead of a simple train-test approach. In n-fold cross-validation, the annotated data set is split into n folds, out of which $n-1$ are used for training and the remaining one for testing. This is repeated until each fold has been employed once as test data. In "stratified" cross-validation, the data is distributed such that the amount of positive and negative instances is the same in each fold. In order to estimate a confidence interval of the measures retrieved in cross-validation, several runs with newly assorted data are performed. This opens different options for the computation of statistical measures describing the classifier performance as outlined in detail by Forman & Scholz (2010).

Generally, precision and recall are defined as $Pr \coloneqq \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and $Re \coloneqq \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. The F-measure of a classifier is then given as

$$F = 2 \cdot \frac{Pr \cdot Re}{Pr + Re}. \tag{1.24}$$

When it comes to cross-validation, one can average over all $F^{(i)}$ of all folds i

$$F_{\mathrm{avg}} = \frac{1}{n} \cdot \sum_{i=1}^{n} F^{(i)}. \tag{1.25}$$

Another option is to first average precision and recall over all folds and then compute the F-measure

$$Pr = \frac{1}{n} \cdot \sum_{i=1}^{n} Pr^{(i)} \tag{1.26}$$

$$Re = \frac{1}{n} \cdot \sum_{i=1}^{n} Re^{(i)} \tag{1.27}$$

$$F_{\mathrm{pr,re}} = 2 \cdot \frac{Pr \cdot Re}{Pr + Re}. \tag{1.28}$$

Similarly, the AUC measure can be determined as average over all folds

$$\mathrm{AUC}_{\mathrm{avg}} = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathrm{AUC}^{(i)} \tag{1.29}$$

or as the area $\mathrm{AUC}_{\mathrm{merge}}$ under the merged ROC curve resulting from all instances of all folds sorted for their output probabilities. This merging implies however that the classifier is assumed to produce well-calibrated probability estimates.

2 Materials and Methods

All developments and computations presented in the next sections were based on original ABUS images that had been acquired in routine clinical care (screening or diagnosis) and were provided to all partners of the ASSURE project. All data was anonymised. The Institutional Review Board waived the need for informed consent and approved the use of anonymised images for the studies performed within the project. In total, 815 ABUS volumes of 104 women acquired either at Radboud University Medical Centre (RUNMC) (Nijmegen, Netherlands) or at Jules-Bordet-Institute (IJB) (Brussels, Belgium) using a Siemens ACUSON S2000 ABVS or U-Systems somo-v device have been provided. The images were organized in subsets as listed in table 2.1. A*, B*, and C* consist of those ABUS volumes from A, B, and C, respectively, that actually contain the nipple. Visibility of the nipple is an important aspect for some evaluations presented in this work and was assessed manually. As will be described in the following sections, each of the data subsets was rated manually by two out of three readers providing the ground truth for this study.

Furthermore, all computations described in this thesis were performed on a Windows 7 machine with an IntelÂő CoreâĎć i7-2627M processor at 2.7 GHz and with 6 GB of RAM. All computing times that are given in the following chapters refer to this specific hardware.

Table 2.1: Data subsets used in this work

Name	*Images*	*Women*	*Institute*	*Scanner*	*Nipple visible*	*Readers*
A	37	14	RUNMC	U-Systems	yes & no	1 & 2
A*	28	14	RUNMC	U-Systems	yes	1 & 2
B	331	23	RUNMC	Siemens	yes & no	1 & 2
B*	312	22	RUNMC	Siemens	yes	1 & 2
C	447	67	IJB	Siemens	yes & no	2 & 3
C*	394	66	IJB	Siemens	yes	2 & 3

2.1 Empirical Analysis of ABUS Artefacts

In the previous chapter, diverse imaging artefacts that can occur in ABUS images have been mentioned. In order to analyse their frequency in clinical practice and their relevance for diagnosis, 368 ABUS volumes were evaluated manually with respect to image quality. Two radiologists ("Reader 1" and "Reader 2") inspected the images visually and annotated them separately, i.e. if they thought a specific image quality criteria was not fulfilled properly, the image was labelled accordingly. Inter-rater agreement analysis was performed to evaluate the ob-

jectivity of the defined criteria. The results of the manual artefact annotation served as basis for the envisaged software development.

There have been publications treating image quality or special artefacts of ultrasound (Keeble et al. 2013), breast ultrasound (Baker et al. 2001), and even automated breast ultrasound (Boehler & Peitgen 2008), but to the author's knowledge, this was the first time that a statistical analysis of specific image artefacts for ABUS was performed. Arleo et al. (2014) reported that in the first month of using ABUS in their institution the recall rate due to BI-RADS 0-incomplete scans was 25 % but trended down to under 13 % in the third month. These numbers show that ABUS has a learning curve, meaning that technicians learn to acquire better images over time and with training. This encourages the implementation of an automated quality assessment system that supports the technicians as proposed in this work.

Manual Annotation Process

In order to facilitate the manual annotation of a substantial amount of ABUS images, a dedicated software tool was developed. The tool allowed easy access to a defined list of ABUS images that were supposed to be classified by the radiologist. It comprised a standard medical viewer, such that artefacts could be detected and analysed in similar environment as for clinical diagnosis. A list of the possible artefacts was provided electronically for simple processing and automatic structured storage of the manual ratings.

Two medical researchers with several years of experience in ABUS image interpretation manually annotated the images of data sets A and B. The readers were explicitly asked to evaluate each single ABUS image individually without considering potentially available other views of one breast. In most cases, a region affected by an artefact can be examined better in another view of the same breast such that the radiologist is still enabled to make a proper diagnosis. However, a classification of the whole study across the different available views is only the next step after rating each ABUS image individually. Having said this, high quality of all images in one examination is an essential prerequisite for both efficient and effective diagnostic image read as well as for CAD algorithms.

Inter-rater Agreement

The inter-rater agreement of the manual annotation was computed as Cohen's κ coefficient (Cohen 1960). It is defined as

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \tag{2.1}$$

with p_0 being the relative observed agreement between the two raters and p_c being the hypothetical probability of random agreement. The latter one expresses the probability that both raters agree by chance and is based on the empirical probability of each rater choosing each category. A complete agreement between the two observers yields $\kappa = 1$. If there is no agreement between the two raters other than what would be expected for random selection, then $p_0 = p_c$ and consequently $\kappa = 0$.

There is no universally accepted interpretation of the κ value, however many diverse guidelines have appeared in the literature. Landis & Koch (1977) stated that one could interpret κ values in the range of 0 to 0.20 as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and 0.81 to 1 as almost perfect agreement. Fleiss (1973) characterized κ values below 0.40 as poor, between 0.40 and 0.75 as fair to good, and above 0.75 as excellent agreement. In conclusion, inter-rater agreement with κ between 0.40 and 0.60 might be acceptable and values above 0.75 express good agreement whereas κ values below 0.40 should be considered sceptically (Greve & Wentura 1997). This latter interpretation will be adopted throughout this work.

Relevance and Frequency of Artefacts

The incidence of the considered artefacts was evaluated and used to estimate the relevance of each single artefact. The higher the relative frequency of a specific artefact, the higher the significance and usefulness of an automated artefact detection are. The outcome of the manual annotation influenced the subsequent software development process in two ways. It defined the relevant artefacts that should be covered by the envisaged image quality assessment tools, and served as training basis for the classifiers that were employed as described in the following sections.

2.2 Computer-aided Analysis of ABUS Artefacts

In the following sections, the methods for an automated detection of diverse image artefacts will be described in detail. Generally, the artefact detection was based on image processing and machine learning, meaning that several image features were extracted and used for classifier training. The feature extraction was performed on different scales depending on the considered problem. For the relative nipple position, the nipple shadow, and the breast contour shape, features describing the image as a whole were used yielding a classification on image level. The workflow diagram describing how these three quality aspects were approached is shown in figure 2.1. For the air artefacts, a sliding window approach was employed leading to a classification on pixel level. The visibility of the nipple was assessed by features based on a set of probability maps.

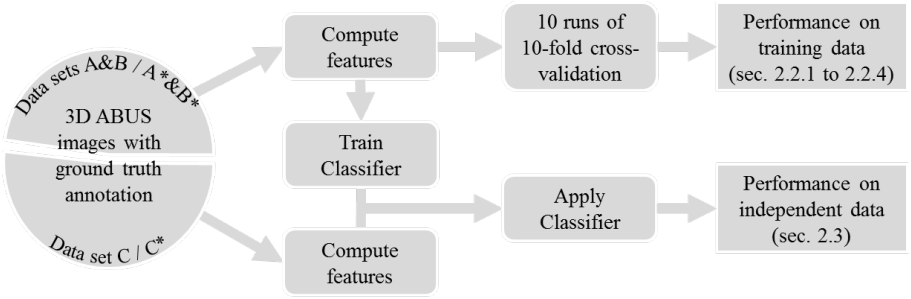


Figure 2.1: Schematic overview of the workflow and data usage for the automatic classification of the relative nipple position, the nipple shadow, the breast contour shape and a combination of these three.

2.2.1 Relative Nipple Position

The position of the nipple relative to the breast contour line is a relevant image quality aspect in ABUS imaging. If the nipple is pushed too close towards the borders of the imaged breast or not supported correctly by the designated cushions, the nipple shadow often gets very prominent and the contour line of the breast cannot be clearly distinguished any more. As a consequence, it is

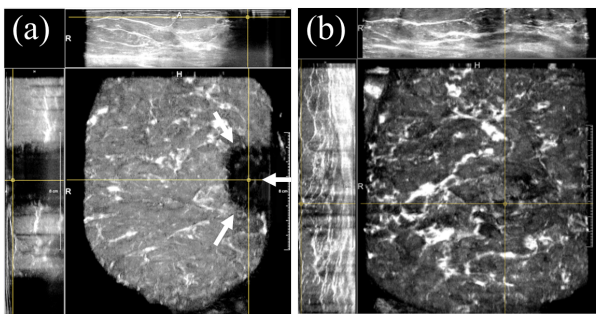


Figure 2.2: Two ABUS of the same breast acquired in different views. Whereas the nipple is pushed to the very edge of the breast in the LAT view image in (a), the AP view image in (b) shows no artefacts at all.

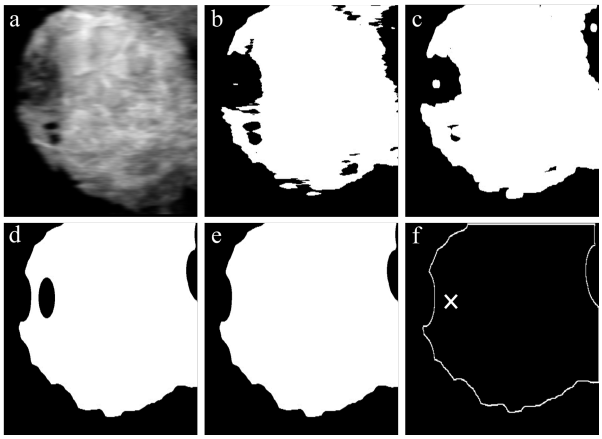


Figure 2.3: Single steps for breast mask computation: (a) smoothed coronal projection, (b) binary mask after applying Otsu's threshold, (c) dilated mask, (d) closed mask, (e) holes are filled, (f) eroded and contoured mask with a marker at the nipple position (set by technician during image acquisition).

often unclear whether all relevant parts of the breast were imaged correctly. A sample case is shown in figure 2.2.

Pre-Processing

The ABUS images were prepared for feature extraction in several pre-processing steps. First, a 2D coronal breast mask was computed similarly to the approach proposed by Tan et al. (2013a). Therefore, a coronal mean projection of a stack of 120 slices close to the skin was performed. However, the top 50 slices from the skin were excluded from the breast mask computation to avoid responses from the skin tissue. The projection image was smoothed using a Gaussian filter with a sigma of 0.2 mm and binarized by applying Otsu's threshold filter (Otsu 1975). In order to close holes within the breast mask or at its edges, the binary image was dilated and holes were filled before it was eroded again. Finally, the breast contour line was computed based on the mask image as shown in figure 2.3.

Following the standard acquisition protocol, the technicians generally pinpoint the actual nipple position $(x_\mathrm{T}, y_\mathrm{T})$ in the coronal view at the end of each acquisition. Therefore, the absolute nipple position coordinates were considered as given and used as input for feature extraction.

Feature Computation

Based on the extracted breast mask and the given nipple position, nine presumably meaningful features were computed.

- c_view:The view of the considered image strongly influences the absolute nipple position and may affect the impact of a nipple being close to the contour line of the breast. Thus, a categorical feature c_view that can be one of the four available standard views (AP, LAT, MED, SUP) was extracted from the information provided in the header of the DICOM file.

- x_T and y_T: The given nipple coordinates (x_T, y_T) were considered as possibly important features since the absolute nipple position in the image may correlate with the position relative to the breast. As the appearance of ABUS images differs a lot depending on the breast size and the transducer position, the absolute nipple position is however not coupled directly to the nipple position relative to the breast image.

- d_{min}: The shortest Euclidean distance d_{min} between the nipple position (x_T, y_T) and the breast mask contour line was computed.

- c_{io}: It was determined whether the nipple was located inside or outside the breast mask. The latter case can occur when the shadow around the nipple is very dark and close to the breast contour such that this region is—by mistake—not included in the breast mask. A categorical feature $c_{io} \in \{1, -1\}$ was included.

- d^*_{min}: The signed distance between nipple position and contour line was computed as $d^*_{min} = d_{min} \cdot c_{io}$.

- A_B: The total 2D physical area of the breast A_B was computed using the pixel size and the number of pixels within the breast mask.

- $A_{B/I}$: The ratio of the physical 2D area of the breast A_B to the total image size was calculated.

- d_{COM}: The centre of masses (x_{COM}, y_{COM}) of the breast area and the Euclidean distance d_{COM} between (x_{COM}, y_{COM}) and (x_T, y_T) was determined.

Classification

The learning step was based on data sets A* and B*, i.e. 342 ABUS volumes actually containing the nipple in order to prevent bias from calculations that were based on incorrect assumptions for the nipple coordinates. As described in section 2.1, the ground truth annotation for classification was done by two medical experts with several years of experience in ABUS image interpretation. The relative nipple position was categorised by each reader separately as "too close to the contour line of the breast", "acceptable", or "good". For classifier training, this rating was transformed to a two class annotation: If both readers agreed that the nipple position was "too close to the contour line of the breast", the case was given the positive class. All other cases were summarized in the negative class.

Information gain ratio feature ranking (see section 1.5) was then performed in order to estimate the relevance and discriminant capacity of the computed attributes. The standalone application WEKA [1] (Hall et al. 2009) was employed for this purpose. Machine learning was implemented using a Random Forest classifier (Breiman 2001) as provided by the OpenCV library [2] (Bradski 2000). Classifier performance was measured in 10-run 10-fold stratified cross-validation.

[1] Version 3.7.11, http://www.cs.waikato.ac.nz/ml/weka/
[2] Version 2.4.10, http://opencv.org

The number of trees in a forest was set to 100, whereas the maximum depth of each tree was set to 15. The minimum sample count required at each node to be split was set to 10 % of the total number of samples, yielding 35. The number m of random features considered at each node for decision tree construction was set to $\log_2(M) + 1$ as proposed by Breiman (2001) with M being the number of features, and thus was 4. The 10 folds for cross-validation were randomly assorted under the constraint of similar class distribution in each fold.

2.2.2 Nipple Shadow

A major issue in sonographic imaging is the strong reflection of ultrasound waves at tissue-air boundaries, which causes regions adjacent to air cavities to be occluded by acoustic shadows on the resulting images. In breast sonography, the nipple being surrounded by a potentially uneven areola and connected to ducts is very difficult to be imaged properly. If an insufficient amount of contact fluid is applied to the areola region or the ducts are filled with air, the tissue behind the nipple cannot be seen on the ultrasound image (see figure 2.4). Another reason for a badly imaged nipple region is inadequate support of the breast (by cushions) in lateral or medial view, which could lead to insufficient contact between transducer and areola.

Although the tissue behind the nipple might be sufficiently visible in other views of one breast, an automated detection of extremely prominent nipple shadows could help to sensitise the technicians to this issue.

Feature Computation

In order to estimate the size of a potential nipple shadow, it was assumed that the shape of the shadow could be approximated by a cylinder around the nipple with the axis going in antero-posterior direction (see figure 2.5). As the nipple is (approximately) a disk in the coronal plane, once it has stopped the US wave it produces a cylindrical shadow region. The nipple position $(x_\mathrm{T}, y_\mathrm{T})$ in coronal plane was obtained from the DICOM header as given by the technician during image acquisition and assumed to be the same for all coronal slices. The size of the potentially dark cylindrical region around the nipple position was estimated by counting cylinder segments (rings) that had low pixel intensity. The radius of the different cylinder segments varied from 4.0 mm to 20.0 mm in steps of 4.0 mm (see figure 4). In AP direction, the height of each cylinder segment was approximately 2.0 mm. The top layer was positioned starting at 6.0 mm below the skin avoiding potentially disturbing high intensity signals due to skin fat or sound reflections within the coupling layers of the transducer. The bottom layer ended at 26.0 mm below the skin. The following seven features were extracted:

- c_view: The view of the considered image affects the absolute nipple position and the possibilities to support the breast properly by cushions. Thus, a

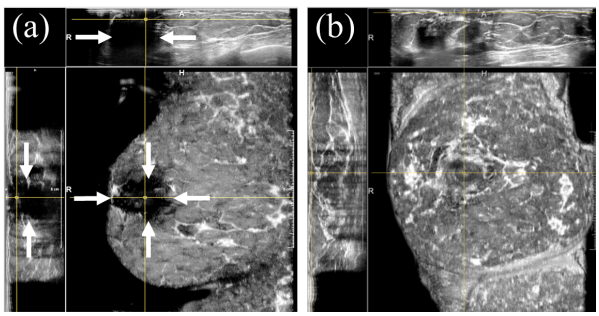


Figure 2.4: Sample case for prominent nipple shadow. (a) shows the MED view of a left breast with very dominant shadow behind the nipple. (b) is the same breast imaged in AP view with nearly no shadow behind the nipple.

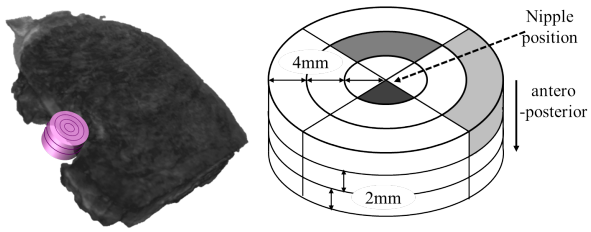


Figure 2.5: Arrangement of cylinder segments that were used to estimate the size of a nipple shadow. The symmetry axis was at the nipple position as shown on the left, which is a segmented 3D view of a sample ABUS with the nipple area marked by the magenta cylinder. Three out of ten used layers and three out of five used rings (radii) are shown on the right.

categorical feature c_{view} that can be one of the four available standard views (AP, LAT, MED, SUP) was introduced.

- $N_{I<50}$ and $N_{I<60}$: The segments showing a lower mean intensity than a specific threshold value were counted. The intensity threshold was set to 50 and 60, respectively, yielding two features, $N_{I<50}$ and $N_{I<60}$, for every image. In the present 8-bit grey scale images, these threshold values yielded reasonable differentiation between tissue and shadow signals.

- N_{Pix}: The amount of pixels N_{Pix} in the cylinder segments that had a mean intensity below 60 was counted. This number accounted for the different sizes of the considered cylinder segments.

- σ^2_{bright}: The variance σ^2_{bright} of brightness in one cylindrical region of 4.0 mm radius around the nipple was calculated since ultrasound shadow signals tend to have a lower variance than signals reflected from structured tissue. The cylinder went from the skin to a depth of 25.0 mm in antero-posterior direction.

- x_T and y_T: The coordinates (x_T, y_T) describing the absolute position of the nipple in coronal plane were included.

Classification

Automatic classification of ABUS images according to the size of a potential nipple shadow was again based on data sets A* and B* and performed similarly to the procedure described in section 2.2.1 for the relative nipple position. The ground truth annotation was provided by two medical experts (see section 2.1) independently rating the nipple shadow of an image as “too prominent”, “acceptable” or “good”. Those images which both radiologists found a “too prominent” nipple shadow in were assigned the positive class. All other images were given the negative class label in order to focus on high specificity of the trained classifier.

Again, information gain ratio feature ranking was performed prior to Random Forest training and testing in 10-fold 10-run stratified cross-validation experiments. The number of trees in a forest was set to 100 whereas the maximum depth of each tree was set to 15. The minimum sample count required at each

node to be split was set to 10% of the total number of samples, yielding 34. The number of random features considered at each node for decision tree construction was set to $\log_2(M) + 1 = 3$, since the number M of attributes was 7.

2.2.3 Breast Contour Shape

The shape of the breast contour is an important indicator of the completeness of an ABUS image. Breast size and shape as well as the distribution of dense and fatty tissue vary strongly among different women. Nevertheless, the typical shape of a breast in the coronal plane of an ABUS scan is round. Whereas large breasts may fill the whole image, small breasts generally result in a rather circular structure with a smooth contour line in the coronal plane. A very irregular, uneven breast contour line is an indicator for incomplete breast coverage during image acquisition caused, e.g., by insufficient support of the breast by cushions. In figure 2.6 two views of the same breast illustrate the described effect: The MED view in figure 2.6a is missing some parts of the breast indicated by the very irregular contour line in the upper right quadrant of the coronal plane. The AP view of this breast in figure 2.6b proves that the breast was of normal shape and could be imaged correctly.

Pre-Processing

In order to extract the breast mask and its contour line, several pre-processing steps were performed. They were similar to those described in section 2.2.1 but with focus on the breast contour line. A 4 mm stack of coronal slices starting at a distance of 7 mm from the skin was used for breast mask generation. The top 7 mm of coronal slices were excluded since they often contain spurious signals caused by sound reflections within contact fluid on parts of the transducer that do not have skin contact. Coronal slices lying deeper than 11 mm were not included in order to avoid signals from the ribs that can already appear from this depth on, depending on the breast size and the transducer positioning. The 4 mm stack of slices was projected to one 2D coronal slice that was binarized using Otsu's threshold filter (Otsu 1975). Morphologic closing with a circular structuring element of 3 mm radius was performed to smooth the contour and potential small irregularities. Holes that were lying completely within the binary breast mask were closed. If more than one connected component were in the image, the largest area was kept as breast mask and the others were ignored. This may happen if large parts of the armpit or the ribcage are imaged.

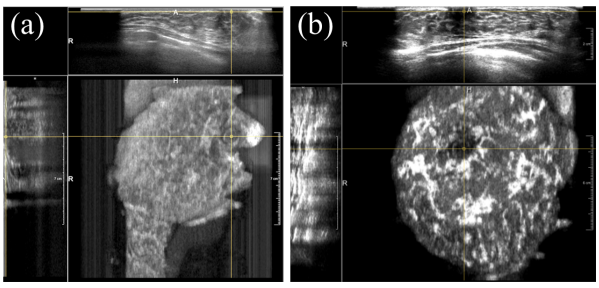


Figure 2.6: Sample case for irregular breast contour shape. (a) shows the MED view of a left breast with very irregular contour line in the upper right quadrant of the coronal image. (b) is the same breast imaged completely in AP view with roundish breast contour line.

Feature Computation

The 2D coronal breast mask and breast contour line were used to extract 17 presumably discriminative features for the detection of prominent background regions and irregular breast contour shapes.

- c_{view}: The view direction was taken into account since breast positioning and cushion support depend on the intended view. The typically available four views of one breast (AP, MED, LAT, SUP) were used as categorical feature c_{view}.
- A_{B}: The physical area A_{B} in 2D coronal view of the breast mask was assessed as a first indicator for the amount of tissue being imaged.
- p_{Mask}: The perimeter p_{Mask} of the breast mask was determined and corresponded to the length of the breast contour line. The higher p_{Mask}, the more curves and irregularities might be in the contour line.
- x_{C} and y_{C}: The position $(x_{\text{C}}, y_{\text{C}})$ of the breast mask centroid was computed as an indicator for the position and "mass distribution" of the breast within the image.
- l_1, l_2, and F: An ellipsoid was fitted to the breast contour line, and the lengths l_1 and l_2 of the ellipsoid axes were determined. The flatness F was computed as the ratio l_2/l_1 to indicate whether the breast contour is extremely elongated in one direction or rather roundish.
- p_{Circle} and r_{Circle}: The perimeter p_{Circle} and the radius r_{Circle} of a circle that has the same surface as the breast mask were computed.
- N_{Border} and p_{border}: The amount of pixels N_{Border} that belong to the breast mask and are touching the edges of the image, as well as the physical length p_{border} of these pixels (perimeter on border) were measured. The higher these measures, the higher is the probability that the imaged breast is very large.
- R_{Border}: The ratio $R_{\text{Border}} = p_{\text{Border}}/p_{\text{Mask}}$ of the breast mask perimeter along the border and the total breast mask perimeter was computed.
- R_{Round}: The roundness $R_{\text{Round}} = p_{\text{Circle}}/p_{\text{Mask}}$ was determined as the inverse ratio between the actual perimeter of the mask and the perimeter of a circle with the same surface. Since the circle is the geometrical shape with the lowest ratio between perimeter and surface, R_{Round} being close to 1 is a strong indicator for a round and smooth breast contour line. If R_{Round} is very small, the determined breast contour line is supposed to be "inefficient", meaning that it has many turns and irregularities.
- p_1 and p_2: The first two principal moments p_1 and p_2 of the breast mask were determined.
- $A_{\text{B/I}}$: The relative size of the breast mask $A_{\text{B/I}} = A_{\text{B}}/A_{\text{Image}}$ compared to the total size of the image was computed. The higher this value, the higher the probability that the breast was imaged completely is.

Classification

Automatic classification of the breast contour shape in ABUS images was based on data sets A and B, i.e. a total of 368 volumes, and conducted similarly to the procedure described in section 2.2.1 for the relative nipple position. The ground truth annotation was provided by two medical experts (see section 2.1) independently rating irregularities in the shape of the breast contour in correlation to predominant background regions as "too irregular", "acceptable" or "not detected". Those images in which both radiologists found the contour to be "too irregular" were assigned the positive class. All other images were given the negative class label in order to focus on high specificity of the trained classifier.

Again, feature ranking based on information gain ratio and subsequent Random Forest classifier training were conducted. The forests consisted of 100 trees, each. The maximum depth of each tree was set to 15, the minimum number of samples for a node to be split was 35, and the number of selected random features was 4. Ten runs of 10-fold stratified cross-validation were performed.

2.2.4 Joint Image Quality Rating

The three ABUS imaging artefacts described in the previous sections—the nipple position, the nipple shadow, and the breast contour shape—are correlated to each other and thus, often appear at the same time in an image. If the ultrasound transducer is not positioned properly or with insufficient pressure, the relative position of the nipple might be inadequate, and, at the same time, the breast contour shape might show irregularities. A prominent nipple shadow, especially if caused by insufficient support of the breast in lateral or medial view, often appears in correlation with an inadequate nipple position (too close to the breast contour in the image), and leads to an irregular breast contour shape. This observation motivates the idea of combining the different computed image features in order to create a single, joint image quality rating. Whereas a dedicated artefact detection method provides more detailed information to the user concerning the origin of an artefact and potential corrective actions, a joint approach might be more stable relying on more features and exploiting the fact that different image artefacts can emerge from the same source.

The three ABUS imaging issues described hitherto were approached by the same method: Features describing the image as a whole were computed and utilized by a classifier to reproduce a manual ground truth annotation. In this part of the work, the previously extracted features were combined and employed to retrace a condensed expert annotation.

The ground truth for a joint image quality rating was extracted from the manual expert annotations introduced in section 2.1. For the distinct consideration of single artefacts, an image was assigned the positive class if both readers agreed to see the specific artefact and that it had the potential to impede diagnosis. This was the case if the nipple was "too close to the breast contour", if the nipple shadow was "too prominent", or if the breast contour shape was "too irregular". In the proposed joint approach, an image was considered to be of low(er) quality if at least one of the three aforementioned image quality aspects was detected concordantly by both readers.

The image features described in the previous sections were combined to a joint feature set describing different image properties. In total, 29 features were used for this classification:

- c_{view}: The view direction was taken into account since it was an important factor for all aforementioned imaging issues. The four standard views (AP, MED, LAT, SUP) were used as categorical feature c_{view}.
- Eight features as derived for the classification of the relative nipple position (without c_{view}).
- Four features describing the intensity distribution around the nipple position (without c_{view}, x_T, and y_T) as described for the nipple shadow classification.
- 16 features characterizing the breast contour shape (without c_{view}).

The machine learning step was performed in the same way as for the three single artefacts described before using data sets A* and B*. 10-run 10-fold stratified cross-validation of Random Forest classifier was conducted to categorise the

quality of an ABUS volume as a whole. There were 100 trees in each forest, the maximum depth of each tree was set to 15, the minimum number of samples at each node to be split was 34, and the number of random features considered at each node was 5.

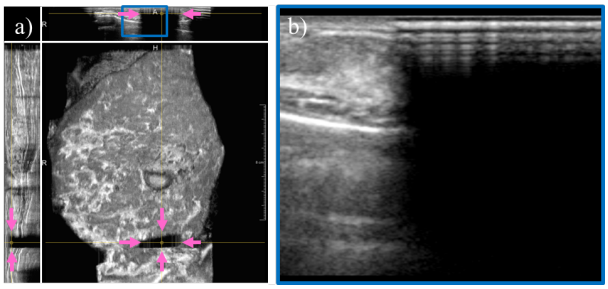


Figure 2.7: (a) Sample ABUS image showing an air artefact. The top part shows one original transversal slice, the left part shows one reconstructed sagittal slice and the mean window shows one reconstructed coronal slice. The red arrows mark the air artefact region. (b) is a magnified view of the artefact as it is visible in transversal view.

2.2.5 Air Artefacts

Ultrasound waves can only be transmitted to the patient's skin if a sufficient amount of contact fluid is applied to the transducer. Otherwise, the ultrasound waves cannot leave the ultrasound probe but are reflected back and forth within the coupling layers (see figure 1.5). This effect can originate from skin folds or air bubbles under the transducer as well as in sparsely applied contact fluid.

The reflection of ultrasound waves on air and subsequently within the coupling layers of the transducer produces a regular pattern of 3–6 bright and dark stripes in the top of the image parallel to the transducer surface (see figure 2.7b top). No further signal reaches the affected transducer elements, which causes fading reverberations in the image and consequently posterior acoustic shadowing of variable extents.

Lesions detected in ABUS screening images are typically very small; reported mean diameters range from 10 mm (Brem et al. 2015) to 14.3 mm (Giuliano & Giuliano 2013). Air bubbles between transducer and skin can be of similar size, illustrating that the posterior shadowing caused by air artefacts may even mask invasive cancers, which otherwise would have been detected by radiologists. Furthermore, the described reverberation pattern could mimic suspicious acoustic shadowing (Baker et al. 2001). Additionally, shadows might be problematic for computerised approaches (Tan et al. 2012; Kuo et al. 2013; Moon et al. 2012), i.e. they affect the intensity profile of the image and can increase the rate of false positive region candidates of automated lesion detection or segmentation.

As described in the following, the properties of air artefacts were analysed in detail in order to develop an automated detection method. Subsequently, features related to the characteristic reverberation pattern were extracted using a sliding window approach yielding 2D parameter maps. This way, a locally operating classifier could be trained.

Data Set

Regarding the relative occurrence of air artefacts, no previous systematic large scale studies exist, but in the present study, all available, i.e. 815 (see table 2.1), ABUS scans were used to generate statistically meaningful output. Each ABUS volume was screened by two of three radiologists for the presence of air artefacts

resulting in 79 "positive" ABUS volumes of 48 women, in which air artefacts were observed by both radiologists. For classification purposes, the data was split into a training set and a test set: The 60 positive volumes of data sets C1 and C2 were chosen as training set, whereas the 19 positive volumes of data sets A and B were assigned to the test set. This way, it was also assured that no overlap of patients or studies existed between training and test set avoiding training bias. The test set was extended by 17 additional volumes without air artefacts. The non-artefact images had been acquired from the same breasts as the artefact images (different view), respectively. These were however not available for all cases.

Properties of Air Artefacts

In those 79 images that had been found to contain air artefacts by both radiologists, a researcher with over two years of experience in ABUS processing annotated the artefact regions in detail by drawing outlines on the edges of the artefact (in 2D coronal reconstruction). These annotations served as ground truth to determine the performance of the proposed algorithms. Artefact regions smaller than $20\,\mathrm{mm}^2$ were considered clinically irrelevant—typical ultrasound-detected lesions are bigger in size—and therefore excluded from the study.

The manually annotated artefacts were examined with respect to their (2D) size in the coronal plane, the number of visible reverberation lines (reflections), the spatial frequency of these lines and the depth of the ultrasound shadow connected to the artefacts.

Pre-Processing

In order to exclude the background as well as irrelevant tissue from further evaluation, a breast mask was computed for each image. As the ultrasound transducer was a linear array, it was assumed that the breast mask was approximately the same in all coronal slices. Therefore, a mean projection image was calculated out of 50 coronal slices. Otsu's threshold filter (see section 1.4.1) was applied to separate the foreground from the background. The largest connected component was kept and decreased by erosion. This eliminates, e.g., parts of the axilla that have been considered as relevant foreground by mistake. The breast masks were designed to cover only the most relevant part of the image leaving out the background and non-breast structures spuriously appearing in the image.

Feature Computation

In this study, features were not computed for each image as a whole, but were extracted using a sliding window approach. Therefore, a window of 3 mm in superior, 3 mm in lateral, and 20 mm in anterior direction was slid over the coronal reconstruction of the image and at each window position, 28 features were calculated. Feature extraction was only performed within the previously computed coronal breast mask to exclude background regions from evaluation. In antero-posterior direction, the sliding window had a total depth of 20 mm (starting at the skin) and was divided into a "small" part V_s including the first

5 mm from the skin and a "large" part V_l comprising the remaining 15 mm. The step size for the sliding window was 1.5 mm in superior and lateral direction (half the window size in coronal plane). Based on the standardized coronal elongation of the ABUS images used in this study, which is 154 mm × 168 mm, at most 11 450 window positions were possible. Due to the excluded background regions, however, the number of sliding window positions, i.e. feature vectors, was much smaller, around 6400.

Two main types of features were used in this study: 1) Statistical features based on the pixel intensities of the volumes V_s and V_l, and 2) Sine fit features as well as Fourier Transform-based features from V_s, which should capture the characteristic periodic stripe pattern present in air artefacts. Feature extraction was performed using Matlab (MATLAB R2011a, The MathWorks, Inc., Natick, Massachusetts, United States) and inbuilt functions. In the following, the different types of features that were extracted from the 3D images to generate 28 2D coronal parameter maps for each image are described.

The discriminant capacity of all 28 computed features was evaluated by information gain ratio feature ranking test that was applied to the entire training data set.

Statistical features Since air artefacts typically cause a deep shadow on the ultrasound image, standard statistic measures were computed to describe the intensity distribution in the two parts of the sliding window, V_s and V_l. The mean μ_s, μ_l, the standard deviation σ_s, σ_l, the median m_s, m_l, and the entropy H_s, H_l of V_s and V_l, as well as the respective ratios of these values μ_s/μ_l, σ_s/σ_l, m_s/m_l, and H_s/H_l were computed.

The entropy H_i was defined as

$$H_i = -\sum_{j=1}^{256} p_j(V_i) \cdot \log_2(p_j(V_i)), \qquad i \in \{s, l\}\,, \tag{2.2}$$

where $p(V_i)$ contains the 256 histogram counts of the sliding window part V_i.

Furthermore, the skewness s_s, s_l, and kurtosis k_s, k_l of V_s and V_l as well as the respective products $s_s \cdot s_l$, $k_s \cdot k_l$ were determined. The skewness s_i was computed as

$$s_i = \frac{\frac{1}{n}\sum_{j=1}^{n} (I_j - \mu_i)^3}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} (I_j - \mu_i)^2}^3}, \tag{2.3}$$

where n is the number of voxels in V_i, and I_j is the intensity of voxel j. s_i is a measure of the asymmetry of the data around the sample mean.

The kurtosis k_i describes how outlier-prone a distribution is and was calculated as

$$k_i = \frac{\frac{1}{n}\sum_{j=1}^{n} (I_j - \mu_i)^4}{\left(\frac{1}{n}\sum_{j=1}^{n} (I_j - \mu_i)^2\right)^2}. \tag{2.4}$$

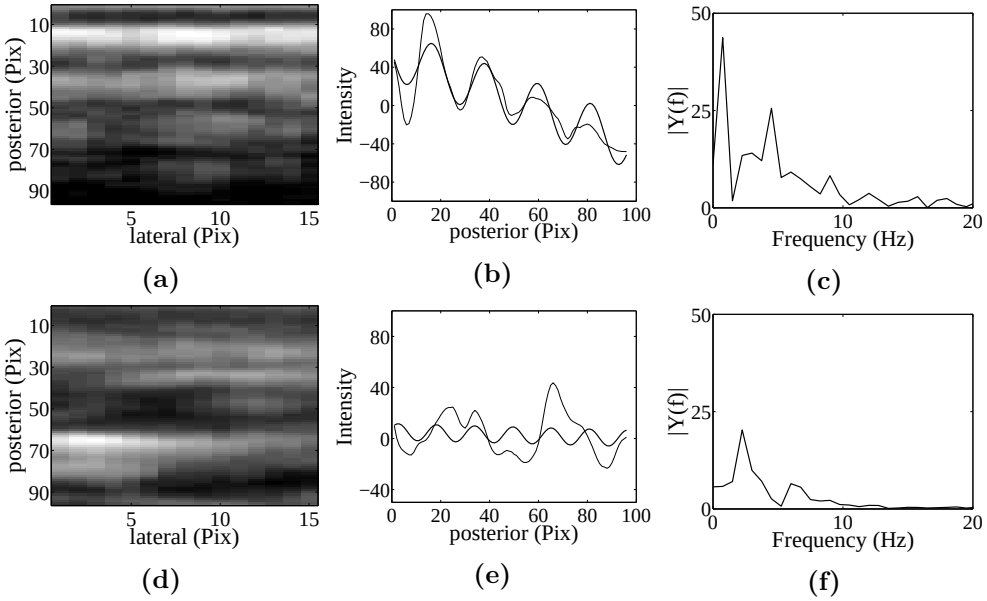


Figure 2.8: (a, d) Examples for transversal projections of the sliding window, (b, e) the corresponding 1D projections (thin line) and sinus fits (bold line), and (c, f) the single sided amplitude spectrum of the Fourier Transform. (a-c) illustrate the characteristic stripe pattern of an artefact-affected instance (positive class), whereas (d-f) represent normal tissue (negative class). Note that the data plotted in (b,e) is the one dimensional projection $\overline{V_s}$ of $V_s - \mu_s$, i.e. it was shifted by the mean value, resulting in partially negative intensity values.

Apart from that, the 70th and 90th percentiles of V_s and V_l were computed.

Sine fit features Other features were extracted from a sine fit to the one-dimensional projection $\overline{V_s}$ of $V_s - \mu_s$ to the antero-posterior axis. The motivation behind this was that, if the sliding window covered an air artefact, the characteristic stripe pattern was expected to be visible as sine curve in this projected view (see figure 2.8). The function to be fitted was

$$f(x) = A \cdot \sin\left(\frac{2\pi x}{p} + \frac{2\pi}{q}\right) + b \cdot x + d \tag{2.5}$$

with amplitude A and period p, allowing for a shift $1/q$ in $x-$direction and d in $y-$direction, respectively. The slope b describes the loss of intensity within the top coronal image slices. Start parameters were set to

$$\begin{aligned} A_0 &= \max\left(\overline{V_s}\right) - \min\left(\overline{V_s}\right) \\ q_0 &= -1 \\ b_0 &= -1 \\ d_0 &= 0. \end{aligned}$$

Start parameter for p was the expected period p_e, empirically determined from the ground truth annotations. Since the reverberation patterns originate from

acoustic reflections within the coupling layers of the transducer, p_e should be characteristic for a specific transducer. The mean fit error Δ was determined as the mean squared distance between original data and fitted curve.

Fourier Transform features Furthermore, Fourier Transform was applied to V_s. Whereas a sine fit can only model an oscillation with one single frequency, Fourier Transform was considered to be more robust against superposition of slightly de-phased signals. Consequently, if the sliding window was at an air artefact, a peak should appear around the empirically determined mean frequency f_e, which was computed as the ratio between the depth of V_s and the measured period p_e. The maximum amplitude of the Fourier spectrum within the range of $f_e \pm 2\sigma_f$ (approximately) was retrieved as feature $Y(f_e)$. σ_f describes the standard deviation of f_e derived from the measured standard deviation σ_p of the period p_e as

$$\sigma_f = \frac{\mathrm{d}f_e}{\mathrm{d}p_e} \cdot \sigma_p. \tag{2.6}$$

As shown in figure 2.8 (c), the continuous signal decay in an artefact region is represented by an additional peak $Y(f_{low})$ below 2 Hz in the Fourier spectrum.

Classification

As described in the following, the derived feature information was combined with the ground truth in order to decide which of these features were most relevant for region classification, i.e. for determining which regions were actual air artefacts. In a second step, the information gained in the training step was applied to test images in order to locate potential air artefact regions.

Again, Random Forest classification was employed. The classification for this two class (artefact or no artefact) problem was separated in two stages: First, 10-fold cross validation was applied to the training data to evaluate the performance on the level of the sliding window (each window position yielding one sample). Secondly, the classifier was trained on the whole training data set and applied to the disjunct test data set containing artefact and non-artefact images. The performance of this train/test step was evaluated in terms of true positive (correctly detected) and false positive instances (on window-level), artefact regions, and images, respectively. The workflow design and data usage of this experiment is depicted in figure 2.9. In contrast to the workflow of the previous section (see figure 2.1), the training and test data included not all available images but only those, that contained air artefacts. Furthermore, the evaluation was done on different scales, i.e. first on window-level and then on region-level, whereas the previous classifications always referred to the images as a whole.

For the cross-validation experiment, the ten folds were sorted such that images of the same visit were in one fold to avoid bias from very similar instances being in different folds. Each Random Forest was built of 100 trees. The maximum depth $d_{\max}$ of a tree was set to the number of available features M, thus $d_{\max} = M = 28$. The minimum number N_{Samples} of samples required at a leaf node for it to be split was set to 1 % of the shadow pixel instances in the data set: $N_{\text{Samples}} = 27$. The number N_{Features} of randomly selected features at each node

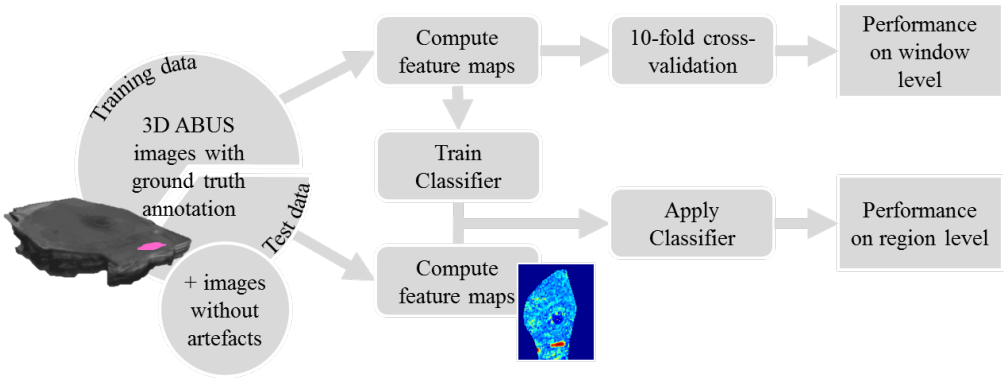


Figure 2.9: Schematic overview of workflow and data usage for the above described air artefact detection.

was set to 5. In a second step, the classifier was trained on the whole training data set and applied to the test data set.

The above described classification yielded a prediction of class on the level of the sliding window. Since air artefacts are generally bigger than the chosen size of the sliding window, the classifier performance was also evaluated on (dynamic) region-level in the test data set. Therefore, pixels that were assigned the positive class (artefact) by the trained algorithm were segmented as connected components, representing air artefact region candidates. This artefact mask produced by the classifier was superimposed to the manual ground truth delineations. Any overlap between both masks was counted as true positive region. Positive regions in the classifier-mask that had no correspondence in the ground truth annotation were considered false positive regions. In order to account for imperfect manual delineation of the artefact regions when creating the ground truth, they were given a margin of 4 mm before comparing them to the classifier outputs.

False positive reduction was achieved by introducing a lower threshold for the region size of the connected components. Artefact region candidates with a size smaller than this threshold were ignored in the evaluation. The physical area of each connected component in coronal plane was computed based on the amount of included pixels and the pixel size. To measure the classifier performance, the number of false positive regions per image as well as the true positive regions were determined for different threshold values.

2.2.6 Automated Assessment of Nipple Visibility

The nipple is an important landmark on ABUS images for the radiologists to localise the quadrants of a breast lesion (Karnan & Thangavel 2007). Furthermore, it serves as reference marker in CAD systems, which gained significant interest in recent years (Tan et al. 2015). Registration algorithms that fuse ABUS data with other imaging modalities, such as mammography, MRI or tomosynthesis, use the nipple position as reference with stable spatial correlation to improve accuracy (Tan et al. 2013a).

In standard ABUS acquisition routine, the nipple position in coronal view is annotated manually by the technicians and saved in the header of the DICOM file. In some cases, however, the manual annotation is not correct, e.g. due to slight inattentiveness or because the nipple is not visible at all. The latter case can occur if the breast is very big or not supported sufficiently by the cushions that are generally used for MED or LAT views. Sample cases are shown in figure 2.10. Automated detection of incorrectly annotated cases has the potential to improve and facilitate any further processing of ABUS images. This section describes a fully automatic approach to assess the correctness of the nipple position made by the technician during the acquisition process. A Random Forest classifier was used to detect the cases with incorrectly marked nipple positions. Feature computation was based on an extension of a previously presented multi-scale Laplacian- and Hessian-based automatic nipple detection method (Wang et al. 2014) by adding prior knowledge encoded in a probabilistic atlas, which accounts for the empirical probability distribution of the nipple position per view.

Data Set

For this study, the images of data sets B and C as well as 7 additional ABUS volumes[3], which had not been included in the previous sections, were used.

If the nipple was not visible on the image, it was manually tagged as class 1[4], otherwise, the centre of the nipple was pinpointed in coronal view. If the

[3] accomplishing data set B, i.e. additional scans of the women already included in data set B and acquired with the same scanner at the same institution

[4] A more rigid manual classification of nipple visibility than in the previous sections was applied here. If the nipple was still partly visible, it was already tagged class 1, whereas the classification of nipple shadow (section 2.2.2) and nipple position (section 2.2.1) were designed to handle those cases regularly.

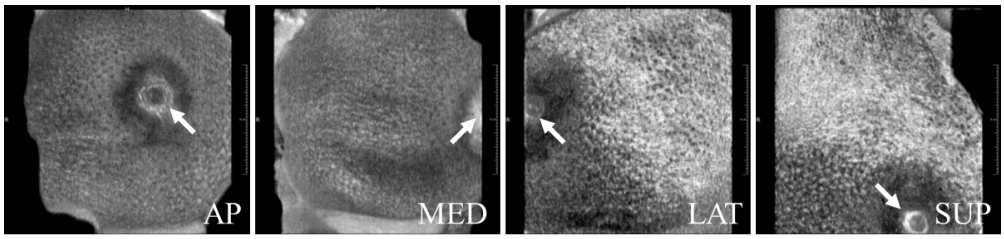


Figure 2.10: ABUS scans of the left breast of one woman, illustrating the nipple positions in different views. Note that in the MED view of this case, the nipple is barely visible.

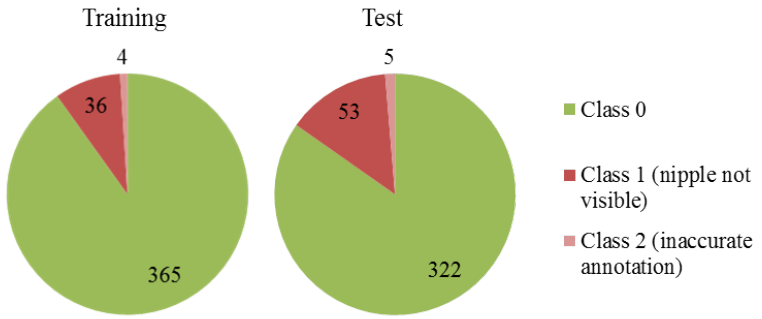


Figure 2.11: Class distribution of the training and test data sets used in this study.

distance between the ground truth annotation and the position marked by the technicians was more than 16 mm, i.e. the original annotation was inaccurate, the image was labelled as class 2. The remaining images were assigned to class 0.

The data set was separated into a training and test data set such that images of the same women were within the same set in order to avoid bias in classifier training[5]. The class distribution is presented in figure 2.11. Atlas generation and classifier training were performed on the training data set. 100, 94, 95, and 26 images from the training data set where used to generate the atlas for AP, MED, LAT, and SUP view, respectively.

Automatic Nipple Detection

Since it is presumed that the nipple itself is located at the skin (anterior-most slices) whereas a potentially adjacent shadow and the ducts are at the same coronal coordinates (x, y) in deeper slices, the nipple position is generally determined on a 2D coronal map derived from the volumetric information of an ABUS image. Wang et al. (2014) previously presented a hybrid automatic nipple detection method using multi-scale Laplacian and Hessian filters to generate a probability map, which incorporated the blobness of the nipple as well as the tubular structure of the nipple shadow. The detection rate of this method was reported to be around 88 % for a tolerance in distance error of 10 mm. The method was extended by incorporating an atlas expressing the empirical probability map for the nipple position in different views. A confidence factor measuring the reliability of the computed nipple position resulting from the original algorithm was implemented. In those cases, where the original algorithm failed to detect the correct nipple position, multiplying the original map with the atlas added empirical information to the joint probability distribution of nipple position. A schematic overview of the proposed workflow is given in figure 2.12.

[5]Note however that patients of data sets B and C were mixed up in this case in order to achieve a more balanced training and test data set.

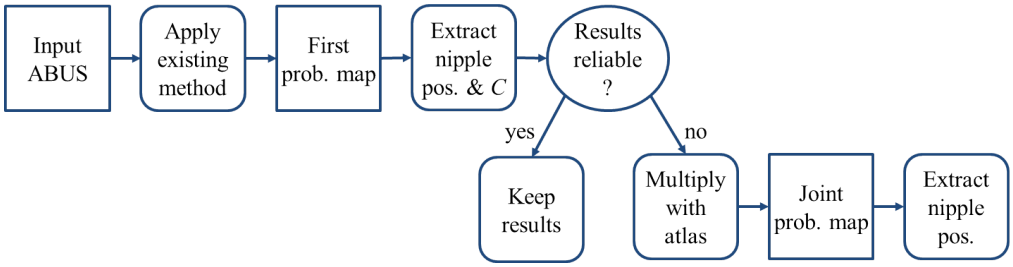


Figure 2.12: Schematic workflow for the proposed joint nipple detection method.

Original nipple detection method As described in detail by Wang et al. (2014), the nipple position is originally retrieved from the multiplication of a 2D Laplacian-based probability map and a 2D Hessian-based probability map. To generate the Laplacian map, a 2D breast mask is computed as minimum intensity projection of a coronal slab of thickness 1.5 mm starting at 1.85 mm from the skin. The nipple position is searched within this breast mask on a nipple slab. This nipple slab is computed from coronal slices starting at 0.35 mm from the skin and going to 1.85 mm. The maximum intensity projection of the nipple slab is first down-sampled and then smoothed by a Gaussian kernel with $\sigma = 3$. The blobness of the nipple (shadow) is measured by a Laplacian of Gaussian (LoG) filter at different scales (Huertas & Medioni 1986; Lindeberg 1998), i.e. σ of the Gaussian kernel ranged from 1.5 mm to 15 mm. The scale that delivers the globally minimal response is selected resulting in a 2D probability map.

A Hessian filter is employed to detect the position of the characteristic tubular acoustic nipple shadow. Eigen values of the Hessian matrix present different specific patterns for various geometrical structures in 3D, e.g. blobs, tubes or disks (Descoteaux et al. 2008). Assuming that the Eigen values are sorted as $\lambda_1 \geq \lambda_2 \geq \lambda_3$, a dark tubular structure will result in a pattern of $\lambda_3 \approx 0$ and $\lambda_1 \approx \lambda_2 \gg 0$. A 3D breast mask is generated by thresholding the ABUS image at its 25^{th} percentile. For each voxel within the mask, the Eigen values are calculated. In order to save computing time, a simplified analysis of Eigen values in introduced: The second Eigen value is summed up over all voxels along the antero-posterior direction resulting in a 2D likelihood map.

Finally, the Laplacian map is inverted and multiplied with the Hessian map such that the most probable nipple position can be extracted as the global maximum of this combined map.

Reliability of original nipple detection method The original automatic nipple detection method described by Wang et al. (2014) determines the local maxima of the Laplacian- and Hessian-based probability map. The local maximum with the highest probability value v_1 in range $[0, 1]$ is considered to be the nipple position. The ratio between v_1 and the second highest local maximum with value $v_2 < v_1$ was chosen as confidence measure C and rescaled to $[0, 100]$ as follows

$$C = \left(\frac{v_1}{v_1 + v_2} - 0.5 \right) \cdot 2 \cdot 100 \tag{2.7}$$

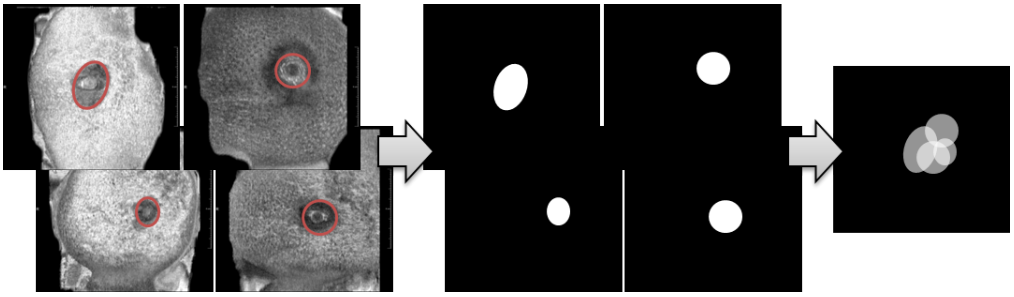


Figure 2.13: Schematic overview of the atlas generation. First, the nipple and areola region were marked in every image by an adapted ellipse (red circles in the left most original images). Then, the arithmetic mean of all binary nipple masks of one view was computed (right most image).

Atlas generation Based on those training images where the nipple was actually visible, one atlas image was created for each of the four available views. Therefore, the nipple and areola region were manually encircled in every single image by an ellipse adapted to the actual size of this region in coronal view. Pixels within the ellipse were given a value of 1, and pixels outside the ellipse were set to 0. The arithmetic mean of all ellipse images was computed for each view separately. Since the images of left and right breast were assumed to be symmetric, they were mirrored along the axial direction, such that all images contributed to the atlas for the left and right breast of each view, respectively. Finally, the atlas images were re-sampled to a spatial resolution of 0.6 mm × 0.6 mm. A schematic overview of the atlas generation is given in figure 2.13.

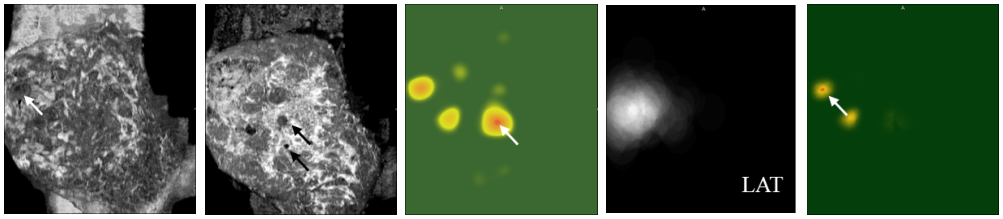
Joint detection method The 2D probability maps of the original nipple detection method were computed for a test data set and re-sampled to the spatial resolution of the atlas images. The most probable nipple position $(x_\mathrm{O}, y_\mathrm{O})$ as well as the confidence measure C were determined as described in equation (2.7). For varying threshold values C^*, the joint detection method was applied to those images with $C < C^*$. In these cases, the existing probability map was multiplied with the corresponding atlas image and the new most probable nipple position $(x_\mathrm{M}, y_\mathrm{M})$ was determined as the global maximum of this joint map (see figure 2.14). Thus, the predicted nipple position $(x_\mathrm{P}, y_\mathrm{P})$ of the joint detection method was

$$(x_\mathrm{P}, y_\mathrm{P}) = \begin{cases} (x_\mathrm{O}, y_\mathrm{O}) & \text{if } C \geq C^* \\ (x_\mathrm{M}, y_\mathrm{M}) & \text{otherwise} \end{cases} \tag{2.8}$$

The root-mean-square distance d in mm from the ground truth nipple position $(x_\mathrm{G}, y_\mathrm{G})$ to the automatically determined nipple position was used as quality measure for the automatic detection method:

$$d = \sqrt{(x_\mathrm{P} - x_\mathrm{G})^2 + (y_\mathrm{P} - y_\mathrm{G})^2}. \tag{2.9}$$

Furthermore, the distribution of nipple detection rates was computed for variable tolerance thresholds of distance errors.



(a) **(b)** **(c)** **(d)** **(e)**

Figure 2.14: Sample case with $C = 12\,\%$ for joint nipple detection method. (a) Coronal image slice, the arrow marks the nipple. (b) Coronal image slice at different depth with misleading structures marked by the arrows. (c) Heat map of probability distribution as computed by the original method. The maximum (arrow) is not the actual nipple position. (d) Corresponding atlas. (e) Multiplication of (c) with (d). The new global maximum (arrow) corresponds to the actual nipple position.

Image Classification

The above described extended nipple detection method was employed to extract 13 features from the ABUS images, which allowed predicting correctness of the manual nipple position annotation made by the technicians during image acquisition.

- The view (AP, MED, LAT, or SUP)
- The distance between the nipple position $(x_\mathrm{T}, y_\mathrm{T})$ that was marked by the technician and the most probable nipple position $(x_\mathrm{P}, y_\mathrm{P})$ determined in the joint detection method described above. If they are very similar, they are most probably correct.
- The distance between $(x_\mathrm{T}, y_\mathrm{T})$ and the most probable nipple position $(x_\mathrm{O}, y_\mathrm{O})$ according to the original detection method.
- The confidence value C produced by the original nipple detection method.
- The value v_T^O of the original probability map at $(x_\mathrm{T}, y_\mathrm{T})$ indicates whether the original algorithm detected structures at this position that could belong to a nipple.
- The value v_T^A of the atlas (=empirical probability map) at $(x_\mathrm{T}, y_\mathrm{T})$ gives an intuition how typical or atypical the marked position is.
- The value v_T^M of the joint probability map at $(x_\mathrm{T}, y_\mathrm{T})$ $(= v_\mathrm{T}^\mathrm{O} \cdot v_\mathrm{T}^\mathrm{A})$.
- The value v_O^O of the original probability map at $(x_\mathrm{O}, y_\mathrm{O})$, which corresponds to the maximum value of the original probability map.
- The value v_O^A of the atlas at $(x_\mathrm{O}, y_\mathrm{O})$. The higher this value, the more typical is the prediction.
- The value v_O^M of the joint probability map at $(x_\mathrm{O}, y_\mathrm{O})$ $(= v_\mathrm{O}^\mathrm{O} \cdot v_\mathrm{O}^\mathrm{A})$. The higher this value, the better might be the prediction.

- The difference between $v_{\mathrm{O}}^{\mathrm{O}}$ and $v_{\mathrm{T}}^{\mathrm{O}}$ indicating which position is more probable.
- The difference between $v_{\mathrm{O}}^{\mathrm{A}}$ and $v_{\mathrm{T}}^{\mathrm{A}}$ indicating which position is more typical for the respective view.
- The difference between $v_{\mathrm{O}}^{\mathrm{M}}$ and $v_{\mathrm{T}}^{\mathrm{M}}$ indicating which position is more probable with regard to the respective view.

A Random Forest classifier was trained on the training data set and applied to the test data set. The Random Forest was built of 100 trees at a maximum depth of 13 layers, using 3 randomly selected features per split. The minimum number of samples at a node for it to be split was 3. For classification, class 1 and 2 images were merged in the positive class, whereas class 0 cases composed the negative class. The trained classifier was evaluated by applying it to the images of the test data set.

2.3 Performance of Automated Image Quality Assessment on disjunct data

The following analysis of the developed algorithms was designed to measure their performance when applied to an unseen data set that was acquired at a different institution and rated manually by a different reader than the training data set. For this purpose, data set C was employed. Note that the algorithms introduced in section 2.2 had been trained on data sets A and B, which are independent from data set C (see table 2.1).

The data was annotated manually by two radiologists regarding the occurrence of inadequate nipple positioning, prominent nipple shadows, or irregular breast contour shapes. Furthermore, the ground truth for the joint image quality rating as described in section 2.2.4 was inferred from this rating. One of the two readers ("Reader 2") had also annotated the ABUS images that were used for algorithm development and classifier training (see section 2.1), and his ratings were thus supposed to be in line with the automatic method. The second reader ("Reader 3") was a senior radiologist with several years of experience in diagnostic breast ultrasound and ABUS image interpretation. He had not been involved in any classifier training steps.

Classification of nipple shadow and relative nipple position requires nipple coordinates as input information. These coordinates were assessed manually in order to evaluate the classifier performance independently of nipple detection or annotation errors, i.e. in a best case scenario. In a second step, the automated joint nipple detection method described in section 2.2.6 was employed to generate the required nipple coordinates yielding a less optimistic evaluation. In both cases, data set C* was used, i.e. 53 images were excluded from the study since the nipple was not visible at all in these scans.

The automatic image quality assessment was designed to produce a probability score between 0 and 1 for each single artefact as well as for the joint score. Specific thresholds to decide whether an image was assigned a positive or a negative rating were derived from classifier training based on the training data sets A and B. From the resulting ROC statistics, the decision thresholds corresponding to distinct specificities were determined (see section 3.2). In this case, a target specificity of 0.97 was sighted, i.e. the decision thresholds were chosen such that a specificity of 0.97 was achieved in the training data.

The manual annotations of the two clinicians were evaluated by measuring the inter-rater agreement using Cohen's κ coefficient. The same metric was applied to compare the results of the automated image quality assessment to those of each single reader. Furthermore, specificity, sensitivity and ROC statistics were used to depict the congruence between automatic and manual image quality rating.

When combining the annotations of two readers to one quality rating per image, there are different options. For classifier training, an image was considered to be affected by a certain artefact if and only if both readers detected this specific artefact in the respective image. This mode was aiming at a classification with high specificity since only those images that clearly contained the artefact were considered as positive. Ambiguous images were assigned the negative class. The same approach was chosen for the present study.

2.4 Clinical Implementation

In order to evaluate three of the developed automated quality assessment (AQUA) algorithms in clinical routine, they were implemented into an already existing framework for data handling and workflow management. The three AQUA modules included in the image processing pipeline were designed to detect images with an inappropriate nipple position, a prominent nipple shadow or with a very irregular breast contour shape. In the following, the technical aspects of the installation at Radboud University Medical Centre (Nijmegen, The Netherlands) as well as the different methods of evaluation will be described. The performance of this prototype was measured in terms of usability for the technicians.

2.4.1 Technical Aspects

The employed software framework, called MIRIAM (Medical Image pRocessIng And Management) consists of a database for medical images and a workflow engine. It comprises several back-end components which deliver the functionality of the medical image database. Diverse front-end components represent the GUI or provide a connection to other software systems.

The complete back-end logic is implemented in Java on the basis of a Java Application Server which is EE 6 (Enterprise Edition) compliant. Amongst others, the back-end contains the workflow engine of the system. It relies on a SOUP (Software Of Unknown Provenance) component called “Activiti” which accomplishes workflow related tasks. BPMN 2.0 (Business Process Model and Notation) workflow processes can be deployed to the workflow engine to fulfil a specific customer need and guide the user through an interactive pipeline. Apart from that, the workflow engine can be configured to execute non-interactive tasks. It can, e.g., send mails or perform automatic data manipulation actions. In addition, it can wrap an AQUA module and execute it in the context of an active workflow. The MIRIAM front-end provides different kinds of secured and structured access to the data stored in the back-end. For example, MIRIAM provides several web based user interfaces to interact with the end user via a modern web browser using the SOUP component “Vaadin”, which is a Java framework to build modern web applications. There are different interfaces for users with and without administrative access rights. The administration user interface provides system relevant operations like administration of users and roles, or system configurations. The simple user interface allows, e.g., executing interactive workflow activities or viewing medical image data.

The framework retrieved the image data from a DICOM node in the clinic where the images were sent directly via network after acquisition. Together with a configuration file in XML-format, the image data was passed through the different AQUA modules, which operated on the images and produced output images, e.g. breast masks, as well as meta information (in XML), e.g. computed features or classifier rating results. These were captured and evaluated by the framework for an image quality rating.

The chosen set up for the integration into clinical routine is depicted in figure 2.15. Generally, the images are acquired by a technician and stored automatically on a server PC. The standard protocol does not include any image quality

checks before the radiologist performs the diagnostic read. If the images are of low quality, proper diagnosis might be hampered or even impossible, which could entail a recall of the patient. This might take several days and cause anxiety to the patient. Following the proposed extended workflow, in case of a low-quality image, the technician was supposed to be alerted and provided with recommendations how to improve the image. With the patient still present in the examination room, it should be no problem for the technician to redo the scan.

Technically, this was achieved by an additional PC in the examination room, which was connected via internal network to the AQUA server PC. After login, the technicians launched the MIRIAM application in the web-browser where the current examination with the corresponding rating results were displayed timely.

2.4.2 Usability

The usability of the above described software prototype was evaluated by seven technicians who were asked to fill in a questionnaire after the first usage of the system. The questionnaire is attached in Appendix A.

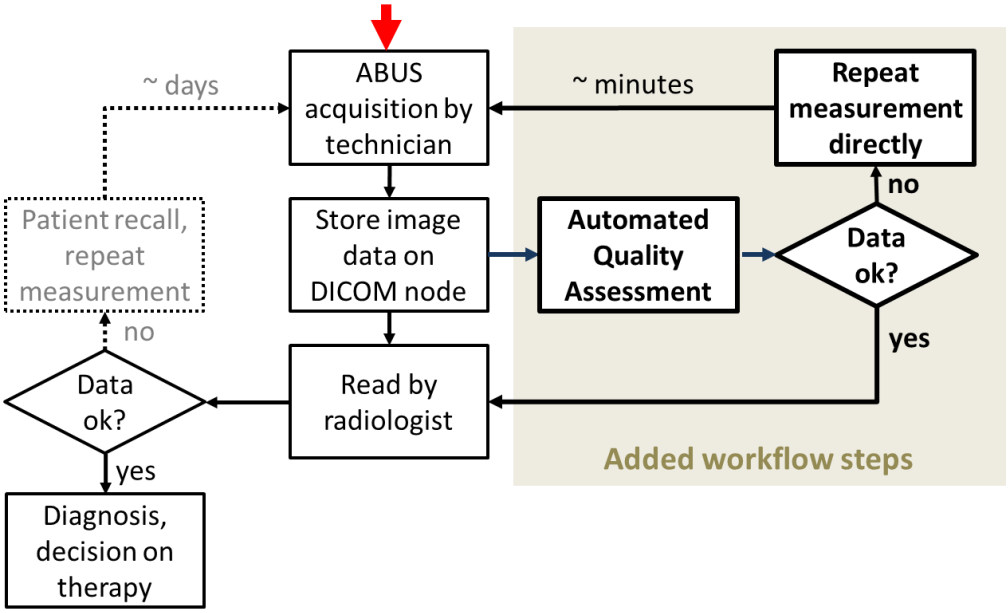


Figure 2.15: Proposed workflow for ABUS AQUA in clinical routine. The current standard procedure does not include any image quality check before the radiologist performs the diagnostic read. The new steps of automated image quality assessment will only take a few minutes.

3 Results

3.1 Empirical Analysis of ABUS Artefacts

In order to facilitate and standardize the manual annotation of a large amount of ABUS images with respect to diverse potential image artefacts, an unpretentious software suite was developed. The tool consisted of a comprehensive medical image viewer, check boxes for all the artefacts that should be considered, and a text box for free comments. The viewer displayed one ABUS image at a time in the three available orthogonal views and provided standard scrolling, zooming, and windowing operations. A screenshot of the rating tool is shown in figure 3.1. Rating results were stored locally in XML format and could be reloaded and changed if necessary.

A data set of 368 ABUS images (A and B) was manually annotated by two medical researchers ("Reader 1" and "Reader 2") with respect to the occurrence of specific image artefacts. The inter-rater agreement was analysed as well as the relative frequency of each artefact.

Inter-rater Agreement

The inter-rater agreement of the two readers is listed for each artefact in table 3.1. The κ value varied strongly between the different image artefacts, indicating that some of them were easier to define objectively than others. The position of the nipple relative to the breast contour line on the image was a clear criterion, which reached a very good agreement with $\kappa = 0.84$. Visibility of at least three ribs, breast contour shape and nipple shadow had κ values between 0.41 and 0.71, which expressed acceptable agreement. The wavy pattern and the air artefacts showed low κ values around 0.3, and the rating of discontinuities yielded a negative κ. The image quality rating based on the latter mentioned artefacts differed significantly between the two raters.

Three sample images that the two readers disagreed upon are shown in figure 3.2. The relative position of the nipple is very close to the contour line of the breast (and to the edges of the image) in figure 3.2a, which made Reader 1 noting an insufficient nipple position down. On the other hand, the nipple area is still visualized completely explaining why Reader 2 did not mark this image. The acoustic shadow caused by the nipple in figure 3.2b was perceived as too prominent by Reader 1, whereas Reader 2 was content with the information that still could be retrieved from the area behind the nipple. The case shown in figure 3.2c is a large breast, which renders breast contour shape rating more difficult. Reader 1 rated this case as positive emphasizing that there is a large background region in the upper left corner albeit the breast is not imaged completely. Reader 2 was focussing on the shape of the contour, which is indeed very smooth in this case and, thus, rated it as negative.

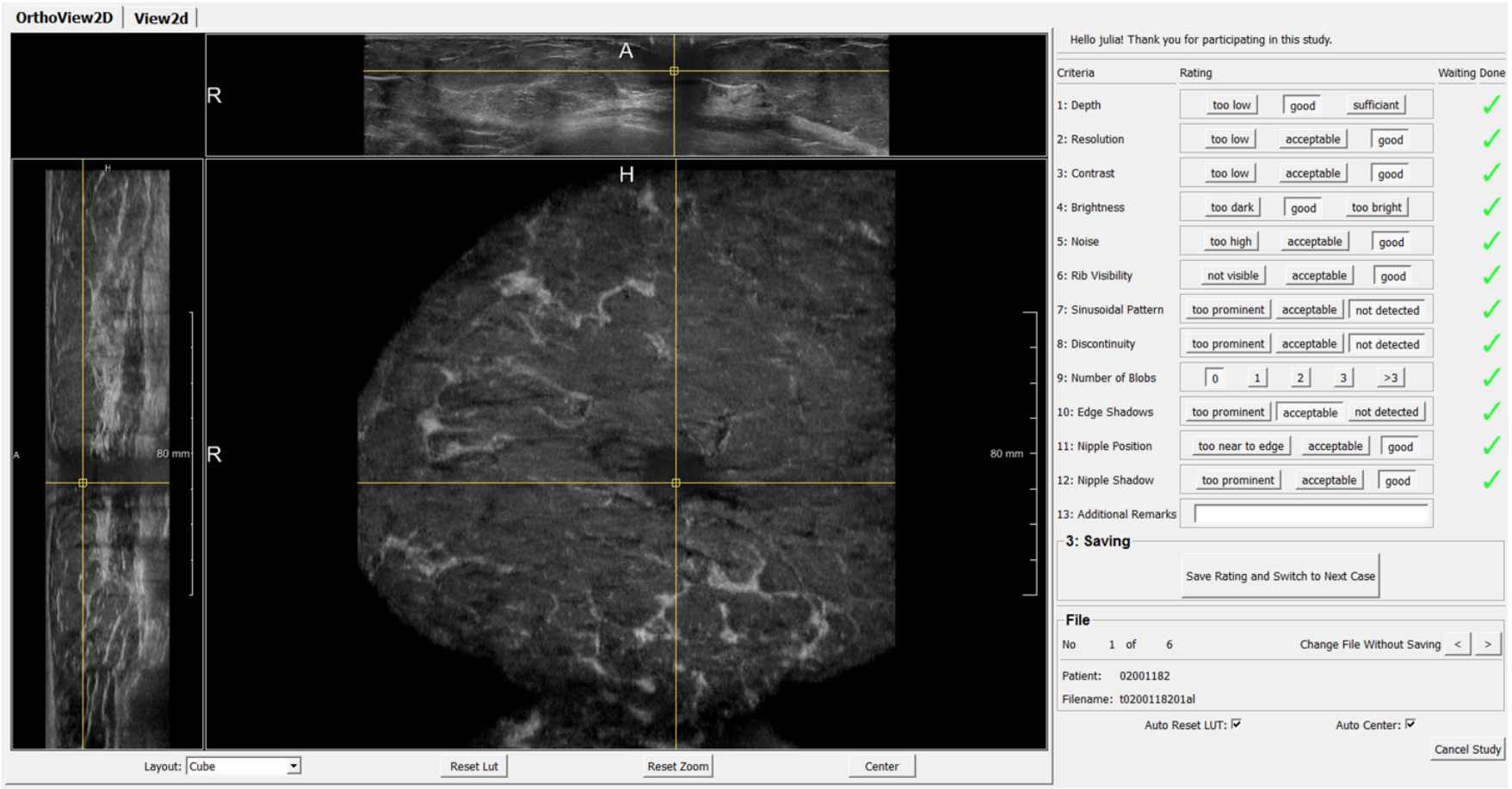
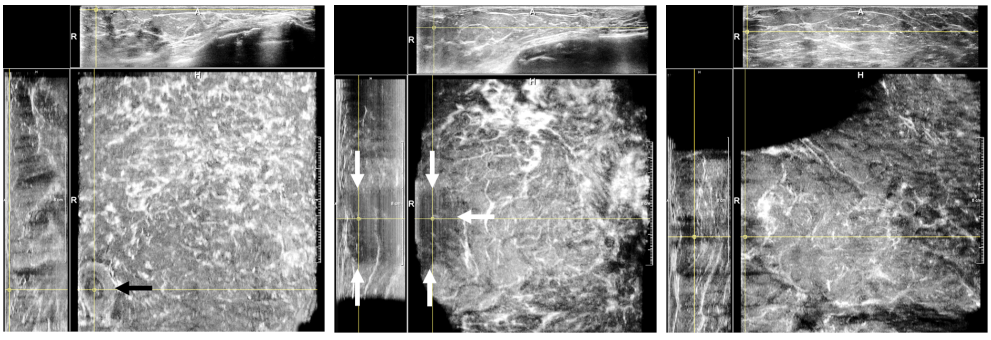


Figure 3.1: Screenshot of the dedicated ABUS artefact annotation software tool. The viewer allows scrolling through the entire volume. The buttons on the right provide all predefined options for possible artefacts.

Table 3.1: Computed inter-rater agreement for manual annotation of specific ABUS artefacts

Artefact	*Cohen's* κ
Relative nipple position	0.84
Visibility of ribs	0.71
Breast contour shape	0.61
Nipple shadow	0.44
Wavy pattern	0.30
Air artefacts	0.32
Discontinuities	< 0



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

Figure 3.2: Sample images that the two readers disagreed upon concerning the classification of (a) the relative nipple position, (b) the nipple shadow, and (c) the breast contour shape.

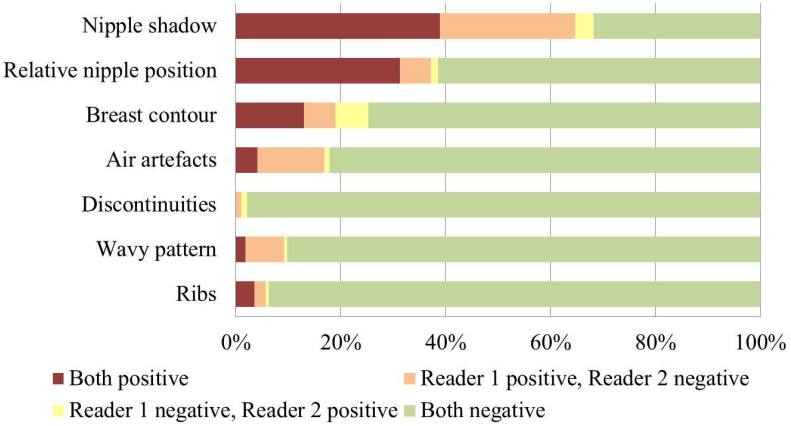


Figure 3.3: Frequency of specific artefacts in the considered data set (368 images) as annotated by two raters. Red bars indicate the number of images that were rated concordantly as showing the artefact.

Relevance and Frequency of Artefacts

The results of the manual annotation process are shown in figure 3.3. The red bars indicate the amount of images that were rated as showing the specific artefact by both readers. The green bars represent the images with no artefacts found, whereas the orange and yellow bars stand for the images which got one positive and one negative vote. Nevertheless, the general trend for the incidence of specific artefacts is clearly visible and similar for both readers. The shadow caused by the nipple is the most prominent artefact which was marked by both observers in almost 40 % of the images. The relative position of the nipple in the ABUS image was found by both raters to be too close to the contour line of the breast in 31 % of the considered images. The third most frequent imaging issue was a very irregular shape of the breast contour line which was annotated in 13 % of the images. The inter-rater agreement for these three artefacts ranged from acceptable to good. The other four considered artefacts were detected concordantly in less than 10 % of the images and yielded generally low inter-rater agreements.

3.2 Computer-aided analysis of ABUS Artefacts

3.2.1 Relative Nipple Position

Feature extraction and classifier training based on 340 manually annotated images (data sets A* and B*) was performed in order to detect those images where the nipple was pushed too far aside during transducer positioning. In total, nine features characterizing the position of the nipple relative to the rest of the breast were computed per image. The average computing time for all features was (3 ± 2) s per volumetric image. 10-run 10-fold stratified cross validation was done to evaluate the reliability of the selected features.

A feature ranking test (see section 1.5) applied to the whole data set revealed the relative relevance of the single features with respect to the classification of the nipple position. The higher a feature was ranked, the higher its discriminant capacity was. The six top ranked features are listed in table 3.2. The values differ clearly from each other and allow a clear sorting of the features. Since the first three features are highly correlated to each other, as can be seen in the feature space plot in figure 3.4a, the 2D feature space spanned by the first and the fourth feature, $d*_{\mathrm{min}}$ and d_{COM}, are plotted in figure 3.4b. One can see that these features already offer a fair distinction between the two classes. This potential was exploited and amplified by applying a Random Forest classifier based on all available features.

In figure 3.5 two ROC curves of the proposed nipple position classification are plotted. Figure 3.5a shows sample ROC curves of one run of 10-fold cross-validation. The thin lines represent the single folds; the bold line is the merged ROC curve. In figure 3.5b the merged ROC curves of the ten runs are plotted. Partly due to the small number of cases within one fold, the ROC curves of the single folds vary clearly. The merged ROC curves, however, are all very similar indicating the stability of the proposed classification model. Furthermore, the sensitivity and specificity of each reader when compared to the other one, respectively, are marked in figure 3.5b. It can be seen that they only performed slightly better than the algorithm, i.e. Reader 1 achieved a sensitivity of 0.95 and a specificity of 0.91 if the rating of Reader 2 is considered ground truth, whereas Reader 2 has a sensitivity of 0.80 and a specificity of 0.98 in the opposite case.

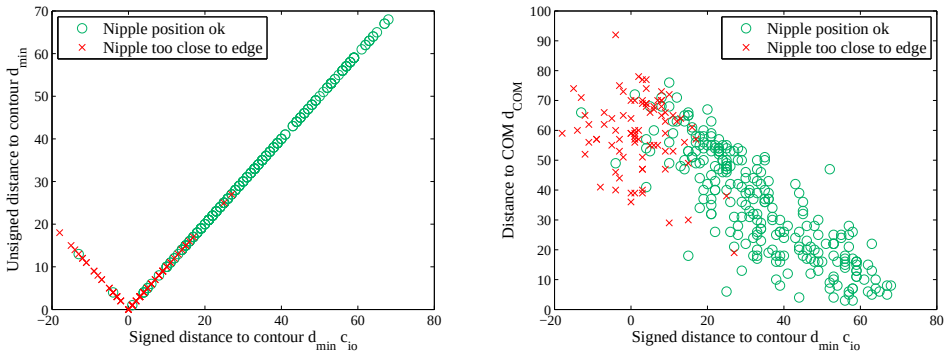
The ten repetitions of 10-fold cross-validation resulted in an $\mathrm{AUC}_{\mathrm{merge}}$ of 0.987 ± 0.002 (95 % CI). Numbers for specificity, sensitivity and F-Measure are given in table 3.3 for different operating points along the ROC curves. The mean value of specificity at the best cut-off point of the ten runs was 0.91 correlating with a sensitivity of 0.93. Other points along the ROC curve could however be chosen, e.g. aiming at a very high specificity to avoid false positive classifications. In this case, the proposed Random Forest classifier achieved a sensitivity of 0.36 at a specificity of 0.99.

In figure 3.6, extreme outlier cases are shown. A false positive case is shown in figure 3.6a where the breast is very large and not completely visible in the image. In this case, the breast mask fails to describe the true contour of the breast. The breast in figure 3.6b is small and skinny which impedes proper ultrasound coupling. As a consequence, a bright rectangle caused by reflections is visible in the upper right corner of the image and breast mask segmentation using Otsu's

Table 3.2: Features for classification of relative nipple position ranked after their information gain ratio

Info gain ratio	*Feature*
0.4492	Signed distance to contour $d*_{\text{min}} = d_{\text{min}} \cdot c_{\text{io}}$
0.4092	Unsigned distance to contour d_{min}
0.3522	Is the nipple inside the breast mask? c_{io}
0.1674	Distance to COM d_{COM}
0.1258	x_{T}
0.0786	c_{view}
0	all others

filter failed. Figure 3.6c shows a false negative case caused by the irregular breast contour shape of the breast, which in turn produces an erroneous breast mask.



(a) **(b)**

Figure 3.4: 2D feature space plots of (a) first and second, and (b) first and fourth features as ranked according to the information gain ratio measure for nipple position classification.

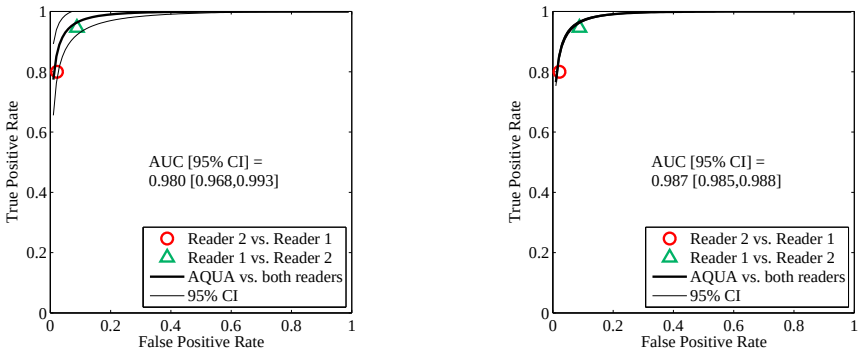


Figure 3.5: Sample (a) and mean (b) ROC curves of the 10 runs of 10-fold cross-validation for the classification of the relative nipple position.

Table 3.3: Performance measures of the proposed classification of the relative nipple position obtained in 10-run 10-fold cross-validation

Spec.	95 % *CI*	*Sens.*	95 % *CI*	F_{avg}	95 % *CI*	$F_{Pr,Re}$	95 % *CI*
0.905	0.009	0.926	0.013	0.882	0.010	0.886	0.010
0.950	0.002	0.780	0.011	0.882	0.010	0.887	0.010
0.969	0.001	0.594	0.024	0.882	0.010	0.887	0.010
0.990	0.002	0.363	0.064	0.697	0.026	0.725	0.025

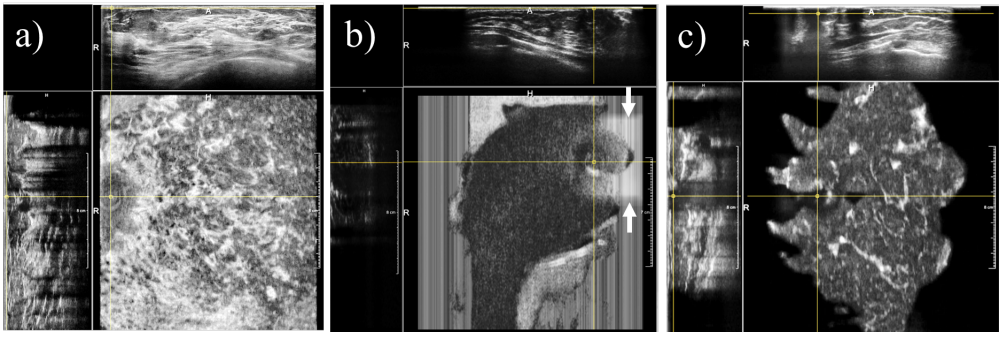


Figure 3.6: Examples for outliers of the nipple position classification. (a) shows a false positive case where a significant part of the breast is not visible in the scan. (b) is a false negative due to an erroneous breast mask caused by intense reflections within the coupling layers of the transducer. (c) shows a false negative case caused by the irregular breast contour shape.

3.2.2 Nipple Shadow

Seven features describing the nipple were extracted from every image. In total, 340 ABUS images (instances) with a clearly visible nipple were included in this study. On average, it took (5 ± 2) s per ABUS image to compute all features.

The information gain ratio feature ranking test was applied to all instances of the data set. The feature ranking is listed in table 3.4. The values are close to each other but still show a clear order, indicating different significance for classification of the nipple shadow. The feature space of the two top ranked features, $N_{I<50}$ and $N_{I<60}$, is shown in figure 3.7 and proves a clear tendency of positive instances towards high numbers of low intensity cylinder segments and of negative instances towards low values of $N_{I<50}$ and $N_{I<60}$. Obviously, a clear distinction between the two classes would not have been possible based only on the top ranked features, which speaks in favour of the applied Random Forest classification.

Automatic classification of the nipple shadow yielded an $\mathrm{AUC}_{\mathrm{merge}}$ of 0.842 ± 0.006. Sample ROC curves of one run of cross-validation as well as of all ten repetitions are plotted in figure 3.8. The best cut-off point on the ROC curve yields a specificity of 0.82 and a sensitivity of 0.73. As can be seen in table 3.5, at a specificity of over 0.99, sensitivity of the nipple shadow classification is still 0.24. If the rating of Reader 1 is considered as ground truth, the sensitivity and specificity of Reader 2 are 0.56 and 0.89, respectively, and therefore very close to the performance of the classifier (see figure 3.8). The same accounts for the reverse case, where Reader 1 achieves a sensitivity of 0.90 and a specificity of 0.56 when compared to Reader 2.

Figure 3.9 shows three sample outlier cases. The false positive case in figure 3.9a is a small and skinny breast with a clearly visible nipple shadow close to the breast contour line. However, it was rated as negative by the readers since it is hardly possible to get better images of such a small breast in the present view and a repeated scan probably would not enhance the image. Figure 3.9b shows a false negative case where the dark region is not directly below the nipple but rather in a half ring around it. In figure 3.9c, the false negative classification was caused by a relatively bright and fuzzy shadow. However, the algorithm was designed to detect very prominent, low intensity nipple shadows as shown in figure 2.4a.

Table 3.4: Features for classification of nipple shadow ranked after their information gain ratio

Info gain ratio	*Feature*
0.3199	Number of segments $N_{I<50}$
0.2331	Number of segments $N_{I<60}$
0.2132	Number of pixels N_{Pix} in segments with $I < 60$
0.1471	Variance $\sigma^2_{\mathrm{bright}}$ in central cylinder segments
0.0873	x_T
0.0680	c_{view}
0	y_T

Table 3.5: Performance measures of the proposed classification of the nipple shadow obtained in 10-run 10-fold cross-validation

Spec.	95 % *CI*	*Sens.*	95 % *CI*	F_{avg}	95 % *CI*	$F_{Pr,Re}$	95 % *CI*
0.817	0.021	0.728	0.022	0.822	0.012	0.830	0.012
0.953	0.002	0.513	0.015	0.764	0.015	0.771	0.015
0.972	5E-18	0.454	0.020	0.676	0.021	0.693	0.019
0.991	0	0.240	0.040	0.664	0.019	0.682	0.018

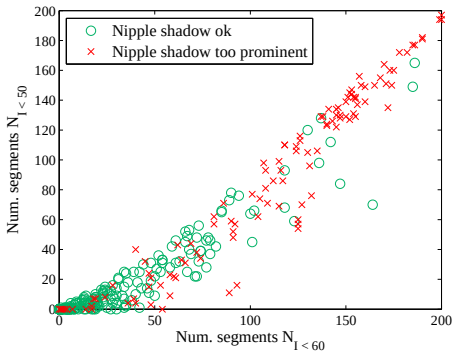
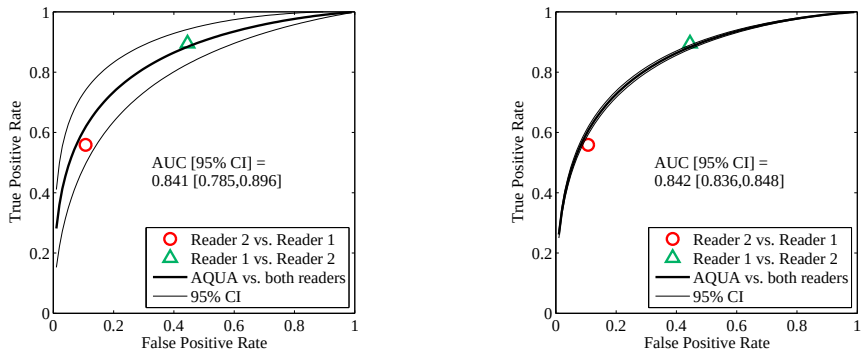


Figure 3.7: 2D feature space plot of the two top ranked features according to the information gain ratio measure for nipple shadow classification.



(a) **(b)**

Figure 3.8: Sample (a) and mean (b) ROC curves of the 10 runs of 10-fold cross-validation for the classification of the nipple shadow.

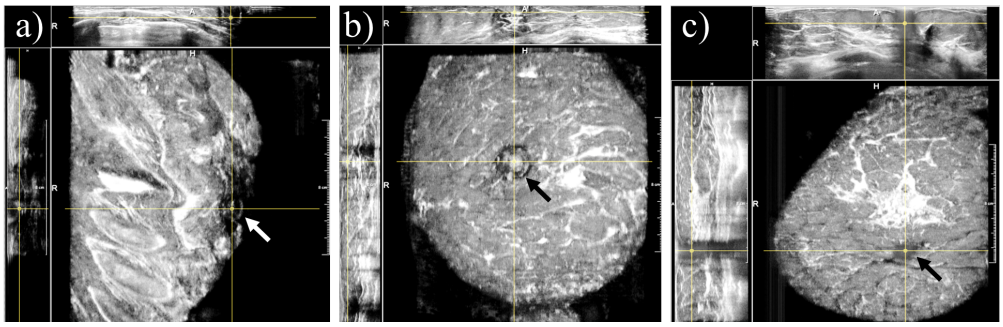


Figure 3.9: Incorrectly classified cases of the nipple shadow. (a) is a false positive case caused by the nipple being very close to the breast contour line. (b) is a false negative with a structured, ring-like nipple shadow. (c) shows a false negative case with fuzzy and bright nipple shadow.

3.2.3 Breast Contour Shape

In order to detect irregularities, e.g. shadows, along the breast contour line in the coronal plane of an ABUS image, 17 features were extracted and used as input for a Random Forest classifier. 10-run 10-fold cross-validation was performed to analyse the discriminant capacity of the computed features based on data sets A and B. The average computing time for feature extraction was (6 ± 4) s.

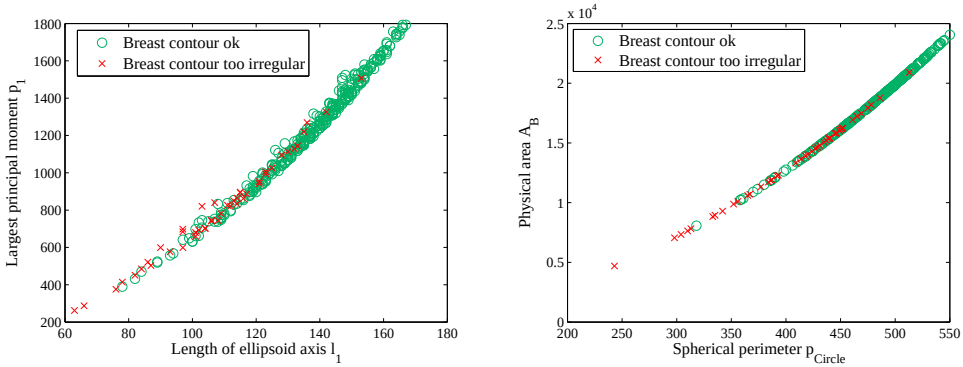
A feature ranking test was applied to all instances of the data set to measure the information gain ratio of the single features with respect to the classification of the breast contour shape (see section 1.5). The 13 features that had an empirical information gain ratio above 0 are listed in table 3.6. The computed values of information gain ratio are very close to each other and all below 0.19, indicating that their discriminant character is not very pronounced. This can also be seen in figure 3.10a, where the 2D feature space spanned by l_1 and p_1 is plotted. A tendency of positive instances (red crosses) towards lower values of p_1 and l_1 is observable but not distinctive. As can be seen in figure 3.10a, l_1 and p_1 are correlated, and thus, they are ranked similarly by the information gain ration test. Features that are totally dependent on each other, e.g. the area of the breast mask and the equivalent spherical radius and perimeter, are given exactly the same value of information gain ratio (see table 3.6).

Random Forest classification was performed under consideration of all available features and yielded the following results. A sample ROC plot of one run illustrating the merged ROC curve as well as the 95 % confidence interval is shown in figure 3.11a. Furthermore, the mean of all merged ROC curves of all runs is plotted in figure 3.11b. One can see that the merged curves are all very similar in shape, i.e. the AUC has a very small confidence interval. The performance of Reader 1 when compared to Reader 2 and vice versa was very similar to the classifier performance, i.e. the sensitivities were 0.69 and 0.68 at specificities of 0.92 and 0.93, respectively. The area under the curve was determined as $\mathrm{AUC}_{\mathrm{merge}} = 0.885 \pm 0.003$ (95 % CI). As can be seen in table 3.7, the specificity at the best cut-off point of the ROC curve is 0.82 yielding a sensitivity of 0.79. Depending on the preferred operation mode of the classifier, different decision thresholds for the computed class probabilities can be chosen. Putting the focus, e.g., on a very high specificity (0.99), the sensitivity decreases to 0.15.

Figure 3.12a shows a sample false positive case. The breast as such is imaged correctly, but parts of the axilla and the arm cause atypical contour lines, which are misinterpreted by the classifier. Figures 3.12b and c show false negative cases where parts of the breast are not imaged correctly. Nevertheless, the breast mask has smooth contours obscuring missing parts and misleading the classifier.

Table 3.6: Features for classification of breast contour shape ranked after their information gain ratio

Info gain ratio	*Feature*
0.1879	Length of ellipsoid axis l_1
0.1837	Largest principal moment p_1
0.1745	Spherical perimeter p_{Circle}
0.1745	Physical area of breast mask A_{B}
0.1745	Spherical radius r_{Circle}
0.1672	Relative breast area $A_{\text{B/I}}$
0.1412	Ellipsoid axis l_2
0.1272	Principal moment p_2
0.1122	Flatness F
0.0883	Centroid coordinate x_{C}
0.0714	Roundness R_{Round}
0.043	Perimeter p_{Mask}
0.0196	c_{view}
0	all others



(a) **(b)**

Figure 3.10: 2D feature space plots of (a) first and second, and (b) third and fourth features as ranked according to the information gain ratio measure for breast contour shape classification.

Table 3.7: Performance measures of the proposed classification of the breast contour shape obtained in 10-run 10-fold cross-validation

Spec.	95 % *CI*	*Sens.*	95 % *CI*	F_{avg}	95 % *CI*	$F_{Pr,Re}$	95 % *CI*
0.822	0.044	0.787	0.042	0.653	0.020	0.688	0.019
0.950	0.001	0.566	0.029	0.653	0.020	0.688	0.020
0.969	6E-4	0.455	0.049	0.653	0.020	0.688	0.019
0.990	9E-4	0.149	0.025	0.436	0.057	0.464	0.059

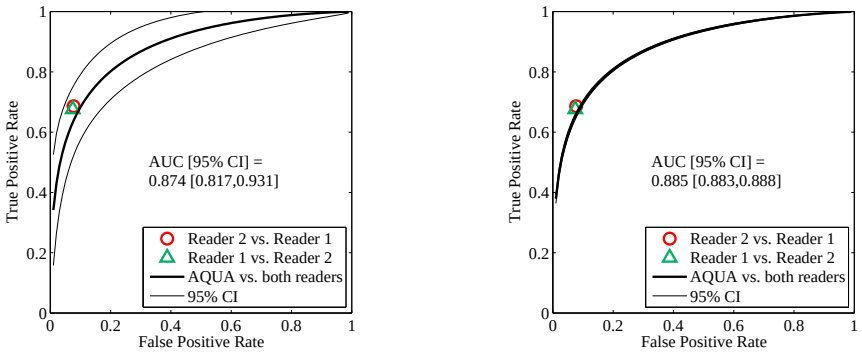


Figure 3.11: Sample (a) and mean (b) curves of the 10 runs of 10-fold cross-validation for the classification of the breast contour shape.

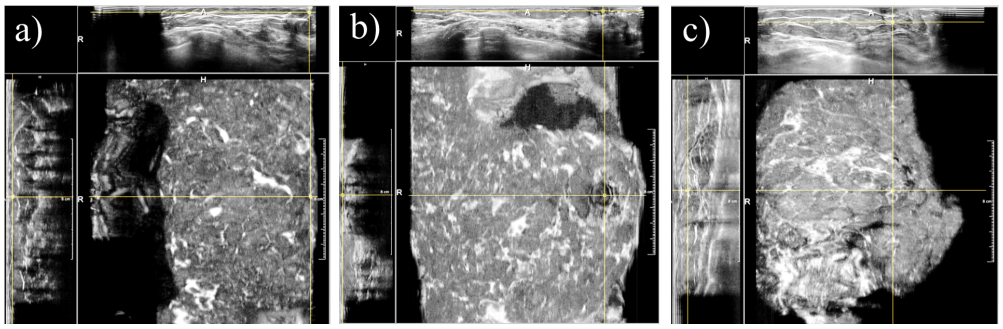


Figure 3.12: Examples for outliers of the breast contour classification: (a) is a false positive case where parts of the axilla and the arm are visible on the image. The false negative cases in (b) and (c) show relatively smooth contours obscuring the fact that parts of the breast are not imaged correctly.

3.2.4 Joint Image Quality Rating

Since the three artefacts described hitherto are correlated to each other with respect to their origin and their appearance, the features characterizing the single issues were combined to establish a joint image quality rating. The simultaneous appearance of these three artefacts was evaluated in data sets A* and B* as shown in figure 3.13. The numbers in the overlapping regions indicate the relative amount of images that showed two or three artefacts at the same time. Out of the 340 included images, 67 showed only one of the three considered imaging issues, whereas 76 were affected by at least two artefacts at the same time. These numbers support the assumption that there is a correlation between these three artefacts and a general, joint image quality measure might be a reasonable implementation. According to the expert annotation, the most relevant overlap was observed for issues related to the nipple: 13 % of the images were afflicted simultaneously by a prominent nipple shadow and an inadequate nipple position. Furthermore, a co-occurrence of all three considered artefacts was detected in 7 %.

A feature ranking test (see section 1.5) was applied to all 29 features that were included in this joint approach. As shown in table 3.8, the top ranked features lie very close to each other indicating that these artefact specific features might not be specific enough for general classification when considered separately from the other features. Altogether, however, they yielded good classifier performance as shown below. Three of the four top ranked features were designed to classify the nipple position with respect to the rest of the breast in the image. This shows how the nipple position is correlated to general image quality. Most other top ranked features originated from the nipple shadow characterisation. It is reasonable that a feature describing the most frequent artefact plays an important role in joint image quality classification.

The ROC curves plotted in figure 3.15 indicate a good performance of the proposed classifier. In the ten runs of cross-validation, an AUC_{merge} of 0.935 ± 0.002 (95 % CI) was measured. Similarly to the ROC curves of the dedicated, artefact specific classifiers presented in the previous sections, the deviations between the different runs of cross-validation are very small. This shows that the proposed method is stable and robust towards random re-sorting of instances and folds. Again, the classifier performance was very similar to that of both readers when compared to each other. Given the rating of Reader 1 as ground truth, Reader 2 achieves a sensitivity of 0.65 at a specificity of 0.95. In the opposite case, Reader 1 has a sensitivity of 0.97 and a specificity of 0.52.

The mean ROC curve of the ten runs depicted in figure 3.15 shows a very steep gradient at high specificities (low false positive rate on the abscissa) up to a sensitivity of 0.6. From that point, the curve has a minor slope, meaning that a further increase in the true positive rate could only be achieved by accepting a strong increase in the false positive rate.

Table 3.8: Features for joint classification ranked according to their information gain ratio

Info gain ratio	*Feature*
0.2537	Signed distance between nipple and contour $d_{\text{min}} \cdot c_{\text{io}}$
0.2482	Unsigned distance between nipple and contour d_{min}
0.2406	Number of segments $N_{I<50}$
0.2311	Is the nipple inside the breast mask? c_{io}
0.2237	Number of segments $N_{I<60}$
0.1507	Distance between nipple and COM d_{COM}
0.1470	Variance σ^2_{bright} in central cylinder segments of niple shadow
0.1457	Number of pixels N_{Pix} in nipple shadow segments with $I < 60$
0.1280	x_{T}
0.0917	Roundness of breast contour R_{Round}
0.0853	c_{view}
0	all others

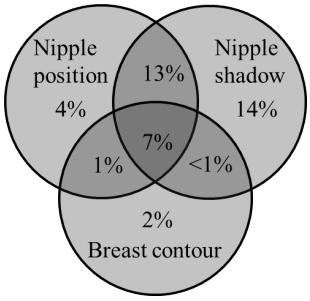


Figure 3.13: Co-occurrence of prominent nipple shadow, irregular breast contour shape and inadequate nipple position in ABUS images of data sets A* and B*.

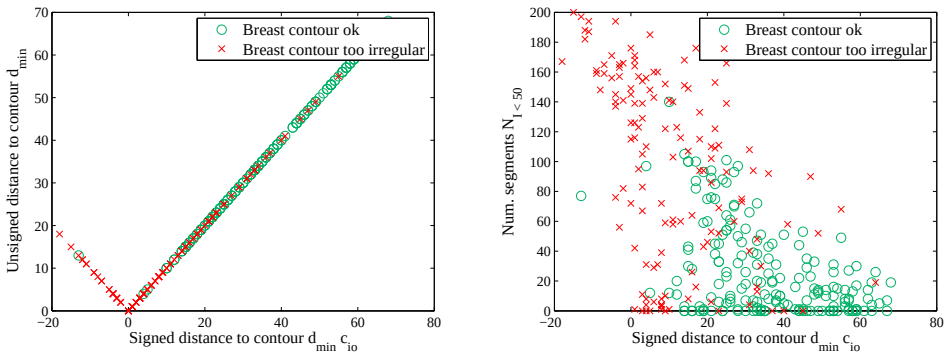


Figure 3.14: 2D feature space plots of (a) first and second, and (b) first and third feature as ranked according to the information gain ratio measure for joint image quality classification.

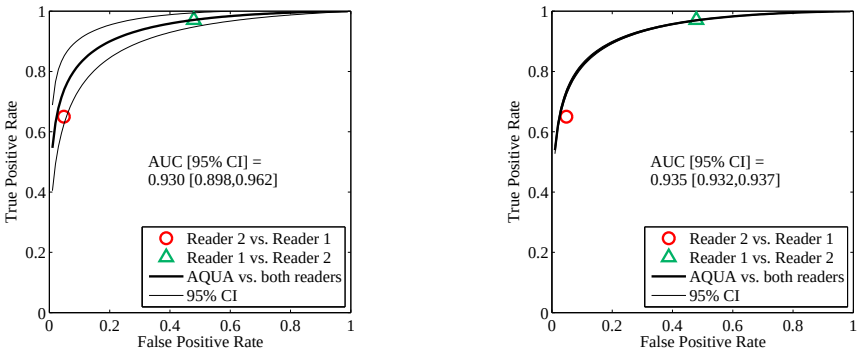


Figure 3.15: Sample (a) and mean (b) curves of the 10 runs of 10-fold cross-validation for the joint classification.

Table 3.9: Performance measures of the proposed joint classification obtained in 10-run 10-fold cross-validation

Spec.	95 % *CI*	*Sens.*	95 % *CI*	F_{avg}	95 % *CI*	$F_{Pr,Re}$	95 % *CI*
0.909	0.012	0.810	0.012	0.866	0.003	0.871	0.003
0.950	0.001	0.734	0.007	0.826	0.008	0.832	0.006
0.970	0.000	0.697	0.012	0.827	0.008	0.832	0.006
0.990	1E-18	0.554	0.036	0.773	0.017	0.793	0.014

3.2.5 Air Artefacts

Properties of Air Artefacts

Based on the manually outlined artefact regions, general properties of air artefacts were examined within the available 79 masked ABUS images. This yielded 126 annotated air artefact regions to be considered. The results are presented in table 3.10 and in the histogram plots of figure 3.16. The mean depth of the stripe pattern was 5 mm and the mean depth of the shadow was 23 mm, which is in good agreement with the chosen depths for V_s and V_l. The mean size of the marked air artefacts was 109 mm^2 in the coronal plane, whereas the smallest annotated artefact was 21 mm^2. This confirmed that a sliding window size of 3 mm × 3 mm in coronal plane, which was approximately half the minimum size of artefacts, was a reasonable choice for feature extraction. The median measured size was 75 mm^2, and 25 out of the 126 artefact regions even exceeded 160 mm^2. Considering the reported mean sizes of ABUS detected lesions[1] of 80 mm^2 to 161 mm^2, it becomes clear that air artefacts indeed could occlude whole lesions. The distance between the bright stripes in air artefact regions had a mean value of 1.2 mm with a low standard deviation of 0.2 mm, confirming that the pattern is very similar in images produced by the same transducer.

Evaluation of the Considered Features

The feature ranking test, when applied to all instances of the training data, showed that there was no single overly discriminant feature, but that there were many features with similar weight. The computed values of information gain ratio are overall very low (the highest being 0.027 for the kurtosis k_l).

Fourier Transform was used to analyse the frequency distribution of V_s. The expected frequency f_e of the one dimensional projection of V_s was $f_e = V_s/p_e = 5\,\text{mm}/1.2\,\text{mm} = 4.2$. As shown in the Fourier spectrum of figure 2.8c, the peak $Y(f_e)$ around 4.2 is pronounced for the characteristic stripe pattern compared to normal tissue signals. As shown in figure 3.17e, the artefact region can be estimated from the $Y(f_e)$ parameter map. Furthermore, the peak $Y(f_{low})$ below

[1] Reported mean diameters d of lesions detected in ABUS screening images range from 10 mm (Brem et al. 2015) to 14.3 mm (Giuliano & Giuliano 2013). Assuming a round shape of lesions, this yields an area of $(d/2)^2 \cdot \pi = 80 \ldots 161\,\text{mm}^2$

Table 3.10: Empirical properties of the 126 considered air artefact regions segmented in 79 masked ABUS images.

	Area in coronal plane in mm^2	*Depth of stripe pattern in* mm	*Depth of shadow in* mm	*Distance of stripes in* mm
Min	21	2.0	4.9	1.0
Max	607	9.5	55.1	2.1
Mean	109	5.3	22.8	1.2
Stddev	92	1.1	9.2	0.2
Median	75	5.2	22.4	1.2

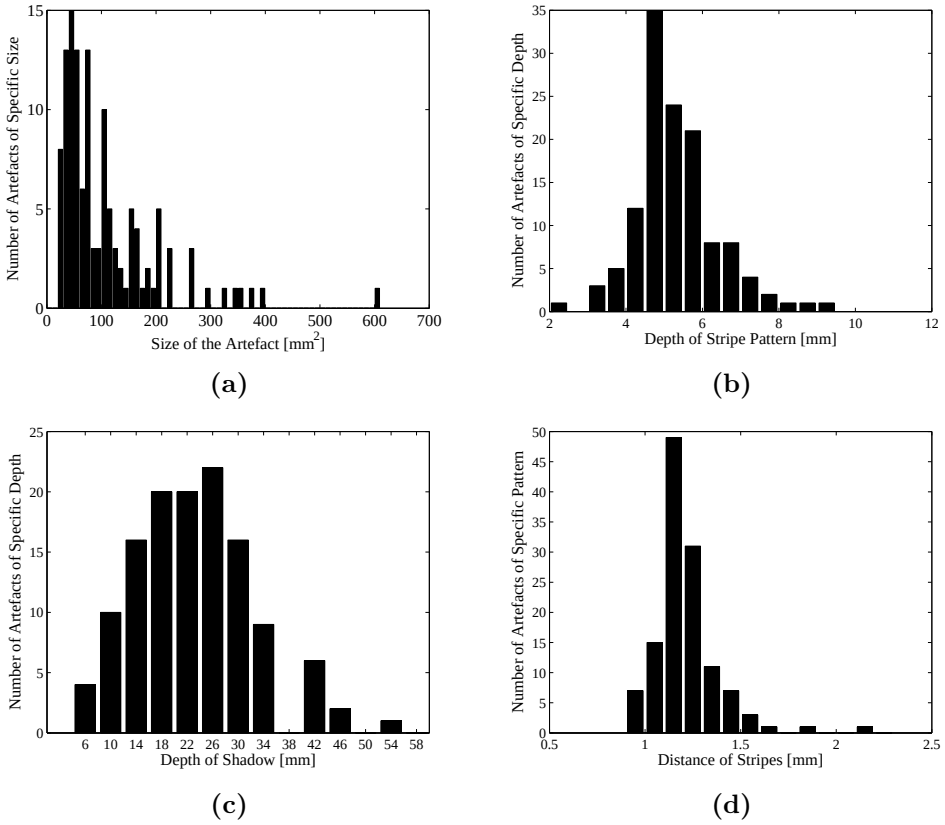


Figure 3.16: Histogram plots for the empirical analysis of air artefacts: (a) artefact size in coronal plane, (b) depth of stripe pattern, (c) depth of acoustic shadow, and (d) distance between single stripes. Only artefacts within the masked breast image and with a minimum region size of 20 mm^2 in the coronal plane were considered.

2 Hz and the correlated slope b of the sine fit also provided valuable information for classification as indicated by the parameter maps in figures 3.17d and 3.17f. Average computing time for feature extraction was (187 ± 30) s per 3D image. The computing times per window position were also analysed for the different types of features as shown in table 3.11.

The feature space plot in figure 3.18 shows the values of μ_l and $Y(f_{low})$ for all training samples. It can be seen that these two features complement each other in separating the two classes, but the positive and negative clusters still overlap, which makes proper classification very difficult.

Classification

10-fold cross validation as well as receiver operating characteristic (ROC) were used to test classifier performance and evaluate the obtained results. The results are analysed considering parts of the image of increasing sizes (windows, regions and full images) in order to show that the algorithm is able not only to detect images with artefacts, but also to point out the precise location of these artefacts inside the images. This is potentially important in clinical practice, as air artefacts happen mostly due to a lack of contact gel between the scanning

Table 3.11: Computing times per window position for the different feature types.

Feature type	*Mean in* ms	*Stdev in* ms
Sine fit	22	9
3^{rd} and 4^{th} order statistics	3.8	0.4
Percentiles	2.1	0.2
1^{st} and 2^{nd} order statistics	0.93	0.09
Entropy	0.4	0.3
Fourier Transform	0.12	0.04

probe and the patient's body. Consequently, the present method can raise a

Table 3.12: Features for air artefact detection ranked after their information gain ratio

Info gain ratio	*Feature*
0.0267	Kurtosis k_l
0.0220	Mean μ_l
0.0190	Skewness s_l
0.0187	m_s/m_l
0.0177	Median m_l
0.0163	μ_s/μ_l
0.0140	70th percentile of V_l
0.0120	Amplitude $Y(f_{low})$ (FT)
0.0113	90th percentile of V_s
0.0100	Mean μ_s
0.0095	Standard deviation σ_s
0.0095	EntropyH_l
0.0089	Mean m_s
0.0077	Entropy H_s
0.0064	H_s/H_l
0.0044	Slope b (Sine fit)
0.0039	70th percentile of V_s
0.0033	σ_s/σ_l
0.0027	Amplitude $Y(f_e)$ (FT)
0.0024	Standard deviation σ_l
0.0020	90th percentile of V_s
0.0020	Skewness s_s
0.0018	$s_s \cdot s_l$
0.0015	$k_s \cdot k_l$
0.0009	Amplitude A (Sine fit)
0.0008	Kurtosis of small window k_s
0.0002	Fit error Δ (Sine fit)
0.0002	Period p (Sine fit)

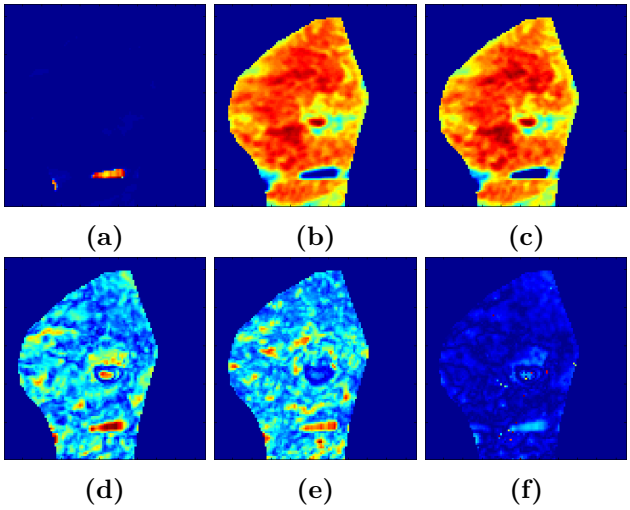


Figure 3.17: Sample 2D coronal parameter maps showing the computed features as heat maps. (a) Ratio of median values m_s/m_l, (b) mean value μ_l, (c) median value m_l, (d) peak $Y(f_{low})$ in Fourier spectrum below 2 Hz, (e) peak $Y(f_e)$ in Fourier spectrum within $f_e \pm 2\sigma_f$, and (f) slope of sine fit.

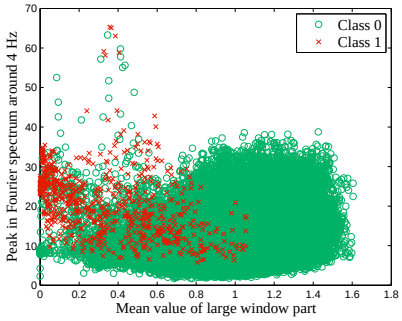


Figure 3.18: Feature space plot illustrating the discriminative power of the mean value of the large window, and the Fourier peak around 4 Hz. Positive instances (sliding window at a specific position) are represented as red crosses, negative instances as green circles.

warning flag if an artefact occurs as well as provide valuable information on how to solve the problem quickly.

The classifier performance on window-level in 10-fold cross-validation applied to the training data set is described in table 3.13 and by the ROC curve in figure 3.19a. The ROC curve proves an overall good performance of the prediction. The area under the merged ROC curve is $\mathrm{AUC}_{\mathrm{merge}} = 0.92 \pm 0.03$ (95 % CI). The numbers in table 3.13 show that a smart choice of decision thresholds for the classifier can yield very high sensitivity (0.95) and specificity (0.99) at the same time. Due to the unbalanced data set (positives/negatives = 2410/390568 = 0.006), it is however important to consider the F-Measure, which incorporates the ratio of true positive and false positive instances. The F-Measure reveals that more true positive instances are traded off by a much higher number of false positives. In conclusion, these numbers show that the proposed method can detect up to 95 % of positive instances (on window-level) and would benefit

Table 3.13: Performance measures of the proposed air artefact classification obtained in 10-fold cross-validation. Note that only one run of cross-validation was performed and that confidence intervals are based on the ten folds.

Spec.	95 % *CI*	*Sens.*	95 % *CI*	F_{avg}	95 % *CI*	$F_{Pr,Re}$	95 % *CI*
0.994	0.002	0.584	0.092	0.436	0.068	0.436	0.068
0.947	0.016	0.953	0.029	0.198	0.067	0.198	0.067

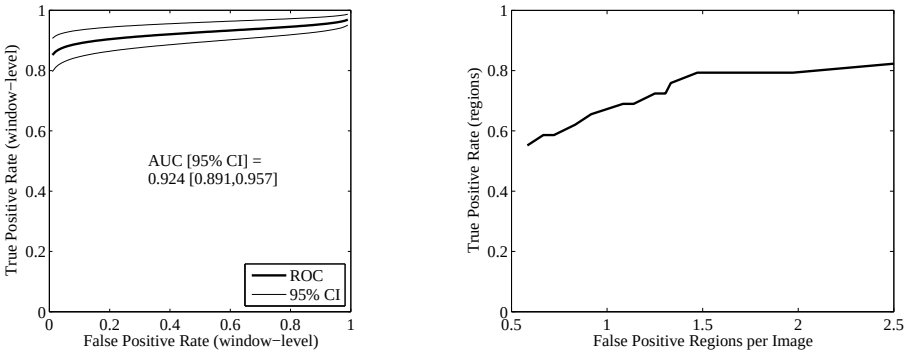
strongly from a more elaborated false positive reduction.

Applied to the test data set, the trained classifier yielded an AUC of 0.96 on window-level. For the region-level evaluation, the true and false positive regions (connected components) in the test data set were counted. The relation between the true positive rate TPR_R (on region level) and false positive regions per image is shown in the free response ROC (FROC) plot in figure 3.19b and in table 3.14 for two thresholds for the minimal accepted artefact size. The minimum size of the manually annotated artefacts was $21\,\mathrm{mm}^2$, consequently one could decide to only accept automatically detected potential artefacts if they exceeded this limit. However, due to the sliding window approach, it is possible that only parts of the actual artefact are denoted as positive instances by the classifier and, thus, the detected region can be smaller than $21\,\mathrm{mm}^2$.

Considering the images as a whole and discarding all candidate artefact regions smaller than $21\,\mathrm{mm}^2$, the classifier correctly identified 15 of the 19 artefact images as positive images whereas 5 of 17 normal images were misclassified. Figure 3.20 illustrates sample output images of the presented method.

Many artefact regions were detected very precisely by the proposed method. The false positive regions shown in figures 3.20c and 3.20f can be explained by the fact that they indeed show all derived image properties of an air artefact. However, they were not annotated as such because they are relatively small, directly adjacent to the breast contour, and thus, not clinically relevant. The missed artefact region (false negative) in figure 3.20f is an example for a very small artefact that was detected only partly by the proposed method and, thus, filtered by the lower threshold of $21\,\mathrm{mm}^2$ for positive regions. The differences between the manual segmentation in figure 3.20h and the classifier result in figure 3.20i are due to a slight lateral expansion of the acoustic shadow caused by this kind of contact artefacts. Whereas in the upper coronal slices, two artefact regions could be distinguished, they melted to one in the deeper slices (as displayed).

The time needed to train the classifier was 340 s, although it is important to notice that this step only needs to be carried out once when setting up the system and not in every potential use. The time needed to classify an image once the classifier had been trained was $(1.0 \pm 0.2)\,\mathrm{s}$ depending on the size of the imaged breast. Together with the time for feature computation, the total time needed to obtain a classification for a new image is $(188 \pm 30)\,\mathrm{s}$.



(a) **(b)**

Figure 3.19: Performance of the Random Forest (RF) classifier for air artefact detection. (a) The fitted merged ROC curve of 10-fold cross validation shows the performance of RF on the level of sliding windows when applied to the training data set. (b) FROC plot for the test data set showing the relation between true positive rate TPR_R (considering connected regions instead of pixels) and the number of false positive regions per image for varying minimum allowed sizes of potential air artefacts.

Table 3.14: True and false positive connected regions counted for two lower thresholds for the allowed size of detected artefacts.

Min. area in coronal (mm^2)	FP regions per image	Abs. counts of FPs	TPR_R	Abs. counts of TPs
0	2.6	93/36	0.83	24/29
21	0.58	21/36	0.55	16/29

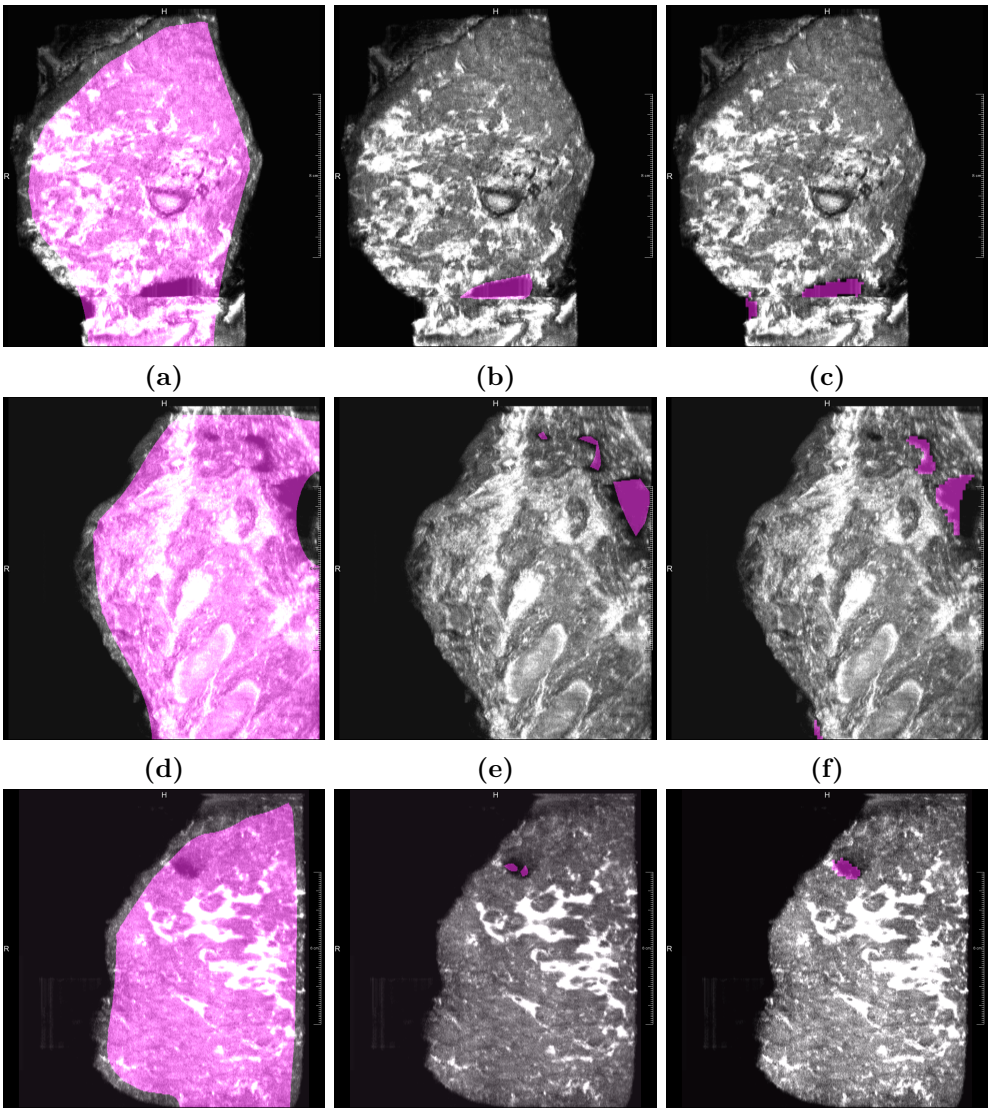


Figure 3.20: Three sample cases showing in the left column the computed breast mask, in the middle column the manually segmented artefact regions, and in the right column the automatically determined artefact regions as magenta overlays, respectively. In (i), the artefact was detected correctly. In (c) and (f), there are small false positive regions (along breast contour), whereas in (f), a very small artefact region was missed by the algorithm.

3.2.6 Automated Assessment of Nipple Visibility

The present study was separated into two parts: First, the improved automatic nipple detection method was evaluated on those test data set images that contained the nipple. Secondly, the new nipple detection algorithm was employed to extract features for the classification of the images with respect to the correctness of the manual nipple annotation. The classifier was trained on all images of the training data set and applied to the full test data set.

Automatic Nipple Detection

The reliability of the existing Laplacian- and Hessian-based nipple detection method was measured with the confidence factor C, which was computed according to equation 2.7 for those 327 test data set images that contained the nipple. Pearson's correlation between C and the distance d was -0.47 ($p < 0.05$, 95% CI $[-0.39, -0.55]$). This significant linear correlation supported the idea of using the confidence measure C as indicator for the reliability of the computed nipple position.

The atlases which represent the empirical probability distribution of the nipple position based on the training images are shown in figure 3.21. There is a clear tendency for the nipple position of each view, respectively. Whereas the nipple tends to be in the centre of the image for the AP view, it is pushed aside for MED and LAT views, as well as towards the bottom of the image for SUP view. The SUP atlas is less smooth than the others due to a lower number of available images for this view.

The mean distance d between ground truth nipple position $(x_\mathrm{G}, y_\mathrm{G})$ and the nipple position $(x_\mathrm{P}, y_\mathrm{P})$ as predicted by the proposed joint method was averaged over all 327 test images that contained the nipple. In figure 3.22, the distance d and the detection rate are plotted over the varying threshold C^*. One can see that the mean distance d decreases from (9 ± 17) mm to (7 ± 12) mm when C^* increases from 0 to 34. For values between 34 and 75, d stays approximately constant but raises slightly for threshold values above 75. The detection rate (counting those cases where d was smaller than 10 mm) reaches a maximum plateau of 0.85 for C^* values between 27 and 42. When the original nipple detection method by Wang et al. (2014) was used alone, the detection rate was 0.82. When the atlases was used for all cases, i.e. a decision threshold of 100

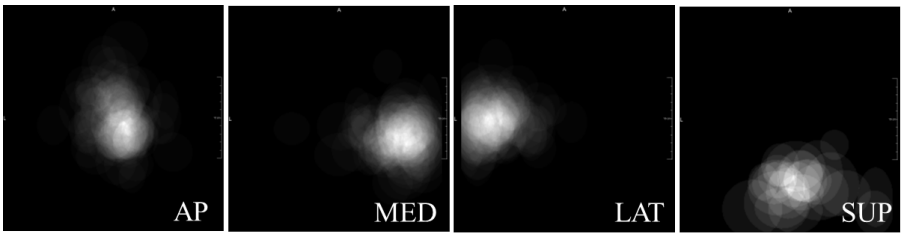


Figure 3.21: Atlas images for the four examined views as derived from the training data set images. The present figures represent the probability distribution of the nipple position for the left breast. They are mirrored along the vertical axis to achieve the atlases for the right breast.

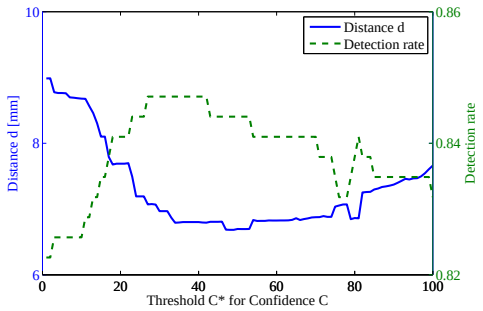


Figure 3.22: Mean distance d and detection rate over threshold values C^* for confidence measure C. With increasing C^*, the distance error d decreases and the detection rate raises until $C^* = 34$.

Table 3.15: Detection rate of the original detection method, two different modes of the joint detection method, as well as for the atlas alone, as measured in the images of the test data set that actually contained the nipple.

Algorithm	*AP*	*MED*	*LAT*	*SUP*	*all*
Original	0.90	0.76	0.85	0.71	0.82
Joint (opt.)	0.90	0.79	0.87	0.80	0.85
Joint (always)	0.87	0.78	0.83	0.83	0.83
Atlas alone	0.24	0.13	0.24	0.06	0.19

was chosen such that the original Laplacian and Hessian method was never used, the detection rate was 0.83 and the distance error was (8 ± 13) mm, which is however not significantly smaller than for the original method (p-value from paired t-test of > 0.1).

These results suggested that the best performance can be achieved by choosing a lower threshold C^* of 34 to trust the original nipple detection method alone, and adding the atlas-based method for all other cases. Detailed performance analysis of the joint nipple detection per view is shown in tables 3.15 and 3.16. The detection rate and the mean distance error d were evaluated for the original method, the proposed joint detection method with the determined optimal lower threshold $C^* = 34$, the proposed joint method with $C^* = 100$, and for the atlas alone. The numbers indicate that especially for the MED and LAT views, the joint nipple detection method that used the atlas where the original method was not reliable helped to improve the results significantly ($p = 0.04$). The original algorithm already performed very well for the AP view images. For the SUP view, the joint method that always combines the probabilistic atlas and the Laplacian-Hessian-method irrespective of C performs best. When the empirical atlas was used alone, the detection rates and the mean distance error were significantly worse than for any of the three other methods.

Image classification

The Random Forest classifier was trained on the training data set and applied to the test data set to discriminate the cases with correctly marked nipple from

Table 3.16: Mean distance error d ($\pm$ stdev) of the original detection method, two different modes of the joint detection method, as well as for the atlas alone, as measured in the images of the test data set that actually contained the nipple. p-values are computed in paired t-test and refer to the original method.

Algorithm	*AP*	*MED*	*LAT*	*SUP*	*all*
Original	5 ± 6	12 ± 20	7 ± 11	19 ± 29	9 ± 17
Joint (opt.)	4 ± 4	9 ± 16	5 ± 4	12 ± 22	7 ± 12
p-value	0.3	0.04	0.04	0.1	0.0007
Joint (always)	6 ± 7	10 ± 15	7 ± 14	9 ± 14	8 ± 13
p-value	0.1	0.1	0.9	0.04	0.1
Atlas	18 ± 10	23 ± 13	19 ± 16	25 ± 10	20 ± 13
p-value	7E-20	5E-5	8E-8	0.2	2E-21

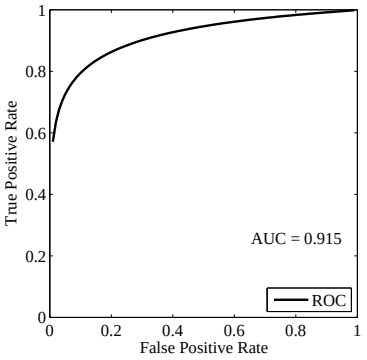


Figure 3.23: ROC plot for the classification of correctly pinpointed nipple positions versus incorrectly annotated cases, e.g. due to invisible nipple. The technicians' annotation was considered to be correct if the distance to the ground truth was less than 16 mm.

those, where the manual annotations by the technicians were wrong, e.g. because the nipple was not in the image at all. The ROC curve of the classification is shown in figure 3.23. The AUC is 0.92. At the point being closest to the upper left corner of the ROC plot, the F-Measure is 0.72 corresponding to a sensitivity of 0.90 and a specificity of 0.89. This means that 52 out of 58 positive cases were detected correctly.

3.3 Performance of Automated Image Quality Assessment on disjunct data

To evaluate the performance of the automated image quality assessment tools described in sections 2.2.1 to 2.2.4, the test data set C, which was not used previously, was annotated manually by two medical experts with respect to the three considered artefacts and processed by the four image quality assessment algorithms focusing on the relative nipple position, the nipple shadow, the breast contour shape, as well as the joint quality rating based on the three single aspects.

The results of the manual annotation are shown in figure 3.24. Counting all images where both readers detected the same artefact, the most frequent artefact was a prominent nipple shadow with 108 affected images (24 %), followed by an inadequate nipple position which was found in 83 images (19 %). An irregular breast contour shape was found by both readers concordantly in only 14 images (3 %). Compared to figure 3.3, the amount of low-quality images is decreased significantly. Whereas, e.g., a prominent nipple shadow was claimed in nearly 40 % of the cases in data sets A and B, in data set C, only 24 % were annotated as such.

The co-occurrence of artefacts in the considered data set is shown in figure 3.25. Overlapping regions indicate the number of images that were affected by two or three issues at the same time. Note that only those cases with actually visible nipple are considered, explaining the seeming discrepancy to the numbers in figure 3.24. Only 1 % of the images were affected by all three artefacts, whereas the highest co-occurrence was measured between nipple shadow and nipple position with 8 % of the images. Compared to figure 3.13, the trend of co-occurring image quality aspects is similar, but the total amount of affected images is reduced. This means that in both data sets the highest co-occurrence was measured for a prominent nipple shadow and an inadequate nipple position whereas the shape of the breast contour seemed a bit decoupled of these two and generally was detected in fewer cases.

The inter-rater agreement was analysed to measure the objectivity of the selected artefacts and to estimate the reliability of the manual read. For the expert annotation of the nipple shadow artefact and the relative nipple position,

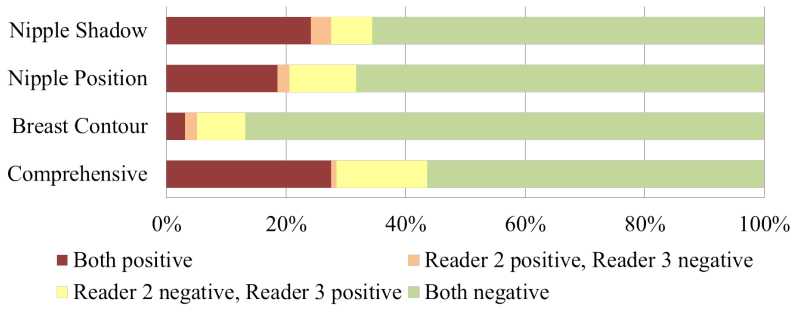


Figure 3.24: Frequency of specific artefacts in the considered data set C (447 images) as annotated by two raters. Red bars indicate the number of images that were rated concordantly as showing the artefact.

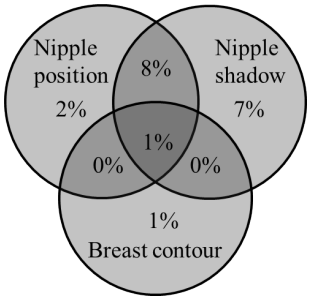


Figure 3.25: Relative co-occurrence of prominent nipple shadow, irregular breast contour shape and inadequate nipple position in data set C* (394), i.e. those 394 images where the nipple was actually visible.

Table 3.17: Computed inter-rater agreement for manual annotation of specific ABUS artefacts

Artefact	*Cohen's* κ
Relative nipple position	0.65
Nipple shadow	0.75
Breast contour shape	0.34
Joint	0.66

the agreement was very good with κ values of 0.75 and 0.65, respectively. The agreement between Reader 2 and Reader 3 was acceptable ($\kappa = 0.50$) for the joint quality rating, but poor for the breast contour shape ($\kappa = 0.34$). Thus, the more frequently an artefact appeared, the more clearly it seemed to be defined to the readers. Furthermore, the shape of the breast contour with respect to the size and shape of the breast as well as diverse shadows is by far more complex than the characteristics of the nipple, which is generally a clearly defined point in the image. Reader 2 marked fewer images as artefact-affected than Reader 3.

The performance of the automated artefact detection tools using the manually determined nipple position coordinates is represented by the ROC plots in figure 3.26 and the values listed in table 3.18. The ROC curves describe the specificity and sensitivity of the developed algorithms with regard to the manual annotations of both readers separately (green and blue curves) as well as to the combination of both ratings (magenta curves). One can see that the ROC curve of Reader 2 has always a (slightly) larger AUC value than the curve of Reader 3. This can be explained by the fact that the classifiers had been trained based on the manual annotations of Reader 2 (and Reader 1), but not of Reader 3. Nevertheless, the discrepancies between the green and blue curves are small, especially in the "clinically relevant" areas of small false positive rates (FPR $< 10\,\%$). It is also evident that the magenta curve, describing the classifier performance when compared to the common rating of both readers, is always very close to the curve of Reader 2. This is due to the fact that Reader 2 annotated more conservatively than Reader 3 and, thus, was always closer to the common rating than Reader 3 (see figure 3.24). The sensitivity and specificity of both readers when compared to each other, respectively, are plotted as op-

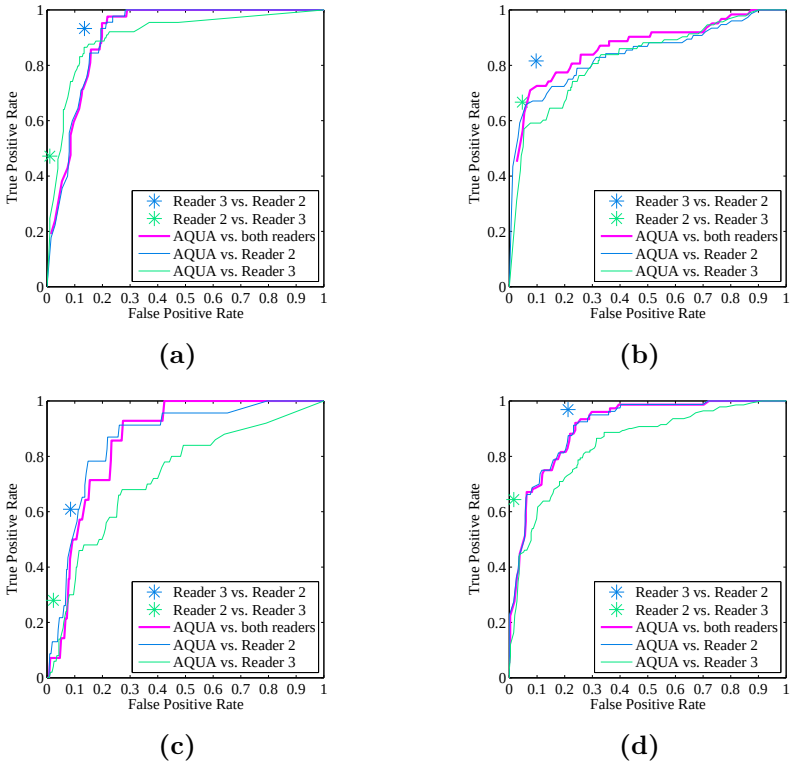


Figure 3.26: ROC plots describing classification results for artefact detection with respect to (a) the relative nipple position, (b) the nipple shadow, (c) the breast contour shape, and (d) the joint approach.

timistic reference points. They perform always slightly better than the AQUA algorithm.

If the joint nipple detection algorithm was employed to infer the nipple position coordinates, the classification of the relative nipple position and the acoustic nipple shadow is less reliable as shown by the ROC plots in figure 3.27. These curves having lower AUC values than the ROC plots in figure 3.26 show that the proposed classifier strongly depends on correct input parameters. The detection rate of the joint nipple detection method was 85 % (see section 3.2.6), whereas the manual annotation of the nipple position coordinates can be considered as ground truth, i.e. providing a detection rate of 100 %.

When the automated artefact detection methods were evaluated at the operating point (decision threshold) that was determined in the training step for a specificity of 0.97, they yielded the performance values listed in table 3.18. The specificity was high for all four considered artefacts, it varied between 0.83 and 0.91. The sensitivities ranged from 0.50 for the breast contour shape to 0.86 for the relative position of the nipple. Cohen's κ was computed for the agreement between the automatic artefact annotation and the combined rating of both readers.

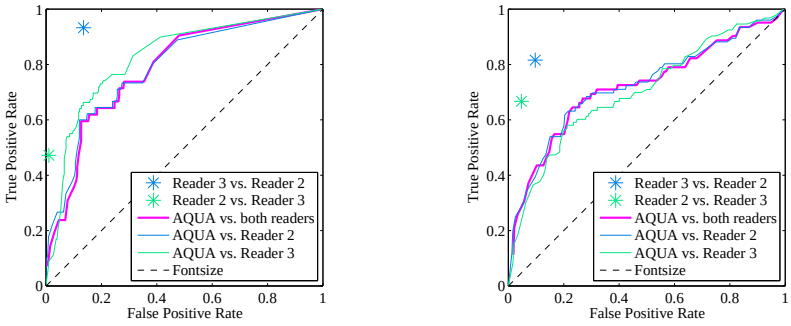


Figure 3.27: ROC plots describing classification results for artefact detection with respect to (a) the relative nipple position, and (b) the nipple shadow, with the nipple position coordinates being inferred from the joint nipple detection algorithm.

Table 3.18: Performance measures for retrospective analysis of automated image quality assessment

	Nipple position	*Nipple shadow*	*Breast contour*	*Joint*
Specificity	0.84	0.83	0.91	0.91
Sensitivity	0.86	0.77	0.50	0.68
Cohen's κ	0.45	0.47	0.19	0.58
AUC Reader 2 & 3	0.91	0.82	0.75	0.91
AUC Reader 2	0.91	0.84	0.86	0.91
AUC Reader 3	0.91	0.82	0.73	0.84

3.4 Clinical Implementation

A software prototype for automated image quality assessment (AQUA) in ABUS based on the already existing framework MIRIAM was built. Every incoming ABUS scan was automatically checked by all modules listed below and the outcome was displayed within two minutes on the MIRIAM dashboard. The following image properties were checked by the AQUA system:

- Protocol Check
  - Was the nipple position marked by the technicians?
  - Is the image size within acceptable limits?
- Overall Image Quality (Joint rating, sec. 3.2.4)
  - What is the general impression of the image?
  - Only if this joint artefact detection fails, i.e. the general image quality is rated low, the images are checked for the following quality aspects. This two-stage procedure was chosen because a high specificity was required.
- Relative Nipple Position (sec. 3.2.1)
  - Is the nipple properly visible on the image or was it pushed far aside, probably covering important other tissue structures?
- Nipple Shadow (sec. 3.2.2)
  - Is the acoustic shadow caused by the nipple too prominent and covering important regions of the breast?
- Breast Contour Shape (sec. 3.2.3)
  - Is the contour line of the breast following a roundish shape on the image? If the contour line is rather irregular and ragged, the breast might not have been supported properly during image acquisition and some parts of the breast might not have been imaged correctly.

3.4.1 Technical Aspects

The GUI designed within this work was supposed to show the image quality rating results of the current examination as simple as possible and at first glance. After a login dialogue that prevents unauthorized access to the image data, the Monitor Board page is displayed automatically (see figure 3.28). If no other examination is selected from the patient browser, the rating results of the current examination are displayed as a traffic light (see figure 3.29) and updated automatically. For each breast, there are four fields, each of which is assigned to one of the standard views that are acquired in one acquisition. Every field shows a sample coronal slice of the respective image as well as brief note on potential image quality issues. Since the images could not be sent directly from the scanner to the AQUA server, a bypass via an existing server within the clinic's DICOM network was implemented. Data transfer of one ABUS scan (approx. 120 MB) took one minute. Image processing and feedback to the network until display of the rating results took another minute, adding up to two

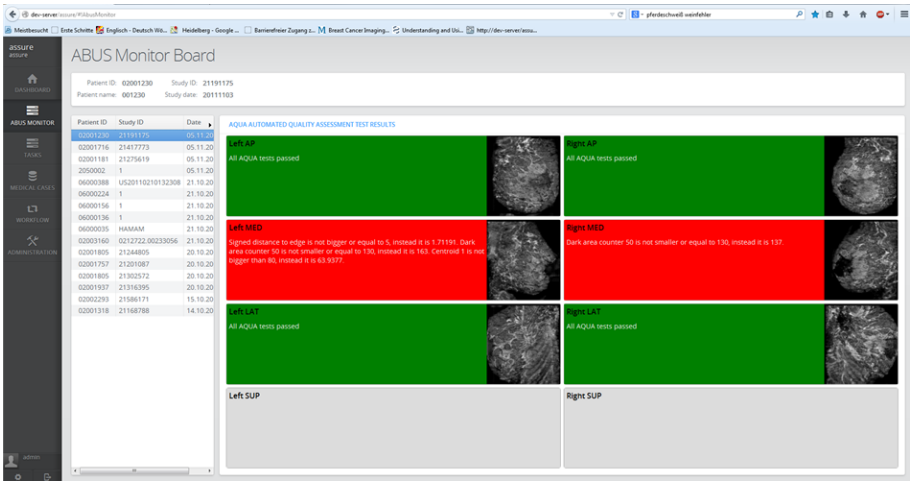


Figure 3.28: The GUI of the AQUA software prototype. It contains a patient browser on the left, shows the patient info of the currently selected patient on the top and displays the rating results of the selected examination in the centre.

Figure 3.29: The rating results for each scan are displayed as traffic light. If the image passed all AQUA tests, the field is green. A red field indicates that the image might be of low quality. Textual information on the detected issue is given.

minutes time delay. The acquisition of one view takes around one minute. Since the technicians were not able to wait for the AQUA rating results of the last views, they were instructed to consider only the rating for the first scans and, if applicable, to repeat these views.

3.4.2 Usability

The answers of seven technicians to the distributed questionnaires after a first contact to the ABUS AQUA software prototype are shown in figure 3.30. The general impression of the technicians was that the proposed software needed some improvement. From the written comments that were provided by the technicians it could be concluded that this was mainly due to an insufficient specificity of the software, i.e. the relative position of the nipple was found to be too close to the breast contour in many cases of large breasts. However, large

breasts often require such positioning to cover the whole breast volume. This was also the reason for the frustration that four of the technicians claimed. As a consequence, all of the technicians would like to have more user interaction with the software, e.g. the possibility to enter comments. They wanted to explain why they did not repeat a specific scan that was rated as low-quality image.

The graphical user interface was rated overall positively, as well as the loading and computing time of the software. As the AQUA tool is running in the background while the next scan is prepared, the impression of long or short computing times depends partly on the number of acquired views and the time that is needed for repositioning of the transducer. The amount of scans that was proposed to be repeated was estimated to 0–1 out of 6 by most of the technicians which corresponds to the amount that was expected from software design.

Due to a short effective test run period of only some weeks, the actual effect of the proposed image quality assessment software could not be evaluated statistically in this clinical set up. Nevertheless, interviews with radiologists and technicians were performed to retrieve qualitative feedback on the software. The radiologists reported a remarkable improvement of overall ABUS image quality over the test run period. The technicians did not repeat many scans, partly due to the time delay between image acquisition and quality rating that precluded the repetition especially of the last views within one exam. However, the clinical technicians stated that they were sensitized to the image quality aspects potentially influencing the radiologists in reading the images.

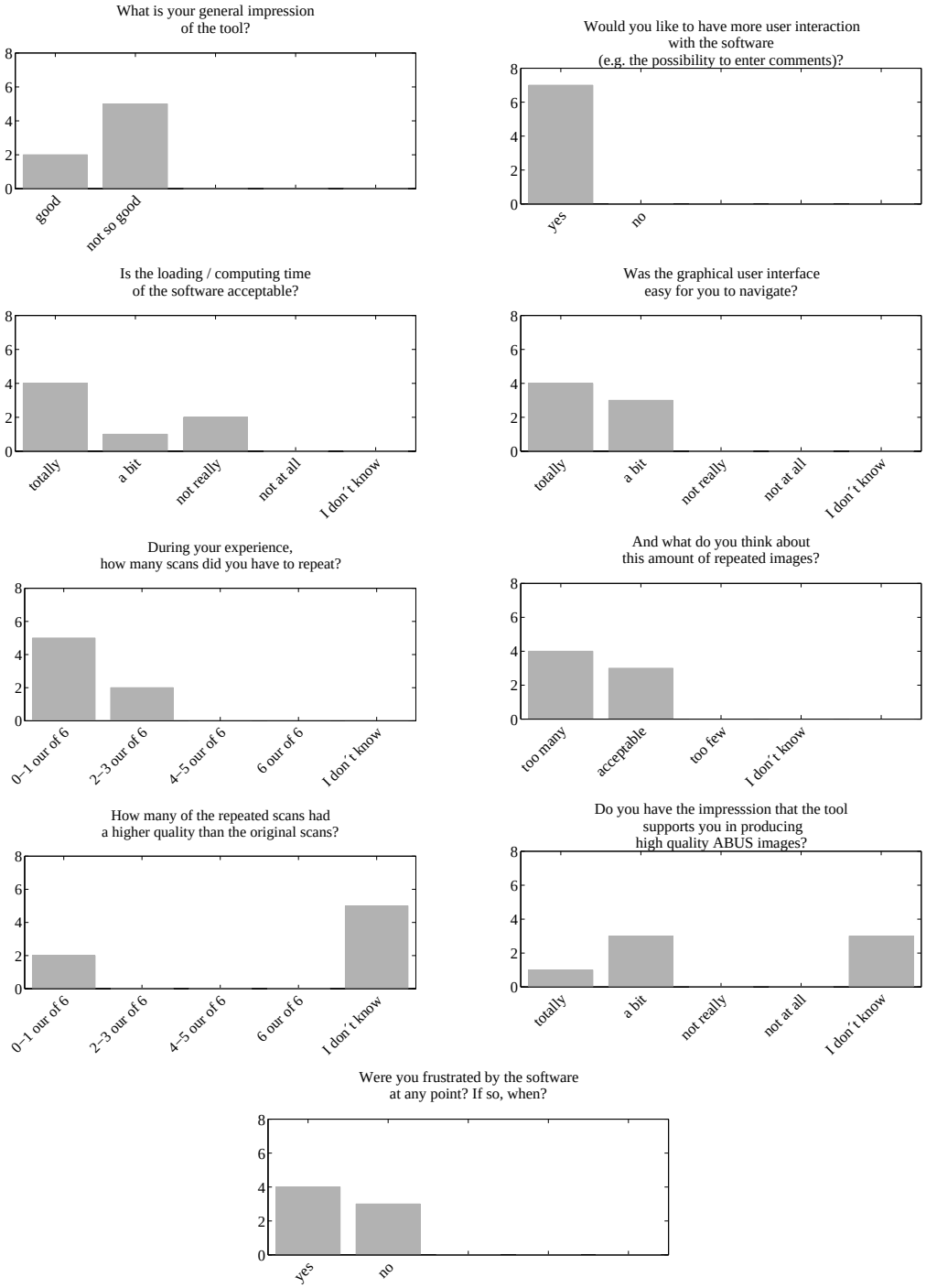


Figure 3.30: Answers of seven technicians to the questionnaire on the usability of the ABUS AQUA prototype after a first usage.

4 Discussion

4.1 Empirical Analysis of ABUS Artefacts

A reader study was initiated in order to evaluate the incidence of diverse imaging artefacts that can appear in ABUS images. The four most frequent imaging issues were a prominent nipple shadow, an inadequate relative nipple position, an irregular breast contour shape and the occurrence of air artefacts. The presented results suggested focusing the automated artefact detection on these four most frequent specific imaging issues.

The inter-rater agreement for the nipple position and shadow, the visibility of ribs as well as the breast contour shape was generally good with κ values ranging from 0.44 to 0.84, whereas the annotation of air artefacts and the wavy pattern only yielded mediocre agreement. The low agreement measured for the air artefact may be due to the fact that many air artefacts only affect a very small region. They might have been ignored by Reader 2 since they did not affect the diagnostic value of the considered image. Basically, the present κ values for artefact detection correspond to reported inter-rater agreement for lesion classification in ABUS images. Kim & Hong (2014) found a substantial agreement on lesion shape ($\kappa = 0.71$), and moderate agreement on other attributes as lesion type, mass orientation, echogenicity, and posterior acoustic features ($0.44 < \kappa < 0.59$).

For all considered quality aspects, there were more cases annotated as artefact affected by Reader 1 than by Reader 2, i.e. Reader 1 was more restrictive than Reader 2. Together with the varying κ values, this indicates that the annotation and classification of image artefacts and image quality is a very subjective task, which depends strongly on the experience and personal preferences of the radiologist rating the image. This may be true for all diagnostic images, but the complexity of 3D ABUS data sets—in comparison to, e.g., 2D mammographic images—enforces this effect. It is clear that lower inter-rater agreement in the ground truth results in poorer performance of a supervised machine learning approach, which is searching for clear rules.

It has to be noted, however, that some disagreement between the readers might also be caused by ill-posed questions, i.e. unclear instructions concerning the annotation process. Whereas one reader might have considered the images as self-contained instances, the other might have had in mind the correlation to the other views of the same study. Although the readers were told to follow the first interpretation, a radiologist might not be able to ignore the images he seen before but intrinsically includes this information in his rating. A display of the images in randomized order might have been beneficial to avoid this effect. The same accounts for the degree of abstraction that the readers were asked for: Whereas one reader might only look for the described image quality aspect, the other might put it into context and make assumptions on the clinical relevance

of the detected artefact.

Having said this, various scenarios for alternative ground truth acquisitions open up: One could rely only on one reader. The consistency within the rating might result in better classification results in the first place, but would lack generalisability. Letting the same reader(s) rate the image quality several times is another option that was evaluated in a very recent experiment (not presented in this work). It turned out that the intra-rater agreement was very similar to the inter-rater agreement, showing the complexity of the tackled problem.

4.2 Computer-aided Analysis of ABUS Artefacts

4.2.1 Nipple Position

The first image quality criterion that is described in this work is the position of the nipple in the image with respect to the rest of the breast tissue. As explained before, proper positioning of the transducer such that the nipple is well visible on the image is an essential factor for the diagnostic value of an ABUS image. The nipple is an important landmark not only for radiologic analysis of the data but also for computerized post-processing approaches. The most relevant feature to detect unsatisfying nipple positioning was the distance between nipple and breast contour, which is an intuitive measure for the problem.

With a true positive rate of 0.78 at a false positive rate of 0.03, the proposed algorithm reached high specificity and sensitivity at the same time. The good performance may be correlated to the fact that the manual rating of the nipple position was essentially driven by the same parameters as the automatic classification. This means that the clinicians marked the nipple being âĂIJtoo close to the edge of the breastâĂİ if the distance between nipple and breast contour line was very small. Exactly the same distance measure was used as feature for classifier training, i.e. the semantic gap between human perception and computed attributes was very small in this case.

Nevertheless, there were some outliers that were not classified correctly. They were generally caused by an erroneous breast mask due to irregular breast contour shapes. In some other cases, the algorithm was not able to reproduce the complex decision process that a human reader performs. Even if the described features were determined as expected, the readers might have anticipated and considered other aspects, e.g. parts of the breast that were not visible in the image, as shown in figure 3.6a.

Some of the selected features in this and the following sections were highly correlated, e.g. the signed distance measure $d^* = d \cdot c_{\text{io}}$. This suggests that might be a bias in the trained classifiers towards the correlated features. However, due to the random selection of features at each node, Random Forest classifiers are generally not biased by correlated features as long as they are used as "black box" for classification purposes (as opposed to feature selection or ranking purposes) (Strobl et al. 2008).

4.2.2 Nipple Shadow

In a next step, automatic classification of the acoustic shadow adjacent to the nipple was developed. The achieved AUC is not as good as for the classifica-

tion of the relative nipple position. This agrees with the higher complexity of the problem and correlated uncertainty of the readers as described in section 3.1. The inter-rater agreement was much lower for the annotation of the nipple shadow ($\kappa = 0.44$) than for the relative position ($\kappa = 0.84$), showing that the readers were less determined about the impact of a specific nipple shadow on the image quality.

Outliers were mainly due to an atypical appearance of the acoustic shadow. A sample case can be seen in figure 3.9c where the region behind the nipple has a relatively high intensity but nevertheless is fuzzy and does not contain any anatomical information.

A prominent nipple shadow can hamper diagnosis dramatically. The size of typical ABUS detected lesions was reported to be very small; mean diameters of 10 mm to 14 mm were found in the literature (Brem et al. 2015; Giuliano & Giuliano 2013). The shadowed region behind the nipple can take several cm^3 of the image and, thus, has the potential to occlude complete lesions. However, it has to be noted that radiologists analyse up to five views per breast and that typically the whole breast—including the region adjacent to the nipple—is visible when combining the information of all views. Whereas experienced radiologists perform this combinatorial analysis intuitively, computer aided detection systems might reach their limits. Registration between different views might be the first idea coming up to imitate this combinatorial task, but to the author's knowledge, no reliably operating full registration between different ABUS volumes has been presented to date. Boehler & Peitgen (2008) proposed to use registration between the single slices of one ABUS volume in order to reduce motion artefacts (discontinuities, wavy patterns). This might be considered as a first step towards full ABUS registration. However, to estimate the quality of a complete examination, a less complex approach combining the ratings obtained for each single view might already be sufficient. If it is assured that the tissue behind the nipple is represented sufficiently in one view of the study, it will not be necessary to repeat another view due to a prominent nipple shadow.

The proposed method of counting low intensity cylinder segments for measuring the size of the acoustic shadow corresponds to a very rough down-sampling of the considered region. More sophisticated approaches were also tested (not shown in this work) but turned out to be less robust against the variable intensity distributions in this region adjacent to the nipple. Specifically, a region growing algorithm had been implemented to evaluate the dimensions of the acoustic shadow connected to the nipple. However, defining the starting point ("seed") for the algorithm automatically was difficult due to the natural extent of the nipple and the coinciding inaccuracy of the nipple position used as input parameter.

4.2.3 Breast Contour Shape

Since the breast generally has a round and smooth shape, irregularities in the imaged contour are strong indicators for an inappropriate image acquisition. Although the discriminative power of each single feature was equally low, reasonable classification results could be achieved, similar to the results for the nipple shadow classification. As shown in the ROC plots of figure 3.11, the

curves are still set clearly apart from the diagonal line indicating proper classification, which also accounts to the two classifier models described above. A very tine confidence interval was retrieved from 10 repetitions of 10-fold cross validation, which proves reproducibility of the presented results. Whereas the classification of the nipple shadow was impeded by the low inter-rater agreement in the ground truth annotation, the characterisation of the breast contour shape was hampered by the low number of positive cases as unbalanced data sets make classifier training harder.

As for the above discussed image quality aspects, the classifier performance strongly depends on the suitability of the selected features. The attributes proposed in this work were retrieved from the physical properties of ultrasound propagation and the resulting appearance of a breast on an ABUS image, i.e. they were relatively intuitive. It is of course possible to compute more abstract features such as SIFT (scale-invariant feature transform) features (Lowe 1999) on different scales, which are generally used for object recognition. First attempts that were performed during this work (not presented) showed that these features were computationally too expensive for the planned real-time application. Furthermore, they could not be combined with the artefact-specific image quality rating that was chosen in this work. They might however be better suitable for unsupervised learning approaches to the image quality rating problem. This coincides with the idea of using much larger amounts of data for classifier training, e.g. for deep learning. Although the required amount of input data increases by several orders of magnitude, this is an active field of research and first works on lesion classification, image annotation or segmentation using deep learning algorithms have been presented.

4.2.4 Joint Image Quality Rating

Motivated by the fact that many images are affected by more than one artefact at a time, the three previously investigated imaging problems were combined into one general image quality rating. Based on the combination of all features dedicated to characterise the images as a whole with respect to specific artefacts, classifier training was performed with the aim to reflect the sum of the expert annotation. Very good results were achieved in the cross-validation experiments, i.e. the sensitivity was 0.70 at a specificity of 0.97. To achieve an even higher sensitivity, it might be beneficial to extend the list of proposed features by more detailed attributes as for example the breast cup size. However, computing the actual 3D volume of the breast based on an ABUS scan is not trivial and, to the authorsâĂŹ knowledge, has not yet been performed completely automatically by any other group. First steps like fully automatic chest wall segmentation have been presented by Tan et al. (2013b), who approximated the chest wall by a cylinder. However, computing time was reported to be 6.5 min per breast image, which would be too slow for the application we were aiming at. Thus, extracting information from 3D images by projecting them to 2D, e.g. the breast mask area, was more reliable, i.e. reproducible, and reduced complexity and computational costs.

If such a joint image quality rating was applied in clinical practice, the technicians would get no information on the exact reason of a failed image quality

check. The results of this work suggest that there is a trade-off between high sensitivity in terms of correctly detected low-quality images and the level of detail in the provided information. It would clearly depend on the preferences of the clinical personnel whether detailed information at the cost of lower sensitivities would be preferred over a more general quality rating.

Due to the characteristics of a trained classifier that works with class probabilities, a compromise between high specificity and high sensitivity must always be found for the final application. The proposed methods allow the user to decide on which level the software should operate. Whereas a high specificity is a very common setting in medical imaging (for example in algorithms that aim at detecting potentially malignant cancerous lesions), for image quality assessment, a false positive case might be less dramatic. At worst, an image rated wrongly as low-quality may mislead the technicians to repeat a view that was already of high quality. In this case, this increases examination time and represents a minor hassle for the patient but, as ABUS is a radiation free technique, no adverse consequences for the patient's health would follow.

Nevertheless, the chosen approach of summarizing those cases that have been rated differently by both readers in the negative class was aiming at high specificities. It was decided that only those cases which could be clearly attributed a specific image quality issue, i.e. which were rated as positive by both readers, were assigned the positive class in order to provide the classifier with a "cleaner" training sample. Other scenarios were also tested in this work, e.g. excluding the differently rated images completely from classifier training and testing yielded significantly better performances—as one would expect—but was also less realistic.

4.2.5 Air Artefacts

In this part of the work, a comprehensive study on air artefacts in ABUS images was conducted. In order to assess the relative importance of the problem, the images were examined and annotated by two expert radiologists. A considerable amount of the images were marked concordantly by both experts as affected by air artefacts. 58 of the 126 annotated artefact regions were bigger than $80\,\mathrm{mm}^2$, which corresponds to the reported average size of lesions detected in ABUS images. This shows how air artefacts—similar to the acoustic shadow possible caused by the nipple—have the potential to obscure or even totally occlude lesions. The decision to exclude artefact regions that were smaller than $20\,\mathrm{mm}^2$ was discussed thoroughly with radiologists who explained that artefacts below that limit were not clinically relevant.

Specific image features were extracted on the level of a sliding window of $3\,\mathrm{mm} \times 3\,\mathrm{mm}$ size in coronal plane and 20 mm depth in AP direction. This window size was leaned on the empirically determined sizes of air artefacts in the provided data set. A multi-scale approach with differently sized windows might add relevant information to the method, but was not yet considered due to already large computing times. Computational costs were also the reason for the chosen workflow of two different train/test scenarios: A Random Forest classifier was first tested in cross-validation experiments. and in the second stage, the trained classifier was applied to the test data set where. The analysis

of the test data set using different levels of detail showed that the proposed algorithm was able not only to raise a warning when air artefacts were present but also to point out which parts of the image were most likely to contain them.

Note that the 815 ABUS images evaluated in this study were randomly selected from the clinical archive and not preselected in any way. They included images acquired from four different views (AP, LAT, MED, SUP), which was however not considered as relevant information at the chosen scale of sliding window. Although the air artefacts of one transducer are generally very similar in terms of stripe distance (see figure 3.16), their appearance also depends on the surrounding tissue and the ultrasound scanner settings, e.g. time gain compensation. Missed artefact regions were either very small or did not exhibit the stripe pattern as clearly as expected, whereas false positive regions were mostly due to an incorrect breast mask including background regions, which often show the characteristic stripe pattern, too. Especially for very small artefact regions, the acoustic shadow can be very small, which makes them hard to detect, but at the same time less clinically relevant since most of the tissue behind them can be displayed properly.

The high amount of false positive regions that indeed present the typical stripe pattern but were not annotated manually again illustrates the problem of bridging the semantic gap between the computer, which can only detect the predefined pattern, and the clinician, who already interprets his findings concerning clinical relevance. The algorithm at its current stage is not yet able to rate the clinical relevance of its findings. A more extensive manual delineation of the artefact regions including also very small regions might have yielded better results at this stage.

Since air artefacts typically only take a small part of an image (if at all), the data that was used for machine learning was very unbalanced, i.e. there were many more negative instances (normal image) than positives (artefacts). Synthetically balancing the data set either by excluding negative instances or synthesizing additional positive instances might improve the performance of the classifier but was beyond the scope of this work.

Processing time for one ABUS image (3D volume) was 188 s, which is hardly acceptable for clinical application of direct feedback to the technician after image acquisition. Analysis of the computing times per window position for the different feature types revealed that the Sine fit features were computationally very expensive when compared to the other feature types (see table 3.11). Although they were the most intuitive measures to detect the stripe pattern, they were not ranked very high by the information gain ratio test (see table 3.14). Furthermore, the intrinsic correlation between Fourier Transform and Sine fit might introduce obsolete information to the classification process. These three facts together suggest that it might be reasonable to exclude the Sine fit features from future versions of an air artefact detection method.

The results of the feature ranking test indeed challenge the chosen approach of characterising the stripe pattern instead of only accounting for the much simpler statistics comparing the histograms of the small and the large windows. It was however found in preceding experiments that such a simplistic approach might easily get trapped, e.g. at the nipple.

To the best of the author's knowledge, there is no previous work the presented

results could be directly compared to since this was the first time that this specific artefact was considered for image quality improvement. Nevertheless, other classifier-based algorithms have been used to work with ABUS images. Specifically, the present algorithm detected 55 % of the artefacts at 0.58 false positive regions per image (FP/image). This is very much in line with similar machine learning applications for ABUS images (Tan et al. 2013a), where a lesion detection rate of 64 % at 1 FP/image was obtained. The highest detection rate of the proposed method was 83 % at 2.6 FP/image (Tan et al. (2013a) obtained a comparable 89 % detection rate at more than 10 FP/image). The clinical and practical consequences of a false positive air artefact are however very different from those of a false positive lesion. Whereas CAD algorithms focus on high sensitivities in order not to miss any suspicious object, an artefact detection tool should operate on very high specificity not bothering the technicians with unnecessary alerts. In conclusion, the proposed method can only be considered for clinical implementation if the specificity and the computing time are increased significantly.

The developed algorithms and the proposed classifier approach could easily be transferred to other ABUS devices. Only the specific properties of air artefacts produced by a different transducer would need to be examined such that the algorithms could be adapted, e.g. to a different mean stripe distance.

4.2.6 Automated Assessment of Nipple Visibility

Since the nipple is an important landmark in all medical breast images, automatic nipple detection is a pre-requisite for a wide range of image processing tasks, i.e. image registration or computer aided diagnosis. The aim of the present work was twofold: A previously described automatic nipple detection algorithm (Wang et al. 2014) was improved by incorporating prior location knowledge using a probabilistic atlas. Secondly, using features computed by this nipple detection method, a novel algorithm that assesses the quality of the manual nipple marks given by technicians was developed.

The generated atlas images clearly depicted the tendency of nipple positioning in the four examined views, e.g. for most AP view cases, the nipple is indeed located close to the centre of the image. Nevertheless, when the atlas was used alone for nipple detection, the mean detection rate was only 0.19 (with a tolerance of 10 mm) and the mean distance error was as large as (20 ± 13) mm, so this empirical information is only useful as adjunct to other methods, which account for the anatomy of the nipple in correlation with the physical properties of ultrasound imaging.

Since all source images originated from the same clinic, no detailed conclusion concerning the robustness of the atlases across different clinics could be made. Further investigations with a larger data set from other centres could be performed to clarify this question. The same accounts for the uniformity of patient positioning when performed by variable technicians. This issue could not be evaluated since the images were anonymised (including the acquiring technician).

By incorporating this prior location knowledge into the automated nipple localisation algorithm, nipple detection rate was increased from 0.82 to 0.85

and the mean distance error was decreased significantly from (9 ± 17) mm to (7 ± 12) mm (p-value from paired t-test was 0.0007). The nipple detection method presented by Kim et al. (2014b) uses the detection of elliptical structures in the coronal slices of an ABUS image. A nipple detection accuracy of (3 ± 4) mm and a detection rate of 94 % were reported, but the algorithm was only applied to 11 AP view images and 7 LAT view images. The joint nipple detection method proposed in this work yielded comparable mean distance errors of (4 ± 4) mm and (5 ± 4) mm as well as detection rates of 90 % and 87 % for 109 AP and 84 LAT view images, respectively.

In a second step, the improved nipple detection method was used to extract meaningful features for image classification according to the correctness of the manual nipple position annotation performed by the technicians. Class 1 (nipple not visible at all) and class 2 (technician's annotation deviated more than 16 mm from ground truth) images were assigned to the positive class since in both cases the manually tagged position should not be used for further image processing. The features used as input for the classifier were equally sensible in both cases. The proposed method was tested on an independent dataset of 380 ABUS volumes, resulting in sensitivity and specificity rates of 0.90 and 0.89, with an AUC of 0.92. This means that 52 of 58 incorrectly annotated nipple positions were detected automatically by the proposed method. The currently used software version of the ABUS scanner obliges the technicians to locate the nipple position manually even if it is not visible at all, which justifies the chosen approach.

4.3 Performance of Automated Image Quality Assessment on disjunct data

In order to evaluate some of the algorithms described in section 2.2, an additional analysis of their performance was conducted on data set C (see beginning of chapter 2), which was completely independent from the training data. The focus of this analysis was put on the relative nipple position, the extent of the nipple shadow, the shape of the breast contour, as well as on the joint quality measure based on these three.

The analysis of the manual annotation results revealed that the relative frequency of distinct artefacts was similar as in section 3.1 irrespective of the data set and the reader (see figures 3.3 and 3.24). It has however to be noticed that the total amount of low-quality images was significantly lower in data set C than in data sets A and B. In general there are two possible reasons for this disagreement: either data set C was indeed of higher image quality than A and B, or the readers of data set C were less restrictive. Since only two readers were available for each data set, statistics are not very strong and it is hard to infer the true reason. Assuming that Reader 2 performed similarly for all data sets, the conclusion would be that the data sets indeed differed in their overall image quality.

With respect to the amount of flagged images of Reader 2, who worked through both data sets, Reader 1 and Reader 3 were generally more restrictive designating more artefact-affected images than Reader 2. This was a consistent trend indicating that the image quality rating is still dependent on the personal

preferences and experience of the reader, albeit the problem was divided to the detection of very specific problems with the aim of increased objectivity and measurability.

Again, the inter-rater agreement was computed as Cohen's κ yielding very good agreement for the nipple position and the nipple shadow, but poor agreement for the annotation of the breast contour shape. This might be due to the lower incidence of irregular breast contour lines and the fact that it is a far more complex task to characterise the breast contour than to describe the nipple position and shadow.

Concordance between manual and automatic image quality rating was assessed for both readers separately as well as for the combination of their rating results. The ROC curves shown in figure 3.26 prove overall good performance of the proposed methods. The shape of the curves and the AUC values correspond well to the results that had been obtained in classifier training (see section 3.2). Apart from the fact that Random Forests are generally not prone to over-fitting, the results presented in this section prove that the proposed methods are also well applicable to independent data sets. However, using the classifier decision thresholds that yielded a specificity of 0.97 in the training step, resulted in lower specificities in this test run, ranging from 0.83 for the nipple shadow to 0.91 for the joint approach. This discrepancy between training and test results shows that the classifiers would benefit from a more diverse training data set, e.g. including images from different clinics and annotated by more readers.

Cohen's correlation was also computed for the agreement between automatic rating results and expert annotation. The κ values follow similar trends for each considered artefact, respectively, indicating once more that the automatic detection and classification of specific artefacts contends with the same uncertainties as human readers.

4.4 Clinical Implementation

For the first time, a software prototype for ABUS image quality assessment was integrated to clinical routine in order to evaluate the usability and reliability of the proposed approach.

The cooperation with the Radboud University Medical Centre offered the exceptional opportunity to get feedback from clinical technicians on the usability of the proposed system at this very early stage of development. Due to the limited evaluation time, it was not possible to adapt the algorithms and repeat the evaluation at a more advanced stage of the software. Nevertheless, the answered questionnaires revealed potential improvements of the envisaged AQUA system for ABUS imaging, e.g. the incorporation of the breast cup size into the classification process, since large breasts require different positioning than small breasts.

The feedback of the technicians showed that they were not very enthusiastic about using the AQUA prototype. It should however be taken into account that people are generally reluctant to changes in their daily routine, especially if the quality of their work is being questioned by a machine.

The successful technical implementation of the prototype motivates further steps towards a comprehensive ABUS quality assessment system. The algo-

rithms and the hardware have potential to be optimised, e.g. by combining the information of all views in one study. This was beyond the scope of this work, but offers interesting tasks for the future. Albeit the implemented software was in a very early prototype version, the image quality was reported consistently by radiologists and technicians to have improved over the test run period. Even though only very few scans were actually repeated, it turned out that the image quality rating of previous scans sensitized and motivated the technicians remarkably.

4.5 Impact

From discussions with radiologists and technicians carried out during this work, it was concluded that automated software that is able to detect ABUS artefacts could find immediate application in clinical practice. ABUS image artefacts can generally not be amended by the technicians once the acquisition has been performed. The only option to correct a defect is to repeat the affected scan. An algorithm running real-time in the background during acquisition procedures can alert technicians performing the acquisition if the image quality is low and potentially limiting the diagnostic evaluation. With the patient still present in the examination room, low-quality acquisitions can easily be replaced by repeating the scan.

To date, no automatic image quality assessment is performed at all before the radiologist reads the images. Thus, the proposed application has high potential to improve the current clinical practice. While a low false positive rate is usually considered essential for medical imaging tools, in this case it is supposed that consequences of false positives (mainly increased scanning time) are less severe than for other applications. On the other hand, false negatives might lead to diminished image quality and impede diagnosis. In any case, the classifier-based approach allows adjusting the sensitivity and specificity of the final application to the user's preferences.

The images used as data base for this thesis had all been acquired in two centres, which are highly involved in research. The personnel in these clinics is very experienced in breast imaging and pay high attention to image quality. It has to be noted that the technicians were already very experienced and well trained, and generally do not suffer from too restrict time constraints. Therefore, it could not be expected that the improvement in image quality upon the usage of the presented AQUA tools would be very high. Having said this, it is likely that the observed numbers of artefact-affected images are lower bounds when comparing to other sites where technicians might be less experienced or working under higher temporal pressure.

Reading time for an ABUS examination has been reported to be 9 min (Skaane et al. 2015), and thus is very high when compared to standard X-ray mammographies that can be read within 2 min (Dang et al. 2014) by an experienced radiologist. Since reading time is very expensive, there is a demand for computerised support, i.e. computer aided detection systems as proposed by Tan et al. (2015). High image quality is not only an essential prerequisite for profound diagnostics but also for any further image processing. Consequently, the image quality aspects discussed in this thesis were inspired not only by the clinical

needs of radiologists but also by the technical requirements for CAD.

According to the manual rating of the three artefacts discussed in sections 4.2.1 to 4.2.3, in 40 out of all 83 annotated examinations, there was no or only one corrupt image, while in 43 examinations, there were two or more corrupt images. This means that an early feedback to the technician after the first scan that showed artefacts might have helped to avoid another image with incorrect settings. However, throughout this work, the ABUS volumes were considered as independent images. Their correlation to the other images of one examination and the consequences for the usefulness of this examination was beyond the scope of this work and will be subject to further studies in the future. Registration between different views of one study could clarify whether the breast was imaged completely. An important prerequisite for registration is the segmentation of characteristic anatomical parts of the breast. Current publications focus on the automatic detection of the nipple (Wang et al. 2014; Kim et al. 2014a), the chest wall (Tan et al. 2013a, 2014) and the pectoralis muscle (Gubern-Mérida et al. 2015), as well as on the correspondence between lesions in different views (Tan et al. 2013a). First attempts to register breast images acquired from different views or even with different modalities have been taken (Georgii et al. 2013), but are still in the fledgling stages. Work is in progress and further steps are taken, e.g. by Gubern-Mérida et al. (2016) reporting to correlate lesions between different views with a distance error of (8 ± 10) mm (mean $\pm$ stdev).

The methods developed in this work are tailored to the needs of ABUS examinations but could easily be extended to other modalities. If nothing else, the general approach of using a manually annotated set of clinical images as basis for machine learning algorithms for image quality classification was tested and approved. The same approach has also already been applied successfully to dynamic contrast enhanced (DCE) breast MR images in order to detect motion between the different volumes of a time series (Wang et al. 2015).

Ultrasound image quality has been an interesting topic since the beginnings of sonography and is gaining importance with increasing complexity of ultrasound techniques and scanners. Various articles treating ultrasound image quality can be found in the literature. Gibson et al. (2001) proposed software that checks image quality automatically, focusing on image resolution, low- and high-contrast penetration depths as well as low- and high-contrast sensitivity. They used phantom images in order to examine the functionality of the ultrasound device as such, i.e. the beam former and the transducer. A similar system was presented by Thijssen et al. (2007) who described a software package for use in a performance testing protocol for medical ultrasound equipment. They used simple test objects (phantoms) and measured spatial resolution, contrast sensitivity, and clutter in fundamental and (tissue) harmonic modes. The "Guidelines for regular quality assurance testing of ultrasound scanners" of the British Medical Ultrasound Society (Dudley et al. 2014) show once more, that there is awareness of ultrasound image quality aspects, but rather from the technical point of view. Nearly all references that were found in the literature deal with the functionality of the equipment, but not with the usage of the system. The AQUA system proposed in this thesis aims at supporting the technicians in their daily routine during image acquisition by detecting application errors.

Manufacturers of ultrasound devices are developing algorithms that aim at

the real-time adjustment of ultrasound parameters to produce high-quality images as the patent of Lin & Seyed-Bolorforosh (2010) shows. El-Zehiry et al. (2013) proposed a method that runs directly on the scanner, analyses the image data stream in real-time, and improves the depth, the focus, and the frequency of the ultrasound device. They measured image quality by computing localized features based on Gabor filters of different scales, which were correlated by machine learning to the manual annotations of an expert clinician. This approach was thus similar to the methods applied in the present thesis, which, however, aimed at a different application. Whereas standard hand-held ultrasound examinations are performed by the clinician who can adjust parameters and adapt the image until a diagnosis is confirmed, ABUS images are acquired by technicians and cannot be altered once the scan has been performed, which entails a remote diagnostic read. Therefore, it is essential that the images are of high quality and all important parts of the breast are imaged correctly.

The integration of the developed algorithms to clinical workflow offered the exceptional opportunity to get direct feedback on the practicability of the proposed methods in a realistic use case. The conclusions and prospects in this work are therefore not only pure hypotheses but based on practical experience and evaluation. Thus, this work shows the complete development process of an image quality assessment software from requirement analysis via exploitation of the physical basics and causes of specific artefacts up to the technical implementation of a software prototype that is tested in clinical routine.

5 Conclusion

In breast cancer screening programmes, high image quality is an essential prerequisite for sound diagnosis and proper functionality of computer aided detection (CAD) systems. The increasing throughput of diagnostic images demands for software solutions that support the clinical personnel in their daily routine. In this work, a novel, fully automatic image quality assessment (AQUA) system for automated 3D breast ultrasound (ABUS) was designed from scratch and advanced such that a prototype could be tested in clinical routine.

Thorough analysis of the most common image quality aspects in ABUS showed that the most critical issues were related more to the acquisition process than to technical configuration of the scanner, which had already been addressed by other groups. With this work, for the first time, a quality assurance was tailored to ABUS acquisition with a focus on correct positioning of the breast, handling of the transducer and application of contact fluid. Approved image processing algorithms and machine learning methods were combined to solve this novel problem, i.e. the classification of ABUS images according to their quality based on specific aspects as for example the acoustic shadow caused by the nipple. The physical properties of ultrasound imaging were deployed to understand specific quality aspects and extract corresponding descriptive image features to bridge the semantic gap between human perception and computer algorithms. The installed prototype proved to sensitise technicians to the relevance of specific image quality aspects for reliable diagnosis such that, according to the radiologists, the overall image quality increased during the test run period. This work is by no means a finished project, but rather a first proof of concept for on-line feedback on image quality. Nevertheless, it resulted in a valuable tool that potentially supports technicians in their daily routine. Further studies need however to be performed to validate these preliminary results.

The basic principles of AQUA tool development described in this thesis open up a variety of further improvements. The accuracy of the single algorithms could be enhanced by incorporating more prior knowledge. Combining information of different views would yield a more realistic rating. Based on the flexible software framework it is now possible to refine existing algorithms and extend the portfolio of AQUA tools, e.g. by including more technical quality aspects. The chosen machine learning approach enables enhancements by expanding the ground truth to images from other clinics and annotated by more readers. The clinical evaluation showed that short run times were a crucial factor for user acceptance. The integration of AQUA tools to the ultrasound device together with general tuning for computational speed could result in real-time feedback clearly increasing the ease of use. The employed software framework was shown to be capable of managing on-line feedback on image quality and, thus, could be easily extended to other modalities or application scenarios.

Appendix A

The following two pages show the questionnaire that was distributed to the technicians for usability evaluation of the software prototype.

Questionnaire on Usability of AQUA tools for ABUS

The following questions are asked to evaluate the user friendliness and usability of the automated quality assessment (AQUA) tools that have been applied to ABUS images in clinical routine for a test run at the Radboud University Medical Center Nijmegen. **The Questionnaire should be filled in after the first use and after eight weeks of using the AQUA system.**

Please, cross the options that you agree with. Comments are optional, in case you want to explain your answer in more detail.

What is your general impression of the tool?

☐ good ☐ not so good

Comment:

Was the graphical user interface easy for you to navigate?

☐ totally ☐ a bit ☐ not really ☐ not at all ☐ I don't know

Comment:

Is the loading / computing time of the software acceptable?

☐ totally ☐ a bit ☐ not really ☐ not at all ☐ I don't know

Comment:

Would you like to have more user interaction with the software (e.g. the possibility to enter comments)?

☐ yes ☐ no

Comment:

During your experience, how many scans did you have to repeat?

☐ 0-1 out of 6 ☐ 2-3 out of 6 ☐ 4-5 out of 6 ☐ 6 out of 6 ☐ I don't know

And what do you think about this amount of repeated images?

☐ too many ☐ acceptabel ☐ too few ☐ I don't know

Comment:

How many of the **repeated** scans had a higher quality than the original scans?

☐ 0-1 out of 6 ☐ 2-3 out of 6 ☐ 4-5 out of 6 ☐ 6 out of 6 ☐ I don't know

Comment:

Do you have the impresssion that the tool supports you in producing high quality ABUS images?

☐ totally ☐ a bit ☐ not really ☐ not at all ☐ I don't know

Comment:

Were you frustrated by the software at any point? If so, when?

☐ yes ☐ no

Comment:

In general, how did you feel when using the tool? (Please, put one cross in every row.)

	totally	**a bit**	**not really**	**not at all**	**I don't know**
confident	☐	☐	☐	☐	☐
encouraged	☐	☐	☐	☐	☐
supported	☐	☐	☐	☐	☐
supervised	☐	☐	☐	☐	☐
disturbed	☐	☐	☐	☐	☐
frustrated	☐	☐	☐	☐	☐

Thank you for taking the time!

If you have any questions, feel free to contact me:

Julia Schwaab
mediri GmbH, Vangerowstr. 18, 69115 Heidelberg
j.schwaab@mediri.com
+49 6221 7256975

Bibliography

An, Y. Y., Kim, S. H., & Kang, B. J. (2015). The image quality and lesion characterization of breast using automated whole-breast ultrasound: A comparison with handheld ultrasound. *European Journal of Radiology*, 84(7), 1232–1235.

Arleo, E. K., Saleh, M., Ionescu, D., et al. (2014). Recall rate of screening ultrasound with automated breast volumetric scanning (ABVS) in women with dense breasts: a first quarter experience. *Clinical Imaging*, 38(4), 439–444.

Baker, J. A., Soo, M. S., & Rosen, E. L. (2001). Artifacts and pitfalls in sonographic imaging of the breast. *American Journal of Roentgenology*, 176(5), 1261–1266.

Bennett, R. L., Sellars, S. J., & Moss, S. M. (2011). Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland. *British Journal of Cancer*, 104(4), 571–577.

Bercoff, J. (2011). Ultrafast Ultrasound Imaging. In I. V. Minin & O. V. Minin (Eds.), *Ultrasound Imaging*.

Berg, W. A. (2008). Combined screening with ultrasound and mammography vs. mammography alone in women at elevated risk of breast cancer. *JAMA*, 299(18), 2151.

Berg, W. A. & Yang, W. T. (2014). *Diagnostic Imaging: Breast*. Salt Lake City and Utah: Amirsys, 2nd edition.

Boehler, T. & Peitgen, H.-O. (2008). Reducing motion artifacts in 3-D breast ultrasound using non-linear registration. In D. Metaxas, L. Axel, G. Fichtinger, & G. Székely (Eds.), *Medical Image Computing and Computer-Assisted Intervention MICCAI 2008*, volume 5241 of *Lecture Notes in Computer Science* (pp. 998–1005).: Springer-Verlag Berlin Heidelberg.

Bradski, G. (2000). The opencv library. *Dr. Dobb's Journal*, 25(11), 120–126.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Brem, R. F., Lenihan, M. J., Lieberman, J., & Torrente, J. (2015). Screening breast ultrasound: Past, present, and future. *American Journal of Roentgenology*, 204(2), 234–240.

Buchbender, S., Obenauer, S., Mohrmann, S., et al. (2013). Arterial spin labelling perfusion MRI of breast cancer using FAIR TrueFISP: Initial results. *Clinical Radiology*, 68(3), e123–e127.

Clevert, D.-A., Jung, E. M., Jungius, K.-P., Ertan, K., & Kubale, R. (2007). Value of tissue harmonic imaging (THI) and contrast harmonic imaging (CHI) in detection and characterisation of breast tumours. *European Radiology*, 17(1), 1–10.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Dang, P. A., Freer, P. E., Humphrey, K. L., Halpern, E. F., & Rafferty, E. A. (2014). Addition of tomosynthesis to conventional digital mammography: Effect on image interpretation time of screening examinations. *Radiology*, 270(1), 49–56.

Delorme, S. & Debus, J. (1998). *Ultraschalldiagnostik: Verstehen, lernen und anwenden*. Duale Reihe. Stuttgart: Hippokrates-Verl.

Descoteaux, M., Collins, D. L., & Siddiqi, K. (2008). A geometric flow for segmenting vasculature in proton-density weighted MRI. *Medical Image Analysis*, 12(4), 497–513.

Deserno, T. M. (2011). *Biomedical image processing*. Biological and medical physics, biomedical engineering. Berlin and Heidelberg and New York: Springer.

Drukteinis, J. S., Mooney, B. P., Flowers, C. I., & Gatenby, R. A. (2013). Beyond mammography: New frontiers in breast cancer screening. *The American journal of medicine*, 126(6), 472–479.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York: Wiley, 2nd edition.

Dudley, N., Russell, S., Ward, B., & Hoskins, P. (2014). The BMUS guidelines for regular quality assurance testing of ultrasound scanners. *Ultrasound*, 22(1), 6–7.

El-Zehiry, N., Yan, M., Good, S., et al. (2013). Learning the manifold of quality ultrasound acquisition. In D. Hutchison, T. Kanade, & J. Kittler (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2013*, volume 8149 of *Lecture Notes in Computer Science* (pp. 122–130). Berlin and Heidelberg: Springer Berlin Heidelberg.

Etzioni, R., Urban, N., Ramsey, S., et al. (2003). The case for early detection. *Nature reviews. Cancer*, 3(4), 243–252.

Ferlay, J., Soerjomataram, I., Ervik, M., et al. (2013). GLOBOCAN 2012 v1.0: Cancer incidence and mortality worldwide: IARC CancerBase No. 11.

Fleiss, J. L. (1973). *Statistical methods for rates and proportions*. A Wiley publication in applied statistics. New York: Wiley.

Forman, G. & Scholz, M. (2010). Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12(1), 49–57.

Georgii, J., Zöhrer, F., & Hahn, H. K. (2013). Model-based position correlation between breast images. In C. L. Novak & S. Aylward (Eds.), *SPIE Medical Imaging*, SPIE Proceedings (pp. 86701U).

Gibson, N. M., Dudley, N. J., & Griffith, K. (2001). A computerised quality control testing system for B-mode ultrasound. *Ultrasound in Medicine & Biology*, 27(12), 1697–1711.

Giuliano, V. & Giuliano, C. (2013). Improved breast cancer detection in asymptomatic women using 3D-automated breast ultrasound in mammographically dense breasts. *Clinical Imaging*, 37(3), 480–486.

Gonzalez, R. C. & Woods, R. E. (2008). *Digital image processing*. Upper Saddle River and N.J: Pearson Prentice Hall, 3rd edition.

Greve, W. & Wentura, D. (1997). *Wissenschaftliche Beobachtung: Eine Einführung*. Weinheim: Beltz, PsychologieVerlagsUnion, 2nd edition.

Gubern-Mérida, A., Tan, T., van Zelst, J. C., Mann, R. M., & Karssemeijer, N. (2016). Automated linking of suspicious findings between automated 3D breast ultrasound volumes: Paper 9785-22: accepted for publication. In *SPIE Medical Imaging*, SPIE Proceedings.

Gubern-Mérida, A., Tan, T., van Zelst, J. C., et al. (2015). Pectoral muscle durface segmentation in automated 3D breast ultrasound using cylindrical transform and atlas information. In M. Harz, T. Mertzanidou, & J. Hipwell (Eds.), *MICCAI-BIA 2015*: Fraunhofer Publica.

Günther, M. (2014). Perfusion imaging. *Journal of Magnetic Resonance Imaging*, 40(2), 269–279.

Hall, M., Frank, E., Holmes, G., et al. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1), 10.

Hoskins, P., Martin, K., & Thrush, A. (2010). *Diagnostic ultrasound: Physics and equipment*. Cambridge medicine. Cambridge and UK and New York: Cambridge University Press, 2nd edition.

Huertas, A. & Medioni, G. (1986). Detection of intensity changes with subpixel accuracy using Laplacian-Gaussian masks. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (5), 651–664.

Karnan, M. & Thangavel, K. (2007). Automatic detection of the breast border and nipple position on digital mammograms using genetic algorithm for asymmetry approach to detection of microcalcifications. *Computer methods and programs in biomedicine*, 87(1), 12–20.

Keeble, C., Wolstenhulme, S., Davies, A. G., & Evans, J. A. (2013). Is there agreement on what makes a good ultrasound image? *Ultrasound*, 21(3), 118–123.

Kelly, K. M., Dean, J., Comulada, W. S., & Lee, S.-J. (2010a). Breast cancer detection using automated whole breast ultrasound and mammography in radiographically dense breasts. *European Radiology*, 20(3), 734–742.

Kelly, K. M., Dean, J., Lee, S.-J., & Comulada, W. S. (2010b). Breast cancer detection: Radiologists' performance using mammography with and without automated whole-breast ultrasound. *European Radiology*, 20(11), 2557–2564.

Kim, E. J., Kim, S. H., Kang, B. J., & Kim, Y. J. (2014a). Interobserver agreement on the interpretation of automated whole breast ultrasonography. *Ultrasonography (Seoul, Korea)*, 33(4), 252.

Kim, H. & Hong, H. (2014). Detection of the nipple in automated 3D breast ultrasound using coronal slab-average-projection and cumulative probability map. In S. Aylward & L. M. Hadjiiski (Eds.), *SPIE Medical Imaging*, SPIE Proceedings (pp. 90351S).

Kim, J. H., Cha, J. H., Kim, N., et al. (2014b). Computer-aided detection system for masses in automated whole breast ultrasonography: Development and evaluation of the effectiveness. *Ultrasonography (Seoul, Korea)*, 33(2), 105–115.

Kolb, T. M., Lichy, J., & Newhouse, J. H. (2002). Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: An analysis of 27,825 patient evaluations. *Radiology*, 225(1), 165–175.

Kompan, I. N. (2015). *Adaption in dynamic contrast-enhanced MRI*. PhD thesis, University Bremen, Bremen.

Kuo, H.-C., Novak, C. L., Aylward, S., et al. (2013). Automatic 3D lesion segmentation on breast ultrasound images. In C. L. Novak & S. Aylward (Eds.), *SPIE Medical Imaging*, SPIE Proceedings (pp. 867025–867031).

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159.

Laub, G. & Kroeker, R. (2006). syngo TWIST for dynamic time-resolved MR angiography. *Magnetom Flash*, 3, 92–95.

Lehman, C. D., Blume, J. D., Weatherall, P., et al. (2005). Screening women at high risk for breast cancer with mammography and magnetic resonance imaging. *Cancer*, 103(9), 1898–1905.

Lin, F. & Seyed-Bolorforosh, M. (2010). System and method for automatic ultrasound image optimization. US-Patent (US 20100305441 A1).

Lindeberg, T. (1998). Feature detection with automatic scale selection. *International journal of computer vision*, 30(2), 79–116.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *The proceedings of the seventh IEEE international conference on Computer vision* (pp. 1150–1157).

Mandelson, M. T., Oestreicher, N., Porter, P. L., et al. (2000). Breast density as a predictor of mammographic detection: Comparison of interval-and screen-detected cancers. *Journal of the National Cancer Institute*, 92(13), 1081–1087.

Mann, R. M., Hoogeveen, Y. L., Blickman, J. G., & Boetes, C. (2007). MRI compared to conventional diagnostic work-up in the detection and evaluation of invasive lobular carcinoma of the breast: a review of existing literature. *Breast Cancer Research and Treatment*, 107(1), 1–14.

McCormack, V. A. & dos Santos Silva, I. (2006). Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer epidemiology, biomarkers and prevention*, 15(6), 1159–1169.

Medved, M., Fan, X., Abe, H., et al. (2011). Non-contrast enhanced MRI for evaluation of breast lesions. *Academic Radiology*, 18(12), 1467–1474.

Moon, W. K., Lo, C.-M., Chang, J. M., et al. (2012). Computer-aided classification of breast masses using speckle features of automated breast ultrasound images. *Medical physics*, 39(10), 6465–6473.

Mourad, M. & Stonestreet, J. (24.03.2015). Oldest evidence of breast cancer found in Egyptian skeleton: A REUTERS press release.

Notomi, Y., Lysyansky, P., Setser, R. M., et al. (2005). Measurement of Ventricular Torsion by Two-Dimensional Ultrasound Speckle Tracking Imaging. *Journal of the American College of Cardiology*, 45(12), 2034–2041.

Ohuchi, N., Suzuki, A., Sobue, T., et al. (2015). Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START): a randomised controlled trial. *The Lancet*.

Olson, J. S. (2002). *Bathsheba's breast: Women, cancer & history*. Baltimore: Johns Hopkins University Press.

Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, 11(285-296), 23–27.

Otten, J. D. M., Karssemeijer, N., Hendriks, J. H. C. L., et al. (2005). Effect of recall rate on earlier screen detection of breast cancers based on the Dutch performance indicators. *Journal of the National Cancer Institute*, 97(10), 748–754.

Platel, B., Mus, R., Welte, T., Karssemeijer, N., & Mann, R. M. (2014). Automated characterization of breast lesions imaged with an ultrafast DCE-MR protocol. *IEEE Transactions on Medical Imaging*, 33(2), 225–232.

Postema, M. & Attenborough, K. (2011). *Fundamentals of medical ultrasonics*. Milton Park and Abingdon and Oxon and New York: Spon Press.

Schopper, D. & Wolf, C. d. (2009). How effective are breast cancer screening programmes by mammography? Review of the current evidence. *European Journal of Cancer*, 45(11), 1916–1923.

Skaane, P., Gullien, R., Eben, E. B., et al. (2015). Interpretation of automated breast ultrasound (ABUS) with and without knowledge of mammography: A reader performance study. *Acta Radiologica*, 56(4), 404–412.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12), 1349–1380.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, 9(1), 307.

Tabár, L., Fagerberg, C. J., Gad, A., et al. (1985). Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *Lancet (London, England)*, 1(8433), 829–832.

Tan, T., Mordang, J.-J., van Zelst, J. C., et al. (2015). Computer-aided detection of breast cancers using Haar-like features in automated 3D breast ultrasound. *Medical physics*, 42(4), 1498–1504.

Tan, T., Platel, B., Hicks, M., Mann, R. M., & Karssemeijer, N. (2013a). Finding lesion correspondences in different views of automated 3D breast ultrasound. In C. L. Novak & S. Aylward (Eds.), *SPIE Medical Imaging*, SPIE Proceedings (pp. 86701N).

Tan, T., Platel, B., Huisman, H., et al. (2012). Computer-aided lesion diagnosis in automated 3-D breast ultrasound using coronal spiculation. *IEEE Transactions on Medical Imaging*, 31(5), 1034–1042.

Tan, T., Platel, B., Mann, R. M., Huisman, H., & Karssemeijer, N. (2013b). Chest wall segmentation in automated 3D breast ultrasound scans. *Medical Image Analysis*, 17(8), 1273–1281.

Tan, T., van Zelst, J. C., Zhang, W., et al. (2014). Chest-wall segmentation in automated 3D breast ultrasound images using thoracic volume classification. In S. Aylward & L. M. Hadjiiski (Eds.), *SPIE Medical Imaging*, SPIE Proceedings (pp. 90351Y).

Thijssen, J., van Wijk, M., & Cuypers, M. (2007). Performance testing of medical echo/Doppler equipment. In *Lemoigne, Caner et al. (Hg.) – Physics for medical imaging applications*, volume 240 (pp. 177–195).

Törnberg, S., Kemetli, L., Ascunce, N., et al. (2010). A pooled analysis of interval cancer rates in six European countries. *European Journal of Cancer Prevention*, 19(2), 87–93.

van Zelst, J. C., Platel, B., Karssemeijer, N., & Mann, R. M. (2015). Multiplanar reconstructions of 3D automated breast ultrasound improve lesion differentiation by radiologists. *Academic Radiology*, 22(12), 1489–1496.

Wang, L., Böhler, T., Zöhrer, F., et al. (2014). A hybrid method towards automated nipple detection in 3D breast ultrasound images. *Proceedings of the IEEE Engineering in Medicine and Biology Society*, 2014, 2869–2872.

Wang, L., Gubern-Mérida, A., Diaz, O., et al. (2015). Automated detection of motion in breast DCE-MRI to assess study quality and prevent unnecessary call-backs: ECR 2015 / C-1845: Electronic Poster.

Witten, I. H. & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann series in data management systems. San Diego and Los Angeles: Elsevier Science & Technology Books and Sony Electronics [Distributor], 2nd edition.

Yaghjyan, L., Colditz, G. A., Collins, L. C., et al. (2011). Mammographic breast density and subsequent risk of breast cancer in postmenopausal women according to tumor characteristics. *Journal of the National Cancer Institute*.

Acknowledgments

Writing this thesis has been a challenge that I would not have been able to master on my own. Retrospectively, the last three years seem to have flown by. Nevertheless, I met many great people being colleagues, friends and my family, who contributed to this work by advising and encouraging me. I am very grateful for their support.

The first one to thank is my supervisor Matthias, who offered me the opportunity to do this PhD and to contribute to the ASSURE project. He trusted in me, gave me scientific freedom, and supported me in all my decisions. His vast knowledge and bright ideas helped my to consider my work from a different perspective whenever I got stuck. Last but not least he afforded me travelling to many conferences and workshops introducing me to the breast imaging community. I appreciate all he has done for me in the past years a lot.

Many thanks go to my second referee Robert, who agreed so straightforward to review my thesis and travel from Girona to Bremen as if it was a matter of course. From the early beginning of my thesis I profited from his experience in image processing and machine learning. During several ASSURE meetings I got him to know as a very warm-hearted and imperturbable person, and I hope that our collaboration will persist beyond ASSURE.

One of the most important persons who contributed considerably to the success of this thesis is Yago. He was sincerely interested in my work, taught me that persistence paid, and always managed to motivate me when I felt lost.

Special thanks go to my colleagues at mediri: Johannes, who entrusted me with the project management of ASSURE resulting in a very educational challenge; Alba, who has grown out of being a master's student and contributed significantly to this work with her programming skills; JÃijrgen, who encouraged me with infinite patience and his throughout positive attitude; Sigurd, who developed the software framework that was used as basis for this work; Ina, who let me benefit from her experiences doing the PhD; DÃűrte, Manuela, Stefan, Joe, who all contribute to the nice working atmosphere at mediri.

There are several guys who taught me a lot in the past three years, finally became friends of mine, and deserve sincere thanks: Albert, who read many of my drafts with dedication and always tried to make me think scientifically; Jan, my personal radiologist, who rated these hundreds of images, and always has a funny line in store; Tao, who helped me to do the first steps with ABUS and was a great company during MICCAI in Nagoya.

Many other colleagues of the ASSURE project supported me enormously and should not be missing on this list: Lei, who helped me to understand how classifiers are working and invested quite a lot of time in our nipple detection algorithm; Fabian, who built the ABUS annotation tool and answered many of my early questions on breast images; Ahmed and André, who rated hundreds of images and gave me an understanding of the radiologists' point of view; Oliver, Arnau, and Joan, who gave me a warm welcome when I was in Girona; Ritse and Bram, who might have suffered from the dozens of mails and questions I sent them; Ashley, who helped me to design the questionnaires; the technicians at Radboud University Medical Centre, who tested the AQUA system and filled in these questionnaires.

Last but not least I want to thank my family and friends, who supported me in all I have done, especially Peter, my safe haven.

Thank you!