Master Thesis Technical Medicine

Jeroen Willemse

Artificial intelligence in breast-specific gamma imaging

Exploring the possibilities of automation of breast cancer detection and classification

Leids Universitair Medisch Centrum

TU Delft

Erasmus MC
Universitair Medisch Centrum Rotterdam

General

Being the most frequently diagnosed type of cancer in women in 2020, breast cancer has remained a hot topic for researchers all over the world. All this research has led to great improvements in overall survival, disease-free survival and overall prognosis for patients suffering from breast cancer.

A key factor in breast cancer management is early and accurate diagnosis to allow patient-specific targeted therapy based on tumor characteristics. A relatively new diagnostic tool in the diagnosis of breast cancer is breast-specific gamma imaging(BSGI), also referred to as molecular breast imaging(MBI). BSGI is a nuclear medicine imaging technique that is based on the uptake of the radiotracer ^{99m}Technetium-sestamibi.

One indication for BSGI is problem solving: the evaluation of indeterminate breast abnormalities or symptoms. In this thesis, we aim to explore the possibilities of automation of breast cancer detection and classification using BSGI images.

In the first part of this thesis, a conducted literature review is described, which focuses on (semi-)quantitative features of BSGI images, and their possible relationships with clinical factors, such as malignancy, molecular subtypes and invasiveness.

The second part of this thesis describes the development and evaluation of different artificial intelligence (AI) algorithms in the detection and classification of breast lesions. For the detection of (malignant) lesions, three convolutional neural networks were designed, while the classification of lesions was performed using publicly available machine learning algorithms.

The final part is the clinical validation of the convolutional neural networks described above. This evaluation includes the possible time-savings of using AI compared to traditional interpretation of BSGI, the agreement between physicians and the AI-algorithm, as well as a possible correlation between the AI output and the BI-RADS scores assigned to images by nuclear medicine physicians.

Thesis contents

PART 1

Literature review

Can features derived from Breast-specific Gamma Imaging/ Molecular Breast Imaging predict malignancy, invasiveness, biological markers, histopathological grading and molecular subtyping?

Author

J.R.J. Willemse

Supervisors

W. Grootjans
Leiden University Medical Center

L.M. Pereira Arias-Bouda
Leiden University Medical Center

F.H.P. van Velden
Leiden University Medical Center

Abstract

A key factor in breast cancer management and increasing survival is early and accurate diagnosis. An emerging diagnostic tool is Breast-specific Gamma Imaging(BSGI), also referred to as Molecular Breast Imaging(MBI). Rather than an anatomical imaging modality, it provides functional information represented by uptake of 99mtc-sestamibi. This review provides an overview of recent studies on (semi-)quantitative analysis of BSGI/MBI images that sought to investigate the clinical parameters associated with image parameters. A total of 10 studies were reviewed. Most significant results were found in the correlation between the tumor-to-background ratios (TBR) found on images, when compared to malignancy rates. Higher TBR were associated with a higher change of malignancy, as well as a higher Ki-67 protein status and higher rates of estrogen receptor(ER) negativity and Progesterone receptor(PR) negativity. In addition, a higher TBR was found to be positively correlated with a higher histological grade of tumors. Other significant positive correlations include the coefficient of variation(COV), which was associated with a higher degree of invasiveness. One study found that irregular shape and a linear/ductal uptake pattern were a predictive factors for malignancy. The results found in the reviewed studies show there is clinical information embedded within MBI/BSGI images, which correlates with (semi-)quantitatively calculated image parameters. Possibly, in the future, Artificial Intelligence(AI), which has not yet been implemented in MBI/BSGI, can be used to aid in detecting these image parameters and provide automatic clinical information.

Introduction

In 2018, a total of 2,088,849 new cases of breast cancer worldwide were diagnosed, resulting in 626,679 deaths, making it the second most frequent type of cancer in both sexes following lung cancer. In females, breast cancer represents an estimated 24% to 30% of newly diagnosed cancers, and accounts for 15% of all cancer related deaths in women[1,2]. Important uncontrollable and controllable risk factors for developing breast cancer include age, family history, menarche history, obesity, smoking, alcohol consumption, breast density, contraceptive pill usage and, in less than 10% of cases, inherited genetic mutations[3,4,5]. Breast cancer incidence is strongly associated with a high Human Development Index(HDI) tier. The highest breast cancer incidence is seen in Belgium, followed by the Netherlands, whereas the lowest incidence is found in South Central Asia and Africa. In these latter areas, however, incidence is increasing , attributable to changes in lifestyle. In North-America and Europe, mortality has seen a decline over the past years, which can be partially explained by accurate diagnosis and improved treatment[6,1].

Breast cancer diagnosis

A key factor in breast cancer management and increasing survival is early and accurate diagnosis. Mammography screening programs showed to reduce mortality by breast cancer with around 40% in women in the age of 40 to 70 years old and an overall reduction in mortality of 19%[7,8,9]. Unfortunately, mammographic screening is associated with high rates of false-positive findings, leading to unnecessary radiation dose, overtreatment and anxiety. False-positive findings are more common in patients of younger age (< 40 years old). Also, dense breasts pose a challenge for mammography screening[7.] Reported sensitivity from mammography alone ranges from 63% to 98%[10,11,12], and even lower sensitivity was reported in patients with dense breasts[13].

After positive findings on mammographic screening, or after symptomatic presentation, additional diagnostic modalities, such as targeted mammography, which is a more extensive mammographic procedure following a mammographic screening, ultrasound(US), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET) and biopsies can be used to rule out false-positives and provide further classification of the tumor[5,14]. When looking at the added value of US and MRI in the diagnostic work-up of breast cancer, Berg et al. found that the use of just US in detecting malignant lesions had a higher sensitivity than using mammography on its own (83% vs. 68%, respectively) and that MRI alone showed an even higher sensitivity of 94%. US alone and MR alone, however, had significantly lower specificities in the detection of malignancies of 34% and 26%, respectively, versus 75% in mammography alone[15].

The role of PET/CT(Computed Tomography) in the primary diagnosis of breast cancer is fairly limited, but can be useful in the detection of distant metastasis and lymph node infiltration[16]. Mainly due to the low spatial resolution of PET, a sensitivity as low as 48% has been reported, with false-positive findings resulting from inflammations, fibroadenomas and gynecomastia. Reported specificity ranges from 73% up to 100%[17]. Recently, novel imaging modalities for the diagnosis of breast cancer have been introduced. Breast-specific gamma imaging(BSGI), also referred to Molecular breast imaging (MBI), depending on the type and number of detectors used, is a nuclear medicine imaging technique, derived from scintimammography, utilizing small-field-of view dedicated gamma cameras for detection of intravenously administered radiotracers[18,19]. BSGI is a functional imaging modality, most often based on the uptake of 99mTc-sestamibi, a radiotracer that is able to show cellular

proliferative activity by mitochondrial uptake and angiogenesis, generally observed in tumors[20,21].

BSGI acquisition protocols slightly vary across institutions and manufacturers, but generally consist of 99mTc-sestamibi administration in the antecubital vein, followed by image acquisition mediolateral-oblique (MLO) and cranialcaudal(CC) perspectives. Initially, administered doses varied from 770-1100 MBq, but more recently published studies have reported equivalent diagnostic accuracy with doses up to 2,5 times lower than original protocols. This was achieved mainly due to improved collimator design[22,23,24].
Images are acquired bilaterally in craniocaudal(CC) and Mediolateral Oblique(MLO) views, with slight breast compression between either a detector and compression paddle, or two detectors, resulting in a complete BSGI set of 4 images[25]. Acquisition times for each scan vary between 8-10 minutes[26,27].

A 2017 meta-analysis by Zhang et al. compared diagnostic performance of BSGI and MRI in breast cancer. Whereas sensitivity of BSGI versus MRI showed comparable results (0.84; 95% CI, 0.79-0.88), and 0.89; 95% CI, 0.84-0.92, respectively), specificity was significantly higher in BSGI (0.82; 95% CI, 0.74-0.88 and 0.39; 95% CI, 0.30-0.49, respectively). This analysis showed areas under the summary receiver operating characteristics (SROC) curves of 0.93 and 0.72 for BSGI and MRI, respectively[28]. Other advantages of using BSGI as an additional diagnostic modality when compared to MRI include acquisition time , interpretation time and costs. Also, BSGI does not require a dedicated room, as BSGI devices are comparable in size to that of an ultrasound device. It does, however, require necessary radiation shielding, which is not the case in MRI . Finally, BSGI is less influenced by breast density, such as in mammography[29]. Disadvantages of BSGI include the whole-body dose distribution of 5 mSv after administration of 600 MBq of 99mTc sestamibi, which is approximately 10 times the dose received from mammography alone[26,30].

In 2012, Conners et al. proposed a lexicon specifically for interpretation of BSGI images, aimed to improve observer agreement and provide standardized assessment of BSGI images[31,32]. Similarly to the Breast Imaging and Reporting Data System (BI-RADS), this lexicon is a guide for visual interpretation by a nuclear medicine phycisiant[33]. Scoring is based on various aspects of BSGI images, such as background uptake, mass- and non-mass uptake, distribution, uptake pattern, symmetry, lesion location and qualitative assessment of lesion-to-background ratio, which is categorized as photopenic (less than subcutaneous fat), minimal-mild uptake(equal or slightly greater uptake), moderate uptake(visually greater uptake than mild, but less than twice as intense as subcutaneous fat) and marked uptake(visually at least twice as intense as subcutaneous fat)[31,32]. Several studies have attempted to describe correlations between (semi-quantitative) MBI/BSGI image features and molecular subtypes of breast cancer, histopathological features, clinical outcomes such as patient management, invasiveness of cancer and prognosis. This review article will assess these studies and provide a comprehensive overview. Articles were found using Pubmed and Google Scholar search.

Breast cancer Classification

As breast cancer encompasses a heterogenous group of tumors, definitive classification is often based on histological features, such as tumor size, staging, histologic grade, differentiation, lymph node status, and hormonal status[34,35].

Histologically, breast cancer can be divided into subgroups such as the stage 0, non-invasive Ductal Carcinoma in Situ (DCIS), or already invasive cancers such as Invasive Ductal Carcinoma (IDC) or Invasive Lobular Cancer (ILC)[36]. IDC is the most common type of invasive breast cancers, accounting

for 50 - 80% of breast cancers, followed by ILC accounting for 5 - 15% of newly reported cases[2,37]. Molecular markers also play an important role in classification of breast cancers, as they overcome certain limitations emerging with imaging techniques. Several molecular, or biological markers have been identified to play a role in breast cancer pathogenesis, breast cancer prognosis, and choice of treatment[38]. These markers include Estrogen Receptor(ER), Progesterone Receptor(PR), Human Epidermal Growth Factor Receptor 2 (HER2) and parameters such as antigen Ki-67[39,40]. ER, PR, HER2 and Ki-67 status can be evaluated through immunohistochemical(IHC) staining on either surgically resected specimens or core needle biopsies(CNB). ER and PR receptor status can be either positive or negative. A specimen is generally considered ER/PR positive if >1% of nuclei is stained positively in 10 high-power fields[41]. Approximately 70% of invasive breast cancers show an overexpression(i.e. positive) ER status[42]. HER2 overexpression exists in an estimated 18 -20% of invasive breast cancers, and is either determined through IHC or fluorescent in situ hybridization(FISH)[43,44].

The protein Ki-67, coded by the MKI-67 gene, is overexpressed in highly proliferative tumors, while down-regulated in G0 cell cycles. It is associated with rapid tumor-growth and thus aggressive cancers. High Ki-67 expression (>14%) is an indicator for poor prognosis, upstaging to invasive cancer and patient outcome[45,46,47]. Based on the expression of these markers, a tumor can be categorized in 4 subtypes; 1) ER positive and/or PR positive, HER2 negative and Ki-67 14%(Luminal A); 2) ER positive and/or PR positive, HER2 negative and Ki-67 > 14%(Luminal B); 3) ER positive and/or PR positive, HER2 positive, regardless of Ki-67 expression and 4) Triple negative breast cancers, also referred to as basal-like subtype[48,49]. When leaving Ki-67 expression out of the equation, breast cancers can be categorized in 3 subtypes, based on ER/PR positiv ity /negativity and HER2 positivity/negativity. Molecular subtyping of breast cancers plays an important role in patient management, prognosis and treatment decision making. ER/PR and HER2 expression are somewhat negatively correlated (i.e. ER/PR positive tumors mostly show a negative HER2 expression), making personalized therapy decision making more viable. ER/PR positive tumors are susceptible to systemic endocrine therapy(Tamoxifen), whereas HER2 positive tumors can be more efficiently treated with HER2 targeting therapy such as Trastuzumab(Herceptin). Triple-negative subtypes are less susceptible to endocrine therapies and/or Herceptin, and are best treated with systemic chemotherapy[50,51].

Tumor-to-background ratio

One of the most widely investigated features of BSGI is the semi-quantitative analysis of the tumor-to background ratio(TBR): The standardized uptake values of 99mTc sestamibi in lesions in comparison to healthy background uptake. In a study by Park et al. it was hypothesized that semi-quantitative calculation of the TBR would be helpful in discriminating benign from malignant breast lesions. In their study, investigating 118 lesions, it was concluded that, with a TBR cutoff value of 1.5, they were able to discriminate benign from malignant lesions with a specificity of 92%, whereas visual analysis alone had a specificity of 82% ($p = 0.008$). Mammography and ultrasound had specificity values of 82% ($p = 0.008$) and 62% ($p < 0.001$) respectively. The number of false-positive findings in these 118 lesions was reduced by six after adding semi-quantitative calculation of TBR to the visual analysis, when compared to visual analysis alone. Park et al. did not report average TBR values for malignant and benign lesions separately[52].

Similarly, in a paper published by Tan et al. in 2014, the added value of semi-quantitative analysis of the TBR for discriminating between benign and malignant was evaluated in 92 lesions. In this study, early (10-15 min) and delayed (90-120 min) images were separately evaluated. Malignant lesions in early images showed a significantly higher TBR than benign lesions (3.18 ± 1.57 vs. 1.53 ± 0.59, $p < 0.05$). In delayed images, similar results were found (2.91 ± 1.91 vs. 1.46 ± 0.54, $p < 0.05$). With visual analysis only, recorded sensitivity was 77.8% and specificity was 81.6%. Using only TBR as a

classification method, analysis of the early images (optimal cutoff: 2.06) resulted in a sensitivity of 81.5% and a specificity of 92.1%. In delayed image TBR analysis (optimal cutoff: 1.77) a sensitivity and specificity of 97.5% and 98.5%, respectively, were reported. When combining visual analysis and semi-quantitative analysis, the authors reported a sensitivity of 85.2% and a specificity of 92.2% in early images, and a sensitivity and specificity of 83.3% and 98.5% respectively in delayed BSGI images[53].

Also aiming to assess whether lesions could be classified as either malignant or benign based on TBR, Choi et al. found that malignant lesions were associated with a higher TBR value. In 372 breast lesions, they found an TBR of 2.2 ± 1.0 and 1.6 ± 0.5 ($p < 0.001$) for malignant and benign lesions, respectively. Overall, with a cut-off TBR of 2.1, a sensitivity of 50% was found, a sensitivity of 85% and an accuracy of 75%. For lesions larger than 1cm, this increased to 70%, 88% and 84% respectively. ROC curve analysis for all lesions gave an area under the curve of 0.728 ($p < 0.001$, 95% CI, 0.625-0.831) and 0.853 ($p=0.047$, 95% CI, 0.761-0.945) for lesions with a diameter larger than 1cm[54].

Meissnitzer et al. conducted a similar study, where the relative uptake factor, analogous to the TBR, was semi-quantitively calculated and compared to biopsy confirmed malignant and benign lesions. In accordance to the previously described studies investigating correlations between TBR and malignancy rates, Meissnitzer et al. found a significantly elevated relative uptake factor in malignant lesions. Malignant lesions showed a ratio of 4.27 ± 0.64, whereas benign lesions only showed a relative uptake factor of 2.37 ± 0.18[55] .

Studies that investigated the difference in TBR between malignant and benign lesions are listed below in table 1.

Table 1 TBR values for malignant and benign lesions

	TBR		
Study	Malignant	Benign	P-value
Park et al. [51]	Na	Na	Na
Tan et al. [52]	3.18 ± 1.57	1.53 ± 0.59	<0.05
Choi et al.[53]	2.2 ± 1.0	1.6 ± 0.5	<0.001
Meissnitzer et al.[55]	4.27 ± 0.64	2.37 ± 0.18	<0.05

Other than investigating the correlation between TBR and malignancy, its correlation with invasiveness has also been studied. In 2015, Yoon et al. published a retrospective study investigating confirmed ductal carcinomas. They found that IDC was associated with a high TBR of 2.5 ± 0.8. Pathologically confirmed pure ductal carcinomas in situ (DCIS-P) and DCIS with micro-invasion (DCIS-Mi) had significantly lower TBR values of 1.8 ± 0.4 and 2.1 ± 0.4, respectively ($p=0.001$)[56]. In accordance with these results, Yoo et al. evaluated the significance of TBR values for determination of upstaging to invasive cancer from DCIS. They found that upstaged invasive cancers had a significantly higher TBR of 2.70 (2.16–3.45) compared to 2.28 (1.86–2.98) in DCIS ($p=0.002$)[57].

In a study published by Tan et al., these results were further confirmed. In their study, published in 2016, a TBR of 2.25 ± 0.14 was found in non-invasive breast cancers, whereas a TBR of 3.15 ± 0.14 was found in invasive cancers ($p=0.005$)[58].

Finally, the correlation of TBR with molecular subtypes and clinicopathological markers has been investigated. Yoon et al. aimed to evaluate semi-quantitative BSGI analysis as a prognostic tool for predicting recurrence based on its relationship with histopathological markers in breast cancer. A

significant correlation was found in ER status, where ER-negative tumors showed a TBR of 3.9 ± 1.5 and ER-positive tumors showed a significantly lower (p=0.03) TBR of 3.4 ± 1.4. HER2-positivity was associated with a significantly (p=0.001) higher TBR (4.2 ± 1.8) when compared to HER2-negative cancers(3.4 ± 1.3). Ki-67 positive cancers (with a threshold of 10%) were significantly (p=0.001) correlated with a higher TBR and showed a value of 3.8 ± 1.6. Ki-67 negative cancers showed a TBR of 3.1 ± 1.1. A significant relationship was also found in Nuclear and histological grades. For nuclear grade, a higher TBR was found in higher nuclear grades(G1: 2.6 ± 1.1, G2 3.5 ± 1.5, G3: 3.8 +1.4, p=0.04). Similar values were recorded for histological grade, with a TBR of 2.9 ± 1.3 for G1, 3.7 ± 1.4 for G2 and 3.8 ± 1.5 for G3(p=0.01). Yoon et al. reported no significant difference in TBR between PR-positive and PR-negative tumors and between triple-negative and non-triple negative subtypes of breast cancer[59].

Lee et al. found a significant correlation of TBR with PR-status, where PR-positive lesions had a reported TBR of 3.3 ± 1.3 versus a TBR of 4.4 ± 2.2 in those with a PR-negative status[60] . Ki-67 status, with a threshold of 14%, was also correlated with TBR values. Negative and low (<14%) expression of Ki-67 was found in lesions with a TBR of 3.2 ± 1.4, whereas positive ki-67 expression was associated with a higher TBR of 4.4 ± 1.9 (p=0.007). Another significant difference was found in luminal A cancers versus non-luminal A cancers (p=0.007), where non-luminal A cancers had a significantly higher TBR of 4.2 ± 1.9 versus a TBR of 3.0 +1.2 in luminal A cancers. Another significant correlation was found in histological grade, with a TBR of 2.9 ± 1.2 for G1, 3.8 ± 1.8 for G2 and 3.9 ± 1.3 for G3 (p= 0.029). No significant correlations of HER2 status, ER status, nuclear grade and TBR were found by Lee et al.[60]. The same authors published a second article in 2018 on a different patient cohort, in which they confirmed their results, with significant correlations between TBR and ER status, PR status, Ki-67 status(14% threshold)histological grade and subtypes[61]. A significant correlation of ER-status and PR-status was also reported by Yoo et al. in 2017. They found that ER- positive lesions had a significantly lower TBR than ER-negative lesions (3.49 ± 1.41 versus 4.41 ± 2.25, p = 0.0049), which was also seen in PR-positive status versus PR-negative status (3.45 ± 1.41 versus 4.56 ± 2.17, p = 0.0006). Similar to the conclusions of lee et al.[60,61], Yoo et al. found significant correlations between histological grade and TBR (G1: 2.99 ± 1.0, G2: 3.60 ± 1.49, G3: 4.20 ± 2.00 (p = 0.004). No significant differences in TBR values were reported for HER2-status, Ki-67 expression (14% threshold) nuclear grade, or triplenegative versus non-triple- negative subtypes[62].

Table 2 TBR values for ER and PR status

	ER			**PR**		
	+	-	P-value	+	-	P-value
YOON ET AL.[59]	3.4 ± 1.4	3.9 ± 1.5	0.03	3.8 ± 1.5	3.4 ± 1.4	0.06
LEE ET AL.[60]	3.4 ± 1.5	4.3 ± 2.1	0.068	3.3 ± 1.3	4.4 ± 2.2	0.036
LEE ET AL.[61]	3.2 ± 1.9	4.0 ± 1.9	0.029	3.0 ± 1.7	4.2 ± 1.9	0.004
YOO ET AL.[62]	3.49 ± 1.41	4.41 ± 2.25	0.0049	3.45 ± 1.41	4.56 ± 2.17	0.0006
TAN ET AL.[58]	3.02 ± 0.17	2.93 ± 0.14	0.740	2.90 ± 0.16	3.12 ± 0.19	0.370

Table 3 TBR values for HER2 and Ki-67 status

	HER2			KI-67		
	+	-	P-value	+	-	P-value
YOON ET AL.[59]	4.2 ± 1.8	3.4 ± 1.3	0.001	3.8 ± 1.6	3.1 ± 1.1	0.001
LEE ET AL.[60]	4.2 ± 2.3	3.6 ± 1.5	0.229	4.4 ± 1.9	3.2 ± 1.4	0.007
LEE ET AL.[61]	4.1 ± 2.1	3.4 ± 1.8	0.145	4.3 ± 2.2	2.9 ± 1.4	0.001
YOO ET AL.[62]	4.13 ± 2.05	3.55 ± 1.51	0.0679	Na	Na	Na
TAN ET AL.[58]	2.76 ± 0.19	2.94 ± 0.14	>0.05	3.00 ± 0.12	2.95 ± 0.28	0.85

Tan et al., contradicting previsously mentioned studies, did not find any significant difference in TBR among ER-status, PR-status, HER2-status, Ki-67 expression(8% threshold), nuclear - and histological grade or molecular subtypes[58]. An overview of biological markers and associations with TBR values are summarized in table 2 and table 3. Nuclear- and histological grade correlations are shown in table 4.

Table 4 TBR values for different nuclear- and histological grades

	NUCLEAR GRADE				HISTOLOGICAL GRADE			
	G1	G2	G3	P-value	G1	G2	G3	P-value
YOON ET AL.[59]	2.6 ± 1.1	3.5 ± 1.5	3.8 +1.4	0.04	2.9 ± 1.3	3.7 ± 1.4	3.8 ± 1.5	0.01
LEE ET AL.[60]	2.3 ± 1.0	3.5 ± 1.5	3.9 ± 1.3	0.104	2.9 ± 1.2	3.8 ± 1.8	3.9 ± 1.3	0.029
LEE ET AL.[61]	Na	Na	Na	Na	3.0 ± 0.9	3.7 ± 2.1	Na	0.035
YOO ET AL.[62]	2.69 ± 0.71	3.55 ± 1.46	4.00 ± 1.93	0.076	2.99 ± 1.00	3.60 ± 1.49	4.20 ± 2.00	0.004
TAN ET AL.[58]	Na	Na	Na	Na	Na	3.07 ± 0.15	3.12 ± 0.17	0.810

Coefficient of variation

The coefficient of variation (COV) can be used to quantify the degree of tumor heterogeneity. Yoo et al., who aimed to determine wheter certain BSGI image features were associated with upstaging in situ carcinomas to invasive carcinomas, calculated the COV in 168 patients. Lesions that were upstaged to invasive carcinomas (n =58) had a significantly ($p < 0.0001$) higher COV when compared to in situ carcinomas (n =117). Reported median COV for invasive (upstaged) carcinomas was 35.9%, versus 23.5% in non-invasive cancers[62] .

Similarly, Yoon et al., calculated the COV for tumors that have been either confirmed to be DCIS P, DCIS-MI, and IDC. They found that tumors with an invasive aspect (DCIS-MI and IDC) had a significantly higher ($p = 0.016$) COV of 14.3 ± 5.5% and 19.5 ± 7.4%, respectively, when compared to pure in situ cancers, which had a COV of 12.6 ± 4.2%[56] .

Tumor shape and distribution

Other important factors in interpretation of BSGI lesions such as shape and distribution can also be used for discrimination between benign and malignant lesions. Choi. et al, albeit visually and not with semi-quantification, aimed to investigate which BSGI features in women with recently diagnosed breast cancer are associated with malignancy. They found that when comparing malignancy rates in oval shaped lesions to malignancies in irregular shapes, the latter was a significantly higher predictive factor for malignancy ($p= 0.004$). Linear/ductal distribution was also highly predictive for malignancy when compared to focal distribution (100% vs. 20.8% respectively, $p=0.001$), regional distribution (100% vs. 9.5% respectively, $p < 0.001$) and segmental distribution (100% vs. 2.9% respectively, $p < 0.001$). No significant different malignancy rates were found in focal vs. regional vs. segmental distribution patterns in non-mass uptake[54] .

Discussion

Most studies that are described in this review indicate that BSGI and MBI image features can be, to some degree, used to discriminate between clinical factors. A total of 10 studies investigated the possible correlations between semi-quantitative tumor-to-background ratios found in BSGI images with clinical aspects. All four studies that sought to determine the possible correlation between uptake in lesions and malignancy rates, found that malignant tumors show a significantly higher uptake than benign lesions[52.53.54.55]. The absolute TBR values that were calculated differed across studies (table 1), because of different calculation methods and acquisition protocols used in each study.

Additionally, three studies concluded that the TBR was significantly correlated with invasiveness of breast lesions[56,57,58]. Of five studies aiming to find the correlation between TBR and biological/histopathological markers in the tumor (ER-status, PR-status, HER2-status, Ki67-expression, Histological grade, Nuclear Grade, subtypes), TBR correlated significantly with ER-status in 3 studies, PR-status in 3 studies and HER2-status in 1 study. Three out of four studies investigating the correlation between TBR and ki-67-expression found that a higher TBR is associated with a higher expression of the protein. Nuclear and Histological tumor grades correlated significantly with TBR in two out of three and four out of five studies, respectively.

Interestingly, Tan et al.[57] , who found no correlations between TBR and clinical parameters, except for invasiveness, used a different approach than other studies when calculating the TBR. Whereas other studies calculated the background uptake by using the average uptake of three specified regions of interest (ROI) in normal breast tissue compared to the maximum uptake within the lesion, Tan et. al used placed the same lesion ROI on a nonlesion area equally distanced from the nipple in the same breast[54,56,57,59,60,61,62]. This alternative approach to TBR calculation might explain the slightly different TBR values calculated by Tan et al. which in turn might have led to non-significant results. Meissnitzer[55] also used a different calculation method, mainly distinguishing itself by smaller regions of interest(20mm2), but, in contrast to Tan et al.[57], found significant differences in malignant and benign lesions.

Semi-quantitative calculations of tumor heterogeneity (COV) in both studies showed that the degree of invasiveness can be extrapolated from BSGI images[56,57].

In General, the studies mentioned in this review indicate that BSGI/MBI image features significantly correlate with clinical parameters. Although most studies cited in this review only investigated the impact of one feature such as TBR, COV or shape, a combination of those could provide even more information about tumor characteristics. A combination of quantitively calculated features would be

analogous to the lexicon provided by Conners et al., but with more standardized interpretation, rather than visual analysis dependent on the interpreter[31,32].

It has to be noted, however, that most studies included in this overview are conducted in Asian medical centers on Asian women. As disparities in breast cancer incidence and biological aspects such as breast density exist among different races/ethnicities, it is not clear whether the results found in these studies are directly applicable on women of other ethnicities. A limitation is that different acquisition methods have been used across the studies described in this article. Different protocols, with different acquisition times and administered doses may lead to different uptake values within breast tissue, and therefore lead to non-comparable calculations across studies. Nonetheless, acquisition protocols within each study itself did not vary for its own patient cohort. The difference in protocols across studies may explain the, for instance, different values found for the TBR or COV. This can be seen, for instance, in table 1, where Meissnitzer et al. found that malignant lesions had an average TBR of 2.2, whereas this value in other studies would be classified as benign. This indicates the need for standardized acquisition protocols, not only within a medical center or study design, but across medical centers. One way to possibly overcome the issue of interobserver variability in BSGI as a result of visual analysis is the use of Artificial Intelligence(AI). AI methods, such as machine learning or deep learning, have been used in other imaging modalities for the detection and classification of breast lesions. These AI algorithms detect or interpret pre-determined image features without the interference of humans and autonomously learn which features are to what extend associated with benign and malignant lesions. In mammography, AI algorithms such as deep-learning and machine learning have been shown to be able to detect malignancies with an accuracy similar to that of radiologists[63,64]. Similarly, machine learn ing algorithms have been shown to be able to distinguish benign from malignant lesions with a promising accuracy in MRI[65] and in US[66] . To the best of our knowledge, the only study on the implementation of deep-learning in BSGI/MBI images is a study by Carter et al., concluding that deep-learning is a feasible method of classifying parenchymal uptake, a known risk factor for breast cancer[67,68].

Conclusion

In conclusion, (semi-)quantitative and visual interpretation of BSGI images could provide valuable information of the lesions clinical nature. Depending on calculation method of different image features, it seems possible to discriminate malignant from benign lesions, tumor invasiveness, and molecular subtyping of lesions found in BSGI images. Standardized automatic interpretation of BSGI images using artificial intelligence has not yet been investigated, but could be a valuable tool in breast cancer diagnosis.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: a cancer journal for clinicians, vol. 68, no. 6, pp. 394–424, 2018.
[2] F. A. Tavassoli, "Pathology and genetics of tumours of the breast and female genital organs," World Hhealth Organization Classification of Tumours, 2003.
[3] K. Rojas and A. Stuckey, "Breast cancer epi demiology and risk factors," Clinical obstetrics and gynecology, vol. 59, no. 4, pp. 651–672, 2016.
[4] W. Majeed, B. Aslam, I. Javed, T. Khaliq, F. Muhammad, A. Ali, and A. Raza, "Breast cancer: major risk factors and recent developments in treatment," Asian Pac J Cancer Prev, vol. 15, no. 8, pp. 3353–3358, 2014.
[5] M. Milosevic, D. Jankovic, A. Milenkovic, and D. Stojanov, "Early diagnosis and detection of breast cancer.," Technology and health care : of ficial journal of the European Society for Engi neering and Medicine, vol. 26 4, pp. 729–759, 2018.
[6] M. Ghoncheh, Z. Pournamdar, and H. Sale hiniya, "Incidence and mortality and epidemiology of breast cancer in the world," Asian Pacific Journal of Cancer Prevention, vol. 17, no. S3, pp. 43–46, 2016.

[7] L. E. Pace and N. L. Keating, "A systematic assessment of benefits and risks to guide breast cancer screening decisions," Jama, vol. 311, no. 13, pp. 1327–1335, 2014.
[8] J. Seely and T. Alhassan, "Screening for breast cancer in 2018—what should we be doing today?," Current Oncology, vol. 25, no. Suppl 1, p. S115, 2018.
[9] H. D. Nelson, R. Fu, A. Cantor, M. Pappas, M. Daeges, and L. Humphrey, "Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 us preventive services task force recommendation," Annals of internal medicine, vol. 164, no. 4, pp. 244–255, 2016.
[10] H. Burhenne, L. W. Burhenne, F. Goldberg, T. Hislop, A. Worth, P. Rebbeck, and L. Kan, "Interval breast cancers in the screening mammography program of british columbia: analysis and classification.," AJR. American journal of roentgenology, vol. 162, no. 5, pp. 1067–1071, 1994.
[11] . L. Robertson, "A private breast imaging practice: medical audit of 25,788 screening and 1,077 diagnostic examinations.," Radiology, vol. 187, no. 1, pp. 75–79, 1993.
[12] K. Kerlikowske, D. Grady, J. Barclay, E. A. Sickles, and V. Ernster, "Effect of age, breast density, and family history on the sensitivity of first screening mammography," Jama, vol. 276, no. 1, pp. 33–38, 1996.
[13] M. T. Mandelson, N. Oestreicher, P. L. Porter, D. White, C. A. Finder, S. H. Taplin, and E. White, "Breast density as a predictor of mammographic detection: comparison of intervaland screen-detected cancers," Journal of the National Cancer Institute, vol. 92, no. 13, pp. 1081– 1087, 2000.
[14] S. H. Jafari, Z. Saadatpour, A. Salmaninejad, F. Momeni, M. Mokhtari, J. S. Nahand, M. Rahmati, H. Mirzaei, and M. Kianmehr, "Breast cancer diagnosis: Imaging techniques and biochemical markers," Journal of cellular physiology, vol. 233, no. 7, pp. 5200–5213, 2018.
[15] W. A. Berg, L. Gutierrez, M. S. NessAiver, W. B. Carter, M. Bhargavan, R. S. Lewis, and O. B. Ioffe, "Diagnostic accuracy of mammogra phy, clinical examination, us, and mr imaging in preoperative assessment of breast cancer," radiology, vol. 233, no. 3, pp. 830–849, 2004.
[16] M. G. Hildebrandt, A. R. Kodahl, D. Teilmann-Jørgensen, O. Mogensen, and P. T. Jensen, "[18f] fluorodeoxyglucose pet/computed tomography in breast cancer and gynecologic cancers: a literature review," PET clinics, vol. 10, no. 1, pp. 89–104, 2015
[17] K. Warning, M. G. Hildebrandt, B. Kristensen, and M. Ewertz, "Utility of 18fdg-pet/ct in breast cancer diagnostics–a systematic review," Dan Med Bull, vol. 58, no. 7, p. A4289, 2011.
[18] E. A. Jones, T. D. Phan, D. A. Blanchard, and A. Miley, "Breast-specific γ-imaging: molecular imaging of the breast using 99mtc-sestamibi and a small-field-of-view γ-camera," Journal of nuclear medicine technology, vol. 37, no. 4, pp. 201– 205, 2009.
[19] M. O'Connor, D. Rhodes, and C. Hruska, "Molecular breast imaging," Expert review of anticancer therapy, vol. 9, no. 8, pp. 1073–1080, 2009.
[20] F. Scopinaro, R. Pani, G. De Vincentis, A. Soluri, R. Pellegrini, and L. M. Porfiri, "High-resolution scintimammography improves the accuracy of technetium-99m methoxyisobutylisonitrile scintimammography:

use of a new dedicated gamma camera," European journal of nuclear medicine, vol. 26, no. 10, pp. 1279–1288, 1999.

[21] Y. Sun, W. Wei, H.-W. Yang, and J.-L. Liu, "Clinical usefulness of breast-specific gamma imaging as an adjunct modality to mammography for diagnosis of breast cancer: a systemic review and meta-analysis," European journal of nuclear medicine and molecular imaging, vol. 40, no. 3, pp. 450–463, 2013.

[22] K. J. Kuhn, J. A. Rapelyea, J. Torrente, C. B. Teal, and R. F. Brem, "Comparative diagnostic utility of low-dose breast-specific gamma imaging to current clinical standard," The Breast Journal, vol. 22, no. 2, pp. 180–188, 2016.

[23] C. B. Hruska, A. L. Weinmann, and M. K. O'Connor, "Proof of concept for low-dose molecular breast imaging with a dual-head czt gamma camera. part i. evaluation in phantoms," Medical physics, vol. 39, no. 6Part1, pp. 3466–3475, 2012.

[24] C. B. Hruska, A. L. Weinmann, C. M. Tello Skjerseth, E. M. Wagenaar, A. L. Conners, C. L. Tortorelli, R. W. Maxwell, D. J. Rhodes, and M. K. O'Connor, "Proof of concept for low-dose molecular breast imaging with a dual-head czt gamma camera. part ii. evaluation in patients," Medical physics, vol. 39, no. 6 Part1, pp. 3476–3483, 2012.

[25] A. I. Huppe, A. K. Mehta, and R. F. Brem, "Molecular breast imaging: a comprehensive review," in Seminars in Ultrasound, CT and MRI, vol. 39, pp. 60–69, Elsevier, 2018

[26] A. A. van Loevezijn, A. C. van Breda Vriesman, P. A. Neijenhuis, and L. M. Pereira Arias-Bouda, "[Breast-specific gamma imaging in breast cancer]," Ned Tijdschr Geneeskd, vol. 160, p. A9610, 2016.

[27] M. K. O'Connor, M. M. Morrow, K. N. Hunt, J. C. Boughey, D. L. Wahner-Roedler, A. L. Conners, D. J. Rhodes, and C. B. Hruska, "Comparison of tc-99m maraciclatide and tc-99m sestamibi molecular breast imaging in patients with suspected breast cancer," EJNMMI research, vol. 7, no. 1, p. 5, 2017

[28] A. Zhang, P. Li, Q. Liu, and S. Song, "Breast-specific gamma camera imaging with tc-mibi has better diagnostic performance than mag netic resonance imaging in breast cancer patients: A meta-analysis," Hellenic Journal of Nuclear Medicine, vol. 20, no. 1, pp. 26–35, 2017.

[29] L. R. Rechtman, M. J. Lenihan, J. H. Lieberman, C. B. Teal, J. Torrente, J. A. Rapelyea, and R. F. Brem, "Breast-specific gamma imaging for the detection of breast cancer in dense versus nondense breasts," American Journal of Roentgenology, vol. 202, no. 2, pp. 293–298, 2014.

[30] FDA, "Cardiolite kit for the preparation of technetium tc99m sestamibi for injection." http://www.accessdata.fda.gov/drugsatfda_ docs/label/2008/019785s018lbl.pdf.

[31] A. L. Conners, C. B. Hruska, C. L. Tortorelli, R. W. Maxwell, D. J. Rhodes, J. C. Boughey, and W. A. Berg, "Lexicon for standardized interpretation of gamma camera molecular breast imaging: observer agreement and diagnostic ac curacy," European journal of nuclear medicine and molecular imaging, vol. 39, no. 6, pp. 971–982, 2012.

[32] A. L. Conners, R. W. Maxwell, C. L. Tortorelli, C. B. Hruska, D. J. Rhodes, J. C. Boughey, and W. A. Berg, "Gamma camera breast imaging lexicon," American Journal of Roentgenology, vol. 199, no. 6, pp. W767–W774, 2012.

[33] E. S. Burnside, E. A. Sickles, L. W. Bassett, D. L. Rubin, C. H. Lee, D. M. Ikeda, E. B. Mendelson, P. A. Wilcox, P. F. Butler, and C. J. D'Orsi, "The acr bi-rads® experience: learning from history," Journal of the American College of Radiology, vol. 6, no. 12, pp. 851–860, 2009.

[34] E. A. Rakha, J. S. Reis-Filho, F. Baehner, D. J. Dabbs, T. Decker, V. Eusebi, S. B. Fox, S. Ichihara, J. Jacquemier, S. R. Lakhani, et al., "Breast cancer prognostic classification in the molecular era: the role of histological grade," Breast Cancer Research, vol. 12, no. 4, pp. 1–12, 2010.

[35] S. J. Schnitt, "Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy," Modern Pathology, vol. 23, no. 2, pp. S60–S64, 2010.

[36] G. N. Sharma, R. Dave, J. Sanadya, P. Sharma, and K. Sharma, "Various types and management of breast cancer: an overview," Journal of advanced pharmaceutical technology & research, vol. 1, no. 2, p. 109, 2010.

[37] S. J. Schnitt and J. Connolly, "Diseases of the breast," 2004.

[38] M. Duffy, N. Harbeck, M. Nap, R. Molina, A. Nicolini, E. Senkus, and F. Cardoso, "Clinical use of biomarkers in breast cancer: Updated guidelines from the european group on tumor markers (egtm)," European journal of cancer, vol. 75, pp. 284–298, 2017.

[39] B. Y. Azizun-Nisa, F. Raza, and N. Kayani, "Comparison of er, pr and her-2/neu (c-erb b 2) reactivity pattern with histologic grade, tumor size and lymph node status in breast cancer," Asian Pac J Cancer Prev, vol. 9, no. 4, pp. 553– 6, 2008.

[40] M. R. Hussein, S. R. Abd-Elwahed, and A. R. Abdulwahed, "Alterations of estrogen receptors, progesterone receptors and c-erbb2 oncogene protein expression in ductal carcinomas of the breast," Cell biology international, vol. 32, no. 6, pp. 698–707, 2008.

[41] M. E. H. Hammond, D. F. Hayes, M. Dowsett, D. C. Allred, K. L. Hagerty, S. Badve, P. L. Fitzgibbons, G. Francis, N. S. Goldstein, M. Hayes, et al., "American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version)," Archives of pathology & laboratory medicine, vol. 134, no. 7, pp. e48–e72, 2010.
[42] H. Joshi and M. F. Press, "Molecular oncology of breast cancer," in The Breast, pp. 282–307, Elsevier, 2018.
[43] D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire, "Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene," science, vol. 235, no. 4785, pp. 177–182, 1987.
[44] A. C. Wolff, M. E. H. Hammond, J. N. Schwartz, K. L. Hagerty, D. C. Allred, R. J. Cote, M. Dowsett, P. L. Fitzgibbons, W. M. Hanna, A. Langer, et al., "American society of clinical oncology/college of american pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer," Archives of pathology & laboratory medicine, vol. 131, no. 1, pp. 18–43, 2007.
[45] X. Sun and P. D. Kaufman, "Ki-67: more than a proliferation marker," Chromosoma, vol. 127, no. 2, pp. 175–186, 2018.
[46] N. A. Soliman and S. M. Yussif, "Ki-67 as a prognostic marker according to breast cancer molecular subtype," Cancer biology & medicine, vol. 13, no. 4, p. 496, 2016
[47] J. Gerdes, L. Li, C. Schlueter, M. Duchrow, C. Wohlenberg, C. Gerlach, I. Stahmer, S. Kloth, E. Brandt, and H.-D. Flad, "Immuno-biochemical and molecular biologic characterization of the cell proliferation-associated nuclear antigen that is defined by monoclonal antibody ki-67.," The American journal of pathology, vol. 138, no. 4, p. 867, 1991.
[48] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, et al., "Molecular portraits of human breast tumours," nature, vol. 406, no. 6797, pp. 747–752, 2000.
[49] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, et al., "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," Proceedings of the National Academy of Sciences, vol. 98, no. 19, pp. 10869–10874, 2001.
[50] A. G. Waks and E. P. Winer, "Breast cancer treatment: a review," Jama, vol. 321, no. 3, pp. 288–300, 2019.
[51] S. M. Fragomeni, A. Sciallis, and J. S. Jeruss, "Molecular subtypes and local-regional control of breast cancer," Surgical Oncology Clinics, vol. 27, no. 1, pp. 95–120, 2018.
[52] K. S. Park, H. W. Chung, Y. B. Yoo, J.-H. Yang, N. Choi, and Y. So, "Complementary role of semiquantitative analysis of breast-specific gamma imaging in the diagnosis of breast cancer," American Journal of Roentgenology, vol. 202, no. 3, pp. 690–695, 2014.
[53] H. Tan, L. Jiang, Y. Gu, Y. Xiu, L. Han, P. Wu, H. Zhang, and H. Shi, "Visual and semi-quantitative analyses of dual-phase breast-specific gamma imaging with tc-99m-sestamibi in detecting primary breast cancer," Annals of nuclear medicine, vol. 28, no. 1, pp. 17–24, 2014.
[54] E. K. Choi, J. J. Im, C. S. Park, Y.-A. Chung, K. Kim, and J. K. Oh, "Usefulness of feature analysis of breast-specific gamma imaging for predicting malignancy," European Radiology, vol. 28, no. 12, pp. 5195–5202, 2018.
[55] T. Meissnitzer, A. Seymer, P. Keinrath, J. Holzmannhofer, C. Pirich, K. Hergan, and M. Meissnitzer, "Added value of semi-quantitative breast-specific gamma imaging in the workup of suspicious breast lesions compared to mammography, ultrasound and 3-t mri," The British journal of radiology, vol. 88, no. 1051, p. 20150147, 2015.
[56] H.-J. Yoon, Y. Kim, and B. S. Kim, "Intratumoral metabolic heterogeneity predicts invasive components in breast ductal carcinoma in situ," European radiology, vol. 25, no. 12, pp. 3648– 3658, 2015.
[57] J. Yoo, B. S. Kim, and H.-J. Yoon, "Predictive significance of breast-specific gamma imaging for upstaging core-needle biopsy-detected ductal carcinoma in situ to invasive cancer," Annals of nuclear medicine, vol. 32, no. 5, pp. 328–336, 2018.
[58] H. Tan, H. Zhang, W. Yang, Y. Fu, Y. Gu, M. Du, D. Cheng, and H. Shi, "Breast-specific gamma imaging with tc-99m-sestamibi in the diagnosis of breast cancer and its semiquantitative index correlation with tumor biologic markers, subtypes, and clinicopathologic characteristics," Nuclear Medicine Communications, vol. 37, no. 8, pp. 792–799, 2016.
[59] H.-J. Yoon, Y. Kim, K.-T. Chang, and B. S. Kim, "Prognostic value of semi-quantitative tumor uptake on tc-99m sestamibi breast-specific gamma imaging in invasive ductal breast cancer," Annals of nuclear medicine, vol. 29, no. 7, pp. 553–560, 2015.
[60] S. J. Lee, Y. Y. Choi, C. Kim, and M. S. Chung, "Correlations between tumor to background ratio on breast-specific gamma imaging and prognostic factors in breast cancer," Journal of Korean medical science, vol. 32, no. 6, pp. 1031– 1037, 2017.

[61] S. J. Lee, M. S. Chung, S.-J. Shin, and Y. Y. Choi, "Correlation of tumor uptake on breast specific gamma imaging and fluorodeoxyglucose pet/ct with molecular subtypes of breast cancer," Medicine, vol. 97, no. 43, 2018.

[62] J. Yoo, H.-J. Yoon, and B. S. Kim, "Prognos tic value of primary tumor suv max on f-18 fdg pet/ct compared with semi-quantitative tumor uptake on tc-99m sestamibi breast-specific gamma imaging in invasive ductal breast cancer," Annals of nuclear medicine, vol. 31, no. 1, pp. 19–28, 2017.

[63] A. Rodriguez-Ruiz, K. L˚ang, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser, T. H. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, et al., "Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists," JNCI: Journal of the National Cancer Institute, vol. 111, no. 9, pp. 916–922, 2019

[64] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, et al., "International evaluation of an ai system for breast cancer screening," Nature, vol. 577, no. 7788, pp. 89–94, 2020.

[65] Y. Ji, H. Li, A. V. Edwards, J. Papaioannou, W. Ma, P. Liu, and M. L. Giger, "Independent validation of machine learning in diagnosing breast cancer on magnetic resonance imaging within a single institution," Cancer Imaging, vol. 19, no. 1, p. 64, 2019.

[66] T. Xiao, L. Liu, K. Li, W. Qin, S. Yu, and Z. Li, "Comparison of transferred deep neural networks in ultrasonic breast masses discrimination," BioMed research international, vol. 2018, 2018.

[67] R. E. Carter, Z. I. Attia, J. R. Geske, A. L. Conners, D. H. Whaley, K. N. Hunt, M. K. O'Connor, D. J. Rhodes, and C. B. Hruska, "Classification of Background Parenchymal Uptake on Molecular Breast Imaging Using a Convolutional Neural Network," JCO Clin Cancer Inform, vol. 3, pp. 1–11, 02 2019.

[68] C. B. Hruska, D. J. Rhodes, A. L. Conners, K. N. Jones, R. E. Carter, R. K. Lingineni, and C. M. Vachon, "Background parenchymal uptake during molecular breast imaging and associated clinical factors," American Journal of Roentgenology, vol. 204, no. 3, pp. W363–W370, 2015.

PART 2

Artificial intelligence in breast specific gamma imaging

The use of artificial intelligence in Breast-specific Gamma Imaging

detection of breast cancer using convolutional neural networks in breast-specific gamma imaging
&
machine Learning based radiomics in the classification of breast cancer

Author

J.R.J. Willemse

Supervisors

W. Grootjans
Leiden University Medical Center

L.M. Pereira Arias-Bouda
Leiden University Medical Center

F.H.P. van Velden
Leiden University Medical Center

Abstract

Introduction

Breast cancer is the most frequently diagnosed type of cancer in women in 2020. Treatment and prognosis of breast cancer is highly dependent on early and accurate diagnosis. In recent years, many studies have evaluated the use of artificial intelligence for the detection of breast cancer in mammographic images. Another imaging modality, breast-specific gamma imaging (BSGI), or molecular breast imaging, has not yet been subject to AI algorithms to detect breast cancer. In this paper, we aim to develop and evaluate convolutional neural networks (CNNs) that can detect malignancies in breast-specific gamma imaging, and evaluate the efficacy of different machine learning classifiers to classify breast tumors based on estrogen receptor (ER) status, progesterone receptor (PR) status and human epidermal growth factor receptor 2 (HER2-neu) status.

Methods

Three CNNs were created and trained and tested on a total of 3,503 BSGI images. The models varied in complexity in terms of convolutional layers and filter sizes. A semi quantitative lesion segmentation was created based on adaptive thresholding and shape and location analysis. Radiomics features were extracted, and univariate feature selection was applied to disregard redundant features. Different machine learning classifiers, which are widely used in literature for binary classification problems, were evaluated.

Results

In detecting malignancies in a dataset containing clean breasts and breast with malignant lesions, the best performing network reached an area under the receiving operating characteristic (AUROC) of 0.93, while an AUROC of 0.88 was achieved when using the same networks in the classification of malignant versus benign lesions. The best performing machine learning classifiers were the linear discriminating analysis (LDA) classifier for ER and PR status, reaching accuracies of 75% in both receptors. In Her2-neu prediction using machine learning, the best accuracy of 69% was achieved by the RF classifier.

Discussion & conclusions

Based on the results presented in this paper, CNNs can accurately detect malignancies in BSGI images, and discriminate malignancies from benign lesions to a certain extent. The combination of radiomics and machine learning, however, is in its current implementation not accurate enough to predict the ER, PR and Her2-neu status in BSGI images. Future research, however, could focus on using a combination of imaging modalities, such as BSGI, MRI and mammography to improve the predictive accuracy of machine learning.

Contents

1 Introduction

1.1 Breast cancer

In 2020, female breast cancer has become the worlds most diagnosed cancer type in women, with nearly 2.3 million newly diagnosed cases, surpassing lung cancer. In the same year, female breast cancer also led to nearly 700,000 deaths worldwide[1]. The highest rates of incidence of breast cancer is found in developed countries, such as Belgium, the Netherlands and Luxembourg. While still low in developing countries, breast cancer incidence in regions such as South Central Asia and Africa is on the rise due to changes in lifestyle, such as alcohol use, smoking and contraceptive pill usage.

Different types and subtypes of breast cancer exist, the latter being defined by immunohistochemistry (IHC) expression of estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (Her2-neu, also referred to as HER2) and the protein ki-67. The most frequent subtype is hormone receptor positive and Her2-neu negative(ER/PR + / Her2-neu -). This subtype also has the best overall survival and disease-free survival due to the availability of endocrine therapies such as tamoxifen targeting ER and PR. Triple positive tumors (ER/PR +, Her2-neu +) can be treated with either tamoxifen or anti-Her2-neu therapy, such as trastuzumab (Herceptin). Inversely, triple negative (ER/PR - / Her2-neu -) tumors have the worst prognosis, as there is no targeted therapy available. For triple negative tumors, the only feasible systemic therapy currently available is chemotherapy[2].

As breast cancer prognosis is also highly dependent on early and accurate diagnosis, several countries have set up screening mammography programs, focused on finding early signs of non-symptomatic breast cancer. These programs have resulted in an overall reduction in mortality of around 19%[3,4,5].

After findings in initial mammographic screening, or after symptomatic presentation, other imaging modalities can be used, such as targeted mammography, ultrasound (US), magnetic resonance imaging (MRI) and finally, positron emission tomography (PET) to evaluate possible findings with screening mammography[6,7]. Based on these additional diagnostic tools, suspicious findings can be further evaluated using fine needle aspiration(FNA) or biopsies.

Mammographic images are scored using BI-RADS (breast imaging-reporting and data System), a standardized protocol to evaluate mammographic images, developed by the American College of Radiology (ACR). BI-RADS score range from 0 to 6, where 0 represents incomplete or inevaluable images and 6 represents a biopsy proven malignancy[8]. Table 1 shows the interpretation of each category.

Table 1 BI-RADS classification protocol for mammographic images

ASSESSMENT CATEGORY	DESCRIPTION
0	Need additional imaging
1	Negative
2	Benign
3	Probably benign
4	Suspicious
5	Highly suggestive of malignancy
6	Proven malignancy

Another imaging modality which has been introduced in the last years, is breast specific gamma imaging (BSGI), also referred to as molecular breast imaging (MBI). Typical indications for BSGI include: 1) excluding multifocality of breast lesions; 2) evaluation of tumor extent and size; 3) evaluation of discrepancies between clinical and radiological findings; 4) evaluation of mammographic BI-RADS 3 lesions and 5) evaluation of bloody nipple discharge with normal or inconclusive radiological findings.

BSGI is a scintigraphic functional imaging technique of the breast based on the uptake of the radioactive tracer ^{99m}Technetium-sestamibi (^{99m}Tc-sestamibi). ^{99m}Tc-sestamibi is a radiotracer that is able to show increased cellular proliferative activity by increased mitochondrial activity and angiogenesis, generally observed in tumors[9,10].

Although BSGI acquisition protocols may vary between institutions and device manufacturers, imaging is generally performed after administration of the radiotracer in the antecubital vein, and acquiring images from the same angles as in mammography acquisition. Administered activity initially varied between 770 and 1,100 MBq, but recent studies have shown a possible dose reduction up to 2.5 times, mainly as a result of improvements in collimator design[11,12,13]. Image acquisition is performed in two perspectives: mediolateral-oblique (MLO) and craniocaudal (CC), and acquisition times of each projection vary between 8 and 10 min.

BSGI has several advantages when compared to other additional diagnostic imaging modalities such as MRI. With a comparable sensitivity in BSGI and MRI in the detection of breast cancer (0.84 ; 95% CI: 0.79 – 0.88 and 0.89; CI: 0.84 – 0.92, respectively), BSGI showed a significantly higher specificity of 0.82 (95% CI: 0.74 – 0.88) than MRI (0.39; CI: 0.30 – 0.49)[14] . Also, because BSGI is a functional imaging modality, rather than anatomical, BSGI images are less influenced by breast density, which on anatomical imaging can influence interpretability of images[15]. Finally, BSGI acquisition is faster, interpretation times are shorter, and costs are lower. The disadvantage of BSGI is the whole-body dose distribution of around 5 mSv, following a dose administration of 600 MBq ^{99m}Tc, approximately a ten-fold increase when compared to mammography alone[16,17].

1.2 Artificial Intelligence

1.2.1 Lesion detection

For many years, studies have been published about computer-aided detection (CAD), where CAD is used as an assistance to radiologists in diagnosis and medical image interpretation. In 2010, around 74% of mammograms in the United States were interpreted with the use of CAD[18]. Although welcomed with much enthusiasm at first, later studies had shown that the benefits of CAD in medical imaging in some cases led to reducing the radiologist's accuracy, longer interpretation times, and lead to higher recall and biopsy rates[19,20,21].

More recently, developments in artificial intelligence have resulted in highly accurate algorithms, which have been tested on different medical aspects, such as skin lesion analysis[22], retinal image analysis[23], brain MRI segmentation[24], or staging of lung cancer[25]. These algorithms are types of deep learning networks, a subtype of machine learning. Machine learning algorithms are complex mathematical models that use big sets of data to learn patterns in that data, which can then be used to make predictions on new, unseen data. Deep learning specifically is an algorithm based on neural networks, in which several layers of nodes are interconnected, simulating interconnected neurons in the visual cortex of animals[26]. One type of neural networks is the convolutional neural network (CNN), which is very well suited for image analysis and image classification. In CNNs, an image is

convolved in convolutional layers, which reduce an image into less complex features, such as edges, shapes and combinations of shapes. These features are then propagated through the neural layers, resulting in a classification for the original input image.

The first aim of this paper is to create a set of different CNNs, which might be able to quickly and automatically detect abnormalities in BSGI images, as well as discriminate malignant from benign lesions.

1.2.2 Lesion Classification

1.2.2.1 Radiomics

In 1973, *Haralick* et al. first proposed the use of textural features of images with the goal of classification[27]. Since then, interest has been growing in the medical field for application of image feature analysis in phenotyping tumors. In turn, this led to the introduction of radiomics, the concept of extracting of a large number of quantitative features from any 2D or set of 2D images, and finding the relationship between this quantitative information and qualitative clinical data, to find the underlying pathophysiology of a tissue.

Radiomics extraction returns a set of numbers representing a quantitative description of a predefined region of interest (ROI). In oncology, this usually includes tumor primary features such as shape, size and intensity, as well as secondary features such as texture analysis[28]. These features have been widely investigated over the last years for their use to characterize tumors noninvasively, based on medical imaging. Most often, anatomical imaging modalities such as CT and MRI are used as a basis for radiomics[29].

Although not yet implemented on BSGI images, several studies have attempted to describe the relationship between quantification of tumors visible on BSGI and tumor classification in terms of invasiveness, tumor subtyping, and prognosis. In 2017, *Lee et al.* correlated the tumor-to-background ratio (TBR), defined as the ratio between the mean uptake in a lesion and the mean uptake value of three circular regions in breast tissue outside the lesion, , and found significant correlations in PR-status of lesions[30]. Similarly, *Yoon et al.* found that a high TBR was significantly correlated with ER/PR negativity, as well as HER2-neu negativity[31].

Although different strategies were used in the studies described above in calculating the TBR, these studies indicate that image features derived from BSGI images might correlate with clinical prognostic factors such as receptor status. Even though scintigraphic images have not widely been used in radiomics analysis, based on these studies BSGI images could be a suitable subject to radiomics analysis.

No studies have yet been published on the application of radiomics in BSGI images to predict ER, PR and Her2-neu status of tumors identified on BSGI images.

1.2.2.2 Machine learning

Similarly to the more specific deep learning, the broader concept of machine learning (ML) has gained traction in the recent years. ML encompasses a wide variety of algorithms able to classify new sets of datapoints into a certain category, based on previously learned patterns in training data. With large numbers of features derived during radiomics extraction, and selecting the features with the highest variance, or discriminating power, machine learning is well suited to train, analyse and predict on datasets containing radiomics features.

The combination of radiomics and ML has been subject to an increasing number of studies over the past years, especially in the field of oncology. In 2018, for instance, *Lu et al.* showed that the

combination of radiomics derived from MRI and ML achieved a 81.8% accuracy in predicting molecular subtypes of gliomas[32]. Also, a study by *Hyun et al.* described the use of radiomics derived from PET images and ML in predicting histological subtyping of lung cancer and achieved an accuracy of 76.9%. [33] Other studies have reported on the combination of radiomics and ML in subtype classification of pancreatic tumors[34], renal tumors[35], meningiomas[36], thymic epithelial tumors and other cancer types[37].

Successful non-invasive tumor classification with a combination of radiomics and ML could pave the way for a reduction of painful, time-consuming biopsies, as well as a reduction in the time from diagnosis to targeted treatment.

The secondary aim of this paper is to (semi-)automatically segment lesions in BSGI images, perform radiomics extraction on these regions of interest (ROI), and apply different ML algorithms, which are used in other research aiming to combine radiomics and ML, to predict ER, PR, and Her2-neu positivity or negativity in breast tumors.

2 Methods

2.1 Data

2.1.1 Data acquisition

Data were acquired retrospectively over the period of January 1st 2014 to January 1st 2021 from the Alrijne Hospital Leiderdorp, The Netherlands. A total of 1,423 patients underwent BSGI for problem-solving in that time frame.

Problem solving is defined as the evaluation of indeterminate breast abnormalities or symptoms. Typical indications are summarized in table 2. This includes the evaluation of discrepancies between clinical and radiological findings, the evaluation of mammographic BI-RADS 3 (probably benign) lesions where patient reassurance is required, or evaluation of (bloody) nipple discharge with normal or inconclusive radiological findings[38].

Table 2 Typical indications for problem-solving with BSGI

	TYPICAL PROBLEM-SOLVING INDICATIONS
1	Evaluation of discrepancies between clinical and radiological findings
2	Evaluation of mammographic BI-RADS 3 (probably benign) lesions where patient reassurance is required
3	Evaluation of (bloody) nipple discharge with normal or inconclusive radiological findings

Based on reporting by trained nuclear medicine physicians, as is clinical practice, scans were classified using NG-BI-RADS[39] scores designed for molecular breast imaging. BSGI scans with a BI-RADS score of >3 were considered to be suspicious for malignancy, and thus were subject to further investigation in the form of biopsies or resection of suspicious tissue. BSGI scans with a BSGI BI-RADS score of 3 or lower were marked as non-suspicious, and therefore these patients were not required to undergo any direct further investigation. It has to be noted, however, that this does not mean that nothing is visible in any of these images, as nuclear medicine physicians generally also include clinical factors, such as symptoms, and other medical imaging modalities in their decision making.

For confirmed malignancies (CM), hormonal receptor status was also collected, and consecutively subcategorized into either a positive or negative ER-status, PR-status and HER2-neu status.

For included patients, both the cradio-caudal (CC) and mediolateral-oblique (MLO) projections of the relevant breast, with a suspected malignancy or abnormality, were extracted and included, in order to create a diverse dataset which responds to different projections. In the inclusion period (2014-2021), two different BSGI devices were used (*Dilon 6800, Dilon Diagnostics* & *Discovery NM 750b, GE Medical Systems*). The single-detector *Dilon 6800* results in fewer images than the dual detector *Discovery NM 750b*, as the dual-detector device creates two separate images for each projections, which can then be combined to create an averaged image of the breast. CC and MLO projections of the relevant breast made by the single detector *Dilon 6800* were included, and the averaged CC and MLO projections of the dual-detector were included.

DICOM images were extracted and labelled accordingly (Suspicious: pathologically confirmed Malignant, Suspicious: pathologically confirmed benign and non-suspicious) . An overview of the data is given in the figure below:

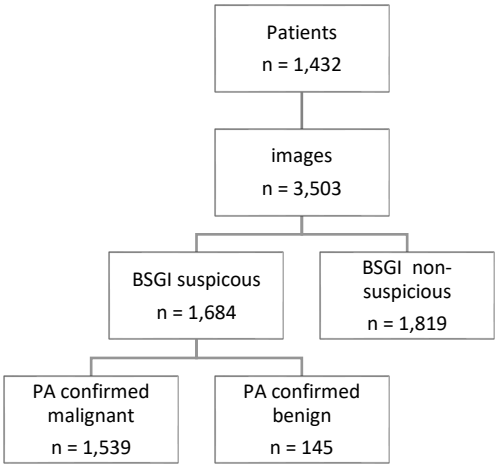


Figure 1: Overview of suspicious: Malignant, suspicious: Benign and non-suspicious data. PA = Pathologically assessed

Out of all 3503 images, 2366 images were acquired using the *Dilon 6800*, while 1137 were acquired using the *Discovery NM 750b*. Image parameters and acquisition settings can be seen in table 3.

Table 3 Image and acquisition parameters of Both BSGI devices

	DEVICE	
	Dilon 6800	*Discovery NM 750b*
AMOUNT	*2366*	*1137*
PIXEL ARRAY SIZE	*80 x 80*	*96 x 96*
PIXEL SPACING	*3,2 mm*	*2,46 mm*
TOTAL COUNTS	*131568 ± 51483*	*71280 ± 32334*
RADIONUCLIDE	*$99^{m}Tc$*	*$99^{m}Tc$*
DOSE	*740 MBq*	*250 -350 MBq*
COLLIMATOR	*LEGP*	*LEHR*
FRAMES	*1*	*8*
ACQUISITION TIME PER FRAME	*480s*	*60s*
ACQUISITION TIME PER PROJECTION	*480s*	*480s*

2.1.2 Data Preparation

Pixel arrays with a 96 x 96 size were downsized to 80 x 80 using bilinear interpolation, to match all sizes of images to be used for training and testing.

The pathologically confirmed malignancies and non-suspicious labelled images were used for initial training and evaluation of the model, and the pathologically confirmed benign lesions were set aside for later evaluation of the models performance in discrimination malignant from benign lesions. These benign images could have been placed within the category of non-suspicious lesions, but the size of the dataset with benign images (n = 145) is too small for the entire dataset to impact the training process.

This image dataset (n = 3,503, malignant and non-suspicious) was randomly split in a ratio of 14 : 86, resulting in a training set of 3,007 images, and a test set of 496 images. The original split ratio of 15 : 85 was changed due to a number of corrupt files in the training set. The ratio between a training and test sets was chosen as this ratio should generally be in the range of 10-20% to achieve the highest accuracy[40]. The training set was further randomly separated into a normal training set and a validation set with a ratio of 80: 20. After each epoch of the training process, a new split in training and validation sets was performed. This overview can be seen in figure 2.

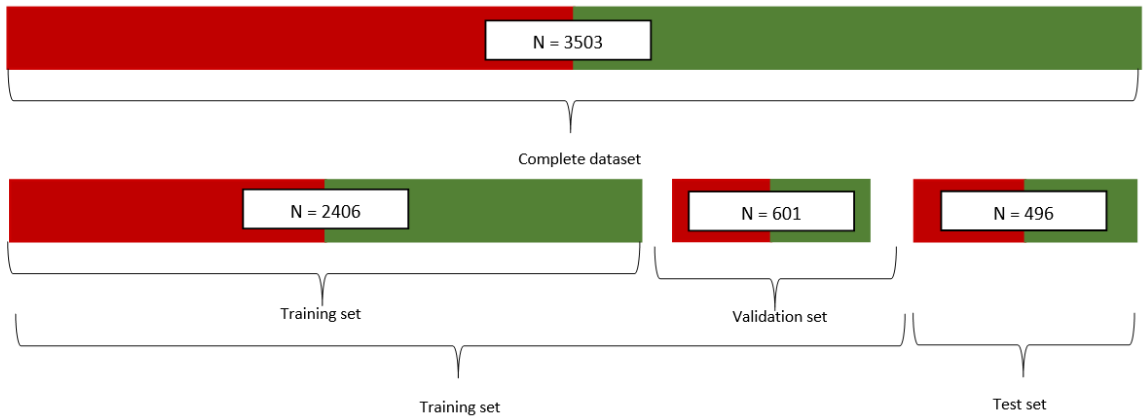


Figure 2 Overview of Training, validation and test dataset size

2.1.3 Data Augmentation

As neural networks generally have an increased performance with more training data, data augmentation was implemented to increase the size of the training and validation sets by a factor 2. Data augmentation works on the principle of artificially generating new data from original data. Each image in the original training set was flipped horizontally to create a new image (figure 3). Because algorithms trained on augmented data will also be tested on non-augmented data, images were not flipped vertically to create a set of four images in the augmented training set. This would mean that MLO projections would be augmented into unrealistic images with an upside-down breast, which would not be present in actual BSGI images.

Step 1) Each Image within the test set is augmented

Step 2) Both the original image and the augmented image are used as inputs for the model

Step 3) The two outcomes are averaged to give a final prediction for the image

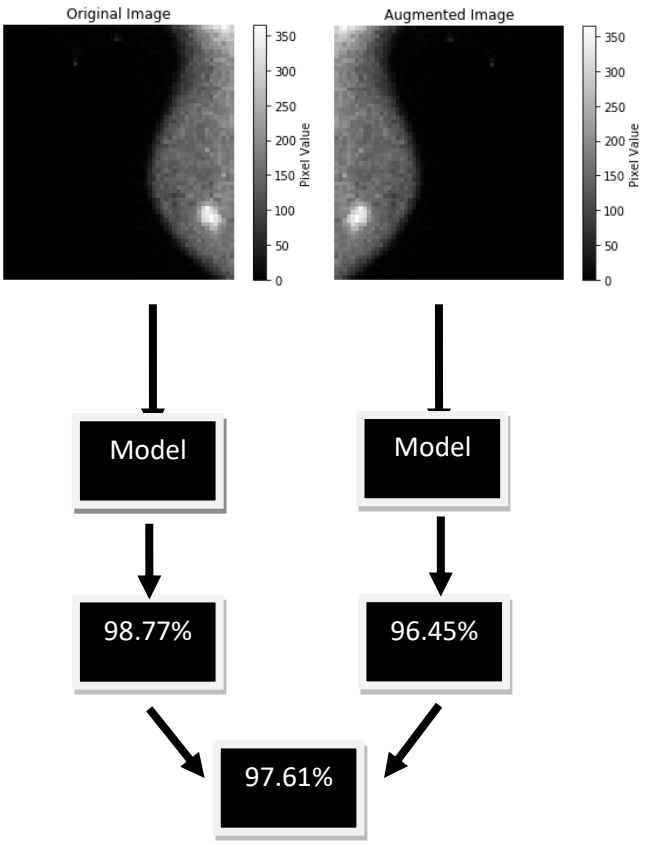


Figure 3 Testing a model using augmented images

Data augmentation resulted in a training and validation set of 4,812 and 1,202 images, respectively.

Testing of the network trained on augmented data was performed on the original test set (n=496) as well on an augmented version of the test set (n = 992), where the outcomes of both the original and the augmented image were averaged to give a final prediction.

2.2 Lesion Detection

2.2.1 CNN Architecture

The principle behind CNNs is the breakdown of an image into less complex features, which are then used to train a model through fully connected layers. The features are extracted using convolutions, and are then run through max-pooling layers in order to reduce the feature map size (for computational purposes) and decrease variance. Pooling is a process in which the size of a given array is reduced by taking the maximum, average or minimum value of a certain region of that array, and creating a new array with those values. For instance, an 4 by 4 array pooled by a 2 by 2 max-pooling layer would result in a new 2 by 2 array containing the maximum values of the four quadrants of the original array. The networks described in this paper use maximum-pooling instead of Average Pooling (AP) or Minimum-Pooling (MP), as max pooling is generally better suited for situations in which higher pixel intensities (e.g. lesions) are sought after, and there is a relatively small-amount of pixel data available, which would be blurred out by average pooling.

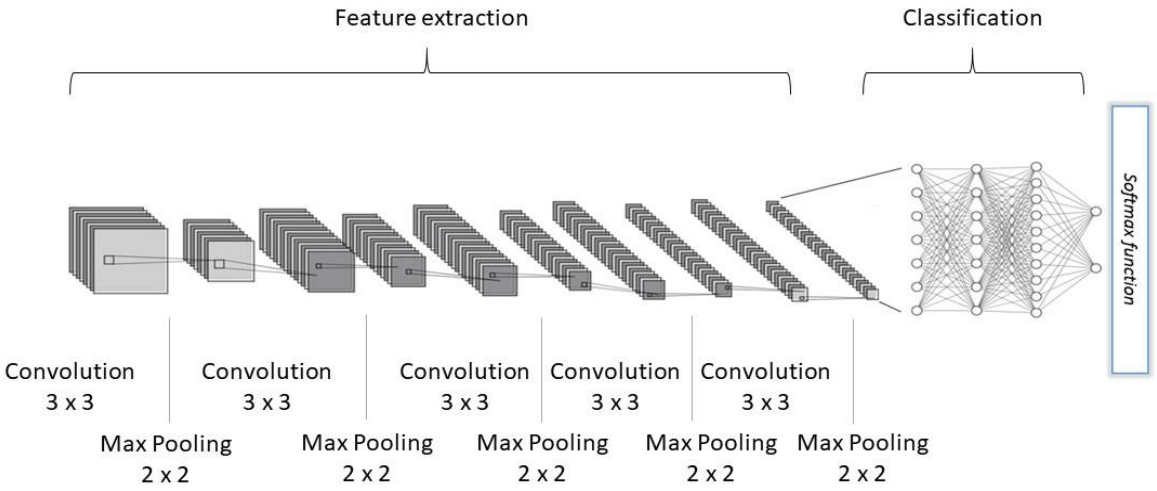


Figure 4 Network architecture of Network 2

Three CNNs were developed using *Pytorch(v.1.8.1).* The networks were all based on the principle of convolutional layers, rectified linear units (ReLU), each followed by a Max Pooling layer. The final feature maps are then concatenated into flattened vectors which are the input values to the two fully connected layers. The three networks, however, varied in the number of convolutional layers (and max-pooling layers), the amount of features, and filter sizes. The Network architecture of Network 2 can be seen in figure 4, and settings can be seen in table 4.

The amount of possible designs for CNNs is theoretically infinite, and designing a network is mostly a trial and error process. As there is no rule of thumb for the amount of convolutional layers and other network parameters in solving a classification problem, the choices of these networks designs were based on general image complexity, which is relatively low in BSGI images. Therefore, designs of the CNNs described in this paper are somewhat arbitrary, but increasingly complex with three convolutional layers for the first network, five convolutional layers for the second network, and seven convolutional layers for the final network. Networks in similar size, though, have been previously been described in the use of deep learning applications in nuclear medicine imaging techniques, such as thyroid scintigraphy[41]. Following the convolutional layers, each of the three networks is concluded by two fully connected layers that provide the actual classification by updating the weights, or connections between nodes, after each epoch.

An epoch is defined as a combination of one forward and backward pass of the entire training set through the algorithm.

Because the output values from the last fully connected layer can be negative or positive, they cannot be interpreted as probabilities. The soft-max activation function uses outputs from the last fully connected layer to provide a probability distribution for a given input image, thus normalizing the outputs to values between 0 and 1. As these two probabilities, the probability of a malignancy or no malignancy, sum up to 1, in this paper only the probability of a malignancy will be communicated.

Table 4 Network configuration parameters of Network 1, Network 2 and Network 3

	NETWORK 1	NETWORK 2	NETWORK 3
BLOCK 1	**Conv.* 1 → 32 (3,3)** **Max – Pool** (2,2)**	**Conv. 1 → 32 (3,3)** **Max - Pool (2,2)**	**Conv. 1 → 16 (3,3)** **Max - Pool (2,2)**
BLOCK 2	**Conv. 32 → 64 (3,3)** **Max - Pool (2,2)**	**Conv. 32 → 32 (3,3)** **Max - Pool (2,2)**	**Conv. 16 → 32 (3,3)** **Max - Pool (2,2)**
BLOCK 3	**Conv. 64 → 128 (3,3)** **Max - Pool (2,2)**	**Conv. 32→ 64 (3,3)** **Max - Pool (2,2)**	**Conv. 32 → 64 (3,3)** **Max - Pool (2,2)**
BLOCK 4		**Conv. 64 → 128 (2,2)** **Max - Pool (2,2)**	**Conv. 64 → 64 (3,3)** **Max - Pool (2,2)**
BLOCK 5		**Conv. 128 → 512 (3,3)** **Max - Pool (2,2)**	**Conv. 64 → 128 (3,3)** **Max - Pool (2,2)**
BLOCK 6			**Conv. 128 → 256 (3,3)** **Max - Pool (2,2)**
BLOCK 7			**Conv. 256 → 512 (3,3)** **Max - Pool (2,2)**
FC*** 1	**In: 128** **Out: 512**	**In: 512** **Out: 1024**	**In: 512** **Out: 1024**
FC 2	**In: 512** **Out: 2**	**In: 1024** **Out: 2**	**In: 1024** **Out: 2**

***CONVOLUTION**
**** MAXIMUM POOLING LAYER**
***** FULLY CONNECTED LAYER**

2.2.2 Training

Training was performed in a two stages. The first stage was to determine the optimal number of epochs required to minimize loss, while preventing the model from overfitting to the training set.

Loss is a measure of mismatch between predicted labels which your model outputs and the actual labels. In the networks described in this paper, the mean squared error (MSE) loss-function was used. In this function, y represents the predicted output and $\tilde{y}$ represents the true label.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

After each epoch, the training- and validation losses should converge and decrease, as the weights of the fully connected layers are updated based on images and their labels that have gone through the network.

In overfitting, the training- and validation losses no longer converge because of an increasing validation loss. Overfitting is a sign of an overcomplexity in a network, with weights adjusted specifically to each training image, and negatively influences the performance on external data sets. Finding the point of overfitting is generally done by running the training data through the network for an extended amount of epochs, and visually finding the point where the losses start to diverge (Appendix 2).

Optimization of the model was performed using the Adam Optimizer, a widely-used stochastic gradient optimizer. This optimizer uses a learning rate to reduce the loss measured by the MSE loss-function. This means that when, after an epoch a certain mismatch is found between the predicted label of an image and the actual label (i.e. malignant or benign), the optimizer will update the weights in the network to reduce the loss in the next epoch. The learning rate determines the extent to which weights in the fully connected layers are updated after each epoch. A high learning rate adjusts the weights more when a mismatch is found when compared to a lower learning rate. By

updating the weights in the network after each epoch, the network reaches an equilibrium where updating the weights have an increasingly smaller effect on accuracy. By adjusting the learning rate downwards further in the training process, smaller adjustments in the weights are possible, which would not have been achievable with the initial learning rate.

To increase model performance, the learning rate was updated after every 5 epochs. These learning rates can be seen in table 5.

Table 5 Learning rates

EPOCH	LEARNING RATE
0 - 5	0.001
6-10	0.0008
11-15	0.0006
16-20	0.0005
21-25	0.0004
> 25	0.0003

After having determined the required number of epochs needed to optimally train the model without overfitting, the model was trained with the settings described in former paragraphs, with a batch size of 100 images, and the optimal number of epochs.

2.2.3 Model Evaluation

Model testing and evaluation was performed on a per-image basis with all three designed models. The test set (n = 496) as well as the augmented test set (n = 992) were run through the trained models and outcomes were recorded. Outcome measures include the area under the receiver operating characteristic (AUROC), accuracy, sensitivity and specificity.

Additionally, a voting-based strategy was implemented to see how a combination of the three models would perform. In this model, the average of the three other networks would provide the predicted label for each image in the test set. This could eliminate specific cases in which one network would perform poorly and provide a wrong predicted label, whereas the other two network would perform properly and average out the poor prediction of the first network. This model will be referred to as the Averaged Model.

Finally, the dataset of 145 pathologically confirmed benign lesions with an addition of 145 random images from the pathologically confirmed malignancies test set and their respective augmented datasets were run through the algorithms to evaluate the model's performance on discriminating benign from malignant lesions. To reduce the effect of the randomized selection of the 145 malignant images, each network was tested three times and the results were averaged to give a representative performance.

2.2.4 Occlusion mapping

Because CNNs are often viewed as 'black boxes' which are unable to show why a certain prediction have been given, interpretability of results can often be difficult, especially in unsupervised algorithms.

Occlusion mapping was used to check the triggering conditions for each of the predicted labels. In occlusion mapping, a small part of the image is occluded iteratively, until all parts of the image have been blocked once. For the heatmaps generated in this algorithm, each set of 2 by 2 pixels, with a stride of 1, in an image was occluded by replacing them with pixels with an intensity of 70% of the average of all non-zero pixels in the whole image. This percentage is based on creating a gray pixel value similar to the background of the breast, excluding the region outside the breast (zero pixel values). Taking a higher percentage, such as 100% of the mean, replaces the occluded pixels with a too high intensity, as the lesion is also responsible for this mean pixel value.

By predicting the image label with each occlusion, it is possible to see when the occlusion of a region leads to changes in the predicted label. From this information, an occlusion heatmap can be generated to assess whether the network makes predictions based on the expected regions (i.e. a prediction of a malignancy should show higher intensities on the heatmap at the lesion location).

2.3 Machine Learning: predicting the receptor status using radiomics

To extract radiomics from a lesion, a segmentation of the lesion has to be made. The radiomics extracted from this segmentation can then be used as inputs to train machine learning models.

2.3.1 Lesion Segmentation

For the segmentation process, only images with a 96 x 96-pixel array size were used (*Discovery NM 750b).* The 80 x 80 images were not used in this case, because each pixel is expected to have less information in it, and because these types of images are no longer acquired for problem-solving purposes.

Lesion segmentation was performed in a three-step process: 1) Adaptive thresholding to find possible lesion regions; 2) Identifying the lesion in the predetermined regions based on sphericity and location and 3) Marching squares lesion-edge delineation.

Before starting the segmentation process, all images were blurred with a Gaussian Filter with $\sigma=3$ (pixels) to remove potential noise in the image which could affect thresholding. A sigma of 3 was found to reduce the noise sufficiently enough, while keeping the lesions shape intact and similar in size and intensity. A lower sigma would not reduce noise sufficiently, while a too high sigma would blur the lesion too much, and could therefore lead to an increased lesion size.

First, the breast was separated from the background by applying region growing. This ensured that breast size (i.e. higher mean intensities in the overall image) would not affect the final segmentation. Also, this causes any possible artefacts with high intensities present with the FOV and outside the breast to be disregarded and not be seen as a potential lesion by the segmentation algorithm.

In order to find possible lesion locations within a BSGI scan, adaptive thresholding was used based on the intensity distribution of the breast region in the image. The threshold for regions of interest were set at 90% of the maximum intensity of the breast region. This 90% would ensure that only the highest intensities, and thus the suspected lesion, would be marked as a ROI. With this thresholding technique, sometimes the thoracic wall would be marked as a region of interest.

To counter situations in which the thoracic wall would be marked as a ROI, all possible lesion locations were analyzed and ranked by their sphericity using the following function:

$$Sphericity = \frac{2\pi R}{P} = \frac{2\sqrt{\pi A}}{P}$$

R = radius of circle with same area as thresholded region

P = perimeter of thresholded region

If one of the regions had a clearly higher level of sphericity, this location was regarded to be the lesion location. In the case of two possible locations having the same level of sphericity (within a ± 20% margin), the lesion with the smallest Euclidean distance to the center of the whole image would be regarded as the lesion.

By cropping the image to a 56 x 56 array, which was found to be large enough to include the entirety of each lesion, with the lesion location in its center, a marching squares Python algorithm was used to delineate the lesion from the surrounding healthy tissue. The iso-level of the marching squares algorithm was mainly based on the intensities of the cropped image, with the iso-level being set at 120% of the mean intensity. This value of 120% was found by trial and error, and ensured proper delineation in 117 of 145 cases. If unsatisfied about the delineation, the iso-level was manually changed to improve the outcome.

The iso-contours resulting from the marching squares algorithm were used to generate a an area over the lesion. This resulted in a final segmentation of the lesion.

The result of each segmentation step can be seen in figure 5.

2.3.2 Feature extraction

After final ROI selecting, radiomics were extracted using *PyRadiomics(v.3.0.1),* which consists of 108 radiomic features. The radiomics included in this extractor include first order features, 2D shape features, 3D shape features, gray level co-occurrence matrix (GLCM) features, Gray Level Size Zone Matrix (GLCM) features, Gray Level Run Length Matrix (GLRLM) features, Neighboring Gray Tone Difference Matrix (NGTDM) features and Gray Level Dependence Matrix (GLDM) features. In addition, the Coefficient of variation (COV), and the following Tumor-to-Background (TBR) ratios were calculated: TBR_{mean}, TBR_{max} and TBR_{high}.

$$TBRmean = \frac{\text{Mean Lesion Intensity}}{\text{Mean Background Intensity}}$$

$$TBRmax = \frac{\text{Maximum Lesion Intensity}}{\text{Mean Background Intensity}}$$

$$TBR_{high} = \frac{\text{Top 30\% Lesion Intensities}}{\text{Mean Background Intensity}}$$

2.3.3 Feature reduction

After manually removing non-unique variables, in which the same feature gives the same value for each lesion, and 3D-image features, 48 radiomic features for each segmentation were labelled with their respective hormonal receptor status (ER +/-, PR +/- and HER2-neu +/-).

In order to reduce data dimensionality, and select the most statistically significant features, univariate feature selection was used to select the 15 and 25 features with the least variance from the 48 radiomic features.

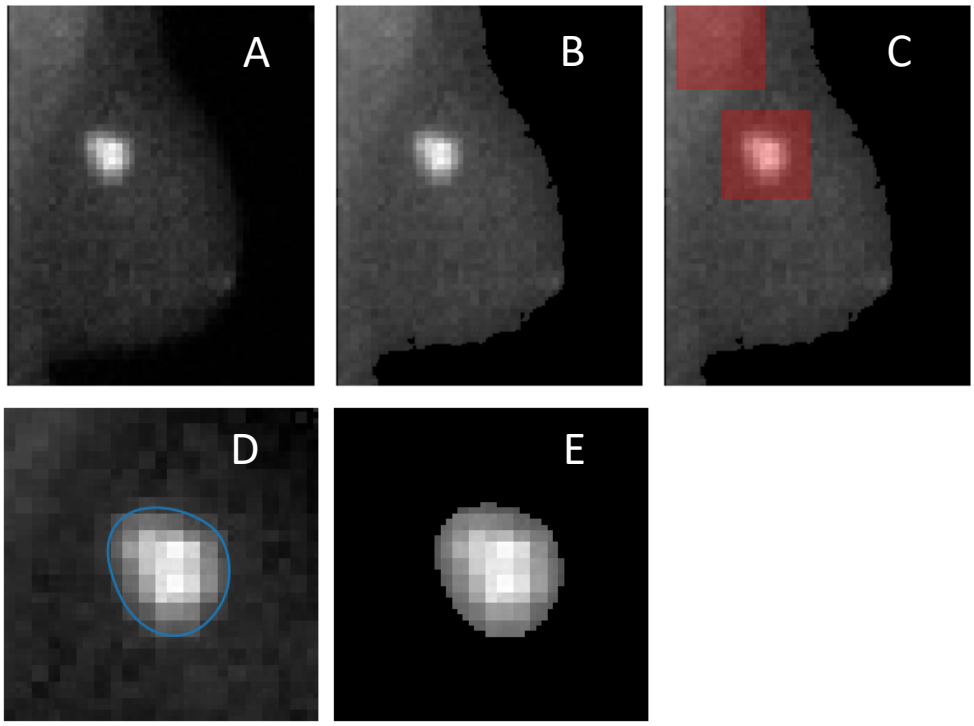


Figure 5 A) Original image with focal uptake. B) Image after morphological operations with background removed. C) Possible lesion locations based on thresholding of image b. D) Lesion segmentation. E) Final segmentation.

2.3.4 Machine Learning

Machine learning implementation was performed using the *Scikit-learn package* (*v. 0.24.2*) in P*ython(v. 3.7).* Training and test sets were split with a ratio of 70:30 resulting in sets of n = 102 and n = 43, respectively. Internal validation dataset size depends on the number of k-folds, where a higher number of k-folds results in a smaller size of the internal validation set size.

Each model was trained and evaluated with different stratified k-folds : 3, 5, 10 and 15. These k-folds were chosen to evaluate a wide range of different splits within the training set, and are common values for evaluating ML models.

A total of 7 widely used machine learning algorithms were evaluated separately: logistic regression (LR), random forest (RF), gaussian naive bayes (GNB), linear discriminant analysis (LDA), decision tree classifier (DTC), support vector machine (SVM) and K-nearest neighbors (KNN). These models are delivered with the *Scikit-learn package*, and can be adjusted with different settings.

For LR, the liblinear solver was used as an optimizer and one-versus-rest classification scheme, which is especially suited for smaller datasets with a binary classification problem such as the one used in this paper.

In the RF model, the number of trees used was set at 50, which was found through trial and error and proved to be the best performing amount of trees. The GNB model provided by *Scikit-learn* was untunable, and was therefore let in default. The LDA model uses the singular value decomposition (SVD) solver, without shrinkage applied. In the DTC model, the GINI impurity index was used for measuring the quality of a split, and a maximum depth of the tree was set at 15, as a tree too deep would encourage overfitting of the model, while a tree with a low depth would result in reduced performance. In SVM, the radial basis function (RBF) kernel was used, with degree of 3 used for this polynomial, as it was expected that datapoints, or the features, would not be linearly separable by a linear kernel. Other settings in the SVM model remained at default. In KNN, the value of *k,* was set at 5, which was found through trial and error, and resulted in the highest accuracy. Also in KNN, the weight function was set at uniform, so that all neighbors of a given sample were weighed equally.

Also, a voting classifier was built based on the five best performing models. The voting-based classifier gives a prediction based on these outcomes of the other models, and selects the most occurring prediction. Selecting just five, and not all of the models, ensures that poor performing models are not included in the decision making, while taking into account most of the other, well-performing models.

3 Results

3.1 Network training

Each network was trained for an excessive number of epochs to visually determine the point of overfitting, where the validation and training losses started diverging. For network 1 and 2 with non-augmented training data, this resulted in 11 epochs for both networks. For network 3, optimal training on non-augmented data was set at 7 epochs. With augmented training data, epochs were set at 10, 10 and 7 for networks 1, 2 and 3, respectively.

All accuracy and loss graphs can be found in appendix 1A and 1B.

3.2 Network Performance

AUROC scores were calculated for each network. The ROCs of the networks trained on non-augmented data with non-augmented test data can be seen in figure 6A. Figure 6B and 6C show the ROCs of the networks trained on the original training sets tested on augmented data and trained on augmented data and tested on original test sets, respectively. Finally, the ROCs of the network trained on augmented data and tested on augmented data can be seen in figure 6D.

Because the model gives an output in the form of a probability between 0 to 1 (or 0% to 100%), the network does not tell whether or not a malignancy is present. One way to convert a probability into an actual conclusion is to take the highest prediction of the network. With a cut-off probability of 0.5 (i.e, everything with a malignancy probability of 0.5 of higher was deemed malignant), accuracies, sensitivities and specificities could be calculated.

The performances of all networks with different combinations of training and testing on non-augmented and augmented datasets are summarized in table 6.

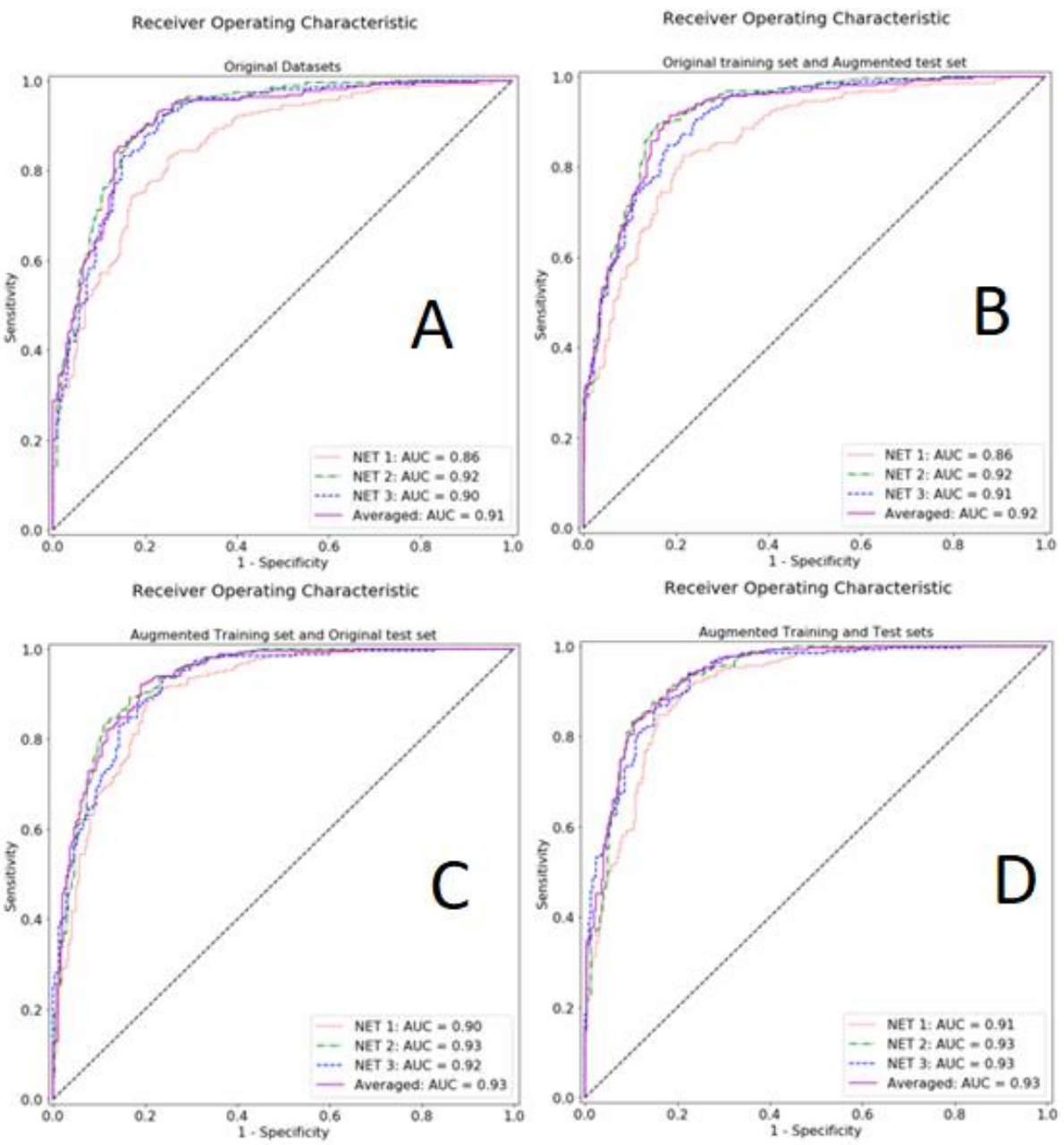


Figure 6 ROC curves of A) Networks trained and tested on original datasets; B) Networks trained on augmented datasets and tested on original datasets; C) Networks trained and tested on augmented datasets and; D) Networks trained on original datasets and tested on augmented datasets.

Table 6 Performances of all networks including the averaged model in detecting malignancies, using different combinations of augmented and non-augmented datasets. In each table, columns represent the results for a given combination of augmented or non-augmented training and test sets. Sensitivity and specificity values were calculated at a cut-off value of 0.5 (or 50%).

Network 1

Training set	non-augmented	Augmented	non-augmented	Augmented
Test set	non-augmented	Augmented	Augmented	Non-augmented
AUC	0.86	0.91	0.86	0.9
Accuracy	77%	84%	79%	79%
Sensitivity	77%	79%	77%	79%
Specificity	78%	89%	82%	88%
TP	190	195	189	196
FP	55	27	44	30
TN	195	223	206	220
FN	57	52	58	51

Network 2

Training set	Non-augmented	Augmented	non-augmented	Augmented
Test set	Non-augmented	Augmented	Augmented	Non-augmented
AUC	0.92	0.93	0.92	0.93
Accuracy	84%	85%	86%	85%
Sensitivity	84%	78%	87%	78%
Specificity	84%	92%	85	91%
TP	208	192	214	192
FP	39	19	37	22
TN	211	231	213	228
FN	39	55	33	55

Network 3

Training set	non-augmented	Augmented	non-augmented	Augmented
Test set	non-augmented	Augmented	Augmented	Non-augmented
AUC	0.9	0.93	0.91	0.92
Accuracy	84%	85%	82%	85%
Sensitivity	74%	77%	73%	76%
Specificity	93%	92%	92%	93%
TP	184	191	180	188
FP	17	19	20	17
TN	233	231	230	233
FN	63	56	67	59

Averaged Model

Training set	non-augmented	Augmented	non-augmented	Augmented
Test set	non-augmented	Augmented	Augmented	Non-augmented
AUC	0.91	0.93	0.92	0.93
Accuracy	85%	86%	86%	86%
Sensitivity	81%	79%	81%	78%
Specificity	88%	93%	90%	93%
TP	201	194	200	192
FP	31	18	25	17
TN	219	232	225	233
FN	46	53	47	55

3.3 Occlusion Mapping

With occlusion mapping, it is possible to map the regions of an image responsible for a certain decision by the AI. The aim was to perform occlusion mapping on the best performing model in terms of AUROC score and accuracy, sensitivity and specificity. Therefore, the second best model, Network 2, was selected (trained and tested on augmented data: AUROC = 0.93, accuracy = 85%, sensitivity = 78% and specificity = 92% at a cut-off probability of 0.5).

In figure 7, examples of true positives, true negatives, false positives and false negatives are shown including corresponding heatmaps. These heatmaps indicate image regions based on which the model mostly based its decision to either predict a malignancy (in the case of a malignancy probability of 0.5 or higher) or no malignancy (in the case of a malignancy probability of lower than 0.5).

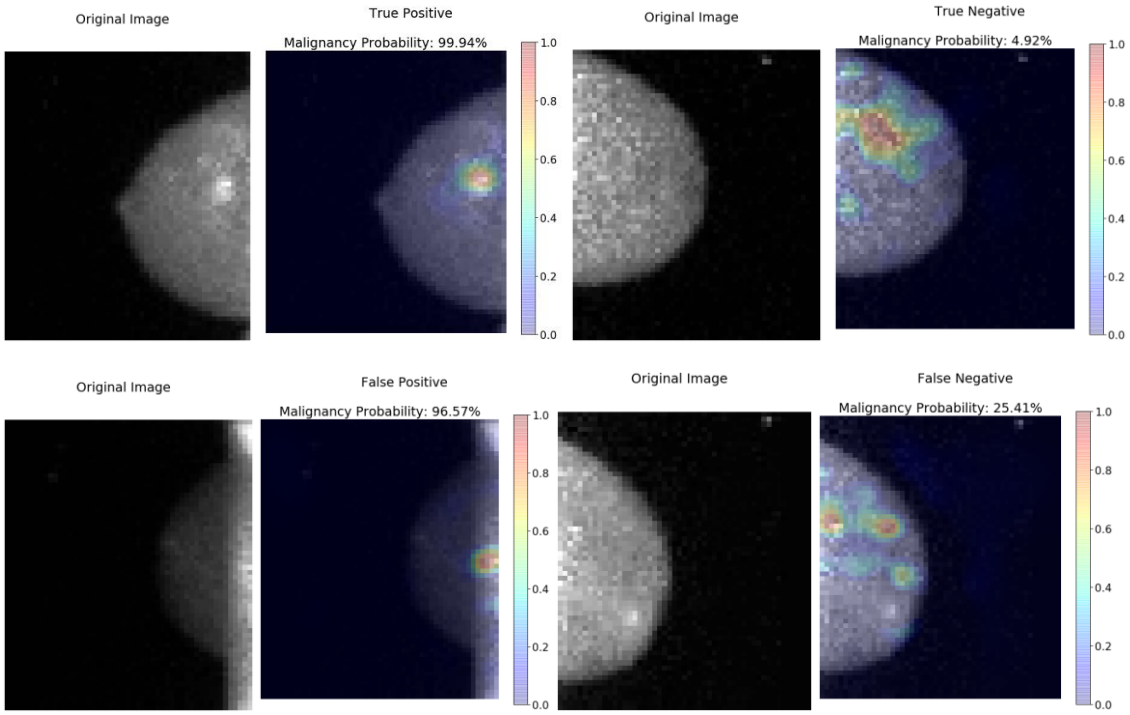


Figure 7 Examples of True Positives(Top left), True Negatives(top right), False Positives (Bottom Left) and False Negatives(Bottom right) including corresponding heatmaps. These are examples of classifications from Network 2.

3.4 Probability Distributions

As each input image results in a prediction by the model between 0 and 1 (or 0% to 100%) for the chance of a malignancy being present in the image, it is possible to evaluate if the model is more 'certain', or outputs higher chances of a presence of a malignancy, in images that contain actual malignancies compared to images that do not contain a malignancy.

With a cut-off probability of 0.5, meaning that a malignancy prediction of 0.5 or higher leads to a malignancy classification, while probabilities below 0.5 result in non-suspicious classification, the averaged model resulted in a sensitivity and specificity of 78% and 93%, respectively. In figure 8 below, the malignancy probabilities can be seen for each category: true negatives(TN), false negatives(FN), false positives(FP) and true positives(TP). The TN had a mean malignancy prediction of 0.17 ± 0.09, the FN had a mean of 0.29 ± 0.11 and the FP and TP resulted in mean malignancy predictions of 0.69 ± 0.14 and 0.91 ± 0.13, respectively. Distributions can be seen in figure 8.

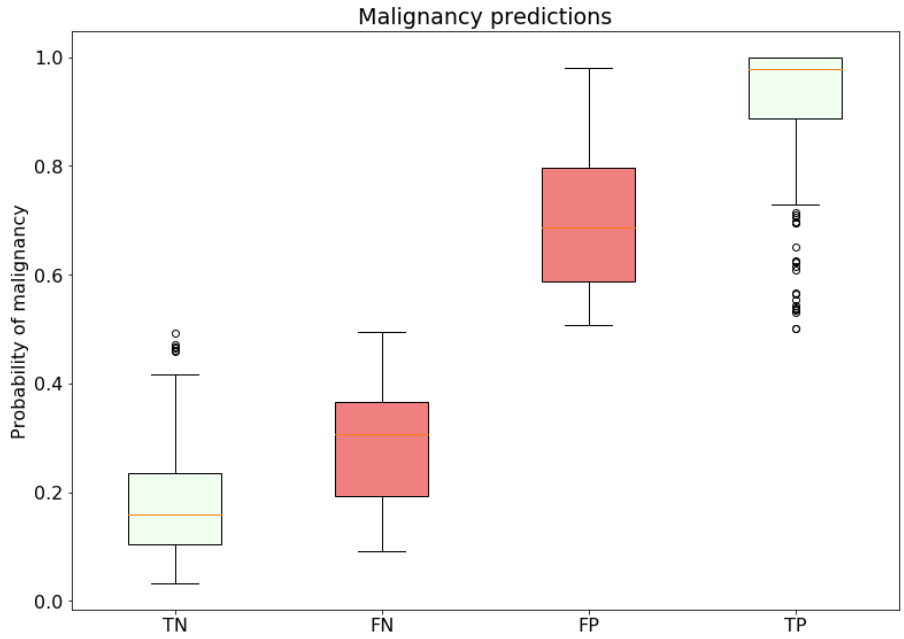


Figure 8 Probability distributions of True Negatives(TN), False Negatives(FN), False Positives(FP) and True Positives(TP) of the averaged model trained and tested on augmented data.

3.5 Misclassifications

3.5.1 False Negatives

Network 2, trained and tested on augmented data, had a better performance (AUROC = 0.93) and the false negatives resulting from this algorithm were also analyzed and categorized. In the case of false negative classifications, there could have been several reasons for the network missing a malignancy in the breast image. These reasons include: 1) No apparent lesion visible in this projection; 2) lesion is present, but too small or insignificant for a malignancy probability of 0.5 or higher; 3) lesion is partially outside of FOV, and therefore not significant enough to provide a malignancy probability of 0.5 or higher and 4) lesion location is close or adjacent to an intense thoracic wall. The amount of false negatives and the corresponding mean malignancy probabilities for each category is shown in table 7.

Table 7 Categories of false negatives in the test set of Network 2, trained and tested on augmented data. Category 1: No apparent lesion, Category 2: Small or insignificant lesion, Category 3: Outside of FOV, Category 4: Close to thoracic wall

FALSE NEGATIVES
NETWORK 2 – AUGMENTED TRAINING AND TEST DATA

CATEGORY	Amount	Mean Malignancy probability(%)
1	20	13.2 ± 8.6
2	25	33.77 ± 19.41
3	4	8.75 ± 0.71
4	6	16.30 ± 14.29
TOTAL	55	22.82 ± 18.15

Examples of each category of false negatives including their corresponding heatmaps are shown in figure 9.

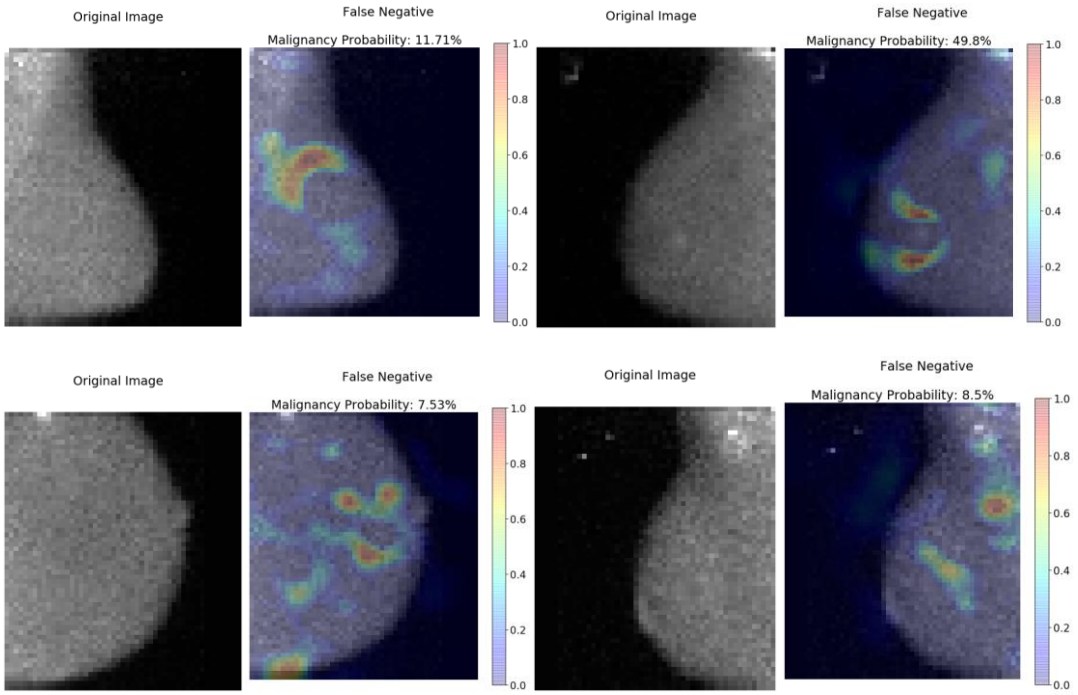


Figure 9 Examples of False Negatives of Network 2; Top Left: category 1, Top Right: category 2, Bottom Left: category 3, Bottom right: category 4

3.5.2 False Positives

False positives resulting from network 2 were also categorized based on the previously mentioned categories: 1) artefacts ; 2) thoracic wall identified as lesion ; 3) Diffuse uptake in breast and 4) intense uptake in the nipple. In table 8, the amount of false positives and the mean malignancy probability of each category is summarized. In figure 10, examples of each category of false positives are shown.

Table 8 Categories of false positives in the test set of Network 2, trained and tested on augmented data. . Category 1: Artefacts, Category 2:Thoracic wall, Category 3: Diffuse uptake, Category 4: Intense nipple

FALSE POSITIVES
NETWORK 2 – AUGMENTED TRAINING AND TEST DATA

CATEGORY	Amount	Mean Malignancy probability(%)
1	1	96.57 ± 0.0
2	5	89.57 ± 8.07
3	11	74.72 ± 17.11
4	2	87.34 ± 12.34
TOTAL	19	81.11 ± 16.18

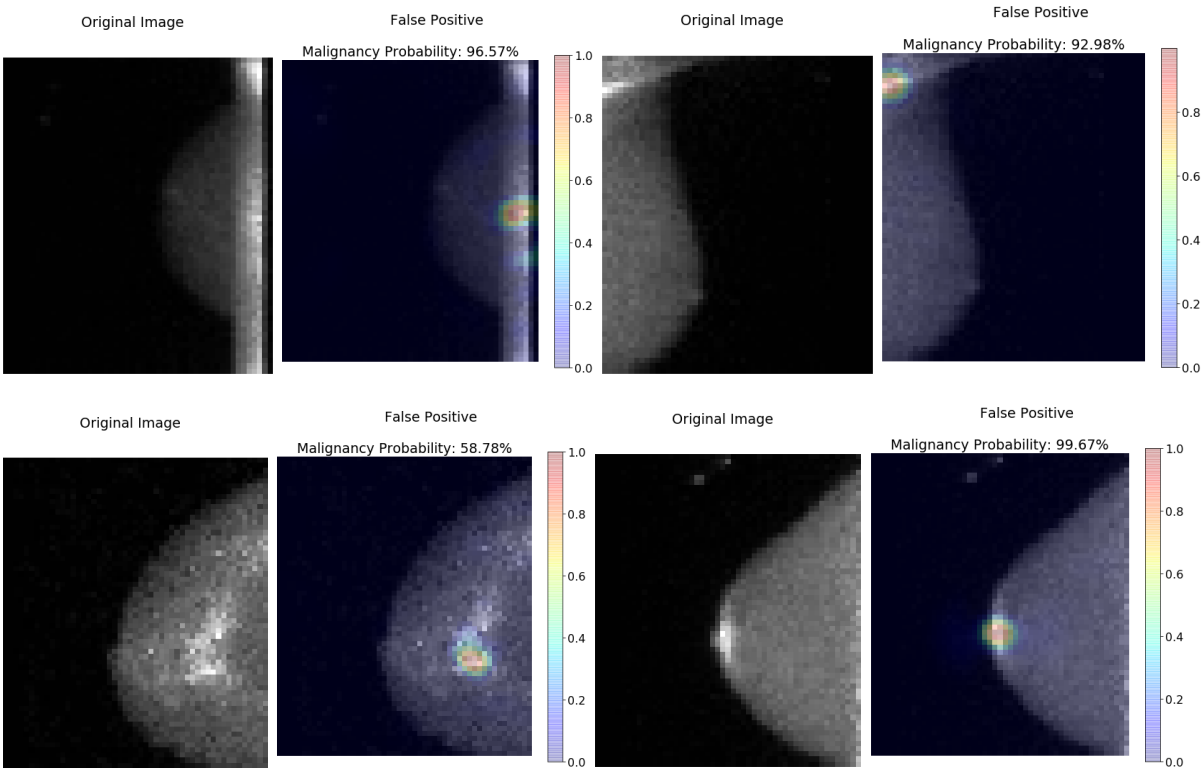


Figure 10 Examples of False Positives of Network 2; Top Left: category 1, Top Right: category 2, Bottom Left: category 3, Bottom right: category 4

3.5 Malignant versus Benign lesions

The same trained models and weighed model were used to evaluate their performance in discriminating between malignant and benign lesions, rather than detecting malignant lesions versus non suspicious BSGI images. This means that lesions that were classified as suspicious for malignancy by a nuclear medicine physician, but later were found to be benign through pathological assessment, were used instead of non-suspicious breasts.

Figure 13 shows all ROC curves for the three networks and the averaged model trained and tested on both original data and augmented data.

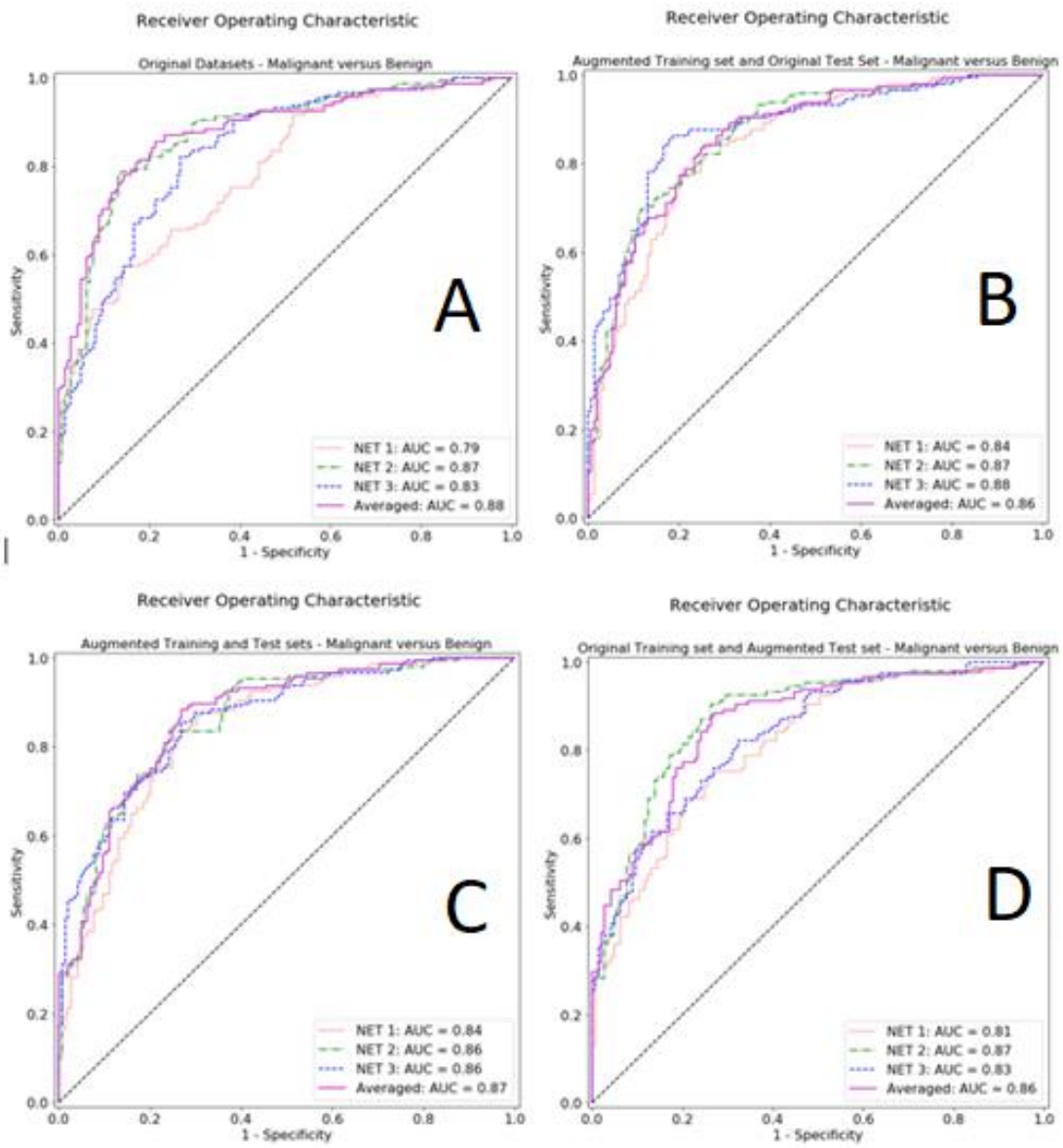


Figure 11 Malignant versus benign - ROC curves of A) Networks trained and tested on original datasets; B) Networks trained on augmented datasets and tested on original datasets; C) Networks trained and tested on augmented datasets and; D) Networks trained on original datasets and tested on augmented datasets.

All accuracies, sensitivities and specificities at a cut-off probability of 0.5 for all networks and data setup combinations can be viewed in table 9.

Table 9 Performances of all networks including the averaged model in discriminating malignant from benign lesions, using different combinations of augmented and non-augmented datasets. In each table, columns represent the results for a given combination of augmented or non-augmented training and test sets. Sensitivity and specificity values were calculated at a cut-off value of 0.5 (or 50%).

	Network 1 Malignant versus Benign			
Training set	non-augmented	Augmented	non-augmented	Augmented
Test set	non-augmented	Augmented	Augmented	Non-augmented
AUC	0.79	0.83	0.81	0.84
Accuracy	69%	73%	73%	75%
Sensitivity	79%	77%	79%	82%
Specificity	60%	70%	68%	68%
TP	114	111	114	119
FP	58	43	46	46
TN	87	102	99	99
FN	31	34	31	26

	Network 2 Malignant versus Benign			
Training set	non-augmented	Augmented	non-augmented	Augmented
Test set	non-augmented	Augmented	Augmented	Non-augmented
AUC	0.87	0.86	0.87	0.87
Accuracy	81%	77%	80%	78%
Sensitivity	87%	77%	86%	78%
Specificity	76%	77%	73%	77%
TP	126	112	125	113
FP	35	33	39	33
TN	110	112	106	112
FN	19	33	20	32

	Network 3 Malignant versus Benign			
Training set	non-augmented	Augmented	non-augmented	Augmented
Test set	non-augmented	Augmented	Augmented	Non-augmented
AUC	0.83	0.85	0.83	0.88
Accuracy	76%	76%	74%	82%
Sensitivity	74%	75%	68%	83%
Specificity	77%	76%	80%	81%
TP	107	109	99	121
FP	33	35	29	28
TN	112	110	116	117
FN	38	36	46	24

	Averaged Model Malignant versus Benign			
Training set	non-augmented	Augmented	non-augmented	Augmented
Test set	non-augmented	Augmented	Augmented	Non-augmented
AUC	0.88	0.87	0.86	0.86
Accuracy	80%	78%	79%	78%
Sensitivity	86%	79%	81%	78%
Specificity	74%	78%	78%	79%
TP	125	114	117	113
FP	38	32	32	31
TN	107	113	113	114
FN	20	31	28	32

3.5.1 Occlusion Mapping

Figure 12 show examples of TP, TN, FP and FN predictions by Network 3 trained on augmented data and tested on original data. (AUROC = 0.93, accuracy = 82%, sensitivity = 83% and specificity = 81% at a cut-off probability of 0.5).

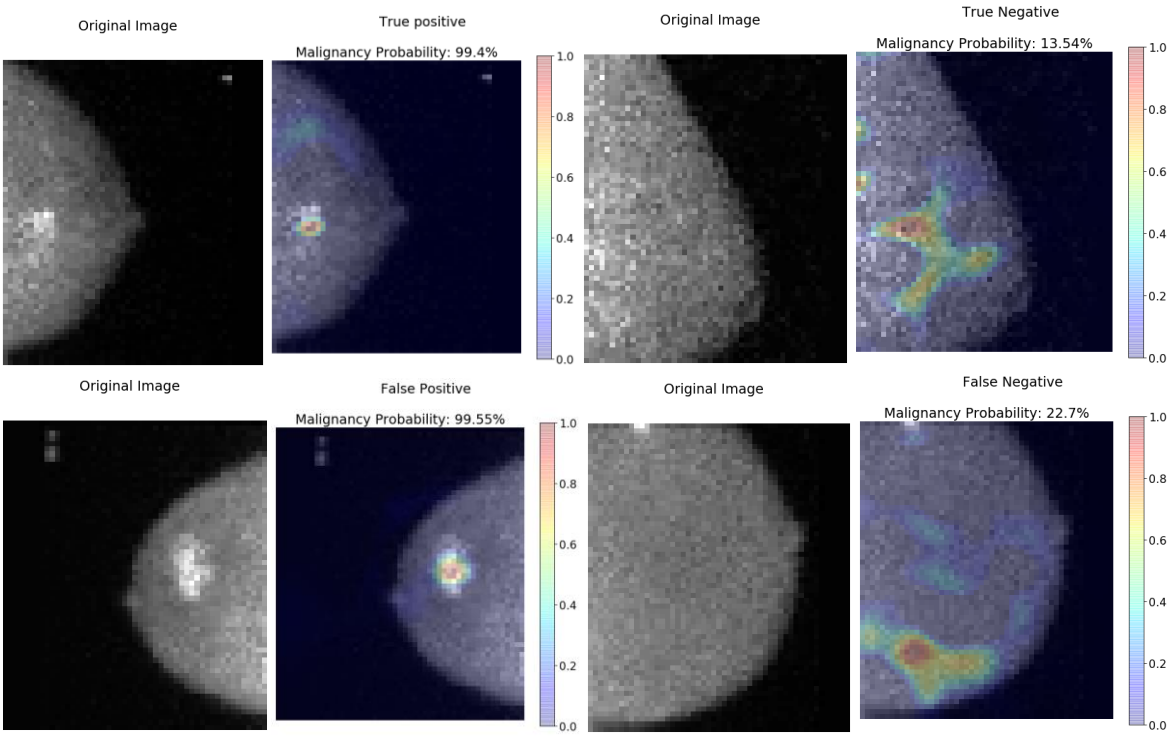


Figure 12 Examples of True Positives(Top left), True Negatives(top right), False Positives (Bottom Left) and False Negatives(Bottom right) including corresponding heatmaps. These are examples of classifications from Network 3

3.5.2 Probability Distributions

Similarly to the malignancy detection results, a boxplot (figure 15) was generated to visualize the distribution of malignancy predictions for each outcome: TN, FN, FP and TP. The cut-off probability value in this case was 0.5.

For TN, mean predicted probability of a present malignancy was 0.21 ± 0.11. For FN and FP, the mean predictions of a malignancy were 0.30 ± 0.11 and 0.77 ± 0.17, respectively. Finally, the mean malignancy prediction for actual malignancies (TP) was 0.93 ± 0.10. These distributions are shown in figure 13.

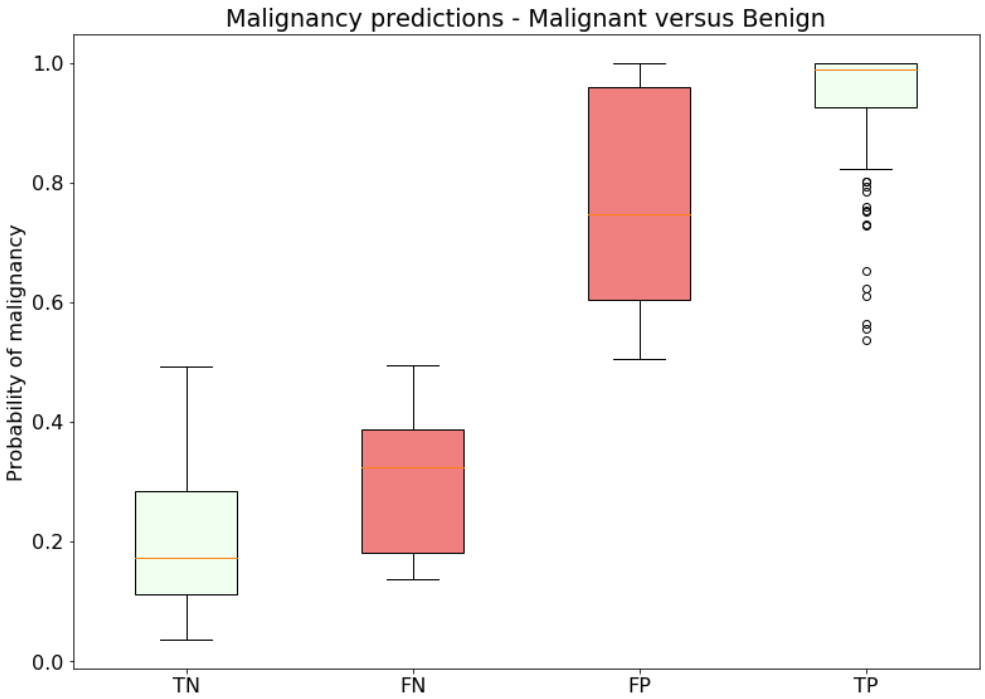


Figure 13 Probability distributions of True Negatives(TN), False Negatives(FN), False Positives(FP) and True Positives(TP) of Network 3 trained on augmented data and tested on a malignant versus benign augmented dataset.

3.5.3 Misclassifications – Malignant Versus Benign

False Negatives

The best performing algorithm on discriminating malignant from benign lesions is Network 3, trained on augmented data and tested on orignal data. Using this algorithm on the test set resulted in a total of 24 false negatives, where a malignancy was missed and the malignancy probability fell below the 0.5 probability threshold. These 24 false negatives were categorized using the following four categories: 1) No lesion visible in this projection; 2) Diffuse uptake pattern not fitting with malignancy; 3) Lesion present, but not focal or intense enough for malignancy prediction and 4) malignancy too close to thoracic wall for identification. The number of false negatives and their corresponding malignancy predictions are shown in table 10, and examples are shown in figure 14.

Table 10 Categories of false negatives (malignant versus benign) in the test set of Network 3, trained on augmented data and tested on original data. Category 1: No apparent lesion, Category 2: Diffuse uptake, Category 3:Lesion insignificant, Category 4: Thoracic wall

FALSE NEGATIVES
NETWORK 3 - MALIGNANT VERSUS BENIGN

CATEGORY	Amount	Mean Malignancy probability(%)
1	5	26.10 ± 6.08
2	6	24.22 ± 8.72
3	8	27.96 ± 8.18
4	5	33.14 ± 10.79
TOTAL	24	0.30 ± 0.11

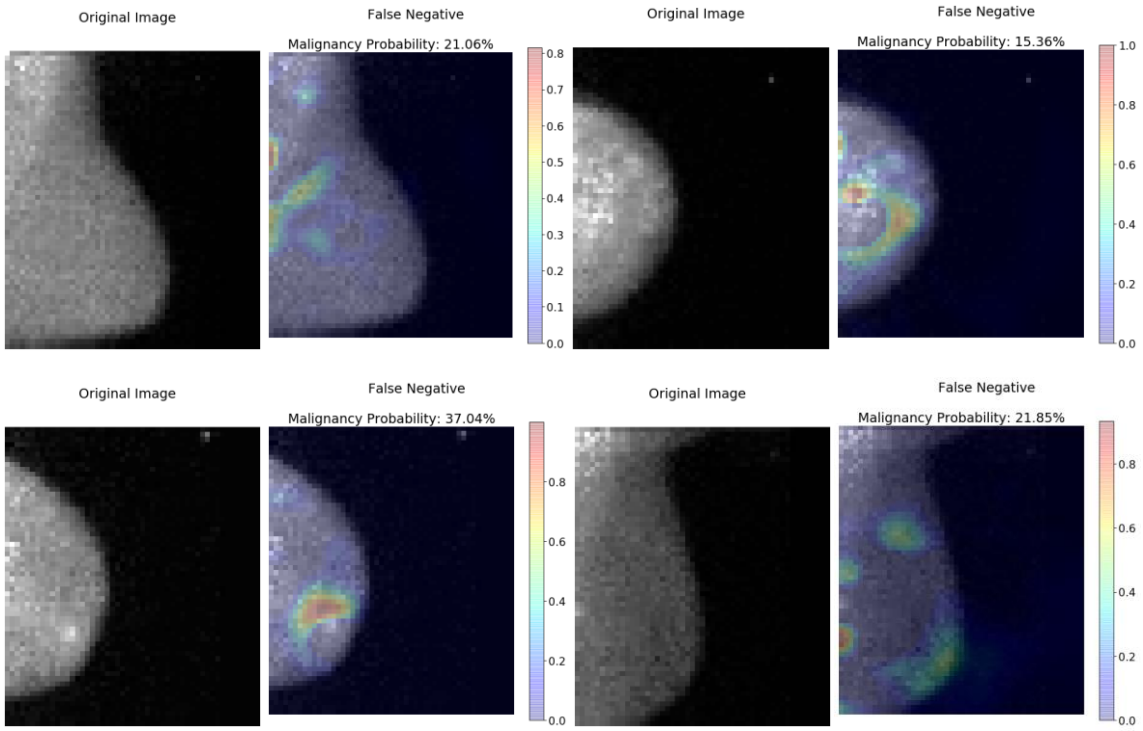


Figure 14 Examples of False Negatives of Network 3; Top Left: category 1, Top Right: category 2, Bottom Left: category 3, Bottom right: category 4

False Positives

This same algorithm led to a total of 28 false positives. Again, these false positives were separated into four categories: 1) Uptake pattern similar to malignancy; 2) Thoracic wall identified as malignancy; 3) Unclear decision basis and 4) Intense uptake in the nipple.

In table 11, the amount of images in each category of false positives is shown, as well as the mean malignancy probability per category. Figure 15 shows examples of each category.

Table 11 Categories of false positives (malignant versus benign) in the test set of Network 3, trained on augmented data and tested on original data. Category 1: Uptake pattern, Category 2: Thoracic wall, Category 3: Unclear, Category 4: Intense nipple

FALSE POSITIVES
NETWORK 3 - MALIGNANT VERSUS BENIGN

CATEGORY	Amount	Mean Malignancy probability(%)
1	6	91.88 ± 6.05
2	13	67.35 ± 8.11
3	5	74.07 ± 10.84
4	4	92.08 ± 4.54
TOTAL	28	0.77 ± 0.17

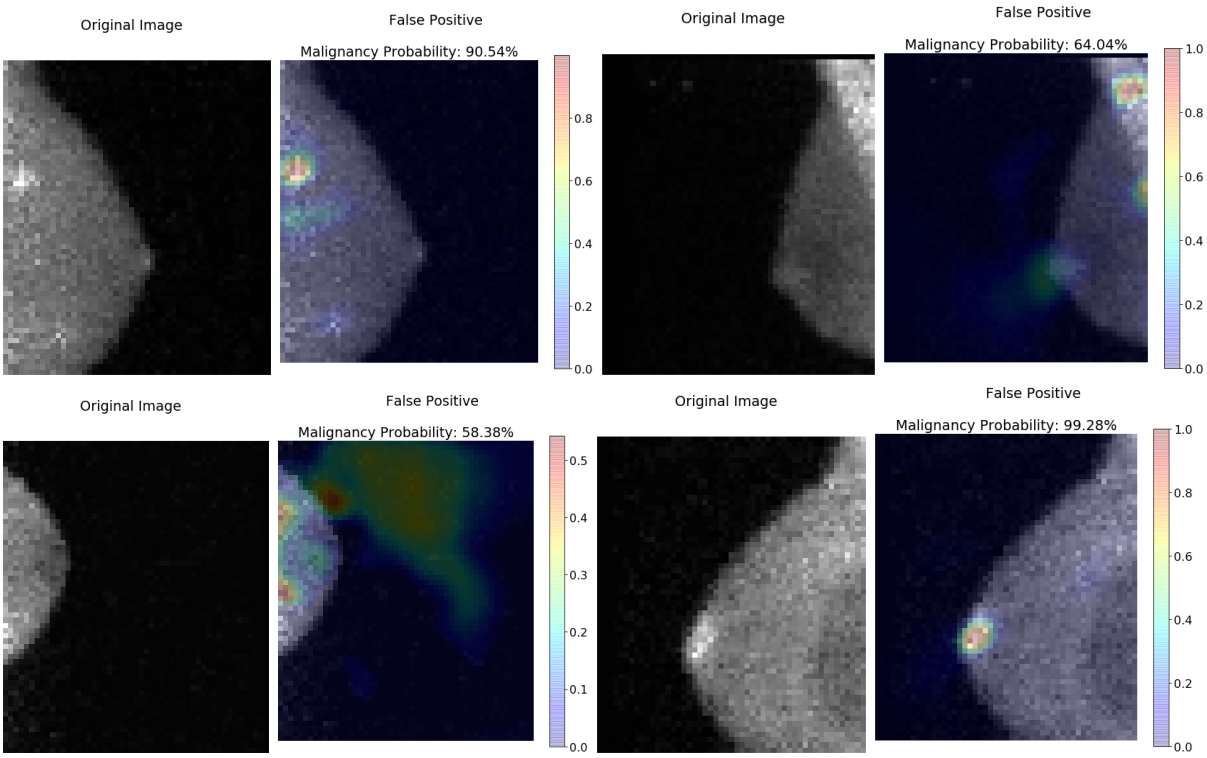


Figure 15 Examples of False Positives of Network 3; Top Left: category 1, Top Right: category 2, Bottom Left: category 3, Bottom right: category 4

3.6 Radiomics & Machine learning: receptor status prediction

3.6.1 Segmentation

Segmentation was performed on BSGI images consisting of 96 x 96 pixels acquired using the *GE Discovery NM 750b* BSGI system. Pixel size for these images was 2.46mm by 2.46mm.

A total of 145 RMLO or LMLO images with focal uptake lesions were used as input for the segmentation algorithm. In all instances, the lesion location was correctly identified as a possible lesion location based on adaptive thresholding. In 116 instances, two or more lesion locations were identified other than the actual lesion location.

With sphericity calculation of the pixel groups identified by means of adaptive thresholding, lesions were correctly identified in 102 out of 116 images with multiple possible lesion locations. In 7 cases, sphericity of the actual lesion was more than 20% higher than the other possible lesion location, thus correctly selecting the lesion based on sphericity.

In 7 cases, sphericity of possible lesions identified by the algorithm fell within the 20% discriminative margin, in which case the smallest Euclidean distance to the center of the whole image gave the decisive lesion location. In 6 of these 7 cases, use of the Euclidean distance correctly identified the lesion. In 1 lesion, the lesion location had to be manually selected, as the Euclidean distance of the other possible lesion location to the image center was closer than the actual lesion location.

From these segmentations, the custom features were calculated, which are summarized in table 12. The P-value is calculated using a two-tailed t-test.

Table 12 Custom calculated lesion parameters and lesion size for each lesion receptor status

	ER			PR			HER2-NEU		
	+ N = 73	- N = 72		+ n = 74	- N = 71		+ N=45	- N=100	
AREA(CM^2)	1.79 ±0.9	2.22 ± 1.1	P = 0.006	1.78 ± 1.0	2.20 ± 1.1	P=0.006	1.91 ± 0.89	1.72 ± 1.19	P = 0.58
TBR_{MEAN}	1.65 ± 0.28	1.67 ± 0.26	P = 0.63	1.64 ± 0.27	1.65 ± 0.23	P= 0.68	1.69 ± 0.27	1.62 ± 0.21	P = 0.26
TBR_{HIGH}	1.98 ± 0.44	2.11 ± 0.46	P = 0.19	2.00 ± 0.44	2.13 ± 0.47	P = 0.19	2.13 ± 0.48	1.94 ± 0.37	P = 0.11
TBR_{MAX}	2.29 ± 0.56	2.54 ± 0.6	P = 0.054	2.31 ± 0.54	2.55 ± 0.07	P = 0.053	2.55 ± 0.63	2.27 ± 0.57	P = 0.10
COV	0.16 ± 0.06	0.2 ± 0.16	P =0.06	0.16 ± 0.05	0.19 ± 0.15	P = 0.07	0.20 ± 0.07	0.16 ± 0.07	P = 0.06

For classification of the Estrogen, Progesterone and HER2-neu receptor status, 15 features were selected used univariate feature selections. The models were also trained with a selection of 25 features. These selected features are shown in appendix 2A.

3.6.2 Estrogen Receptor

The training set consisted of 52 ER-positive lesions and 50 ER-negative lesions, while the external test set consisted of 21 ER-positive lesions and 22 ER-negative lesions. Accuracies in predictions of the test set of all different combinations of ML classifiers varied between 51% and 72% when using 15 radiomic features. With 25 features, these accuracies ranged from 53% to 74%. The best overall performance, with an accuracy of 74%, was achieved with 25 radiomics features, using the LDA classifier at 3 k-folds. This resulted in a sensitivity of 81.0% and a specificity of 63.3%. The confusion matrix for these results is shown in figure 16.

All results from other classifiers using 15 and 25 features are shown in appendix 3A.

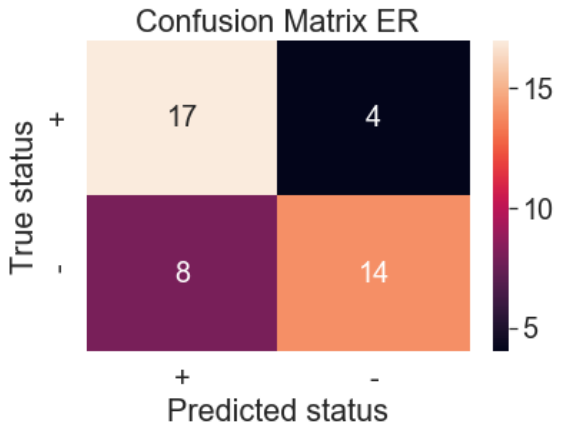


Figure 16 Confusion matrix for the classification of ER status using the LDA classifier

3.6.3 Progesterone Receptor

With the 15 features selected using univariate selection (table 18), the same models were trained and tested, now labelled with either PR-positivity or PR-negativity. The training set consisted of 51 PR-positive lesions and 51 PR-negative lesions. The test set consisted of 23 PR-positive lesions and 20

PR-negative lesions. Accuracies ranged from 51% to 74% with 15 features, and from 53% to 72% with 25 radiomics features.

Similarly to the classification of ER, the best overall performance was found in the LDA classifier, using 15 radiomics features. This classifier resulted in an accuracy of 74%, with a sensitivity and specificity of 78.3% and 59.1%, respectively. The confusion matrix can be seen in figure 17.

All results from other classifiers using 15 and 25 features are shown in appendix 3B.

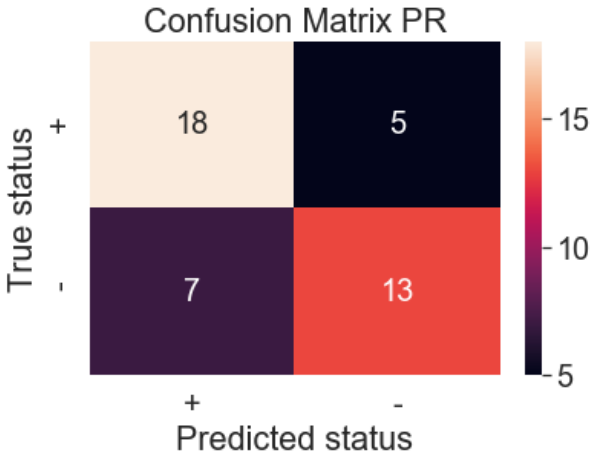


Figure 17 Confusion matrix for the classification of PR status using the LDA classifier

3.6.4 HER2-Neu Receptor

The HER2-Neu receptor data-set was in imbalanced dataset with in total 100 HER2-neu negative lesions and 45 HER2-neu positive lesions. After splitting the dataset into a training and test set, the training set consisted of 31 HER2-neu positive lesions and 71 HER2-neu negative lesions. The test set consisted of 14 HER2-neu positive lesions and 29 HER2-neu negative lesions. The selected features are shown in table 18 (appendix 2).

Prediction of HER2-neu receptor using the given machine learning models resulted test set accuracies of 51% to 69% with 15 and 51% to 67% with 25 features. The best overall performance was achieved using the RF classifier and 15 radiomics features. With an accuracy of 69%, a sensitivity of 60% and a specificity of 75%, the RF classifier slightly outperformed other classifiers.

All results from other classifiers using 15 and 25 features are shown in appendix 3C.

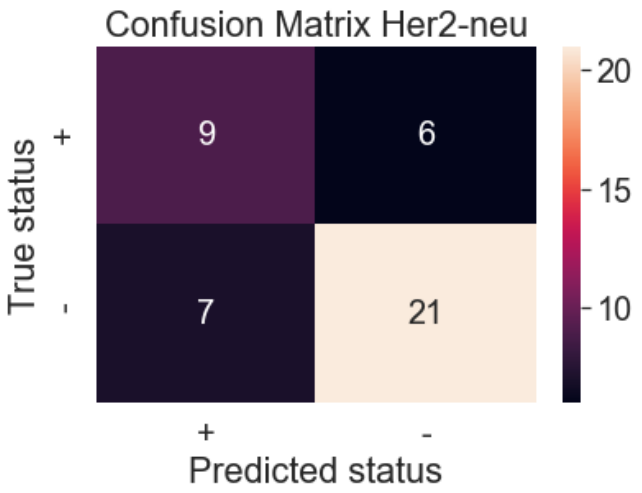


Figure 18 Confusion matrix for the classification of ER status using the RF classifier

4. DISCUSSION

4.1 Deep Learning

4.1.1 Lesion Detection

CNNs used on mammographic images has been widely investigated and proven to be of additional diagnostic value in detection and classification of breast lesions[42]. CNNs developed for breast cancer detection in mammograms have become more complex over time, with the one of the earliest CNNs presented the *CifarNet*[43] with three convolutional layers, up to the more widely used and better known *VGGNet*[44], which showed the effects of increasingly deep neural networks with up to 19 convolutional layers.

This, however, was to the authors knowledge the first study to investigate the possible merits of using AI, and specifically deep learning, on BSGI, meaning there is no benchmark to which the results presented in this paper can be objectively compared.

Several studies have previously successfully implemented deep-learning algorithms on scintigraphic images, such as in bone scintigraphy[45,46,47], thyroid scintigraphy[40], and myocardial perfusion scintigraphy[48]. As a result of the relative simple morphological image features resulting from scintigraphic images, when compared to traditional anatomic imaging modalities such as CT or MRI, scintigraphic images are regarded to be more amendable to AI applications, such as deep learning. In applications of AI in scintigraphic images described in the studies mentioned above, neural networks consisted of three to five convolutional layers, varying in convolutional kernel size, max-pooling filter size, and the order of convolutions and pooling operators. These studies have shown AUCs varying from 0.85 to 0.96 in bone scinigraphy[43,44,45], 0.85 to 0.90 in thyroid scintigraphy and 0.89 to 0.95 in myocardial perfusion scintigraphy[47], similar to the best performing network presented in this paper.

Based on the results of this study, a combination of BSGI and AI, especially CNNs, can accurately detect the presence of malignancies in patients suspected to suffer from breast cancer with AUCs comparable to those of previously described CNN applications in mammographic images and scintigraphic images.

Our models used a balanced dataset consisting of 1,432 BSGI images, which is also similar in size to the other studies describing the use of AI in scintigraphy, but smaller than the datasets used in most studies on AI in mammography. As a result of the relatively simple nature of these functional images, datasets are generally allowed to be smaller than when using functional imaging modalities, as there is more variation in functional images, thus requiring more examples in a dataset for the CNNs to adapt to patterns. Increasing our dataset with data augmentation generally had a positive impact on the network performances and generalizability.

Comparing the best and worst performing networks, network 1 and 2 leads to the insight that the differences in performance are attributable to decreases in both the number of false positives and false negatives, which were 55 and 57, respectively, for network 1, and 39 and 39, respectively, for network 2. Network 3, however, resulted in a large increase in the number of false negatives (63), while having a strong decrease in false positives (17).

This means that with a number of convolutional layers higher than two, the network is better equipped to detect ‘clean’ BSGI images, but has an increased chance of classifying a malignancy as a non-suspicious finding, therefore missing out on malignancies. Apparently, if the number of convolutional layers becomes too high, the algorithm also extracts features that are not influenced by the presence of a possible lesion, thereby reducing the impact of features of actual lesions, such as edges, shapes and combinations of these features.

Across all networks, reported sensitivity was lower than the specificity. Network 2 resulted in a sensitivity and specificity of 78% and 92%, respectively. These performances are similar to the sensitivity and specificity in general, without the use of AI, in the detection of malignancies. In 2016, *Yu et al.,* reported a sensitivity of 80% and a specificity of 83% in detecting malignancies on BSGI images[49]. A direct comparison between the results of this deep learning algorithm and the results in previous studies, however, is difficult as it is unclear how many patients who received a BSGI BI-RADS score <4 by a nuclear medicine physician, and thus deemed not suspicious, in this paper were false negatives.

The occlusion maps confirm that the networks make their classification based on lesions, and not on other parts of the image. Lack of explainability of AI algorithms in healthcare settings is often raised as an obstacle for actual implementation in a clinical setting, and has raised medical, legal, ethical and societal questions.[50] Providing occlusion heatmaps is one method of giving transparency into the decision making process of neural networks, and could potentially increase trust in these algorithms' decision making and increase willingness for implementation of more AI applications in healthcare. Occlusion maps also expose some of the weaknesses of the networks, with missed lesions partially outside the FOV, intense uptake in the nipple classified as a lesion, and in some cases the thoracic wall identified as a lesion. This again stresses the need for explainability of AI models used in healthcare; not only is it needed to verify correct classifications, but also to easily identify misclassifications in case they happen[51].

4.1.2 Malignant versus benign

The best performing network in specifically distinguishing malignant from benign lesions was the averaged model, with an accuracy, sensitivity and specificity of 80%, 86% and 74%, respectively. In this case, the specificity is lower than the sensitivity, instead of higher. As the benign lesions in the test data set were previously identified by nuclear medicine physicians as suspicious, it is to be expected that the algorithm trained to detect malignancies also classifies a number of these benign lesions as malignant, thus increasing the FP rate, and lowering the specificity.

As slight differences in tracer uptake distribution, intensity and uptake locations could mean a difference in a BSGI BI-RADS 3 or 4 classification by a nuclear medicine physician, these differences could potentially also be utilized by deep learning based algorithms to discriminate benign from malignant lesions. Improvements to a benign vs. malignant classification algorithm could potentially be made with a larger dataset of histopathologically confirmed benign lesions. With the relatively small number of images containing confirmed benign lesions in this dataset (n = 145), however, training the networks on benign images would not have been feasible, even when data augmentation would have been applied. Also, because differences between malignant and benign lesions on BSGI images are sometimes difficult to identify, new deep learning networks trained for this specific purpose might have to be enlarged in terms of increasing the number of convolutional layers and pooling layers to convolve images into smaller and more specific image features. This might also explain why the best performance of a standalone network (excluding the averaged model) was achieved in network 3, the deepest network with the most convolutional layers.

Despite promising results about the CNNs abilities to detect malignancies in BSGI images, several limitations exist. The main limitation is the per-image basis on which the model evaluation was performed. As BSGI acquisition is performed in different perspectives, serval images are made of each breast. A nuclear medicine physician would take all images in a certain acquisition into consideration when appointing a BI-RADS score to a breast. In BSGI image series of a breast, if a

lesion is less visible in one projection, the nuclear medicine physician would still consider the breast suspicious for the presence of a malignancy, whereas the AI in this case would result in one wrong classification (false negative), while correctly predicting the other projections. Even though this effect of using a per-image evaluation instead of a per-patient evaluation potentially affects representation of the results, it has to be noted that this effect could also work inversely, where only one projection is predicted correctly, while the other projections are classified wrongly. This latter effect, however, is expected to be smaller than the first effect, given the overall good performance in detecting malignancies. Another limitation is that there is no further clinical information on which the AI is trained and tested, but only relies on information provided through the BSGI images. This also is different from clinical practice, in which nuclear medicine physicians take more information into account when interpreting BSGI images, such as results from other diagnostic tools, patient characteristics and medical history.

A future feasibility study will be conducted on a per-patient basis, and evaluate the correlation between predicted outcomes of network 2 and BSGI BI-RADS scores, as well as evaluate the time saving capabilities of using AI.

4.2 Lesion Classification

The secondary aim of this study to apply ML algorithms on radiomics features, which were extracted on a semi-automatically segmented lesions in BSGI images, to predict ER, PR and Her2-neu status.

In the recent years, several studies have attempted to use a combination of radiomics and ML for predicting hormonal and Her2-neu status prediction in breast cancer using different imaging modalities. *Ma et al.* published a paper in 2018 attempting to use mammographic images for the classification of breast tumors using 331 subjects, reaching accuracies of 80% in the prediction of ER/PR status, and 75% in the prediction of Her2-neu status, using the Naive Bayes ML classifier[52]. In 2019, *Xie et al.* described the same goal using MRI instead of mammography. In 190 patients, classification accuracies using different ML classifiers (DT, LDA, SVM, KNN, ensemble) in the prediction in triple-negative versus non-triple negative tumors varied between 72% and 91%.

In contrast to anatomical imaging techniques, functional imaging such as BSGI have not been as extensively investigated for the efficacy of applying radiomics and ML in predicting subtypes of cancer. *Han et al.*, however, described in 2021 the use of PET-only radiomics (no CT) in predicting histological subtypes of non-small cell lung cancer (NSCLC) using a variety of ML techniques. The authors concluded that ML could help differentiate the histological subtypes of NDCLC, with accuracies of up to 79% (for both the LDA and SVM classifiers). To the best of our knowledge, no papers have been published on the combination of radiomics and ML in BSGI.

Several studies, however, performed semi-quantitative analysis of BSGI images and found correlations with the TBR and hormonal status and Her2-neu status. In accordance with those studies[29,30,53], we found that our calculated TBR_{max}, TBR_{high} and TBR_{mean} also was lower on average in ER and PR positive tumors than in ER and PR negative tumors. The same similarity was seen in Her2-neu status, where Her2-neu positivity was associated with a higher TBR.

The best accuracies in predicting ER, PR and Her2-neu status we found were 74%, 74% and 67%, respectively. These accuracies lower than accuracies found in the previously mentioned studies that used anatomical imaging on which radiomics and ML were applied. There could be several explanations for this relatively poor discriminative performance. First, many radiomic features that are extracted are based on texture analysis of the ROI. Texture features have been shown to be

increasingly sensitive with increasing spatial resolution[54], and with the relatively low spatial resolution of BSGI compared to for instance MRI and mammography (3.5 mm (GE Discovery NM750b)[55], 0.5 - 1mm (3T MRI)[56], 0.1mm (digital mammography[57])) , BSGI texture analysis is a more challenging task, as a low spatial resolution limits the number of discriminative image features. Also, the limited amount of pixels present in the segmented ROIs in BSGI images compared to MRI, CT and mammography result in relatively large alterations on extracted features if extra pixels, or less pixels are included in the ROI after segmentation. This is especially troublesome in lesions with less focality, leading to more uncertainty about the correctness of the segmentation. Future scintigraphic imaging modalities, however, could increase the potential for combining ML and radiomics, if spatial resolution and pixel sizes are reduced significantly.[58]

With ER and PR positivity found in approximately 87% and 71% cases of breast cancer, respectively, the prediction of hormonal receptor status using ML should be treated with caution[59]. For rare occurrences, such as ER and PR negative tumors, which are bad prognostic factors and should be ruled out, a test should generally have a high specificity. With specificities of 64% and 59%, however, ML seems to be insufficient for proper detection of ER and PR negative tumors. HER2-neu positivity occurs in approximately 9.9% of cases, and is a poor prognostic factor, which should thus be ruled out by a high sensitivity. The sensitivity of 57% does not meet the requirements for sufficiently ruling out the more risky HER2-neu positive lesions.

Especially since the major clinical implications of subtyping tumors, IHC or pathological assessment should remain the gold standard, as receptor status prediction based on radiomics and ML in BSGI result in inferior specificities and sensitivities.

Future research directions could focus on combining different imaging modalities, such as mammography and/or MRI, to non-invasively predict hormonal and Her2-neu status, as a combination of anatomical and functional imaging might contain sufficient data on a textural feature level, as well as information on metabolic activities, such as high TBR levels which are associated with poor prognostic factors such as ER/PR negativity and HER2-neu positivity.

5. Conclusion

In this paper, the aim was to create and validate CNNs designed with the purpose of accurately detecting abnormalities in BSGI images, as well as discriminate malignant from benign lesions. According to standards in literature[60], the detection of malignancies of these novel networks can be considered outstanding, and the specific task of discriminating malignancies from lesions that were deemed possibly malignant at first, but turned out to be benign lesion can be considered excellent.

In patients undergoing BSGI for problem-solving, AI could potentially aid in quickly identifying patients with a high chance of suffering from breast cancer, or identify those with negligible risk. Also, it could potentially be used as a decision-support tool for lesions that are one the border of receiving a BI-RADS 3 or 4 classification, by providing the nuclear medicine physician with an extra parameter: the AI prediction.

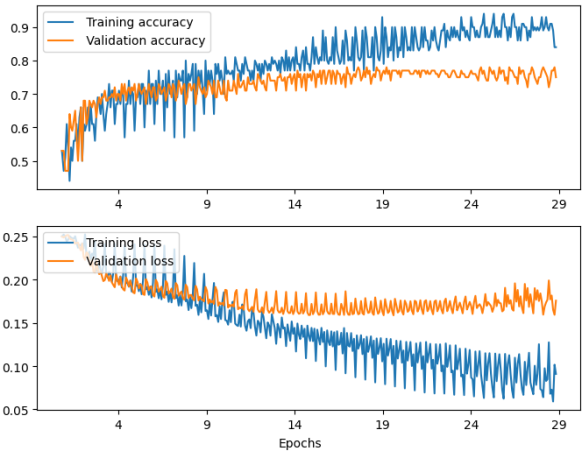
The combinations of radiomics and ML is currently not sufficiently accurate for reliable classification of breast tumors based on ER/PR and Her2-neu status. Future research should focus on expanding the scope of combining radiomics and ML in the classification of lesions. This could potentially be achieved by the addition of radiomics derived from mammographic and/or MRI images.

Appendices

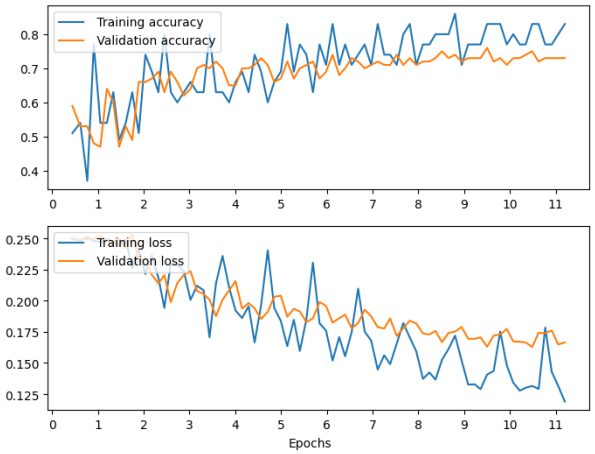
Appendix 1A

Loss-accuracy graphs of Network 1, Network 2, and Network 3 trained on original, non-augmented data. Left column: graphs with an excessive amount of epochs, used to determine the point of divergence for training and validation curves. Right column: final graphs, where models are trained until this previously determined point of divergence.

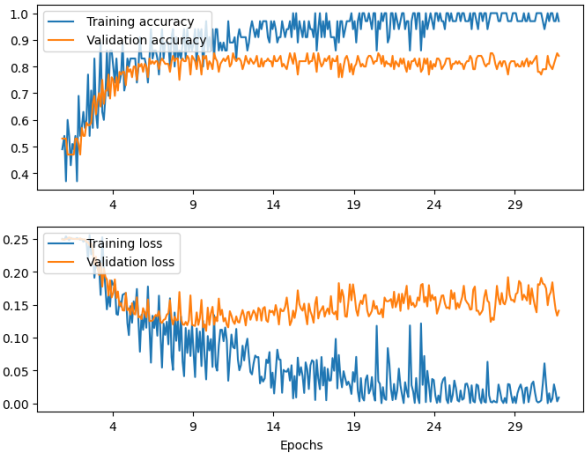
Network 1



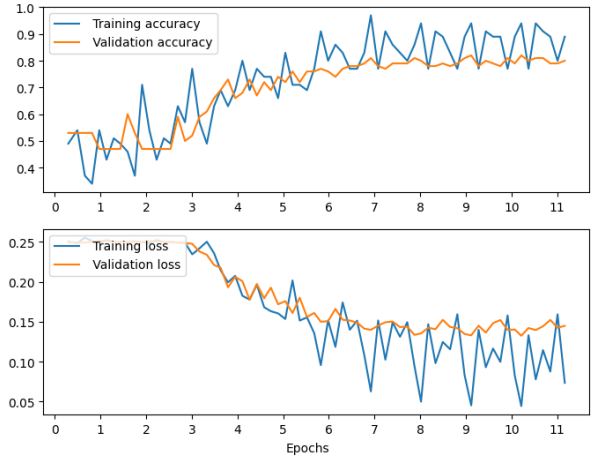
Network 1



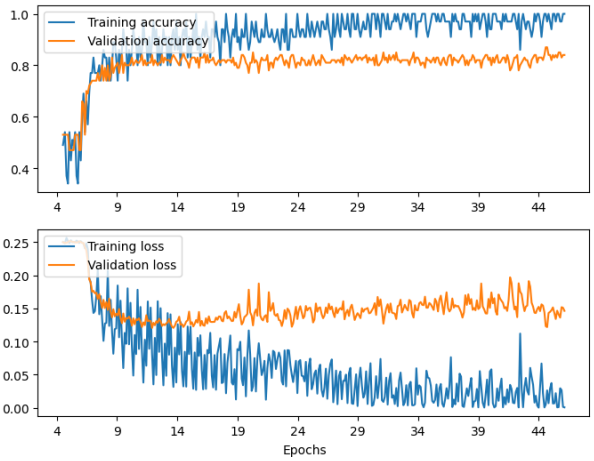
Network 2



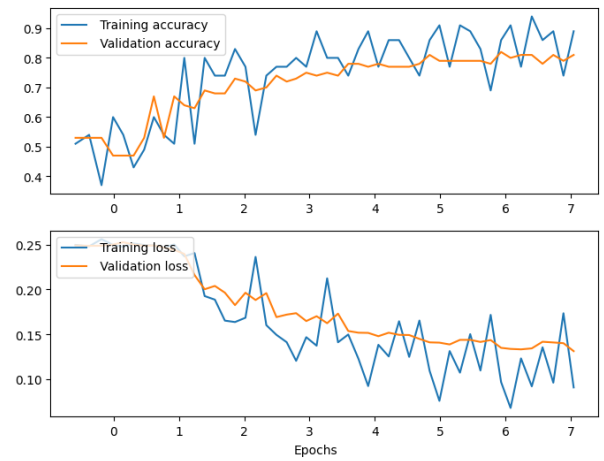
Network 2



Network 3

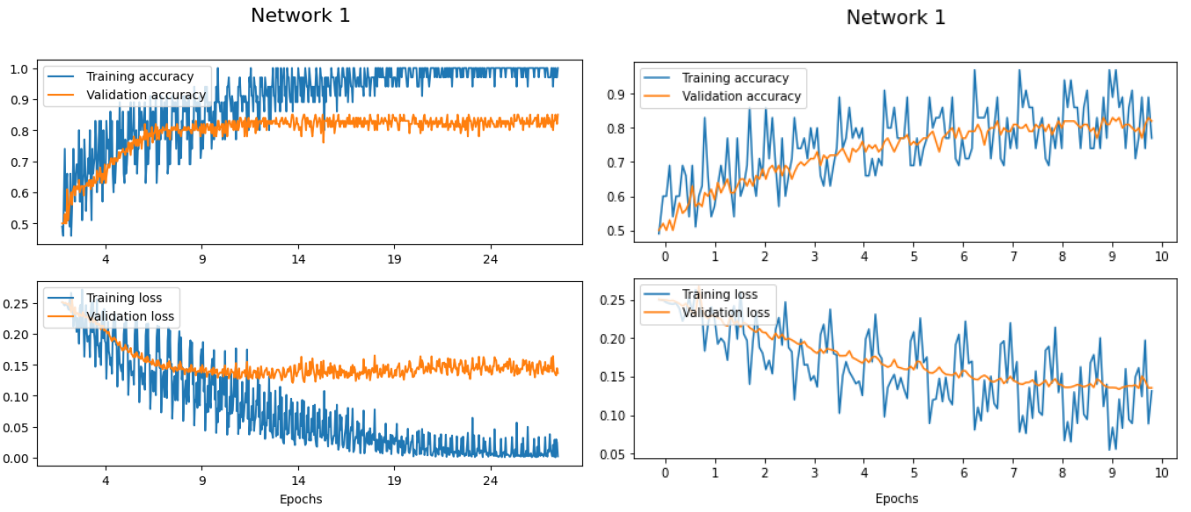


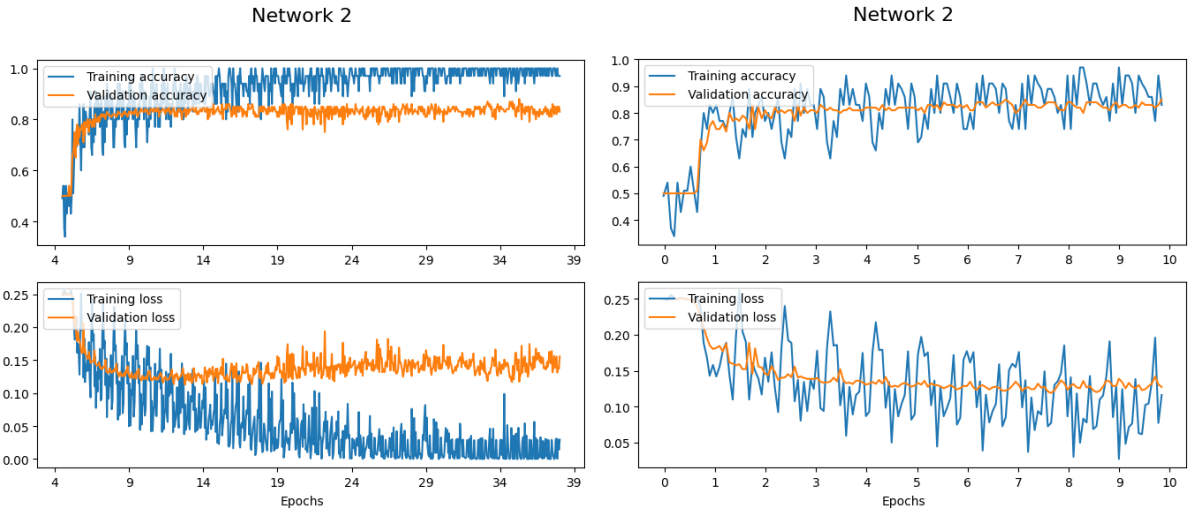
Network 3

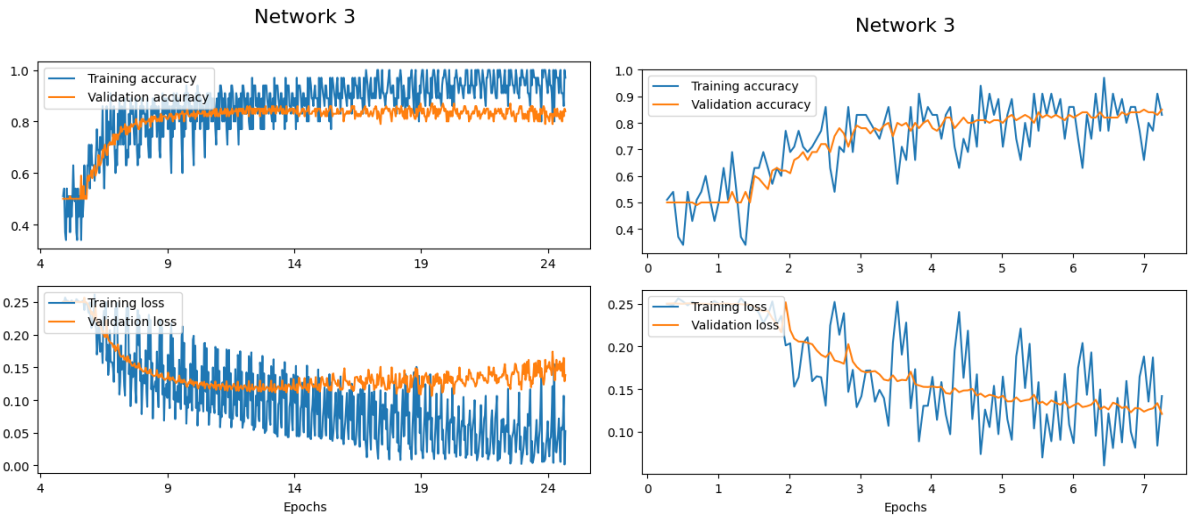


Appendix 1B

Loss-accuracy graphs of Network 1, Network 2, and Network 3 trained on augmented data. Left column: graphs with an excessive amount of epochs, used to determine the point of divergence for training and validation curves. Right column: final graphs, where models are trained until this previously determined point of divergence.







Appendix 2

Selected Radiomics features using univariate feature selection

#	FEATURE NAME		
	Estrogen Receptor Classification	Progesterone Receptor Classification	HER2-neu Classification
1	Maximum 2D Diameter Row	Maximum 2D Diameter Row	Maximum 2D Diameter Row
2	Maximum 2D Diameter Column	Maximum 2D Diameter Column	Minor Axis length
3	Minor Axis length	Surface Area	Surface Area
4	Major Axis Length	First Order Energy	First Order Energy
5	Surface Area	First Order Total Energy	First Order Skewness
6	First Order Energy	GLDM Dependence Non Uniformity	First Order Total Energy
7	First Order Total Energy	GLDM Gray level Non Uniformity	GLDM Dependence Non Uniformity
8	GLDM Dependence Non Uniformity	GLRLM Gray Level Non Uniformity	GLDM Dependence Variance
9	GLDM Dependence Variance	GLRLM Long Run Low Gray Level Emphasis	GLDM Gray level Non Uniformity
10	GLDM Gray level Non Uniformity	GLRLM Run Length Non Uniformity	GLDM Large Dependence Low Gray Level Emphasis
11	GLRLM Gray Level Non Uniformity	First Order Skewness	GLRLM Long Run Low Gray Level Emphasis
12	GLRLM Long Run Low Gray Level Emphasis	GLDM Dependence Entropy	GLRLM Run Variance
13	GLRLM Run Length Non Uniformity	GLDM Dependence Variance	GLSZM Large Area Low Gray Level Emphasis
14	GLRLM Run Variance	GLRLM Run Variance	TBR_{high}
15	GLSZM Large Area Low Gray Level Emphasis	GLSZM Large Area Low Gray Level Emphasis	TBR_{max}
16	First Order Minimum	Major Axis Length	Shape Elongation
17	First Order Range	First Order 10 Percentile	Maximum 2D Diameter Column
18	First Order Skewness	First Order 90 Percentile	First Order 10 Percentile
19	GLDM Dependence Entropy	First Order Mean	First Order Interquartile Range
20	GLDM Large Dependence Low Gray Level Emphasis	First Order Median	First Order Kurtosis
21	GLRLM Run Entropy	First Order Minimum	GLDM Dependence Entropy
22	GLRLM Short Run Emphasis	First Order Root Mean Squared	GLDM Dependence Non Uniformity Normalized
23	TBR_{high}	GLDM Large Dependence Low Gray Level Emphasis	GLRLM Gray Level Non Uniformity
24	TBR_{max}	TBR_{mean}	GLRLM Shor tRun Low Gray Level Emphasis
25	COV	COV	COV

Appendix 3A

Estrogen receptor results from different ML classifiers, using 15 and 25 features.

15 FEATURES								
MODEL	K-fold							
	3		5		10		15	
	CV	Test	CV	Test	CV	Test	CV	Test
LR	72% ± 2%	72%	67% ± 8%	67%	65 ± 5%	65%	63% ± 2%	67%
RF	67% ± 4%	67%	67% ± 9%	65%	69% ± 9%	67%	65% ± 11%	65%
NB	66% ± 12%	65%	53% ± 11%	56%	61% ± 2%	65%	58% ± 9%	56%
LDA	70% ± 7%	72%	65% ± 11%	65%	69% ± 12%	72%	67% ± 5%	63%
DT	56% ± 2%	51%	55% ± 6%	56%	59%± 4%	65%	60% ± 6%	56%
SVM	61% ± 5%	56%	65% ± 2%	65%	66% ± 8%	65%	63% ± 5%	65%
KNN	69% ± 10%	67%	71% ± 6%	67%	70% ± 12%	72%	69% ± 3%	67%
ENSEMBLE	66% ± 8%	67%	65% ± 5%	65%	66% ± 10%	67%	63% ± 8%	65%
25 FEATURES								
MODEL	K-fold							
	3		5		10		15	
	CV	Test	CV	Test	CV	Test	CV	Test
LR	71% ± 5%	72%	68% ± 4%	67%	73% ± 2%	72%	69% ± 3%	65%
RF	65% ± 9%	63%	69% ± 4%	67%	67% ± 5%	65%	65% ± 7%	65%
NB	64% ± 4%	63%	60% ± 5%	58%	65% ± 13%	58%	59% ± 3%	56%
LDA	74% ± 12%	74%	72% ± 8%	70%	72% ± 7%	70%	66% ± 5%	63%
DT	63% ± 2%	63%	59% ± 4%	53%	62% ± 5%	58%	55% ± 3%	53%
SVM	69% ± 4%	65%	71% ± 5%	65%	68% ± 5%	65%	68% ± 9%	67%
KNN	71% ± 14%	70%	74% ± 10%	70%	71% ± 12%	70%	63% ± 10%	60%
ENSEMBLE	65% ± 4%	67%	67% ± 3%	67%	69% ± 6%	65%	64% ± 4%	56%

Appendix 3B

Progesterone receptor prediction results from different ML classifiers, using 15 and 25 features.

15 FEATURES								
MODEL	K-fold							
	3		5		10		15	
	CV	Test	CV	Test	CV	Test	CV	Test
LR	71% ± 11%	70%	68% ± 4%	67%	68% ± 3%	67%	62% ± 7%	65%
RF	65% ± 9%	67%	70% ± 5%	65%	65% ± 11%	67%	69% ± 7%	67%
NB	66% ±0.12%	65%	54% ± 4%	56%	63% ± 8%	67%	59% ± 6%	56%
LDA	75% ± 11%	74%	69% ± 10%	67%	71% ± 5%	70%	66% ± 6%	67%
DT	54% ± 3%	51%	54% ± 6%	56%	64% ± 5%	65%	61% ± 8%	58%
SVM	61% ± 2%	56%	69% ± 8%	67%	64% ± 3%	65%	62% ± 9%	63%
KNN	67% ± 5%	65%	70% ± 12%	67%	72% ± 9%	69%	67% ± 9%	67%
ENSEMBLE	66% ± 8%	67%	64% ± 6%	63%	65% ± 14%	65%	64% ± 5%	63%
25 FEATURES								
MODEL	K-fold							
	3		5		10		15	
	CV	Test	CV	Test	CV	Test	CV	Test
LR	70% ± 7%	72%	69% ± 8%	67%	71% ± 8%	70%	68% ± 7%	67%
RF	66% ± 11%	65%	69% ± 3%	67%	69% ± 10%	67%	67% ± 4%	65%
NB	62% ± 8%	60%	59% ± 6%	58%	62% ± 12%	60%	54% ± 7%	53%
LDA	73% ± 12%	72%	71% ± 6%	67%	70% ± 7%	70%	68% ± 6%	65%
DT	62% ± 7%	60%	56% ± 9%	53%	60% ± 2%	60%	59% ± 3%	56%
SVM	65% ± 4%	58%	70% ± 7%	65%	66% ± 9%	67%	65% ± 6%	65%
KNN	70% ± 10%	67%	73% ± 11%	70%	71% ± 5%	70%	65% ± 9%	63%
ENSEMBLE	67% ± 6%	67%	69% ± 5%	67%	67% ± 9%	67%	65% ± 3%	60%

Appendix 3C

HER2-receptor prediction results from different ML classifiers, using 15 and 25 features.

15 FEATURES								
MODEL	K-fold							
	3		5		10		15	
	CV	Test	CV	Test	CV	Test	CV	Test
LR	61% ± 8%	58%	66% ± 13%	65%	64% ± 6%	65%	57% ± 9%	53%
RF	61% ± 4%	63%	62% ± 9%	63%	69% ± 7%	69%	62% ± 8%	65%
NB	52% ± 10%	53%	50% ± 4%	53%	56% ± 2%	63%	55% ± 3%	56%
LDA	64% ± 2%	63%	62% ± 6%	65%	70% ± 7%	56%	67% ± 2%	63%
DT	53% ± 7%	53%	53% ± 9%	51%	57% ± 4%	67%	52% ± 5%	56%
SVM	62% ± 6%	65%	67% ± 4%	65%	65% ± 5%	65%	63% ± 5%	60%
KNN	58% ± 3%	56%	63% ± 5%	53%	63% ± 7%	65%	65% ± 7%	65%
ENSEMBLE	62% ± 5%	65%	63% ± 8%	58%	61% ± 8%	67%	58% ± 6%	56%
25 FEATURES								
MODEL	K-fold							
	3		5		10		15	
	CV	Test	CV	Test	CV	Test	CV	Test
LR	63% ± 7%	60%	67% ± 7%	63%	65% ± 4%	63%	59% ± 4%	58%
RF	63% ± 5%	63%	62% ± 6%	63%	66% ± 5%	67%	64% ± 2%	60%
NB	54% ± 7%	56%	56% ± 9%	53%	58% ± 5%	56%	53% ± 5%	51%
LDA	62% ± 9%	60%	64% ± 3%	63%	71% ± 5%	67%	67% ± 4%	65%
DT	55% ± 4%	56%	61% ± 4%	58%	56% ± 8%	51%	56% ± 3%	53%
SVM	66% ± 5%	63%	65% ± 7%	65%	66% ± 3%	63%	66% ± 8%	65%
KNN	58% ± 6%	53%	67% ± 8%	67%	66% ± 8%	65%	68% ± 8%	65%
ENSEMBLE	64% ± 7%	63%	62% ± 5%	58%	65% ± 7%	60%	62% ± 9%	56%

References

[1] Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J Clin. 2021;71(3):209-249.
[2] Onitilo AA, Engel JM, Greenlee RT, Mukesh BN. Breast cancer subtypes based on er/pr and her2 expression: comparison of clinicopathologic features and survival. Clinical Medicine & Research. 2009;7(1-2):4-13.
[3] Pace LE, Keating NL. A systematic assessment of benefits and risks to guide breast cancer screening decisions. JAMA. 2014;311(13):1327.
[4] Seely JM, Alhassan T. Screening for breast cancer in 2018—what should we be doing today? Current Oncology. 2018;25(11):115-124.
[5] Nelson HD, Fu R, Cantor A, Pappas M, Daeges M, Humphrey L. Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 u. S. Preventive services task force recommendation. Ann Intern Med. 2016;164(4):244.
[6] H. Burhenne, L. W. Burhenne, F. Goldberg, T. Hislop, A. Worth, P. Rebbeck, and L. Kan, “Interval breast cancers in the screening mammography program of british columbia: analysis and classiffcation.," AJR.American journal of roentgenology, vol. 162, no. 5, pp. 1067-1071,1994.
[7] C. L. Robertson, “A private breast imaging practice: medical audit of 25,788 screening and 1,077 diagnostic examinations." Radiology, vol. 187, no. 1, pp. 75-79, 1993.
[8] Spak DA, Plaxco JS, Santiago L, Dryden MJ, Dogan BE. BI-RADS ® fifth edition: A summary of changes. Diagnostic and Interventional Imaging. 2017;98(3):179-190.
[9] F. Scopinaro, R. Pani, G. De Vincentis, A. Soluri, R. Pellegrini, and L. M. Porfiri ,”High-resolution scintimammography improves the accuracy of technetium-99m methoxy- isobutylisonitrile scintimammography: use of a new dedicated gamma camera,"European journal of nuclear medicine, vol. 26, no. 10, pp. 1279 -1288, 1999.
[10] Y. Sun, W. Wei, H.-W. Yang, and J.-L. Liu, “Clinical usefulness of breast-specific gamma imaging as an adjunct modality to mammography for diagnosis of breast cancer: a systemic review and meta-analysis," European journal of nuclear medicine and molecular imaging, vol. 40, no. 3, pp. 450-463, 2013.
[11] Kuhn KJ, Rapelyea JA, Torrente J, Teal CB, Brem RF. Comparative diagnostic utility of low-dose breast-specific gamma imaging to current clinical standard. Breast J. 2016;22(2):180-188.
[12] Hruska CB, Weinmann AL, O’Connor MK. Proof of concept for low-dose molecular breast imaging with a dual-head CZT gamma camera. Part I. Evaluation in phantoms: Low-dose MBI in phantoms. Med Phys. 2012;39(6Part1):3466-3475.
[13] Hruska CB, Weinmann AL, Tello Skjerseth CM, et al. Proof of concept for low-dose molecular breast imaging with a dual-head CZT gamma camera. Part II. Evaluation in patients: Low-dose MBI in patients. Med Phys. 2012;39(6Part1):3476-3483.
[14] Zhang A, Li P, Liu Q, Song S. Breast-specific gamma camera imaging with 99mTc-MIBI has better diagnostic performance than magnetic resonance imaging in breast cancer patients: A meta-analysis. Hell J Nucl Med. 2017;20(1):26-35.
[15] Rechtman LR, Lenihan MJ, Lieberman JH, et al. Breast-specific gamma imaging for the detection of breast cancer in dense versus nondense breasts. AJR Am J Roentgenol. 2014;202(2):293-298.
[16] A. A. van Loevezijn, A. C. van Breda Vriesman, P. A. Neijenhuis, and L. M. Pereira Arias-Bouda, “Breast-specific gamma imaging in breast cancer" Ned Tijdschr Geneeskd, vol. 160, p. A9610, 2016.
[17] Huppe AI, Mehta AK, Brem RF. Molecular breast imaging: a comprehensive review. Semin Ultrasound CT MR. 2018;39(1):60-69.
[18] Rao VM, Levin DC, Parker L, Cavanaugh B, Frangos AJ, Sunshine JH. How widely is computer-aided detection used in screening and diagnostic mammography? Journal of the American College of Radiology. 2010;7(10):802-805.
[19] Tchou PM, Haygood TM, Atkinson EN, et al. Interpretation time of computer-aided detection at screening mammography. Radiology. 2010;257(1):40-46.

[20] Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med. 2007;356(14):1399-1409.
[21] Gilbert FJ, Astley SM, Gillan MGC, et al. Single reading with computer-aided detection for screening mammography. N Engl J Med. 2008;359(16):1675-1684.
[22] El-Khatib H, Popescu D, Ichim L. Deep learning–based methods for automatic diagnosis of skin lesions. Sensors. 2020;20(6):1753.
[23] Badar M, Haris M, Fatima A. Application of deep learning for retinal image analysis: A review. Computer Science Review. 2020;35:100203.
[24] Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain mri segmentation: state of the art and future directions. J Digit Imaging. 2017;30(4):449-459.
[25] Baskar S, Shakeel PM, Sridhar KP, Kanimozhi R. Classification system for lung cancer nodule using machine learning technique and ct images. In: 2019 International Conference on Communication and Electronics Systems (ICCES). IEEE; 2019:1957-1962.
[26] Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. J Physiol 195:215–243
[27] Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst, Man, Cybern. 1973;SMC-3(6):610-621.
[28] Crivelli P, Ledda RE, Parascandolo N, Fara A, Soro D, Conti M. A new challenge for radiologists: radiomics in breast cancer. BioMed Research International. 2018;2018:1-10.
[29] Rizzo S, Botta F, Raimondi S, et al. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp. 2018;2(1):36.
[30] Lee SJ, Choi YY, Kim C, Chung MS. Correlations between tumor to background ratio on breast-specific gamma imaging and prognostic factors in breast cancer. J Korean Med Sci. 2017;32(6):1031-1037.
[31] Yoon H-J, Kim Y, Chang K-T, Kim BS. Prognostic value of semi-quantitative tumor uptake on Tc-99m sestamibi breast-specific gamma imaging in invasive ductal breast cancer. Ann Nucl Med. 2015;29(7):553-560.
[32] Lu C-F, Hsu F-T, Hsieh KL-C, et al. Machine learning–based radiomics for molecular subtyping of gliomas. Clin Cancer Res. 2018;24(18):4429-4436.
[33] Hyun SH, Ahn MS, Koh YW, Lee SJ. A machine-learning approach using pet-based radiomics to predict the histological subtypes of lung cancer. Clin Nucl Med. 2019;44(12):956-960.
[34] Kaissis G, Ziegelmayer S, Lohöfer F, et al. A machine learning algorithm predicts molecular subtypes in pancreatic ductal adenocarcinoma with differential response to gemcitabine-based versus FOLFIRINOX chemotherapy. Real FX, ed. PLoS ONE. 2019;14(10):e0218642.
[35] Uhlig J, Leha A, Delonge LM, et al. Radiomic features and machine learning for the discrimination of renal tumor histological subtypes: a pragmatic study using clinical-routine computed tomography. Cancers. 2020;12(10):3010.
[36] Park YW, Oh J, You SC, et al. Radiomics and machine learning may accurately predict the grade and histological subtype in meningiomas using conventional and diffusion tensor imaging. Eur Radiol. 2019;29(8):4068-4076.
[37] Hu J, Zhao Y, Li M, et al. Machine-learning-based computed tomography radiomic analysis for histologic subtype classification of thymic epithelial tumors. European Journal of Radiology. 2020;126:108929.
[38] van Loevezijn AA, van Breda Vriesman AC, Neijenhuis PA, Pereira Arias-Bouda LM. [Breast-specific gamma imaging in breast cancer]. Ned Tijdschr Geneeskd. 2016;160:A9610.
[39] Conners AL, Maxwell RW, Tortorelli CL, et al. Gamma camera breast imaging lexicon. AJR Am J Roentgenol. 2012;199(6):W767-774.
[40] Barry-Straume, Jostein, et al. "An evaluation of training size impact on validation accuracy for optimized convolutional neural networks." *SMU Data Science Review* 1.4 (2018): 12.
[41] Qiao T, Liu S, Cui Z, et al. Deep learning for intelligent diagnosis in thyroid scintigraphy. J Int Med Res. 2021;49(1):030006052098284.

[42] Hamidinekoo, A.; Denton, E.; Rampun, A.; Honnor, K.; Zwiggelaar, R. Deep learning in mammography and breast histology, an overview and future trends. Med. Image Anal. 2018, 47, 45–67
[43] Krizhevsky, A., Hinton, G., 2009. Learning multiple layers of features from tiny images.
[44] Simonyan, K. , Zisserman, A. ,2014. Very deep convolutional networks for large scale image recognition. In: International Conference on Learning Representations . arXiv preprint
[45] Zhao Z, Pi Y, Jiang L, et al. Deep neural network based artificial intelligence assisted diagnosis of bone scintigraphy for cancer bone metastasis. Sci Rep. 2020;10(1):17046.
[46] Cheng D-C, Hsieh T-C, Yen K-Y, Kao C-H. Lesion-based bone metastasis detection in chest bone scintigraphy images of prostate cancer patients using pre-train, negative mining, and deep learning. Diagnostics. 2021;11(3):518.
[47] Aoki Y, Nakayama M, Nomura K, et al. The utility of a deep learning-based algorithm for bone scintigraphy in patient with prostate cancer. Ann Nucl Med. 2020;34(12):926-931.
[48] Arvidsson I, Overgaard NC, Astrom K, et al. Prediction of obstructive coronary artery disease from myocardial perfusion scintigraphy using deep neural networks. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE; 2021:4442-4449.
[49] Yu X, Hu G, Zhang Z, et al. Retrospective and comparative analysis of 99mTc-Sestamibi breast specific gamma imaging versus mammography, ultrasound, and magnetic resonance imaging for the detection of breast cancer in Chinese women. BMC Cancer. 2016;16(1):450.
[50] Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310.
[51] Burkart N, Huber MF. A survey on the explainability of supervised machine learning. jair. 2021;70:245-317.
[52] Ma W, Zhao Y, Ji Y, et al. Breast cancer molecular subtype prediction by mammographic radiomic features. Academic Radiology. 2019;26(2):196-201.
[53] Yoo J, Yoon H-J, Kim BS. Prognostic value of primary tumor SUVmax on F-18 FDG PET/CT compared with semi-quantitative tumor uptake on Tc-99m sestamibi breast-specific gamma imaging in invasive ductal breast cancer. Ann Nucl Med. 2017;31(1):19-28.
[54] Mayerhoefer ME, Szomolanyi P, Jirak D, Materka A, Trattnig S. Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: An application-oriented study: Effects of MRI acquisition parameters on texture analysis. Med Phys. 2009;36(4):1236-1243.
[55] Moadel RM. Breast cancer imaging devices. Seminars in Nuclear Medicine. 2011;41(3):229-241.
[56] Rahbar H, Partridge SC, DeMartini WB, Thursten B, Lehman CD. Clinical and technical considerations for high quality breast MRI at 3 tesla. J Magn Reson Imaging. 2013;37(4):778-790.
[57] Karssemeijer N, Frieling JTM, Hendriks JHCL. Spatial resolution in digital mammography: Investigative Radiology. 1993;28(5):413-419.
[58] Rizzo S, Botta F, Raimondi S, et al. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp. 2018;2(1):36.
[59] van Dooijeweert C, Deckers IAG, Baas IO, van der Wall E, van Diest PJ. Hormone- and HER2-receptor assessment in 33,046 breast cancer patients: a nationwide comparison of positivity rates between pathology laboratories in the Netherlands. Breast Cancer Res Treat. 2019;175(2):487-497.
[60] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

PART 3

Clinical validation

The use of artificial intelligence in Breast-specific Gamma Imaging

Clinical validation

Author

J.R.J. Willemse

Supervisors

W. Grootjans
Leiden University Medical Center

L.M. Pereira Arias-Bouda
Leiden University Medical Center

F.H.P. van Velden
Leiden University Medical Center

Abstract

Introduction

Over the years, many applications of artificial intelligence(AI) in healthcare settings have been proposed. A small proportion of those, however, fails to be implemented clinically, despite good performances in research settings. For clinical implementation, however, AI algorithms must be validated and evaluated. In this paper, we aim to: 1) find a possible correlation by the predictions made by an AI algorithm and assessments by nuclear medicine physicians in breast-specific gamma imaging(BSGI) and 2) determine possible time-saving capabilities of using an AI algorithm when compared to standard clinical routine.

Methods

13 patients who underwent BSGI for a problem-solving indication were included, and BI-RADS classifications made by physicians were compared to predictions made by a previously build convolutional neural network(CNN). For this part, the Pearsons correlation coefficient and Cohen's kappa agreement scores were calculated. Also, the BSGI interpretation times by nuclear medicine physicians were compared to the time required by the CNN to make a prediction for each individual patient.

Results

Out of 26 breasts, only 4 were classified as BI-RADS 4 or higher, resulting in follow-up diagnosis and pathological assessment. Correlation coefficients of $r = 0.71$ and an $r^2 = 0.51$ were found between the average prediction per breast made by the CNN and the BI-RADS classification by the physician. For BI-RADS scores of 1 and 2, a kappa agreement score of 0.53 was found, while a kappa score of 0.25 for BI-RADS 3 and 4 and 0.64 for BI-RADS 5 were found. Average total interpretation time for a single patient by a nuclear medicine physician was 255 seconds, while the CNN made predictions in 31 seconds.

Discussion and conclusion

Despite a positive correlation that was found between BI-RADS scores and predictions made by the AI, the small number of patients with suspicious lesions, and the large variability of AI predictions within each BI-RADS category indicate the necessity for further research. It is therefore advised that similar research with larger cohorts are conducted to confirm the correlation found in this study.

Contents

Introduction

Artificial Intelligence (AI) applications in healthcare settings have been widely investigated and evaluated in the recent years, with many applications showing high efficacy and efficiency in clinical research. Few AI applications however, have made it through to actual clinical practice[1,2], with many proposed AI-based algorithms remaining in the prototype phase, despite promising results.

Several reasons exist for this gap between research and clinical practice. Reasons for hesitancy in adopting AI-based clinical decision making solutions include lack of explainability, legal and ethical issues, and pressure on medical professionals on continuously having to adapt to new technologies, resulting in a low number of studies of studies prospectively monitoring AI performance in real-world clinical environments[3,4]. This research-practice gap is also referred to as the 'AI-chasm', where potentially sound algorithms are not further evaluated in clinical practice[5,6]. To cross the AI-chasm, theoretical algorithms have to be tested on real-world data in order to see whether generalizability is achievable, to compare the physicians' outcomes with the algorithms' outcomes, and to evaluate the effect of AI-implementation on the physicians' workflow.

This phase in the development cycle of AI in healthcare is, the clinical validation phase, is an essential step in finding failure modes and possible impacts on workflows, as well as finding hidden biases in the model. Based on findings from the clinical validation phase, models can be adjusted to specific needs for the clinical use case[7,8].

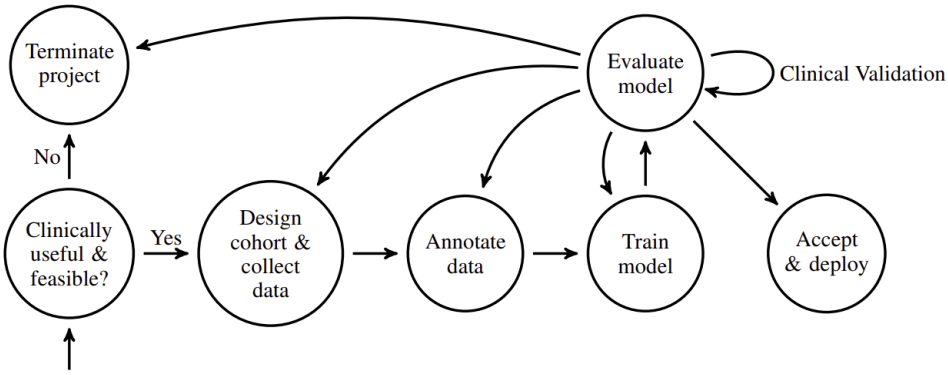


Figure 1 Development cycle of clinical AI models[4]

Recently, we have creating a deep-learning based convolutional neural networks (CNNs) for the detection of malignancies in breast-specific gamma imaging(BSGI). BSGI is a scintigraphic nuclear medicine imaging technique used in the diagnosis, therapy response monitoring and screening in patients with, or patients suspected of having breast cancer[9]. One of the indications for using BSGI as an adjunct imaging modality is problem solving. Problem solving is defined as the evaluation of indeterminate breast abnormalities or symptoms, where there are discrepancies between clinical and radiological findings, where evaluation of mammographic BI-RADS 3 (Breast Imaging Reporting and Data System) lesions is required for patient reassurance, or where (bloody) nipple discharge with normal or inconclusive radiological findings is evaluated.[10]

The best performing algorithm reached an area under the receiving operating characteristic (AUC) of 0.93 with an accuracy, sensitivity and specificity of 85%, 78% and 92%, respectively, at a cut-off malignancy probability of 0.5(50%). One main limitation of that study, however, was the evaluation on a per-image basis, instead of on a per-patient basis, which is the clinical routine.

BSGI images are interpreted using a BI-RADS lexicon specifically proposed for BSGI. This BI-RADS scoring system is based on uptake patterns, associated findings(axillary uptake, nipple uptake, vessel uptake), lesion location, qualitative intensity of lesion uptake and lesion size[11]. The entire BI-RADS lexicon for BSGI images, proposed by Conners et al.[11], can be found in appendix 1. Following image assessment, each breast is given a BI-RADS score from 0 to 6 (1- Negative (no lesion found), 2- Benign(no malignant features), 3- Probably benign(very low probability of cancer), 4- Suspicious (intermediate probability of cancer) or 5- Highly suggestive of malignancy). Categories 1 and 6 are assigned to incomplete imaging (wrong acquisition, artefacts, etc.), and known biopsy-proven malignancies respectively. Because the extent to which suspected malignancies in this lexicon are based on image characteristics, we hypothesize that predictions from our deep-learning application positively correlate with BI-RADS scores assigned by nuclear medicine physicians.

In this paper, we aim to evaluate the deep-learning algorithm on a per-patient basis in patients undergoing BSGI for a problem-solving indication(with no pathologically proven diagnosis yet), find possible correlations between the predictions of the algorithm and assigned BI-RADS scores, quantify the agreement between the AI and nuclear medicine physicians and determine the time-saving capabilities of using AI compared to clinical routine.

Methods

To evaluate the performance of the model in a clinical setting, patients were included from June to August 2021 in the *Alrijne Hospital Leiderdorp, The Netherlands*, who underwent BSGI for a problem-solving indication with the *Discovery NM 750b (GE Healthcare, Chicago, USA)*. Evaluation of BSGI images acquired for these patients was performed according to local clinical routine protocols, and patients were assigned BI-RADS scores for each breast, resulting to either scheduled follow-up (BI-RADS 1-3) or further investigation (BI-RADS >3).

To evaluate time saving capabilities, nuclear medicine physicians annotated interpretation times for each patient including time required for: 1) preparation and reading up on clinical data and ; 2) interpretation and reporting times. Preparation and reading up on clinical data includes viewing previously acquired imaging.

As model inputs, DICOM images from the cranio-caudal (CC) and mediolateral-oblique(MLO) BSGI projections were used, as these are the same projections that were used to train the model. Because initial evaluation of the model also resulted in the best performance using data-augmentation of the test set, this technique was also applied on the images used in this clinical validation. This resulted in 2 predictions(one for the original image and one for the augmented version) by the model per DICOM image, which were averaged to acquire a percentage on the possible presence of a malignancy (malignancy prediction).

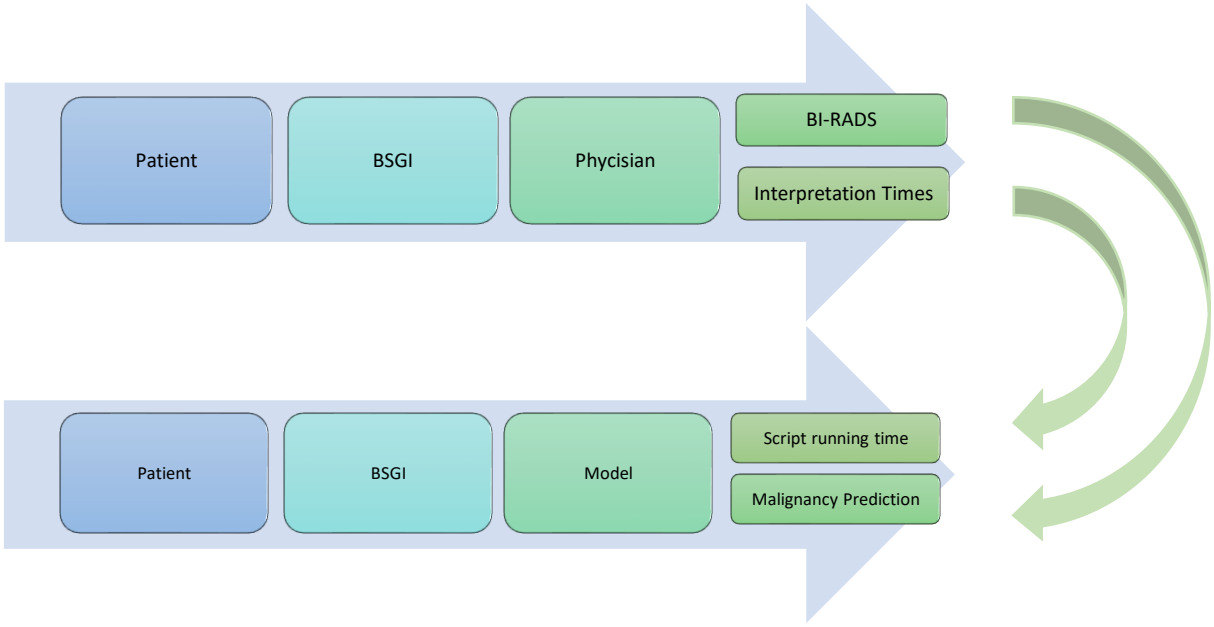


Figure 2 Workflow for the clinical validation phase. Bi-RADS scores are compared to malignancy predictions, and interpretation times are compared to the running time of the script.

To evaluate results, the average malignancy prediction per breast was correlated with its assigned BI-RADS score. The correlation was calculated using Pearson's correlation coefficient Also, Cohen's Kappa scores were calculated to find the agreement scores between nuclear medicine physicians and the model. To do so, the average malignancy predictions were categorized into 3 categories: 0 – 25%, 26 – 50% and >50%. We used three categories to make results comparable to a previously published study by Conners et al.[11], who also used three categories (Negative or benign, probably benign, and suspicious) to asses interobserver agreement among physicians in the interpretation of BSGI images. For instance, if a breast that was given a BI-RADS 2 score results in an average prediction of the model (across all projections of that breast) of 0-25%, both would be in agreement.

The AI script was run on a computer with 64-bit *Windows 10 64-bit OS (v.21H1)*and an *Intel® Core™ i7-770HQ CPU @ 1.80GHz*, with 8GB of RAM.

Results

Patients

A total of 13 female patients were included, all of whom had both breasts investigated with BSGI. Mean patient age was 56 at the time of image acquisition. Out of 26 breast investigated, 12 received a BI-RADS 1 score by the nuclear medicine physician, 7 received a BI-RADS 2 score, while 3, 4 and 1 breast received BI-RADS 3, 4 and 5 scores, respectively. These characteristics are summarized in table 1.

Table 1 BI-RADS scores assigned to BSGI scans

VARIABLES	
PATIENTS	N = 13
AGE(YEARS) MEAN	56 (37-74)
BI-RADS 1	N = 12
BI-RADS 2	N = 7
BI-RADS 3	N = 3
BI-RADS 4	N = 3
BI-RADS 5	N = 1

Model outputs

For each patient, BSGI images of both breast were run through the model, generating 8 predictions per breast (projections: AVG CC, GEO CC, CC TOP, CC BOT, AVG MLO, GEO MLO, MLO TOP, MLO BOT). Figure 3 shows the average malignancy prediction by the algorithm for each BI-RADS score assigned to the same series of projections. Also, the average predictions of each BI-RADS category are shown. For BI-RADS 1, this average malignancy prediction across all breasts in this category was 13,17%, 21,51% for BI-RADS 2 lesions, and 34,4%, 49,61% and 92,67% for BI-RADS 3, 4 and 5, respectively. Between these observations (average per BI-RADS omitted), the Pearson correlation coefficient was 0.71 ($p<0.05$), and r^2 =0.51.

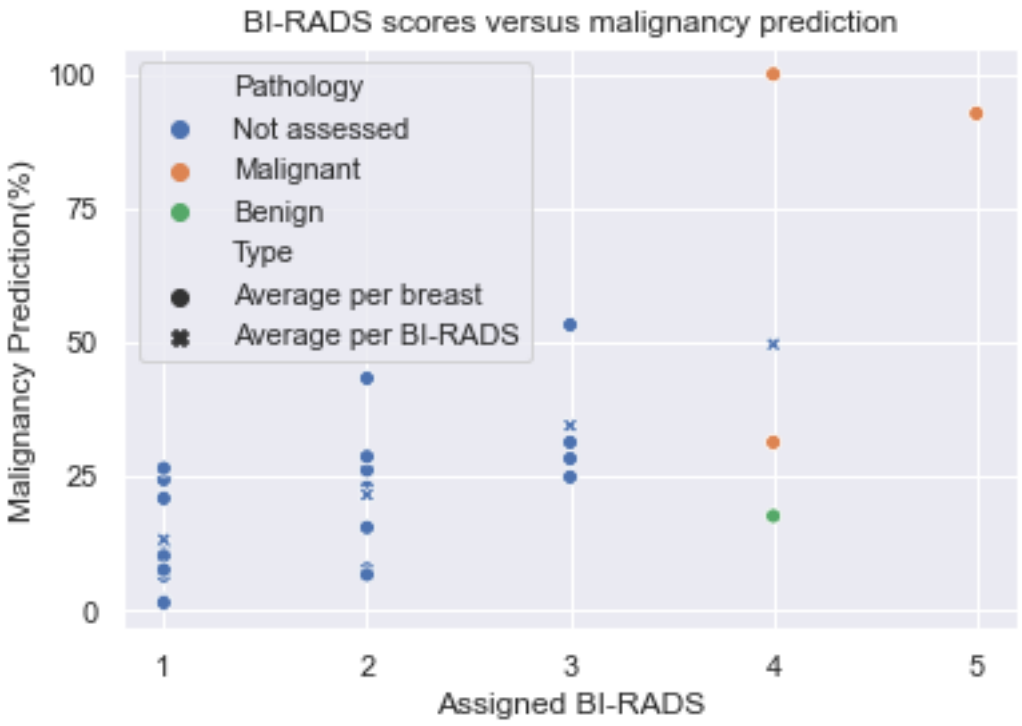


Figure 3 Average malignancy predictions by the model for each BI-RADS category

Agreement scores

Cohen's Kappa scores were calculated to determine the agreement between the model and the nuclear medicine physician across three categories separately. We found that for breasts being classified with either a BI-RADS 1 or 2 score, the kappa score was 0.53. For breasts with either a BI-RADS 3 or 4 score, the agreement between the nuclear medicine physicians and the AI resulted in an agreement of 0.28, while the agreement for breasts with a high BI-RADS score was 0.64.

The agreement results are summarized in table 2.

Table 2 Agreement between Nuclear medicine physicians and the AI-model

PHYSICIANS INTERPRETATION	**AI-PREDICTION** **PREDICTED PROBABILITY OF A MALIGNANCY PRESENT**			
	0 – 25%	>25% - 50%	>50%	Kappa score
BI-RADS 1 & 2	16	3	1	0.53
BI-RADS 3 & 4	2	2	2	0.25
BI-RADS 5	0	0	1	0.64

Time savings

Times were noted in 8 instances by nuclear medicine physicians. The average total time, a combination of time spent on reading up on relevant patient specific clinical data (reading) and the actual interpretation (reporting) time, was 255 seconds on average, with a standard deviation of 61 seconds. Average reading time was 191s (± 67s), while reporting times were 65s on average (±19s).

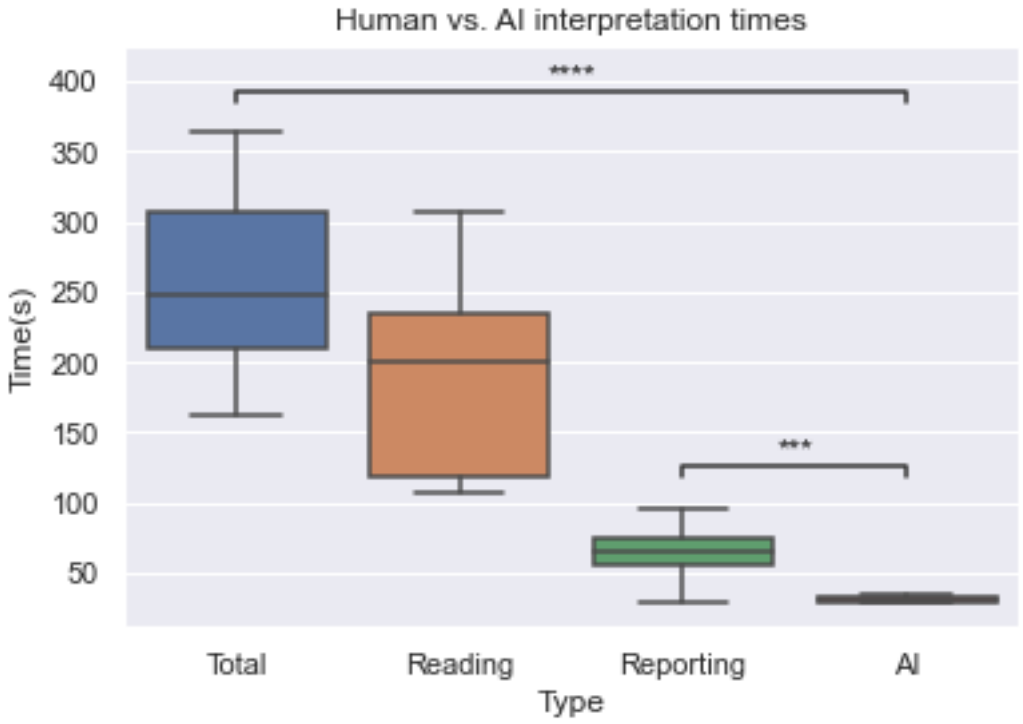


*Figure 4 Time spent on preparing image interpretation (reading), The actual reporting (reporting), their combined time(total) and the time needed for the AI-model to return results for a patient. ***: 1.00e-04 < p <= 1.00e-03 , ****: p <= 1.00e-04, paired t-test.*

Average time for the AI model to give predictions for each patient (left and right breast) was 31s ± 2s. The AI had a significantly shorter time to interpret all images for a patient compared to clinical routine, with p-values well under 0.05. The average reduction in time spent on interpretation was reduced from 255 seconds to 31 seconds, a 87.8% reduction.

Discussion

Interpretation

In this clinical validation study, we aimed to evaluate: 1) a possible correlation between the output of the model and assigned BI-RADS scores by a nuclear medicine phycisian; 2) the agreement between phycisians and the models' output and finally: 3) The potential time-reduction for interpretation of BSGI images with AI compared to current clinical routine.

As hypothesized, a positive correlation ($r=0.71$, $r^2=0.51$) can be seen between the average predicted chances of a malignancy by the AI-model and the BI-RADS scores assigned to those same breasts. There appear to be some parralells in the interpretation by the physician and the model. This can be explained by the fact that BI-RADS scores are based upon similar features that would be of incluence to the activation of neural nodes in the deep-learning model. In the case of high focal uptake, for instance, the AI-model has shown to predict higher probabilities of a malignancy being present, which is also a subject of the BI-RADS lexicon. Inversely, diffuse uptake patterns have been shown to decrease the predicted malignancy probability of the AI-model, and are also subject of the BI-RADS lexicon.

The agreement between the AI-model and the BI-RADS scores vary across different BI-RADS scores. The highest agreement can be found in the BI-RADS 5 category. However, only one breast from the 13 patients included received a BI-RADS 5 score. Therefore, the reliability of this single measurement, although correct, must still be doubted and can not lead to any conclusions. The lowest agreement was found in BI-RADS 3 and BI-RADS 4 scores. This low score (0.25) is can be mainly attributed to some BI-RADS 2 breasts were just above the 25% threshold used in these calculations, while one BI-RADS 3 breast was slightly above the 50% threshold. This low agreement, however, is in line with a prior study by Conners et al[11], where an interobserver agreement in the assessment of BSGI images in isolation (without additional clinical information) of 0.15 was found in BI-RADS 3 versus BI-RADS 1 and BI-RADS 2. *Conners et al.* also found that with the addition of mammograms, interobserver agreement increased the kappa to 0.3, indicating more concensus is reached with more information available. According to *Landis et al.*[12], this kappa score of 0.15 implies agreement, albeit very limited. A moderate agreement in BI-RADS 1 and 2 lesions compared to other lesions was sginificantly lower than the same value reported by *Conners et al.* (0.53 versus 0.91 respectively) this lower score found in this paper was mainly caused by the fact that several BI-RADS 2 classified breasts resulted in predictions slightly above 25%, one BI-RADS 1 breast resulted in a predicted probability of slightly above 25%, and one BI-RADS 3 breast resulted in a predicted probability of slightly below 25%. Furthermore, in one of of the breasts in these categories, high intensity uptake was measured in the nipple, an effect that already has caused false positive findings in the network before, as it had not seen enough similar examples to distinguish this benign finding from malignant uptake patterns.

The speed of the AI-model in putting out results has been shown to be significantly higher than that of nuclear medicine phycisians. The apparent correlation between BI-RADS scores and outputs by the model could potentially pave the way for decision-support AI-models in the workflow of nuclear medicine phycisians. A combination of pre-assessed images by the algorithm and clinical information

obtained by the nuclear medicine phycisian could speed up the diagnostic process, saving time and money, while freeing time for more patients.

The major limitation of this clinical validation phase is the small number of patients included. As a result, a relatively small number of BI-RADS 3, 4, and 5 classifications were tested for correlations and agreement with the outputs of the AI-model. Especially the correlation with BI-RADS 3 and 4 lesions is insufficient to demonstrate with this small cohort, even though these classifications are possibly the most interesting in terms of decision-making support using AI, as a shift from a BI-RADS 3 to 4 or inversely has a direct impact on further diagnosis and treatment for patients.

Another limitation is the fact that BI-RADS 3 breasts(or lesions) are classified as 'probably benign', and not defintively benign, meaning that no further diagnostic testing is performed. Although unlikely, there is a possibility that one of the overclassifications by the AI-model was in fact not a false-positive, but a true positive. All BI-RADS 4 and 5 classifications were eventaully proven to be malignant, except for one BI-RADS 4 classification which the AI-model predicted to be not malignant. This lesion was eventually found to be benign after pathological assessment.

Conclusion

In conclusion, there appears to be a positive correlation between BI-RADS scores and predictions made by the AI-model. This correlation, in combination with slight to substantial agreement between nuclear medicine physicians and the speed of AI in estimating the chance of a mlaignancy indicates that this study could pave the way for future AI-based decision making algorithms in the diagnosis of breast cancer in patients undergoing BSGI. However, future studies should include more patients, especially with BI-RADS scores of 3 and higher.

Appendices

Appendix 1 – BSGI BI-RADS Lexicon by Conners et al.[11]

Indication: Describe clinical problems (if any), history of biopsies (date and results), risk factors, indicate if patient is pre- (last menstrual period [LMP] less than one month ago), peri- (LMP more than one month ago and less than 12 months ago), or postmenopausal (LMP at least 1 year ago), phase of menstrual cycle (if relevant), and any use of selective estrogen receptor modulators or medications with estrogenic or progestogenic activity			
Comparison: Prior breast imaging, including prior gamma camera breast imaging studies (if any) should be reviewed, with the dates and types of prior studies reported			
Technical Factors: Report dose (MBq) and type of tracer injected and duration of circulation phase (time from injection to imaging). If additional views beyond routine CC and MLO projections were obtained, these should be detailed			
Limitations: Describe any suboptimal positioning, motion, pixel dropout, "hot pixels", electronic, or other artifacts which are felt to affect image interpretation			
Background	Describe degree of radiotracer uptake in background normal parenchyma, which may be uniform (homogeneous) or patchy (heterogeneous)		
	Photopenic	Less than subcutaneous fat	
	Minimal-Mild	Equal to or slightly greater than subcutaneous fat	
	Moderate	Visually greater than mild, but less than twice as intense as subcutaneous fat	
	Marked	Visually at least twice as intense as subcutaneous fat	
Findings: Categories and Terms			**Description**
Mass	Uptake which has convex outward borders, no interspersed normal uptake, and is seen on two projections (if location is amenable)		
Non-Mass Uptake	Uptake distinct from the surrounding tissue that does not fit criteria for a mass and which usually contains interspersed areas of normal glandular tissue		
	Distribution	Focal area	<25% of a quadrant or < 2 cm in diameter in a confined area
		Segmental	Uptake in linear or triangular region or cone with apex pointing toward nipple that suggests (but is not specific for) intraductal pathology
		Regional	Uptake in a large volume of tissue, ≥ 2 cm in diameter, not conforming to a ductal distribution; may be geographic
		Multiple regions	Uptake in at least two large volumes of tissue;

			more than one area of geographic uptake
		Diffuse	Uptake distributed throughout the breast
	Internal pattern of uptake	Homogeneous	Confluent, uniform uptake
		Heterogeneous/Patchy	Variable, nonuniform uptake
	Symmetry	Symmetric	Similar uptake pattern in both breasts
		Asymmetric	More uptake in one breast compared to the other
Associated Findings		Axillary uptake	Uptake in the axilla, usually thought to be a lymph node which may or may not be pathologic
		Nipple uptake	Radiotracer uptake within the nipple, a physiologic finding if not associated with other suspicious uptake
		Vessel uptake	Serpiginous linear uptake corresponding with a vessel
Location	Breast	Right, left or bilateral	
	In-breast location	Quadrant or clock-face location, or specifically in the subareolar or central breast or axillary tail	
	Depth/Distance from the nipple	Anterior, central or posterior third or measured distance from the nipple	Measurement is made from the center of the finding and recorded in centimeters
Qualitative intensity of uptake in lesion[a]	Photopenic	Uptake in lesion is less than surrounding background parenchyma	
	Mild	Uptake which appears to be less than 50% of background	
	Moderate	Uptake which appears to be at least 50% of background but not twice as intense as background	
	Marked	Uptake which appears to be at least twice background uptake	
Lesion size	X	Longest measurement of the lesion, made on whichever image best depicts the lesion	

	Y	Measurement orthogonal to X, made using the same image used to define X
	Z	If the lesion is visible on both projections, Z should be an orthogonal measurement made on the projection (CC or MLO) not used to define X/Y
Assessment Categories		
Incomplete Assessment	0 - Incomplete	Additional imaging is needed before a final assessment can be rendered
Final Assessment	1 - Negative	No lesion found (routine follow-up)
	2 - Benign	No malignant features; e.g., photopenia (routine follow-up)
	3 - Probably benign	Very low probability of cancer (follow-up MBI examination is recommended in 6 months if targeted diagnostic mammogram and ultrasound are negative)
	4 - Suspicious	Intermediate probability of cancer (biopsy is recommended)
	4a-Low suspicion	Used for a finding which requires intervention but is of low suspicion for malignancy
	4b-Intermediate suspicion	Used for a finding which is judged to be of intermediate suspicion for malignancy
	4c-Moderate suspicion (but not classic)	Used for a finding which is judged to be of moderate suspicion for malignancy
	5 - Highly suggestive of malignancy	High probability of malignancy (biopsy is recommended)
	6 - Known biopsy-proven malignancy	Appropriate action should be taken

References

[1] Panch T, Mattie H, Celi LA. The “inconvenient truth” about AI in healthcare. npj Digit Med. 2019;2(1):77.
[2] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195.
[3] Morley J, Machado CCV, Burr C, et al. The ethics of AI in health care: A mapping review. Social Science & Medicine. 2020;260:113172.
[4] the Precise4Q consortium, Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310.
[5] Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. npj Digital Med. 2018;1(1):40, s41746-018-0048-y.
[6] McCradden MD, Stephenson EA, Anderson JA. Clinical research underlies ethical integration of healthcare artificial intelligence. Nat Med. 2020;26(9):1325-1326.
[7] Lu C, Strout J, Gauriau R, et al. An overview and case study of the clinical ai model development life cycle for healthcare systems. arXiv:200307678 [cs, eess]. Published online March 26, 2020.
[8] Bini SA. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care? The Journal of Arthroplasty. 2018;33(8):2358-2361.
[9] Muzahir S. Molecular breast cancer imaging in the era of precision medicine. American Journal of Roentgenology. 2020;215(6):1512-1519.
[10] van Loevezijn AA, van Breda Vriesman AC, Neijenhuis PA, Pereira Arias-Bouda LM. [Breast-specific gamma imaging in breast cancer]. Ned Tijdschr Geneeskd. 2016;160:A9610.
[11] Conners, Amy Lynn, et al. "Lexicon for standardized interpretation of gamma camera molecular breast imaging: observer agreement and diagnostic accuracy." *European journal of nuclear medicine and molecular imaging* 39.6 (2012): 971-982.
[12] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Acknowledgements

The past year has been a true adventure and a wonderful experience, in which I learned a lot, got to meet a lot of helpful and fun people, and in which I got to know myself better, especially my strengths and weaknesses.

I would like to express my sincere gratitude towards my supervisors: Willem, Lenka and Floris. You have given me the opportunity to develop myself, while keeping a finger on the pulse concerning the progress I made. Thank you for all the help and support throughout this year, it wouldn't have been possible without you!

Secondly, I would like to thank all the staff of the nuclear medicine departments of the LUMC and Alrijne hospital, for the warm welcome on your departments. Also, I would like to thank Anneke Zeillemaker for letting me join you on your daily routine.

Also, a big thank you to the research group of the LUMC: Wyanne, Maaike, Fleur, Pim, Timo, Dennis and Marijn, for the occasional drinks, digital 'coffee-meetings' and the laughter. Also, a big thank you to Alina for introducing me to the Alrijne hospital, and helping me finding my way around there.

Finally, I would like to thank my parents and stepparents, as well a my brother and the rest of my family and friends for their (mental) support. You have all made this past year much easier for me!

Jeroen