Global radiomic features for improving the process of curating educational mammographic test sets

Somphone Siviengphanom

Bachelor of Business (Computing and Information System)
Master of Education (Management and Leadership) Merit

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Discipline of Medical Imaging Science
Sydney School of Health Sciences
Faculty of Medicine and Health
The University of Sydney

July 2024

Candidate's declaration

I, Somphone Siviengphanom, hereby declare that the work contained within this thesis is my own and has not been submitted for any other degree at any other institution.

I affirm that I was the principal researcher of all the work included in this thesis. I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Ethical approvals from the University of Sydney Human Ethics Committee were granted for all relevant studies presented in this thesis.

I acknowledge that if a higher degree is awarded for this thesis, the thesis will be lodged with the University of Sydney Library and made available for immediate use.

Somphone Siviengphanom
15 July 2024

Authorship attribution statement

Chapter 2 of this thesis is published as Siviengphanom, S., Z. Gandomkar, S.J. Lewis, and P.C. Brennan, *Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs.* Academic Radiology, 2021. **29**(8): p. 1228–1247. I conducted the literature search, screened and assessed each article, analysed the data, summarised the findings, developed themes, designed the study, and wrote the drafts of, reviewed and edited the manuscript.

Chapter 3 of this thesis is published as Siviengphanom, S., S.J. Lewis, P.C. Brennan, and Z. Gandomkar, *Computer-extracted global radiomic features can predict the radiologists' first impression about the abnormality of a screening mammogram.* British Journal of Radiology, 2024. **97**: p. 168–179. I acquired, analysed, and interpreted data, conceptualised and designed the study, interpreted the results, wrote the drafts of, reviewed, and edited the manuscript.

Chapter 4 of this thesis is published as Siviengphanom, S., Z. Gandomkar, S.J. Lewis, and P.C. Brennan, *Global radiomic features from mammography for predicting difficult-to-interpret normal cases.* Journal of Digital Imaging, 2023. **36**: p. 1541–1552. I acquired, analysed, and interpreted data, conceptualised and designed the study, interpreted the results, wrote the drafts of, reviewed, and edited the manuscript.

Appendix 1.3 of this thesis is published as Siviengphanom, S., S.J. Lewis, P.C. Brennan, Z. Gandomkar. *Predicting the gist of breast cancer on a screening mammogram using global radiomic features*. in *SPIE Medical Imaging 2024: Image Perception, Observer Performance, and Technology Assessment*. 2024. San Diego, California, United States: SPIE. I acquired, analysed, and interpreted data, conceptualised and designed the study, interpreted the results, wrote the drafts of, reviewed, and edited the manuscript.

Appendix 1.5 of this thesis is published as Siviengphanom, S., Z. Gandomkar, S.J. Lewis, and P.C. Brennan. *Global mammographic radiomic signature can predict radiologists' difficult-to-interpret normal cases*. in *SPIE Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment*. 2023. San Diego, California, United States: SPIE. I acquired, analysed, and interpreted data, conceptualised and designed the study, interpreted the results, wrote the drafts of, reviewed, and edited the manuscript.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Somphone Siviengphanom
15 July 2024

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Patrick Brennan
15 July 2024

Acknowledgements

I would like to acknowledge and express my sincere gratitude to all the fantastic individuals who supported me kindly during my PhD journey and helped to make this research a reality.

First and foremost, I would like to give my high and respectful appreciation to my lead supervisor, Professor Patrick Brennan. My PhD has been an amazing experience and I thank him wholeheartedly, not only for his enormous academic support, but also for giving me so many wonderful expert advice, encouragement, and opportunities during my PhD journey.

My heartfelt appreciation for my auxiliary/former lead supervisor, Professor Sarah Lewis, is also tremendous. Her outstanding guidance, mentorship, and invaluable support have played a vital role in shaping the trajectory of my work, and I am truly appreciative of the time and effort she has invested in my academic and professional development.

Similarly, no words can describe how much I appreciate Dr Ziba Gandomkar, my truly dedicated auxiliary supervisor, mentor, and colleague, who shares her expertise so enthusiastically. Her superb supervision, inspiration, and advice have helped me greatly throughout my PhD journey, without which I would not be here today. Thank you countlessly for believing in me and my work and encouraging me to push on.

I would like to also express my genuine gratitude to Dr Phuong Dung (Yun) Trieu, my final auxiliary supervisor, for her motivation and insightful knowledge of the BreastScreen REader Assessment STrategy (BREAST) program. I am very grateful for the opportunities she provided for intellectual discussions and her patience in addressing my queries and concerns.

My profound appreciation also goes to Dr Robert Heard, an exceptional biostatistician, who has the kindest heart and greatest patience in providing me with his time and brilliant statistical guidance, shaping many of my work in this thesis and making statistics so easy to understand.

A special thank you to Dr Kriscia Tapia, BREAST and Medical Image Optimisation and Perception Group (MIOPeG) team for their incredible support and knowledge sharing throughout my PhD.

I would like to sincerely thank the Australian Government for their support in offsetting the tuition fees of my PhD through the Research Training Program Fees Offset. Without this assistance, I would not have been able to pursue my studies.

A heartfelt thanks also to my colleagues and friends at the University of Sydney, University of Technology Sydney and around the world for their unwavering support and belief in me.

Finally, and most importantly, to my wonderful family, thank you in particular, my husband, Sai, my boys, Maxson and Deason, and my parents, Khamfong and Mounly, for their unconditional love, relentless support, and encouragement throughout these years. Thank you for taking care of me and giving me the time and support I need to fulfill my dream. This thesis holds as much significance to me as it does to my husband, boys, and parents.

Table of Contents

Table of Figures

Table of Tables

Publications and presentations

Numerous parts of this thesis have been published and/or presented in the following forums:

Peer reviewed journal articles

- **Siviengphanom, S.**, Z. Gandomkar, S.J. Lewis, and P.C. Brennan, *Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs.* Academic Radiology, 2021. **29**(8): p. 1228–1247. https://doi.org/10.1016/j.acra.2021.09.025

- **Siviengphanom, S.**, S.J. Lewis, P.C. Brennan, and Z. Gandomkar, *Computer-extracted global radiomic features can predict the radiologists' first impression about the abnormality of a screening mammogram.* British Journal of Radiology, 2024. **97**: p. 168–179. https://doi.org/10.1093/bjr/tqad025

- **Siviengphanom, S.**, Z. Gandomkar, S.J. Lewis, and P.C. Brennan, *Global radiomic features from mammography for predicting difficult-to-interpret normal cases.* Journal of Digital Imaging, 2023. **36**: p. 1541–1552. https://doi.org/10.1007/s10278-023-00836-7

- **Siviengphanom, S.**, P.C. Brennan, S.J. Lewis, P.D. Trieu, Z. Gandomkar. A machine learning model based on *global mammographic radiomic features can predict which normal mammographic cases radiology trainees find most difficult*. Journal of Imaging Informatics in Medicine. 2024 Under review.

Conference papers

- **Siviengphanom, S.**, Z. Gandomkar, S.J. Lewis, and P.C. Brennan. *Global mammographic radiomic signature can predict radiologists' difficult-to-interpret normal cases*. in *SPIE Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment*. 2023. San Diego, California, United States: SPIE. https://doi.org/10.1117/12.2645377

- **Siviengphanom, S.**, S.J. Lewis, P.C. Brennan, Z. Gandomkar. *Predicting the gist of breast cancer on a screening mammogram using global radiomic features*. in *SPIE Medical Imaging 2024: Image Perception, Observer Performance, and Technology Assessment.* 2024. San Diego, California, United States: SPIE. https://doi.org/10.1117/12.3005470

Oral presentations

The University of Sydney, MIOPeG Meeting 2022

- **Siviengphanom, S.**, Z. Gandomkar, S.J. Lewis, and P.C. Brennan, *Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs.* Academic Radiology, 2021. **29**(8): p. 1228–1247.

The University of Sydney, Faculty of Medicine and Health Higher Degree by Research Conference 2022

- **Siviengphanom, S.**, Z. Gandomkar, S.J. Lewis, and P.C. Brennan, *Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs.* Academic Radiology, 2021. **29**(8): p. 1228–1247.

SPIE Medical Imaging Conference 2023

- **Siviengphanom, S.**, Z. Gandomkar, S.J. Lewis, and P.C. Brennan. *Global mammographic radiomic signature can predict radiologists' difficult-to-interpret normal cases*. in *SPIE Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment*. 2023. San Diego, California, United States: SPIE. https://doi.org/10.1117/12.2645377

SPIE Medical Imaging Conference 2024

- **Siviengphanom, S.**, Z. Gandomkar, S.J. Lewis, and P.C. Brennan. *Predicting the gist of breast cancer on a screening mammogram using global radiomic features.* in *SPIE Medical Imaging 2024: Image Perception, Observer Performance, and Technology Assessment*. 2024. San Diego, California, United States: SPIE. https://doi.org/10.1117/12.3005470

RSNA Annual Meeting 2024

- **Siviengphanom, S.**, P.C. Brennan, S.J. Lewis, P. D. Trieu and Z. Gandomkar. *Enhancing radiology resident training in mammogram interpretation: leveraging global mammographic radiomic features to identify difficult cases*. in *RSNA Annual Meeting.* 2024, 1-5 December. McCormick Place, Chicago, IL, United States.

Poster presentations

NSW Cancer Conference 2022

- **Siviengphanom, S.**, Z. Gandomkar, S.J. Lewis, and P.C. Brennan, *Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs.* Academic Radiology, 2021. **29**(8): p. 1228–1247.

NSW Cancer Conference 2023

- **Siviengphanom, S.**, Z. Gandomkar, S.J. Lewis, and P.C. Brennan, *Global radiomic features from mammography for predicting difficult-to-interpret normal cases.* Journal of Digital Imaging, 2023. **36**: p. 1541–1552.

Awards

- SPIE Medical Imaging 2024 Student Conference Scholarship.
- Robert F. Wagner All-Conference Best Student Paper Award Finalist for outstanding paper. SPIE Medical Imaging, 2023.

Abstract

This thesis aims to explore the value of global mammographic radiomic features for improving the process of curating educational mammographic test sets. It first examines existing literature to identify knowledge gaps. It then studies the value of global radiomic features in predicting global gist signals (radiologists' first impression about an image), and difficult normal cases of radiologists (widely varying experience) and radiology trainees (limited skills).

Three original research studies with ethics approval were included. **First study** employed 4191 mammograms to collect gist signals, from 13 radiologists, which were averaged and given to each image. Images were grouped into *high-* and *low-gist*. 130 global radiomic features per image were extracted and used to build eight separate machine learning classifiers for characterizing *high-gist* images. **Second study**, 361 normal cases interpreted by 537 readers were classed into highly *difficult-* and *easy-to-interpret*. 34 global radiomic features per image were extracted and used to create three classifiers to predict *difficult-to-interpret* cases. **Third study** involved grouping 280 normal cases, interpreted by 137 radiology trainees, into *hardest-* and *easiest-to-interpret*. 34 global radiomic features for each case were retrieved and used to construct a classifier for predicting *hardest-to-interpret* cases. The models were trained and validated using 10-fold cross-validation (**first study**) and leave-one-out-cross-validation approach (**second and third study**). Their performances were assessed by the area under receiver operating characteristic curve (AUC). Important features were identified via exploratory factor analysis (**all studies**).

The classifiers achieved AUCs of up to 0.84 for characterizing *high-gist* images (**first study**) and 0.71 for predicting *difficult-to-interpret* normal cases of radiologists (**second study**), and 0.75 AUC for predicting *hardest-to-interpret* normal cases of radiology trainees (**third study**). Overall, eleven important features were *cluster shade, sum entropy*, *difference variance*, *correlation*, *homogeneity, cluster prominence, standard deviation, skewness, kurtosis, range,* and *coarseness*.

The findings offer novel insights into possible appearances of global mammographic radiomic features that underlie radiological errors and difficult normal cases when interpreting screening mammograms. This data could be used to enhance educational mammographic test sets.

Chapter 1 – Introduction

1.1 Background

Breast cancer is the most commonly occurring and the primary cause of cancer-related death in women around the world [1]. In 2020, more than 2.2 million new cases of breast cancer were diagnosed and almost 700,000 deaths were reported [1-4]. Despite the notable evolution in screening techniques, technological advancement in image acquisition, and innovative therapies for breast cancer, there is a persistent and concerning trend of increased incidence and mortality rates across numerous countries globally [5-12]. Early detection of breast cancer through screening mammography, however, has proven to be one of the most effective ways to reduce the breast cancer mortality rate [13-17]. In the rest of this section, a brief background about screening mammography and its challenges will be presented, and the role of educational materials and self-assessment modules to improve diagnostic accuracy in interpreting mammograms will be the focus.

1.1.1 Screening mammography: challenges and strategies for improving interpretation accuracy

For many decades, screening mammography has been the only prevalent screening method of breast imaging for early detection and management of breast cancer, proven to significantly reduce mortality by up to 40% [8, 12, 18-20]. Detecting breast cancer early via screening mammography therefore remains a crucial cancer control strategy as it contributes to better outcomes in treated patients with potential survival gain of at least 5 years after the diagnosis [13-15].

Whilst screening mammography has many benefits, like most screening tests, it is not completely accurate as its efficacy depends greatly on the mammographic interpretation and perceptions of radiologists which can vary widely [21-23]. Differences in skill and experience levels among radiologists can lead to errors in mammographic interpretation, potentially undermining the efficacy of breast cancer screening outcomes while incurring significant clinical and economic consequences [23-25]. Numerous studies have reported considerable prevalence of missed cancer cases (i.e., up to 50% of false negative errors) alongside instances where normal cancer-free mammograms were incorrectly identified as cancers (i.e., up to 16% of false positive recall rates and 63% of cumulative chance of a false positive reading following ten rounds of biennial breast screening) [23, 26-32]. In addition, false negative errors may lead

to treatment difficulties and decreased overall survival owing to the delayed detection of breast cancer [33, 34]. False positive misdiagnoses, on the other hand, can cause enormous yearly expenses to the healthcare system (such as over US$2.5 billion in the United States) [35] while increasing patient depression and concerns due to unnecessary invasive procedures and/or overdiagnoses [36-38]. Although various strategies like double or second readers (external radiologist or Artificial Intelligence/AI) have been proposed to reduce errors in mammography interpretation, these strategies still face issues with variability in readers' experience, leading to significant inter-reader disagreement and dismissal of AI marked lesions [23, 39]. It is therefore imperative to consistently evaluate and enhance radiologists' proficiency in accurately interpreting screening mammographic images so that diagnostic errors can be drastically minimised.

Although clinical audit programs are commonly used to evaluate cancer detection performance of radiologists engaged in mammography screening, radiologists often face delays in receiving timely feedback on the efficacy of their interpretation [25, 40-42]. The low prevalence rate of breast cancer in the screened population also restricts radiologists' exposure to interpreting cancer cases while extending the time (to 2 years) required to accumulate sufficient data for meaningful assessment of their performances [42-44]. The challenge of defining the total number of cancer and normal cases within audit programs also hampers the precise determination of individual radiologist's sensitivity and specificity [41, 45, 46]. Hence, there is a serious demand for more immediate and responsive systems to better evaluate radiologists' performance when interpreting screening mammograms.

One effective approach to regularly monitor and evaluate radiologists' performance is through self-assessment educational test sets. Numerous educational and training programs such as Breastscreen REader Assessment STrategy (BREAST), Personal Performance in Mammographic Screening (PERFORMS), and DetectED-X have been established to assess and improve clinical radiologists' performances supplemental to the clinical audits [25, 47-51]. These educational platforms generally rely on providing self-assessment educational test sets for radiologists seeking to monitor and identify reasons for mammographic errors while offering immediate feedback upon completion of the test set to minimise these mistakes [25, 47, 48, 51]. Self-guided educational approaches such as self-assessment educational test sets have proven to be an effective educational tool and are popular among radiologists and radiology trainees (i.e. radiology residents) [25, 41, 49, 52-55]. As an example, since the

establishment of BREAST in 2011, over 80% of Australian radiologists (including breast physicians and radiology trainees) across all states and territories as well as radiologists from New Zealand, Asia, Middle East, and Europe have participated in the program [41, 49].

1.2 Deficiencies in literature

1.2.1 Mammogram interpretation: optimising mammography education

It has been shown that the educational test set-based programs can enhance radiologists' proficiency in interpreting mammograms [25, 41, 47, 49, 51-56]. Conversely, presently, the process of curating test sets still heavily relies on expert radiologists' subjective and manual evaluation and selection of difficult/easy cases based on the traditional human-extracted Breast Imaging and Reporting Data System (BI-RADS) breast density category and benign features [52, 57-62]. Recent studies have demonstrated that expert radiologists-determined case difficulty or high BI-RADS breast density-based case difficulty is **not a good descriptor of the actual case difficulty experienced by readers** in the mammographic educational test set [59, 63]. Furthermore, when assessing performance of radiology trainees with less mammography interpretation experience, prior work discovered that although expert-determined difficulty can predict radiology trainees' actual false negative errors, **it cannot reliably predict their false positive errors** [64, 65].

When looking at human-extracted mammographic features associated to case difficulty and radiological errors more specifically, several studies were conducted to identify what human-extracted mammographic appearances make cases difficult for radiologists [66-71]. One study recognised that at reduced recall rates, cancer cases with non-specific densities features and high BI-RADS breast density were more likely to be missed while cancers with stellate masses and low breast density were more likely to be detected/recalled [69]. However, other work [66] established the reverse indicating that lesions with low density were more likely to be missed by readers (i.e., difficult) than lesions with high density, similar to the above mentioned studies [59, 63]. Another research [68] investigated correlation between various image features (e.g., BI-RADS breast density, lesion size, and different types of shape and texture) and detection rate of breast masses. It was found that features related mainly to lesion shape (e.g., lesions with abnormal margins) and size can notably contribute to declined cancer detection [68]. Earlier work focusing on analysing morphology features also revealed that shape features of

mammographic masses were associated with challenging to classify cases like spiculated benign masses and circumscribed malignant tumours [67]. When examining error in detecting mammographic masses, two studies uncovered potential results of using BI-RADS features (such as, mass shape, mass margin, and mass density) to predict individual error made by radiology trainees or cases that they were more likely to miss when detecting mammographic masses (area under receiver operating characteristic curves/AUCs of up to 0.61) [70, 71].

Based on the above findings [59, 63-71], it could be suggested that the uniform test set curation approach based on expert-, BI-RADS breast density- or benign features-determined difficulty alone may be compromised in accurately assessing the actual difficulty of cases based on the individual's performance in the test set. Hence, it could ultimately result in a **less than optimal fulfillment of their training needs**. Moreover, the above studies focus largely on missed cancer cases while the **knowledge around what makes a normal case difficult and what causes false positive errors has not been adequately researched**. According to the holistic processing model, mammogram interpretation expertise development involves building an internal representation of normality and to have a high interpretative accuracy in addition to cancer detection rate, it is necessary to ensure the normal representation/features have been well trained and learned by readers [72-75]. Given the fact that expert-determined case difficulty is not a good predictor for false positive errors [64, 65], understanding what features constitute normal case difficulty and diagnostic errors in radiology when interpreting mammographic images is therefore critically indispensable, particularly for improving radiology training programs to advance radiologists performance. Accordingly, there is an urgent need to address this knowledge gap so that reliable and accurate test set difficulty generation can be developed while individual specific errors can be better examined and improved.

1.2.2 Radiomics: insufficient global radiomics analysis

Radiomics is a rapid emerging field in medical imaging. It focusses on utilising computer-extracted features from medical images to reveal valuable hidden quantitative image-based information which is not optimally recognised or assessed by the human eye [76]. Radiomics analysis pipeline involves extracting and analysing a vast array of quantitative features from medical images and converting them into multi-dimensional mineable data for decision support and improving accuracy of breast cancer management [77-79]. Numerous radiomics studies have also been conducted in breast cancer showing the potential usefulness of radiomics in

improving the accuracy of cancer detection, diagnosis, and prediction [76-80]. Yet, most radiomics studies predominantly address cancer prediction, diagnosis, and forecasting patient outcomes or cancer recurrence (more detailed discussion in Chapter 2 including studies summarised in Table 1) using local radiomic features (describing image sub-regions or patches). **The importance of global radiomic features (explaining the whole image) has not been well explored** [81-83].

Furthermore, only a limited number of studies concentrated on employing radiomic features (i.e., computer-extracted features) to predict the difficulty of cases. These existing studies focus primarily on using computer algorithms to predict and model error making patterns of radiology trainees [64, 70, 71, 84-88]. For example, a user model (or a computer algorithm) was developed to predict false negative errors of trainees interpreting digital breast tomosynthesis (DBT) images [84]. The model first segmented the lesions, extracted local radiomic features, and used a multivariate logistic regression classifier to predict the possibility of trainees missing an abnormal location. This approach exhibited a significant improvement in distinguishing detected and missed lesions compared to random selection. Another work related to predicting missed mammographic masses made by trainees using computer-extracted features demonstrated the existence of error making patterns among trainees [85]. Additionally, relationship between case difficulty and error rates was investigated, finding that increased difficulty, whether self-assessed or expert-assessed, correlated with lower performance in mammography interpretation among trainees [64]. Using BI-RADS features to predict mass detection errors of radiology trainees were also explored, showing that individual models could effectively predict the likelihood of the errors [70] while examining error patterns in local BI-RADS descriptor assessments highlighted the potential for personalised educational materials based on identified error patterns [71, 88].

It is important to note that all the above-mentioned studies concentrate chiefly on false negative errors [64, 70, 71, 84, 85, 88] though high performance in image interpretation heavily relies on reducing false positive errors as well. In this area, an algorithm to model radiology trainees' false positive error making patterns in reading mammograms using computer vision-based segmentation, local radiomic features, and a random forest classifier was suggested [86]. Another computer vision-based algorithm to predict their false positive errors in interpreting DBT cases based on local radiomic features and machine learning was also developed [87]. Their approach involved identifying difficult locations, segmenting potential abnormalities,

extracting local features, and building an individualised classifier to predict error for each trainee. This method proved to be effective in identifying error prone locations, significantly improving over random location selection [86, 87].

Overall, the studies employing radiomics [84-87] **largely give emphasis to local radiomic features and radiology trainees using a limited number of dataset** (n = up to 29 trainees and 3 expert radiologists, and up to 100 mammographic and 60 DBT cases). Most of the existing research also [64, 70, 71, 84-88] pay more attention to false negative errors while **false positive errors/difficult normal cases have not been well analysed.**

1.2.3 Current AI tools lacking focus on more difficult normal cases and global radiomic features

In an effort to improve mammography education, numerous studies investigated the diagnostic ability of using AI tools, including using radiomics approach, to predict readers' difficult cases and error making patterns (e.g., false positive location errors and missed mammographic masses) [85, 86, 89-92]. For instance, in the preliminary experimental work, an individualised adaptive computer-aided radiology training system based on user modelling utilising machine learning and statistical techniques was proposed [89, 90]. It was shown that machine learning user specific models were useful in predicting the level of difficulty of the biopsy-proven mass cases that radiology trainees were likely to experience based on two BI-RADS descriptors (parenchyma density and mass margin). Moreover, between these two BI-RADS features, mass margin was demonstrated to be the most useful features in predicting individual trainee errors [90].

In terms of predicting error making patterns of readers, studies employed radiomic analysis exhibited capability of local radiomic features models to predict radiology trainees' error better than chance in the detection of mammographic masses (AUCs up to 0.63) [85, 91] and their suspicious false positive errors locations (up to 40% accuracy) [86]. Another interesting study [92] exhibited that by combining visual gaze features with image features (BI-RADS and local radiomic features) using machine learning techniques, diagnostic error making pattern of readers could be modelled. It was discovered that group-based model performed reasonably well (AUC of 0.79) while personalised modelling of readers' error was more feasible for more experienced radiologists (AUC of 0.84) than for naïve radiology trainees (AUC of 0.67) [92].

Mass margin and mass shape of BI-RADS features were also found to be important descriptors for human error [92], similar to the previous study [90].

The above mentioned pilot and exploratory work, nevertheless, [85, 86, 89-92] **did not emphasise on more difficult normal cases** to supplement readers' mammogram reading ability, alert readers about the case for enhancing reading strategy (e.g., double/single reading, senior/junior reader) or improve educational test set generation more specifically [39, 60, 93, 94]. Furthermore, these studies yet again mainly concentrated on radiology trainees with a small number of readers (n = up to 10 trainees and 3 expert radiologists) and mammographic cases (n = up to 100 cases) used [85, 86, 89-92]. Once more, those studies [85, 86, 91, 92] that employed radiomics analysis for their predictive models purely put emphasis to local radiomic features and **not global radiomic features** [81-83].

1.2.4 Medical image perception: reliable gist predictor is required

Additionally, it has been established that experienced radiologists can perceive the gist of breast cancer based on their holistic quick initial impression on a mammogram after a half-second image viewing (referred to as "global gist signal") [95, 96]. More recently, evidence shows that experienced radiologists can identify the global gist signal, in prior normal mammograms of women who were diagnosed with breast cancer at a later date, suggesting potential for breast cancer risk prediction [26, 97, 98]. They can also detect the global gist signal in current normal mammograms, containing no apparent localised lesion sign, contralateral to the malignancy breast [99]. Other studies also exhibit the usefulness of global gist signal in mammogram expertise development and interpretation when reading difficult cancer cases (e.g., undetected retrospective visible cancers) and **deficiency of gist global image processing skill can result in diagnostic errors** [100-103]. Despite gist is a noisy signal and can be inconsistent both within the same observer and among different observers, the above findings evidence the great efficacy of the global gist signal in mammographic interpretation and expertise development in detecting and predicting a breast cancer [104, 105]. Consequently, **a good predictor for the global gist signal is needed**.

Given the pivotal roles played by both global gist signal in guiding visual searches and diagnostic decision-makings as well as radiomics in mammography, it is critical to explore a unique method of global radiomic analysis in this thesis for improving the process of curating educational mammographic test sets [26, 81, 95, 97, 98]. Yet there seems to be no study

conducted using global mammographic radiomic features to explain the characteristics of global gist signal, in other words if global radiomic features are the driver of global gist signal. Once this is established, further study is required to assess if these global radiomic features, driving radiologists' initial sense about the case, are drivers of more difficult normal cases readers (in general) experienced in the mammographic test sets, since gist has been exhibited to be important for interpreting difficult cancer cases and expertise development [100-103, 106]. If they are, it would also be helpful to explore if these features are predictors of more difficult normal cases for specific cohort like radiology trainees providing they have inadequate ability in processing global images and interpreting mammograms and there is a value of gist in their expertise development [26, 72, 85, 86, 95, 100, 105]. This is so that the process of curating the educational test sets can be enhanced considering nearly 70% of the total cases in many test sets are normal cancer-free [25, 54, 56] - Figure 1.2.

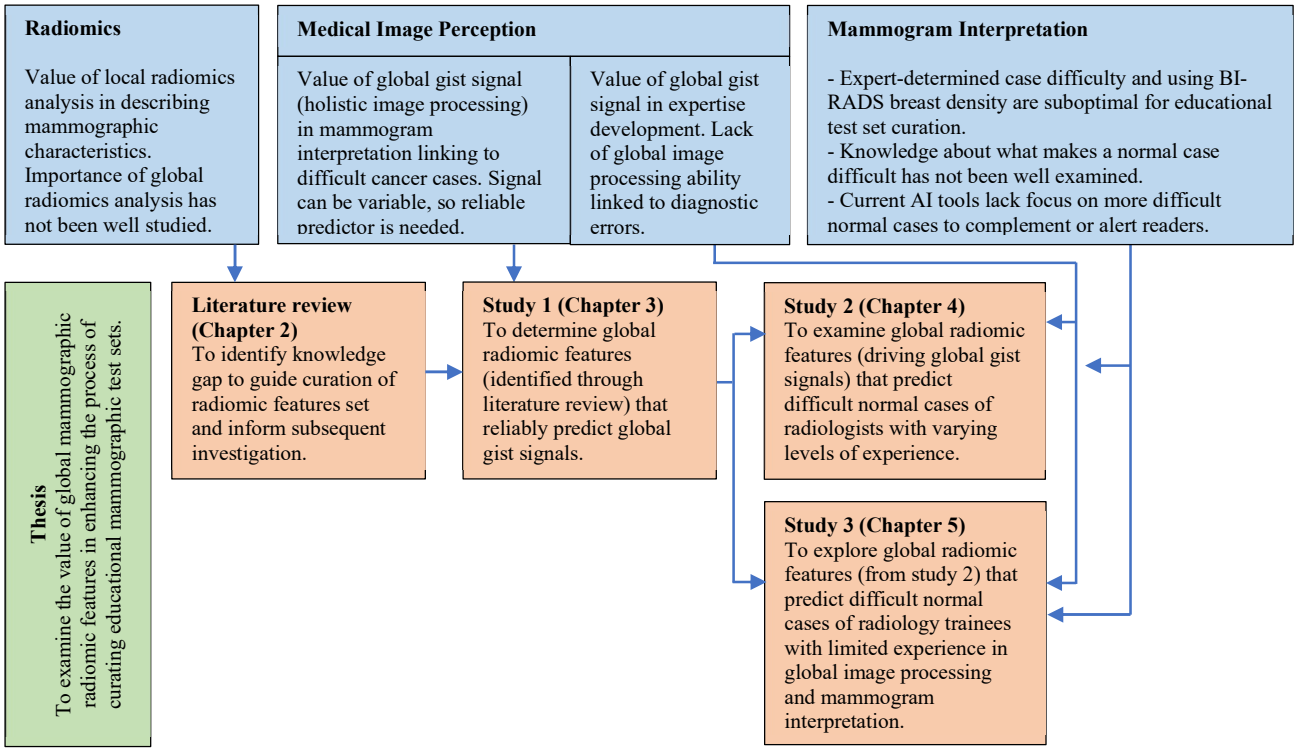


Figure 1.2. Thesis flowchart. Based on the existing knowledge and gaps (blue boxes), the thesis' aim (green) is to be achieved through the objectives of each study/chapter (orange). AI = Artificial Intelligence, BI-RADS = Breast Imaging and Reporting Data System.

1.3 Thesis' aim and objectives

Based on all the above identified knowledge deficiencies in the existing literature, the overarching aim of this thesis is thus to investigate the value of global mammographic features derived from radiomics analysis in enhancing the process of curating educational mammographic test sets (Figure 1.2).

The thesis contains four objectives.

- **Objective 1** is to examine the existing mammographic literature regarding the efficiency of radiomics analysis in breast cancer in order to identify knowledge deficiencies to direct curation of radiomic features set and inform subsequent investigations. Objective 1 is addressed and covered in Chapter 2.

- **Objective 2** is to explore the association between global mammographic radiomic features and the global gist signal and to identify important features that drive the signals. Objective 2 is addressed in Chapter 3 of this thesis.

- Once connection between global radiomic features and global gist signal (objective 2) is established, **objective 3** is to evaluate the ability of global radiomic features in predicting highly difficult normal cases of radiologists of widely varying experience, and to recognise if there are most valuable features in describing these cases while determining if BI-RADS breast density categories are correlated with difficulty scores of these cases. Objective 3 is addressed in Chapter 4.

- After objective 3 is achieved, **objective 4** is to investigate if global mammographic radiomic features can also predict greatly difficult normal cases specifically for radiology trainees who appear to have limited experience in processing global images and interpreting mammograms. It also aims to uncover whether any of these features are best in depicting these difficult cases, and if there are any correlations between BI-RADS breast density and difficulty scores of the cases in this cohort. Objective 4 of the thesis is addressed in Chapter 5.

1.4 Significance of the thesis

The outcomes of the thesis are critical in providing better understanding of the global mammographic image characteristics that drive radiological errors and difficult normal cases when interpreting screening mammographic images. These could be helpful for enhancing and generating more predictable and personalised radiology educational mammographic test sets for radiologists and particularly radiology trainees who lack adequate experience in global image processing and mammography interpretation. The findings may also be valuable for clinical mammographic screening programs in offering reliable and effective means of objectively measuring the performance of each reader and monitoring the efficacy of the screening program over time.

1.5 Thesis outline

This thesis contains six chapters and is structured in the following manner:

- **Chapter 1** (current chapter) introduces the thesis by providing a brief background on the key aspects of the topic, summarises the gaps in the literature, and explains the objectives and significance of the thesis.

- **Chapter 2** provides a detailed literature review and summary of the mammographic radiomic studies' findings related to the potential usefulness of mammography-based (including digital mammography, DBT, and contrast-enhanced mammography) radiomic features in predicting factors that describe breast cancer characteristics, recurrence, and survival. This study informs the subsequent investigation and guides the curation of a set of radiomic features from mammograms.

- **Chapter 3** shows the correlations between global radiomic features derived from mammograms and the global gist signal (i.e., accurately classifying *high-* from *low-gist* images) and identifies important features that drive the strong (*high-gist* images) from weak gist (*low-gist* images) signal.

- **Chapter 4** indicates the efficacy of global mammographic radiomic features in grouping highly *difficult-* from *easy-to-interpret* normal cases of overall population of radiologists including features that are helpful in describing these cases. It also reveals

whether there is significant difference in difficulty level between *low-BD* (breast density) and *high-BD* normal cases.

- **Chapter 5** specifically presents the extended analysis of the investigation on the significance of mammography-based global radiomic features in categorising radiology trainees' *hardest-* vs *easiest-to-interpret* normal cases with valuable features in depicting these cases identified. Furthermore, it shows whether breast density (*low-density* and *high-density*) of these two categories cases differ significantly or not.

- Finally, **Chapter 6** exhibits the holistic discussion based on the overall findings from the three original studies (Chapter 3-5) within the thesis including the implications, limitation, future directions, and conclusion of the thesis.

A bridging section/chapter introduction is included at the beginning of each chapter to show the connection between each chapter. Each chapter also concludes with a chapter findings section at the end to summarise main discovery of the chapter. Where applicable, each chapter has its own list of references. Appendices at the end of this thesis contain conference presentations and abstract corresponding to each of the chapter.

1.6 References

1. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA: A Cancer Journal for Clinicians, 2021. **71**(3): p. 209-249.
2. Ginsburg, O., et al., *The global burden of women's cancers: a grand challenge in global health.* Lancet (London, England), 2017. **389**(10071): p. 847-860.
3. Ferlay, J., et al., *Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018.* European Journal of Cancer, 2018. **103**: p. 356-387.
4. Arnold, M., et al., *Current and future burden of breast cancer: Global statistics for 2020 and 2040.* The Breast, 2022. **66**: p. 15-23.
5. Morrell, S., et al., *Mammography screening and breast cancer mortality in Australia: an aggregate cohort study.* Journal of Medical Screening, 2012. **19**(1): p. 26-34.
6. Marmot, M.G., et al., *The benefits and harms of breast cancer screening: an independent review.* British journal of cancer, 2013. **108**(11): p. 2205-2240.
7. Carioli, G., et al., *Trends and predictions to 2020 in breast cancer mortality: Americas and Australasia.* Breast, 2018. **37**: p. 163-169.
8. Njor, S., et al., *Breast cancer mortality in mammographic screening in Europe: a review of incidence-based mortality studies.* 2012. **19**(Supplement 1): p. 33-41.
9. Malvezzi, M., et al., *European cancer mortality predictions for the year 2019 with focus on breast cancer.* Annals of Oncology, 2019. **30**(5): p. 781-787.
10. Stephen, M., et al., *Mammography service screening and breast cancer mortality in New Zealand: a National Cohort Study 1999–2011.* British Journal of Cancer, 2017.
11. Nounou, M.I., et al., *Breast Cancer: Conventional Diagnosis and Treatment Modalities and Recent Patents and Technologies.* Breast cancer : basic and clinical research, 2015. **9**(Suppl 2): p. 17-34.
12. Paci, E., *Summary of the Evidence of Breast Cancer Service Screening Outcomes in Europe and First Estimate of the Benefit and Harm Balance Sheet.* Journal of Medical Screening, 2012. **19**(1_suppl): p. 5-13.
13. Australian Institute of Health Welfare (AIHW), *Cancer in Australia 2019*. 2019, Australian Institute of Health and Welfare (AIHW): Cat. no. CAN 123. Canberra.
14. Ginsburg, O., et al., *Breast cancer early detection: A phased approach to implementation.* Cancer, 2020. **126**(S10): p. 2379-2393.
15. Grimm, L.J., et al., *Benefits and Risks of Mammography Screening in Women Ages 40 to 49 Years.* J Prim Care Community Health, 2022. **13**: p. 21501327211058322.
16. Tabár, L., et al., *Reduction in mortality from breast cancer after mass screening with mammography.* The Lancet (British edition), 1985. **1**(8433): p. 829-832.
17. Tabár, L., et al., *Swedish Two-County Trial: Impact of Mammographic Screening on Breast Cancer Mortality during 3 Decades.* Radiology, 2011. **260**(3): p. 658-663.
18. Heywang-Köbrunner, S.H., A. Hacker, and S. Sedlacek, *Advantages and Disadvantages of Mammography Screening.* Breast care (Basel, Switzerland), 2011. **6**(3): p. 199-207.
19. Seely, J.M. and T. Alhassan, *Screening for breast cancer in 2018-what should we be doing today?* Current oncology (Toronto, Ont.), 2018. **25**(Suppl 1): p. S115-S124.
20. Welfare, A.I.o.H.a., *BreastScreen Australia monitoring report 2018*. 2018, AIHW: Canberra.
21. Woodard, D.B., et al., *Performance assessment for radiologists interpreting screening mammography.* Statistics in Medicine, 2007. **26**(7): p. 1532-1551.

22. Elmore, J.G., et al., *Variability in Radiologists' Interpretations of Mammograms.* New England Journal of Medicine, 1994. **331**(22): p. 1493-1499.
23. Ekpo, E.U., M. Alakhras, and P. Brennan, *Errors in Mammography Cannot be Solved Through Technology Alone.* Asian Pacific journal of cancer prevention : APJCP, 2018. **19**(2): p. 291-301.
24. Brady, A.P., *Error and discrepancy in radiology: inevitable or avoidable?* Insights Imaging, 2017. **8**(1): p. 171-182.
25. Suleiman, M.E., M. Rickard, and P.C. Brennan, *Perfecting detection through education.* Radiography, 2020. **26**: p. S49-S53.
26. Brennan, P.C., et al., *Radiologists can detect the 'gist' of breast cancer before any overt signs of cancer appear.* Scientific reports, 2018. **8**(1): p. 1-12.
27. Castells, X., et al., *Risk of Breast Cancer in Women with False-Positive Results according to Mammographic Features.* Radiology, 2016. **280**(2): p. 379-386.
28. Lewis, S.J., et al. *Characteristics of frequently recalled false positive cases in screening mammography*. in *SPIE 11513, 15th International Workshop on Breast Imaging*. 2020. Leuven, Belgium: SPIE.
29. Hovda, T., et al., *Radiological review of prior screening mammograms of screen-detected breast cancer.* Eur Radiol, 2021. **31**(4): p. 2568-2579.
30. Moberg, K., H. Grundström, and H. Lundquist, *Radiological review of incidence breast cancers.* Journal of Medical Screening, 2000. **7**(4): p. 177-183.
31. Vlahiotis, A., et al., *Analysis of utilization patterns and associated costs of the breast imaging and diagnostic procedures after screening mammography.* ClinicoEconomics and Outcomes Research, 2018. **Volume 10**: p. 157-167.
32. Krupinski, E.A., *Perceptual Factors in Reading Medical Images*, in *The Handbook of Medical Image Perception and Techniques*, E. Samei and E.A. Krupinski, Editors. 2018, Cambridge University Press: Cambridge. p. 95-106.
33. Li, Y., et al., *The influence on survival of delay in the treatment initiation of screening detected non-symptomatic breast cancer.* Scientific Reports, 2019. **9**(1): p. 1-7.
34. Wadhwa, A., J.R. Sullivan, and M.B. Gonyo, *Missed Breast Cancer: What Can We Learn?* 2016. **45**(6): p. 402-419.
35. Ong, M.-S. and K.D. Mandl, *National Expenditure For False-Positive Mammograms And Breast Cancer Overdiagnoses Estimated At $4 Billion A Year.* Health Affairs, 2015. **34**(4): p. 576-583.
36. Ho, P.J., et al., *Factors associated with false-positive mammography at first screen in an Asian population.* PLOS ONE, 2019. **14**(3): p. e0213615.
37. Alcusky, M., et al., *The Patient Burden of Screening Mammography Recall.* Journal of Women's Health, 2014. **23**(S1): p. S-11-S-19.
38. Skaane, P., *How Can We Reduce Unnecessary Procedures after Screening Mammography?* Radiology, 2019. **291**(2): p. 318-319.
39. Brennan, P.C., et al., *Benefits of Independent Double Reading in Digital Mammography.* Academic Radiology, 2019. **26**(6): p. 717-723.
40. Soh, B.P., et al., *Screening Mammography: Test Set Data Can Reasonably Describe Actual Clinical Reporting.* Radiology, 2013. **268**(1): p. 46-53.
41. Trieu, P.D., et al., *Look how far we have come: BREAST cancer detection education on the international stage.* Frontiers in Oncology, 2023. **12**.
42. Soh, B.P., et al., *Certain performance values arising from mammographic test set readings correlate well with clinical audit.* Journal of Medical Imaging and Radiation Oncology, 2015. **59**(4): p. 403-410.

43. Tapia, K., W. Lee, and P.C. Brennan, *Breast Screen Reader Assessment Strategy: Transforming Breast Cancer Diagnosis Globally: A Mini Review.* International Journal of Radiology & Radiation Therapy, 2017. **4**(1).
44. Australia, B., *Breastscreen Australia National Accreditation Standards*. 2022, Breastscreen Australia. p. 1-204.
45. Qenam, B.A., et al., *The roles of clinical audit and test sets in promoting the quality of breast screening: a scoping review.* Clinical Radiology, 2020. **75**(10): p. 794.e1-794.e6.
46. Brennan, P.C. and B.P. Soh, *The challenge of assessing reader performance in mammography.* Clin Radiol, 2012. **67**(2): p. 192.
47. Gale, A. and Y. Chen, *A review of the PERFORMS scheme in breast screening.* The British Journal of Radiology, 2020. **93**(1112): p. 20190908.
48. Brennan, P.C., et al. *BREAST: a novel method to improve the diagnostic efficacy of mammography*. in *SPIE 8673, Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment*. 2013. Lake Buena Vista (Orlando Area), Florida, United States: SPIE.
49. Brennan, P.C., et al., *BREAST: A Novel Strategy to Improve the Detection of Breast Cancer*, in *International Workshop on Digital Mammography 2014*, H. Fujita, T. Hara, and C. Muramatsu, Editors. 2014, Springer International Publishing Switzerland: Cham. p. 438-443.
50. Chen, Y. and A. Gale, *Performance Assessment Using Standardized Data Sets: The PERFORMS Scheme in Breast Screening and Other Domains*, in *The Handbook of Medical Image Perception and Techniques*, E. Samei and E.A. Krupinski, Editors. 2018, Cambridge University Press: Cambridge. p. 328-342.
51. Brancato, B., et al., *Mammography self-evaluation online test for screening readers: an Italian Society of Medical Radiology (SIRM) initiative.* Eur Radiol, 2022. **32**(3): p. 1624-1633.
52. Trieu, P.D., et al., *Improving radiologist's ability in identifying particular abnormal lesions on mammograms through training test set with immediate feedback.* Scientific Reports, 2021. **11**(9899): p. 1-8.
53. Trieu, P.D., et al., *Improvement of Cancer Detection on Mammograms via BREAST Test Sets.* Academic Radiology, 2019. **26**(12): p. e341-e347.
54. Suleiman, W.I., et al., *Impact of Breast Reader Assessment Strategy on mammographic radiologists' test reading performance.* Journal of Medical Imaging and Radiation Oncology, 2016. **60**(3): p. 352-358.
55. Brennan, P., W. Lee, and K. Tapia, *Breast Screen Reader Assessment Strategy (BREAST): A Research Infrastructure with a Translational Objective*, in *The Handbook of Medical Image Perception and Techniques*, E. Samei and E.A. Krupinski, Editors. 2018, Cambridge University Press: Cambridge. p. 343-356.
56. Chen, Y., et al., *The Relationship between Mammography Readers' Real-Life Performance and Performance in a Test Set–based Assessment Scheme in a National Breast Screening Program.* Radiology: Imaging Cancer, 2020. **2**(5): p. e200016.
57. Dong, L., Y. Chen, and A. Gale. *Breast screening: understanding case difficulty and the nature of errors*. in *SPIE 8673, Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment*. 2013. Lake Buena Vista (Orlando Area), Florida, United States: SPIE.
58. Scott, H.J. and A.G. Gale. *Breast screening technologists: When is a difficult case truly difficult and for whom?* in *SPIE 5749, Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment*. 2005. Bellingham, WA: SPIE.

59. Wong, D., et al., *Mammographic image characteristics in educational test sets: the effect on expert difficulty ratings and subsequent reader performance.* Asia Pacific Journal of Oncology, 2024. **Under review**.
60. Ang, Z.Z., et al., *Classification of normal screening mammograms is strongly influenced by perceived mammographic breast density.* Journal of Medical Imaging and Radiation Oncology, 2017. **61**(4): p. 461-469.
61. Darker, I., Y. Chen, and A. Gale. *Health professionals' agreement on density judgements and successful abnormality identification within the UK Breast Screening Programme*. in *SPIE Medical Imaging*. 2011. SPIE.
62. Sun, S., et al., *Individualised training to address variability of radiologists' performance*. Medical Imaging. Vol. 6917. 2008: SPIE.
63. Wong, D.J., *Association of diagnostic errors in interpreting screening mammograms with image and reader characteristics*. 2023.
64. Grimm, L.J., et al., *Radiology Resident Mammography Training.* Academic Radiology, 2014. **21**(7): p. 888-892.
65. Mazurowski, M.A. *Difficulty of mammographic cases in the context of resident training: preliminary experimental data*. in *SPIE 8673, Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment* 2013. Lake Buena Vista (Orlando Area), Florida, United States: SPIE.
66. Goergen, S.K., et al., *Characteristics of breast carcinomas missed by screening radiologists.* Radiology, 1997. **204**(1): p. 131-135.
67. Rangayyan, R.M., N.R. Mudigonda, and J.E.L. Desautels, *Boundary modelling and shape analysis methods for classification of mammographic masses.* Medical and Biological Engineering and Computing, 2000. **38**(5): p. 487-496.
68. Rawashdeh, M.A.M., et al., *Quantitative Measures Confirm the Inverse Relationship between Lesion Spiculation and Detection of Breast Masses.* Academic radiology, 2013. **20**(5): p. 576-580.
69. Mohd Norsuddin, N., et al., *An investigation into the mammographic appearances of missed breast cancers when recall rates are reduced.* British Journal of Radiology, 2017. **90**(1076): p. 20170048.
70. Grimm, L.J., et al., *Predicting error in detecting mammographic masses among radiology trainees using statistical models based on BI-RADS features.* Medical Physics, 2014. **41**(3): p. 031909.
71. Mazurowski, M.A., et al., *Identifying Error-making Patterns in Assessment of Mammographic BI-RADS Descriptors among Radiology Residents Using Statistical Pattern Recognition.* Academic Radiology, 2012. **19**(7): p. 865-871.
72. Gandomkar, Z. and C. Mello-Thoms, *Visual search in breast imaging.* The British Journal of Radiology, 2019. **92**(1102): p. 20190057.
73. Sheridan, H. and E.M. Reingold, *The Holistic Processing Account of Visual Expertise in Medical Image Perception: A Review.* Frontiers in Psychology, 2017. **8**.
74. Kundel, H., *How to minimize perceptual error and maximize expertise in medical imaging*. Medical Imaging. Vol. 6515. 2007: SPIE.
75. Kundel, H.L., et al., *Holistic Component of Image Perception in Mammogram Interpretation: Gaze-Tracking Study.* Radiology, 2007. **242**: p. 396-402.
76. Parekh, V. and M.A. Jacobs, *Radiomics: a new application from established techniques.* Expert Review of Precision Medicine and Drug Development, 2016. **1**(2): p. 207-226.
77. Gillies, R.J., P.E. Kinahan, and H. Hricak, *Radiomics: Images Are More than Pictures, They Are Data.* Radiology, 2016. **278**(2): p. 563-577.

78. Lambin, P., et al., *Radiomics: Extracting more information from medical images using advanced feature analysis.* European Journal of Cancer, 2012. **48**(4): p. 441-446.
79. Kumar, V., et al., *Radiomics: the process and the challenges.* Magnetic Resonance Imaging, 2012. **30**(9): p. 1234-1248.
80. Conti, A., et al., *Radiomics in breast cancer classification and prediction.* Seminars in Cancer Biology, 2020. **72**: p. 238-250.
81. Siviengphanom, S., et al., *Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs.* Academic Radiology, 2021. **29**(8): p. 1228-1247.
82. Kabbai, L., M. Abdellaoui, and A. Douik, *Image classification by combining local and global features.* The Visual Computer, 2019. **35**(5): p. 679-693.
83. Roy, S., et al., *Optimal co-clinical radiomics: Sensitivity of radiomic features to tumour volume, image noise and resolution in co-clinical T1-weighted and T2-weighted magnetic resonance imaging.* EBioMedicine, 2020. **59**: p. 102963.
84. Wang, M., et al., *Predicting false negative errors in digital breast tomosynthesis among radiology trainees using a computer vision-based approach.* Expert Systems with Applications, 2016. **56**: p. 1-8.
85. Zhang, J., et al., *Using computer-extracted image features for modeling of error-making patterns in detection of mammographic masses among radiology residents.* Medical Physics, 2014. **41**(9): p. 091907.
86. Zhang, J., J.I. Silber, and M.A. Mazurowski, *Modeling false positive error making patterns in radiology trainees for improved mammography education.* Journal of Biomedical Informatics, 2015. **54**: p. 50-57.
87. Wang, M., et al., *A computer vision-based algorithm to predict false positive errors in radiology trainees when interpreting digital breast tomosynthesis cases.* Expert Systems with Applications, 2016. **64**: p. 490-499.
88. Mazurowski, M. and G. Tourassi. *Modeling error in assessment of mammographic image features for improved computer-aided mammography training: initial experience*. in *SPIE Medical Imaging*. 2011. SPIE.
89. Mazurowski, M., J. Lo, and G. Tourassi. *User modeling for improved computer-aided training in radiology: initial experience*. in *SPIE Medical Imaging*. 2010. SPIE.
90. Mazurowski, M.A., et al., *Individualized computer-aided education in mammography based on user modeling: Concept and preliminary experiments.* Medical Physics, 2010. **37**(3): p. 1152-1160.
91. Mazurowski, M., et al. *Modeling resident error-making patterns in detection of mammographic masses using computer-extracted image features: preliminary experiments*. in *SPIE Medical Imaging*. 2014. SPIE.
92. Voisin, S., et al., *Predicting diagnostic error in radiology via eye-tracking and image analytics: Preliminary investigation in mammography.* Medical Physics, 2013. **40**(10): p. 101906.
93. Taylor-Phillips, S. and C. Stinton, *Double reading in breast cancer screening: considerations for policy-making.* The British Journal of Radiology, 2020. **93**(1106): p. 20190610.
94. Gandomkar, Z., et al., *Can eye-tracking metrics be used to better pair radiologists in a mammogram reading task?* Medical Physics, 2018. **45**(11): p. 4844-4856.
95. Evans, K.K., et al., *The gist of the abnormal: Above-chance medical decision making in the blink of an eye.* Psychonomic Bulletin & Review, 2013. **20**(6): p. 1170-1175.

96. Mugglestone, M.D., et al. *Diagnostic performance on briefly presented mammographic images*. in *Proc.SPIE 2436, Medical Imaging 1995: Image Perception*. 1995. San Diego, CA, United States: SPIE.
97. Gandomkar, Z., et al. *Detection of the abnormal gist in the prior mammograms even with no overt sign of breast cancer*. in *SPIE 10718, 14th International Workshop on Breast Imaging (IWBI 2018)*. 2018. Atlanta, Georgia, United States: SPIE.
98. Evans, K.K., A.-M. Culpan, and J.M. Wolfe, *Detecting the "gist" of breast cancer in mammograms three years before localized signs of cancer are visible.* The British Journal of Radiology, 2019. **92**(20190136): p. 1-10.
99. Evans, K.K., et al., *A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast.* Proceedings of the National Academy of Sciences, 2016. **113**(37): p. 10292-10297.
100. Park, H.-B., et al., *Contributions of global and local processing on medical image perception.* Journal of Medical Imaging, 2023. **10**(S1): p. S11911.
101. Nodine, C.F., et al., *Blinded Review of Retrospectively Visible Unreported Breast Cancers: An Eye-Position Analysis.* Radiology, 2001. **221**(1): p. 122-129.
102. Gandomkar, Z., et al., *Reliability of radiologists' first impression when interpreting a screening mammogram.* PLOS ONE, 2023. **18**(4): p. 1-19.
103. Gandomkar, Z., et al. *False-negative diagnosis might occur due to absence of the global radiomic signature of malignancy on screening mammograms*. in *SPIE Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment*. 2023. San Diego, California, United States: SPIE.
104. Gandomkar, Z., et al. *An end-to-end deep learning model can detect the gist of the abnormal in prior mammograms as perceived by experienced radiologists*. in *SPIE Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*. 2021. San Diego, California, United States: SPIE.
105. Gandomkar, Z., et al., *Global processing provides malignancy evidence complementary to the information captured by humans or machines following detailed mammogram inspection.* Scientific Reports, 2021. **11**(20122): p. 1-12.
106. Gandomkar, Z., et al. *Does the strength of the gist signal predict the difficulty of breast cancer detection in usual presentation and reporting mechanisms?* in *SPIE 10952 Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment*. 2019. San Diego, California, United States: SPIE.

Chapter 2 – Literature review

2.1 Chapter introduction

Chapter 2 (current chapter) focusses on **objective 1** of the thesis (Chapter 1) which is to conduct a literature review on the efficiency of radiomics analysis of mammographic images (including digital mammography, DBT, and contrast-enhanced mammography) in describing breast cancer's characteristics, recurrence, and survival. The purpose of this review is to identify knowledge deficiencies to guide curation of radiomic features set and inform subsequent investigation.

The literature review in this chapter has been published as:

Siviengphanom, S., Z. Gandomkar, S.J. Lewis, and P.C. Brennan, *Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs.* Academic Radiology, 2021. **29**(8): p. 1228–1247. https://doi.org/10.1016/j.acra.2021.09.025

Presentations on this paper were made at the following meetings:

- The University of Sydney, Medical Image Optimisation and Perception Group (MIOPeG), Sydney, Australia, April 2022 (Appendix 1.1).
- The University of Sydney, Faculty of Medicine and Health Higher Degree by Research Conference, Sydney, Australia, July 2022 (Appendix 1.1).
- New South Wales (NSW) Cancer Conference, Sydney, Australia, September 2022 (Appendix 1.2).

In addition to the above literature review, a recent literature search has been conducted on 02 July 2024 to confirm that no studies have been done comparable to this thesis using criteria similar to the published review above. Articles with any study design published between 20 December 2020 and 02 July 2024 were searched through the PubMed database using the following keywords: (radiomic and breast cancer) OR (radiomic and breast imaging) OR (radiomic and breast mammography) OR (radiomic and breast tomosynthesis) OR (breast mammography and texture analysis) OR (breast mammography and quantitative analysis) OR (breast tomosynthesis and texture analysis) OR (breast tomosynthesis and quantitative analysis). 935 articles were retrieved, and their titles, abstracts and full texts were screened and

examined, the search found no new studies published with original knowledge similar to the studies in this thesis, indicating this thesis is novel.

PDF of original paper published in Academic Radiology. Siviengphanom, S., Z. Gandomkar, S.J. Lewis, and P.C. Brennan, Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs. Academic Radiology, 2021. 29(8): p. 1228–1247. https://doi.org/10.1016/j.acra.2021.09.025

Special Review

Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs

Somphone Siviengphanom, MEd, Ziba Gandomkar, PhD, Sarah J. Lewis, PhD, Patrick C. Brennan, PhD

Rationale and Objectives: Breast cancer is a highly complex heterogeneous disease. Current validated prognostic factors (e.g., histological grade, lymph node involvement, receptor status, and proliferation index), as well as multigene tests (e.g., Oncotype DX and PAM50) are helpful to describe breast cancer characteristics and predict the chance of recurrence risk and survival. Nevertheless, they are invasive and cannot capture a complete heterogeneity of the entire breast tumor resulting in up to 30% of patients being either over- or under-treated for breast cancer. Furthermore, multigene testings are time consuming and expensive. Radiomics is emerging as a reliable, accurate, non-invasive, and cost-effective approach of using quantitative image features to classify breast cancer characteristics and predict patient outcomes. Several recent radiomics reviews have been conducted in breast cancer, however, specific mammography-based radiomics studies have not been well discussed. This scoping review aims to assess and summarize the current evidence on the potential usefulness of mammography-based (i.e., digital mammography, digital breast tomosynthesis, and contrast-enhanced mammography) radiomics in predicting factors that describe breast cancer characteristics, recurrence, and survival.

Materials and Methods: PubMed database and eligible text reference were searched using relevant keywords to identify studies published between 2015 and December 19, 2020. Studies collected were screened and assessed based on the inclusion and exclusion criteria.

Results: Eighteen eligible studies were included and organized into three main sections: radiomics predicting breast cancer characteristics, radiomics predicting breast cancer recurrence and survival, and radiomics integrating with clinical data. Majority of publications reported retrospective studies while three studies examined prospective cohorts. Encouraging results were reported, suggesting the potential clinical value of mammography-based radiomics. Further efforts are required to standardize radiomics approaches and catalogue reproducible and relevant mammographic radiomic features. The role of integrating radiomics with other information is discussed.

Conclusion: The potential role of mammography-based radiomics appears promising but more efforts are required to further evaluate its reliability as a routine clinical tool.

Key Words: Radiomics; mammography; breast cancer; prognostic factors.

Abbreviations: AI artificial intelligence, **ALNM** axillary lymph node metastasis, **BC** Breast cancer, **CC** craniocaudal, **CEM** contrast-enhanced mammography, **CESM** contrast-enhanced spectral mammography, **DBT** digital breast tomosynthesis, **DCE-MRI** dynamic contrast-enhanced magnetic resonance imaging, **DL** deep learning, **DLR** deep learning radiomics, **DM** digital mammography, **ER** estrogen receptor, **FOS** first order statistics, **GLCM** gray level co-occurrence matrix, **GLRLM** gray level run length matrix, **HER2** human epidermal growth factor receptor 2, **HR** hormone receptor, **iDFS** invasive disease free survival, **ML** machine learning, **MLO** mediolateral oblique, **MRI** magnetic resonance imaging, **PR** progesterone receptor, **ROI** region of interest, **TILs** tumour infiltrating lymphocytes, **TN** triple negative, **TNBC** triple negative breast cancer, **TP** triple positive, **VOI** volume of interest

INTRODUCTION

Breast cancer (BC) is a highly complex heterogeneous disease and includes a range of many subtypes with distinct biological features that leads to diverse clinical appearances, treatment responses and survival outcomes (1,2). The classical factors that describe BC characteristics such as histological grade, lymph node invasion status, immunohistochemical receptor status, and proliferation index are well established and considered important clinical features for predicting the likelihood of recurrence risk, and response to therapy for patients with BC (3,4). However, these classifications are limited since they cannot accurately capture a complete heterogeneity of the entire breast tumor (5), which cause a substantial number (i.e., 30%) of patients being either over- or under-treated for BC (6,7). Numerous multigene tests (e.g., Oncotype DX) have also been developed to better predict BC outcomes. These genomic assays provide many

Acad Radiol 2022; 29:1228–1247

From the Discipline of Medical Imaging Sciences, Sydney School of Health Sciences, Faculty of Medicine and Health, University of Sydney, Level 7, Susan Wakil Health Building D18, Sydney, NSW 2006, Australia.. Received August 10, 2021; revised September 14, 2021; accepted September 26, 2021. **Address correspondence to:** S.S. e-mail: ssiv6387@uni.sydney.edu.au

https://doi.org/10.1016/j.acra.2021.09.025

benefits in predicting the chance of distant recurrence and the magnitude of chemotherapy benefit for patients with BC (3,8,9). Nevertheless, they rely on invasive selective tissue samples of malignancy obtained through biopsy or surgery, which is still incapable of capturing comprehensive information on the entire tumor (10,11). These assays are also expensive (e.g., oncotype DX costs around AUD$5000) and time consuming with a several week delay from the time of sample biopsy or surgery to test results (12–15). There is therefore a strong claim for the development of an alternative method that is reliable, accurate, non-invasive, and cost-effective to better describe BC characteristics and accurately predict the chance of recurrence and survival.

Radiomics, a growing research interest in extracting and analysing large number of quantitative features from medical images converting them into multi-dimensional mineable data, shows good potential for stratifying BC characteristics and predicting BC patient outcome variables (16–19). These data can be used to improve patient management as they allow an in-depth evaluation of the clinical and pathophysiology characteristics of breast tumors (17–22). Radiomics, in a non-invasive and cost-effective manner, reveals tumor heterogeneity and underlying alterations in the tumor histological anatomy, that are difficult to quantitatively identify by the human eye (17,19,21,23). Radiomic features (relating to tumor size, shape, location, intensity histogram, and texture) uncover tumor phenotype connected to its genotypic characteristics, which can be used to improve the accuracy of tumor diagnosis, prognosis, and prediction of recurrence (21,24–26). A combined set of radiomic features (i.e., radiomic signature of tumor) therefore can provide a comprehensive tumor characterization which may be powerful enough to transform clinical management of BC (17,27–33).

Radiomics methodology, either conventional or deep learning (DL) techniques, is a complex process and involves several essential steps - Figure 1. Detailed reviews of the radiomics process are available in Gillies et al. (17,34), Lee et al. (16), Tagliafico et al. (27), Lambin et al. (19), Conti et al. (35), van Timmeren et al. (36), Kocak et al. (37), and Pang et al. (38). Briefly, radiomics have been developed broadly via conventional or handcrafted features approach which mainly extracts and selects features from two-dimensional (2D) region of interest (ROI) or three-dimensional (3D) volume of interest (VOI) after acquiring high-quality images from single or multiple medical imaging modalities, and segmenting the lesion. The lesion segmentation step can be performed either manually by expert radiologists, or otherwise using semi- or fully-automatic feature segmentation algorithms (i.e., artificial intelligence/AI algorithms such as conventional machine learning/ML and DL). The ROI/VOI can include either whole tumor or subregions/subvolumes within the breast tumor.

Extracted handcrafted radiomic features, broadly categorized into four types: (1) *Morphology*, (2) *First order statistics (FOS)*, (3) *Texture,* and (4) *Transformed-based* features, are then selected or grouped and used to construct a model to predict outcomes either alone or integrating with additional information such as clinical or genomic data. Although the handcrafted features methods may use some of the conventional AI algorithms for some of the tasks (e.g., automatic lesion segmentation and classification), they generally rely on experts' predefined features. In DL radiomics (DLR), however, newer end-to-end techniques are driven by input data (e.g., medical images) which can fully and automatically segment, extract, select, and learn features (i.e., deep features) directly from images themselves, and perform prediction simultaneously generating predictive outcomes (i.e., output data) through AI DL algorithms (such as convolutional neural networks) without a need for any human involvement (16,17,19,20,27,34–40).

Whilst application of the computer extracted quantitative features from medical images to detect abnormalities dated back to around 5 decades ago (39–42), the term radiomics was introduced for oncology studies in 2012 (17–21,36). Recent rapid technological advancements especially in the field of AI applied to image analysis (such as ML and DL algorithms), and availability of large imaging dataset have made

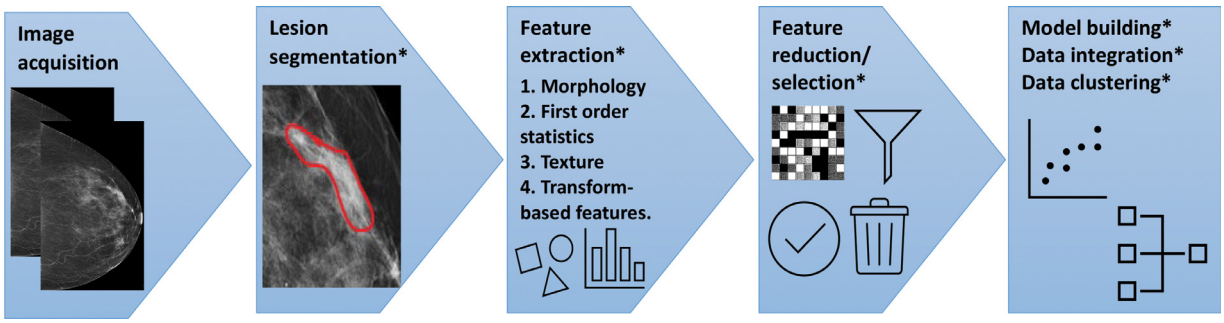


Figure 1. Radiomics workflow. In the conventional radiomics approaches, the first step of radiomics process is to acquire high-quality images from single or multiple medical imaging modalities. Next, a region of interest (ROI) or volume of interest (VOI) is outlined and segmented either manually, semi-, or fully-automatically. Handcrafted radiomic features such as morphology, first order statistics, texture, and transform-based, are then computed and extracted from the ROI/VOI. The extracted handcrafted features are then selected or grouped and used to construct a model to predict outcomes either alone or integrating with additional information such as clinical or genomic data. *In deep learning radiomics, the last four steps after image acquisition (i.e., input data) are automatically performed through artificial intelligence deep learning algorithms to extract and learn features from images and produce predictive outcomes (i.e., output data). (Color version of figure is available online.)

radiomics analysis more feasible (17-19,37). Radiomics has been applied to lung and head-and-neck imaging since 2014 and more recently actively applied to breast imaging (43-46). Whilst radiomics is still at a relatively early stage in BC, it presents potential for facilitating the advancement towards cancer personalized medicine (21,35,47). Breast imaging techniques such as digital mammography (DM), digital breast tomosynthesis (DBT), and contrast-enhanced mammography (CEM)/contrast-enhanced spectral mammography (CESM) have played pivotal role in BC screening, detection, diagnosis, and staging (48,49). Although previous studies have focused on using computer extracted quantitative features from DM, DBT, and CEM/CESM for BC detection, many articles dedicated to reviewing recent advances, were mainly driven by DL-based (50-54) and traditional computer-aided detection tools relying on handcrafted radiomic features (55-59). Several recent radiomics reviews have also been conducted in BC (16,24,27,35,60-62) including a recent study on the role of radiomics in BC risk assessment and a future BC prediction (63). However, the potential role of radiomic analysis based on mammography (i.e., DM, DBT, and CEM/CESM) specifically around predicting BC characteristics, risk of recurrence, and survival has not been well discussed. This scoping review paper therefore aims to systematically evaluate and summarize the current knowledge in this area as well as identify existing knowledge gaps to highlight and inform future research needs.

MATERIALS AND METHODS

This scoping review was conducted within the guidelines of the scoping review methodology framework and recommendations, and the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist (64-68). Due to a diverse range of study designs and methods used in the literature of mammography-based (i.e., DM, DBT, and CEM/CESM) radiomic analysis in BC, this scoping review was undertaken to address this wide study area (67,68).

The focus of this review was to synthesize and summarize the current evidence on the value of mammography-based (i.e., DM, DBT, and CEM/CESM) radiomics in predicting BC characteristics, risk of recurrence, and survival. Articles published between January 1, 2015 and December 19, 2020 with any study design were searched through the PubMed database using the following keywords: (radiomic and breast cancer) OR (radiomic and breast imaging) OR (radiomic and breast mammography) OR (radiomic and breast tomosynthesis) OR (breast mammography and texture analysis) OR (breast mammography and quantitative analysis) OR (breast tomosynthesis and texture analysis) OR (breast tomosynthesis and quantitative analysis). To ensure optimal inclusiveness of the search strategy, we also searched through eligible text reference to identify additional relevant cited studies for potential inclusion that were not identified in our initial search.

Studies were included if they meet the following inclusion criteria: (1) application of mammography-based (i.e., DM, DBT, and CEM/CESM) radiomics to predict BC prognostic factors that describe, (2) BC characteristics, (3) recurrence, and (4) survival. Studies were ineligible if they used imaging modalities other than mammography such as magnetic resonance imaging (MRI), or if they were not purely related to radiomics and BC classification or prediction specifically, including radiomics and BC diagnosis in differentiating malignancy and benign lesions. Indeed, the role of mammographic radiomics for diagnosis and detection has been well reviewed elsewhere in Conti et al. (35), Lee et al. (16), Crivelli et al. (24), Pang et al. (38), and Tagliafico et al. (27). Comments, conference proceedings, editorials, non-English articles, review articles, and case reports were ineligible for inclusion.

To identify potentially eligible studies, the literature search was performed by one author (S.S.), and titles and abstracts screening were performed by two authors (S.S. and Z.G.) in consensus. Selection of eligibility studies based on the above-mentioned criteria is presented in Figure 2.

To summarize and compare the study characteristics and findings, the following relevant data were extracted into an evidence table from each study (where available): reference (first author and publication year), study designation, population size, type of imaging modality, software used for radiomic features extraction, number of radiomic features extracted, radiomic features class extracted, number of radiomics features selected, prediction category, and key findings the study. Two authors (S.S and Z.G.) analysed and interpreted the data in agreement. The data gathered in the evidence table were used to define the main research themes for this scoping review.

RESULTS

A total of 18 eligible original articles (Figure 2) were included in this review and a summary of their characteristics, radiomics approach implementation, and findings is presented in Table 1. One study by Tamez-Peña et al. (69) examined radiogenomics, and eight studies focused on combining radiomics with clinical data - Table 2. The majority of studies (except La Forgia et al. (29) study design not available) were of a single institutional (n = 15) retrospective design (n = 14) using publicly available or institutional image datasets and small sample size (n = 40 to 394), however there were exceptions: Mao et al. (70) reported on a retrospective cohort study from two hospitals; Tagliafico et al. (71) and Tamez-Peña et al. (69) focused a single institutional prospective study; Tagliafico et al. (26) described a multi-centred trial study. All 18 studies were based on handcrafted features approaches using 2D single modality images from DM (n = 11), DBT (n = 3) and CEM/CESM (n = 4).

Based on the aim of this scoping review, we organized articles according to the below three main themes. Some literatures have overlap between two or more categories:

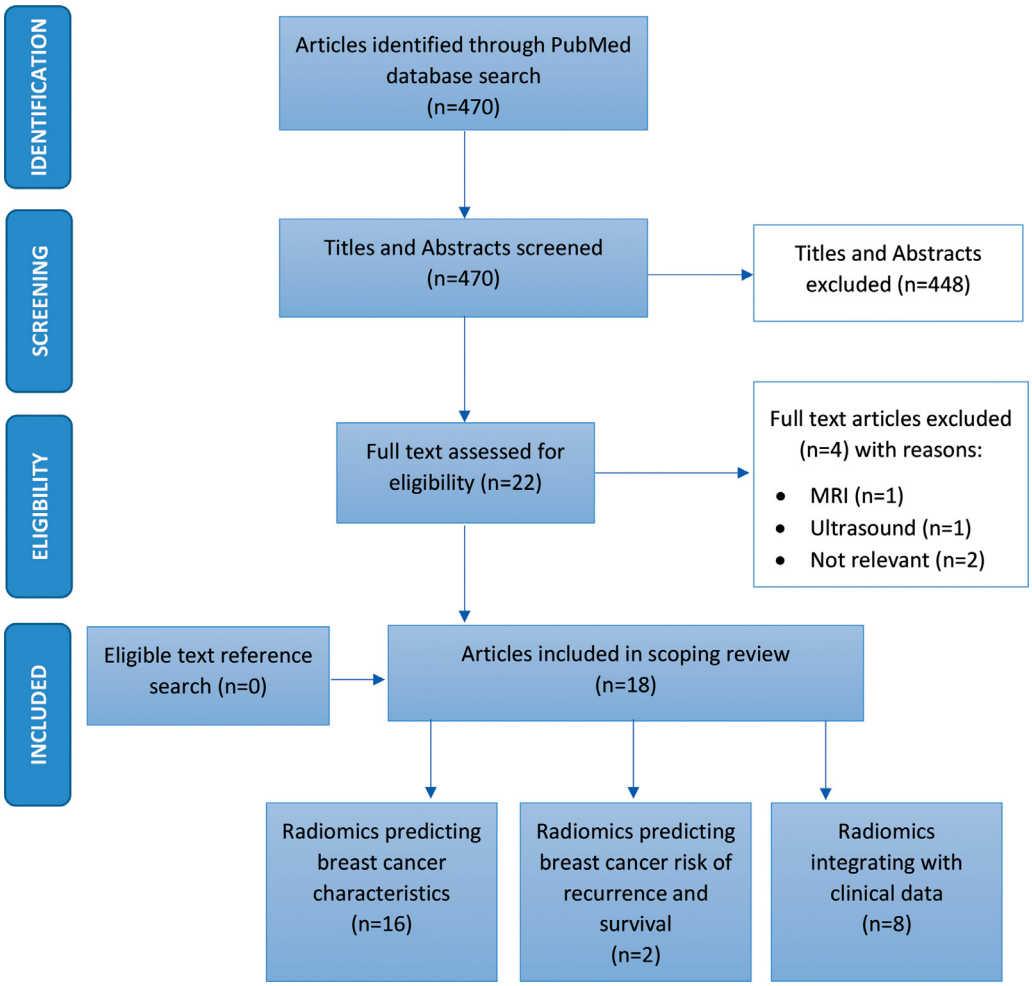


Figure 2. Results of literature search flow diagram. (Color version of figure is available online.)

- Sixteen articles assessed radiomics ability in predicting BC characteristics such as histological grade (n = 4), axillary lymph node metastasis/ALNM (n = 3), invasiveness (n = 3), receptor status (n = 5), Ki-67 (n = 2), molecular subtypes (n = 4), tumor infiltrating lymphocytes/TILs (n = 1) - Table 1;
- Two studies examined the radiomic value in predicting BC risk of recurrence (n = 1) and survival (n = 1) - Table 1; and
- Eight studies presented in Table 2 investigated the added value of radiomics when combined with clinical data in predicting factors such as histological grade (n = 1), ALNM (n = 3), invasiveness (n = 1), molecular subtypes (n = 1), recurrence (n = 1), and survival (n = 1).

Radiomics Predicting Breast Cancer Characteristics

Histological Grade

Significant correlation between a *combination* of radiomic features and BC histological grade has been observed (29,31,72,73). The prediction accuracies, however, varied from 63% to 100% depending on the radiomic signatures and feature selection methods used (31,73). In particular, a joint set of CEM radiomic features primarily based on gray level co-occurrence matrix (GLCM) with a mutual information coefficients feature selection method achieved impressive accuracies of 90% and 100% for differentiating grade 1 from grade 2 and 3 cancers for invasive cancers and non-invasive cancers, respectively (31). Conversely, the levels of accuracy were reduced when different combination of radiomic features and feature selection method were employed (31,73). These findings suggest that the choice of radiomic signature and feature selection method will drive the success of radiomics when predicting histological grade.

In addition, various *individual* radiomic features also showed significant correlation with BC grading, indicating their potential role for tumor grade prediction (29,72). La Forgia et al. (29) found three individual radiomic features (*variation coefficient, entropy*, and *relative smoothness* - Table 3) calculated on ROIs of recombined low- and high-energy CESM images, to be statistically correlated with BC grading (p value < 0.05 and area under the receiver operating characteristic curve/AUC = 80%). Cui et al. identified ten

TABLE 1. Summary of Mammography-based Radiomic Research in Breast Cancer Included in this Review

Reference (author, year)	Study Design	Patient(no.)	Image Acquisition	Segmentation Method	Handcrafted Radiomic Features Class Extracted	Software Used for Extraction	Features Extracted (no.)	Feature Selection Method	Features Selected (no.)	Classification Method	Validation Method	Prediction/ Performances	Findings
Cui et al. 2019 (72)‡	R	80	2D DM (CC and MLO)	Manual	FOS, and texture (GLCM and GLRLM)	Artificial Intelligence Kit software	435/image	Spearman correlation analysis and LOG	2	LOG	Internal validation cohort	Histological grade: 0.73 AUC radiomics alone 0.84 AUC radiomics + clinical data	Mammography-based texture analysis showed great potential in differentiating phyllodes tumors Grade 1 from Grade 2/Grade 3.
Jiang et al. 2020 (28) ‡	R	200	2D DM (CC and MLO)	Manual	Texture (GLCM, NGLDM, GLRLM, and GLZLM)	LIFEx v5.10	136/patient	LASSO	14	Cox regression	Internal validation cohort	Survival: 3.052 Hazard ratio radiomics alone 0.94 C-index radiomics + clinical data	Radiomic nomogram combining Rad-score and pN stage showed favorable performance in the prediction of iDFS in patients with early-stage TNBCs.
La Forgia et al. 2020 (29)	N/A	52	2D CESM (LE and RC)	Manual	FOS	Matlab 2018a	14/ROI	N/A	N/A	LDA	LOOCV	Histological grade: 0.80 AUC Receptor status: 0.84 AUC ER +/ER- 0.76 AUC PR +/PR- 0.91 AUC HER2 +/HER2- 0.77 AUC TN/NTN Ki-67: 0.85 AUC	CESM-based radiomic features could satisfactory and accurately distinguish BC histological outcomes and some molecular subtypes of tumors.
Li et al. 2019 (75) ‡	R	362	2D DM (CC and MLO)*	Manual	FOS, morphology, texture, and wavelet	Matlab 2014a	569/patient	L^1-LOG, L^1-SVM, EN-LOG, XGBoost, and X^2	9	SVM-Linear, SVM-RBF, L^2-LOG, XGBoost, GBDT, KNN, and RF	10-fold cross-validation	Invasiveness: 0.59 AUC radiomics alone 0.72 AUC radiomics + clinical data	The combined predictive model using radiomic signatures and clinical imaging characteristics had potential in classifying an understaging of invasive BC.

(*continued on next page*)

TABLE 1. (Continued)

Reference (author, year)	Study Design	Patient(no.)	Image Acquisition	Segmentation Method	Handcrafted Radiomic Features Class Extracted	Software Used for Extraction	Features Extracted (no.)	Feature Selection Method	Features Selected (no.)	Classification Method	Validation Method	Prediction/ Performances	Findings
Ma et al. 2019 (30)	R	331	2D DM (CC and MLO)	Manual	Morphology, FOS, and texture (GLCM)	N/A	39/lesion	LASSO	12	Naive Bayes	10-fold cross-validation	Molecular subtypes: 0.75 AUC Luminal A+B 0.78 AUC HER2-enriched 0.87 AUC TNBC	Digital mammographic radiomic features were associated with BC subtypes.
Mao et al. 2020 (70) ‡	R (M)	394	2D CESM (LE and RC)*	Manual	FOS, form factor matrix, GLCM, GLSZM, Haralick matrix, and GLRLM	Artificial Intelligence Kit software	396/image	LASSO	5	LASSO, LOG	10-fold cross-validation, internal and external validation sets	ALNM: Radiomics alone 0.74 AUC internal validation 0.68 AUC external validation Radiomics + clinical data 0.77 AUC internal validation 0.79 AUC external validation	CESM-based radiomic nomogram (integrating radiomic signature and the CESM-reported lymph node status) showed good potential for noninvasive preoperative prediction of ALNM in BC.
Marino et al. 2020 (73)	R	48	2D CEM (LE) and 3D T1W DCE-MRI	Manual	FOS, GLCM, GLRLM, absolute gradient, autoregressive model, wavelet, and lesion geometry	MaZda	N/A	Fisher, POE +ACC, and MI coefficient	N/A	LDA, KNN with LOOCV	KNN with LOOCV	Histological grade: 64% to 78% accuracies Invasiveness: 88% to 92% accuracies Receptor status: 87% to 96% accuracies HR+/HR-	Radiomic analysis of both DCE-MRI and CEM indicated potential for noninvasive assessment of tumor-invasiveness, hormone receptor status, and tumor grade.

(*continued on next page*)

TABLE 1. (Continued)

Reference (author, year)	Study Design	Patient(no.)	Image Acquisition	Segmentation Method	Handcrafted Radiomic Features Class Extracted	Software Used for Extraction	Features Extracted (no.)	Feature Selection Method	Features Selected (no.)	Classification Method	Validation Method	Prediction/ Performances	Findings
Marino et al. 2019 (31)	R	100	2D CEM (LE)	Manual	FOS, absolute gradient, GLRLM, GLCM, autoregressive model, wavelet, and lesion geometry	MaZda	300 in total	Fisher, POE +ACC, and MI coefficient	N/A	LDA, KNN with LOOCV	LDA, KNN with LOOCV	Histological grade: 63% to 100% accuracies Invasiveness: 80% to 87% accuracies Receptor status: 74% to 78% accuracies HR +/HR- 90% to 97% accuracies HER2/HR 81% to 100% accuracies TN/TP 79% to 82% accuracies TN/HR	CEM-based radiomic analysis showed potential for non-invasive differentiation of BC invasiveness, hormone receptor status, and tumor grade.
Son et al. 2020 (76) ‡	R	365	2D Synthetic mammography from DBT(CC and MLO)	Manual	FOS, and texture (GLCM, GLRLM, and GLSZM)	Pyradiomics	129 in total	Elastic-net	20, 18, 66, 6, 34, 43, 17, 34, 42	Elastic-net	Independent validation cohort	Molecular subtypes: Radiomics alone 0.65 AUC Luminal A+B 0.56 AUC HER2-enriched 0.84 AUC TNBC Radiomics + clinical data 0.68 AUC Luminal A+B 0.58 AUC HER2-enriched 0.87 AUC TNBC	Radiomic signature from synthetic mammography reconstructed from DBT images showed high performance in distinguishing TNBC.

(*continued on next page*)

TABLE 1. (Continued)

Reference (author, year)	Study Design	Patient(no.)	Image Acquisition	Segmentation Method	Handcrafted Radiomic Features Class Extracted	Software Used for Extraction	Features Extracted (no.)	Feature Selection Method	Features Selected (no.)	Classification Method	Validation Method	Prediction/ Performances	Findings
Tagliafico et al. 2019 (71)	P	70	2D DBT	Manual	Morphology, FOS, and texture	3D Slicer 4.7	106 in total	LASSO	5	LASSO	N/A	Ki-67: 0.65 to 0.70 AUC	Radiomic features extracted from DBT images were found to be associated with BC Ki-67 expression.
Tagliafico et al. 2018 (26)	P (M)	40	2D DBT	Manual	FOS, and texture (GLCM)	3D Slicer 4.7	104 in total	N/A	6	Correlation analysis and Univariate linear regression	N/A	Receptor status: -0.46 r value	Entropy radiomic feature derived from DBT were associated with ER.
Tamez-Peña et al. 2018 (69) †,‡	P	71	2D DM (CC and MLO)	Manual	Morphology, FOS, and texture	CiPAS and Kitware, Insight Toolkit	1078/ROI	Bootstrap and LASSO	13, 71	LASSO	Stringent cross-validation	Recurrence risk: Radiomics alone $\rho = 0.83$ OncotypeDX $\rho = 0.78$ PAM50 Radiomics + clinical data N/A	Mammography-based radiomic features were correlated to BC recurrence score of OncotypeDX and PAM50.
Tan et al. 2020 (74) ‡	R	216	2D DM (CC)	Manual	Morphology, FOS, texture, and Gabor	Matlab 2018b	4395/patient	ANOVA F-value and LASSO	12	SVM and LOG	Internal validation cohort	ALNM: Internal validation 0.862 AUC radiomics alone 0.863 AUC radiomics + clinical data	Mammography-based radiomic nomogram (incorporating radiomic signature and clinical features) were helpful in preoperative prediction of ALNM.

(*continued on next page*)

TABLE 1. (Continued)

Reference (author, year)	Study Design	Patient(no.)	Image Acquisition	Segmentation Method	Handcrafted Radiomic Features Class Extracted	Software Used for Extraction	Features Extracted (no.)	Feature Selection Method	Features Selected (no.)	Classification Method	Validation Method	Prediction/ Performances	Findings
Wang et al. 2020 (78)	R	112	2D DM (CC and MLO)	Manual	FOS, morphology, and texture (GLCM, GLRLM and GLSZM)	Artificial Intelligence Kit software v3.2.2	396 in total	mRMR and backward stepwise LOG selection	3	Likelihood ratio test	100 times cross-validation	Molecular subtypes: 0.84 AUC TNBC	Digital mammography-based radiomics could discriminate round-like TNBC from non-TNBC in invasive ductal carcinoma.
Yang et al. 2019 (32) ‡	R	147	2D DM (CC and MLO)	Semi-automatic	FOS, texture (GLCM), morphology, and wavelet	Matlab R2016b	299 in total	LASSO	10	Gaussian kernel SVM	Gaussian kernel SVM	ALNM: Internal validation 0.88 AUC radiomics alone 0.81 AUC radiomics + clinical data	Mammography-based radiomic nomogram (combining radiomic signature with clinicopathologic and immunohistochemical risk factors, including T stage, age, tumor location, ER and PR status, US ALN status) was able to non-invasively predict preoperative ALN status.
Yu et al. 2020 (79)	R	43	2D DM (CC and MLO)	Manual	Morphology, FOS, and texture	Pyradiomics	204 in total	Mann-Whitney U test and PCC	6	N/A	N/A	TILs: $p < 0.05$	Mammographic radiomic features were found to be useful for distinguishing between high and low TIL levels in TNBC patients.
Zhang et al. 2019 (77)	R	120	2D DM (CC and MLO)	Manual	Morphology, FOS, and texture	Omni-kinetics software	12/ROI	N/A	N/A	Mann-Whitney U or t test	N/A	Molecular subtypes: 0.64 to 0.75 AUC TNBC	Radiomic features from mammography may be helpful to discriminating TNBC from non-TNBC.

(*continued on next page*)

TABLE 1. (Continued)

Reference (author, year)	Study Design	Patient(no.)	Image Acquisition	Segmentation Method	Handcrafted Radiomic Features Class Extracted	Software Used for Extraction	Features Extracted (no.)	Feature Selection Method	Features Selected (no.)	Classification Method	Validation Method	Prediction/ Performances	Findings
Zhou et al. 2019 (33)	R	306	2D DM (CC and MLO)	Manual	FOS, morphology, and texture	N/A	186 in total	LASSO	40	SVM and LOG model	10-fold cross-validation	Receptor status: 0.79 AUC HER2	Radiomic features extracted from digital mammography showed potential as an efficient and non-invasive tool for preoperative assessment of HER2 status in patients with BC.

Abbreviations: ALN, axillary lymph node; ALNM, axillary lymph node metastasis; AUC, area under the receiver operating characteristic curve; BC, breast cancer; CC, craniocaudal; CEM, contrast-enhanced mammography; CESM, contrast-enhanced spectral mammography; C-index, concordance index; DBT, digital breast tomosynthesis; DCE-MRI, dynamic contrast-enhanced magnetic resonance imaging; DM, digital mammography; EN-LOG, elastic net-logistic regression; ER, estrogen receptor; FOS, first order statistics; GBDT, gradient boosting decision tree; GLCM, gray level co-occurrence matrix; GLRLM, gray level run length matrix; GLSZM, gray level size zone matrix; GLZLM, gray level zone length matrix; HER2, human epidermal growth factor receptor 2; HR, hormone receptor; iDFS, invasive disease free survival; KNN, k-nearest neighbour; LASSO, least absolute shrinkage and selection operator; LDA, linear discriminant analysis; LE, low-energy images; LOG, logistic regression; LOOCV, leave-one-out cross-validation; L^1-LOG, L^1-logistic regression; L^1-SVM, L^1-support vector machine; L^2-LOG, L^2-logistic regression; M, multi-centred; MI, mutual information; MLO, mediolateral oblique; mRMR, maximum correlation minimum redundancy; N/A, not available; NGLDM, neighbourhood gray level difference matrix; P, prospective; PCC, pearson correlation coefficient; POE + ACC, probability of error and average correlation; PR, progesterone receptor; R, retrospective; RE, recombined low- and high-energy images; RF, random forest; ROI, region of interest; SVM-Linear, support vector machine with linear kernel; SVM-RBF, support vector machine with radial-based function kernel; TILs, tumor infiltrating lymphocytes; TN, triple negative; TNBC, triple negative breast cancer; TP, triple positive; US, ultrasound; XGBoost, extreme gradient boosting; X^2, Chi-square test.

*Conducted image pre-processing, n = 2.

†Radiogenomics study, n = 1.

‡Combining radiomics with clinical data studies, n = 8.

TABLE 2. Summary of Radiomics with Clinical Data Studies

Prediction	Reference (author, year)	Radiomic Signature Performance	Radiomics with Clinical Data Performance	Difference in Prediction Performance‡	Outcomes
Histological grade	Cui et al. 2019 (72)	0.73 AUC	0.84 AUC*	0.11†	Integrating mammography-based texture features with pre-biopsy clinical features (size and shape) indicated improved performance in discriminating phyllodes tumors grade 1 from grade 2/3.
Axillary lymph node metastasis	Mao et al. 2020 (70)	Internal and external validation 0.74 and 0.68 AUC	Internal and external validation 0.77* and 0.79* AUC	Internal and external validation 0.02 and 0.11†	CESM-based radiomic nomogram (integrating radiomic signature and pre-biopsy CESM-reported lymph node status) showed slightly better performance than radiomic signature alone in predicting ALNM.
	Tan et al. 2020 (74)	Validation cohort 75% sensitivity 77% specificity 76% accuracy 0.862 AUC	Validation cohort 75% sensitivity 81% specificity* 79% accuracy* 0.863 AUC*	0% sensitivity 4% specificity 3% accuracy 0.001 AUC	Mammography-based radiomic nomogram (incorporating radiomic signature and post-biopsy clinical factors such as PR and molecular subtype) was overall more superior than radiomic signature alone in predicting ALNM.
	Yang et al. 2019 (32)	Validation cohort 82% sensitivity 80% specificity 80% accuracy* 0.88 AUC/C-index*	Validation cohort N/A N/A 73% accuracy 0.81 AUC/C-index	N/A N/A 7% accuracy 0.07 AUC/C-index	Mammography-based radiomic signature alone outperformed radiomic nomogram (combining radiomic signature with post-biopsy clinicopathologic and immunohistochemical risk factors, including T stage, age, tumor location, ER and PR status, US ALN status) in predicting ALN status.
Invasiveness	Li et al. 2019 (75)	0.59 AUC	0.72 AUC*	0.13†	Prediction model using both mammographic radiomics (62 features) and clinical imaging characteristics (3 pre-biopsy features: age, asymmetry or mass prevalence, and breast density, and 1 post-biopsy feature: nuclear grade) performed better than radiomic signature alone.
Molecular subtypes	Son et al. 2020 (76)	Validation cohort AUC 0.65 Luminal vs. non-luminal 0.56 HER2 vs. non-HER2	Validation cohort AUC 0.68 Luminal vs. non-luminal* 0.58 HER2 vs. non-HER2*	0.03 0.02 0.03	Combining radiomic signature from synthetic mammography from DBT and pre-biopsy clinical features (patient age, tumor size and qualitative imaging findings)

(*continued*)

TABLE 2. (Continued)

Prediction	Reference (author, year)	Radiomic Signature Performance	Radiomics with Clinical Data Performance	Difference in Prediction Performance‡	Outcomes
		0.84 TN vs. non-TN	0.87 TN vs. non-TN*		showed higher AUC than radiomic signature alone.
Recurrence	Tamez-Peña et al. 2018 (69)	N/A OncotypeDX (ρ = 0.83) PAM50 (ρ = 0.78)	N/A	N/A	The prediction models were improved when combining mammographic radiomics and post-biopsy clinical data (triple negative status and PR).
Survival	Jiang et al. 2020 (28)	Hazard Ratio 3.052 $p < 0.001$	Validation 0.94 C-index	N/A	Radiomic nomogram combining mammographic radiomic signature and post-biopsy pN stage achieved C-index of 0.94 in the validation cohort.

Abbreviations: ALN, axillary lymph node; ALNM, axillary lymph node metastasis; AUC, area under the receiver operating characteristic curve; CESM, contrast-enhanced spectral mammography; C-index, concordance index; DBT, digital breast tomosynthesis; ER, estrogen receptor; N/A, not available; PR, progesterone receptor; US, ultrasound.

*Greater performance.

†Significant improvement in performance.

‡Difference in prediction performance between radiomic signature and radiomics with clinical data.

important DM-based texture features indicating significant differences in grading phyllodes tumors (PTs), with two particular features related to *GLCM correlation* and *cluster prominence* (Table 3) being independent factors in distinguishing PTs grade 1 from grade 2 or 3 (p value < 0.05 and AUC = 73%) (72).

Lymph Node Status

Axillary lymph node metastasis.—Breast-based radiomic signature for predicting ALNM have demonstrated favorable results of up to 0.88 AUC, 80% accuracy, 82% sensitivity and 80% specificity (32,70,74). Notably, a radiomic signature based on a group of 10 selected features (1 FOS and 9 wavelet features) extracted from both craniocaudal (CC) and mediolateral oblique (MLO) view of mammograms indicated better ALNM prediction performance (AUC of 0.88 in the validation cohort) (32) than radiomic signatures from CC view alone (AUC of 0.86) (74) or from low-energy and recombined low- and high-energy CESM images (AUCs of 0.74 internal and 0.68 external validation) (70). However, whilst mammogram-based radiomic signatures (32,74) demonstrated greater ALN status predictive ability than CESM-based radiomic signature (70), the studies (32,74) were based on a single institution data with no external validation conducted unlike the CESM study (70).

Although all the above studies (32,70,74) applied least absolute shrinkage and selection operator logistic regression method to select significant ALNM related radiomic features, their ROI segmentation methods were different; highlighting that semi-automatic segmentation approach performed in the mammographic work of Yang et al. (32) may be more useful than manual segmentation (70,74) and could enhance the ability of radiomic signature in ALNM prediction in patients with BC.

Invasiveness.—Radiomic analysis has also shown its ability to predict BC invasiveness (31,73,75). However, current findings (31,73,75) demonstrate high dependency on the modality or approach used: CEM-based radiomic signature from combining GLCM and FOS features, or GLCM alone was able to differentiate BC invasiveness with high accuracies of 87% and 92% respectively (31,73), whereas the classification performance of DM-based radiomic signature from a collective group of morphology and FOS features, was found to be less satisfactory (AUC of 59%) (75). More precisely, the CEM-based radiomic signature using a group of GLCM features (73) manifested superior outcome (accuracy of 92%) compared to CEM-based radiomic signature using the combined features from GLCM and FOS (accuracy of 87%) (31). These findings suggest that these group of GLCM features contained within CEM images may be more valuable than other approaches or image types (31,75) in the prediction of BC invasiveness. It is important to note however that these findings arising from Marino et al. study (73) are still preliminary since a smaller sample size (48 patients) than the other studies (100 and 362 patients) (31,75) was employed.

It is interesting to explore why the mammographic results were relatively disappointing. In the mammographic work by Li et al. (75), the number of invasive cancer case (n = 110) was less than the non-invasive cancers (n = 252) while the numbers of invasive cancers case (n = 45 and 91) were more

TABLE 3. List of Radiomic Features Cited in this Review

Radiomic Feature Name	Feature Class	Prediction	Description
Cluster prominence (72)	Texture	Histological grade	Measuring asymmetry of a given distribution. High values represent more asymmetry of the image while low values indicate a small peak for the image gray level value and less gray level variation.
Concavity (30)	Morphology	Molecular subtype	Reflecting the irregularity degree of a lesion boundary/shape; the smaller the concavity value, the more regular the shape of the tumor is.
Entropy (26,29)	First order statistics	Histological grade, receptor status (ER), Ki-67	Describing the randomness in the gray level distribution of the image.
GLCM autocorrelation (79)	Texture	TILs	Measuring the degree of texture fineness and coarseness of texture.
GLCM correlation (30,72,79)	Texture	Histological grade, molecular subtype, TILs	Measuring the similarity or linear dependency of the gray level values to their corresponding pixels/voxels in the GLCM; representing the smoothing gradient of pattern in quantitative image with a larger value indicating a smoother pattern.
GLDM low gray level emphasis (79)	Texture	TILs	GLDM quantifies gray level dependencies in an image. GLDM low gray level emphasis measures the distribution of low gray level values, with higher value showing greater concentration of low gray values in the image, which represents the brightness in the mammographic image.
Gray mean (30)	First order statistics	Molecular subtype	Representing brightness of a lesion in the image; the larger the gray mean value, the brighter the lesion is in the image.
NGTDM contrast (79)	Texture	TILs	NGTDM quantifies the differences between a pixel/voxel gray value and the mean gray value of its neighbours within a predefined distance. NGTDM contrast measures the spatial intensity change dependent on the overall gray level dynamic range.
Relative smoothness (29)	First order statistics	Histological grade, receptor status (HER2), Ki-67	Describing the sharpness of the histogram of the image.
Roundness (30)	Morphology	Molecular subtype	Representing shape of a lesion; the larger the roundness value, the rounder the lesion is.
Skewness (29)	First order statistics	Receptor status (PR)	Describing the asymmetry of the histogram of the image.
Uniformity (79)	First order statistics	TILs	Measuring the sum of squares of each intensity value, which reflects the regularity of the image array, where a greater uniformity implies a greater consistency or a smaller range of discrete intensity values.
Variance (79)	First order statistics	TILs	Representing the mean of the squared distances of each intensity value from the mean value, which is a measure of the mean distribution.
Variation coefficient (29)	First order statistics	Histological grade, receptor status (ER), Ki-67	Describing the variability of the gray level distribution.
Variation range (29)	First order statistics	Receptor status (ER), Ki-67	Describing the difference of maximum and minimum of gray level values.

Abbreviations: ER, estrogen receptor; GLCM, gray level co-occurrence matrix; GLDM, gray level dependence matrix; NGTDM, neighbourhood gray tone difference matrix; PR, progesterone receptor; TILs, tumor infiltrating lymphocytes.

than the non-invasive cancers (n = 4 and 12) in Marino et al. (31,73). Also, both mass and asymmetry cases with calcification lesions were included in Li et al. (75) but excluded in the CEM studies of Marino et al. (31,73). The results suggest that these particular features may provide limited relevant lesion characteristics. However, one cannot rule out methodological factors: due to the small sample size of invasive cancers (n = 110) of the single retrospective study of Li et al. (75).

The radiomic analyses in the above studies were not consistent (31,73,75). Computer-based image pre-processing method of correcting intensity and filtering noise on the mammographic images was performed by Li et al. (75) but not on the CEM images by Marino et al. (31,73). Although segmentation method of ROI was conducted manually by expert radiologists in all studies (31,73,75), radiomic feature selection and classification methods were contrasting. Feature selection such as mutual information coefficients and fisher, and classification method such as linear discriminant analysis and k-nearest neighbour classification were used by Marino et al. (31,73). However, ML cross combinations approach of feature selection (e.g., L^1-support vector machine) and classification (e.g., random forest classifier) were employed by Li et al. (75). This further emphasises that varying techniques such as feature selection or the ML and classification approach, substantially effects predictive abilities (31,73,75).

Receptor Status

The effectiveness of breast-based radiomic analysis for predicting receptor status has yielded impressive results (96% accuracy for hormone receptor/HR, 91% AUC for human epidermal growth factor receptor 2/HER2, 97% accuracy for HER2/HR, 100% accuracy for triple negative/TN) (26,29,31,33,73). When predicting HR, HER2, HER2/HR and TN expression specifically, radiomic features extracted from CEM images (31,73) and CESM images (29) were able to accurately differentiate HR+ from HR- (accuracy 96%) (73), HER2+ from HER2- (AUC of 91%) (29), HER2+/HR- from HER2-/HR+ (accuracy of 97%) (31), and TN from triple positive/TP (accuracy of 100%) (31), respectively. More precisely, in discriminating between different HR expressions, Marino et al. showed that CEM-based radiomics using gray level run length matrix (GLRLM) features demonstrated greater prediction performance (accuracy of 96% for HR+ vs. HR-) (73), than other combinations of radiomic features based on CEM (accuracies 74% to 91% HR+ vs. HR-) (31,73) or CESM images (AUCs of 84% estrogen receptor positive/ER+ vs. ER-, and 76% progesterone receptor positive/PR+ vs. PR-) (29). La Forgia et al. using CESM images (29) and Tagliafico et al. using DBT images (26) also indicated significant relationships between CESM-based radiomic features (i.e., *variation coefficient* and *variation range*) and DBT-based radiomic feature (i.e., *entropy*) with ER (AUC of 84% (29), r value = -0.46 (26)), as well as relationships between CESM-based radiomic feature (i.e., *skewness*) and PR status (AUC of 76%) (29) – Table 3. Nevertheless, the discriminative ability in the CESM study by La Forgia et al. as mentioned above was less promising (AUCs of 84% ER+ vs. ER-, and 76% PR+ vs. PR-) (29) than the accuracy of 96% using CEM-based radiomic features derived from GLRLM in the relatively small study of Marino et al. (73).

In terms of specifically discriminating HER2+ from HER2-, radiomic features extracted from CESM images by La Forgia et al. (AUC of 91%) (29) outperformed radiomic features from both CC and MLO views of mammograms as described by Zhou et al. (AUC of 79%) (33). Notably, the *relative smoothness* (Table 3) radiomic feature derived from low-energy CESM was discovered to be significantly correlated to HER2 (p value < 0.05) (29), whereas no radiomic features from recombined low- and high-energy of CESM (29) or from DBT images studied by Tagliafico et al. (26) were found to be associated with HER2. It should be acknowledged that the outcomes presented by the CESM work of La Forgia et al. (29), are only based on the initial results of an observational retrospective study involving only 52 individuals and 68 lesions in contrast to the larger work of Zhou et al. (n = 306) (33), and the multi-centred trial of Tagliafico et al. (26).

With regard to the classification of multifaceted BC receptors (i.e., HER2/HR combinations) or TN, CEM-based radiomic features presented by Marino et al. performed well (31). Specifically, CEM-based radiomics with combined set of GLRLM and wavelet features, and set of GLCM, wavelet, and FOS features were able to separate HER2+/HR- from HER2-/HR+ with an accuracy of 97%, and TN from TP with an accuracy of 100%, respectively (31). These predictive powers were better than other radiomic features combinations based on CEM (accuracies of 90% to 96% HER2+/HR- vs. HER2-/HR+, 79% to 82% TN vs. HR+, and 81% to 88% TN vs. TP) (31) or CESM images (AUC of 77% TN vs. non TN) (29). However, the preliminary findings of the CEM single institution study of Marino et al. (31) are yet again limited by fairly small number of patients (n = 100) and cancer cases (n = 4 for HER2+/HR- and n = 16 for TN).

Ki-67 Proliferation Status

Significant association between quantitative radiomic features of mammography (such as CESM and DBT) and Ki-67 proliferation status has been illustrated (29,71). Specifically, when differentiating Ki-67+ from Ki-67-, CESM-based radiomic features identified by La Forgia et al. (29) yielded greater prediction performance of 85% AUC, than DBT-based radiomic features presented by Tagliafico et al. (AUCs 65% to 70%, low vs. high Ki-67) (71). More explicitly, four FOS radiomic features derived from recombined low- and high-energy of CESM: *variation coefficient, variation range, entropy*, and *relative smoothness* – Table 3, were observed to have substantial relationships with the proliferation activity of Ki-67 ($p < 0.05$), while no radiomic features from CESM low-energy images were significantly related to Ki-67 (29). Despite the fact that five DBT-based radiomic features including two FOS features were shown to be significant in the classification of Ki-67 by Tagliafico et al. ($p \leq 0.001$)

(71), their aforementioned prognostic abilities were less encouraging than those identified by La Forgia et al. (29).

Molecular Subtypes

When characterizing BC molecular subtypes of luminal (A and B), HER2-enriched, and triple negative breast cancer (TNBC), radiomic features derived from both CC and MLO views of digital mammograms described by Ma and colleagues demonstrated an ability to distinguish luminal from non-luminal, HER2-enriched from non-HER2-enriched, and TNBC from non-TNBC subtypes, with statistically significant AUC values of 75%, 78% and 87%, respectively (30). These achievements were greater than the results described in the radiomics studies of Son et al. (76), Zhang et al. (77), and Wang et al. (78) which focused on synthetic mammography (SM) constructed from DBT (AUCs of 65%, 56%, and 84% for luminal vs. non-luminal, HER2-enriched vs. non-HER2-enriched, and TNBC vs. non-TNBC, correspondingly (76)), and DM (AUCs of 64% to 75% (77), and 84% (78) for TNBC vs. non-TNBC), respectively. Importantly, four mammographic radiomic features were shown to be particularly useful by Ma et al. (30): *roundness* and *concavity* of morphology; *gray mean* of FOS from MLO view; and *GLCM correlation* of texture feature from CC view images – Table 3. TNBCs appear for example to be rounder (*roundness*) and more regular (*concavity*) in shape, and have higher brightness (*gray mean*) on x-ray images than other subtypes whilst luminal lesions may have smoother gradient of texture pattern (*GLCM correlation*) than HER2-enriched and TNBCs. Interestingly, although texture features had added value to the prediction of molecular subtypes in Ma et al. (30), Son et al. (76) and Wang et al. (78), no texture feature was found to be helpful when distinguishing between TNBC and non-TNBC in the work of Zhang et al. (77).

Whilst similar overall observations and associations between radiomic features and molecular subtypes were reported in the SM study of Son et al. (76) and the mammographic study of Zhang et al. (77) and Wang et al. (78) (as described above), their aforementioned predictive values (76-78) were lower than those described by Ma et al. (30). It must be acknowledged however that unlike the larger SM analysis of Son et al. (n = 365) (76), no independent validation was carried out in the smaller single institutional/vendor mammographic studies of Ma et al. (n = 331) (30), Zhang et al. (n = 120) (77) and Wang et al. (n = 112) (78).

When examining possible reasons behind somewhat inadequate results of luminal and HER2-enriched in the SM study by Son et al., it is fascinating to note that these results are despite the fact that larger numbers of selected radiomic features were included for the prediction of luminal (n = 66) and HER2-enriched (n = 18) (76), comparing to only 12 selected radiomic features used in the mammographic work of Ma et al. (30). Furthermore, due to the widely known morphologic traits of architectural distortion (i.e., mammographically non-visible masses) and microcalcifications of luminal and HER2-enriched subtypes, even the large numbers of selected radiomic features in Son et al. (76) may not provide the complete representative of these subtypes characteristics. It is clear therefore that a comprehensive library of relevant radiomic features tailored to cancer subtypes is needed if we are to understand which are the best predictive features.

Tumor Infiltrating Lymphocytes

Using radiomic approaches, a mammographic study by Yu et al. (79) based on CC and MLO views of digital mammograms was able to identify six mammographic features that could differentiate high and low TIL levels ($p < 0.05$): two FOS (*variance* from CC and *uniformity* from MLO) and four texture features (*GLCM autocorrelation, GLDM low gray level emphasis* from CC, *GLCM correlation* and *NGTDM contrast* from MLO views) - Table 3. However, it has to be noted that this is the only radiomic study to date on TILs which involved only 43 TNBC patients from single institution.

Radiomics Predicting Breast Cancer Risk of Recurrence and Survival

Gene expression biomarker of breast tumors has become a new paradigm for predicting BC risk of recurrence (12,80,81). However, a radiogenomics approach (predicting gene expression from radiomics) may have value with radiomic signatures based on both CC and MLO views of digital mammograms being significantly associated with two popular BC multigene assays recurrence scores: OncotypeDX (Pearson $\rho = 0.83$) and PAM50 ($\rho = 0.78$) (69), suggesting that radiomic features may be able to predict BC recurrence. Precisely, features derived from fractals and wavelets were of particular usefulness. Nevertheless, it should be acknowledged that the work to date is based on one study with one vendor's images of 71 cancer patients.

With reference to invasive disease free survival (iDFS) of early stage TNBC patients, DM derived radiomic texture features (e.g., GLCM and GLRLM) from both CC and MLO views were shown to have substantial relationships with iDFS in TNBC patients (28). This result implies that mammographic texture features were able to independently predict iDFS and identify patients with increased risk of having poor survival outcomes/iDFS. However, again these data rise from only a single study.

Radiomics Integrating with Clinical Data

Finally, it is important to consider if integrating clinical data with radiomics could add value when predicting factors such as histological grade, ALNM, invasiveness, molecular subtypes, recurrence, and survival (28,32,69,70,72,74-76) - Table 2. So far, some improvements have been shown when compared to radiomic analysis alone, with significant improvements in AUCs of 11% for predicting histological

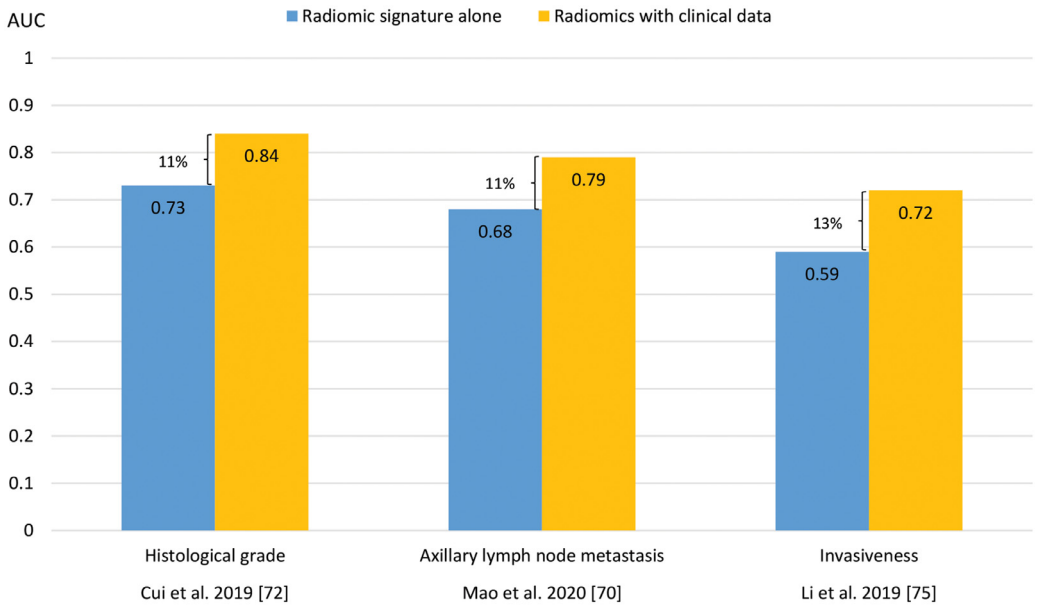


Figure 3. Performance improvements of radiomics with clinical data studies compared to radiomics alone. When integrating radiomics with clinical data, three studies show significant improvements of 11% in AUC for predicting histological grade (72) and axillary lymph node metastasis (70), and 13% in AUC for predicting cancer invasiveness (75), compared to radiomic signature alone. AUC = area under the receiver operating characteristic curve. (Color version of figure is available online.)

grade (combining DM-based radiomics with pre-biopsy tumor size and shape data) (72) and ALNM (incorporating CESM-based radiomics with pre-biopsy CESM-reported lymph node/LN status) (70), as well as 13% AUC improvement for predicting cancer invasiveness (DM-based radiomics integrating with pre-biopsy features: age, asymmetry/mass prevalence, and breast density, and post-biopsy: nuclear grade) (75) - Figure 3. However, due to the limited works and results available in these areas and the lack of consistency in the performance metrics used in some studies, firm conclusions at this time are difficulty to reach.

DISCUSSION

In this scoping review, 18 published studies between January 1, 2015 and December 19, 2020 were identified as evidence for evaluating the potential role of mammography-based (i.e., DM, DBT, and CEM/CESM) radiomic analysis in predicting BC characteristics, recurrence, and survival. Studies have shown that radiomics analysis based on combinations of or single radiomics features contained within mammographic images has large potential for reliable, accurate, non-invasive, and cost-effective assessment of BC characteristics (histological grade, ALNM, invasiveness, receptor status, Ki-67 proliferation status, molecular subtypes, and TILs) as well as predicting recurrence and survival. The vast majority of the current evidence focuses on describing various features of BC (n = 16) rather than predicting outcomes (n = 2). In particular, radiomic features from CEM/CESM generally seem to be more useful than those from DM and DBT in predicting histological grade, invasiveness, receptor status, and Ki-67, while radiomics features from DM appear to be better in predicting ALNM and molecular subtypes. Interestingly, although DBT may be superior to DM in invasive breast lesion detection and may be adapted to replace DM in screening and diagnosis (49,62,82,83), its ability in radiomics analysis remains a key question that needs further investigation as DBT radiomics work to date is scarce with only three studies (26,71,76) available in this review.

Additionally, whilst the above BC factors may be known at the time of diagnosis through histopathological assessment of breast tissue, due to the heterogeneity nature of BC, the current method of biopsy tissue sampling is invasive/expensive and incapable of capturing complete information on the whole breast tumor. Radiomics, however, in a non-invasive and cost-effective way, can quantitatively extract these features directly from clinical medical images capturing the entire tumor heterogeneity and genotypic characteristics which can be used to enhance the accuracy of BC diagnosis, prognosis, and prediction. The potential value and capability of radiomics as alternative means for non-invasive and cost-effective assessment of BC have been shown by the studies in this review and many other reviews (16,24,35,38,62), specifically on its potential role around predicting factors that describe BC characteristics as well as recurrence and survival as mentioned above. Some key values of radiomics include, the possibility of serving as a non-invasive and effective biomarker for preoperative prediction of ALNM (32,74) and tumor invasiveness (31,73), detecting aggressive cancer early such as TNBC (30,76), and predicting therapeutic effect (69) and patients survival outcomes (28) which are essential for guiding therapy decisions including eliminating unnecessary surgery and/or biopsy.

To date, however, all published studies in this review only analysed handcrafted features and the efficacy of these features depends greatly on image acquisition, segmentation, features selection and classification methods. Thus, there are no uniform algorithms for handcrafted feature extraction used for analysis of various breast lesion characteristics such as architectural distortions, masses, asymmetries and calcifications potentially leading to different conclusions (75,76) - Table 1. In this regard, more research efforts are needed to attain a level of standardization and validation before adapting radiomics into clinical practice. Furthermore, the usefulness of mammographic deep radiomic features in describing BC characteristics, as well as predicting survival and recurrence has not yet been examined by the studies and needs to be explored in further research. DLR have become popular and emerged as a promising approach because of its advanced ability to automatically learn to extract the relevant predictive features and predict outcomes directly from images via deep neural networks, with no experts intervention required unlike the handcrafted features radiomics (84). Although DLR faces some major challenges such as the imbalanced nature of typical medical imaging datasets (e.g., less cancer and more normal cases) limiting its predictive efficacy, as well as interpretability issues due to the complex nature of deep neural networks (38,85), some recent studies have shown promising DLR performances around its ability to improve predictive performances when combined with handcrafted or clinical features (53,86-92).

In this review, further added value has also been observed when integrating mammographic radiomics with clinical data (e.g., pre-biopsy features: tumor size, shape, CESM-reported LN status, age, asymmetry/mass prevalence, breast density, and post-biopsy: nuclear grade) in predicting histological grade, ALNM, invasiveness, molecular subtypes, recurrence, and survival - Table 2 and Figure 3. However, more work is necessary as so far there are only a few published studies in this area. Moreover, almost all studies only evaluated single 2D modality images and did not incorporate features from 3D images, different imaging modalities or other omics data such as genomics (i.e., radiogenomics, or multi-omics analysis) - Table 1. Those studies that had a multifaceted focus, demonstrated the value of 3D (93-95), multiparametric (90,95,96) and multimodality images (97-100), radiogenomics (22,62,101) and multi-omics (102-104) in radiomics analysis to increase predictive accuracy. BC is a temporal and spatial complex genetic disease connecting to different levels of molecules and its management often involves various imaging types and/or genetic assays. In fact, in the era of radiomics in precision medicine, integrating radiomics with other features/data is believed to be a crucial direction for future studies which could increase the possibility for better disease management including tailored treatment strategies and improved patient outcomes.

When comparing the use of radiomics in mammography to other imaging modalities such as breast MRI, although both modalities have played vital roles in BC screening, detection, diagnosis and staging, the majority of BC radiomics works to date has been predominantly in MRI (16). MRI may have higher sensitivity compared to mammography when assessing breast disease extent, it is however not able to detect calcifications nor easily accessible for all patients due to its high cost and time consuming process, unlike DM and DBT. When predicting ALNM, for example, mammographic radiomics studies (32,70,74) in our review have exhibited reasonably good results (up to 0.88 AUC) compared to MRI radiomics studies (up to 0.91 AUC) in a review by Ye et al. (60). Moreover, CEM/CESM as cutting edge technologies at lower cost and shorter procedure time than MRI, have been reported to be comparable to MRI (105,106). Recent radiomics study has also revealed similar to higher diagnostic accuracies between radiomics analysis of CEM and dynamic contrast-enhanced MRI (DCE-MRI) in examining BC invasiveness (accuracies CEM 92% vs. DCE-MRI 90%), hormone receptor status (accuracies CEM 96% vs. DCE-MRI 83%), and tumor grade (accuracies CEM 78% vs. DCE-MRI 78%) (73) indicating potential role of CEM radiomics as accessible prognostic factor for BC management complementary to MRI. Therefore, if radiomics analysis can be performed on more routinely accessible DM, DBT, or more cost- and time-efficient CEM (compare to MRI), it would be of clinical value and the evidence in this review has shown the feasibilities.

Our scoping review has some limitations. First, the findings were mostly retrospective and preliminary based on small sample size from single institution or images from single vendor with no external validation (n = 17). The small sample size and lack of multi-centre comparison also limits the prognostic accuracy and generalizability of the results. Second, there was a paucity of papers on issues such as TILs, recurrence, and survival. Therefore, more and larger prospective multi-institutional studies are required, using images from a variety of vendors. These should be followed by external validation to not only confirm some of the very exciting initial results, but also to provide a more comprehensive understanding of radiomic effectiveness. Lastly, radiomic approaches used in the studies varied widely making comparisons between studies or performing meta-analyses difficult, highlighting the need for approach standardization.

In conclusion, our scoping review demonstrates promising early results, indicating a potential clinical role of applying radiomic analysis to mammography (i.e., DM, DBT, CEM/CESM) when describing BC characteristics as well as predicting recurrence and survival. Nonetheless, more efforts are required to address the aforementioned challenges and knowledge gaps, and further evaluate its reliability as a routine clinical tool. The role of integrating radiomics with other information is also highlighted as an important future direction and more research in this area is required.

REFERENCES

1. Turashvili G, Brogi E. Tumor Heterogeneity in Breast Cancer. Frontiers in Medicine 2017; 4(227):1–11.

2. Zardavas D, Irrthum A, Swanton C, Piccart M. Clinical management of breast cancer heterogeneity. Nature Reviews Clinical Oncology 2015; 12(7):381–394.
3. Wu Y, Sahin AA. Prognostic and Predictive Factors of Invasive Breast Cancer. Breast Disease: Diagnosis and Pathology. Springer International Publishing Switzerland, 2016:187–206.
4. Cao SS, Lu CT. Recent perspectives of breast cancer prognosis and predictive factors (Review). Oncology Letters 2016; 12(5):3674–3678.
5. Leithner D, Bernard-Davila B, Martinez DF, Horvat JV, Jochelson MS, Marino MA, Avendano D, Ochoa-Albiztegui RE, Sutton EJ, Morris EA, Thakur SB, Pinker K. Radiomic Signatures Derived from Diffusion-Weighted Imaging for the Assessment of Breast Cancer Receptor Status and Molecular Subtypes. Molecular Imaging and Biology 2020; 22:453–461.
6. Mamounas EP, Russell CA, Lau A, Turner MP, Albain KS. Clinical relevance of the 21-gene Recurrence Score® assay in treatment decisions for patients with node-positive breast cancer in the genomic era. npj Breast Cancer 2018; 4(27):1–6.
7. Chitalia RD, Kontos D. Role of texture analysis in breast MRI as a cancer biomarker: A review. Journal of Magnetic Resonance Imaging 2019; 49(4):927–938.
8. Ovcaricek T, Takac I, Matos E. Multigene expression signatures in early hormone receptor positive HER 2 negative breast cancer. Radiology and Oncology 2019; 53(3):285–292.
9. Kwa M, Makris A, Esteva FJ. Clinical utility of gene-expression signatures in early stage breast cancer. Nature Reviews Clinical Oncology 2017; 14(10):595–610.
10. Woodard GA, Ray KM, Joe BN, Price ER. Qualitative Radiogenomics: Association between Oncotype DX Test Recurrence Score and BI-RADS Mammographic and Breast MR Imaging Features. Radiology 2018; 286(1):60–70.
11. Pinker K, Chin J, Melsaether AN, Morris EA, Moy L. Precision Medicine and Radiogenomics in Breast Cancer: New Approaches toward Diagnosis and Treatment. Radiology, 2018; 287(3):732–747.
12. Lo Gullo R, Daimiel I, Morris EA, Pinker K. Combining molecular and imaging metrics in cancer: radiogenomics. Insights Imaging 2020; 11 (1):1–17.
13. Rummel SK, Lovejoy LA, Turner CE, Shriver CD, Ellsworth RE. Should Genetic Testing for Cancer Predisposition Be Standard-of-Care for Women with Invasive Breast Cancer? The Murtha Cancer Center Experience. Cancers 2020; 12(234):1–10.
14. Harowicz MR, Robinson TJ, Dinan MA, Saha A, Marks JR, Marcom PK, Mazurowski MA. Algorithms for prediction of the Oncotype DX recurrence score using clinicopathologic data: a review and comparison using an independent dataset. Breast Cancer Research and Treatment 2017; 162(1):1–10.
15. Epstein MS, Lagios MD, Silverstein MJ. Ductal Carcinoma in Situ of the Breast. In: Bland KI, Copeland EM, Klimberg VS, Gradishar WJ, eds. The Breast Comprehensive Management of Benign and Malignant Diseases, Fifth Edition, Philadelphia, PA: Elsevier; 2018:562–575.e4.
16. Lee S-H, Park H, Ko ES. Radiomics in Breast Imaging from Techniques to Clinical Applications: A Review. Korean Journal of Radiology 2020; 21(7):779–792.
17. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016; 278(2):563–577.
18. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, Forster K, Aerts HJWL, Dekker A, Fenstermacher D, Goldgof DB, Hall LO, Lambin P, Balagurunathan Y, Gatenby RA, Gillies RJ. Radiomics: the process and the challenges. Magnetic Resonance Imaging, 2012; 30 (9):1234–1248.
19. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RGPM, Granton P, Zegers CML, Gillies R, Boellard R, Dekker A, Aerts HJWL. Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer 2012; 48 (4):441–446.
20. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, Bellomi M. Radiomics: the facts and the challenges of image analysis. European radiology experimental 2018; 2(36):1–8.
21. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, De Jong EEC, Van Timmeren J, Sanduleanu S, Larue RTHM, Even AJG, Jochems A, Van Wijk Y, Woodruff H, Van Soest J, Lustberg T, Roelofs E, Van Elmpt W, Dekker A, Mottaghy FM, Wildberger JE, Walsh S. Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology 2017; 14(12):749–762.
22. Lin P, Liu WK, Li X, Wan D, Qin H, Li Q, Chen G, He Y, Yang H. MRI-based radiogenomics analysis for predicting genetic alterations in oncogenic signalling pathways in invasive breast carcinoma. Clinical Radiology 2020; 75(7). 561.e1-561.e11.
23. Liu J, Sun D, Chen L, Fang Z, Song W, Guo D, Ni T, Liu C, Feng L, Xia Y, Zhang X, Li C. Radiomics Analysis of Dynamic Contrast-Enhanced Magnetic Resonance Imaging for the Prediction of Sentinel Lymph Node Metastasis in Breast Cancer. Frontiers in Oncology 2019; 9 (980):1–8.
24. 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: 1–10. 2018.
25. Valdora F, Houssami N, Rossi F, Calabrese M, Tagliafico AS. Rapid review: radiomics and breast cancer. Breast Cancer Research and Treatment 2018; 169(2):217–229.
26. Tagliafico AS, Valdora F, Mariscotti G, Durando M, Nori J, Forgia DLa, Rosenberg I, Caumo F, Gandolfo N, Houssami N, Calabrese M. An exploratory radiomics analysis on digital breast tomosynthesis in women with mammographically negative dense breasts. The Breast 2018; 40:92–96.
27. Tagliafico AS, Piana M, Schenone D, Lai R, Massone AM, Houssami N. Overview of radiomics in breast cancer diagnosis and prognostication. The Breast 2020; 49:74–80.
28. Jiang X, Zou X, Sun J, Zheng A, Su C. A Nomogram Based on Radiomics with Mammography Texture Analysis for the Prognostic Prediction in Patients with Triple-Negative Breast Cancer. Contrast Media & Molecular Imaging 2020: 1–10. 2020.
29. La Forgia D, Fanizzi A, Campobasso F, Bellotti R, Didonna V, Lorusso V, Moschetta M, Massafra R, Tamborra P, Tangaro S, Telegrafo M, Pastena MI, Zito A. Radiomic Analysis in Contrast-Enhanced Spectral Mammography for Predicting Breast Cancer Histological Outcome. Diagnostics 2020; 10(708):1–11.
30. Ma W, Zhao Y, Ji Y, Guo X, Jian X, Liu P, Wu S. Breast Cancer Molecular Subtype Prediction by Mammographic Radiomic Features. Academic Radiology 2019; 26(2):196–201.
31. Marino MA, Pinker K, Leithner D, Sung J, Avendano D, Morris EA, Jochelson M. Contrast-Enhanced Mammography and Radiomics Analysis for Noninvasive Breast Cancer Characterization: Initial Results. Molecular Imaging and Biology 2019; 22:780–787.
32. Yang J, Wang T, Yang L, Wang Y, Li H, Zhou X, Zhao W, Ren J, Li X, Tian J, Huang L. Preoperative Prediction of Axillary Lymph Node Metastasis in Breast Cancer Using Mammography-Based Radiomics Method. Scientific Reports 2019; 9(4429):1–11.
33. Zhou J, Tan H, Bai Y, Li J, Lu Q, Chen R, Zhang M, Feng Q, Wang M. Evaluating the HER-2 status of breast cancer using mammography radiomics features. European Journal of Radiology 2019; 121:108718.
34. Gillies RJ, Schabath MB. Radiomics Improves Cancer Screening and Early Detection. Cancer Epidemiology Biomarkers & Prevention 2020; 29(12):2556–2567.
35. Conti A, Duggento A, Indovina I, Guerrisi M, Toschi N. Radiomics in breast cancer classification and prediction. Seminars in Cancer Biology 2020; 72:238–250.
36. Van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. Insights into Imaging 2020; 11(91):1–16.
37. Kocak B, Durmaz ES, Ates E, Kilickesmez O. Radiomics with artificial intelligence: a practical guide for beginners. Diagnostic and Interventional Radiology 2019; 25(6):485–495.
38. Pang T, Wong JHD, Ng WL, Chan CS. Deep learning radiomics in breast cancer with different modalities: Overview and future. Expert Systems with Applications 2020; 158:1–14.
39. Haralick RM, Shanmugam K, Dinstein IH. Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics 1973; SMC-3(6):610–621.
40. Galloway MM. Texture analysis using gray level run lengths. Computer Graphics and Image Processing 1975; 4(2):172–179.
41. Winsberg F, Elkin M, Macy J, Bordaz V, Weymouth W. Detection of Radiographic Abnormalities in Mammograms by Means of Optical Scanning and Computer Analysis. Radiology 1967; 89(2):211–215.
42. Giger ML, Chan H-P, Boone J. Anniversary Paper: History and status of CAD and quantitative image analysis: The role ofMedical Physicsand AAPM. Medical Physics 2008; 35(12):5799–5820.
43. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D,

Hoebers F, Rietbergen MM, Leemans CR, Dekker A, Quackenbush J, Gillies RJ, Lambin P. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications 2014; 5(4006):1–8.
44. Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, Liang C, Tian J, Liang C. Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non—Small Cell Lung Cancer. Radiology 2016; 281(3):947–957.
45. Wilson R, Devaraj A. Radiomics of pulmonary nodules and lung cancer. Translational lung cancer research 2017; 6(1):86–91.
46. Rahbar H, McDonald ES, Lee JM, Partridge SC, Lee CI. How Can Advanced Imaging Be Used to Mitigate Potential Breast Cancer Overdiagnosis? Academic Radiology 2016; 23(6):768–773.
47. Parekh V, Jacobs MA. Radiomics: a new application from established techniques. Expert Review of Precision Medicine and Drug Development 2016; 1(2):207–226.
48. Iranmakani S, Mortezazadeh T, Sajadian F, Ghaziani MF, Ghafari A, Khezerloo D, Musa AE. A review of various modalities in breast imaging: technical aspects and clinical outcomes. Egyptian Journal of Radiology and Nuclear Medicine 2020; 51(57):1–22.
49. Gilbert FJ, Pinker-Domenig K. Diagnosis and Staging of Breast Cancer: When and How to Use Mammography, Tomosynthesis, Ultrasound, Contrast-Enhanced Mammography, and Magnetic Resonance Imaging. In: Hodler J, Kubik-Huch RA, von Schulthess GK, eds. Diseases of the Chest, Breast, Heart and Vessels 2019–2022, Cham, Switzerland: Springer International Publishing; 2019:155–166.
50. Bahl M. Detecting Breast Cancers with Mammography: Will AI Succeed Where Traditional CAD Failed? Radiology 2019; 290(2):315–316.
51. Gao Y, Geras KJ, Lewin AA, Moy L. New Frontiers: An Update on Computer-Aided Diagnosis for Breast Imaging in the Age of Artificial Intelligence. American Journal of Roentgenology 2019; 212 (2):300–307.
52. Geras KJ, Mann RM, Moy L. Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives. Radiology 2019; 293(2):246–259.
53. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Scientific Reports 2019; 9(12495):1–12.
54. Batchu S, Liu F, Amireh A, Waller J, Umair M. A Review of Applications of Machine Learning in Mammography and Future Challenges. Oncology 2021; 99(8):483–490.
55. Noble M, Bruening W, Uhl S, Schoelles K. Computer-aided detection mammography for breast cancer screening: systematic review and meta-analysis. Archives of Gynecology and Obstetrics 2009; 279 (6):881–890.
56. Danala G, Patel B, Aghaei F, Heidari M, Li J, Wu T, Zheng B. Classification of Breast Masses Using a Computer-Aided Diagnosis Scheme of Contrast Enhanced Digital Mammograms. Annals of Biomedical Engineering 2018; 46(9):1419–1431.
57. Patel BK, Ranjbar S, Wu T, Pockaj BA, Li J, Zhang N, Lobbes M, Zhang B, Mitchell JR. Computer-aided diagnosis of contrast-enhanced spectral mammography: A feasibility study. Eur J Radiol 2018; 98:207–213.
58. Jochelson MS, Lobbes MBI. Contrast-enhanced Mammography: State of the Art. Radiology 2021; 299(1):36–48.
59. Fusco R, Piccirillo A, Sansone M, Granata V, Rubulotta MR, Petrosino T, Barretta ML, Vallone P, Di Giacomo R, Esposito E, Di Bonito M, Petrillo A. Radiomics and Artificial Intelligence Analysis with Textural Metrics Extracted by Contrast-Enhanced Mammography in the. Breast Lesions Classification. Diagnostics 2021; 11(5):815.
60. Ye D-M, Wang H-T, Yu T. The Application of Radiomics in Breast MRI: A Review. Technology in Cancer Research & Treatment 2020; 19:1–16.
61. Sollini, M., L. Cozzi, G. Ninatti, L. Antunovic, L. Cavinato, A. Chiti, and M. Kirienko, PET/CT radiomics in breast cancer: Mind the step. Methods, 2021. 188: p. 122-132.
62. Grimm LJ, Mazurowski MA. Breast Cancer Radiogenomics: Current Status and Future Directions. Acad Radiol 2020; 27(1):39–46.
63. Mendes J, Matela N. Breast Cancer Risk Assessment: A Review on Mammography-Based Approaches. Journal of Imaging 2021; 7(6):98.
64. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MDJ, Horsley T, Weeks L, Hempel S, Akl EA, Chang C, McGowan J, Stewart L, Hartling L, Aldcroft A, Wilson MG, Garritty C, Lewin S, Godfrey CM, Macdonald MT, Langlois EV, Soares-Weiser K, Moriarty J, Clifford T, Tunçalp Ö, Straus SE. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Annals of Internal Medicine 2018; 169(7):467–473.
65. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology 2005; 8(1):19–32.
66. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Medical Research Methodology 2018; 18(143):1–7.
67. Peters MD, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015; 13(3):141–146.
68. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, McInerney P, Godfrey CM, Khalil H. Updated methodological guidance for the conduct of scoping reviews. JBI Evidence Synthesis 2020; 18 (10):2119–2126.
69. Tamez-Peña J-G, Rodriguez-Rojas J-A, Gomez-Rueda H, Celaya-Padilla J-M, Rivera-Prieto R-A, Palacios-Corona R, Garza-Montemayor M, Cardona-Huerta S, Treviño V. Radiogenomics analysis identifies correlations of digital mammography with clinical molecular signatures in breast cancer. PLOS ONE 2018; 13(3):e0193871.
70. Mao N, Yin P, Li Q, Wang Q, Liu M, Ma H, Dong J, Che K, Wang Z, Duan S, Zhang X, Hong N, Xie H. Radiomics nomogram of contrast-enhanced spectral mammography for prediction of axillary lymph node metastasis in breast cancer: a multicenter study. European Radiology 2020; 12:6732–6739.
71. Tagliafico AS, Bignotti B, Rossi F, Matos J, Calabrese M, Valdora F, Houssami N. Breast cancer Ki-67 expression prediction by digital breast tomosynthesis radiomics features. European Radiology Experimental 2019; 3(36):1–6.
72. Cui WJ, Wang C, Jia L, Ren S, Duan SF, Cui C, Chen X, Wang ZQ. Differentiation Between G1 and G2/G3 Phyllodes Tumors of Breast Using Mammography and Mammographic Texture Analysis. Frontiers in Oncology 2019; 9(433):1–9.
73. Marino MA, Leithner D, Sung J, Avendano D, Morris EA, Pinker K, Jochelson MS. Radiomics for Tumor Characterization in Breast Cancer Patients: A Feasibility Study Comparing Contrast-Enhanced Mammography and Magnetic Resonance Imaging. Diagnostics 2020; 10(492):1–11.
74. Tan H, Wu Y, Bao F, Zhou J, Wan J, Tian J, Lin Y, Wang M. Mammography-based radiomics nomogram: a potential biomarker to predict axillary lymph node metastasis in breast cancer. British Journal of Radiology 2020; 93(1111):1–11.
75. Li J, Song Y, Xu S, Wang J, Huang H, Ma W, Jiang X, Wu Y, Cai H, Li L. Predicting underestimation of ductal carcinoma in situ: a comparison between radiomics and conventional approaches. International Journal of Computer Assisted Radiology and Surgery 2019; 14(4):709–721.
76. Son J, Lee SE, Kim E-K, Kim S. Prediction of breast cancer molecular subtypes using radiomics signatures of synthetic mammography from digital breast tomosynthesis. Scientific Reports 2020; 10(21566):1–11.
77. Zhang H-X, Sun Z-Q, Cheng Y-G, Mao G-Q. A pilot study of radiomics technology based on X-ray mammography in patients with triple-negative breast cancer. Journal of X-Ray Science & Technology 2019; 27 (3):485–492.
78. Wang L, Yang W, Xie X, Liu W, Wang H, Shen J, Ding Y, Zhang B, Song B. Application of digital mammography-based radiomics in the differentiation of benign and malignant round-like breast tumors and the prediction of molecular subtypes. Gland Surgery 2020; 9(6):2005–2016.
79. Yu H, Meng X, Chen H, Han X, Fan J, Gao W, Du L, Chen Y, Wang Y, Liu X, Zhang L, Ma G, Yang J. Correlation Between Mammographic Radiomics Features and the Level of Tumor-Infiltrating Lymphocytes in Patients With Triple-Negative Breast Cancer. Frontiers in Oncology 2020; 10(412):1–9.
80. Mazurowski MA. Radiogenomics: what it is and why it is important. Journal of the American College of Radiology 2015; 12(8):862–866.
81. Naoi, Y. and S. Noguchi, Multi-gene classifiers for prediction of recurrence in breast cancer patients. 2016. 23(1): p. 12-18.
82. Skaane P, Bandos AI, Niklason LT, Sebuødegård S, Østerås BH, Gullien R, Gur D, Hofvind S. Digital Mammography versus Digital Mammography Plus Tomosynthesis in Breast Cancer Screening: The Oslo Tomosynthesis Screening Trial. Radiology 2019; 291(1):23–30.
83. Boroumand G, Teberian I, Parker L, Rao VM, Levin DC. Screening Mammography and Digital Breast Tomosynthesis: Utilization Updates. American Journal of Roentgenology 2018; 210(5):1092–1096.

84. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. Journal of Magnetic Resonance Imaging 2019; 49 (4):939–954.
85. Parekh VS, Jacobs MA. Deep learning and radiomics in precision medicine. Expert Review of Precision Medicine and Drug Development 2019; 4(2):59–72.
86. Sun Q, Lin X, Zhao Y, Li L, Yan K, Liang D, Sun D, Li Z-C. Deep Learning vs. Radiomics for Predicting Axillary Lymph Node Metastasis of Breast Cancer Using Ultrasound Images: Don't Forget the Peritumoral Region. Frontiers in Oncology 2020; 10(53):1–12.
87. Yang X, Wu L, Zhao K, Ye W, Liu W, Wang Y, Li J, Li H, Huang X, Zhang W, Huang Y, Chen X, Yao S, Liu Z, Liang C. Evaluation of human epidermal growth factor receptor 2 status of breast cancer using preoperative multidetector computed tomography with deep learning and handcrafted radiomics features. Chin J Cancer Res 2020; 32(2):175–185.
88. Li Z, Wang Y, Yu J, Guo Y, Cao W. Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. Scientific Reports 2017; 7(5467):1–11.
89. Zheng X, Yao Z, Huang Y, Yu Y, Wang Y, Liu Y, Mao R, Li F, Xiao Y, Wang Y, Hu Y, Yu J, Zhou J. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nature Communications 2020; 11(1236):1–9.
90. Liu W, Cheng Y, Liu Z, Liu C, Cattell R, Xie X, Wang Y, Yang X, Ye W, Liang C, Li J, Gao Y, Huang C, Liang C. Preoperative Prediction of Ki-67 Status in Breast Cancer with Multiparametric MRI Using Transfer Learning. Academic Radiology 2021; 28(2):E44–E53.
91. Jiang M, Zhang D, Tang S-C, Luo X-M, Chuan Z-R, Lv W-Z, Jiang F, Ni X-J, Cui X-W, Dietrich CF. Deep learning with convolutional neural network in the assessment of breast cancer molecular subtypes based on US images: a multicenter retrospective study. European Radiology, 2021; 31(6):3673–3682.
92. Guo X, Liu Z, Sun C, Zhang L, Wang Y, Li Z, Shi J, Wu T, Cui H, Zhang J, Tian J, Tian J. Deep learning radiomics of ultrasonography: Identifying the risk of axillary non-sentinel lymph node involvement in primary breast cancer. EBioMedicine 2020; 60(103018):1–11.
93. Choudhery S, Gomez-Cardona D, Favazza CP, Hoskin TL, Haddad TC, Goetz MP, Boughey JC. MRI Radiomics for Assessment of Molecular Subtype, Pathological Complete Response, and Residual Cancer Burden in Breast Cancer Patients Treated With Neoadjuvant Chemotherapy. Academic Radiology 2020: 1–10. S1076-6332(20)30607-3 In press.
94. Yang L, Yang J, Zhou X, Huang L, Zhao W, Wang T, Zhuang J, Tian J. Development of a radiomics nomogram based on the 2D and 3D CT features to predict the survival of non-small cell lung cancer patients. European Radiology 2019; 29(5):2196–2206.
95. Zhou J, Tan H, Li W, Liu Z, Wu Y, Bai Y, Fu F, Jia X, Feng A, Liu H, Wang M. Radiomics Signatures Based on Multiparametric MRI for the Preoperative Prediction of the HER2 Status of Patients with Breast Cancer. Academic Radiology 2020; 28(10):1352–1360.
96. Jacobs MA, Umbricht CB, Parekh VS, El Khouli RH, Cope L, Macura KJ, Harvey S, Wolff AC. Integrated Multiparametric Radiomics and Informatics System for Characterizing Breast Tumor Characteristics with the OncotypeDX Gene Assay. Cancers 2020; 12(10):2772.
97. Wu M, Zhong X, Peng Q, Xu M, Huang S, Yuan J, Ma J, Tan T. Prediction of molecular subtypes of breast cancer using BI-RADS features based on a "white box" machine learning approach in a multi-modal imaging setting. Eur J Radiol 2019; 114:175–184.
98. Li Y, Liu Y, Zhang M, Zhang G, Wang Z, Luo J. Radiomics With Attribute Bagging for Breast Tumor Classification Using Multimodal Ultrasound Images. Journal of Ultrasound in Medicine 2020; 39 (2):361–371.
99. Chen S, Guan X, Shu Z, Li Y, Cao W, Dong F, Zhang M, Shao G, Shao F. A New Application of Multimodality Radiomics Improves Diagnostic Accuracy of Nonpalpable Breast Lesions in Patients with Microcalcifications-Only in Mammography. Medical Science Monitor 2019; 25:9786–9793.
100. Huang S-Y, Franc BL, Harnish RJ, Liu G, Mitra D, Copeland TP, Arasu VA, Kornak J, Jones EF, Behr SC, Hylton NM, Price ER, Esserman L, Seo Y. Exploration of PET and MRI radiomic features for decoding breast cancer phenotypes and prognosis. npj Breast Cancer 2018; 4(24):1–13.
101. Castaldo R, Pane K, Nicolai E, Salvatore M, Franzese M. The Impact of Normalization Approaches to Automatically Detect Radiogenomic Phenotypes Characterizing Breast Cancer Receptors Status. Cancers 2020; 12(518):1–25.
102. Gallivanone F, Cava C, Corsi F, Bertoli G, Castiglioni I. In Silico Approach for the Definition of radiomiRNomic Signatures for Breast Cancer Differential Diagnosis. International Journal of Molecular Sciences 2019; 20(5825):1–16.
103. Lu M, Zhan X. The crucial role of multiomic approach in cancer research and clinically relevant outcomes. EPMA Journal 2018; 9(1):77–102.
104. Gatta R, Depeursinge A, Ratib O, Michielin O, Leimgruber A. Integrating radiomics into holomics for personalised oncology: from algorithms to bedside. European Radiology Experimental 2020; 4(11):1–9.
105. Li L, Roth R, Germaine P, Ren S, Lee M, Hunter K, Tinney E, Liao L. Contrast-enhanced spectral mammography (CESM) versus breast magnetic resonance imaging (MRI): A retrospective comparison in 66 breast lesions. Diagn Interv Imaging 2017; 98(2):113–123.
106. Lewin J. Comparison of Contrast-Enhanced Mammography and Contrast-Enhanced Breast MR Imaging. Magn Reson Imaging Clin N Am 2018; 26(2):259–263.

2.9 Chapter findings

The findings from this literature review (Chapter 2) revealed a potential role of mammography-based local radiomics analysis in describing breast cancer characteristics and predicting patient outcomes. Radiomic features such as Gray Level Co-Occurrence Matrix (GLCM) and First Order Statistics (FOS) were discovered to be the frequently used features for describing the mammographic characteristics. Studies found in the literature searches (including additional search conducted on 02 July 2024, as per section 2.1) also uncovered a scarcity of research on using mammographic features derived from global radiomics analysis in predicting global gist signals and difficult normal cases of radiologists and radiology trainees.

Chapter 3 – Computer-extracted global radiomic features can predict the radiologists' first impression about the abnormality of a screening mammogram

3.1 Chapter introduction

From the knowledge deficiencies and the **objective 2** described in the introduction (Chapter 1) as well as the findings from the literature review (Chapter 2), it was identified that there was a potential value of using local radiomics analysis in describing mammographic appearances. However, limited studies (based on the literature searches) had been conducted using global mammographic radiomic features to explain the characteristics of the global gist signal which are imperative for radiologists' mammogram interpretation and skill development relating to challenging cancer cases and diagnostic errors. The literature review also showed that radiomic features such as texture-based GLCM and FOS were commonly used for describing breast cancer's characteristics, recurrence, and survival. Following these findings, this chapter (Chapter 3) therefore looks at whether a set of global radiomic features from mammograms involving GLCM and FOS features can predict the global gist signal (i.e., differentiating *high-gist* when readers detect a very strong gist from *low-gist* images when the gist of abnormal was not/poorly perceived). Important features that drive the signals are also explored.

The first original research study in this chapter has been published as:

Siviengphanom, S., S.J. Lewis, P.C. Brennan, and Z. Gandomkar, *Computer-extracted global radiomic features can predict the radiologists' first impression about the abnormality of a screening mammogram.* British Journal of Radiology, 2024. **97**: p. 168–179. https://doi.org/10.1093/bjr/tqad025

Conference proceeding/presentation based on a smaller dataset related to the above study has been published as/presented at:

- Siviengphanom, S., S.J. Lewis, P.C. Brennan, Z. Gandomkar. *Predicting the gist of breast cancer on a screening mammogram using global radiomic features*. in *SPIE Medical Imaging 2024: Image Perception, Observer Performance, and Technology Assessment*. 2024. San Diego, California, United States: SPIE. https://doi.org/10.1117/12.3005470 (Appendix 1.3).

- SPIE Medical Imaging Conference 2024, San Diego, CA, United States, February 2024 (Appendix 1.4).

BJR, 2024, **97**, 168–179
https://doi.org/10.1093/bjr/tqad025
Advance access publication: 12 December 2023
Research Article

Computer-extracted global radiomic features can predict the radiologists' first impression about the abnormality of a screening mammogram

Somphone Siviengphanom , MEd[1,*], Sarah J. Lewis, PhD[1], Patrick C. Brennan, PhD[1], Ziba Gandomkar, PhD[1]

[1]Medical Image Optimisation and Perception Group, Discipline of Medical Imaging Science, Sydney School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW 2006, Australia

*Corresponding author: Somphone Siviengphanom, MEd, Medical Image Optimisation and Perception Group, Discipline of Medical Imaging Science, Sydney School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, Susan Wakil Health Building D18, Sydney, NSW, 2006, Australia (ssiv6387@uni.sydney.edu.au)

Abstract

Objective: Radiologists can detect the gist of abnormal based on their rapid initial impression on a mammogram (ie, global gist signal [GGS]). This study explores (1) whether global radiomic (ie, computer-extracted) features can predict the GGS; and if so, (ii) what features are the most important drivers of the signals.

Methods: The GGS of cases in two extreme conditions was considered: when observers detect a very strong gist (*high-gist*) and when the gist of abnormal was not/poorly perceived (*low-gist*). Gist signals/scores from 13 observers reading 4191 craniocaudal mammograms were collected. As gist is a noisy signal, the gist scores from all observers were averaged and assigned to each image. The *high-gist* and *low-gist* categories contained all images in the fourth and first quartiles, respectively. One hundred thirty handcrafted global radiomic features (GRFs) per mammogram were extracted and utilized to construct eight separate machine learning random forest classifiers (*All*, *Normal*, *Cancer*, *Prior-1*, *Prior-2*, *Missed*, *Prior-Visible*, and *Prior-Invisible*) for characterizing *high-gist* from *low-gist* images. The models were trained and validated using the 10-fold cross-validation approach. The models' performances were evaluated by the area under receiver operating characteristic curve (AUC). Important features for each model were identified through a scree test.

Results: The *Prior-Visible* model achieved the highest AUC of 0.84 followed by the *Prior-Invisible* (0.83), *Normal* (0.82), *Prior-1* (0.81), *All* (0.79), *Prior-2* (0.77), *Missed* (0.75), and *Cancer* model (0.69). *Cluster shade*, *standard deviation*, *skewness*, *kurtosis*, and *range* were identified to be the most important features.

Conclusions: Our findings suggest that GRFs can accurately classify *high-* from *low-gist* images.

Advances in knowledge: Global mammographic radiomic features can accurately predict *high-* from *low-gist* images with five features identified to be valuable in describing *high-gist* images. These are critical in providing better understanding of the mammographic image characteristics that drive the strength of the GGSs which could be exploited to advance breast cancer (BC) screening and risk prediction, enabling early detection and treatment of BC thereby further reducing BC-related deaths.

Keywords: radiomics; screening mammograms; gist; breast cancer; machine learning; global radiomic features; current mammograms; prior mammograms.

Introduction

Clinicians such as radiologists, the experts in the interpretation of medical images, often report that they have an initial impression about an abnormality of a medical image or its "gist" after momentarily glancing at an image. Several observer studies provided evidence suggesting that in less than a half-second, radiologists are capable of distinguishing abnormal mammograms[1] or chest X-rays from the normal ones.[2] A study focusing on mammographic images indicated that radiologists can distinguish abnormal from normal mammograms, containing no obvious localized sign of cancer, contralateral to the malignancies.[3] It was also shown that based on a half-second image presentation, the gist of the abnormal can be detected in prior normal breast images belonging to women who were later diagnosed with breast cancer (BC) (ie, missed cancers based on prior images from a previous screening round).[4-6] Moreover, previous studies showed that women with a false positive (FP) result on screening mammogram or who have had a cancer in one breast have elevated risk of developing a future/contralateral BC.[7,8] These findings imply that the initial impression of a radiologist does not necessarily have a localized source within the image but is possibly driven by the mammographic textural features and not associated with other risk factors such as breast density.[9-12]

In medical image perception literatures, the holistic perceptual processing of radiological instantaneous view of an image's abnormality features is often called the gist signal, which has been extensively studied to better understand radiological errors and enhance screening performance.[3-6,11-21] From the human visual system, one may claim that gist contains both global and local source of information about "what" and "where" the abnormality in an image may be, correspondingly.[12,22,23] A large body of gist literature, studying gist on images without a localized cancer sign such as contralateral mammograms, revealed that gist is associated with global "what" information occurring without prior location of that information.[3-5,13-15,23,24] Views from the basic vision science domains suggest that gist information can be rapidly extracted (eg, using just as small as a

Received: 31 May 2023; **Revised:** 7 August 2023; **Accepted:** 25 October 2023

half-second of an image viewing time) from the entire global image properties in a "non-selective" manner, while identification of the information occurs from the local properties (ie, the "selective" pathway).[16,17,25,26] With that, gist is evidently a global rapid signal which will be referred to as the "global gist signal" (GGS) in this article.

Numerous gist studies focusing on the process of understanding natural scenes suggest that only after a brief sight of an image, observers are capable of categorizing these images into classes such as indoor or outdoor.[27,28] The human visual system relies on texture features in recognizing the global gist of the natural scenes,[29-31] demonstrating that texture analysis can provide strong holistic cues for quick scene recognition. Previous studies have shown that first and second order (Gray Level Co-occurrence Matrix [GLCM]-based) statistical features are relevant for texture analysis in typical viewing modes.[32-34] However, it is unclear whether these features are also relevant for the perceived GGS of a mammogram, which is extracted briefly after image onset based on a half-second image presentation. Our hypothesis is that a set of global radiomic features (GRFs), including both first order statistics (FOS) and GLCM-based, can differentiate between mammographic textures that result in strong gist signals (ie, *high-gist*) and those that result in weak ones (ie, *low-gist*). Therefore, this experimental study aims to (i) explore whether a set of GRFs comprising FOS and GLCM-based statistics extracted from mammograms can predict *high-* from *low-gist* images and (ii) identify the most important GRFs that drive the GGS. We focus only on these two extreme sides of the gist spectrum (highest and lowest gist images) to ensure that a clear difference between them at their maximum level is achieved. Moreover, mammographic images' textural features have been linked to the presence of BC[35,36] and elevated BC risk.[37] Therefore, we included images of normal cancer-free and cancer-containing cases, and high-risk individuals, and considered each class separately to ensure that our model is not capturing the signal related to the presence of BC or elevated BC risk, but rather capturing the gist signal.

In addition to enhancing our comprehension of the gist signal's nature when dealing with complex textures, the present study could yield insights into radiological errors in screening mammograms. Identifying features on normal images which signal a strong gist of the abnormal (FP) and missed cancer cases which signal a weak gist (false negative [FN]) could help in educating radiologists about possible appearances of FP and FN cases so that their diagnostic errors could be better monitored and drastically reduced. Since BC is the major cause of cancer-related death in females worldwide and the efficiency of mammographic screening interpretation depends largely on the visual expertise of clinician radiologists, errors such as FP and FN are significant.[38] While FP errors can trigger huge annual healthcare expenses of almost US$3 billion (an example from the United States) and an increased patient anxiety and depression, FN mistakes could result in increased treatment complications and reduced overall survival due to delayed detection of BC.[39,40] Hence, providing a better understanding about the underlying reasons for diagnostic errors in the interpretation of screening mammograms is crucially important.

Materials and methods

Institutional ethics approval of all experiment procedures was obtained from the Human Research Ethics Committee of the University of Sydney (protocol no. 2020/324). Informed consent from all subjects participated in the study was acquired prior to the data collection.

Participants

Seventeen Australian and New Zealand radiologists and breast physicians were invited to participate in this study. These readers were identified as the "gist expert", through their participation in the previous gist experiment[15] where they demonstrated superior ability to detect reliable and accurate gist of the abnormal in mammographic images. Full details of their selection and identification were described in a previous study[15] which also showed that the superior ability of these "gist experts" was not associated with factors such as their years of experience reading mammograms or number of mammograms interpreted each week. A total of 13 readers (11 radiologists and 2 breast physicians) were recruited.

Prior to the data collection, the 13 participants' characteristics and general workload details (Table 1) were collected via an online questionnaire. Majority of the participants were female (77%) radiologists (85%) working part-time (69%) as screen readers (92%) for the national screening program BreastScreen Australia and did not undertake a fellowship (62%) but had association with a university or educational institute (54%). At the time of participating in the gist experiment, on average, the participants spent 9.5 h per week reading mammograms and read 237 screening cases per week. With that, 70% of their time was dedicated to reading breast images, while 24.6% of their time was dedicated to reading diagnostic mammograms. The participants also had a median of 17.5 years being a certified BreastScreen reader, 21 years registered as a screen reader, and 21.5 years reading mammograms. Of those who had a fellowship training in mammography reading, the median duration of their training program was 6 months in which they completed it on average of 10 years ago. Lastly, on average, the participants performed 250 breast biopsy examinations in the last 12 months with a median of four times in a month correlating/reviewing radiology-pathology findings for biopsy cases, and three times in a month participating in a Multi-Disciplinary Meeting.

Mammographic images

For the purpose of conducting this gist experiment and radiomics analysis, mammographic images and binary (black and white) masks were required. A dataset of de-identified 4191 craniocaudal (CC) unilateral digital imaging and communications in medicine (DICOM) mammograms were obtained from the archive of BreastScreen Australia, which were acquired from asymptomatic women aged between 50 and 75. These images were obtained using a variety of vendors, that is, Siemens (Munich, Germany), Hologic (Hologic, Inc., Marlborough, MA, USA), Sectra (Sectra, Linköping, Sweden), Fujifilm (Fujifilm Corporation, Minato City, Tokyo, Japan), Philips (Philips Healthcare, Amsterdam, the Netherlands), and Konica Minolta (Marunouchi, Chiyoda, Tokyo, Japan). Next, binary masks were generated from 4191 DICOM images to extract breast area from its background using a standard thresholding gray level intensity value of 100. Due to computational efficiency reasons (to save computational resources, ie, time and memory), DICOM images and masks were then converted to TIFF

Table 1. Participants' characteristics and general workload details at the time of participating in the gist experiment ($n = 13$).

Characteristics			Count (%)
Gender (female, male)			10 (77%), 3 (23%)
Discipline (radiologist, breast physician)			11 (85%), 2 (15%)
Working full-time (yes, no)			4 (31%), 9 (69%)
Being a screen reader (yes, no)			12 (92%), 1 (8%)
Fellowship-trained in mammography reading (yes, no)			5 (38%), 8 (62%)
Whether affiliated with a university or educational institute (yes, no)			7 (54%), 6 (46%)
	Median	**Min-Max**	**25th-75th percentile**
No. of hours per week currently spent in reading mammograms	9.5	1.0-30.0	3.0-19.0
No. of screening cases read per week	237	5-800	150-400
Percentage of time dedicated to reading breast images	70	5-95	45-78
Percentage of time dedicated to reading diagnostic mammograms	24.6	0-90.2	9.0-35.0
No. of years certified as a BreastScreen reader	17.5	0.5-22	7.5-20
No. of years registered as a screen reader	21	2-30	12-25
No. of years reading mammograms	21.5	0-34	8.5-25
Duration of fellowship training in mammography reading (months)	6	6-12	6-12
No. of years since completing fellowship training in mammography reading	10	5-16	0-20.0
No. of breast biopsy examinations performed in the last 12 months	250	30-100	77-400
Frequency of correlating/reviewing radiology-pathology findings for biopsy cases in a month	4	1-100	4-40
Frequency of participating in a Multi-Disciplinary Meeting in a month	3	1-6	3-4

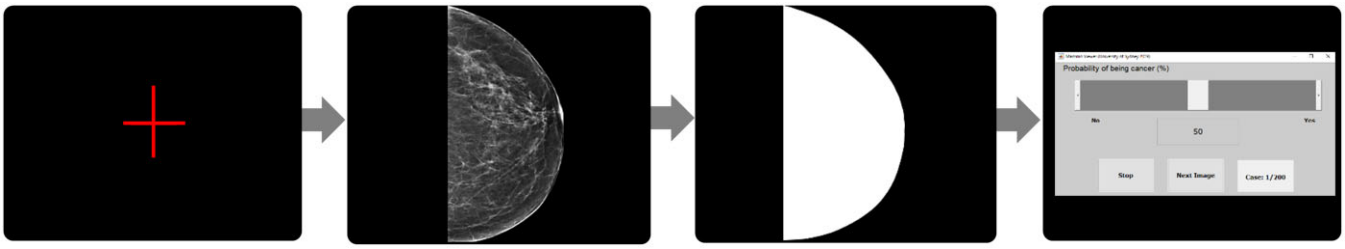


Figure 1. Gist experiment. For a half-second each, a red cross sign was firstly displayed in the centre of the screen, followed by a mammogram and then a corresponding white mask representing the breast area appeared to ensure conclusion of image visual processing. Lastly, without a time limit, a rating screen was shown to collect a gist score (ie, the probability of the image being abnormal) based on a scale of 0-100 (0 means completely confidence that the image was normal while100 means absolute confidence that the image was abnormal).

format with all the right CC images being flipped to the left for consistent left side chest wall on all images. The TIFF images and binary masks were then cropped based on the maximum breast region size and used for the gist experiment (Fig. 1) and radiomics analysis (Fig. 2).

Since the gist of the abnormal has been reported to be presented in both current and prior images[3-6,11,13-16] and to include various levels of gist strength in our analysis, we included all the eight types of the mammograms in our dataset. The eight types were "Normal" ($n = 1409$), "Cancer" ($n = 923$), "Prior-1" ($n = 1438$), "Prior-2" ($n = 419$), "Prior-3" ($n = 2$), "Missed" ($n = 375$), "Prior-Visible" ($n = 308$), and "Prior-Invisible" ($n = 755$). The "Normal" category included current mammograms of women reported and confirmed to be cancer-free normal by at least two independent expert radiologists (over 20 years of experience) with negative follow up mammograms acquired 2 years after. The "Cancer" had current cancer images contained biopsy proven malignancy. The "Prior-1", "Prior-2", and "Prior-3" contained prior no visible cancer signs mammograms, obtained 2, 4, and 6 years before current cancer mammograms, correspondingly. The "Missed", "Prior-Visible", and "Prior-Invisible" were categorized by two expert radiologists or a third expert radiologist as an arbitrator if disagreement occurred between the two expert radiologists,[41] using "Prior-1" category images. These three categories, respectively, included prior mammograms with actional (ie, recall), non-actional (ie, no recall), and no visible cancer signs that were reported as normal, but a later screen showed a biopsy-proven cancer.

Collecting gist scores

To collect the gist signals/scores from the 13 participants, a previously developed multi-observer gist experimental protocol[4] was used. Participants were asked to assess the 4191 CC images in 11 batches (381 images per batch selected randomly) and provide the probability of the image being abnormal/gist score based on a scale of 0 to 100 (0 = completely confidence the image was normal while 100 = absolute confidence that the image was abnormal). An extensive description of the data collection process (Fig. 1) was provided in a previous study.[42] Considering gist has low intra and inter-observer variability, the collected gist scores from the 13 radiologists (who were experts and consistent with each other in their gist scores) were then averaged and used for the analysis, obtaining one gist score per image and eliminating the noise of the gist signal.[11,14]

Categorizing high- vs low-gist images

To identify *high-* vs *low-gist* images, the averaged gist score per image was used to classify images into *high-* and *low-gist* based on the 75th and 25th percentiles of the images

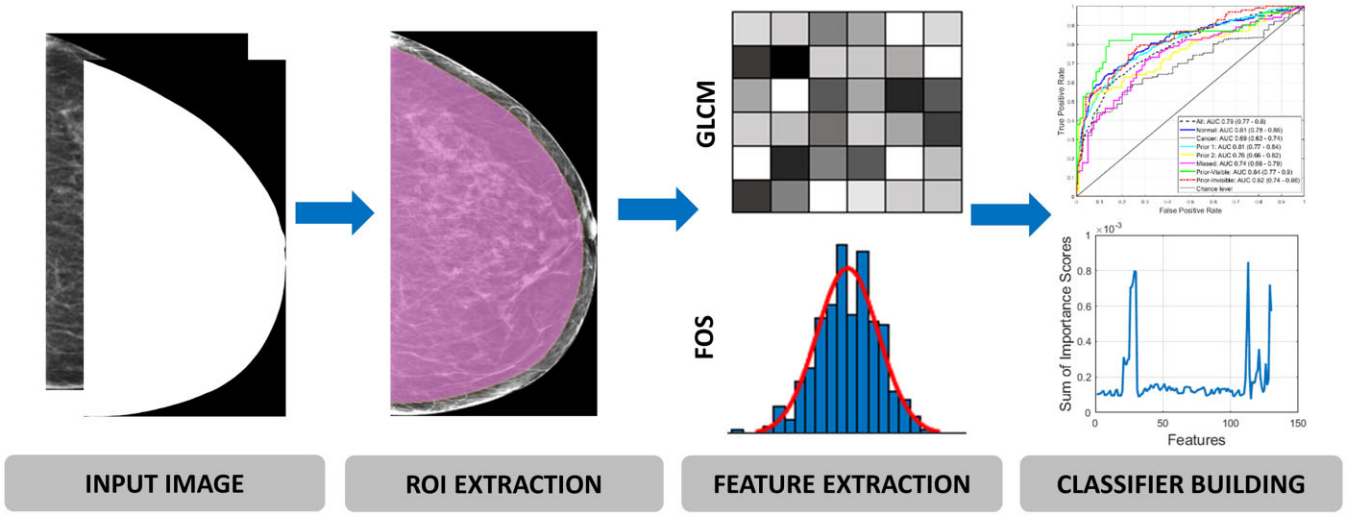


Figure 2. Radiomics analysis workflow. The first step was to obtain input images (unilateral craniocaudal mammograms and masks). Next, a ROI (pink region) was identified using MATLAB's image erosion algorithm to omit skin-air region. A set of 130 (110 GLCM-based and 20 FOS-based) global radiomic features per image were then computed and finally used to build eight machine learning classification models for predicting *high-gist* from *low-gist* images. Abbreviations: FOS = first order statistics; GLCM = gray level co-occurrence matrix; ROI = region of interest.

containing the highest and lowest gist scores, correspondingly. This gave a total of 1048 *high-* and 1048 *low-gist* images (namely "All" image type, including "Prior-3" images which had insufficient number of images, $n = 2$, to form its own group), 162 *high-* and 497 *low-gist* for "Normal", 545 *high-* and 71 *low-gist* for "Cancer", 268 *high-* and 370 *low-gist* for "Prior-1", 73 *high-* and 110 *low-gist* for "Prior-2", 117 *high-* and 80 *low-gist* for "Missed", 59 *high-* and 68 *low-gist* for "Prior-Visible", 92 *high-* and 222 *low-gist* for "Prior-Invisible" image type (Table 2). To accentuate an obvious difference between the highest and lowest gist images, only images in these upper and lower quartiles were used in the analysis.

Radiomics and statistical analysis

Using the 4191 images and masks, a set of 130 GRFs per image was extracted using the handcrafted features approach[43] with our in-house MATLAB-based programs. This is based on the region of interest identified through MATLAB's binary erosion algorithm using disk-shaped structuring element of 100 radius (distance from the origin to the edge of the disk),[44-46] to exclude the vendor post-processed background skin-air region in the mammogram in order to obtain the true breast tissue region.[47] Erosion algorithm was chosen because it is one of most reliable methods of the mathematical morphological image processing which has been widely used for exacting ROI and eliminating the noises in an image.[46] These features included 110 GLCM-based texture and 20 FOS-based features which have been shown to be valuable descriptors of mammographic appearances in measuring the contrast values of spatial inter-relationships between neighbouring pixels[48-50] and the distribution of single pixel intensity values within the image region of interest, correspondingly[49,50] (Table 3).

To predict *high-* from *low-gist* images, the 130 GRFs, derived from the corresponding image types ("All", "Normal", "Cancer", "Prior-1", "Prior-2", "Missed", "Prior-Visible", and "Prior-Invisible"—Table 2), were then used to build eight separate machine learning (ML) random forest classifiers (*All*, *Normal*, *Cancer*, *Prior-1*, *Prior-2*, *Missed*, *Prior-Visible*, and *Prior-Invisible*) based on the ensemble of 500 decision trees using bootstrap aggregation method.[51,52] The random forest with bootstrap aggregation technique was chosen because of its built-in feature selection function that can minimize feature-overfitting and imbalanced data issues (ie, oversampling unique observations from the minority class to balance the data), and automatic estimation of feature importance for generating interpretable classifier.[53,54] Useful features for each predictive model were also recognized through feature important analysis of their importance scores estimated using MATLAB's predictor importance algorithm. To determine the useful GRFs based their total importance scores from each model, a scree test of exploratory factor analysis[55] was utilized. Importance scores show the usefulness of each feature in constructing the decision trees within the model, with larger values mean more impact on the predictions made by the model.

In evaluating the performance of the models, we trained and validated each model using the 10-fold cross-validation approach,[53,56] a reliable, unbiased, and accurate validation method for estimating the model's generalization performance. This involved randomly splitting the dataset into 10 groups, and while the first group was used once to test the predictive performance of the model, the rest of the groups were used to train the model. This process was repeated 10 times until each group was used once as a test set. The area under the receiver operating characteristic curve (AUC) was then used to evaluate the overall performance of the models for distinguishing *high-* from *low-gist* images.

MATLAB R2022a (MathWorks, Natick, MA, USA) was used to complete all radiomics and statistical analysis.

Results

GRFs predicting high- vs low-gist images

The overall performance of the eight models (*All*, *Normal*, *Cancer*, *Prior-1*, *Prior-2*, *Missed*, *Prior-Visible*, *Prior-Invisible*) is shown in Fig. 3. When differentiating *high-* from *low-gist* images, the *Prior-Visible* model reached the highest AUC of 0.84 (95% CI, 0.77-0.91) followed by the *Prior-*

Table 2. Image types included in the study ($n = 8$).

No.	Image type	Description	No. of high-gist images	No. of low-gist images	Total
1	All	Included all the image categories	1048	1048	2096
2	Normal	Current mammograms of women reported and confirmed to be cancer-free normal by at least two independent expert radiologists (20+ years of experience) with negative follow-up mammograms acquired two years after	162	497	659
3	Cancer	Current cancer mammograms of women contained biopsy proven malignancy	545	71	616
4	Prior-1	Prior mammograms with no visible cancer signs acquired two years before current cancer mammograms	268	370	638
5	Prior-2	Prior mammograms with no visible cancer signs acquired four years before current cancer mammograms	73	110	183
6	Missed	Prior-1 mammograms with actionable visible cancer signs (ie, recall) that were reported as normal, but a later screen showed a biopsy-proven cancer	117	80	197
7	Prior-visible	Prior-1 mammograms with non-actionable (ie, not recall) yet visible cancer signs that were reported as normal, but a later screen showed a biopsy-proven cancer	59	68	127
8	Prior-invisible	Prior-1 mammograms with no visible cancer signs that were reported as normal, but a later screen showed a biopsy-proven cancer	92	222	314

Invisible (0.83, 0.77-0.87), *Normal* (0.82, 0.79-0.86), *Prior-1* (0.81, 0.79-0.85), *All* (0.79, 0.77-0.81), *Prior-2* (0.77, 0.68-0.83), *Missed* (0.75, 0.70-0.81), and *Cancer* model (0.69, 0.65-0.75).

Important GRFs

Table 3 and Fig. 4 show the feature importance scores of the 130 GRFs derived from each of the eight models. When comparing each feature's importance score, ranked from high to low, in each individual model (Fig. 5), five features appear to be significant for most of the models. These features were *cluster shade* (feature no. 26-30, calculated across five orientations) of the GLCM,[48-50] *standard deviation* (feature no. 112), *skewness* (feature no. 113), *kurtosis* (feature no. 114), and *range* (feature no. 129 and 130, calculated across two dimensions) of the FOS.[49,50] *Cluster shade* is a measure of the symmetry distribution in an image. *Standard deviation* describes the variation or spreading of gray level intensity from the mean value. *Skewness* is a measure of the unevenness of the distribution. *Kurtosis* measures the extreme values of the distribution in the image. Also, *range* describes the difference between 95th and 5th percentile (feature no. 129), and 99th and 1st percentile of image gray level values (feature no. 130).

Discussion

For almost 50 years, research has shown that the holistic rapid visual impression of an image (ie, GGS) can convey a great amount of useful information about an image abnormality in the medical image interpretation and diagnostic process, aiming to significantly reduce FN and FP errors.[2,18,26] Recent medical image perception studies have recognized the existence of GGS shown to be valuable for providing information about the presence of BC to facilitate the improvement of early BC detection and future risk prediction.[3-6,13-16] However, little is known about which global computer extracted image features are correlated to the intensity of the GGS.[21] Here, for the first time, using radiomics approach,[43] we built upon the previous works and evaluated GRFs that drive the strength of the gist signals. We hypothesized that GRFs of *high-gist* mammograms are different from *low-gist* ones, based on the efficacy of radiomics in mammography[43] and the GGS, independent of lesion characteristics (eg, breast density), in identifying image abnormality.[3-5,11,13,14,16,24] We also examined which features are more important for distinguishing between *high-gist* mammograms from *low-gist* ones.

The results obtained from our eight ML classification models (*All*, *Normal*, *Cancer*, *Prior-1*, *Prior-2*, *Missed*, *Prior-Visible*, *Prior-Invisible*) demonstrated that the GRFs had the ability to accurately differentiate *high-* from *low-gist* images with as high as an AUC of 0.84 (Fig. 3). Interestingly, although the GGS has been shown to be helpful for predicting cancer images,[4,13] the *Cancer* model in our study achieved smaller AUC value of 0.69 in discriminating *high-* from *low-gist* cancer mammograms when compared to other seven models (AUCs of 0.75-0.84). This could be because of the large imbalanced dataset ($n = 474$) between the *high-gist* ($n = 545$) and *low-gist* ($n = 71$) images in the "Cancer" image category, making it difficult for the *Cancer* model to optimally learn the difference, thereby limiting its predictive capability. Nevertheless, our overall promising findings (ie, up to 0.84 AUC) indicate the significance of the GRFs in predicting *high-gist* images which could potentially be used to guide

Table 3. List of global radiomic features ($n = 130$) with feature importance scores.

No.	Feature name	Importance scores from each model							
		All	Normal	Cancer	Missed	Prior-Vis	Prior-Invis	Prior1	Prior2
1	Autocorrelation_1	0.0001	0.0003	0.0003	0.0010	0.0015	0.0005	0.0003	0.0008
2	Autocorrelation_3	0.0001	0.0003	0.0003	0.0009	0.0015	0.0006	0.0003	0.0009
3	Autocorrelation_5	0.0001	0.0003	0.0003	0.0008	0.0013	0.0006	0.0004	0.0009
4	Autocorrelation_9	0.0001	0.0003	0.0003	0.0009	0.0015	0.0005	0.0003	0.0007
5	Autocorrelation_11	0.0001	0.0003	0.0003	0.0008	0.0016	0.0006	0.0004	0.0009
6	Contrast_1	0.0001	0.0003	0.0003	0.0015	0.0007	0.0008	0.0003	0.0018
7	Contrast_3	0.0001	0.0002	0.0003	0.0013	0.0013	0.0005	0.0003	0.0012
8	Contrast_5	0.0001	0.0002	0.0002	0.0009	0.0018	0.0006	0.0003	0.0007
9	Contrast_9	0.0001	0.0002	0.0003	0.0009	0.0019	0.0007	0.0003	0.0008
10	Contrast_11	0.0001	0.0002	0.0003	0.0009	0.0021	0.0006	0.0004	0.0007
11	Correlation_m_1	0.0001	0.0003	0.0002	0.0009	0.0009	0.0007	0.0004	0.0017
12	Correlation_m_3	0.0001	0.0002	0.0003	0.0008	0.0013	0.0006	0.0003	0.0009
13	Correlation_m_5	0.0001	0.0002	0.0003	0.0010	0.0016	0.0005	0.0003	0.0008
14	Correlation_m_9	0.0001	0.0002	0.0002	0.0009	0.0016	0.0005	0.0003	0.0010
15	Correlation_m_11	0.0001	0.0002	0.0003	0.0013	0.0020	0.0006	0.0004	0.0007
16	Correlation_p_1	0.0001	0.0003	0.0003	0.0010	0.0008	0.0007	0.0004	0.0019
17	Correlation_p_3	0.0001	0.0002	0.0003	0.0007	0.0013	0.0006	0.0003	0.0009
18	Correlation_p_5	0.0001	0.0002	0.0003	0.0010	0.0014	0.0005	0.0003	0.0008
19	Correlation_p_9	0.0001	0.0002	0.0002	0.0009	0.0016	0.0005	0.0004	0.0010
20	Correlation_p_11	0.0001	0.0002	0.0003	0.0013	0.0020	0.0006	0.0004	0.0007
21	Cluster_prominence_1	0.0003	0.0011	0.0006	0.0051	0.0161	0.0027	0.0016	0.0043
22	Cluster_prominence_3	0.0003	0.0013	0.0006	0.0050	0.0136	0.0028	0.0015	0.0040
23	Cluster_prominence_5	0.0003	0.0013	0.0005	0.0055	0.0124	0.0026	0.0017	0.0048
24	Cluster_prominence_9	0.0003	0.0012	0.0005	0.0055	0.0139	0.0035	0.0018	0.0052
25	Cluster_prominence_11	0.0003	0.0013	0.0006	0.0050	0.0121	0.0029	0.0017	0.0044
26	Cluster_shade_1	0.0008*	0.0030*	0.0009*	0.0055	0.0249*	0.0071*	0.0035*	0.0119*
27	Cluster_shade_3	0.0008*	0.0033*	0.0010*	0.0051	0.0262*	0.0068*	0.0035*	0.0126*
28	Cluster_shade_5	0.0008*	0.0034*	0.0009*	0.0051	0.0277*	0.0075*	0.0036*	0.0138*
29	Cluster_shade_9	0.0009*	0.0038*	0.0009*	0.0066	0.0303*	0.0088*	0.0040*	0.0148*
30	Cluster_shade_11	0.0009*	0.0040*	0.0009*	0.0059	0.0266*	0.0092*	0.0039*	0.0119*
31	Dissimilarity_1	0.0001	0.0003	0.0003	0.0011	0.0008	0.0007	0.0003	0.0018
32	Dissimilarity_3	0.0001	0.0003	0.0002	0.0010	0.0008	0.0005	0.0003	0.0013
33	Dissimilarity_5	0.0001	0.0003	0.0002	0.0011	0.0015	0.0005	0.0003	0.0011
34	Dissimilarity_9	0.0001	0.0002	0.0003	0.0012	0.0011	0.0007	0.0003	0.0011
35	Dissimilarity_11	0.0001	0.0002	0.0003	0.0011	0.0012	0.0007	0.0004	0.0007
36	Energy_1	0.0001	0.0003	0.0002	0.0013	0.0012	0.0008	0.0004	0.0013
37	Energy_3	0.0001	0.0004	0.0003	0.0013	0.0011	0.0008	0.0004	0.0016
38	Energy_5	0.0001	0.0004	0.0003	0.0011	0.0013	0.0007	0.0003	0.0015
39	Energy_9	0.0001	0.0006	0.0003	0.0009	0.0019	0.0007	0.0003	0.0012
40	Energy_11	0.0001	0.0004	0.0002	0.0013	0.0016	0.0007	0.0004	0.0016
41	Entropy_1	0.0002	0.0003	0.0004	0.0012	0.0024	0.0010	0.0006	0.0012
42	Entropy_3	0.0002	0.0003	0.0004	0.0015	0.0017	0.0007	0.0004	0.0013
43	Entropy_5	0.0001	0.0003	0.0003	0.0015	0.0022	0.0008	0.0004	0.0011
44	Entropy_9	0.0002	0.0004	0.0004	0.0013	0.0018	0.0007	0.0004	0.0012
45	Entropy_11	0.0002	0.0004	0.0003	0.0015	0.0025	0.0008	0.0004	0.0017
46	Homogeneity_m_1	0.0002	0.0003	0.0004	0.0009	0.0021	0.0009	0.0004	0.0016
47	Homogeneity_m_3	0.0001	0.0003	0.0004	0.0013	0.0017	0.0006	0.0003	0.0011
48	Homogeneity_m_5	0.0001	0.0003	0.0003	0.0019	0.0019	0.0006	0.0004	0.0010
49	Homogeneity_m_9	0.0001	0.0003	0.0004	0.0017	0.0013	0.0006	0.0004	0.0015
50	Homogeneity_m_11	0.0001	0.0003	0.0003	0.0023	0.0013	0.0007	0.0004	0.0012
51	Homogeneity_1	0.0002	0.0003	0.0003	0.0008	0.0021	0.0009	0.0004	0.0019
52	Homogeneity_3	0.0001	0.0003	0.0004	0.0016	0.0018	0.0007	0.0003	0.0012
53	Homogeneity_5	0.0001	0.0003	0.0003	0.0018	0.0020	0.0007	0.0004	0.0010
54	Homogeneity_9	0.0002	0.0003	0.0004	0.0014	0.0018	0.0006	0.0004	0.0012
55	Homogeneity_11	0.0001	0.0003	0.0003	0.0025	0.0018	0.0006	0.0004	0.0011
56	Maximum_probability_1	0.0001	0.0003	0.0003	0.0013	0.0014	0.0010	0.0004	0.0013
57	Maximum_probability_3	0.0001	0.0004	0.0003	0.0013	0.0013	0.0008	0.0003	0.0015
58	Maximum_probability_5	0.0001	0.0004	0.0002	0.0011	0.0013	0.0008	0.0004	0.0014
59	Maximum_probability_9	0.0001	0.0005	0.0003	0.0011	0.0017	0.0007	0.0003	0.0012
60	Maximum_probability_11	0.0001	0.0004	0.0002	0.0012	0.0018	0.0007	0.0004	0.0014
61	Sum_of_sqaures_variance_1	0.0001	0.0003	0.0003	0.0009	0.0013	0.0005	0.0003	0.0009
62	Sum_of_sqaures_variance_3	0.0001	0.0003	0.0003	0.0009	0.0016	0.0005	0.0003	0.0010
63	Sum_of_sqaures_variance_5	0.0001	0.0003	0.0003	0.0009	0.0014	0.0005	0.0003	0.0009
64	Sum_of_sqaures_variance_9	0.0001	0.0003	0.0003	0.0009	0.0017	0.0004	0.0003	0.0009

(continued)

Table 3. (continued)

No.	Feature name	Importance scores from each model							
		All	Normal	Cancer	Missed	Prior-Vis	Prior-Invis	Prior1	Prior2
65	Sum_of_sqaures_variance_11	0.0001	0.0003	0.0003	0.0010	0.0015	0.0004	0.0003	0.0009
66	Sum_average_1	0.0001	0.0004	0.0003	0.0011	0.0012	0.0006	0.0003	0.0015
67	Sum_average_3	0.0001	0.0004	0.0004	0.0010	0.0013	0.0007	0.0004	0.0014
68	Sum_average_5	0.0001	0.0004	0.0004	0.0010	0.0014	0.0006	0.0004	0.0013
69	Sum_average_9	0.0001	0.0004	0.0004	0.0009	0.0013	0.0006	0.0004	0.0012
70	Sum_average_11	0.0001	0.0004	0.0003	0.0010	0.0017	0.0006	0.0003	0.0014
71	Sum_variance_1	0.0001	0.0003	0.0003	0.0009	0.0017	0.0006	0.0003	0.0010
72	Sum_variance_3	0.0001	0.0003	0.0003	0.0010	0.0015	0.0005	0.0003	0.0011
73	Sum_variance_5	0.0001	0.0003	0.0003	0.0011	0.0015	0.0006	0.0003	0.0011
74	Sum_variance_9	0.0001	0.0003	0.0003	0.0010	0.0017	0.0005	0.0004	0.0011
75	Sum_variance_11	0.0001	0.0003	0.0003	0.0011	0.0018	0.0006	0.0003	0.0012
76	Sum_entropy_1	0.0001	0.0006	0.0003	0.0015	0.0020	0.0009	0.0004	0.0017
77	Sum_entropy_3	0.0001	0.0006	0.0004	0.0015	0.0014	0.0007	0.0004	0.0019
78	Sum_entropy_5	0.0001	0.0006	0.0003	0.0017	0.0013	0.0010	0.0004	0.0014
79	Sum_entropy_9	0.0001	0.0006	0.0003	0.0018	0.0020	0.0007	0.0004	0.0015
80	Sum_entropy_11	0.0001	0.0006	0.0003	0.0017	0.0019	0.0008	0.0005	0.0020
81	Difference_variance_1	0.0001	0.0003	0.0003	0.0015	0.0008	0.0008	0.0003	0.0017
82	Difference_variance_3	0.0001	0.0002	0.0003	0.0015	0.0011	0.0005	0.0003	0.0011
83	Difference_variance_5	0.0001	0.0002	0.0002	0.0009	0.0018	0.0006	0.0004	0.0007
84	Difference_variance_9	0.0001	0.0002	0.0002	0.0010	0.0019	0.0006	0.0003	0.0009
85	Difference_variance_11	0.0001	0.0002	0.0003	0.0009	0.0021	0.0006	0.0004	0.0008
86	Difference_entropy _1	0.0001	0.0003	0.0003	0.0012	0.0013	0.0008	0.0004	0.0017
87	Difference_entropy _3	0.0001	0.0003	0.0003	0.0010	0.0010	0.0007	0.0003	0.0017
88	Difference_entropy _5	0.0001	0.0003	0.0002	0.0012	0.0012	0.0006	0.0004	0.0010
89	Difference_entropy _9	0.0001	0.0002	0.0003	0.0015	0.0011	0.0008	0.0004	0.0013
90	Difference_entropy _11	0.0001	0.0002	0.0003	0.0017	0.0011	0.0008	0.0005	0.0008
91	Information_measure_of_correlation1_1	0.0001	0.0003	0.0003	0.0011	0.0016	0.0010	0.0004	0.0028
92	Information_measure_of_correlation1_3	0.0001	0.0003	0.0003	0.0008	0.0018	0.0008	0.0003	0.0014
93	Information_measure_of_correlation1_5	0.0001	0.0003	0.0003	0.0011	0.0014	0.0007	0.0004	0.0011
94	Information_measure_of_correlation1_9	0.0001	0.0002	0.0003	0.0011	0.0009	0.0009	0.0004	0.0016
95	Information_measure_of_correlation1_11	0.0001	0.0003	0.0004	0.0017	0.0009	0.0007	0.0004	0.0008
96	Information_measure_of_correlation2_1	0.0001	0.0004	0.0003	0.0013	0.0017	0.0009	0.0004	0.0019
97	Information_measure_of_correlation2_3	0.0001	0.0004	0.0003	0.0011	0.0020	0.0008	0.0004	0.0015
98	Information_measure_of_correlation2_5	0.0001	0.0003	0.0003	0.0013	0.0017	0.0006	0.0004	0.0016
99	Information_measure_of_correlation2_9	0.0001	0.0002	0.0003	0.0012	0.0011	0.0009	0.0004	0.0017
100	Information_measure_of_correlation2_11	0.0001	0.0002	0.0003	0.0017	0.0014	0.0005	0.0004	0.0010
101	Inverse_difference_normalized _1	0.0001	0.0003	0.0003	0.0011	0.0009	0.0007	0.0004	0.0019
102	Inverse_difference_normalized _3	0.0001	0.0003	0.0002	0.0011	0.0011	0.0006	0.0003	0.0014
103	Inverse_difference_normalized _5	0.0001	0.0003	0.0003	0.0011	0.0016	0.0004	0.0003	0.0011
104	Inverse_difference_normalized _9	0.0001	0.0002	0.0003	0.0012	0.0015	0.0007	0.0004	0.0012
105	Inverse_difference_normalized _11	0.0001	0.0003	0.0003	0.0013	0.0011	0.0006	0.0004	0.0007
106	Inverse_difference_moment_normalized _1	0.0001	0.0004	0.0003	0.0014	0.0008	0.0007	0.0003	0.0020
107	Inverse_difference_moment_normalized _3	0.0001	0.0002	0.0003	0.0013	0.0013	0.0005	0.0003	0.0011
108	Inverse_difference_moment_normalized _5	0.0001	0.0002	0.0002	0.0009	0.0015	0.0005	0.0003	0.0007
109	Inverse_difference_moment_normalized _9	0.0001	0.0002	0.0003	0.0011	0.0019	0.0006	0.0003	0.0009

(continued)

Table 3. (continued)

No.	Feature name	Importance scores from each model							
		All	Normal	Cancer	Missed	Prior-Vis	Prior-Invis	Prior1	Prior2
110	Inverse_difference_moment_normalized _11	0.0001	0.0002	0.0003	0.0010	0.0019	0.0006	0.0004	0.0008
111	Mean	0.0002	0.0008	0.0004	0.0013	0.0019	0.0010	0.0005	0.0018
112	Standard_deviation	0.0005	0.0029*	0.0006	0.0085*	0.0262*	0.0050*	0.0025	0.0069
113	Skewness	0.0009*	0.0014	0.0006	0.0056	0.0114	0.0047*	0.0024	0.0065
114	Kurtosis	0.0002	0.0004	0.0008*	0.0019	0.0017	0.0014	0.0006	0.0012
115	Minimum	0.0001	0.0001	0.0003	0.0005	0.0004	0.0002	0.0001	0.0004
116	5th_percentile	0.0002	0.0007	0.0006	0.0016	0.0025	0.0009	0.0005	0.0009
117	10th_percentile	0.0002	0.0008	0.0004	0.0016	0.0015	0.0015	0.0006	0.0014
118	15th_percentile	0.0002	0.0007	0.0004	0.0021	0.0020	0.0021	0.0006	0.0017
119	20th_percentile	0.0002	0.0009	0.0004	0.0017	0.0026	0.0023	0.0007	0.0018
120	25th_percentile	0.0003	0.0011	0.0004	0.0017	0.0032	0.0020	0.0007	0.0024
121	Median	0.0004	0.0018	0.0003	0.0022	0.0027	0.0019	0.0010	0.0026
122	75th_percentile	0.0002	0.0008	0.0003	0.0018	0.0013	0.0015	0.0006	0.0011
123	80th_percentile	0.0002	0.0005	0.0003	0.0015	0.0016	0.0011	0.0004	0.0010
124	85th_percentile	0.0001	0.0004	0.0004	0.0016	0.0018	0.0010	0.0005	0.0013
125	90^{th}_percentile	0.0001	0.0006	0.0004	0.0024	0.0036	0.0013	0.0006	0.0015
126	95th_percentile	0.0003	0.0017	0.0005	0.0048	0.0094	0.0026	0.0012	0.0032
127	Maximum	0.0002	0.0003	0.0003	0.0013	0.0031	0.0007	0.0006	0.0010
128	Range_all (maximum less minimum)	0.0002	0.0004	0.0003	0.0017	0.0032	0.0011	0.0006	0.0021
129	Range5 (95^{th} percentile less 5th percentile)	0.0008*	0.0017	0.0011*	0.0093*	0.0318*	0.0053*	0.0035*	0.0107*
130	Range2 (99th percentile less 1st percentile)	0.0006	0.0033*	0.0008*	0.0087*	0.0165	0.0060*	0.0025	0.0082

Feature nos. 1-110 are GLCM-based features[49-51] with feature parameters as 1, 3, 5, 9, and 11 being pixel distance between the pixel of interest and its neighbour. Feature nos. 111-130 are FOS-based features.[50,51]
* and highlighted in grey = important features based on their importance scores from the machine learning model. The larger the importance scores, the more important the features were for the model.
Abbreviations: FOS = first order statistics; GLCM = gray level co-occurrence matrix.

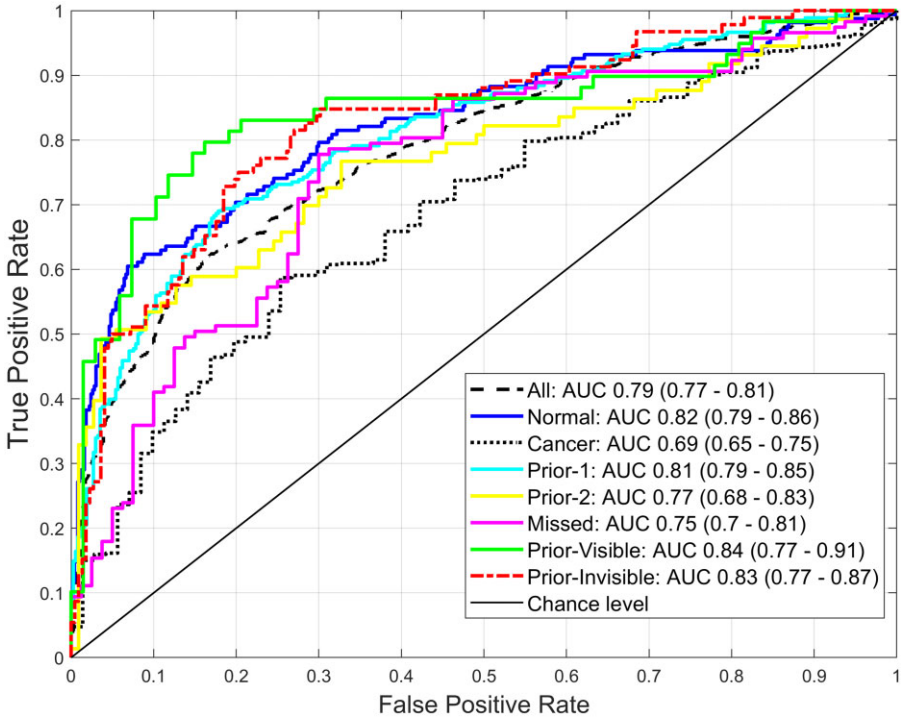


Figure 3. Area under the receiver operating characteristic curve (AUC) for eight classifiers. When differentiating *high-* from *low-gist* images, all eight models using global radiomic features from each image category performed relatively well with the *Prior-Visible* images category model achieved the highest AUC of 0.84 (95% CI, 0.77-0.91) while the *Cancer* model had an AUC of 0.69 (95% CI, 0.65-0.75).

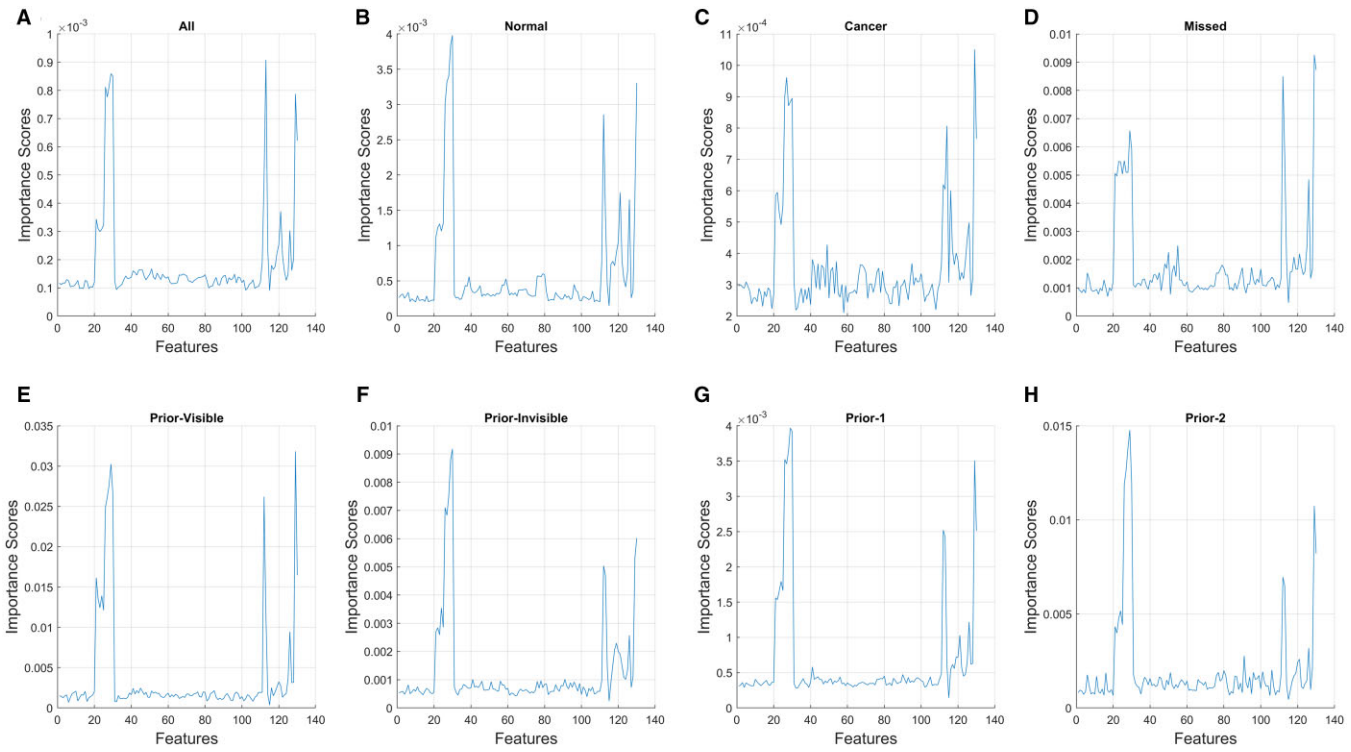


Figure 4. Line plot showing feature importance scores of the 130 features (no. 1-130) in each classifier (A)-(H). As shown, pattern of important features appears to be similar across the eight classifiers.

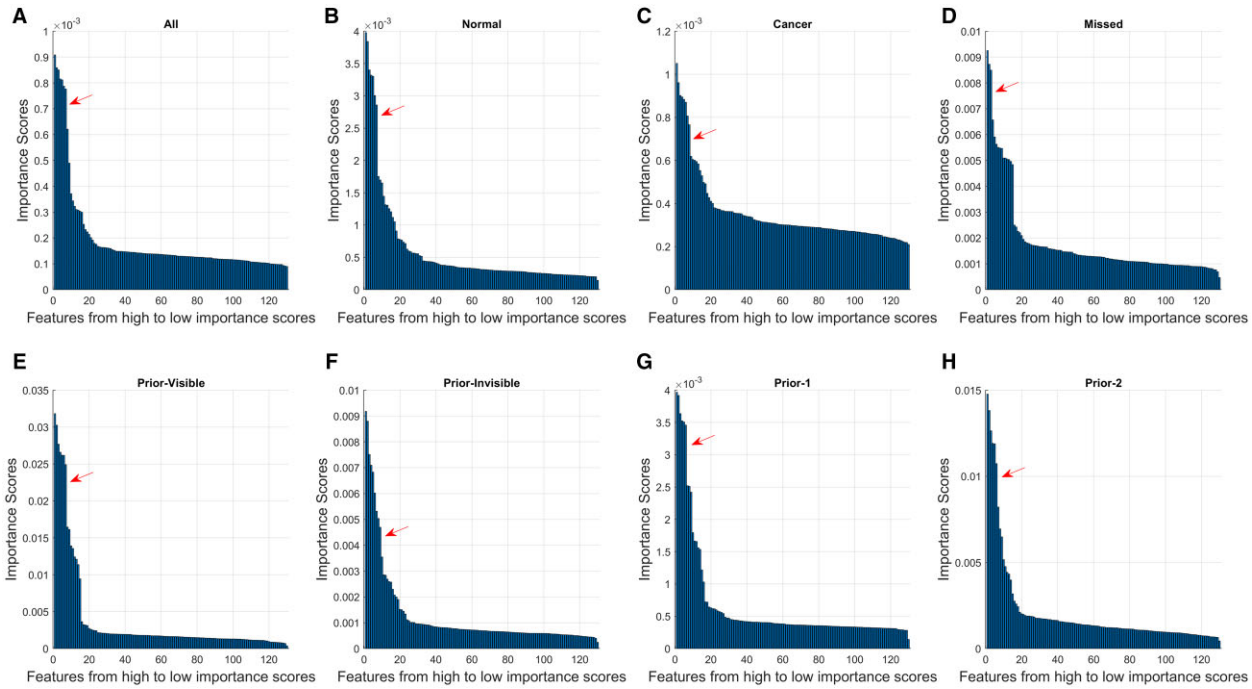


Figure 5. Scree bar chart for each of the eight classifiers showing important features (ranked from high to low importance scores) determined based on the identified steep slope (as shown in red arrow). Seven, seven, eight, three, seven, nine, six, and six most important features were identified for *All* (A), *Normal* (B), *Cancer* (C), *Missed* (D), *Prior-Visible* (E), *Prior-Invisible* (F), *Prior-1* (G), *Prior-2* (H) model, respectively.

the proper use of the GGS to support BC detection and prediction process.

Moreover, if the GGS can be considered as a BC risk factor, similar to breast density or genetic risk factor, our work may be applied as a supplemented artificial vision intelligence tool to signal clinician about an elevated risk of a positive cancer finding in mammographic screening.[3,5,13] It might also be used to augment the recent deep learning model[57] to further advance the accuracy of BC detection in screening mammography especially since the GGS has been shown to

be helpful in improving the performance of the deep learning model in detecting BC.[14] This could not only assist in lessening the possibility of FP and FN errors but also optimizing personalized screening strategy (eg, shorter screening interval) and increasing the chance of BC being detected early with improved therapy outcomes and survivorship. Conversely, more research efforts are needed to further evaluate and validate its clinical usefulness.

Interestingly, although studies[3,58] indicated that GGS of future cancerous abnormalities can be detected in fine details of the parenchymal texture of high-pass filtered mammograms, other works revealed that radiologists extract the GGS using low-resolution peripheral vision prior to detailed foveal processing of any image areas when separating abnormal from normal mammograms.[1,10,24] Our findings supported the relevance of the FOS and GLCM-based second-order textural features to the GGS, extracted by radiologists' peripheral vision. Prior studies[3,13] showed that observers failed to localize BC using such gist signal, indicating that the gist signal in the context of medical images is concordant with findings from basic vision sciences where observers struggle with tasks involving distinguishing phase differences[59] and show uncertainty in localizing bisection.[60] To move beyond qualitative evidence, a specific proposal for peripheral "texture processing" and its relevance to mammography are required. The current best theory for a model of capturing the appearance of textures is the Texture Tiling Model[61,62] which implicates the visual system with the summary statistics computed using the second-order statistics of primary visual cortex over local pooling regions.

In the context of basic vision science, the importance of texture features from non-medical images in identifying the global gist of natural scenes was discovered.[29-31] Our work based on breast medical images also showed similar results, in that, five useful features involving one GLCM-based textural and four FOS-based features (Table 3) from the ML models were found. In concordance with the suggestions from basic vision science studies,[33,34,63] features that are related to noticeable differences in the first order luminance, that is, *range* and *standard_deviation*, were among important features in the classification tasks. In particular, *range5* (feature no. 129) seems to be valuable in all the model except *Normal*, signifying that a high GGS in most image types may be motivated by the overall low range (difference between the 95th and 5th percentiles) of the image gray level value, resembling cancer characteristics. Similarly, among the second-order features, *cluster shade* (feature no. 26-30) exhibited importance in all tasks except for one, suggesting that the "uniformity of mammographic texture" could contribute to the radiologists' initial impression about the abnormality of a case. In this case, these features may be used as an indicator of whether a strong gist of the abnormal is present in an image, which could contribute to the early detection and risk prediction of BC. However, larger research is required to further examine the effectiveness of these features.

There were a few limitations in our study. First, this was the first exploratory study investigating the capability of handcrafted GRFs (ie, 110 GLCM-based texture and 20 FOS-based features) derived from 130 CC unilateral view mammograms[64] in predicting *high-gist* from *low-gist* images based on the gist scores of thirteen "gist experts".[15] The mediolateral oblique (MLO) view was not used in this study. While the entire images were viewed by the readers in the gist experiment, the background skin-air region of the images had to be excluded using erosion algorithm, because it has been enhanced by the vendors, to obtain the true breast region for GRF extraction. The results may be only applicable to the images and gist experts with characteristic/background (eg, being a screen reader with high screening workload) similar to the sample used in the study. Secondly, useful GRFs were determined based on their importance scores using a scree test[55] (Fig. 4), which may be less explicit for some models, for instance, the *Missed* model when compared to the *Normal* model. Also, only "Prior-1" images (attained two years prior to current cancer mammograms) were used to categorize "Missed", "Prior-Visible", and "Prior-Invisible" images. As with all model building, the output of the model is constrained by the parameters entered into the model. So, forthcoming studies should examine other prior mammograms acquired beyond 2 years before, MLO, raw[47] and high-pass filtered bilateral mammograms,[58] digital breast tomosynthesis images,[65] and images from other vendors (eg, GE, GE Healthcare, Chicago, IL, USA). These images may also contain valuable information to suggest an abnormality and/or an increased future malignancy risk, and therefore may result in different gist scores and GRFs. Also, gist experts from other location and using other morphological image processing methods (such as corrosion and closing[66]) may produce different results. Lastly, other potential important GRFs (eg, *cluster prominence*, feature no. 21-25) and types of GRFs (eg, deep learning features[11]) should also be analysed.

Conclusions

To summarize, this study suggests that a set of quantitative GRFs derived from mammographic images can accurately predict *high-* from *low-gist* images with five useful GRFs (one GLCM-based texture and four FOS-based features) discovered and emphasized. These findings are critically important in providing better understanding of the mammographic image characteristics that drive the strength of the GGS, which could be exploited to advance BC screening and risk prediction, enabling early detection and treatment of BC thereby further reducing BC-related deaths.

Acknowledgements

We would like to express gratitude to the National Health and Medical Research Council (NHMRC), National Breast Cancer Foundation (NBCF) and Cancer Institute New South Wales (CINSW) for funding support of this work. We also would like to thank the radiologists and breast physicians for participating in this experiment, the gist research team and Dr Robert Heard for helpful discussion regarding the findings and statistical methods as well as the reviewers for their rigorous evaluation of this article.

Author contributions

Conceptualisation, validation, and writing – review & editing, all authors. Data curation, formal analysis, investigation, methodology, and software, S.S and Z.G. Supervision and funding acquisition, Z.G., S.L., and P.B. Visualisation and writing – original draft, S.S.

Funding

This research was supported by the National Health and Medical Research Council (NHMRC), grant number APP1162872, and National Breast Cancer Foundation (NBCF), grant number IIRS-18-089. Z.G. is supported by the Cancer Institute New South Wales Early Career Fellowship (2022/ECF1426). The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this article.

Conflicts of interest

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this article. P.B. and Z.G., as well as an academic at the University of Sydney, serve as CEO and Co-founder (P.B.)/employee (Z.G.) of DetectedX, a company that aims to educate radiologists. The article will not impact on DetectedX's activities. All other authors of this article, other than being employee (S.L.)/student (S.S.) of the University of Sydney/employee (S.S.) of the University of Technology Sydney, declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Data availability

The dataset that supports the findings of this study is not publicly available as it contains proprietary information that the authors acquired through a license. Data are, however, available from the authors upon reasonable request.

Ethical approval

All procedures performed and reported in this article involving human participants were in accordance with the ethical standards of the University of Sydney's Human Research Ethics Committee [protocol no. 2020/324].

Consent

All individual participants in this study provided their written informed consent for participation and publication.

References

1. Gandomkar Z, Mello-Thoms C. Visual search in breast imaging. *Br J Radiol*. 2019;92(1102):20190057.
2. Kundel HL, Nodine CF. Interpreting chest radiographs without visual search. *Radiology*. 1975;116(3):527-532.
3. Evans KK, Haygood TM, Cooper J, Culpan A-M, Wolfe JM. A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast. *Proc Natl Acad Sci USA*. 2016;113(37):10292-10297.
4. Brennan PC, Gandomkar Z, Ekpo EU, et al. Radiologists can detect the 'gist' of breast cancer before any overt signs of cancer appear. *Sci Rep*. 2018;8(1):8717-8712.
5. Evans KK, Culpan A-M, Wolfe JM. Detecting the "gist" of breast cancer in mammograms three years before localized signs of cancer are visible. *Br J Radiol*. 2019;92(1099):20190136-20190110.
6. Gandomkar Z, Ekpo E, Lewis S, Evans K, Tapia K, Trieu P-D, et al. eds. Detection of the abnormal gist in the prior mammograms even with no overt sign of breast cancer. *SPIE 10718, 14th International Workshop on Breast Imaging (IWBI 2018)*; 2018; Atlanta, Georgia, United States: SPIE.
7. Henderson LM, Hubbard RA, Sprague BL, Zhu W, Kerlikowske K. Increased risk of developing breast cancer after a false-positive screening Mammogram. *Cancer Epidemiol Biomarkers Prev*. 2015;24(12):1882-1889.
8. Ramin C, Withrow DR, Davis Lynn BC, Gierach GL, Berrington De González A. Risk of contralateral breast cancer according to first breast cancer characteristics among women in the USA, 1992–2016. *Breast Cancer Res*. 2021;23(1):24.
9. Gastounioti A, Conant EF, Kontos D. Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. *Breast Cancer Res*. 2016;18(1):91-12.
10. Gandomkar Z, Ekpo E, Chen Z, Suleiman M, Li T, Tapia K, et al. eds. Expert radiologist performance does not appear to impact upon their capability in perceiving the gist of the abnormal on mammograms. *15th International Workshop on Breast Imaging (IWBI2020)*. SPIE; 2020:.
11. Gandomkar Z, Ekpo EU, Lewis SJ, Suleiman ME, Siviengphanom S, Li T, et al. eds. An end-to-end deep learning model can detect the gist of the abnormal in prior mammograms as perceived by experienced radiologists. *SPIE Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*. SPIE; 2021.
12. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology*. 2007;242(2):396-402.
13. Evans KK, Georgian-Smith D, Tambouret R, Birdwell RL, Wolfe JM. The gist of the abnormal: above-chance medical decision making in the blink of an eye. *Psychon Bull Rev*. 2013;20(6):1170-1175.
14. Gandomkar Z, Siviengphanom S, Ekpo EU, et al. Global processing provides malignancy evidence complementary to the information captured by humans or machines following detailed mammogram inspection. *Sci Rep*. 2021;11(1):20122-20112.
15. Gandomkar Z, Siviengphanom S, Suleiman Ma, Wong D, Reed W, Ekpo E, et al. eds. The reliability of radiologists' first impression interpreting a screening mammogram. *SPIE Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment*; 4 April; San Diego, California, United States: SPIE; 2022.
16. Raat EM, Farr I, Wolfe JM, Evans KK. Comparable prediction of breast cancer risk from a glimpse or a first impression of a mammogram. *Cogn Res Princ Implic*. 2021;6(1):72.
17. Wolfe JM, Evans KK, Drew T. The first moments of medical image perception. In: E Samei, EA Krupinski, eds. *The Handbook of Medical Image Perception and Techniques*. 2nd ed. Cambridge University Press; 2018:188-196.
18. Waite S, Grigorian A, Alexander RG, et al. Analysis of perceptual expertise in radiology—current knowledge and a new perspective. *Front Hum Neurosci*. 2019;13:213.
19. Sheridan H, Reingold EM. The holistic processing account of visual expertise in medical image perception: a review. *Front Psychol*. 2017;8:1620.
20. Wolfe JM, Evans KK, Drew T, Aizenman A, Josephs E. How do radiologists use the human search engine? *Radiat Prot Dosimetry*. 2016;169(1-4):24-31.
21. Donovan T, Litchfield D, Crawford TJ. Editorial: medical image perception: how much do we understand it? *Front Psychol*. 2017;8:2072.
22. Carrigan AJ, Wardle SG, Rich AN. Finding cancer in mammograms: if you know it's there, do you know where? *Cogn Res Princ Implic*. 2018;3(1):10-14.
23. Ungerleider L, Mishkin M. Two cortical visual systems. In: Ingle DJ, Goodale MA, Mansfield RJ, eds. *Analysis of Visual Behavior*. The MIT press; 1982:549-586.
24. Gandomkar Z, Ekpo E, Lewis S, Evans K, Tapia K, Trieu P, et al. eds. Does the strength of the gist signal predict the difficulty of

breast cancer detection in usual presentation and reporting mechanisms? *SPIE 10952 Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment; 2019; San Diego, California, United States*; SPIE; 2019.
25. Wolfe JM, Võ ML, Evans KK, Greene MR. Visual search in scenes involves selective and nonselective pathways. *Trends Cogn Sci*. 2011;15(2):77-84.
26. Drew T, Evans K, Võ ML, Jacobson FL, Wolfe JM. Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images? *Radiographics*. 2013;33 (1):263-274.
27. Joubert OR, Rousselet GA, Fize D, Fabre-Thorpe M. Processing scene context: fast categorization and object interference. *Vision Res*. 2007;47(26):3286-3297.
28. Rousselet G, Joubert O, Fabre-Thorpe M. How long to get to the "gist" of real-world natural scenes? *Visual Cognition*. 2005;12 (6):852-877.
29. Renninger LW, Malik J. When is scene identification just texture recognition? *Vision Res*. 2004;44(19):2301-2311.
30. Oliva A, Torralba A. Modeling the Shape of the scene: a holistic representation of the spatial envelope. *Int J Comput Vision*. 2001; 42(3):145-175.
31. Oliva A, Torralba A. Building the gist of a scene: the role of global image features in recognition. In: Martinez-Conde S, Macknik SL, Martinez LM, Alonso JM, Tse PU, editors. *Progress in Brain Research*. Vol. 155: Elsevier; 2006:23-36.
32. Nisbett WH, Kavuri A, Das M. On the correlation between second order texture features and human observer detection performance in digital images. *Sci Rep*. 2020;10(1):13510.
33. Chubb C, Nam J-H, Bindman DR, Sperling G. The three dimensions of human visual sensitivity to first-order contrast statistics. *Vision Res*. 2007;47(17):2237-2248.
34. Julesz B. Texture and visual perception. *Sci Am*. 1965;212 (2):38-48.
35. Gandomkar Z, Tay K, Ryder W, Brennan PC, Mello-Thoms C. iCAP: an individualized model combining gaze parameters and image-based features to predict radiologists' decisions while reading mammograms. *IEEE Trans Med Imaging*. 2016;36(5):1066-1075.
36. Joseph S, Balakrishnan K, eds. Local binary patterns, haar wavelet features and haralick texture features for mammogram image classification using artificial neural networks. *Advances in Computing and Information Technology: First International Conference, ACITY 2011, Chennai, India, July 15-17, 2011 Proceedings*. Springer; 2011.
37. Gandomkar Z, Li T, Shao Z, Tang L, Xiao Q, Gu Y, et al. eds. Breast cancer risk prediction in Chinese women based on mammographic texture and a comprehensive set of epidemiologic factors. *Fifteenth International Workshop on Breast Imaging;Leuven, Belgium*. SPIE; 2020.
38. Krupinski EA. Current perspectives in medical image perception. *Atten Percept Psychophys*. 2010;72(5):1205-1217.
39. Ong M-S, Mandl KD. National expenditure for false-positive mammograms and breast cancer overdiagnoses estimated at $4 billion a year. *Health Aff (Millwood)*. 2015;34(4):576-583.
40. Li Y, Zhou Y, Mao F, et al. The influence on survival of delay in the treatment initiation of screening detected non-symptomatic breast cancer. *Sci Rep*. 2019;9(1):10158-10157.
41. Gandomkar Z, Lewis S, Siviengphanom S, et al. *Classification and Reviewing of Prior Screening Mammograms from Screen-Detected Breast Cancer Cases: SPIE*; 2022.
42. Gandomkar Z, Siviengphanom S, Suleiman MA, et al. Reliability of radiologists' first impression when interpreting a screening mammogram. *PLoS One*. 2023;18(4):e0284605.
43. Siviengphanom S, Gandomkar Z, Lewis SJ, Brennan PC. Mammography-based radiomics in breast cancer: a scoping review of current knowledge and future needs. *Acad Radiol*. 2021;29 (8):1228-1247.
44. TM Inc. imerode—Erode image: The MathWorks Inc.; 2023. Accessed October 17, 2023. https://au.mathworks.com/help/images/ref/imerode.html
45. TM Inc. strel—Morphological structuring element: The MathWorks Inc.; 2023. Accessed October 17, 2023. https://au.mathworks.com/help/images/ref/strel.html
46. Said KAM, Jambek AB, eds. *Analysis of Image Processing Using Morphological Erosion and Dilation. International Conference on Biomedical Engineering (ICoBE 2021);UniMAP, Malaysia (Virtual)*. IOP Publishing Ltd; 2021.
47. Keller BM, Nathan DL, Wang Y, et al. Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation. *Med Phys*. 2012;39(8):4903-4917.
48. Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. *IEEE Trans Syst, Man, Cybern*. 1973;SMC-3 (6):610-621.
49. Scapicchio C, Gabelloni M, Barucci A, Cioni D, Saba L, Neri E. A deep look into radiomics. *Radiol Med*. 2021;126(10):1296-1311.
50. Costaridou L. *Medical Image Analysis Methods*. Taylor & Francis; 2005.
51. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
52. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123-140.
53. Couronné R, Probst P, Boulesteix A-L. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinform*. 2018;19(1):270-214.
54. TM Inc. fitcensemble—Fit ensemble of learners for classification The MathWorks Inc.; 2023. Accessed July 28, 2023. https://au.mathworks.com/help/stats/fitcensemble.html
55. Ledesma RD, Valero-Mora P, Macbeth G. The Scree test and the number of factors: a dynamic graphics approach. *Span J Psychol*. 2015;18:E11.
56. Bengio Y, Grandvalet Y. No unbiased estimator of the variance of K-fold cross-validation. *The J Mach Learn Res*. 2004;5 (2004):1089-1105.
57. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. *Sci Rep*. 2019;9(1):12495-12412.
58. Raat EM, Evans KK. Early signs of cancer present in the fine detail of mammograms. *PLoS One*. 2023;18(4):e0282872.
59. Bennett PJ, Banks MS. The effects of contrast, spatial scale, and orientation on foveal and peripheral phase discrimination. *Vision Res*. 1991;31(10):1759-1786.
60. Levi DM, Klein SA. Sampling in spatial vision. *Nature*. 1986;320 (6060):360-362.
61. Freeman J, Simoncelli EP. Metamers of the ventral stream. *Nat Neurosci*. 2011;14(9):1195-1201.
62. Rosenholtz R, Huang J, Raj A, Balas BJ, Ilie L. A summary statistic representation in peripheral vision explains visual search. *J Vis*. 2012;12(4):14.
63. Julesz B. Visual pattern discrimination. *IEEE Trans Inform Theory*. 1962;8(2):84-92.
64. Gandomkar Z, Ekpo EU, Lewis SJ, Evans KK, Tapi K, Li T, et al. eds. The strength of the gist of the abnormal in the unilateral and bilateral mammograms. *Medical Image Perception Society (MIPS) XVIII Conference; Salt Lake City, Utah, USA*. Medical Image Perception Society; 2019.
65. Wu CC, D'Ardenne NM, Nishikawa RM, Wolfe JM. Gist processing in digital breast tomosynthesis. *J Med Imaging (Bellingham)*. 2020;7(2):022403.
66. Zhang F, Wu Y, Xu M, Liu S, Peng C, Gao Z. A morphological image segmentation algorithm for circular overlapping cells. *Intell Autom Soft Comput*. 2022;32(1):301-321.

3.9 Chapter findings

It was observed from this Chapter 3 that global radiomic features derived from mammographic images could accurately predict global gist signal (i.e., classifying *high-* from *low-gist* images) with AUCs of up to 0.84. Five most useful features for the classification tasks were also identified: *cluster shade*, *standard deviation*, *skewness*, *kurtosis*, and *range*.

Chapter 4 – Global radiomic features from mammography for predicting difficult-to-interpret normal cases

4.1 Chapter introduction

The previous chapter (Chapter 3) demonstrates correlation between global radiomic features and global gist signals which are important for guiding visual searches and making diagnostic decisions when interpreting difficult breast cancer cases linking to diagnostic errors and readers' expertise development. Nevertheless, studies found in the literature searches (Chapter 2) revealed scarce research exploring if these global radiomic features, directing radiologists' initial impression about the case, are correlated to normal case difficulty (high false positive errors made by average radiologists) in the mammographic test sets. Since around 70% of cases in many educational-style test sets and 99.5% of screening mammograms are normal [1-5], it is essential to examine the relationship between global radiomic features and these cases, and if a significant correlation exists, what are the most valuable features in describing these cases. Moreover, since higher BI-RADS category of breast density has often been claimed to signify more difficult cases when generating educational mammographic test sets, it is important to also investigate if BI-RADS breast density correlates to difficulty scores of these normal cases. This chapter (Chapter 4) tries to answer these questions in order to better understand the nature of human difficulty through the global radiomics features for enhancing the process of curating mammographic education test sets in accordance with the **objective 3** and knowledge gaps recognised in the introduction (Chapter 1) and literature review (Chapter 2).

The second original research study in this chapter has been published as:

Siviengphanom, S., Z. Gandomkar, S.J. Lewis, and P.C. Brennan, *Global radiomic features from mammography for predicting difficult-to-interpret normal cases.* Journal of Digital Imaging, 2023. **36**: p. 1541–1552. https://doi.org/10.1007/s10278-023-00836-7

In this paper, three random forest machine learning classifiers were built (*CC*, *MLO*, and *CC+MLO*). The *CC* and *MLO* were constructed using the global radiomic features from corresponding craniocaudal (CC) and mediolateral oblique (MLO) images for distinguishing *difficult-* from *easy-to-interpret* normal cases. The *CC+MLO* model was built using a late fusion method [6] which involved taking the median predictive scores from both *CC* and *MLO* models. One of the main reasons of employing the late fusion method was that it allows fusing and leveraging the outputs of already constructed individual *CC* and *MLO* models without needing build a separate model as each model captures different aspects of the data, and fusing

their outputs reduces individual model biases, leading to a more accurate and robust prediction [6].

Conference proceeding/presentation based on smaller data related to the above study has been published as/presented at:

- Siviengphanom, S., Z. Gandomkar, S.J. Lewis, and P.C. Brennan. *Global mammographic radiomic signature can predict radiologists' difficult-to-interpret normal cases*. in *SPIE Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment*. 2023. San Diego, California, United States: SPIE. https://doi.org/10.1117/12.2645377 (Appendix 1.5).
- SPIE Medical Imaging Conference 2023, San Diego, CA, United States, February 2023 (Appendix 1.6).

Journal of Digital Imaging (2023) 36:1541–1552
https://doi.org/10.1007/s10278-023-00836-7

Global Radiomic Features from Mammography for Predicting Difficult-To-Interpret Normal Cases

Somphone Siviengphanom[1] · **Ziba Gandomkar**[1] · **Sarah J. Lewis**[1] · **Patrick C. Brennan**[1]

Received: 28 November 2022 / Revised: 5 April 2023 / Accepted: 13 April 2023 / Published online: 30 May 2023

Abstract
This work aimed to investigate whether global radiomic features (GRFs) from mammograms can predict *difficult-to-interpret* normal cases (NCs). Assessments from 537 readers interpreting 239 normal mammograms were used to categorise cases as 120 *difficult-to-interpret* and 119 *easy-to-interpret* based on cases having the highest and lowest difficulty scores, respectively. Using lattice- and squared-based approaches, 34 handcrafted GRFs per image were extracted and normalised. Three classifiers were constructed: (i) *CC* and (ii) *MLO* using the GRFs from corresponding craniocaudal and mediolateral oblique images only, based on the random forest technique for distinguishing *difficult-* from *easy-to-interpret* NCs, and (iii) *CC* + *MLO* using the median predictive scores from both *CC* and *MLO* models. Useful GRFs for the *CC* and *MLO* models were recognised using a scree test. The *CC* and *MLO* models were trained and validated using the leave-one-out-cross-validation. The models' performances were assessed by the AUC and compared using the DeLong test. A Kruskal–Wallis test was used to examine if the 34 GRFs differed between *difficult-* and *easy-to-interpret* NCs and if difficulty level based on the traditional breast density (BD) categories differed among 115 *low-BD* and 124 *high-BD* NCs. The *CC* + *MLO* model achieved higher performance (0.71 AUC) than the individual *CC* and *MLO* model alone (0.66 each), but statistically non-significant difference was found (all $p > 0.05$). Six GRFs were identified to be valuable in describing *difficult-to-interpret* NCs. Twenty features, when compared between *difficult-* and *easy-to-interpret* NCs, differed significantly ($p < 0.05$). No statistically significant difference was observed in difficulty between *low-* and *high-BD* NCs ($p = 0.709$). GRF mammographic analysis can predict *difficult-to-interpret* NCs.

Keywords Radiomics · Mammography · Difficult normal cases · Breast cancer · Machine learning · Gist

Introduction

Among women globally, breast cancer is the most prevalent cancer type and leading cause of cancer-related death [1]. While mammography screening has played a major role in the early detection of breast cancer for decades [2–4], mammographic image interpretation is an extremely complex process with reported false-positive (FP) error rates of up to 16% and estimated cumulative risk of FP results of up to 63% in women 50–69 years of age after ten biennial mammographic screening rounds [5, 6]. A study reported that 80% of women having abnormal mammographic findings at screening were recalled for further diagnostic procedures that eventually resolved with normal outcomes and 40% of those went through invasive biopsy which subsequently followed in benign findings [7]. These unnecessary recalls (i.e., FP findings) have placed huge burdens on patients and healthcare systems, causing considerable psychological distress and anxiety as well as unwarranted medical costs (e.g., USD$2.8 billion each year in the USA) [8–11]. Considering that approximately 99.5% of cases in screening programs are normal, the ability of radiologists to interpret and detect normal cases (NCs) correctly is extremely important [12, 13].

Educational test sets are powerful web-based platforms shown to improve radiologists' performance in interpreting mammograms [14–22]. Since NCs often constitute almost 70% of the cases in these educational test sets, the capability to identify normal mammographic features can be well assessed [19, 20, 23]. However, a consistent difficulty for normal mammographic cases across different test sets is important in order to facilitate a reliable and efficient

✉ Somphone Siviengphanom
ssiv6387@uni.sydney.edu.au

[1] Medical Image Optimisation and Perception Group, Discipline of Medical Imaging Science, Sydney School of Health Sciences, Faculty of Medicine and Health, the University of Sydney, Sydney, NSW 2006, Australia

objective measurement of FP errors and the performance of individual reader, since test set based education is widely used worldwide as a means of monitoring clinical performance [16–20, 22–25]. This implies that if a radiologist's performance changes over time, it should be a function of the radiologist, not the test set conditions. Considering readers' performance in educational test sets can largely signify their clinical performance; measured by clinical audits, a consistent difficulty of NCs in test set based education is also vital for clinical screening programs. It can not only enable uniform empirical measurement of the screening program's recall rate performance over time, but also facilitate the establishment of national standards of recall rate that could be used to compare with the recall rate performance of individual breast reader as part of a quality improvement [15, 20, 22–25].

To date, there appears to be no reliable method of standardising test set difficulty for NCs since test set generation currently relies on experts' assessment of case difficulty, where suitable NCs are selected manually by expert radiologists, often based on the traditional breast imaging and reporting data system (BI-RADS) breast density (BD) category and inclusion of benign features [18, 26–28]. Studies, however, have revealed that the difficulty of NCs subjectively defined by experts or perceived by individuals based on the BD classification rarely aligns to the actual difficulty based on readers' performance in the test set [28–30]. Although a retrospective assessment of normal case difficulty is performed by calculating the numbers of errors made by readers engaging with the test set, this is rare, and indeed when it occurs, it relies on the efficacy of the readers of that test set which is still not a good descriptor of the actual difficulty of NCs [18, 26, 28, 29]. Consequently, there is an urgent need for an alternative standardised method of predicting NCs' difficulty so that readers' performance and individual specific FP errors can be better monitored, identified, addressed, and improved.

This study aims to provide a reliable method for automatically describing the difficulty of NCs (i.e., an objective measure of difficulty based on readers' error rate [31]) using global image features [32] derived from radiomics (extracting and analysing a large number of quantitative features from medical images and converting them into multi-dimensional data) [33–38]. Radiomics in mammography, to date, has focused on predicting readers' error making patterns using local radiomic features of specific tumour/FP areas, achieving up to 0.61 AUC [39, 40]. Previous studies have also shown that radiologists detect the gist of "abnormal" based on overall image appearance immediately after image onset [41–43]. Considering the importance of the initial global impression in guiding a visual search and making a diagnostic decision [41–45] as well as the value of radiomics in mammography [38], we hypothesised that global image statistics (i.e., global radiomic features/GRFs) of *difficult-to-interpret* NCs differed from *easy-to-interpret* ones. However, little attempt has been made to predict *difficult-to-interpret* NCs based on GRFs. This study addresses this deficiency. Moreover, since test set curation often depends on the traditional BI-RADS BD classification and higher BD cases have previously been reported to be more difficult to interpret [28, 30, 46], this work also investigates if there is a difference in difficulty level between *low-BD* (BI-RADS BD category A = fatty and B = fibroglandular) vs *high-BD* (BI-RADS BD category C = heterogeneous and D = extremely dense) NCs [47].

Materials and Methods

Ethical approval was obtained for this study from the Human Research Ethics Committee of the University of Sydney [2019/013 and 2017/028] with informed consent acquired from each reader. This study only involved retrospective analysis of existing and non-identifiable data collected as part of the BreastScreen Reader Assessment Strategy (BREAST) databank (https://breast-australia.sydney.edu.au/) [48, 49], which is funded by the Commonwealth of Australia to improve observer performance in BreastScreen Australia.

Readers

A total of 537 readers were included in this study, with expert breast Australian and New Zealand radiologists ($n = 382$), breast physicians ($n = 18$, medical practitioners specializing in interpreting breast images such as mammograms, and diagnosis and management of benign and malignant breast disease, in partnership with breast surgeons, radiologists, oncologists, pathologists and geneticists), and radiology trainees ($n = 137$) who engaged with the BREAST program from 04 September 2014 to 17 March 2021. These readers were recruited either through a radiology and/or breast cancer workshop/conference or a direct registration via the online BREAST platform. Additionally, radiologists may be referred to complete the BREAST test set by their lead BreastScreen Australia radiologist. The number of readers who read each test set varied from 33 to 206, in accordance with the time release of the test sets between 2014 to 2021. There was a mean value of 130 readers per test set, and each reader had an average of 10 years' experience in reading mammograms. Out of 537, 471 readers provided their age details, giving a mean age of 46. While the number of female (44%) and male (41%) readers (15% did not specify their gender) and the number of readers who spent less than 4 h (50%) and 4–30 h per week (48%) reading mammograms were similar, 46% of readers read 20–200 cases of mammograms per week, 26% completed a fellowship which lasted for 3–6 months, and 42% currently read for breast screening program (Table 1).

Normal Cases and Reading Conditions

Given the aim of the study is to predict the difficulty of NCs using GRFs, only NCs were included.NCs came from a screening population; hence, patients were all females, aged between 50 to 75, and generally asymptomatic. Truth for each normal case was validated by the consensus reading of at least two expert radiologists, followed by a subsequent negative screen outcome confirmed by two expert radiologists [14]. A total of 361 normal mammography cases from all the nine existing BREAST test sets were used, with each case including bilateral craniocaudal (CC) and mediolateral oblique (MLO) views—a total of four DICOM images per case. The NCs of each test set were viewed by the 537 readers via the BREAST platform, either at a workshop/conference in a simulated reading room ($n = 320$), or online at their usual clinical setting ($n = 217$). Reading conditions in both settings were comparable with DICOM images being displayed on a pair of 5-megapixel medical standard monitors and recommended ambient light levels range of 20–40 lux [17, 19, 48, 50, 51]. The mammogram examinations were performed using mammography provided by Hologic (Hologic, Inc., Marlborough, MA, USA), GE (GE Healthcare, Chicago, IL, USA), Siemens (Munich, Germany), Sectra (Sectra, Linköping, Sweden), Philips (Philips Healthcare, Amsterdam, the Netherlands), and Fujifilm (Fujifilm Corporation, Minato City, Tokyo, Japan).

Table 1 Readers' details at the time of completing a test set

Total no. of readers	537
Mean no. of years reading mammograms	10
Mean age (from 471 readers with age details while 66 readers did not provide their age details)	46
Gender	
Female	238 (44%)
Male	219 (41%)
Not specified	80 (15%)
No. of mammographic cases reading per week	
< 20	230 (43%)
20–200	248 (46%)
> 200	59 (11%)
No. of hours per week reading mammograms	
< 4	267 (50%)
4–30	258 (48%)
> 30	12 (2%)
No. of readers completed a fellowship lasting 3 to 6 months	
Completed fellowship	138 (26%)
Did not completed fellowship	386 (72%)
Not specified	2 (13%)
No. of readers currently read for breast screening program	
Yes	223 (42%)
No	314 (58%)

Difficult Normal Cases and Their Breast Density

To identify the difficulty of NCs, difficulty scores (i.e., percentage of incorrect reports per case defined as an objective measure of difficulty [31]) were firstly calculated for each of the 361 NCs using the Royal Australian and New Zealand College of Radiologists (RANZCR) rating (scale of 2 to 5 while 1 represents a normal case [48]) provided by each of the 537 readers. Similar to the American College of Radiology's numbered BI-RADS mammographic assessment categories, a RANZCR score of 2 was given when readers thought the case was benign lesion, while a score of 3, 4, or 5 was given when readers believed a malignant lesion was present with greater number indicating a higher confidence of malignancy. For cases with more than one rating from the same reader (this happened when a reader believed there were multiple abnormalities), only the highest rating was used. Difficulty scores were derived by dividing the number of readers who gave an incorrect report (i.e., rated a normal case as cancer with a rating of 3, 4, or 5) by the number of total readers who read the test set. The difficulty scores were then categorised into three classes, giving 120 *difficult-to-interpret*, 122 *intermediate-to-interpret*, and 119 *easy-to-interpret* NCs based on one-third of the total 361 cases having the highest, intermediate, and lowest difficulty scores, respectively. However, only cases (total of 239) in the *difficult-* ($n = 120$) and *easy-to-interpret* ($n = 119$) categories were included for the analysis. The 122 intermediate NCs were excluded to enable a clear distinction between the highest and lowest difficulty NCs. These two categories also resulted in a total of 115 *low-BD* (58 *difficult-* and 57 *easy-to-interpret*) and 124 *high-BD* NCs (62 *difficult-* and 62 *easy-to-interpret*).

Radiomic Analysis

Extraction of Global Radiomic Features

The first step in the radiomics analysis pipeline was to obtain input mammographic images and binary (black and white) masks. In total, 956 mammography DICOM images (4 images per case × 239 NCs: 120 *difficult-* and 119 *easy-to-interpret*) were acquired. Using a standard gray level thresholding value of 100 (all pixel intensities above the threshold value were converted to white while the rest were converted to black), binary masks were generated from 956 DICOM images to extract the breast region from the surrounding image background, removing undesirable labels, markers, and other artifacts. Where necessary, a manual adjustment on the thresholding value was applied to ensure the binary masks were created accurately. Next, for computational

efficiency purposes, the DICOM images and masks were converted to a TIFF file format, and all the right CC and MLO views of mammograms were flipped to the left so that all mammograms had consistent chest wall on the left side. The TIFF images and binary masks were then cropped based on the maximum size of breast region and used as input images for the radiomic analysis (Fig. 1).

Extraction of GRFs was performed using the handcrafted features approach [38] based on region of interests defined through two methods: (1) multiple regions of interests defined by the lattice-based approach covering the entire breast image (ROI) [52, 53] and (2) largest square rectangular box inscribed within breast (SQ) [54–56]. Unlike the previous studies' techniques [52–56], we included the pectoral muscle for our radiomic analysis as many of our cases had FP annotations near the pectoral muscle area.

A total of 34 GRFs per image (gray level co-occurrence matrix/GLCM-based Haralick texture features ($n=30$) [57, 58], first-order statistics/FOS features ($n=2$) [33, 58], and neighbourhood gray tone difference matrix/NGTDM texture features ($n=2$) [58, 59], Table 2) were extracted using our in-house MATLAB programs. These features were selected as they have been shown to be valuable descriptors of mammographic appearances in measuring the contrast values of spatial inter-relationships between neighbouring pixels of the image (GLCM and NGTDM) and the deviation of single pixel intensity value within the image region of interest (FOS) [33, 57–62]. The extracted features were then normalised using a z-score algorithm to calibrate image gray intensity mean to 0 and standard deviation values to 1 [63–65]. The ROI features were analysed using MATLAB distinct block processing method with block size 214×214 pixels and summarised using the standard deviation approach.

Model Building

Bivariate analysis comparing potential GRFs to the *difficult-* and *easy-to-interpret* NCs was conducted.

To predict *difficult-* from *easy-to-interpret* NCs, three binary machine learning (ML) classification models were built: (1) *CC*, (2) *MLO*, and (3) *CC*+*MLO* model. Firstly, in constructing the *CC* and *MLO* classification models, the 34 GRFs, obtained from the corresponding right and left CC and MLO view images only, were fed into MATLAB ensemble of 500 decision trees (i.e., random forest) enhanced with LogitBoost ensemble aggregation algorithm (i.e., adaptive logistic regression) [66–68]. We selected the random forest technique because a random feature selection is embedded within the model which can reduce the features overfitting problem while producing interpretable

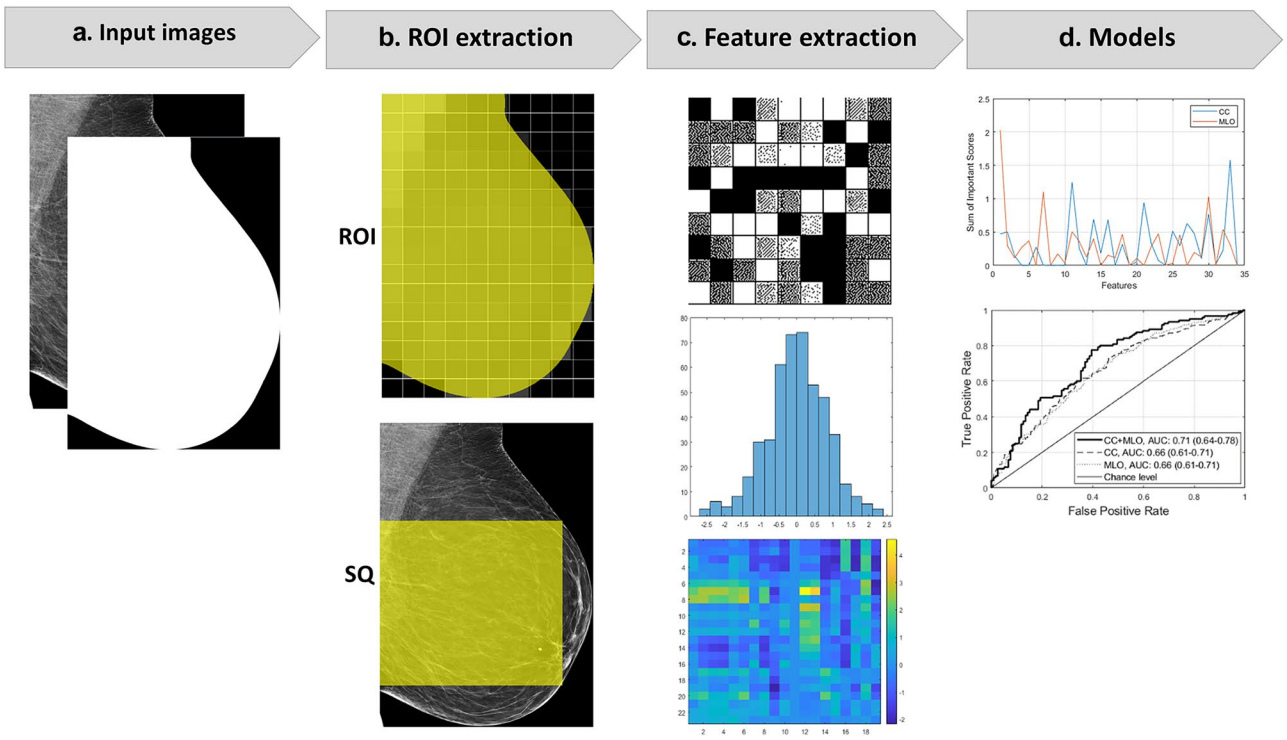


Fig. 1 Study workflow. First, images and masks were acquired and used as input images (**a**). Using the lattice- (ROI) and squared-based (SQ) approaches (**b**), 34 global radiomic features (GRFs) per image were then extracted from the region of interest (yellow region) and normalised (image gray intensity mean = 0 and standard deviation = 1) (**c**). Lastly, three machine learning models for predicting *difficult-to-interpret* mammographic normal cases were built and evaluated (**d**)

Table 2 List of extracted global radiomic features for this study (n = 34)

No.	Name	*p*-value	Importance score from CC model	Importance score from MLO model	Feature class	Parameters
1	ROI_Std_Correlation_3	0.000[a]	0.47	2.03[b,c]	GLCM [57, 58]	3 pixel distance between the pixel of interest and its neighbour
2	ROI_Std_Difference_entropy _3	0.169	0.5	0.29		
3	ROI_Std_Dissimilarity_3	0.245	0.14	0.11		
4	ROI_Std_Difference_variance_3	0.69	0	0.27		
5	ROI_Std_Energy_3	0.025[a]	0	0.37		
6	ROI_Std_Entropy_3	0.022[a]	0.28	0		
7	ROI_Std_Homogeneity_3	0.343	0	1.10[b]		
8	ROI_Std_Information_measure_of_correlation1 _3	0.87	0	0		
9	ROI_Std_Information_measure_of_correlation2 _3	0.582	0	0.17		
10	ROI_Std_Maximum_probability_3	0.029[a]	0	0.03		
11	ROI_Std_Sum_entropy_3	0.000[a]	1.25[b]	0.5		
12	ROI_Std_Sum_variance_3	0.649	0.25	0.36		
13	ROI_Std_Coarseness_3	0.945	0	0.13	NGTDM [58, 59]	3 × 3 neighbourhood size
14	ROI_Std_Cluster_prominence_3	0.398	0.69	0.4	GLCM [57, 58]	3 pixel distance between the pixel of interest and its neighbour
15	ROI_Std_Cluster_shade_3	0.373	0.18	0		
16	ROI_Std_Range_all	0.006[a]	0.68	0.15	FOS [33, 58]	n/a
17	ROI_Std_Sum_of_squares_variance_9	0.215	0	0.12	GLCM [57, 58]	9 pixel distance between the pixel of interest and its neighbour
18	SQ_Correlation_9	0.006[a]	0.31	0.46		
19	SQ_Difference_entropy _9	0.000[a]	0	0		
20	SQ_Dissimilarity_9	0.000[s]	0.05	0.11		
21	SQ_Difference_variance_9	0.000[a]	0.94[b]	0		
22	SQ_Energy_9	0.000[a]	0.33	0.28		
23	SQ_Entropy_9	0.000[a]	0.07	0.47		
24	SQ_Homogeneity_9	0.000[a]	0	0		
25	SQ_Information_measure_of_correlation1 _9	0.000[a]	0.51	0.03		
26	SQ_Information_measure_of_correlation2 _9	0.000[a]	0.3	0.46		
27	SQ_Maximum_probability_9	0.000[a]	0.63	0		
28	SQ_Sum_entropy_9	0.001[a]	0.48	0.19		
29	SQ_Sum_variance_9	0.000[a]	0.11	0.13		
30	SQ_Coarseness_9	0.535	0.77	1.03[b]	NGTDM [58, 59]	9 × 9 neighbourhood size
31	SQ_Cluster_prominence_9	0.511	0.02	0	GLCM [57, 58]	9 pixel distance between the pixel of interest and its neighbour
32	SQ_Cluster_shade_9	0.182	0.22	0.54		
33	SQ_Range_all	0.000[a]	1.57[b,c]	0.31	FOS [33, 58]	n/a
34	SQ_Sum_of_squares_variance_9	0.000[a]	0	0	GLCM [57, 58]	9 pixel distance between the pixel of interest and its neighbour

CC craniocaudal, *FOS* first order statistics, *GLCM* gray level co-occurrence matrix, *MLO* mediolateral oblique, *NGTDM* neighbourhood gray tone difference matrix, *n/a* not applicable, *ROI* multiple regions of interests defined by the lattice-based approach covering the entire breast image, *SQ* largest square rectangular box inscribed within breast, *Std* standard deviation

[a]20 features with statistically significant difference between *difficult-* and *easy-to-interpret* normal cases, p-value < 0.05

[b]Top 6 and [c]top 2 useful features from the machine learning CC/MLO models. The higher the value, the more useful the features were for the predictive models

models with automatic estimation of feature importance [68]. To recognise useful GRFs for the predictive *CC* and *MLO* models, importance scores for each feature were computed through a feature importance analysis using MATLAB's predictor importance algorithm. Importance scores indicate how useful each feature was in the construction of the decision trees within the model. The higher the value, the more useful the features were in making the prediction for the model. Finally, in building the *CC*+*MLO* model, we utilized a late fusion method [69] which consisted of taking the median predictive scores from the constructed *CC* and *MLO* models (Fig. 2).

Statistical Analysis and Validation

To assess the performance of the models on our dataset, we needed to measure how well the predictions made by the model matched the observed data. Thus, we trained and validated the *CC* and *MLO* models using the leave-one-out-cross-validation approach, a reliable, unbiased, and accurate validation method for evaluating the performance of a machine learning model, assuring no data leakage from the training to the testing phase [62, 70]. This means that each time the model was trained, one case (all images belonging to that case) was left out and used once as a test set to validate the predictive performance of the model, while the remaining cases were used to train the model. This process was repeated for all cases until each case was left out once. The performances of the three models for differentiating *difficult-* from *easy-to-interpret* NCs were evaluated by the area under receiver operating characteristic curve (AUC). A DeLong test [71, 72] was used to determine if there was a statistically significant difference in the performance between the three models (*CC* vs *CC*+*MLO*, *MLO* vs *CC*+*MLO*, and *CC* vs *MLO* model). A two-sided p value of <0.05 was used as the criterion of statistically significant difference.

A scree test of exploratory factor analysis [73] was used to determine useful GRFs based on their total importance scores from the *CC* and *MLO* models.

Using a Kruskal–Wallis test, we explored if the 34 GRFs differed between *difficult-* vs *easy-to-interpret* NCs and if the median difficulty level differed among *low-* vs *high-BD* NCs (*difficult-* and *easy-to-interpret*). A p value of <0.05 was considered statistically significant.

All radiomics analysis pipeline and statistical analysis were conducted using MATLAB R2021a (MathWorks, Natick, MA, USA) apart from the DeLong test which was performed using pROC package [74] in RStudio 2021.09.0.351 (RStudio Team, Boston, MA, USA)/R 4.1.2 (R Core Team, Vienna, Austria).

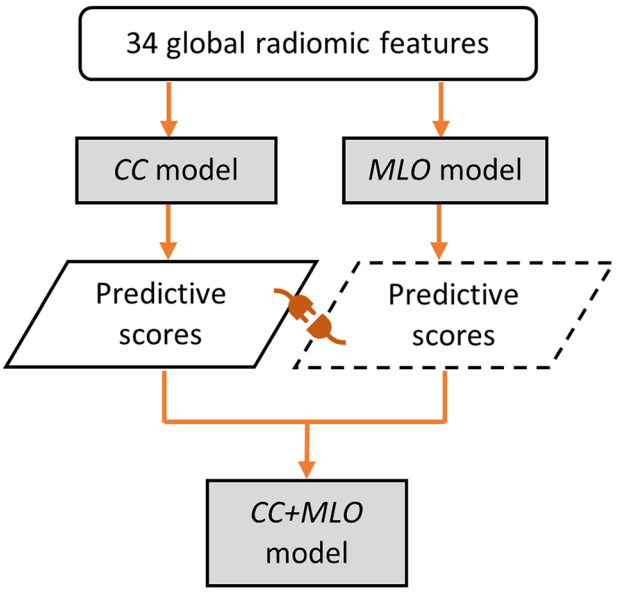


Fig. 2 Model building. First, the *CC* and *MLO* model were built using the 34 global radiomic features extracted from the responding CC and MLO view images only. The *CC*+*MLO* model was the median predictive scores of the *CC* and *MLO* models. CC = craniocaudal, MLO = mediolateral oblique

Results

Radiomics Predicting Difficult Normal Cases

Figure 3 demonstrates the comparisons of the overall classification performance. When only GRFs from either CC or MLO views were included, corresponding *CC* and *MLO* models achieved similar 0.66 AUCs (95% confidence interval/CI: 0.61–0.71). However, when GRFs from both CC and MLO views were combined using the median predictive scores from both *CC* and *MLO* models, the *CC*+*MLO* model reached a higher AUC value of 0.71 (95% CI: 0.64–0.78). Nevertheless, the DeLong test shows that there was no statistically significant difference in the performance between the three models ($Z=-1.561$, $p=0.119$, *CC* vs *CC*+*MLO*; $Z=-1.846$, $p=0.065$, *MLO* vs *CC*+*MLO*; and $Z=0.198$, $p=0.843$, *CC* vs *MLO* model).

Significant Global Radiomic Features

Among the 34 extracted GRFs, 20 features (Table 2) showed a statistically significant difference between *difficult-* vs *easy-to-interpret* NCs ($p<0.05$). When the combined 34 features were used for the ML predictive models, six significant features from predictive *CC* and *MLO* models demonstrated highest importance scores when compared between the 34 features based on the scree test (Fig. 4). The three most important features from *CC* model were

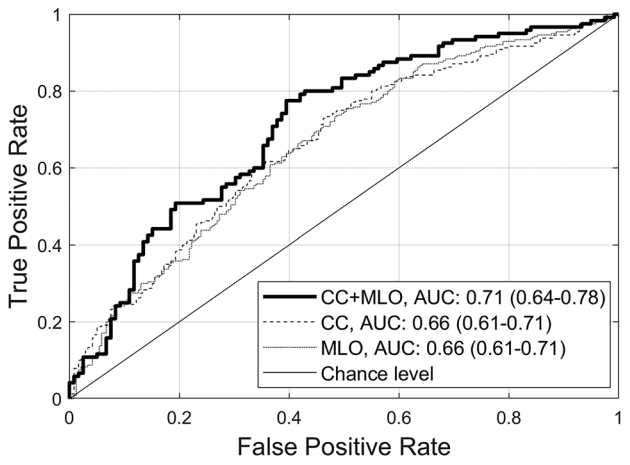


Fig. 3 Area under the receiver operating characteristic curve (AUC) for three classifiers. When differentiating *difficult-* from *easy-to-interpret* normal cases, using GRFs from CC or MLO view images only, the performance of *CC* and *MLO* models achieved similar AUCs of 0.66 (95% CI: 0.61–0.71). However, when fusing GRFs from both CC and MLO images using the median predictive scores from both *CC* and *MLO* models, the combined *CC*+*MLO* model yielded a higher AUC of 0.71 (95% CI: 0.64–0.78). CC = craniocaudal, CI = confidence interval, MLO = mediolateral oblique

ROI_Std_Sum_entropy_3, *SQ_Difference_variance_9*, and *SQ_Range_all* with mean importance score of 1.25. The three most important features from *MLO* model were *ROI_Std_Correlation_3*, *ROI_Std_Homogeneity_3*, and *SQ_Coarseness_9* including 1.39 mean importance score (Table 3).

Breast Density of Difficult Normal Cases

Figure 5 illustrates the comparison of the traditional BD-based difficulty level between *low-* vs *high-BD* NCs (*difficult-* and *easy-to-interpret*). The result from the Kruskal–Wallis test showed no statistically significant difference in the difficulty level between the two groups ($H(1) = 0.14$, $p = 0.709$).

Discussion

This first exploratory study investigated if by employing a radiomic approach to describe the global mammographic features, *difficult-to-interpret* NCs can be predicted. Our work based on the three ML models (*CC*, *MLO*, and *CC*+*MLO*) exhibited that GRFs had the ability to accurately differentiate *difficult-* from *easy-to-interpret* NCs (AUCs up to 0.71). A few earlier studies indicated the value of local radiomic features from specific tumor/FP regions in predicting radiology trainees' error making patterns (AUCs up to 0.61) [39, 40]. However, we focused on GRFs' efficacy in predicting *difficult-to-interpret* NCs (i.e., an objective measure of difficulty based on readers' error rate [31]) based on the significance of radiomics in describing mammographic characteristics [38] and global image processing (i.e., global gist/radiological first impression) in diagnostic decision making [41–44]. Furthermore, our combined *CC*+*MLO* model particularly integrating features from both CC and MLO views of mammograms (using the median predictive scores from both *CC* and *MLO* models) showed higher performance (0.71 AUC, 0.64–0.78) when compared to *CC* or *MLO* model alone (0.66 AUCs, 0.61–0.71) using only features from corresponding CC or MLO images (Fig. 3). This finding suggests that GRFs from both CC and MLO views of mammographic images can provide complementary useful information in predicting *difficult-to-interpret* NCs. Nonetheless, more and larger future works are required to confirm this finding. While the results of comparing *CC* to *MLO* model showed a clearer indication of no statistically significant difference between the model's performance ($Z = 0.198$, $p = 0.843$), the comparisons of *CC* and *MLO* to the combined *CC*+*MLO* model were indefinite ($Z = -1.561$, $p = 0.119$ and $Z = -1.846$, $p = 0.065$, respectively) which suggests more conclusive answers might need a larger sample.

Previous studies reported that the global gist signal without any localised target information could be used to classify normal from abnormal mammograms [41, 42]. An earlier study [75] highlighted the importance of the global gist signal in guiding a visual search and making a diagnosis decision when interpreting difficult cancer cases (unreported retrospectively identified cancers on mammograms). Thus, we hypothesised that *difficult-to-interpret* NCs have global image characteristics which may resemble an abnormal appearance and differ from *easy-to-interpret* images. If a difficult normal image seems to be abnormal at first, this often results in prolonged interpretations, and subsequently, the chance of FP errors may increase since FP errors are often reported at the later stages of prolonged visual searches on normal mammograms [76].

Our work suggests that the GRFs (describing the entire image and largest square rectangular inscribed within breast) contained within mammographic images can help to identify *difficult-to-interpret* NCs. To date the process of assessing difficulty, NCs tends to depend heavily on a manual and subjective evaluation of radiological experts with high-BD cases often perceived as more difficult [19, 26–28, 30]. However, similar to previous studies [28, 29], our finding (Fig. 5) demonstrated no statistically significant difference ($H(1) = 0.14$, $p = 0.709$) observed in the difficulty level based on the traditional BD class between *low-* and *high-BD* NCs, suggesting that BD classifications may not be suitable for estimating difficult NCs. Conversely, our GRFs can provide important details for recognising mammographic features of difficult

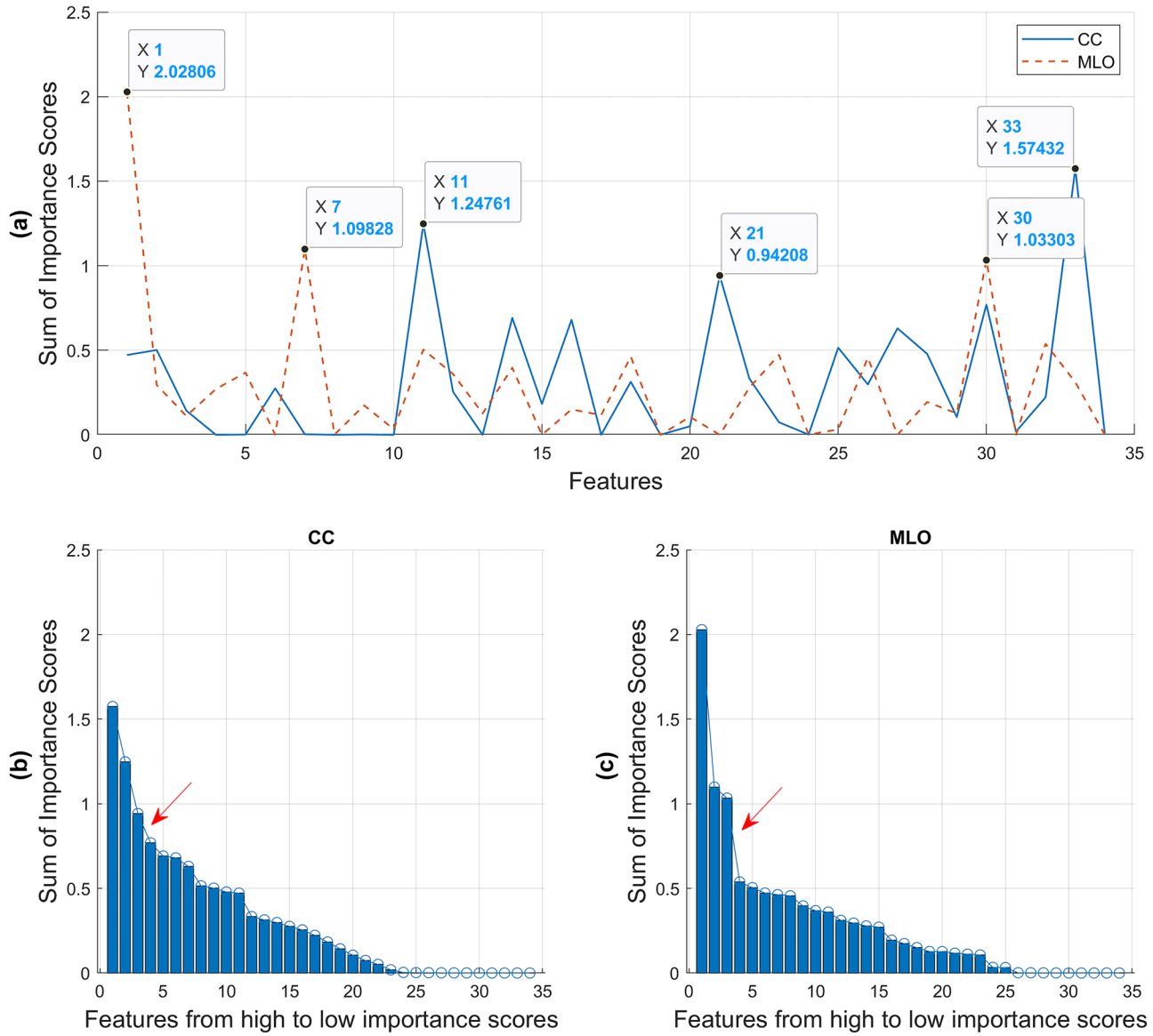


Fig. 4 Significant global radiomic features based on their importance scores derived from the predictive *CC* and *MLO* models. **a** Among the 34 global radiomic features, three features with highest importance scores from *CC* model were *ROI_Std_Sum_entropy_3* (features no. 11) and *SQ_Difference_variance_9* (features no. 21) of GLCM, and *SQ_Range_all* (features no. 33) of FOS, determined based on **b** the scree plot indicating the steep slope occurred after the third highest features (shown by the red arrow). Similarly, three highest importance score features from *MLO* model were *ROI_Std_Correlation_3* (features no. 1) and *ROI_Std_Homogeneity_3* (features no. 7) of GLCM, and *SQ_Coarseness_9* (features no. 30) of NGTDM according to **c** the scree plot with the red arrow showing the steep slope occurred after the third highest features. CC = craniocaudal, FOS = first order statistics, GLCM = gray level co-occurrence matrix, MLO = mediolateral oblique, NGTDM = neighbourhood gray tone difference matrix, ROI = multiple regions of interests defined by the lattice-based approach covering the entire breast image, SQ = largest square rectangular box inscribed within breast, Std = standard deviation

NCs that could be used to facilitate an automatic and consistent approach for evaluating difficult NCs, building more predictable test sets and consistent educational materials for radiologists [14–16, 21]. It can also facilitate a reliable and effective objective measurement of the recall rate performance of each reader as well as screening program over time given that readers' performance in test set based education can correspond well with their clinical performance [15, 20, 22–25].

Additionally, our approach may be used as an augmented clinical artificial intelligence (AI) tool to notify clinical managers about *difficult-to-interpret* NCs so that more experienced readers can be assigned to the case or an appropriate strategy in pairing readers could be employed

Table 3 List of important global radiomic features based on the machine learning models

Feature no.	Name	Models	Feature class	Importance scores	Mean scores	Description
11	ROI_Std_Sum_entropy_3	CC	GLCM [57, 58]	1.25	1.25	Describing the sum of randomness in the gray level distribution of the image
21	SQ_Difference_variance_9	CC	GLCM [57, 58]	0.94		Measuring heterogeneity that places higher weights on differing intensity level pairs that deviate more from the mean
33	SQ_Range_all	CC	FOS [58, 59]	1.57[a]		Describing the difference between maximum and minimum of image gray level values
1	ROI_Std_Correlation_3	MLO	GLCM [57, 58]	2.03[a]	1.39	Measuring similarity or linear dependency of gray level values between the two neighbouring pixels and representing the image's smoothing gradient of texture pattern
7	ROI_Std_Homogeneity_3	MLO	GLCM [57, 58]	1.10		Measuring the closeness of the distribution of elements in the GLCM to the GLCM diagonal
30	SQ_Coarseness_9	MLO	NGTDM [58, 59]	1.03		Measuring average difference between the center pixel and its neighbourhood and is an indication of the spatial rate of change

CC craniocaudal, *FOS* first order statistics, *GLCM* gray level co-occurrence matrix, *MLO* mediolateral oblique, *NGTDM* neighbourhood gray tone difference matrix, *ROI* multiple regions of interests defined by the lattice-based approach covering the entire breast image, *SQ* largest square rectangular box inscribed within breast, *Std* standard deviation

[a]Features with highest importance score based on the *CC* / *MLO* models

to optimise diagnostic accuracy. A single reading strategy, instead of a double reading, can also be exploited for *easy-to-interpret* NCs, reducing readers' workload since NCs constitute about 99.5% of cases in screening programs [13, 77–79]. Nonetheless, more work is required to further explore and verify such an AI tool's effectiveness and reliability before incorporating it into a clinical practice.

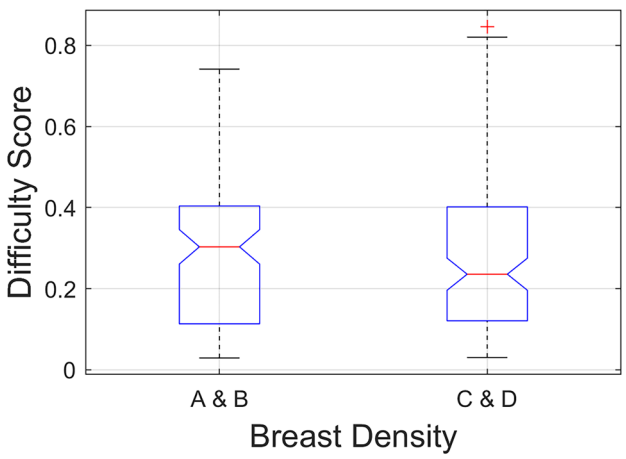


Fig. 5 Comparison of difficulty level between ***low-BD*** (A&B) vs ***high-BD*** (C&D) normal cases**.** No statistically significant difference in the difficulty level was found between the two groups ($H(1) = 0.14$, $p = 0.709$). A = Fatty: 0–25% glandular, B = Fibroglandular: 25–50% glandular, BD = breast density, C = Heterogenous: 51–75% glandular, D = Extremely dense: 76–100% glandular

Our ML approach based on GRFs indicated that the overall mammographic appearance is an important factor in analysing *difficult-to-interpret* NCs. From the total of 34 features, 20 features showed statistically significant difference between *difficult-* vs *easy-to-interpret* NCs ($p < 0.05$, Table 2). Nevertheless, when the combined 34 features were used for the ML-based predictive models, more redundant features were discovered, resulted in six helpful GRFs in describing *difficult-to-interpret* NCs based on their importance scores (Fig. 4 and Table 3). The MLO- rather than the CC-derived parameters appeared to be more valuable which may be linked to the MLO-based GRFs involving more breast tissue including the pectoral muscle. This finding, though, needs further investigation by larger future studies.

Some limitations of this study should be acknowledged. The prevalence of NCs in the BREAST test sets used in this study was lower than the real clinical practice. Also, this first proof-of-concept study only examined 34 mammography-based handcrafted GRFs' ability in predicting *difficult-to-interpret* NCs based on the average difficulty scores of all readers. Moreover, determining important GRFs based on the scree test [73] (Fig. 4b, c) was less clear cut for the *CC* model when compared to the *MLO*. Future works should explore other potential useful CC-based GRFs (e.g., feature no. 30, 14, 16, and 27) and type of radiomic features (e.g., local handcrafted or deep learning features) and using

equipment from other imaging vendors (e.g., Canon, Canon Medical Systems Corporation, Otawara, Tochigi, Japan). Finally, separate models for predicting difficult cases of readers at various experience levels should be investigated.

Conclusions

Our findings suggested that quantitative GRFs extracted from mammograms are helpful for accurately distinguishing *difficult-* from *easy-to-interpret* NCs, denoting the importance of global gist in making a diagnostic decision. Six important features from the classification models were recognised and highlighted. Our findings could be useful for radiology educations and mammography screening programs in creating more predictable test sets, and reliable and effective objective measurement of the recall rate performance of each reader and screening program over time, as well as optimising single and double reading practice.

Abbreviations AI: Artificial intelligence; AUC: Area under receiver operating characteristic curve; BD: Breast density; BI-RADS: Breast Imaging and Reporting Data System; BREAST: BreastScreen Reader Assessment Strategy; CC: Craniocaudal; CI: Confidence interval; DICOM: Digital Imaging and Communications in Medicine; FOS: First order statistics; FP: False positive; GLCM: Gray level co-occurrence matrix; GRFs: Global radiomic features; ML: Machine Learning; MLO: Mediolateral oblique; NCs: Normal cases; NGTDM: Neighbourhood gray tone difference matrix; RANZCR: Royal Australian and New Zealand College of Radiologists; ROI: Multiple regions of interests defined by the lattice-based approach covering the entire breast image; SQ: Largest square rectangular box inscribed within breast

Acknowledgements The authors thank the BreastScreen Reader Assessment Strategy (BREAST) program at the University of Sydney for access and use of their data for the study. BREAST is funded/supported by the Commonwealth Government Department of Health, with their images kindly provided by BreastScreen Australia, the National Breast Cancer Foundation, and the Cancer Institute New South Wales. We would like to also acknowledge the funding support for this research project provided by the National Breast Cancer Foundation [grant ID: IIRS-22-087] and Cancer Institute New South Wales [grant ID: 2022/ECF1426]. The authors wish to thank Dr Robert Heard for his kind statistical advice as well as the reviewers for their rigorous evaluation of this manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This work is supported by the National Breast Cancer Foundation [grant ID: IIRS-22-087] and Cancer Institute New South Wales [grant ID: 2022/ECF1426].

Data Availability The data that support the findings of this study are available from BREAST but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.Data are however available from the authors upon reasonable request and with permission of BREAST.

References

1. Sung H, Ferlay J, Siegel RL et al (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians 71:209–249
2. Seely JM, Alhassan T (2018) Screening for breast cancer in 2018-what should we be doing today? Current oncology (Toronto, Ont) 25:S115-S124
3. Njor S, Nystrom L, Moss S et al (2012) Breast cancer mortality in mammographic screening in Europe: a review of incidence-based mortality studies. 19:33-41
4. Welfare AIoHa (2020) BreastScreen Australia monitoring report 2020. AIHW, Canberra
5. Castells X, Torá-Rocamora I, Posso M et al (2016) Risk of Breast Cancer in Women with False-Positive Results according to Mammographic Features. Radiology 280:379-386
6. Román M, Castells X, Hofvind S, Euler-Chelpin M (2016) Risk of breast cancer after false-positive results in mammographic screening. Cancer Medicine 5:1298-1306
7. Ekpo EU, Alakhras M, Brennan P (2018) Errors in Mammography Cannot be Solved Through Technology Alone. Asian Pacific journal of cancer prevention : APJCP 19:291-301
8. Skaane P (2019) How Can We Reduce Unnecessary Procedures after Screening Mammography? Radiology 291:318-319
9. Ong M-S, Mandl KD (2015) National Expenditure For False-Positive Mammograms And Breast Cancer Overdiagnoses Estimated At $4 Billion A Year. Health Affairs 34:576-583
10. Vlahiotis A, Griffin B, Stavros Md FAT, Margolis J (2018) Analysis of utilization patterns and associated costs of the breast imaging and diagnostic procedures after screening mammography. ClinicoEconomics and Outcomes Research Volume 10:157-167
11. Alcusky M, Philpotts L, Bonafede M, Clarke J, Skoufalos A (2014) The Patient Burden of Screening Mammography Recall. Journal of Women's Health 23:S-11-S-19
12. Lee W, Peters G (2013) Mammographic screening for breast cancer: A review. Journal of Medical Radiation Sciences 60:35-39
13. Peintinger F (2019) National Breast Screening Programs across Europe. Breast Care 14:354-358
14. Brennan PC, Trieu PD, Tapia K, Ryan J, Mello-Thoms C, Lee W (2014) BREAST: A Novel Strategy to Improve the Detection of Breast Cancer. In: Fujita H, Hara T, Muramatsu C, (eds) International Workshop on Digital Mammography 2014. Springer International Publishing Switzerland, Cham, pp 438-443
15. Gale A, Chen Y (2020) A review of the PERFORMS scheme in breast screening. The British journal of radiology 93:20190908
16. Suleiman ME, Rickard M, Brennan PC (2020) Perfecting detection through education. Radiography 26:S49-S53
17. Trieu PD, Tapia K, Frazer H, Lee W, Brennan P (2019) Improvement of Cancer Detection on Mammograms via BREAST Test Sets. Academic Radiology 26:e341–e347
18. Trieu PD, Lewis SJ, Li T et al (2021) Improving radiologist's ability in identifying particular abnormal lesions on mammograms

through training test set with immediate feedback. Scientific reports 11:1-8
19. Suleiman WI, Rawashdeh MA, Lewis SJ et al (2016) Impact of Breast Reader Assessment Strategy on mammographic radiologists' test reading performance. Journal of Medical Imaging and Radiation Oncology 60:352–358
20. Chen Y, James JJ, Cornford EJ, Jenkins J (2020) The Relationship between Mammography Readers' Real-Life Performance and Performance in a Test Set–based Assessment Scheme in a National Breast Screening Program. Radiology: Imaging Cancer 2:e200016
21. Chen Y, Gale A (2018) Performance Assessment Using Standardized Data Sets: The PERFORMS Scheme in Breast Screening and Other Domains. In: Samei E, Krupinski EA, (eds) The Handbook of Medical Image Perception and Techniques. Cambridge University Press, Cambridge, pp 328-342
22. Brancato B, Peruzzi F, Saieva C et al (2022) Mammography self-evaluation online test for screening readers: an Italian Society of Medical Radiology (SIRM) initiative. Eur Radiol 32:1624-1633
23. Soh BP, Lee W, McEntee MF et al (2013) Screening Mammography: Test Set Data Can Reasonably Describe Actual Clinical Reporting. Radiology 268:46-53
24. Soh BP, Lee WB, Mello-Thoms C et al (2015) Certain performance values arising from mammographic test set readings correlate well with clinical audit. Journal of Medical Imaging and Radiation Oncology 59:403-410
25. Qenam BA, Li T, Brennan PC (2021) Test set participation may impact positively on clinical mammographic performanceSPIE 11599, Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment. SPIE, Online, pp 1–8
26. Dong L, Chen Y, Gale A (2013) Breast screening: understanding case difficulty and the nature of errorsSPIE 8673, Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment. SPIE, Lake Buena Vista (Orlando Area), Florida, United States, pp 1-8
27. Scott HJ, Gale AG (2005) Breast screening technologists: When is a difficult case truly difficult and for whom?SPIE 5749, Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment. SPIE, Bellingham, WA, pp 557-565
28. Wong D, Gandomkar Z, Reed W, Lewis SJ, Brennan P (2022) The effect on reader performance of case image characteristics including expert determined difficulty ratings using a mammography test set. Under review
29. Grimm LJ, Kuzmiak CM, Ghate SV, Yoon SC, Mazurowski MA (2014) Radiology Resident Mammography Training. Academic Radiology 21:888-892
30. Ang ZZ, Rawashdeh MA, Heard R, Brennan PC, Lee W, Lewis SJ (2017) Classification of normal screening mammograms is strongly influenced by perceived mammographic breast density. Journal of Medical Imaging and Radiation Oncology 61:461-469
31. Mazurowski MA (2013) Difficulty of mammographic cases in the context of resident training: preliminary experimental dataSPIE 8673, Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment SPIE, Lake Buena Vista (Orlando Area), Florida, United States, pp 1–6
32. Kabbai L, Abdellaoui M, Douik A (2019) Image classification by combining local and global features. The Visual Computer 35:679-693
33. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: Images Are More than Pictures, They Are Data. Radiology 278:563-577
34. Gillies RJ, Schabath MB (2020) Radiomics Improves Cancer Screening and Early Detection. Cancer Epidemiology Biomarkers & Prevention 29:2556-2567
35. Lee S-H, Park H, Ko ES (2020) Radiomics in Breast Imaging from Techniques to Clinical Applications: A Review. Korean Journal of Radiology 21:779-792
36. Tagliafico AS, Piana M, Schenone D, Lai R, Massone AM, Houssami N (2020) Overview of radiomics in breast cancer diagnosis and prognostication. The Breast 49:74-80
37. Conti A, Duggento A, Indovina I, Guerrisi M, Toschi N (2020) Radiomics in breast cancer classification and prediction. Seminars in Cancer Biology 72:238-250
38. Siviengphanom S, Gandomkar Z, Lewis SJ, Brennan PC (2021) Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs. Academic Radiology 29:1228-1247
39. Zhang J, Lo JY, Kuzmiak CM, Ghate SV, Yoon SC, Mazurowski MA (2014) Using computer-extracted image features for modeling of error-making patterns in detection of mammographic masses among radiology residents. Medical Physics 41:091907
40. Zhang J, Silber JI, Mazurowski MA (2015) Modeling false positive error making patterns in radiology trainees for improved mammography education. Journal of Biomedical Informatics 54:50-57
41. Evans KK, Georgian-Smith D, Tambouret R, Birdwell RL, Wolfe JM (2013) The gist of the abnormal: Above-chance medical decision making in the blink of an eye. Psychonomic Bulletin & Review 20:1170-1175
42. Gandomkar Z, Ekpo E, Lewis S et al (2018) Detection of the abnormal gist in the prior mammograms even with no overt sign of breast cancerSPIE 10718, 14th International Workshop on Breast Imaging (IWBI 2018). SPIE, Atlanta, Georgia, United States, pp 1-6
43. Brennan PC, Gandomkar Z, Ekpo EU et al (2018) Radiologists can detect the 'gist' of breast cancer before any overt signs of cancer appear. Scientific reports 8:8717-8717
44. Gandomkar Z, Siviengphanom S, Ekpo EU et al (2021) Global processing provides malignancy evidence complementary to the information captured by humans or machines following detailed mammogram inspection. Scientific reports 11:1-12
45. Gandomkar Z, Mello-Thoms C (2019) Visual search in breast imaging. The British journal of radiology 92:20190057
46. Lehman CD, White E, Peacock S, Drucker MJ, Urban N (1999) Effect of age and breast density on screening mammograms with false-positive findings. American Journal of Roentgenology 173:1651-1655
47. Holland K, Van Gils CH, Mann RM, Karssemeijer N (2017) Quantification of masking risk in screening mammography with volumetric breast density maps. Breast Cancer Research and Treatment 162:541-548
48. Brennan P, Lee W, Tapia K (2018) Breast Screen Reader Assessment Strategy (BREAST): A Research Infrastructure with a Translational Objective. In: Samei E, Krupinski EA, (eds) The Handbook of Medical Image Perception and Techniques. Cambridge University Press, Cambridge, pp 343–356
49. Brennan PC, Tapia K, Ryan J, Lee W (2013) BREAST: a novel method to improve the diagnostic efficacy of mammographySPIE 8673, Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment. SPIE, Lake Buena Vista (Orlando Area), Florida, United States, pp 1–5
50. Tapia KA, Rickard MT, McEntee MF, Garvey G, Lydiard L, C Brennan P (2020) Impact of breast density on cancer detection: observations from digital mammography test sets. International Journal of Radiology & Radiation Therapy 7:36–41
51. McEntee M, Brennan P, Evanoff M, Phillps P, O Connor W, Manning D (2006) Optimum ambient lighting conditions for the viewing of softcopy radiological images. SPIE
52. Zheng Y, Keller BM, Ray S et al (2015) Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment. Medical Physics 42:4149-4160
53. Zheng Y, Wang Y, Keller B, Conant E, Gee J, Kontos D (2013) A fully-automated software pipeline for integrating breast density

and parenchymal texture analysis for digital mammograms: parameter optimization in a case-control breast cancer risk assessment studySPIE 8670, Medical Imaging 2013: Computer-Aided Diagnosis SPIE Lake Buena Vista (Orlando Area), Florida, United States 1 7
54. Wei J, Chan H-P, Wu Y-T et al (2011) Association of Computerized Mammographic Parenchymal Pattern Measure with Breast Cancer Risk: A Pilot Case-Control Study. Radiology 260:42-49
55. Wu Y-T, Sahiner B, Chan H-P et al (2008) Comparison of mammographic parenchymal patterns of normalsubjects and breast cancer patientsSPIE 6915, Medical Imaging 2008: Computer-Aided Diagnosis. SPIE, San Diego, California, United States, pp 1-8
56. Gandomkar Z, Suleiman M, Demchig D, Brennan P, McEntee M (2019) BI-RADS density categorization using deep neural networksSPIE 10952, Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment. SPIE, pp 1–7
57. Haralick RM, Shanmugam K, Dinstein IH (1973) Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics SMC-3:610–621
58. Costaridou L (2005) Medical image analysis methods. Taylor & Francis, Boca Raton
59. Amadasun M, King R (1989) Textural features corresponding to textural properties. IEEE Transactions on Systems, Man, and Cybernetics 19:1264-1274
60. Cui WJ, Wang C, Jia L et al (2019) Differentiation Between G1 and G2/G3 Phyllodes Tumors of Breast Using Mammography and Mammographic Texture Analysis. Frontiers in Oncology 9:1-9
61. Jiang X, Zou X, Sun J, Zheng A, Su C (2020) A Nomogram Based on Radiomics with Mammography Texture Analysis for the Prognostic Prediction in Patients with Triple-Negative Breast Cancer. Contrast Media & Molecular Imaging 2020:1-10
62. La Forgia D, Fanizzi A, Campobasso F et al (2020) Radiomic Analysis in Contrast-Enhanced Spectral Mammography for Predicting Breast Cancer Histological Outcome. Diagnostics 10:1-11
63. Castaldo R, Pane K, Nicolai E, Salvatore M, Franzese M (2020) The Impact of Normalization Approaches to Automatically Detect Radiogenomic Phenotypes Characterizing Breast Cancer Receptors Status. Cancers 12:1-25
64. Kotsiantis SB, Kanellopoulos D, Pintelas PE (2007) Data Preprocessing for Supervised Leaning. International Journal of Computer and Information Engineering 1:4091-4096
65. Haga A, Takahashi W, Aoki S et al (2019) Standardization of imaging features for radiomics analysis. The Journal of Medical Investigation 66:35-37
66. Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). The Annals of Statistics 28:337-407
67. Breiman L (2001) Random Forests. Machine Learning 45:5-32
68. Couronné R, Probst P, Boulesteix A-L (2018) Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 19:1-14
69. Lee SL, Zare MR, Muller H (2019) Late fusion of deep learning and handcrafted visual features for biomedical image modality classification. IET Image Processing 13:382-391
70. Molinaro AM, Simon R, Pfeiffer RM (2005) Prediction error estimation: a comparison of resampling methods. Bioinformatics 21:3301-3307
71. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 44:837-845
72. Sun X, Xu W (2014) Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. IEEE Signal Processing Letters 21:1389-1393
73. Ledesma RD, Valero-Mora P, Macbeth G (2015) The Scree Test and the Number of Factors: a Dynamic Graphics Approach. The Spanish Journal of Psychology 18
74. Robin X, Turck N, Hainard A et al (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77
75. Nodine CF, Mello-Thoms C, Weinstein SP et al (2001) Blinded Review of Retrospectively Visible Unreported Breast Cancers: An Eye-Position Analysis. Radiology 221:122-129
76. Mello-Thoms C (2006) How Does the Perception of a Lesion Influence Visual Search Strategy in Mammogram Reading? Academic Radiology 13:275-288
77. Taylor-Phillips S, Stinton C (2020) Double reading in breast cancer screening: considerations for policy-making. The British journal of radiology 93:20190610
78. Brennan PC, Ganesan A, Eckstein MP et al (2019) Benefits of Independent Double Reading in Digital Mammography. Academic Radiology 26:717-723
79. Gandomkar Z, Tay K, Brennan PC, Kozuch E, Mello-Thoms C (2018) Can eye-tracking metrics be used to better pair radiologists in a mammogram reading task? Medical Physics 45:4844-4856

4.9 Chapter findings

This Chapter 4 indicated that global radiomic features extracted from mammograms had the ability to discriminate *difficulty*- from *easy-to-interpret* normal cases of radiologists with varying experience in mammographic interpretation (AUCs of up to 0.71). Six important features for predicting these cases were identified and these features were *sum entropy, difference variance, range, correlation, homogeneity,* and *coarseness*. It was also reported that correlation between breast density and difficulty scores of these cases was not statistically significant.

4.10 References

1. Suleiman, W.I., et al., *Impact of Breast Reader Assessment Strategy on mammographic radiologists' test reading performance.* Journal of Medical Imaging and Radiation Oncology, 2016. **60**(3): p. 352-358.
2. Chen, Y., et al., *The Relationship between Mammography Readers' Real-Life Performance and Performance in a Test Set–based Assessment Scheme in a National Breast Screening Program.* Radiology: Imaging Cancer, 2020. **2**(5): p. e200016.
3. Suleiman, M.E., M. Rickard, and P.C. Brennan, *Perfecting detection through education.* Radiography, 2020. **26**: p. S49-S53.
4. Peintinger, F., *National Breast Screening Programs across Europe.* Breast Care, 2019. **14**(6): p. 354-358.
5. Lee, W. and G. Peters, *Mammographic screening for breast cancer: A review.* Journal of Medical Radiation Sciences, 2013. **60**(1): p. 35-39.
6. Lee, S.L., M.R. Zare, and H. Muller, *Late fusion of deep learning and handcrafted visual features for biomedical image modality classification.* IET Image Processing, 2019. **13**(2): p. 382-391.

Chapter 5 – A machine learning model based on global mammographic radiomic features can predict which normal mammographic cases radiology trainees find most difficult

5.1 Chapter introduction

The findings from the previous Chapter 4 indicate that there was a value in mammography-based global radiomic features (driving the global gist signals - Chapter 3) for predicting *difficult-to-interpret* normal cases of average readers with diverse levels of mammographic reading experience (peer reviewed journal article). Similarly, the effectiveness of these global mammographic radiomic features in describing *difficulty-to-interpret* normal cases for expert radiologists with more than 20 years of experience reading mammograms, was also uncovered (conference proceedings - Appendix 1.5). Nonetheless, their ability in explaining highly difficult normal cases specifically for radiology trainees who possess limited skills in processing global images and experience in reading mammograms, has had limited study (as per the results of the literature review - Chapter 2), when there is a value of gist for readers' expertise development.

Building on the above findings, the purpose of this Chapter 5 is to examine if *hardest-to-interpret* normal cases of radiology trainees can be predicted using global mammographic radiomic features (same features used in Chapter 4) and if there are any best helpful features at explaining these cases for this cohort. Association between BI-RADS breast density and difficulty scores of these normal cases is also examined since BI-RADS breast density category has often been used as the predictor of case difficulty when generating educational test sets. This study is particularly crucial for enhancing and expediting radiology trainees' learning so that personalised mammographic education test sets curation for radiology trainees can be developed according to the final **objective 4** in the introduction (Chapter 1).

The paper in this chapter has been submitted for publication in the Journal of Imaging Informatics in Medicine and is currently under review.

Abstract has been accepted and will be presented at Radiological Society of North America (RSNA) 2024 Annual Meeting, 1-5 December, McCormick Place, Chicago, IL, United States (Appendix 1.7).

A machine learning model based on global mammographic radiomic features can predict which normal mammographic cases radiology trainees find most difficult

Abstract

This study aims to investigate whether global mammographic radiomic features (GMRFs) can distinguish *hardest-* from *easiest-to-interpret* normal cases for radiology trainees (RTs). Data from 137 RTs were analysed, with each interpreting seven educational self-assessment test sets comprising 60 cases (40 normal and 20 cancer). The study only utilized normal cases (n=280). Difficulty scores were computed based on the percentage of readers who incorrectly classified each case, leading to their classification as *hardest-* or *easiest-to-interpret* based on whether their difficulty scores fell within and above the 75th or within and below the 25th percentile, respectively (resulted in 140 cases in total used). 59 *low-density* and 81 *high-density* cases were identified. 34 GMRFs were extracted for each case. A random forest machine learning model was trained to differentiate between *hardest-* and *easiest-to-interpret* normal cases and validated using leave-one-out-cross-validation approach. The model's performance was evaluated using the area under receiver operating characteristic curve (AUC). Significant features were identified through feature importance analysis. Difference between *hardest-* and *easiest-to-interpret* cases among 34 GMRFs and in difficulty level between *low-* and *high-density* cases was tested using Kruskal-Wallis. The model achieved 0.75 AUC with *cluster prominence* and *range* emerging as the most useful features. 15 GMRFs differed significantly (p<0.05) between *hardest-* and *easiest-to-interpret* cases. Difficulty level among *low-* vs *high-density* cases did not differ significantly (p=0.12). GMRFs can predict *hardest-to-interpret* normal cases for RTs, underscoring the importance of GMRFs in identifying most difficult normal cases for RTs and facilitating customized training programs tailored to trainees learning needs.

1. Introduction

Female breast cancer (BC) stands as one of the leading causes of cancer-related mortality worldwide [1]. Mammographic screening programs, aimed at early detection of BC, have been shown to significantly reduce BC mortality rate by up to 40% [2]. However, interpreting mammograms is a challenging task susceptible to errors, and relying primarily on technological advancements is insufficient to mitigate or address these challenges [3]. Previous studies have reported that correctly interpretating normal cancer-free mammograms can also be highly difficult for readers with inadequate mammographic reading experience such as radiology trainees (RTs) [4, 5]. The capability of readers including RTs to accurately interpret screening mammograms plays a pivotal role in the efficacy of breast screening programs. Despite efforts to propose criteria for screen reader certification aimed at reducing diagnostic variability among readers, none of the current guidelines for these criteria serve as strong predictors of reader performance [6, 7]. As a result, wide inter-reader variability persists, highlighting the critical need to prioritize education and training for readers, especially for less experienced readers like RTs, in this domain [6-9]. Given that normal cases comprise over 99% of all cases in breast screening programs [10, 11], along with the fact that false positive screening results cause huge healthcare costs of nearly US$3 billion yearly (in the United States as an example) and psychological distress to patients due to unnecessary recalls involving invasive procedures [12], it is imperative that education strategies are prioritized to give extra attention to normal cases.

Studies from medical image perception and basic vision science demonstrate that exposing RTs to large mammographic normal cases via test set training programs, often containing nearly 70% of normal cases in a test set, can enhance their expertise development in interpreting normal cases [13-15]. Still, the present test set curation process using experts' or individual's perception of case difficulty via the breast imaging and reporting data system (BI-RADS) breast density and/or benign features are unreliable predictors of normal cases' difficulty among RTs [11, 16-18]. Therefore, there is a strong demand for a more systematic approach of predicting RTs' difficult normal cases so that customized mammographic training test sets for RTs can be facilitated while their false positive errors can be better monitored and improved.

The process of interpreting medical images necessitates the development of capabilities for globally processing images. Earlier studies showed that individuals with less experience such as RTs, often lack the ability to swiftly process the global clues indicative of abnormalities [19]. To investigate this further, we designed an experiment protocol to gather the initial impressions of radiologists regarding the abnormality of a case, known as the gist signal [19-21]. This signal not only indicates a current cancer case [22] but also serves as a predictor of future BC [21]. Moreover, recent research in radiomics has shown that quantitative global mammographic radiomic (or computer extract image) features align closely with the holistic perceptual information gleaned from radiologists' rapid first impressions about the presence of an image abnormality or global gist signal [20, 23]. This signal has proven valuable in guiding visual searches and aiding in making diagnostic decisions when interpreting challenging mammographic cases [11, 22, 24].

Radiomic features capture various aspects of the mammogram, including texture, intensity, and spatial relationships, providing comprehensive information about the underlying tissue characteristics [25, 26]. Radiomic features have been the subject of systematic investigation in previous works, where local radiomic features were utilized to identify missed mammographic masses and false positive location errors of RTs, achieving an area under receiver operating characteristic curves (AUC) of up to 0.61 [4, 5]. More recent studies [11, 16] have revealed that using global mammographic radiomic features (GMRFs), most difficult/easiest normal cases for readers in general and experienced radiologists can be accurately predicted (AUCs of up to 0.73). However, to date, there has been no study exploring whether GMRFs can specifically predict *hardest-* from *easiest-to-interpret* normal cases for RTs. Considering the fact that RTs are still developing their holistic processing capabilities compared to experienced radiologists [13, 21, 22, 24], this paper explores whether GMRFs can serve as reliable predictors specifically for *hardest-to-interpret* normal cases for RTs. It also seeks to recognise if there are any specific significant GMRFs that denote the *hardest-to-interpret* normal cases for RTs. Knowing these could be valuable not only in the training programs' test set curation process of most difficult normal cases for RTs but also facilitating their expertise development in recognising normal image features that could be most difficult for them often resulting in false positive errors. This is so that their specific skills in perceiving the global signature of normality can be well developed since RTs are generally not as good at recognising the normal features as compared to experienced radiologists [13, 24, 27].

Furthermore, since the current approach to curating test sets relies on experts' opinions regarding case difficulty, primarily using breast density as a surrogate metric [11, 16-18], we also examine if a significant difference exists in difficulty level between *low-density* (BI-RADS category A = fatty and B = fibroglandular) and *high-density* (category C = heterogeneous and D = extremely dense) normal cases.

2. Materials and Methods

2.1 Readers

Ethical approvals of this research were acquired from the University of Sydney's Human Research Ethics Committee [protocol no. 2019/013 and 2017/028], which included obtained written informed consent from each reader participating in the study.

A total of 137 Australian and New Zealand RTs who participated in the BREAST (Breastscreen REader Assessment STrategy) program [28] from 04 September 2014 to 23 February 2021 were included in the study. The mean age of the readers was 33 based on the age details provided by 116 RTs while 21 did not provide their age details. In total, 19% of the readers engaged in the program through a radiology/BC conference/workshop in a simulated reading room while 81% completed the program online through the BREAST platform (https://breast-australia.sydney.edu.au/) at their usual clinical workplace. Reading conditions in both settings were equivalent with cases being presented on two 5-megapixel medical standard monitors including ambient light levels range of 20-40 lux [29]. RTs generally had less than 1 year experience reading mammograms (59%), read less than 20 mammographic cases per week (79%), and spent less than 4 hour per week reading mammograms (96%). Also, 16 readers (12%) completed a fellowship in breast imaging that lasted for 3 to 6 months and had any experience reading images for a national breast screening program (Table 1).

2.2 Mammographic normal cases

This retrospective study includes data from readers that completed seven BREAST test sets, each comprising 60 de-identified full-field digital mammographic (FFDM) cases, with 40 cases being cancer-free and 20 cases indicating cancer. For the purpose of this investigation, only normal cancer-free cases were included. Thus, a total of 280 normal FFDM cases (comprised of four images per case: two bilateral craniocaudal/CC and two mediolateral oblique/MLO views) acquired from screening asymptomatic women aged between 40 to 75 (mean/median = 57, standard deviation = 7, and range = 33 based on the age details of 130 women/cases while ages of 10 women were not available due to anonymization process), were utilized. These cases were collected from various mammography machines, including Fujifilm (Fujifilm Corporation, Minato City, Tokyo, Japan), GE (GE Healthcare, Chicago, IL, USA), Hologic (Hologic, Inc., Marlborough, MA, USA), Philips (Philips Healthcare, Amsterdam, the Netherlands), Sectra (Sectra, Linköping, Sweden), and Siemens (Munich, Germany). Each normal case' ground truth underwent confirmation of cancer free status by at least two independent expert radiologists, each possessed over 20 years of experience, with validation through subsequent negative screen outcomes.

2.3 Categorising *hardest-* vs *easiest-to-interpret* normal cases

To categorise normal cases as *hardest-* vs *easiest-to-interpret*, we initially calculated difficulty scores for each of the 280 normal cases using the Royal Australian and New Zealand College of Radiologists (RANZCR) scoring system [28] provided by the 137 RTs. A RANZCR score of 1 represents a normal case while a score of 2 indicates that RTs considered the case as benign. Scores of 3, 4, or 5 signify an equivocal or malignant finding observed by the readers with higher numbers indicating a greater confidence of disease presence. Difficulty scores were computed by dividing the number of RTs who misclassified a normal case as cancer (provided a rating of 3, 4, or 5) by the total number RTs who read the test set. Subsequently, the 280 normal cases were then classified as *hardest-* and *easiest-to-interpret* cases based on the 75th and 25th percentiles of the cases containing the highest and lowest difficulty scores, correspondingly. This resulted in 70 *hardest-* and 70 *easiest-to-interpret* normal cases, totalling of 140 normal cases/560 DICOM images combined. To ensure a clear distinction between the most and least difficult normal cases in order to enable optimal features learning for our machine learning model, only images in the upper and lower quartiles were used for further analysis. Normal cases in these two categories also resulted in a total of 59 *low-density* (35 *hardest-* and 24 *easiest-to-interpret*) and 81 *high-density* normal cases (35 *hardest-* and 46 *easiest-to-interpret*).

2.4 Global radiomic feature/score per case

To obtain a global radiomic feature/score for each of the 140 normal cases, a total of 34 quantitative GMRFs of 560 images belong to 140 cases extracted from the previous study [11] were used. Details about the GMRFs extraction method can be found in the previous study [11]. Briefly, from the 560 DICOM images, we first generated 560 binary masks using a standardised gray level threshold value set at 100 (manual adjustments on the value were also performed where needed) to extract the required breast region from its background while eliminating unwanted artefacts and labels. After that we converted the DICOM images and masks to a TIFF file format, flipped all the right CC and MLO images to the left to have consistent left side chest wall on all images, and cropped the TIFF images and masks based on the maximum size of breast region. These cropped TIFF images and masks were used as input for the radiomic analysis - Figure 1. We then outlined the region of interest on the images using the lattice-based/ROI (multiple regions of interests covering the entire breast image) [30] and squared-based/SQ (largest square rectangular box inscribed within breast) [31] approaches. The 34 handcrafted GMRFs/scores per image were afterward extracted based on the region of interest delineated using our in-house MATLAB platforms and then normalised to calibrate image intensity mean to zero and standard deviation to one using a common z-score normalisation method [32]. The extracted GMRFs of 4 images (i.e., 2 CC and 2 MLO images) belonging to that case were then combined and averaged to obtain one global radiomic feature/score per case.

The 34 GMRFs consisted of 30 gray level co-occurrence matrix/GLCM-based Haralick texture features [33], 2 neighbourhood gray tone difference matrix/NGTDM-based texture features [34], and 2 first-order statistics/FOS-based features [25] (Table 2 and Figure 1). These features were chosen because of their usefulness in describing mammographic appearances in measuring the contrast values of spatial inter-relationships between neighbouring pixels (GLCM and NGTDM) and the distribution of single pixel intensity value within the image region of interest (FOS) [25, 33].

2.5 Model building

For the task of differentiating between normal cases that are *hardest-* and those of *easiest-to-interpret* for RTs, a random forest machine learning model was built using the averaged 34 GMRFs feeding through MATLAB ensemble of decision trees boosted with the adaptive logistic regression method (i.e., LogitBoost). The random forest with boosting approach was chosen due to its ability to produce explainable model with automatic estimation of feature importance and its built-in feature selection approach which could minimise features overfitting and potential selection bias problems [35]. Important GMRFs were recognised based on the analysis of feature importance scores using MATLAB's predictor importance algorithm, which also helps in identifying and mitigating redundant and biased features. Feature importance scores indicate how significant each feature was in the dataset when building the predictive model with larger value means larger effect the features had on the model in predicting *hardest-* from *easiest-to-interpret* normal cases of RTs (Figure 1).

2.6 Statistical analysis and validation

In order to examine the performance of the model on our data and to maximise the use of the available data for both training and validation, we trained and validated the model using the resampling leave-one-out-cross-validation (LOOCV) approach (an unbiased, dependable, and accurate validation method for assessing a machine learning model' generalisation performance) [36]. Each time when training the model, all cases were used, except one case which was left out and used once as a test set to validate the model's predictive performance. We repeated this process 140 times (as per the total number of the normal cases we had) until each case was left out and used once as a test set. The model's performance was then evaluated based on how accurately it predicted the left-out/unseen case in each repetition, providing effectively estimation of how well the model generalises to the unseen cases. The overall performance of the model for discriminating *hardest-* from *easiest-to-interpret* normal cases of RTs was assessed using the AUC.

A Kruskal-Wallis test was employed to investigate if the 34 GMRFs differed between *hardest-* and *easiest-to-interpret* normal cases of RTs, and if difficulty level differed among *low-* vs *high-density* normal cases of RTs. A *p*-value of less than 0.05 was considered statistically significant.

Moreover, a scree test of exploratory factor analysis [37] was used to determine the usefulness of the GMRFs based the sum of their importance scores from the model.

Radiomics and statistical analysis were performed using MATLAB R2022a (MathWorks, Natick, Massachusetts, USA).

3. Results

3.1 Global mammographic radiomic features predicting difficult normal cases of radiology trainees

The overall performance of the machine learning classification model for differentiating *hardest-* from *easiest-to-interpret* normal cases of RTs using GMRFs is shown in Figure 2, which indicates an AUC of 0.75 (95% confidence interval: 0.67-0.83).

3.2 Significant global mammographic radiomic features and breast density of difficult normal cases

Table 2 shows that 15 out of the total 34 GMRFs had a statistically significant difference between *hardest-* vs *easiest-to-interpret* normal cases of RTs with $p < 0.05$.

Table 2 and Figure 3 show the significant GMRFs based on their importance scores derived from the predictive machine learning model. Out of the 34 features, when comparing each feature's total importance scores ranking from the highest to the lowest, two important features were recognized. These features were *SQ_Cluster_prominence_9* (feature# 31) of GLCM, and *ROI_Std_Range_all* (feature# 16) of FOS. *Cluster prominence* measures the skewness and asymmetry of the GLCM [33]. *Range* describes the difference between maximum and minimum of image gray level values [25].

Figure 4 revealed that the difficulty of fatty or *low-density* normal cases does not significantly differ from that of dense or *high-density* cases (*p*-value = 0.12), with *low-density* cases appearing to be more challenging for RTs, albeit non-significant.

4. Discussion

In this proof-of-concept study, we explored 1) if using a set of global handcrafted radiomic or computer-extracted image features derived from mammograms, *hardest-* vs *easiest-to-interpret* normal cancer-free cases of RTs can be classified, and 2) whether any of these features are useful in this classification task.

Previous works [4, 5] recognized the usefulness of local radiomic features in identifying missed mammographic masses and false positive location errors of RTs, obtaining up to 0.61 AUC. Other later studies [11, 16] indicated the value of GMRFs in predicting most *difficult-to-interpret* normal cases of general and experienced radiologists (over 20 years of experience reading mammograms), attaining up to 0.73 AUC. Building on these recent research outputs [11, 16], we explored feasibility of using GMRFs in describing RTs' *hardest-* vs *easiest-to-interpret* normal cases based on the hypothesis that the image characteristics of these two categories' cases are diverse. In addition, the results from our machine learning classifier revealed similar findings in that GMRFs had the ability to distinguish between *hardest-* from *easiest-to-interpret* normal cases of RTs, achieving an AUC of 0.75 (95% CI: 0.67-0.83) - Figure 2. This finding affirms the importance of GMRFs in predicting *hardest-to-interpret* normal cases for less experience readers like RTs (Table 1).

Mammographic breast density using BI-RADS category are often employed as a surrogate measure in assessing difficult normal cases for RTs, where cases with higher density are regularly considered to be more difficult to interpret [17, 38]. However, studies have shown that breast density category details alone may not be completely reliable for predicting difficult/easy normal cases of radiologists including RTs [11, 39]. In line with these findings, the results from our dataset confirmed that BI-RADS density categories were not strongly correlated with difficulty scores - Figure 4. Our GMRFs relating to *hardest-to-interpret* normal cases of RTs therefore may be used to provide supplementary clinical image information (in addition to BI-RADS breast density class) about most difficult features that RTs are likely to commit a false positive error. Also, since RTs find recognising normal image features challenging [13, 14], our findings could therefore aid their screening mammographic interpretation expertise development in recognising hardest and easiest breast features of normal cases which could then facilitate automatic hardest and easiest normal cases test set curation for customized education/training programs. Using the proposed model, we can create customized test sets tailored to the specific learning needs of trainees at different stages of their careers. For example, early-stage trainees could be presented with easier cases to build foundational skills, while more advanced trainees could be challenged with harder cases that reflect real-world complexities. This customized test set curation would allow for targeted education and training programs that

align with the trainees' current skill levels, helping them to progress more effectively through their learning journey.

Interestingly, when comparing two groups based on a single GMRF, 15 out of 34 features showed a statistically significant difference between *hardest-* and *easiest-to-interpret* normal cases of RTs ($p < 0.05$). However, when examining the 34 features together through our machine learning classifier, only two features (*cluster prominence* and *range*) were found to be most useful in describing RTs' *hardest-* from *easiest-to-interpret* normal cases (Table 2 and Figure 3). *Range*, in particular, has also been discovered to be a valuable feature in previous studies [11, 20] highlighting the significance of *range* in various classification tasks among different groups of readers (e.g., categorising *hardest-* from *easiest-to-interpret* normal cases of readers in general as well as RTs). While *cluster prominence* was not found to be an important feature in the previous work [11, 20], in this dataset, we observed that more difficult cases, which resulted in false positive errors more frequently, exhibited higher *cluster prominence* which is a measure of the skewness and asymmetry of the GLCM. This suggests that *cluster prominence* may be an important marker of highly challenging cases for RTs. A higher value of *cluster prominence* implies more asymmetry about the mean in the GLCM, while a lower value indicates a peak near the mean value and less variation about the mean. Intuitively, *cluster prominence* refers to the extent to which clusters of pixels within an image stand out from their surroundings. In other words, it measures the distinctiveness of clusters within the image with a higher *cluster prominence* value indicates that clusters are more salient or prominent within the image. In the context of our study, normal cases with higher *cluster prominence* value pose greater difficulty for trainees. Trainees may struggle with these cases because the heightened prominence of clusters within the image makes it more challenging to distinguish between normal tissue and potential abnormalities, resulting in false positive annotations. Although the features may be abstract in nature and not directly visible to humans, they could be quantitatively measured using our radiomics approach and used as educational tools to teach and/or alert RTs that the higher the value of *cluster prominence*, the harder the case could be for RTs, increasing the chance of them making a false positive error. Furthermore, creating an alert system based on these features and educating RTs about these correlations can improve their ability to recognize difficult normal cases and enhance decision-making in clinical practice. Nonetheless, more studies are needed to further examine the effectiveness and reliability of these features.

This study also has a few limitations that should be recognised. It was a proof-of-concept study examining the efficacy of 34 handcrafted GMRFs [11] in identifying the *hardest-* from *easiest-to-interpret* confirmed cancer-free normal cases of 137 RTs (mostly with less than 1 year of experience reading mammograms) specifically. Only normal cases were included in the study which contained only those that had difficulty scores of greater than or equal to the 75th percentile (*hardest-to-interpret*) and less than or equal to the 25th percentile (*easiest-to-interpret*). LOOCV approach [36] was used to train and validate the model in order to maximise the use of the available data. The GMRFs were extracted from the region of interest delineated using the lattice-based [30] and squared-based [31] approaches. Future research should explore other approaches of identifying region of interest (e.g., retroareolar region), validating the model (e.g., holdout method [36]), employing other difficult normal cases with difficulty scores falling outside the included range (e.g., scores that are less than the 75th percentile and greater than 25th percentile), and radiomic analysis involving local and/or deep learning features [26, 40].

5. Conclusions

In conclusion, in addition to previous research [11, 20], our work highlights the importance of GMRFs for classifying *hardest-* from *easiest-to-interpret* normal cases of RTs with two important features (*cluster prominence* and *range*) being identified. Our findings could be useful for developing RTs' interpretation expertise in identifying most difficult image features of normal breast cases as well as enabling automatic and tailored test set curation of the most difficult cases for RTs' education/training programs.

6. Abbreviations

AUC	Area under receiver operating characteristic curves
BC	Breast cancer
BI-RADS	Breast imaging and reporting data system
BREAST	Breastscreen REader Assessment STrategy
CC	Craniocaudal
FOS	First-order statistics
GLCM	Gray level co-occurrence matrix
GMRFs	Global mammographic radiomic features
MLO	Mediolateral oblique
NGTDM	Neighbourhood gray tone difference matrix
RANZCR	Royal Australian and New Zealand College of Radiologists
ROI	Multiple regions of interests defined by the lattice-based approach covering the entire breast image
RTs	Radiology trainees
SQ	Largest square rectangular box inscribed within breast

7. Acknowledgements

The authors would like to express appreciation to the BreastScreen Reader Assessment Strategy (BREAST) program at the University of Sydney for access and use of the data. BREAST is funded/supported by the Commonwealth Government Department of Health and Aged Care and the National Breast Cancer Foundation (NBCF), with their images provided by BreastScreen Australia and the Cancer Institute New South Wales (CINSW). The authors also would like to acknowledge the funding support for this research provided by the National Health and Medical Research Council [grant ID: APP1162872] and NBCF [grant ID: IIRS-18-089]. Z.G. is supported by the CINSW Early Career Fellowship [grant ID: 2022/ECF1426]. The authors wish to thank Dr Robert Heard for his statistical advice and the reviewers for their rigorous assessment of this manuscript.

8. Table and Figure Legends

- **Table 1. Characteristics of readers (n=137)**
- **Table 2. Extracted global mammographic radiomic features (n=34)**
- **Figure 1. Radiomics workflow. 1).** Input mammographic images and masks were acquired. **2).** Region of interest (yellow colour region) was delineated using lattice- and squared-based approaches. **3).** A set of 34 global mammographic radiomic features (GMRFs)/scores per image were extracted from the yellow region of interest, then standardised to have image intensity mean equal to zero and standard deviation equal to one. The extracted GMRFs of four images belonging to each of the 140 cases were then averaged to obtain one GMRF per case **4).** Finally, using the averaged GMRFs, a random forest machine learning model for differentiating radiology trainees' *hardest-* from *easiest-to-interpret* normal cases was constructed and assessed.
- **Figure 2. Model performance.** For discriminating *hardest-* from *easiest-to-interpret* normal cases of radiology trainees, 0.75 AUC (95% confidence interval: 0.67 - 0.83) was achieved by the random forest machine learning model.
- **Figure 3. Scree plot showing significant global mammographic radiomic features.** From 34 features, two important features were identified, indicating in the scree plot with the steep slope occurred after the second highest features (as shown by the red arrow). Those important features, ranked from the highest to the lowest important scores, were: 1). *SQ_Cluster_prominence_9* (feature# 31) of GLCM, and 2). *ROI_Std_Range_all* (feature# 16) of FOS. FOS = first order statistics, GLCM = gray level co-occurrence matrix, ROI = multiple regions of interests defined by the lattice-based approach covering the entire breast image, SQ = largest square rectangular box inscribed within breast.
- **Figure 4. Breast density comparison.** Non statistically significant difference ($p = 0.12$) in the difficulty level was found between *low-density* and *high-density* normal cases of radiology trainees.

9. References

1. Sung H, et al.: Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians 71:209-249, 2021
2. Paci E: Summary of the Evidence of Breast Cancer Service Screening Outcomes in Europe and First Estimate of the Benefit and Harm Balance Sheet. Journal of Medical Screening 19:5-13, 2012
3. Ekpo EU, Alakhras M, Brennan P: Errors in Mammography Cannot be Solved Through Technology Alone. Asian Pac J Cancer Prev 19:291-301, 2018
4. Zhang J, Lo JY, Kuzmiak CM, Ghate SV, Yoon SC, Mazurowski MA: Using computer-extracted image features for modeling of error-making patterns in detection of mammographic masses among radiology residents. Medical Physics 41:091907, 2014
5. Zhang J, Silber JI, Mazurowski MA: Modeling false positive error making patterns in radiology trainees for improved mammography education. Journal of Biomedical Informatics 54:50-57, 2015
6. Wong DJ, et al.: Do reader characteristics affect diagnostic efficacy in screening mammography? A systematic review. Clinical Breast Cancer, 2023
7. Gandomkar Z, Lewis SJ, Li T, Ekpo EU, Brennan PC: A machine learning model based on readers' characteristics to predict their performances in reading screening mammograms. Breast Cancer 29:589-598, 2022
8. Hofvind S, et al.: Audit feedback on reading performance of screening mammograms: An international comparison. Journal of Medical Screening 23:150-159, 2016
9. Suleiman ME, Rickard M, Brennan PC: Perfecting detection through education. Radiography 26:S49-S53, 2020
10. Lee W, Peters G: Mammographic screening for breast cancer: A review. Journal of Medical Radiation Sciences 60:35-39, 2013
11. Siviengphanom S, Gandomkar Z, Lewis SJ, Brennan PC: Global radiomic features from mammography for predicting difficult-to-interpret normal cases. Journal of Digital Imaging 36:1541–1552, 2023
12. Vlahiotis A, Griffin B, Stavros Md FAT, Margolis J: Analysis of utilization patterns and associated costs of the breast imaging and diagnostic procedures after screening mammography. ClinicoEconomics and Outcomes Research Volume 10:157-167, 2018
13. Gandomkar Z, Mello-Thoms C: Visual search in breast imaging. Br J Radiol 92:20190057, 2019
14. Nodine CF, et al.: How experience and training influence mammography expertise. Acad Radiol 6:575-585, 1999
15. Buist DS, et al.: Influence of annual interpretive volume on screening mammography performance in the United States. Radiology 259:72-84, 2011
16. Siviengphanom S, Gandomkar Z, Lewis SJ, Brennan PC: Global mammographic radiomic signature can predict radiologists' difficult-to-interpret normal cases. Proc. SPIE Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment: City
17. Ang ZZ, Rawashdeh MA, Heard R, Brennan PC, Lee W, Lewis SJ: Classification of normal screening mammograms is strongly influenced by perceived mammographic breast density. Journal of Medical Imaging and Radiation Oncology 61:461-469, 2017
18. Darker I, Chen Y, Gale A: Health professionals' agreement on density judgements and successful abnormality identification within the UK Breast Screening Programme. Proc. SPIE Medical Imaging: City
19. Gandomkar Z, et al.: Reliability of radiologists' first impression when interpreting a screening mammogram. PLOS ONE 18:1-19, 2023
20. Siviengphanom S, Lewis SJ, Brennan PC, Gandomkar Z: Computer-extracted global radiomic features can predict the radiologists' first impression about the abnormality of a screening mammogram. British Journal of Radiology 97:168–179, 2024
21. Brennan PC, et al.: Radiologists can detect the 'gist' of breast cancer before any overt signs of cancer appear. Sci Rep 8:1-12, 2018
22. Gandomkar Z, et al.: Global processing provides malignancy evidence complementary to the information captured by humans or machines following detailed mammogram inspection. Sci Rep 11:1-12, 2021
23. Siviengphanom S, Lewis SJ, Brennan PC, Gandomkar Z: Predicting the gist of breast cancer on a screening mammogram using global radiomic features. Proc. SPIE Medical Imaging 2024: Image Perception, Observer Performance, and Technology Assessment: City
24. Evans KK, Georgian-Smith D, Tambouret R, Birdwell RL, Wolfe JM: The gist of the abnormal: Above-chance medical decision making in the blink of an eye. Psychonomic Bulletin & Review 20:1170-1175, 2013
25. Gillies RJ, Kinahan PE, Hricak H: Radiomics: Images Are More than Pictures, They Are Data. Radiology 278:563-577, 2016

26. Siviengphanom S, Gandomkar Z, Lewis SJ, Brennan PC: Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs. Academic Radiology 29:1228-1247, 2021
27. Gandomkar Z, et al.: An end-to-end deep learning model can detect the gist of the abnormal in prior mammograms as perceived by experienced radiologists. Proc. SPIE Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment: City
28. Brennan P, Lee W, Tapia K: Breast Screen Reader Assessment Strategy (BREAST): A Research Infrastructure with a Translational Objective, Cambridge: Cambridge University Press, 2018
29. McEntee M, Brennan P, Evanoff M, Phillps P, O Connor W, Manning D: Optimum ambient lighting conditions for the viewing of softcopy radiological images: SPIE, 2006
30. Zheng Y, et al.: Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment. Medical Physics 42:4149-4160, 2015
31. Gandomkar Z, Suleiman M, Demchig D, Brennan P, McEntee M: BI-RADS density categorization using deep neural networks. Proc. SPIE 10952, Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment: City
32. Haga A, et al.: Standardization of imaging features for radiomics analysis. The Journal of Medical Investigation 66:35-37, 2019
33. Haralick RM, Shanmugam K, Dinstein IH: Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics SMC-3:610-621, 1973
34. Amadasun M, King R: Textural features corresponding to textural properties. IEEE Transactions on Systems, Man, and Cybernetics 19:1264-1274, 1989
35. Breiman L: Random Forests. Machine Learning 45:5-32, 2001
36. Molinaro AM, Simon R, Pfeiffer RM: Prediction error estimation: a comparison of resampling methods. Bioinformatics 21:3301-3307, 2005
37. Ledesma RD, Valero-Mora P, Macbeth G: The Scree Test and the Number of Factors: a Dynamic Graphics Approach. The Spanish Journal of Psychology 18, 2015
38. A Tapia K, Rickard MT, McEntee MF, Garvey G, Lydiard L, C Brennan P: Impact of breast density on cancer detection: observations from digital mammography test sets. International Journal of Radiology & Radiation Therapy 7:36-41, 2020
39. Grimm LJ, Kuzmiak CM, Ghate SV, Yoon SC, Mazurowski MA: Radiology Resident Mammography Training. Academic Radiology 21:888-892, 2014
40. Pertuz S, Torres GF, Tamimi R, Kamarainen J: Open Framework for Mammography-based Breast Cancer Risk Assessment. Proc. 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI): City

Table 1. Characteristics of readers (n=137)		
Mean years of age (derived from 116 radiology trainees' age details while 21 radiology trainees did not provide their age details)	33	
	N	**%**
Country		
Australia	135	99
New Zealand	2	1
Location at the time of engaging with the BREAST program		
At a workshop or conference	26	19
At a clinic	111	81
# of years reading mammograms		
Less than 1	81	59
1 to 5	53	39
6 to 12	3	2
# of mammographic cases reading per week		
Less than 20	108	79
20 to 200	28	20
More than 200	1	1
# of hours per week spent reading mammograms		
Less than 4	131	96
4 to 30	5	4
More than 30	1	1
# of readers completed a fellowship lasted for 3 to 6 months		
Completed fellowship	16	12
Did not complete fellowship	119	87
Not specified	2	1
# of readers read for breast screening program		
Yes	16	12
No	121	88

Table 2: Extracted global mammographic radiomic features (n=34)

Feature no.	Feature name	*p*-value	Feature importance score	Feature category	Parameters
1	ROI_Std_Correlation_3	0.207	0.229	GLCM	3-pixel distance between the pixel of interest and its neighbour
2	ROI_Std_Difference_entropy_3	0.775	0.104		
3	ROI_Std_Dissimilarity_3	0.218	0.134		
4	ROI_Std_Difference_variance_3	0.688	0.166		
5	ROI_Std_Energy_3	0.475	0.038		
6	ROI_Std_Entropy_3	0.621	0.112		
7	ROI_Std_Homogeneity_3	0.209	0.238		
8	ROI_Std_Information_measure_of_correlation1_3	0.654	0.129		
9	ROI_Std_Information_measure_of_correlation2_3	0.663	0.234		
10	ROI_Std_Maximum_probability_3	0.462	0.217		
11	ROI_Std_Sum_entropy_3	0.086	0.410		
12	ROI_Std_Sum_variance_3	0.558	0.039		
13	ROI_Std_Coarseness_3	0.869	0.409	NGTDM	3x3 neighbourhood size
14	ROI_Std_Cluster_prominence_3	0.756	0.493	GLCM	3-pixel distance between the pixel of interest and its neighbour
15	ROI_Std_Cluster_shade_3	0.905	0.353		
16	ROI_Std_Range_all	0.003^	0.759*	FOS	n/a
17	ROI_Std_Sum_of_squares_variance_9	0.555	0.225	GLCM	9-pixel distance between the pixel of interest and its neighbour
18	SQ_C orrelation_9	0.028^	0.132		
19	SQ_Difference_entropy_9	0.000^	0.083		
20	SQ_Dissimilarity_9	0.000^	0.290		
21	SQ_Difference_variance_9	0.000^	0.128		
22	SQ_Energy_9	0.000^	0.171		
23	SQ_Entropy_9	0.000^	0.098		
24	SQ_Homogeneity_9	0.000^	0.049		
25	SQ_Information_measure_of_correlation1_9	0.001^	0.270		
26	SQ_Information_measure_of_correlation2_9	0.000^	0.284		
27	SQ_Maximum_probability_9	0.000^	0.414		
28	SQ_Sum_entropy_9	0.901	0.458		
29	SQ_Sum_variance_9	0.000^	0.296		
30	SQ_Coarseness_9	0.002^	0.333	NGTDM	9x9 neighbourhood size
31	SQ_Cluster_prominence_9	0.325	0.778*	GLCM	9-pixel distance between the pixel of interest and its neighbour
32	SQ_Cluster_shade_9	0.068	0.124		
33	SQ_Range_all	0.002^	0.479	FOS	n/a
34	SQ_Sum_of_squares_variance_9	0.000^	0.059	GLCM	9-pixel distance between the pixel of interest and its neighbour

^features (n=15) showing statistically significant difference (*p*-value < 0.05) between *hardest-* and *easiest-to-interpret* normal cases of radiology trainees.
* and highlighted in grey = important features (n=2) from the machine learning classifier. The higher importance score means the more useful the features were for the classifier.

Abbreviations:
FOS = first order statistics, GLCM = gray level co-occurrence matrix, NGTDM = neighbourhood gray tone difference matrix, n/a = not applicable, ROI = multiple regions of interests defined by the lattice-based approach covering the entire breast image, SQ = largest square rectangular box inscribed within breast, Std = standard deviation.

Figure 1

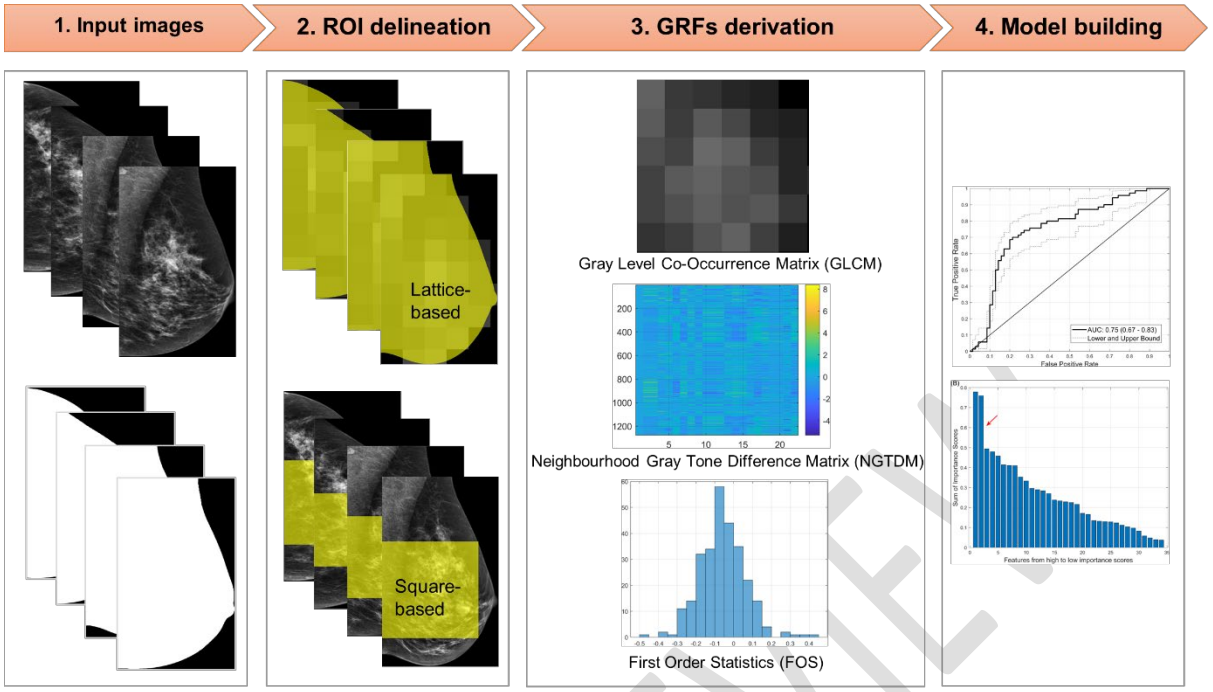


Figure 2

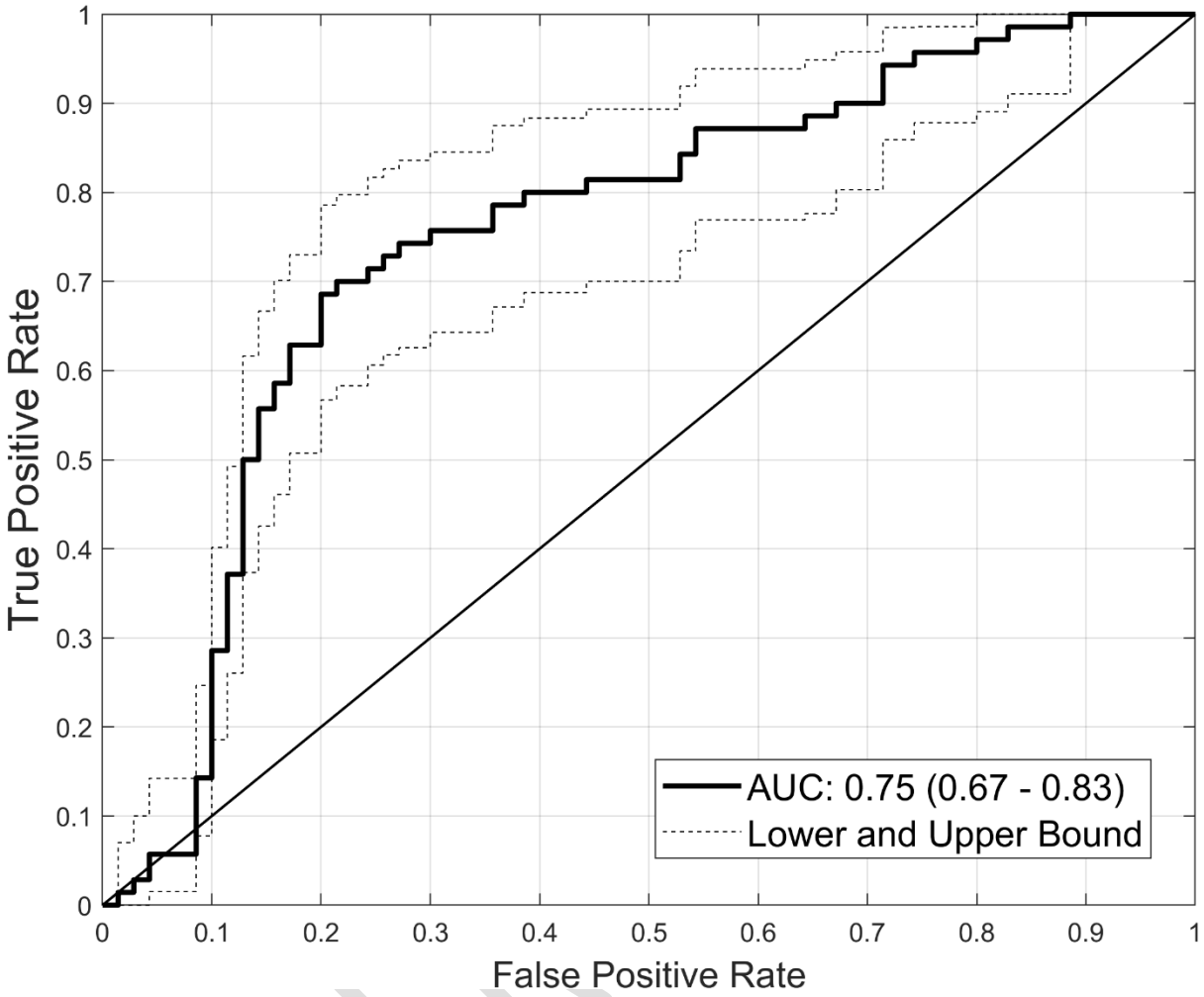


Figure 3

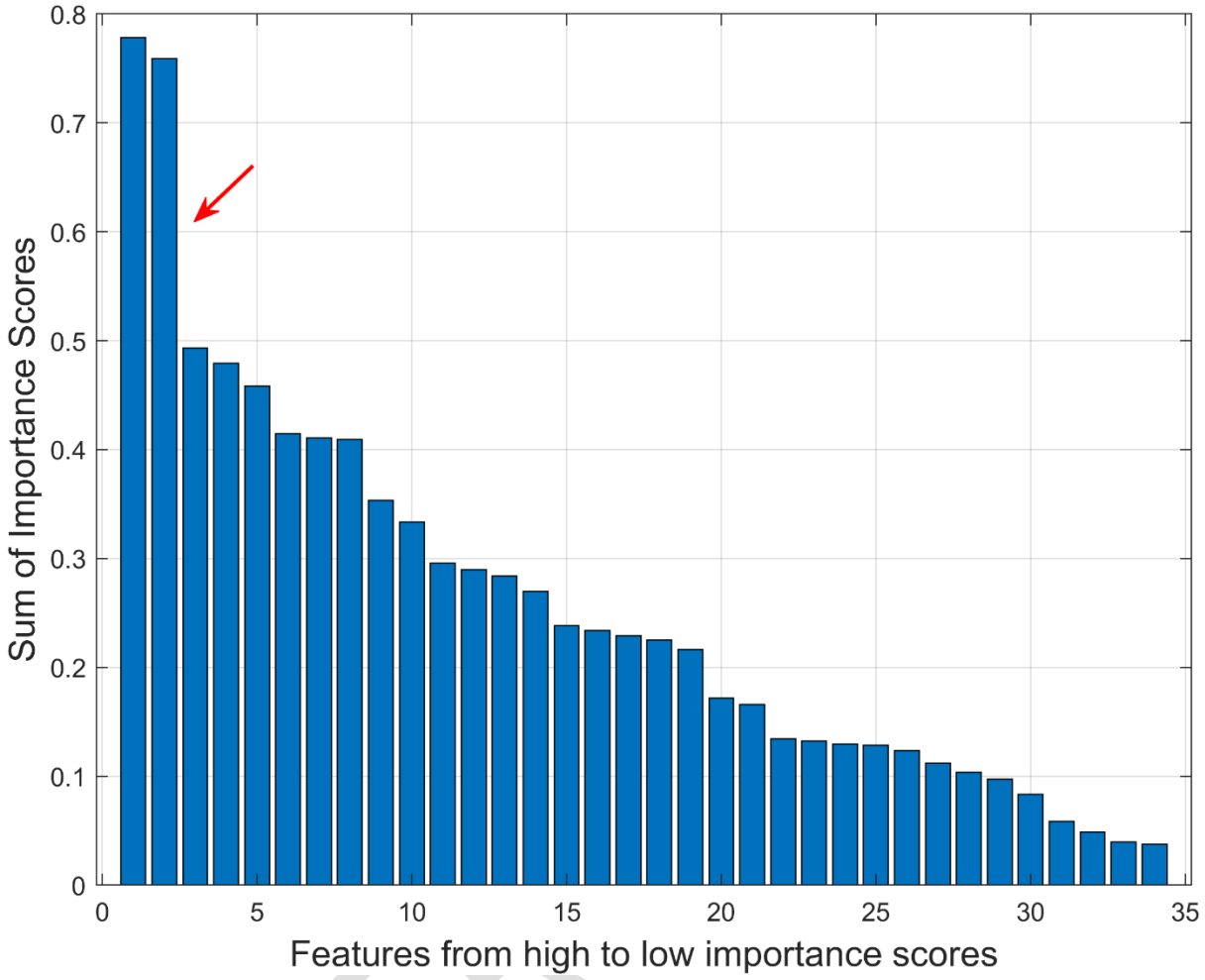
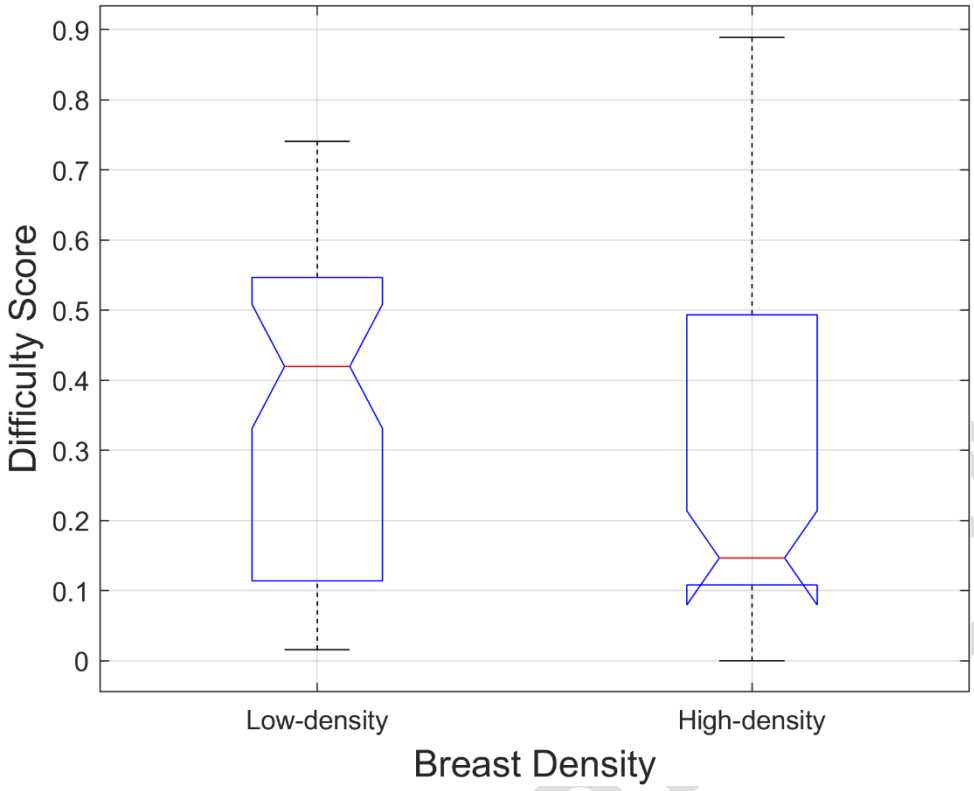


Figure 4



5.9 Chapter findings

The results from this Chapter 5 underlined the efficacy of global mammographic radiomic features in distinguishing between *hardest-* and *easiest-to-interpret* normal cases of radiology trainees, achieving an AUC of 0.75 with *cluster prominence* and *range* recognised as helpful features. In addition, no statistically significant relationship between breast density and difficulty scores of this cohort's cases was observed.

Chapter 6 – Discussion and conclusion

6.1 Chapter introduction and overview of the thesis

For recent decades, educational and self-assessment training programs have been widely utilised globally for monitoring and enhancing clinical radiologists' (including radiology trainees) skills and performances in interpreting breast screening images [1-7]. However, the current process of curating difficult normal cases for educational mammographic test sets still relies heavily on the expert radiologists' manual and subjective assessment using the classic breast density BI-RADS category and benign features [8-16]. This standard method appears to be inconsistent and unreliable as it fails to provide comprehensive descriptors of the actual normal case difficulty and false positive errors of the individual based on their performance in the test sets, leading to inadequate achievement of radiologists' training requirements [11, 13, 14, 17-19]. It is thus extremely necessary to understand what makes a normal case difficult and causes false positive errors to occur for these readers, not only for improving the curation of the educational test sets but also enhancing and expediting their skills in recognising normality features.

Additionally, extensive evidence highlights the significance of radiomics analysis (computer-extracted image features) in mammography and global gist signal (radiologists' holistic quick impression about a mammogram's abnormality) in guiding visual searches and informing diagnostic decisions when reading difficult cancer cases [20-23]. Global gist signal has also been shown to be important for mammogram readers' expertise development [24] and lacking the ability to capture this signal or in some cases where this signal is poorly perceived could result in diagnostic errors [25, 26]. Albeit global gist signal appears to have poor intra- and inter-observer reliability, its value in mammographic interpretation is considerable [24-33]. Therefore, with the value of radiomics in describing mammographic characteristics [34-38], there is a strong need for a reliable predictor for global gist signal using radiomics technique.

For enhancing training in radiology, a small number of studies were conducted using computer algorithms to model and predict error making patterns of readers [18, 39-45] but they predominantly concentrate on false negative errors and false positive errors or difficult normal cases have inadequate investigation. A few AI tools including using radiomics analysis were also proposed to predict error making patterns (such as missed mammographic masses and false positive location errors) of readers and their difficult cases (e.g., biopsy-proven masses) [40, 43, 46-49]. These tools, nonetheless, lack focus on more difficult normal cases to

complement readers' mammographic interpretation skills, alert readers about the difficult cases for suitable reading strategy such as double vs single reading and/or senior vs junior reader, or improve mammographic educational test sets especially [12, 50-52]. Studies that dedicated to mammographic radiomics method [40, 43, 49] also only focus on radiology trainees using a small dataset (e.g., up to 10 trainees and 3 experts, and 100 mammographic cases) and only employing local radiomic features while global radiomics features have been insufficiently examined [20, 53, 54].

To date there appears to be no study conducted in examining the efficacy of mammographic features derived from a novel global radiomic analysis for improving the process of generating educational mammographic test sets. Hence, for the first time, this thesis aims to address this knowledge gap by **firstly** exploring if global mammographic radiomic features can predict the global gist signal, and if so, which of these features best predict the signals. **Secondly**, it seeks to evaluate whether these global radiomic features from mammograms (driving radiologists' initial impression about the case) can distinguish highly difficult from easy normal cases of radiologists with diverse levels of experience, considering almost 70% of cases in the test sets are normal [4, 5, 55]. It also attempts to recognise the most valuable features in depicting these cases as well as assessing the link between BI-RADs breast density and difficulty scores of these normal cases since BI-RADS breast density has been claimed to be the predictor of difficult cases when generating educational test sets [11, 18, 19]. **Lastly**, the final aim of the thesis is to investigate further whether, using global mammographic radiomic features radiology trainees' more difficult normal cases can be separated from the easy ones since this cohort has insufficient experience in processing global images and interpreting mammograms [28, 29, 31, 33, 40, 43]. Plus, it tries to reveal whether any of these features are best in distinguishing the cases, and if BI-RADS breast density categories have any associations with the cases in this cohort.

The thesis' objectives were attained via the **three original research studies,** presented in Chapter 3, 4 and 5. Prior to these original research studies, the literature review (Chapter 2) was conducted to identify knowledge gap in the literature guiding curation of radiomic features and informing the subsequent three original research investigations.

The first study (Chapter 3) recruited 13 "gist expert" radiologists (exceptional capability to detect reliable and accurate gist of the abnormal) [25] to assess 4191 BreastScreen Australia screening craniocaudal mammograms and provide global gist signals/scores (i.e., probability

of image abnormality). The collected gist scores were averaged to obtain one gist score per image. Images were categorized into *high-* and *low-gist* images based on the respective 75th and 25th percentile of images having the highest and lowest gist scores. A set of 130 handcrafted global radiomic features were extracted and used to build machine learning classifiers to discriminate *high-* from *low-gist* images.

The second study (Chapter 4) obtained 120 *difficult-* and 119 *easy-to-interpret* normal cases based on one-third of cases having the respective highest and lowest difficulty scores (percentage of false positives) provided by 537 BREAST readers of widely varying experience, interpreting 361 cases. Cases in these two categories also yielded 115 *low-BD* (BI-RADS category A = fatty and B = fibroglandular) and 124 *high-BD* (C = heterogeneous and D = extremely dense). A set of 34 global radiomic features per image were extracted and used to build machine learning models for separating *difficult-* from *easy-to-interpret* normal cases of radiologists.

The third study (Chapter 5) acquired 70 *hardest-* and 70 *easiest-to-interpret* normal cases based on the 75th and 25th percentiles of the cases containing the corresponding highest and lowest difficulty scores provided by 137 BREAST radiology trainees of limited mammogram reading experience, interpreting 280 normal cases. 59 of these cases were recognised as *low-density* (BI-RADS category A = fatty and B = fibroglandular) and 81 were *high-density* (C = heterogeneous and D = extremely dense). The 34 global radiomic features of images (from the second study [14]) belonging to these two categories cases were used and to obtain one global radiomic feature per case, four features of images belonging to that case were averaged. These averaged features were then employed to build a machine learning classifier for discriminating between *hardest-* and *easiest-to-interpret* normal cases of radiology trainees.

6.2 Key findings of the thesis

From the literature review, it has been found that limited studies had been conducted using global radiomic features derived from mammograms to predict global gist signal and difficult normal cases of radiologists and radiology trainees. Radiomic features such as GLCM and FOS were commonly used for describing breast cancer's characteristics, recurrence, and survival. These findings had guided the curation of global radiomic features set for and the investigations of the subsequent three original studies with some key findings as followed (Figure 6.2).

The **first key finding** of this thesis (Chapter 3) was that a set of global radiomic features derived from screening mammograms demonstrated the ability to accurately predict *high-* (e.g., normal images containing a strong gist of the abnormal/false positive) from *low-gist* images (e.g., missed cancer cases having a weak gist/false negative) with AUCs of up to 0.84. From the eight machine learning models (*All*, *Normal*, *Cancer*, *Prior-1*, *Prior-2*, *Missed*, *Prior-Visible*, and *Prior-Invisible*), five most important features (*cluster shade* of texture GLCM*, standard deviation, skewness, kurtosis, and range* of FOS) were identified based on their importance scores. In particular, two features, appearing to be the most valuable in predicting *high-* vs *low-gist* images in almost all the models, were *cluster shade* (measuring the symmetry distribution in an image) and *range* (describing the difference between 95th and 5th percentile of image gray level values). Consistent with previous basic vision science studies [56-61], this finding affirmed the importance of texture GLCM and FOS features from mammograms in identifying *high-* vs *low-gist* images while offering unique insights into the methodology advancement of extracting image features using global radiomics analysis.

The **second noteworthy discovery** in this thesis (Chapter 4) mirrored the first (Chapter 3), in that global mammographic radiomic features (directing radiologists' initial impression about the case) were able to discriminate highly *difficult-* from *easy-to-interpret* normal cases of clinician radiologists. Six features (*sum entropy*, *difference variance*, *correlation*, and *homogeneity* of GLCM, *range* of FOS, and *coarseness* of NGTDM) were the most helpful in describing *difficult-to-interpret* normal cases. Two most useful features, more specifically, were *correlation* (measuring similarity of gray level values between the two neighbouring pixels) and *range* (describing the difference between maximum and minimum of image gray level values). Although previous mammographic studies [40, 43, 49] focused on local radiomic features of specific tumour or false positive areas to predict error making patterns of radiology trainees (AUCs up to 0.63), the reported findings in the thesis emphasised the significance of the novel global radiomic analysis in describing *difficult-to-interpret* normal cases of radiologists in general (AUCs of up to 0.71). Furthermore, the findings revealed no evidence of BI-RADS breast density classification being the complete/sole predictor of difficult cases for radiologists, similar to previous studies [11, 18, 19], when curating educational mammographic test sets.

Correspondingly, the **final significant findings** based on Chapter 5 of the thesis confirmed further that radiomics-driven global mammographic features had the capability to distinguish

radiology trainees' *hardest-* from *easiest-to-interpret* normal cases (0.75 AUC). Prior work [40, 43, 49] found the value of local mammographic radiomic features in describing error making patterns of radiology trainees (AUCs up to 0.63), but they only used a small number of readers (up to 10 trainees) and cases (up to 100) and specific important features from their models were not studied. The results from this work (using larger dataset, 137 trainees and 140 cases) provided important discovery of the two best global radiomic features in explaining radiology trainees' *hardest-* vs *easiest-to-interpret* normal cases. These features were *cluster prominence* (measuring the skewness and asymmetry of the GLCM) of GLCM and *range* (describing the difference between maximum and minimum of image gray level values) of FOS. Interestingly, *range* has also been found to be an important feature for predicting difficult normal cases of general radiologists in the second study (chapter 4) [14], but *cluster prominence* has not. This indicates a novel discovery of *cluster prominence* feature as an important indicator of highly challenging normal cases for naïve radiology trainees. Moreover, the study's findings also aligned with the previous research [11, 14, 18, 19] verifying that BI-RADS breast density categories were not significantly related to the difficulty radiology trainees experienced in the test sets.

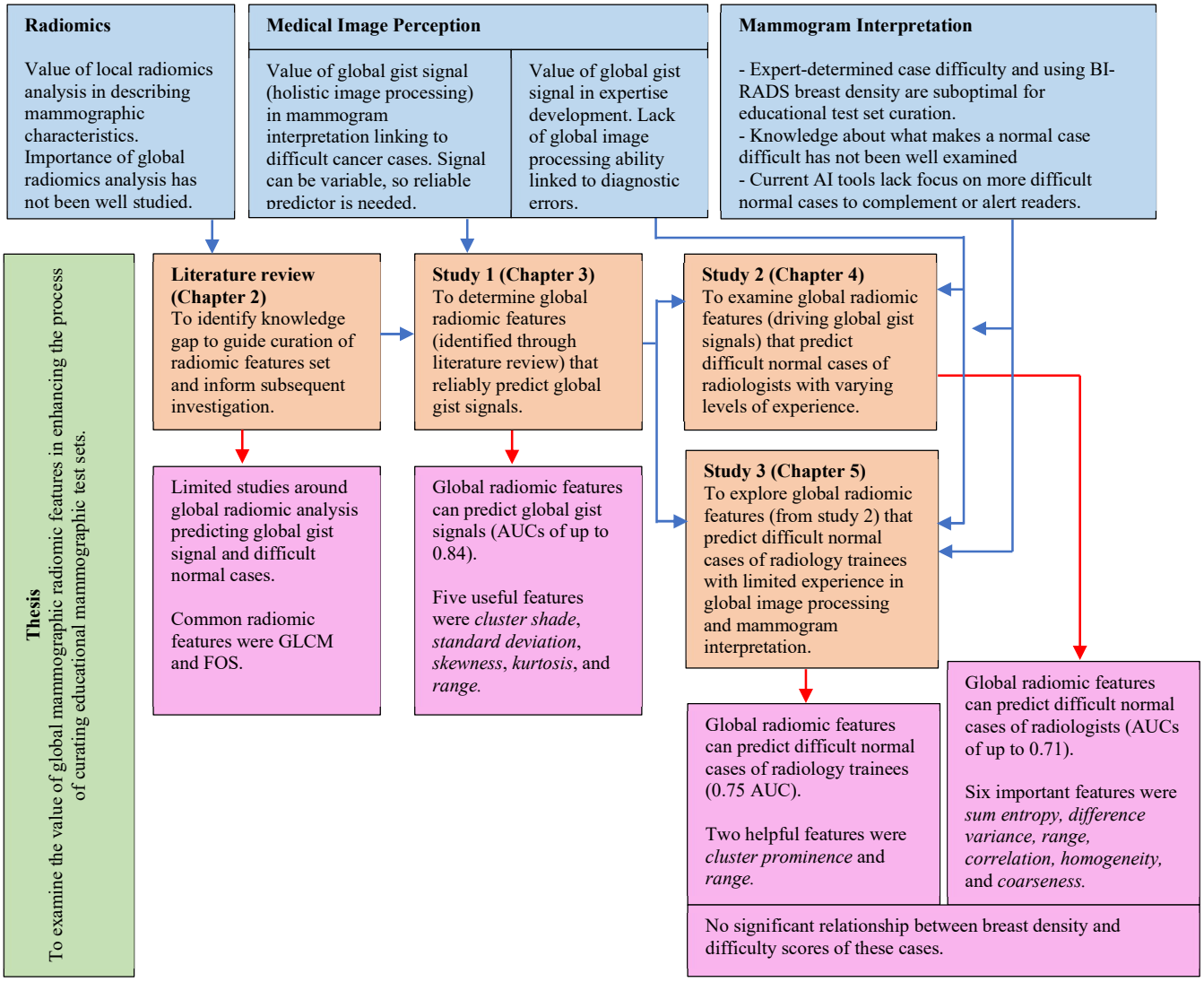


Figure 6.2. Summary of thesis' key findings. To achieve the overall aim of the thesis (green box), based on the existing knowledge and gaps (blue), literature review and three original studies were conducted (orange) with some key results indicated in the pink boxes. AUC = area under receiver operating characteristic curve, BI-RADS = Breast Imaging and Reporting Data System, FOS = First Order Statistics, GLCM = Gray Level Co-Occurrence Matrix.

6.3 Clinical implications

The novel and valuable methods of global radiomic analysis and innovative discovery of specific useful features have provided great contributions to the research field of radiomics analysis, medical image perception, radiology education, and screening mammography. Precisely, the **first findings** (Chapter 3) established the correlation between global radiomic features and global gist signals, suggesting that a set of features (describing the whole image using the novel global radiomic technique) inherent within mammographic images could

provide better understanding about the process behind medical image perception. The global characteristics of image abnormality perceived by radiologists can be recognized and described through the novel method of global radiomic analysis. The discovery about the five useful features (*cluster shade* of GLCM*, standard deviation, skewness, kurtosis, and range* of FOS), particularly *cluster shade* and *range*, has generated new important details about the possible computer-extracted mammogram features that direct the strength of the global gist signal. These findings are crucial for advancing radiomics technique and could be integrated into radiology education and breast screening programs regarding the possible features of false positives (e.g., normal images holding strong gist of the abnormal) and false negatives (e.g., missed cancer/cancer images having weak gist of the abnormal). Moreover, considering the global gist signal has been shown to be effective in enhancing the performance of the modern deep learning model for detecting breast cancer in mammography screening, this work could serve as a valuable supplement to further refine the model's accuracy [28, 62]. It could also be utilized as an added artificial intelligence tool to alert clinician radiologists to an increased risk of both negative and positive cancer findings in screening mammograms if global gist signal can be regarded as a breast cancer risk predictor similar to breast density [30-32]. These findings could then be used to support and advance early detection and diagnosis of breast cancer screening and prediction of future breast cancer risk while reducing diagnostic errors and optimising therapy options and survival.

The **second and third discovery** (Chapter 4 and 5) also advocates that a set of quantitative global mammographic radiomic features (describing the whole image and largest square rectangular inscribed within breast) is valuable for radiology education and breast screening programs. In particular, from the **second findings**, the highlighted six helpful features (*sum entropy*, *difference variance*, *correlation*, and *homogeneity* of GLCM, *range* of FOS, and *coarseness* of NGTDM) especially *correlation* and *range*, could offer new and deeper comprehensions on the probable image features that could best explain radiologists' highly difficult normal cases. This is crucially important since BI-RADS breast density class presented as being an unreliable predictor of normal cases' difficulty [11, 18, 19]. Hence for radiology education benefits, these findings may be utilized to facilitate an automatic and consistent approach to assessing most difficult normal cases and to curate more predictable and consistent test sets educational materials for radiologists in general. The important features could also be incorporated into clinician radiology education, particularly the test set-based ones, to teach readers about the potential difficult normal cases' features that could likely lead to false

positive errors. In terms of screening mammography programs, these novel findings could be used as an augmented clinical artificial intelligence tool to notify clinical managers about difficult/easy normal cases so that appropriate reading strategies (e.g., double/single reading, senior/junior reader) could be well implemented [12, 50-52]. Additionally, they can also enable a dependable and efficient objective assessment of the false positive recall rate performance of each reader and the screening programs longitudinally. This is especially pertinent as the correlation between readers' performance in test set-based education and their clinical performance suggests a valuable metric for ongoing evaluation [3, 4, 7, 63-65].

More importantly, radiology trainees with less experience and exposure in interpreting mammographic images and processing global images often find it highly difficult to correctly interpret normal cancer-free mammograms leading to false positive errors [17, 18, 40, 43]. The **third/last results** (Chapter 5) thus further suggest that global mammographic radiomic features are useful for personalised radiology education exclusively for radiology trainees. The discovery of two significant features (*cluster prominence* of GLCM especially, and *range* of FOS) could provide innovative and meaningful insights into possible appearances of highly challenging normal cases that could be used to train radiology trainees in recognizing features commonly associated with false positive errors. These features can also be used as auxiliary clinical image data supplementary to BI-RADS breast density class since BI-RADS alone does not completely represent the actual difficulty of the individual within the test set [11, 14, 18, 19]. This knowledge is useful not only for objective assessment of greatly difficult normal cases for the purpose of personalised test set generation of radiology trainees' training programs but also for facilitating their specific expertise development in interpreting screening normal mammograms [33, 66-70]. This is so that their false positive errors can be better examined, addressed, and improved while their skills in perceiving the global features of normality can be developed.

6.4 Limitations and future directions

Some limitations about the thesis should be noted. The prevalence of cancer and normal images/cases used in the three original studies (Chapter 3-5) from the BreastScreen Australia (Study 1, Chapter 3)/BREAST test sets (Study 2-3, Chapter 4-5) was respectively higher and lower than the usual clinical practice. In terms of methodology, the studies analysed mostly GLCM texture- and FOS-based global mammographic radiomic features extracted using the

handcrafted approach. Region of interests for feature extraction were delineated using the erosion algorithm (Study 1 Chapter 3), or lattice- and squared-based approaches (Study 2-3, Chapter 4-5). Models were built using random forest technique boosted with bootstrap aggregation (Study 1, Chapter 3) or adaptive logistic regression method (Study 2-3, Chapter 4-5). Also, although studies 2 and 3 (Chapter 4 and 5) extracted the global radiomic features from both CC and MLO images of normal cancer-free mammograms, the first study (Chapter 3) only examined the CC view of both cancer and normal mammograms but not MLOs. In addition, sample readers and images (13 readers and 4191 images - Study 1, 537 readers and 239 cases/956 images - Study 2, and 137 readers and 140 cases/560 images - Study 3) were mainly from Australia or New Zealand. The results from the models were also limited by the above factors inputted into them.

Future research should therefore examine other types of images e.g., MLO/bilateral/cancer mammograms, and DBT (since gist has been found to be recorded in DBT), other approaches to building the models (e.g., XGBoost and neural networks) and defining the region of interest, e.g., retroareolar area, and other types of radiomics texture features/analysis, e.g., gray level run length matrix and local features and/or deep learning analysis [20, 71-73]. Readers with different levels of mammographic reading experience and images from other countries may result in different gist scores, difficulty scores and global radiomic features which should also be investigated.

Lastly, since gist has been claimed to contain both global and local information about an image abnormality [25, 70, 74-76], future research questions arising from the current work could be to explore whether combining handcrafted/deep learning global and local features could improve the model's performance. It may also be worth investigating if further added value could be observed when integrating mammographic radiomic features with clinical features such as breast density, given that this integration has been found to increase predictive accuracy [20, 48].

6.5 Conclusion

In conclusion, the results from this thesis have provided original knowledge about the value of global radiomic features in offering better understanding around image features associated to radiological errors and difficult normal cases of radiologists and radiology trainees when interpretating screening mammograms. Eleven important features (six GLCM: *cluster shade,*

sum entropy, *difference variance*, *correlation*, *homogeneity,* and *cluster prominence;* four FOS: *standard deviation, skewness, kurtosis, and range*; and one NGTDM: *coarseness*) were identified and highlighted.

These novel findings are crucial for refining educational mammographic test set curation for radiologists and radiology trainees and understanding around the process behind medical image perception in recognising computer-extracted image features that drive the global gist signals. In addition, the findings can be helpful for breast screening programs by enabling reliable and effective objective measurement of individual's and screening programs' performance over time. These outcomes can then contribute to advancing early detection, diagnosis, and risk prediction of breast cancer while minimizing errors, optimising reading strategies and treatment choices and ultimately enhancing survival rates.

6.6 References

1. Brennan, P.C., et al. *BREAST: a novel method to improve the diagnostic efficacy of mammography*. in *SPIE 8673, Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment*. 2013. Lake Buena Vista (Orlando Area), Florida, United States: SPIE.
2. Brennan, P.C., et al., *BREAST: A Novel Strategy to Improve the Detection of Breast Cancer*, in *International Workshop on Digital Mammography 2014*, H. Fujita, T. Hara, and C. Muramatsu, Editors. 2014, Springer International Publishing Switzerland: Cham. p. 438-443.
3. Brancato, B., et al., *Mammography self-evaluation online test for screening readers: an Italian Society of Medical Radiology (SIRM) initiative.* Eur Radiol, 2022. **32**(3): p. 1624-1633.
4. Chen, Y., et al., *The Relationship between Mammography Readers' Real-Life Performance and Performance in a Test Set–based Assessment Scheme in a National Breast Screening Program.* Radiology: Imaging Cancer, 2020. **2**(5): p. e200016.
5. Suleiman, M.E., M. Rickard, and P.C. Brennan, *Perfecting detection through education.* Radiography, 2020. **26**: p. S49-S53.
6. Chen, Y. and A. Gale, *Performance Assessment Using Standardized Data Sets: The PERFORMS Scheme in Breast Screening and Other Domains*, in *The Handbook of Medical Image Perception and Techniques*, E. Samei and E.A. Krupinski, Editors. 2018, Cambridge University Press: Cambridge. p. 328-342.
7. Gale, A. and Y. Chen, *A review of the PERFORMS scheme in breast screening.* The British Journal of Radiology, 2020. **93**(1112): p. 20190908.
8. Dong, L., Y. Chen, and A. Gale. *Breast screening: understanding case difficulty and the nature of errors*. in *SPIE 8673, Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment*. 2013. Lake Buena Vista (Orlando Area), Florida, United States: SPIE.
9. Trieu, P.D., et al., *Improving radiologist's ability in identifying particular abnormal lesions on mammograms through training test set with immediate feedback.* Scientific Reports, 2021. **11**(9899): p. 1-8.
10. Scott, H.J. and A.G. Gale. *Breast screening technologists: When is a difficult case truly difficult and for whom?* in *SPIE 5749, Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment*. 2005. Bellingham, WA: SPIE.
11. Wong, D., et al., *Mammographic image characteristics in educational test sets: the effect on expert difficulty ratings and subsequent reader performance.* Asia Pacific Journal of Oncology, 2024. **Under review**.
12. Ang, Z.Z., et al., *Classification of normal screening mammograms is strongly influenced by perceived mammographic breast density.* Journal of Medical Imaging and Radiation Oncology, 2017. **61**(4): p. 461-469.
13. Siviengphanom, S., et al. *Global mammographic radiomic signature can predict radiologists' difficult-to-interpret normal cases*. in *SPIE Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment*. 2023. San Diego, California, United States: SPIE.
14. Siviengphanom, S., et al., *Global radiomic features from mammography for predicting difficult-to-interpret normal cases.* Journal of Digital Imaging, 2023. **36**: p. 1541–1552.

15. Sun, S., et al., *Individualised training to address variability of radiologists' performance*. Medical Imaging. Vol. 6917. 2008: SPIE.
16. Darker, I., Y. Chen, and A. Gale. *Health professionals' agreement on density judgements and successful abnormality identification within the UK Breast Screening Programme*. in *SPIE Medical Imaging*. 2011. SPIE.
17. Mazurowski, M.A. *Difficulty of mammographic cases in the context of resident training: preliminary experimental data*. in *SPIE 8673, Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment* 2013. Lake Buena Vista (Orlando Area), Florida, United States: SPIE.
18. Grimm, L.J., et al., *Radiology Resident Mammography Training.* Academic Radiology, 2014. **21**(7): p. 888-892.
19. Wong, D.J., *Association of diagnostic errors in interpreting screening mammograms with image and reader characteristics*. 2023.
20. Siviengphanom, S., et al., *Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs.* Academic Radiology, 2021. **29**(8): p. 1228-1247.
21. Gillies, R.J. and M.B. Schabath, *Radiomics Improves Cancer Screening and Early Detection.* Cancer Epidemiology Biomarkers & Prevention, 2020. **29**(12): p. 2556-2567.
22. Mao, N., et al., *Added Value of Radiomics on Mammography for Breast Cancer Diagnosis: A Feasibility Study.* Journal of the American College of Radiology, 2019. **16**(4): p. 485-491.
23. Nodine, C.F., et al., *Blinded Review of Retrospectively Visible Unreported Breast Cancers: An Eye-Position Analysis.* Radiology, 2001. **221**(1): p. 122-129.
24. Park, H.-B., et al., *Contributions of global and local processing on medical image perception.* Journal of Medical Imaging, 2023. **10**(S1): p. S11911.
25. Gandomkar, Z., et al., *Reliability of radiologists' first impression when interpreting a screening mammogram.* PLOS ONE, 2023. **18**(4): p. 1-19.
26. Gandomkar, Z., et al. *False-negative diagnosis might occur due to absence of the global radiomic signature of malignancy on screening mammograms*. in *SPIE Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment*. 2023. San Diego, California, United States: SPIE.
27. Gandomkar, Z., et al. *An end-to-end deep learning model can detect the gist of the abnormal in prior mammograms as perceived by experienced radiologists*. in *SPIE Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*. 2021. San Diego, California, United States: SPIE.
28. Gandomkar, Z., et al., *Global processing provides malignancy evidence complementary to the information captured by humans or machines following detailed mammogram inspection.* Scientific Reports, 2021. **11**(20122): p. 1-12.
29. Brennan, P.C., et al., *Radiologists can detect the 'gist' of breast cancer before any overt signs of cancer appear.* Scientific reports, 2018. **8**(1): p. 1-12.
30. Evans, K.K., A.-M. Culpan, and J.M. Wolfe, *Detecting the "gist" of breast cancer in mammograms three years before localized signs of cancer are visible.* The British Journal of Radiology, 2019. **92**(20190136): p. 1-10.
31. Evans, K.K., et al., *The gist of the abnormal: Above-chance medical decision making in the blink of an eye.* Psychonomic Bulletin & Review, 2013. **20**(6): p. 1170-1175.
32. Evans, K.K., et al., *A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast.* Proceedings of the National Academy of Sciences, 2016. **113**(37): p. 10292-10297.

33. Gandomkar, Z. and C. Mello-Thoms, *Visual search in breast imaging.* The British Journal of Radiology, 2019. **92**(1102): p. 20190057.
34. Lambin, P., et al., *Radiomics: Extracting more information from medical images using advanced feature analysis.* European Journal of Cancer, 2012. **48**(4): p. 441-446.
35. Kumar, V., et al., *Radiomics: the process and the challenges.* Magnetic Resonance Imaging, 2012. **30**(9): p. 1234-1248.
36. Gillies, R.J., P.E. Kinahan, and H. Hricak, *Radiomics: Images Are More than Pictures, They Are Data.* Radiology, 2016. **278**(2): p. 563-577.
37. Parekh, V. and M.A. Jacobs, *Radiomics: a new application from established techniques.* Expert Review of Precision Medicine and Drug Development, 2016. **1**(2): p. 207-226.
38. Conti, A., et al., *Radiomics in breast cancer classification and prediction.* Seminars in Cancer Biology, 2020. **72**: p. 238-250.
39. Wang, M., et al., *Predicting false negative errors in digital breast tomosynthesis among radiology trainees using a computer vision-based approach.* Expert Systems with Applications, 2016. **56**: p. 1-8.
40. Zhang, J., et al., *Using computer-extracted image features for modeling of error-making patterns in detection of mammographic masses among radiology residents.* Medical Physics, 2014. **41**(9): p. 091907.
41. Grimm, L.J., et al., *Predicting error in detecting mammographic masses among radiology trainees using statistical models based on BI-RADS features.* Medical Physics, 2014. **41**(3): p. 031909.
42. Mazurowski, M.A., et al., *Identifying Error-making Patterns in Assessment of Mammographic BI-RADS Descriptors among Radiology Residents Using Statistical Pattern Recognition.* Academic Radiology, 2012. **19**(7): p. 865-871.
43. Zhang, J., J.I. Silber, and M.A. Mazurowski, *Modeling false positive error making patterns in radiology trainees for improved mammography education.* Journal of Biomedical Informatics, 2015. **54**: p. 50-57.
44. Wang, M., et al., *A computer vision-based algorithm to predict false positive errors in radiology trainees when interpreting digital breast tomosynthesis cases.* Expert Systems with Applications, 2016. **64**: p. 490-499.
45. Mazurowski, M. and G. Tourassi. *Modeling error in assessment of mammographic image features for improved computer-aided mammography training: initial experience*. in *SPIE Medical Imaging*. 2011. SPIE.
46. Mazurowski, M., J. Lo, and G. Tourassi. *User modeling for improved computer-aided training in radiology: initial experience*. in *SPIE Medical Imaging*. 2010. SPIE.
47. Mazurowski, M.A., et al., *Individualized computer-aided education in mammography based on user modeling: Concept and preliminary experiments.* Medical Physics, 2010. **37**(3): p. 1152-1160.
48. Voisin, S., et al., *Predicting diagnostic error in radiology via eye-tracking and image analytics: Preliminary investigation in mammography.* Medical Physics, 2013. **40**(10): p. 101906.
49. Mazurowski, M., et al. *Modeling resident error-making patterns in detection of mammographic masses using computer-extracted image features: preliminary experiments*. in *SPIE Medical Imaging*. 2014. SPIE.
50. Brennan, P.C., et al., *Benefits of Independent Double Reading in Digital Mammography.* Academic Radiology, 2019. **26**(6): p. 717-723.

51. Taylor-Phillips, S. and C. Stinton, *Double reading in breast cancer screening: considerations for policy-making.* The British Journal of Radiology, 2020. **93**(1106): p. 20190610.
52. Gandomkar, Z., et al., *Can eye-tracking metrics be used to better pair radiologists in a mammogram reading task?* Medical Physics, 2018. **45**(11): p. 4844-4856.
53. Kabbai, L., M. Abdellaoui, and A. Douik, *Image classification by combining local and global features.* The Visual Computer, 2019. **35**(5): p. 679-693.
54. Roy, S., et al., *Optimal co-clinical radiomics: Sensitivity of radiomic features to tumour volume, image noise and resolution in co-clinical T1-weighted and T2-weighted magnetic resonance imaging.* EBioMedicine, 2020. **59**: p. 102963.
55. Suleiman, W.I., et al., *Impact of Breast Reader Assessment Strategy on mammographic radiologists' test reading performance.* Journal of Medical Imaging and Radiation Oncology, 2016. **60**(3): p. 352-358.
56. Renninger, L.W. and J. Malik, *When is scene identification just texture recognition?* Vision Research, 2004. **44**(19): p. 2301-2311.
57. Oliva, A. and A. Torralba, *Building the gist of a scene: the role of global image features in recognition*, in *Progress in Brain Research*, S. Martinez-Conde, et al., Editors. 2006, Elsevier. p. 23-36.
58. Oliva, A. and A. Torralba, *Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope.* International Journal of Computer Vision, 2001. **42**(3): p. 145-175.
59. Chubb, C., et al., *The three dimensions of human visual sensitivity to first-order contrast statistics.* Vision Research, 2007. **47**(17): p. 2237-2248.
60. Julesz, B., *Texture and Visual Perception.* Scientific American, 1965. **212**(2): p. 38-49.
61. Julesz, B., *Visual Pattern Discrimination.* IEEE Transactions on Information Theory, 1962. **8**(2): p. 84-92.
62. Shen, L., et al., *Deep Learning to Improve Breast Cancer Detection on Screening Mammography.* Scientific Reports, 2019. **9**(12495): p. 1-12.
63. Soh, B.P., et al., *Screening Mammography: Test Set Data Can Reasonably Describe Actual Clinical Reporting.* Radiology, 2013. **268**(1): p. 46-53.
64. Soh, B.P., et al., *Certain performance values arising from mammographic test set readings correlate well with clinical audit.* Journal of Medical Imaging and Radiation Oncology, 2015. **59**(4): p. 403-410.
65. Qenam, B.A., T. Li, and P.C. Brennan. *Test set participation may impact positively on clinical mammographic performance*. in *SPIE 11599, Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*. 2021. Online: SPIE.
66. Nodine, C.F., et al., *Nature of expertise in searching mammograms for breast masses.* Academic radiology, 1996. **3**(12): p. 1000-1006.
67. Nodine, C.F., et al., *How experience and training influence mammography expertise.* Acad Radiol, 1999. **6**(10): p. 575-85.
68. Ekpo, E.U., M. Alakhras, and P. Brennan, *Errors in Mammography Cannot be Solved Through Technology Alone.* Asian Pacific journal of cancer prevention : APJCP, 2018. **19**(2): p. 291-301.
69. Buist, D.S., et al., *Influence of annual interpretive volume on screening mammography performance in the United States.* Radiology, 2011. **259**(1): p. 72-84.
70. Kundel, H.L., et al., *Holistic Component of Image Perception in Mammogram Interpretation: Gaze-Tracking Study.* Radiology, 2007. **242**: p. 396-402.

71. Pertuz, S., et al. *Open Framework for Mammography-based Breast Cancer Risk Assessment*. in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. 2019. Chicago, IL, USA: IEEE.
72. Wei, J., et al., *Association of Computerized Mammographic Parenchymal Pattern Measure with Breast Cancer Risk: A Pilot Case-Control Study.* Radiology, 2011. **260**(1): p. 42-49.
73. Wu, C.C., et al., *Gist processing in digital breast tomosynthesis.* J Med Imaging (Bellingham), 2020. **7**(2): p. 022403.
74. Carrigan, A.J., S.G. Wardle, and A.N. Rich, *Finding cancer in mammograms: if you know it's there, do you know where?* Cognitive Research: Principles and Implications, 2018. **3**(1): p. 1-14.
75. Ungerleider, L. and M. Mishkin, *Two Cortical Visual Systems*, in *Analysis of visual behavior*, D.J. Ingle, M.A. Goodale, and R.J. Mansfield, Editors. 1982, The MIT press: London, England. p. 549-586.
76. Siviengphanom, S., et al., *Computer-extracted global radiomic features can predict the radiologists' first impression about the abnormality of a screening mammogram.* British Journal of Radiology, 2024. **97**(1153): p. 168–179.

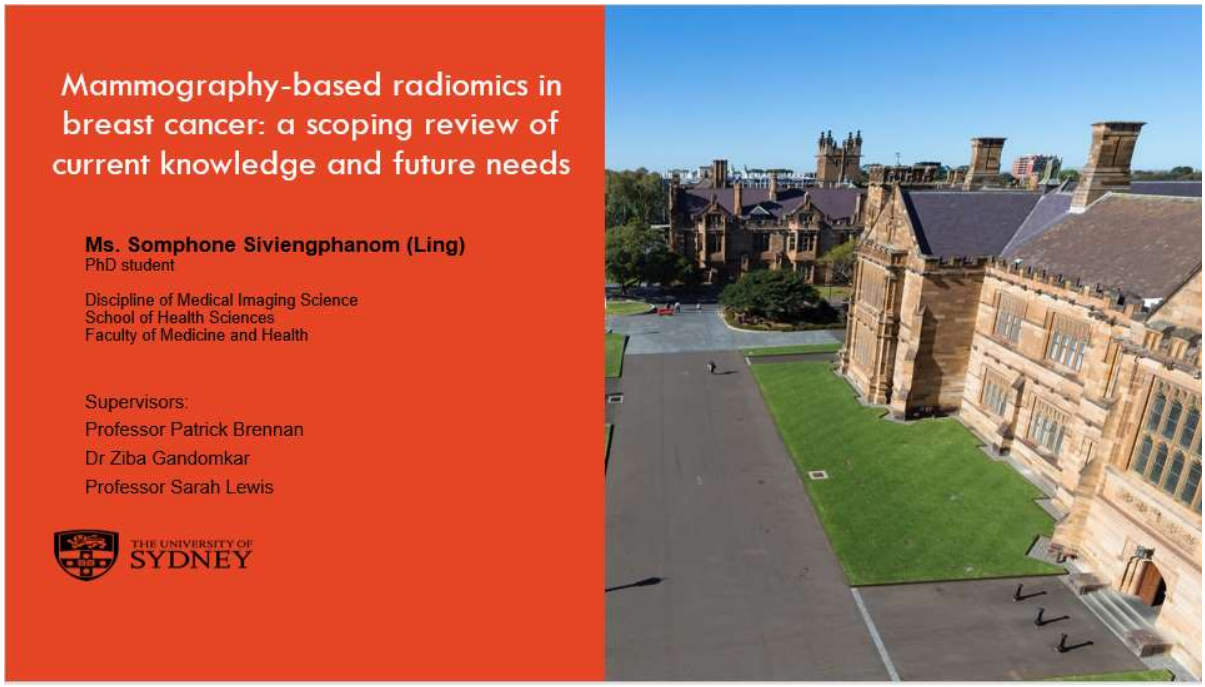
Appendix 1

Presentations and abstract

This appendix consists of:

- Appendix 1.1 Oral presentation on *mammography-based radiomics in breast cancer: a scoping review of current knowledge and future needs*.
- Appendix 1.2 Poster presentation on *mammography-based radiomics in breast cancer: a scoping review of current knowledge and future needs*.
- Appendix 1.3 Conference proceedings on *predicting the gist of breast cancer on a screening mammogram using global radiomic features.*
- Appendix 1.4 Oral presentation on *predicting the gist of breast cancer on a screening mammogram using global radiomic features*.
- Appendix 1.5 Conference proceedings on *global mammographic radiomic signature can predict radiologists' difficult-to-interpret normal cases*.
- Appendix 1.6 Oral presentation on *global mammographic radiomic signature can predict radiologists' difficult-to-interpret normal cases*.
- Appendix 1.7 Abstract on *enhancing radiology resident training in mammogram interpretation: leveraging global mammographic radiomic features to identify difficult cases*.

Appendix 1.1 Oral presentation on *mammography-based radiomics in breast cancer: a scoping review of current knowledge and future needs*

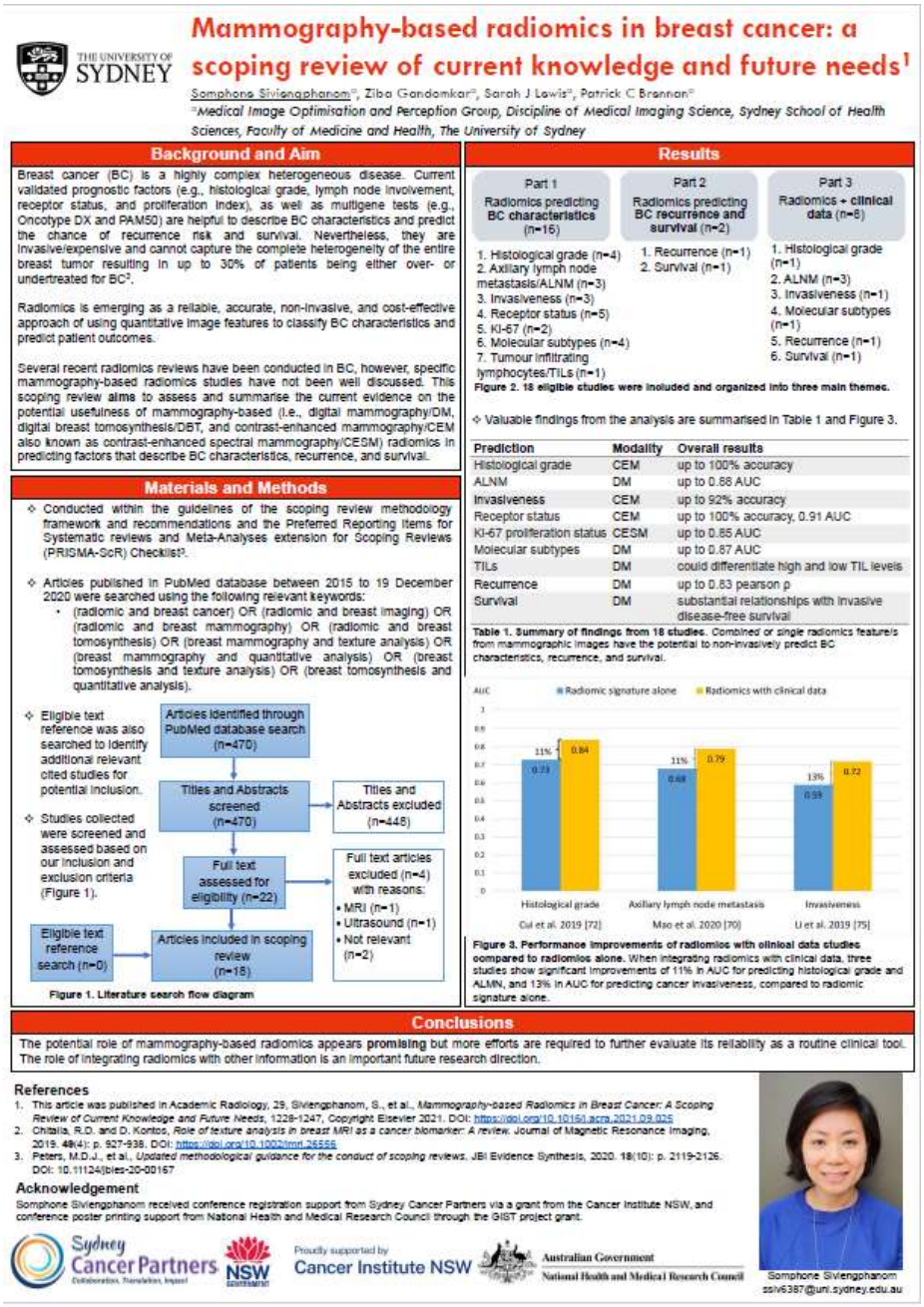


This oral presentation was given at:

- The University of Sydney, Medical Image Optimisation and Perception Group (MIOPeG), Sydney, Australia, April 2022.
- The University of Sydney, Faculty of Medicine and Health Higher Degree by Research Conference, Sydney, Australia, July 2022.

This presentation was related to the paper "Mammography-based radiomics in breast cancer: a scoping review of current knowledge and future needs" as described in Chapter 2.

Appendix 1.2 Poster presentation on *mammography-based radiomics in breast cancer: a scoping review of current knowledge and future needs*



Mammography-based radiomics in breast cancer: a scoping review of current knowledge and future needs[1]

Somphone Sivengphanom[a], Ziba Gandomkar[a], Sarah J Lewis[a], Patrick C Brennan[a]
[a]Medical Image Optimisation and Perception Group, Discipline of Medical Imaging Science, Sydney School of Health Sciences, Faculty of Medicine and Health, The University of Sydney

Background and Aim

Breast cancer (BC) is a highly complex heterogeneous disease. Current validated prognostic factors (e.g., histological grade, lymph node involvement, receptor status, and proliferation index), as well as multigene tests (e.g., Oncotype DX and PAM50) are helpful to describe BC characteristics and predict the chance of recurrence risk and survival. Nevertheless, they are invasive/expensive and cannot capture the complete heterogeneity of the entire breast tumor resulting in up to 30% of patients being either over- or undertreated for BC[2].

Radiomics is emerging as a reliable, accurate, non-invasive, and cost-effective approach of using quantitative image features to classify BC characteristics and predict patient outcomes.

Several recent radiomics reviews have been conducted in BC, however, specific mammography-based radiomics studies have not been well discussed. This scoping review **aims** to assess and summarise the current evidence on the potential usefulness of mammography-based (i.e., digital mammography/DM, digital breast tomosynthesis/DBT, and contrast-enhanced mammography/CEM also known as contrast-enhanced spectral mammography/CESM) radiomics in predicting factors that describe BC characteristics, recurrence, and survival.

Materials and Methods

- Conducted within the guidelines of the scoping review methodology framework and recommendations and the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist[3].
- Articles published in PubMed database between 2015 to 19 December 2020 were searched using the following relevant keywords:
  - (radiomic and breast cancer) OR (radiomic and breast imaging) OR (radiomic and breast mammography) OR (radiomic and breast tomosynthesis) OR (breast mammography and texture analysis) OR (breast mammography and quantitative analysis) OR (breast tomosynthesis and texture analysis) OR (breast tomosynthesis and quantitative analysis).
- Eligible text reference was also searched to identify additional relevant cited studies for potential inclusion.
- Studies collected were screened and assessed based on our inclusion and exclusion criteria (Figure 1).

Figure 1. Literature search flow diagram

Results

Part 1	Part 2	Part 3
Radiomics predicting BC characteristics (n=16)	Radiomics predicting BC recurrence and survival (n=2)	Radiomics + clinical data (n=8)
1. Histological grade (n=4) 2. Axillary lymph node metastasis/ALNM (n=3) 3. Invasiveness (n=3) 4. Receptor status (n=5) 5. KI-67 (n=2) 6. Molecular subtypes (n=4) 7. Tumour infiltrating lymphocytes/TILs (n=1)	1. Recurrence (n=1) 2. Survival (n=1)	1. Histological grade (n=1) 2. ALNM (n=3) 3. Invasiveness (n=1) 4. Molecular subtypes (n=1) 5. Recurrence (n=1) 6. Survival (n=1)

Figure 2. 18 eligible studies were included and organized into three main themes.

- Valuable findings from the analysis are summarised in Table 1 and Figure 3.

Prediction	Modality	Overall results
Histological grade	CEM	up to 100% accuracy
ALNM	DM	up to 0.88 AUC
Invasiveness	CEM	up to 92% accuracy
Receptor status	CEM	up to 100% accuracy, 0.91 AUC
KI-67 proliferation status	CESM	up to 0.85 AUC
Molecular subtypes	DM	up to 0.87 AUC
TILs	DM	could differentiate high and low TIL levels
Recurrence	DM	up to 0.83 pearson p
Survival	DM	substantial relationships with invasive disease-free survival

Table 1. Summary of findings from 18 studies. Combined or single radiomics feature/s from mammographic images have the potential to non-invasively predict BC characteristics, recurrence, and survival.

Figure 3. Performance improvements of radiomics with clinical data studies compared to radiomics alone. When integrating radiomics with clinical data, three studies show significant improvements of 11% in AUC for predicting histological grade and ALNM, and 13% in AUC for predicting cancer invasiveness, compared to radiomic signature alone.

Conclusions

The potential role of mammography-based radiomics appears **promising** but more efforts are required to further evaluate its reliability as a routine clinical tool. The role of integrating radiomics with other information is an important future research direction.

References

1. This article was published in Academic Radiology, 29, Sivengphanom, S., et al., *Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs*, 1228-1247, Copyright Elsevier 2021. DOI: https://doi.org/10.1016/j.acra.2021.09.025
2. Chitalia, R.D. and D. Kontos, *Role of texture analysis in breast MRI as a cancer biomarker: A review.* Journal of Magnetic Resonance Imaging, 2019. 49(4): p. 927-938. DOI: https://doi.org/10.1002/jmri.26556
3. Peters, M.D.J., et al., *Updated methodological guidance for the conduct of scoping reviews.* JBI Evidence Synthesis, 2020. 18(10): p. 2119-2126. DOI: 10.11124/jbies-20-00167

Acknowledgement

Somphone Sivengphanom received conference registration support from Sydney Cancer Partners via a grant from the Cancer Institute NSW, and conference poster printing support from National Health and Medical Research Council through the GIST project grant.

Somphone Sivengphanom
ssiv6387@uni.sydney.edu.au

This poster presentation was given at New South Wales (NSW) Cancer Conference, Sydney, Australia, September 2022.

This presentation was related to the paper "Mammography-based radiomics in breast cancer: a scoping review of current knowledge and future needs" as described in Chapter 2.

Appendix 1.3 Conference proceedings on p*redicting the gist of breast cancer on a screening mammogram using global radiomic features*

Predicting the gist of breast cancer on a screening mammogram using global radiomic features

Somphone Siviengphanom*[1], Sarah J. Lewis[1], Patrick C. Brennan[1], Ziba Gandomkar[1]
[1]Medical Image Optimisation and Perception Group, Discipline of Medical Imaging Science, Faculty of Medicine and Health, the University of Sydney, Sydney, NSW 2006, Australia

ABSTRACT

This study explored if using a set of global radiomic (i.e., computer-extracted) features derived from mammograms could predict the gist of breast cancer (holistic perceptual information provided from radiologists' first impression about the presence of an image abnormality after a brief sight of the image). A retrospective de-identified dataset was used to collect the gist of breast cancer (i.e., gist scores) from 13 readers interpreting 1100 screening craniocaudal mammograms (659 current "normal" cancer-free images, and 441 "prior" no visible signs of cancer images acquired two years before current cancer mammograms). The collected gist scores from all readers were averaged to eliminate the noise of the gist signal, giving one gist score per image. The images were grouped into *high-* and *low-gist* based on the 75th and 25th percentiles of the images containing the highest and lowest gist scores, respectively. A set of 130 handcrafted global radiomic features per image were extracted and used to construct two machine learning random forest classifiers: 1). *Normal* and 2). *Prior* based on the corresponding features computed from the "normal" and "prior" images, for distinguishing *high-* from *low-gist* images. The classifiers were trained and validated using the 10-fold cross-validation approach and their performances were measured by the area under the receiver operating characteristic curve (AUC). The *Normal* and *Prior* classifiers resulted in AUCs of 0.83 (95% CI: 0.77-0.85) and 0.84 (95% CI: 0.80-0.87) respectively, suggesting that the global mammographic radiomic features can predict the gist of breast cancer on a screening mammogram.

Keywords: radiomics, screening mammograms, gist, breast cancer, normal mammograms, prior mammograms, machine learning, global radiomic features

1. INTRODUCTION

In the field of medical image perception, it has been shown that medical experts (for instance, radiologists) can surprisingly perceive an initial impression about the presence of a medical image abnormality or its "gist" after a quick glimpse on an image [1-3]. They can even detect an image abnormality in prior normal mammograms belonging to women who were later diagnosed with breast cancer (i.e., missed cancers based on prior images) [1; 4] as well as in current normal mammograms, containing no apparent lesion sign, contralateral to the malignancies [5]. These findings evidence that the gist of breast cancer without a localised source of malignancy, possibly driven by the mammographic textural features (not associated with other risk factors like breast density), may contain useful information that could be used to improve a detection of breast cancer and prediction of a future breast cancer risk event [1; 4; 6; 7].

Gist, from the human visual system, has been claimed to be linked to both "what" and "where" information about the global and location of image abnormality, respectively [8; 9]. However, most of the recent gist mammographic studies showed that the gist of breast cancer is rather about the "what" information extracted quickly after a half-second image display for example, from the entire global image properties in a "non-selective" manner [1; 4; 5; 10; 11]. Here, we aim at focusing only the "what" or global gist of breast cancer information that is the holistic perceptual information of radiological instantaneous view of an image's abnormality features captured in as little as a half-second image presentation.

*ssiv6387@uni.sydney.edu.au

Medical Imaging 2024: Image Perception, Observer Performance, and Technology Assessment,
edited by Claudia R. Mello-Thoms, Yan Chen, Proc. of SPIE Vol. 12929, 129290J

Several studies in medical image perception have widely examined the presence of the global gist of breast cancer to better understand radiological errors in breast screening performances [1; 4; 5; 12]. Breast cancer in females is the primary cause of cancer-related death globally [13]. With screening mammography shown to reduce breast cancer related death through early detection of disease, interpreting mammographic images correctly is immensely critical [14]. However, this is a highly challenging task with studies reported considerable 16% false positive recall rates and 63% cumulative risk of a false positive outcome after 10 breast screening rounds [15; 16], as well as up to 50% false negative rates [1]. False positives could trigger large yearly healthcare costs (for example, US$2.8 billion was reported in the United States) and increased patient psychological distress [17; 18], while false negative mistakes might lead to delayed detection of disease, increased treatment complications and poor survival [19; 20]. Moreover, textural features (involving first- and second-order gray level co-occurrence matrix/GLCM-based statistics) have been shown to be helpful in identifying the global gist of natural scenes [21-24]. However, it is unknown if a set of these global features from mammograms using radiomics approach [25] can distinguish the strongly perceived (i.e., *high-gist*) from the poorly perceived gist of breast cancer images (i.e., *low-gist*). We assume that there are differences in the global radiomic features between the highest and lowest of gist images (concentrating only on these categories to enable a strong distinction between them). This paper, therefore, aims to tackle this research gap so that the characteristics of false positives (normal mammograms with strongly perceived gist of breast cancer) and false negatives (missed cancer on prior mammograms with poorly perceived gist) in the interpretation of screening mammograms can be better understood. This could yield insights into radiological errors in screening mammograms and radiology education about possible mammographic appearances of false positive and false negative cases, facilitating early breast cancer detection and diagnosis, and optimizing treatment selections and survivorship.

2. MATERIALS AND METHODS

2.1 Readers

17 Australian and New Zealand radiologists and breast physicians who were identified to be the "super-gisters" (demonstrated superior ability to detect reliable and accurate gist of the abnormal in mammographic images) from their participation in the previous gist experiment [26], were invited to participate in this study. This superior ability of the "super-gisters" showed to have no association with factors like their number of mammograms interpreted each week or years of experience reading mammograms. In total, 13 readers (11 radiologists and 2 breast physicians) were recruited for the experiment.

2.2 Mammograms

A retrospective de-identified dataset of 1100 craniocaudal (CC) unilateral Digital Imaging and Communications in Medicine (DICOM) mammograms were obtained from the BreastScreen Australia archives, acquired from asymptomatic women aged between 50 and 75. Using these DICOM images with a standard thresholding gray level intensity value of 100, binary masks were generated to extract the breast region from its background. After that, the DICOM images and masks were converted to a TIFF file format and cropped based on the maximum breast region size including all the right CC images being flipped to the left for consistent left side chest wall on all images. These TIFF images and binary masks were then used for the gist experiment (section 2.3) and radiomics analysis (section 2.5).

Two types of mammograms were included in this study: 1). current "normal" (n=1409) and 2). "prior" (n=1063) mammograms since the gist of breast cancer has been reported to be presented in these images [1; 4; 5]. The current "normal" mammograms were images of women reported and confirmed to be cancer-free by at least two independent expert radiologists (20+ year experience) with negative follow up mammograms obtained two years later. The "prior" mammograms included prior images with no visible cancer signs, obtained two years before their current biopsy-proven cancer mammograms.

2.3 Collecting gist of breast cancer/scores

Using previously developed multi-observer gist experimental protocol [26], the gist of breast cancer (i.e., gist scores) were collected from the 13 participants who were asked to assess the 1100 CC images. For a half-second each, first a red cross sign was presented to ensure readers fixated at the centre of the screen, followed by a mammographic image and then a corresponding white mask signifying the breast area was shown to ensure conclusion of image visual processing. Finally, without a time limit, a rating interface was displayed requesting readers to provide a gist score (i.e., the probability of the image being abnormal) based on a scale of 0 to 100 (0 means readers were completely confident that the image was normal while 100 represents readers' absolute confidence that the image was abnormal).

Due to the gist of breast cancer being a noisy signal [27], the collected gist scores from the 13 readers were then averaged to produce one gist score per image.

2.4 Classifying high- vs low-gist images

To classify *high*- and *low-gist* images, the averaged gist score per image was used to group images as *high*- and *low-gist* based on the 75th and 25th percentiles of the images containing the highest and lowest gist scores, correspondingly, resulting in 162 *high*- and 497 *low-gist* for the "normal", and 151 *high*- and 290 *low-gist* for the "prior" category. To ensure an obvious distinction between the highest and lowest gist images at their maximum level is achieved, only images in these upper and lower quartiles were included for further analysis, and 750 and 622 intermediate gist images in the "normal" and "prior" category respectively were eliminated.

2.5 Extracting global radiomic features

A set of 130 handcrafted global radiomic features were extracted using our in-house MATLAB-based programs based on the region of interest delineated using MATLAB's erosion algorithm, to exclude the skin-air region in the mammogram (Figure 1) [28].

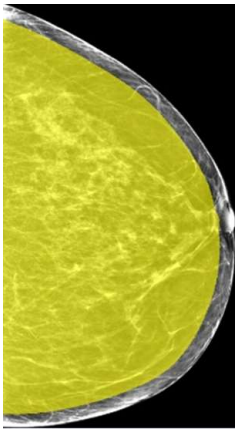


Figure. 1. The region of interest (yellow area) delineated using MATLAB's image erosion algorithm to omit skin-air region. Global radiomic features were computed from this yellow region of the image.

These global radiomic features consisted of 110 GLCM-based textural features and 20 first-order statistics based features (Table 1) which have been shown to be valuable descriptors of mammographic appearances in measuring the contrast values of spatial inter-relationships between neighbouring pixels and the distribution of single pixel intensity values within the image region of interest, correspondingly [29; 30].

Table 1: List of global radiomic features derived from mammograms (n=130)

Feature No.	Feature Name	Feature class	Parameters
1-5	Autocorrelation	Gray level co-occurrence matrix	1-, 3-, 5-, 9-, and 11-pixel distance between the pixel of interest and its neighbour
6-10	Contrast		
11-15	Correlation_m		
16-20	Correlation_p		
21-25	Cluster_prominence		
26-30	Cluster_shade		
31-35	Dissimilarity		
36-40	Energy		

41-45	Entropy		
46-50	Homogeneity_m		
51-55	Homogeneity		
56-60	Maximum_probability		
61-65	Sum_of_sqaures_variance		
66-70	Sum_average		
71-75	Sum_variance		
76-80	Sum_entropy		
81-85	Difference_variance		
86-90	Difference_entropy		
91-95	Information_measure_of_correlation1		
96-100	Information_measure_of_correlation2		
101-105	Inverse_difference_normalized		
106-110	Inverse_difference_moment_normalized		
111	Mean	First-order statistics	n/a
112	Standard_deviation		
113	Skewness		
114	Kurtosis		
115	Minimum		
116	5th_percentile		
117	10th_percentile		
118	15th_percentile		
119	20th_percentile		
120	25th_percentile		
121	Median		
122	75th_percentile		
123	80th_percentile		
124	85th_percentile		
125	90th_percentile		
126	95th_percentile		
127	Maximum		
128	Range_all (Maximum less Minimum)		
129	Range5 (95th percentile less 5th percentile)		
130	Range2 (99th percentile less 1st percentile)		

2.6 Building machine learning classifier

In discriminating *high-* vs *low-gist* images, the 130 global radiomic features derived from the corresponding image types ("normal" and "prior") were then used to construct two separate machine learning random forest classifiers (*Normal* and *Prior*) using bootstrap aggregation method based on the ensemble of 500 decision trees [31]. We chose the random forest bootstrapping approach for the classification task because it has automatic estimation of feature importance for producing an explainable classifier while offering integrated feature selection function for reducing feature-overfitting issues [32].

2.7 Statistical analysis and validation

The *Normal* and *Prior* classifiers were then trained and validated using the 10-fold cross-validation approach [33] which comprised of randomly dividing the dataset into 10 groups. We used 9 out of the 10 groups to train the classifier while reserving one group for testing the classifier's predictive performance. We replicated this process 10 times until each group was used once as a test set. The overall performance of the classifiers for distinguishing *high-* from *low-*gist images was then assessed using the area under the receiver operating characteristic curve (AUC).

Radiomics and statistical analysis were preformed using MATLAB R2022a (MathWorks Inc., Natick, MA, USA).

3. RESULTS

Figure. 2 demonstrated the results of the overall performance from the machine learning random forest classifiers (*Normal* and *Prior*) in categorizing *high-* from *low-gist* images, attaining AUCs of 0.83 (95% CI: 0.77-0.85) and 0.84 (95% CI: 0.80-0.87), correspondingly.

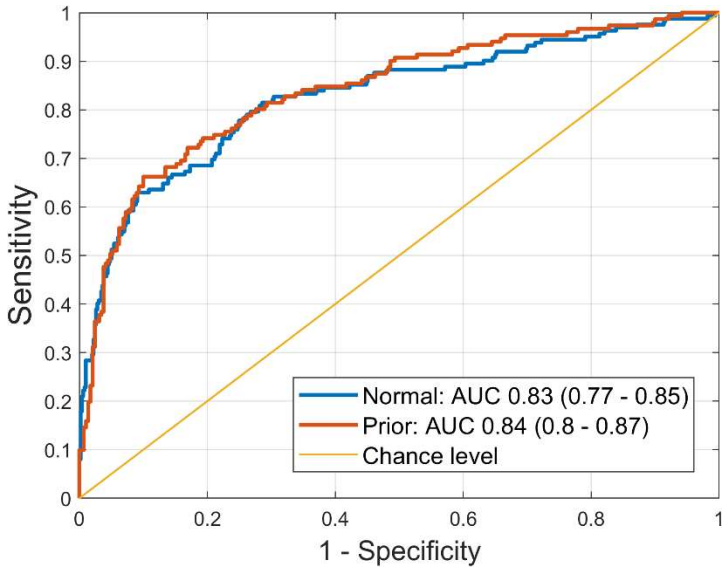


Figure. 2. Overall performance of the machine learning random forest classifiers (*Normal* and *Prior*) in categorizing *high-* from *low-gist* images based on the area under the receiver operating characteristic curve (AUC).

4. CONCLUSION

The overall promising results of the classifiers indicated the ability of global radiomic features contained within mammographic images in predicting the gist of breast cancer (*high-* from *low-gist* images) on a screening mammogram. This could generate important understandings around radiological errors in screening mammography and radiology education about possible features of false positives (normal images with strong gist) and false negatives (missed cancer on prior images with weak gist), supporting early detection and diagnosis of breast cancer while optimizing therapy choices and survivorship.

REFERENCES

1 Brennan PC, Gandomkar Z, Ekpo EU et al (2018) Radiologists can detect the 'gist' of breast cancer before any overt signs of cancer appear. Scientific reports 8:8717-8717

2 Gandomkar Z, Mello-Thoms C (2019) Visual search in breast imaging. The British journal of radiology 92:20190057

3 Evans KK, Georgian-Smith D, Tambouret R, Birdwell RL, Wolfe JM (2013) The gist of the abnormal: Above-chance medical decision making in the blink of an eye. Psychonomic Bulletin & Review 20:1170-1175

4 Evans KK, Culpan A-M, Wolfe JM (2019) Detecting the “gist” of breast cancer in mammograms three years before localized signs of cancer are visible. The British journal of radiology 92:20190136

5 Evans KK, Haygood TM, Cooper J, Culpan A-M, Wolfe JM (2016) A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast. Proceedings of the National Academy of Sciences 113:10292-10297

6 Gandomkar Z, Li T, Shao Z et al (2020) Breast cancer risk prediction in Chinese women based on mammographic texture and a comprehensive set of epidemiologic factorsFifteenth International Workshop on Breast Imaging. SPIE, Leuven, Belgium

7 Gastounioti A, Conant EF, Kontos D (2016) Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. Breast Cancer Research 18:1-12

8 Drew T, Evans K, Võ ML, Jacobson FL, Wolfe JM (2013) Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images? RadioGraphics 33:263-274

9 Carrigan AJ, Wardle SG, Rich AN (2018) Finding cancer in mammograms: if you know it’s there, do you know where? Cognitive Research: Principles and Implications 3:1-14

10 Gandomkar Z, Siviengphanom S, Ekpo EU et al (2021) Global processing provides malignancy evidence complementary to the information captured by humans or machines following detailed mammogram inspection. Scientific reports 11:1-12

11 Wolfe JM, Võ ML, Evans KK, Greene MR (2011) Visual search in scenes involves selective and nonselective pathways. Trends Cogn Sci 15:77-84

12 Raat EM, Farr I, Wolfe JM, Evans KK (2021) Comparable prediction of breast cancer risk from a glimpse or a first impression of a mammogram. Cognitive Research: Principles and Implications 6:1-14

13 Sung H, Ferlay J, Siegel RL et al (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians 71:209-249

14 Seely JM, Alhassan T (2018) Screening for breast cancer in 2018-what should we be doing today? Current oncology (Toronto, Ont) 25:S115-S124

15 Castells X, Torá-Rocamora I, Posso M et al (2016) Risk of Breast Cancer in Women with False-Positive Results according to Mammographic Features. Radiology 280:379-386

16 Vlahiotis A, Griffin B, Stavros Md FAT, Margolis J (2018) Analysis of utilization patterns and associated costs of the breast imaging and diagnostic procedures after screening mammography. ClinicoEconomics and Outcomes Research Volume 10:157-167

17 Ong M-S, Mandl KD (2015) National Expenditure For False-Positive Mammograms And Breast Cancer Overdiagnoses Estimated At $4 Billion A Year. Health Affairs 34:576-583

18 Ho PJ, Bok CM, Ishak HMM et al (2019) Factors associated with false-positive mammography at first screen in an Asian population. PLOS ONE 14:e0213615

19 Li Y, Zhou Y, Mao F et al (2019) The influence on survival of delay in the treatment initiation of screening detected non-symptomatic breast cancer. Scientific reports 9:1-7

20 Wadhwa A, Sullivan JR, Gonyo MB (2016) Missed Breast Cancer: What Can We Learn? 45:402-419

21 Oliva A, Torralba A (2001) Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. International Journal of Computer Vision 42:145-175

22 Renninger LW, Malik J (2004) When is scene identification just texture recognition? Vision Research 44:2301-2311

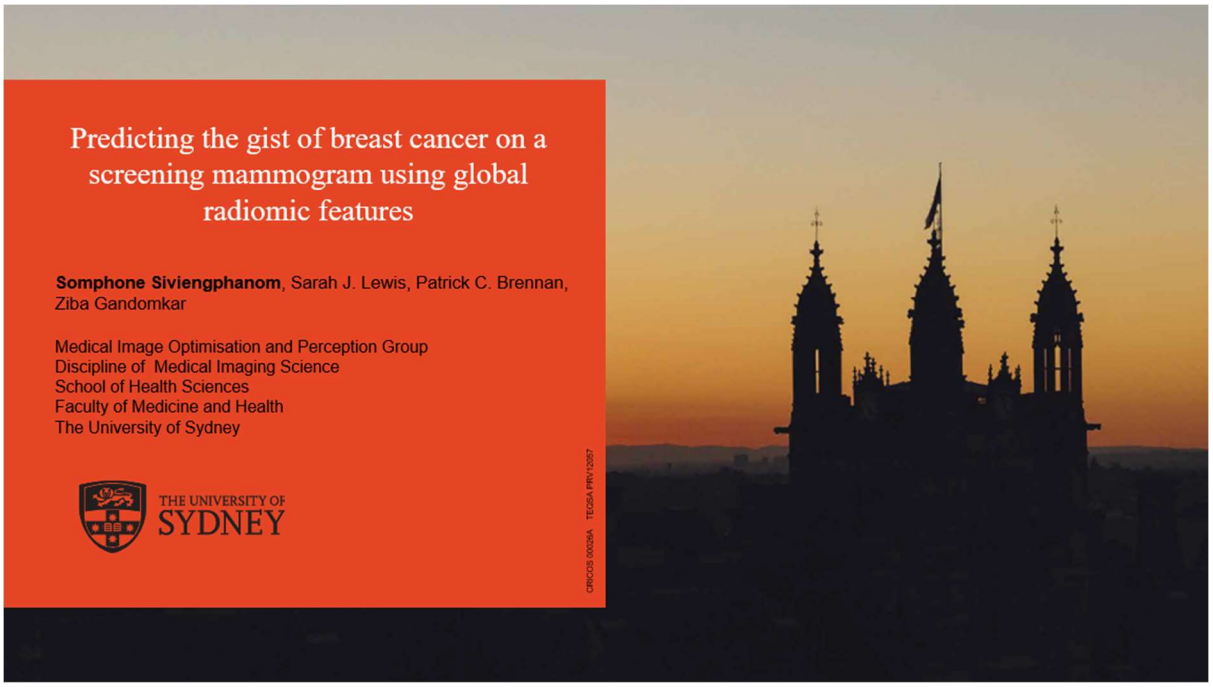
23 Julesz B (1981) Textons, the elements of texture perception, and their interactions. Nature 290:91-97

24 Nisbett WH, Kavuri A, Das M (2020) On the correlation between second order texture features and human observer detection performance in digital images. Scientific reports 10

25 Siviengphanom S, Gandomkar Z, Lewis SJ, Brennan PC (2021) Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs. Academic Radiology 29:1228-1247

26 Gandomkar Z, Siviengphanom S, Suleiman Ma et al (2022) The reliability of radiologists' first impression interpreting a screening mammogramSPIE Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment. SPIE, San Diego, California, United States
27 Gandomkar Z, Ekpo EU, Lewis SJ et al (2021) An end-to-end deep learning model can detect the gist of the abnormal in prior mammograms as perceived by experienced radiologistsSPIE Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment. SPIE
28 van den Boomgaard R, van Balen R (1992) Methods for fast morphological image transforms using bitmapped binary images. CVGIP: Graphical Models and Image Processing 54:252-258
29 Haralick RM, Shanmugam K, Dinstein IH (1973) Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics SMC-3:610-621
30 Scapicchio C, Gabelloni M, Barucci A, Cioni D, Saba L, Neri E (2021) A deep look into radiomics. La radiologia medica 126:1296-1311
31 Breiman L (2001) Random Forests. Machine Learning 45:5-32
32 Couronné R, Probst P, Boulesteix A-L (2018) Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 19:1-14
33 Bengio Y, Grandvalet Y (2004) No Unbiased Estimator of the Variance of K-Fold Cross-Validation. The Journal of Machine Learning Research 5:1089–1105

Appendix 1.4 Oral presentation on *predicting the gist of breast cancer on a screening mammogram using global radiomic features*



This presentation was given at SPIE Medical Imaging Conference 2024, San Diego, CA, United States, February 2024.

This presentation was related to the paper "Computer-extracted global radiomic features can predict the radiologists' first impression about the abnormality of a screening mammogram" as described in Chapter 3.

Appendix 1.5 Conference proceedings on g*lobal mammographic radiomic signature can predict radiologists' difficult-to-interpret normal cases*

Global mammographic radiomic signature can predict radiologists' difficult-to-interpret normal cases

Somphone Siviengphanom*[1], Ziba Gandomkar[1], Sarah J. Lewis[1], Patrick C. Brennan[1]
[1]Medical Image Optimisation and Perception Group, Discipline of Medical Imaging Science, Faculty of Medicine and Health, the University of Sydney, Sydney, NSW 2006, Australia

ABSTRACT

This study investigated whether a global radiomic signature (i.e., a set of global radiomic features) from mammograms can predict radiologists' *difficult-to-interpret* normal cases. Retrospective non-identifiable data collected from 342 radiologists interpreting 81 normal mammograms were used to group cases as *difficult-to-interpret* (41 cases) and *easy-to-interpret* (40 cases) based on one-third of cases having the correspondingly highest and lowest difficulty scores. A set of 34 global radiomic features per image were extracted based on regions of interests delineated using lattice- and squared-based approaches, and normalised. Three machine learning classification models were constructed: 1). *CC*, using the 34 global radiomic features derived from craniocaudal images only, and 2). *MLO*, using the features from mediolateral oblique images only, both based on a random forest method for differentiating *difficult-to-interpret* from *easy-to-interpret* normal cases, and 3). *CC+MLO* model using the median predictive scores from both *CC* and *MLO* models. We trained and validated the models using leave-one-out-cross-validation approach. Performances of the models were measured by the area under the receiver operating characteristic curve (AUC). The *CC+MLO* model outperformed (0.73 AUC, 0.62-0.83) the *CC* (0.70 AUC, 0.62-0.78) and *MLO* (0.68 AUC, 0.60-0.76) models. The results showed that the global mammographic radiomic signature has the ability to predict radiologists' *difficult-to-interpret* normal cases.

Keywords: radiomics, mammography, difficult normal cases, breast cancer, machine learning model, easy normal cases, gist, global radiomic features

1. INTRODUCTION

Breast cancer in women is the most common cancer worldwide [1]. Mammography has long been the only prevalent screening method of breast imaging for an early detection of breast cancer [2]. However, previous studies have reported that mammographic interpretation is not an easy task with many normal mammographic cases have been incorrectly diagnosed as cancers, resulting in up to 16% false positive error rates and up to 63% anticipated cumulative risk of false positive results (in women aged 50-69 after undergoing ten biennial screenings) [3-5]. Of those who were recalled, 80% had to go through unnecessarily further assessment and 40% of those had unwarranted biopsy with benign outcomes [6]. These false positive results have not only caused enormous medical costs on healthcare systems (for example, around USD$3 billion per year in the United State of America) but also significant mental distress and anxiety on patients [7]. Accordingly, ongoing efforts to assess and improve radiologists' performance in interpreting cancer-free normal mammograms through effective educational programs are essential so that false positive findings can be significantly reduced.

Test set based educational programs, often containing almost 70% of the normal cases, have been shown to improve the ability of radiologists when interpreting and identifying normal mammograms [8, 9]. However, the current process of test set curation for difficult normal cases seems to depend greatly on expert radiologists' manual and subjective measurement based on the conventional breast density classification of the Breast Imaging and Reporting Data System (BI-RADS) system and benign features [10, 11]. Studies have shown that the subjective breast density-based evaluation of normal cases difficulty is not a good descriptor of the true difficulty based on individual radiologist's performance in the test set [10, 12]. This demonstrates the need for an alternative uniform approach for predicting normal cases' difficulty so that individuals' performance including false positive errors can be more accurately assessed and improved.

*ssiv6387@uni.sydney.edu.au

Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment, edited by Claudia R. Mello-Thoms, Yan Chen, Proc. of SPIE Vol. 12467, 1246708
 · 1605-7422 · doi: 10.1117/12.2645377

This study aims to investigate if radiologists' *difficult-to-interpret* normal cases can be automatically and objectively described using a global quantitative mammographic signature (i.e., a set of image features) derived from radiomics analysis - the process of extracting, evaluating, and mining high-dimensional image features [13]. Given the value of radiomics in mammography [13] and the importance of global image processing or gist (i.e., radiologist first impression) in guiding visual search and making diagnostic decisions when interpreting and detecting difficult cancer cases [14], it can be assumed that global radiomic signature (i.e., global image statistics) of *difficult-to-interpret* differs from *easy-to-interpret* normal cases. Although previous studies have examined the value of local radiomic signature for predicting radiology trainees' specific tumour or false positive areas [15, 16], minimal effort has been devoted to distinguish radiologists' *difficult-to-interpret* from *easy-to-interpret* normal cases using global radiomic signature. Hence, we attempt to address this deficiency.

2. MATERIALS AND METHODS

1.1 Readers and normal cases

Retrospective non-identifiable data gathered from 342 Australian and New Zealand radiologists who engaged with the BREAST program (https://breast-australia.sydney.edu.au/) [17] from 04 September 2014 to 14 March 2021, interpreting 120 normal mammographic cases of three largest volume of readers BREAST test sets were used. There were 40 normal cases in each test set where each case had four DICOM images (right and left, bilateral craniocaudal/CC and mediolateral oblique/MLO views). These cases were confirmed to be cancer-free normal cases by at least two independent radiologists with negative follow up mammograms obtained 2 years later.

1.2 Difficult normal cases

Based on the objective measure of difficulty (defined as a percentage of incorrect reports per case [18]), we identified the difficulty for our 120 normal cases by firstly calculating difficulty scores for each case using the Royal Australian and New Zealand College of Radiologists rating (scale of 2-5 where 1 represents a normal case [17]) given by the 342 radiologists. A rating of 2 denoted benign lesion while a rating of 3, 4, or 5 indicated malignant lesion with a greater number indicating a higher confidence of malignancy. If the same radiologist provided more than one ratings for the same case, only the highest rating was used. Difficulty scores were computed by dividing the number of incorrect reports (i.e., the number of radiologists who rated a normal case incorrectly as malignancy giving a 3, 4, or 5 rating) by the total number of radiologists of that test set. Next, using one-third of the total 120 normal cases having highest and lowest difficulty scores, 41 cases were classified as difficult-to-interpret and 40 cases were easy-to-interpret, correspondingly. In order to make a clear difference between the highest and lowest difficulty normal cases, only these two categories' cases (a total of 81 cases) were included in the study and 39 cases with intermediate difficulty scores were excluded.

1.3 Radiomic analysis

1.3.1 Global radiomic features extraction

Before global radiomic features extraction were performed, 324 mammography DICOM images of the 81 normal cases (4 images per case) were firstly obtained and used to create binary masks (i.e., black and white) using a standard thresholding value of 100 (including manual adjustment where required) to extract breast region from its background. To achieve computational effectiveness, DICOM images and binary masks were then transformed to TIFF format with all the right CC and MLO images flipped to the left to have a uniform chest wall on the left side. Next, the TIFF images and masks were cropped based on the breast region's maximum size and used for global radiomic features extraction.

Global radiomic features extraction was conducted based on regions of interests delineated using 1). lattice-based/ROI (multiple regions of interests covering the entire breast image) [19], and 2). squared-based/SQ (largest square rectangular box inscribed within breast) [20] approaches (Fig. 1). Given that many of our cases had false positive annotations around the pectoral muscle area, we included pectoral muscle in our radiomic analysis, which is different from the previous approaches [19, 20] where pectoral muscle was excluded.

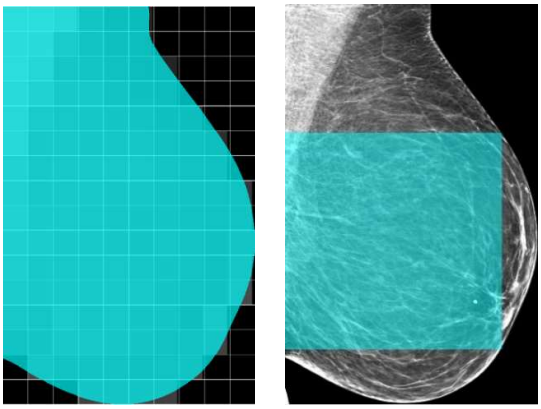


Fig. 1. Left: Lattice-based approach. Right: Squared-based approach. Global radiomic features were extracted from the cyan colour area of the image.

Using our in-house MATLAB-based programs, 34 global radiomic features per image (Table 1) were extracted from 324 images and then normalised using a z-score normalisation technique (standardising image intensity mean to zero and standard deviation to one) [21]. ROI features were computed using MATLAB image distinct block processing technique (block size 214 x 214 pixels) and summed up using the standard deviation method. The extracted global radiomic features can be broadly categorised into three groups: Gray Level Co-Occurrence Matrix/GLCM based Haralick (n=15) [22], First Order Statistics/FOS (n=1) [23], and Neighbourhood Gray Tone Difference Matrix/NGTDM (n=1) [24].

Table 1: Global mammographic radiomic features (n=34)

No.	**Name**	**Feature class**	**Parameters**
1	ROI_Std_Correlation_3	GLCM	3 pixel distance between the pixel of interest and its neighbour
2	ROI_Std_Difference_entropy _3		
3	ROI_Std_Dissimilarity_3		
4	ROI_Std_Difference_variance_3		
5	ROI_Std_Energy_3		
6	ROI_Std_Entropy_3		
7	ROI_Std_Homogeneity_3		
8	ROI_Std_Information_measure_of_correlation1 _3		
9	ROI_Std_Information_measure_of_correlation2 _3		
10	ROI_Std_Maximum_probability_3		
11	ROI_Std_Sum_entropy_3		
12	ROI_Std_Sum_variance_3		
13	ROI_Std_Coarseness_3	NGTDM	3x3 neighbourhood size
14	ROI_Std_Cluster_prominence_3	GLCM	3 pixel distance between the pixel of interest and its neighbour
15	ROI_Std_Cluster_shade_3		
16	ROI_Std_Range_all	FOS	n/a

17	ROI_Std_Sum_of_squares_variance_9	GLCM	9 pixel distance between the pixel of interest and its neighbour
18	SQ_Correlation_9		
19	SQ_Difference_entropy _9		
20	SQ_Dissimilarity_9		
21	SQ_Difference_variance_9		
22	SQ_Energy_9		
23	SQ_Entropy_9		
24	SQ_Homogeneity_9		
25	SQ_Information_measure_of_correlation1 _9		
26	SQ_Information_measure_of_correlation2 _9		
27	SQ_Maximum_probability_9		
28	SQ_Sum_entropy_9		
29	SQ_Sum_variance_9		
30	SQ_Coarseness_9	NGTDM	9x9 neighbourhood size
31	SQ_Cluster_prominence_9	GLCM	9 pixel distance between the pixel of interest and its neighbour
32	SQ_Cluster_shade_9		
33	SQ_Range_all	FOS	n/a
34	SQ_Sum_of_squares_variance_9	GLCM	9 pixel distance between the pixel of interest and its neighbour

Abbreviations:

CC = craniocaudal, FOS = first order statistics, GLCM = gray level co-occurrence matrix, MLO = mediolateral oblique, NGTDM = neighbourhood gray tone difference matrix, n/a = not applicable, ROI = multiple regions of interests defined by the lattice-based approach covering the entire breast image, SQ = largest square rectangular box inscribed within breast, STD = standard deviation.

1.3.2 Model construction

Three binary machine learning classification models (*CC*, *MLO*, *CC+MLO*) for predicting *difficult-to-interpret* from *easy-to-interpret* normal cases were built. *CC* and *MLO* models were built based on the extracted 34 global radiomic features, derived from the respective right and left CC and MLO images only, using MATLAB's random forest approach (ensemble of 500 decision trees boosted with LogitBoost ensemble aggregation algorithm) [25]. The random forest method was selected for this classification task as it has embedded random feature selection in the model which can reduce feature-overfitting issues, producing interpretable models [26]. The last *CC+MLO* model was built based on a late fusion method [27] which involved acquiring the median predictive scores from both *CC* and *MLO* models.

1.4 Statistical analysis and validation

In evaluating the models' performance on our dataset, the models were trained and validated using the leave-one-out-cross-validation method [28], implying that every time the model was trained, all cases were used, except one case which was used as a test set once to validate the predictive performance of the model. This method was repeated 81 times (as we had 81 cases) until every single case was used as a test set once. The area under the receiver operating characteristic curve (AUC) was used to assess the models' overall performance for discriminating *difficult-to-interpret* from *easy-to-interpret* normal cases.

All statistical and radiomics analysis in this study were performed using MATLAB R2021a (MathWorks Inc., Natick, MA, USA).

3. RESULTS

The overall classification performance of the three models (*CC*, *MLO*, *CC+MLO*) is shown in Fig. 2. When combining global radiomic features from both CC and MLO images using the median value of the predictive scores from both *CC* and *MLO* models, the *CC+MLO* model outperformed the *CC* and *MLO* models alone, achieving AUCs of 0.73 (95% Confidence Interval/CI: 0.62-0.83) vs 0.70 (95% CI: 0.62-0.78) and 0.68 (95% CI: 0.60-0.76), correspondingly.

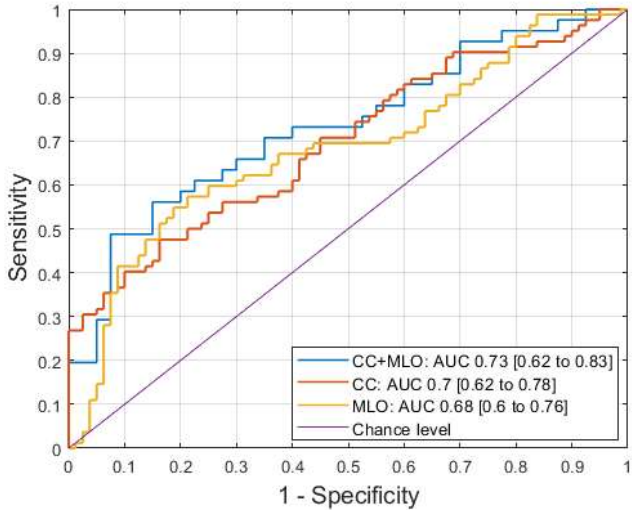


Fig. 2. Performance of the three machine learning models: *CC+MLO*, *CC*, and *MLO* based on the area under the receiver operating characteristic curve (AUC).

4. CONCLUSION

The classifier's performance implied that global mammographic radiomic signature had the ability to predict *difficult-to-interpret* from *easy-to-interpret* normal cases, showing the importance of global image processing in interpreting mammograms.

REFERENCES

[1] Sung H, Ferlay J, Siegel RL et al (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians 71:209-249

[2] Seely JM, Alhassan T (2018) Screening for breast cancer in 2018-what should we be doing today? Current oncology (Toronto, Ont) 25:S115-S124

[3] Román M, Castells X, Hofvind S, Euler-Chelpin M (2016) Risk of breast cancer after false-positive results in mammographic screening. Cancer Medicine 5:1298-1306

[4] Castells X, Torá-Rocamora I, Posso M et al (2016) Risk of Breast Cancer in Women with False-Positive Results according to Mammographic Features. Radiology 280:379-386

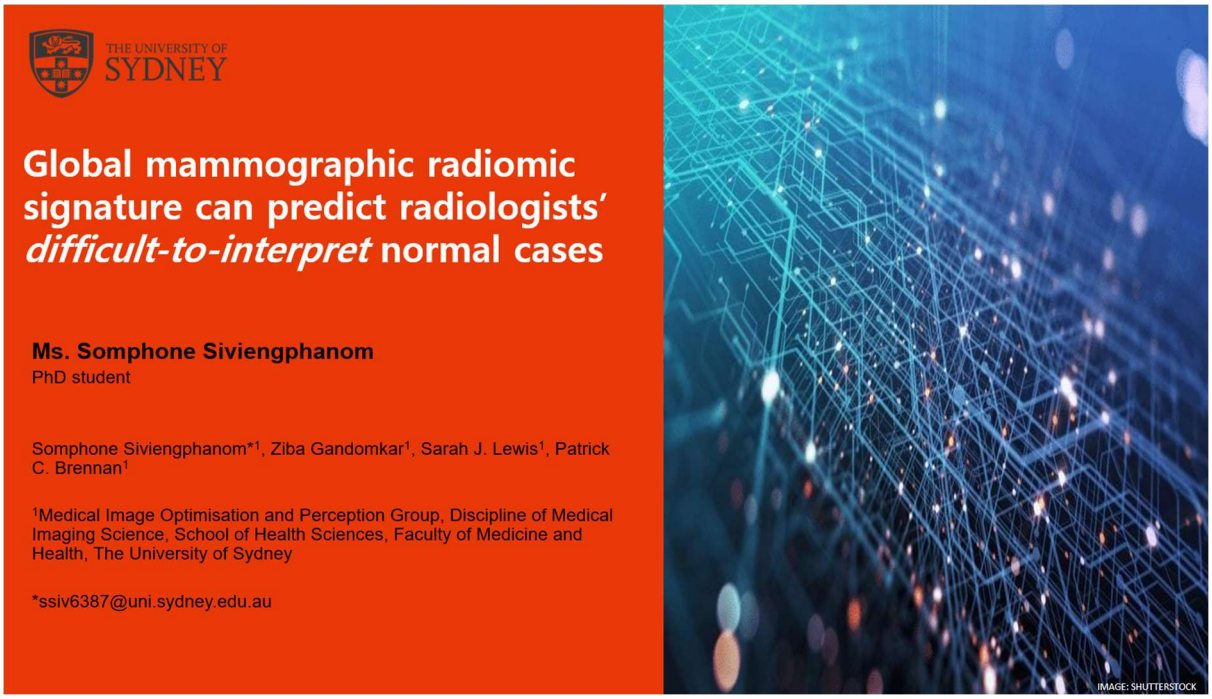
[5] Vlahiotis A, Griffin B, Stavros Md FAT, Margolis J (2018) Analysis of utilization patterns and associated costs of the breast imaging and diagnostic procedures after screening mammography. ClinicoEconomics and Outcomes Research Volume 10:157-167

[6] Ekpo EU, Alakhras M, Brennan P (2018) Errors in Mammography Cannot be Solved Through Technology Alone. Asian Pacific journal of cancer prevention : APJCP 19:291-301

[7] Ong M-S, Mandl KD (2015) National Expenditure For False-Positive Mammograms And Breast Cancer Overdiagnoses Estimated At $4 Billion A Year. Health Affairs 34:576-583

[8] Gale A, Chen Y (2020) A review of the PERFORMS scheme in breast screening. The British journal of radiology 93:20190908
[9] Suleiman ME, Rickard M, Brennan PC (2020) Perfecting detection through education. Radiography 26:S49-S53
[10] Wong D, Gandomkar Z, Reed W, Lewis SJ, Brennan P (2022) The effect on reader performance of case image characteristics including expert determined difficulty ratings using a mammography test set. Under review
[11] Darker I, Chen Y, Gale A (2011) Health professionals' agreement on density judgements and successful abnormality identification within the UK Breast Screening Programme. SPIE
[12] Grimm LJ, Kuzmiak CM, Ghate SV, Yoon SC, Mazurowski MA (2014) Radiology Resident Mammography Training. Academic Radiology 21:888-892
[13] Siviengphanom S, Gandomkar Z, Lewis SJ, Brennan PC (2021) Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs. Academic Radiology 29:1228-1247
[14] Gandomkar Z, Ekpo E, Lewis S et al (2019) Does the strength of the gist signal predict the difficulty of breast cancer detection in usual presentation and reporting mechanisms? Proc. SPIE 10952 Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment, 1095203. SPIE, San Diego, California, United States, pp 1-8
[15] Zhang J, Lo JY, Kuzmiak CM, Ghate SV, Yoon SC, Mazurowski MA (2014) Using computer-extracted image features for modeling of error-making patterns in detection of mammographic masses among radiology residents. Medical Physics 41:091907
[16] Zhang J, Silber JI, Mazurowski MA (2015) Modeling false positive error making patterns in radiology trainees for improved mammography education. Journal of Biomedical Informatics 54:50-57
[17] Brennan P, Lee W, Tapia K (2018) Breast Screen Reader Assessment Strategy (BREAST): A Research Infrastructure with a Translational Objective. In: Samei E, Krupinski EA, (eds) The Handbook of Medical Image Perception and Techniques. Cambridge University Press, Cambridge, pp 343-356
[18] Mazurowski MA (2013) Difficulty of mammographic cases in the context of resident training: preliminary experimental data. Proc. SPIE 8673, Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment 86730W. SPIE, Lake Buena Vista (Orlando Area), Florida, United States, pp 1-6
[19] Zheng Y, Keller BM, Ray S et al (2015) Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment. Medical Physics 42:4149-4160
[20] Gandomkar Z, Suleiman M, Demchig D, Brennan P, McEntee M (2019) BI-RADS density categorization using deep neural networks. Proc. SPIE 10952, Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment, 109520N. SPIE, pp 1-7
[21] Haga A, Takahashi W, Aoki S et al (2019) Standardization of imaging features for radiomics analysis. The Journal of Medical Investigation 66:35-37
[22] Haralick RM, Shanmugam K, Dinstein IH (1973) Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics SMC-3:610-621
[23] Costaridou L (2005) Medical image analysis methods. Taylor & Francis, Boca Raton
[24] Amadasun M, King R (1989) Textural features corresponding to textural properties. IEEE Transactions on Systems, Man, and Cybernetics 19:1264-1274
[25] Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). The Annals of Statistics 28:337-407
[26] Couronné R, Probst P, Boulesteix A-L (2018) Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 19:1-14
[27] Lee SL, Zare MR, Muller H (2019) Late fusion of deep learning and handcrafted visual features for biomedical image modality classification. IET Image Processing 13:382-391
[28] Molinaro AM, Simon R, Pfeiffer RM (2005) Prediction error estimation: a comparison of resampling methods. Bioinformatics 21:3301-3307

Appendix 1.6 Oral presentation on *global mammographic radiomic signature can predict radiologists' difficult-to-interpret normal cases*



This presentation was given at SPIE Medical Imaging Conference 2023, San Diego, CA, United States, February 2023.

This presentation was related to the paper "Global radiomic features from mammography for predicting difficult-to-interpret normal cases" as described in Chapter 4.

Appendix 1.7 Abstract on *enhancing radiology resident training in mammogram interpretation: leveraging global mammographic radiomic features to identify difficult cases*

Title:

Enhancing radiology resident training in mammogram interpretation: leveraging global mammographic radiomic features to identify difficult cases.

Purpose: Radiology residents (RRs) have limited capability for global image processing. To accelerate and enhance trainee learning, identifying global image features in an image that make a normal case difficult to interpret for a trainee is of significant interest to radiology educators. This study therefore investigates whether difficult-to-interpret normal, cancer-free cases for RRs can be distinguished from easy-to-interpret cases using global mammographic radiomic features (GMRFs).

Methods and Materials: A set of retrospective de-identified reading data consisting of 280 normal screening mammographic cases from 137 RRs was utilized. These cases were categorized into difficult- vs. easy-to-interpret normal cases (70 cases each) based on the 75th and 25th percentiles of the cases containing the respective highest and lowest difficulty scores (i.e., percentage of incorrect reports per case). Thirty-four quantitative global radiomic features of images, extracted from a previous study, were investigated for the 140 difficult- and easy-to-interpret categories of normal cases. The GMRFs were then used to build a random forest machine learning classifier for predicting difficult- from easy-to-interpret normal cases of radiology trainees. The model was trained and validated using leave-one-out-cross-validation approach. The performance of the model was assessed using the area under the receiver operating characteristic curve (AUC).

Results:
The model achieved an AUC of 0.75 (95% CI: 0.67-0.83), indicating the effectiveness of GMRFs in distinguishing difficult from easy-to-interpret cases for RRs. Significant differences in 15 out of 34 features were observed, with cluster prominence and range emerging as key predictors.

Conclusion:
GMRFs show promise in improving and accelerating RRs' interpretation skills and facilitating personalized training programs for mammogram interpretation. This approach may reduce false positive errors and enhance RRs' expertise in breast cancer screening.

Clinical Relevance:

Global mammographic radiomic features offer a valuable tool for facilitating the tailoring radiology resident's education and help build required global image processing skills.

This abstract's presentation will be given at RSNA 2024 Annual Meeting, 1-5 December, McCormick Place, Chicago, Illinois, United States.

This presentation was related to the paper “A machine learning model based on global mammographic radiomic features can predict which normal mammographic cases radiology trainees find most difficult” as described in Chapter 5.