Added benefits of computer-assisted analysis of Hematoxylin-Eosin stained breast histopathological digital slides

Ziba Gandomkar, MSc

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Faculty of Health Sciences

The University of Sydney

December, 2017

Candidate's Statement

I, Ziba Gandomkar, hereby declare that the work contained within this thesis is my own and has not been submitted to any other university or institution as a part or a whole requirement for any higher degree.

I, Ziba Gandomkar, hereby declare that I was the principal researcher of all work included in this thesis, including work published with multiple authors.

In addition, ethical approval from the University of Sydney Human Ethics Committee was granted for the three studies presented in this thesis. Participants were required to read a participant information document and informed consent was gained prior to data collection.

I, Ziba Gandomkar, understand that if I am awarded a higher degree for my thesis entitled "**Added benefits of computer-assisted analysis of Hematoxylin-Eosin stained breast histopathological digital slides**" being lodged herewith for examination, the thesis will be lodged in the University Library and be available immediately for use. I agree that the University Librarian (or in the case of a department, the Head of Department) may supply a photocopy or microform of the thesis to an individual for research or study or to library.

Name Ziba Gandomkar

Date 10/12/2017

Preface

The University of Sydney allows thesis containing publication. This thesis consists of ten chapters and encompasses the candidate's published papers, papers under consideration for publication, the bridging chapters for the papers, along with introduction, discussion and conclusion chapters, as instructed in the University of Sydney guideline for thesis containing publications. Each chapter can be read independently as it is presented as a self-reliant section and includes its own references. The thesis layout is shown below.

- **Chapter 1** is an introduction to the thesis. It provides an overview of the added benefits of whole slide imaging to breast pathology, summarizes the 'gaps' in the existing literature and explains the objectives to be addressed by this thesis.
- **Chapter 2** presents a detailed review of the literature on computer‑based image analysis in breast pathology. It discusses, compares and contrasts the previous studies to find key remaining challenges in computer-assisted analysis of breast histopathological images. This chapter was published as the review paper "Computer-based image analysis in breast pathology" in the Journal of Pathology Informatics, 7:43, 2016.
- **Chapter 3** serves as a bridging chapter for the paper presented in chapter 4 and provides a detailed background about mitotic figures, their importance, and the magnitude of disagreement among pathologists in the recognition of mitotic figures. It also briefly justifies the necessity of study presented in Chapter 4.
- **Chapter 4** presents the published journal paper "Determining image processing features describing the appearance of challenging mitotic figures and miscounted nonmitotic objects," which was published in the Journal of Pathology Informatics, 8:34, 2017. It explores the relationship between image-based features and the difficulties in the recognition of mitotic figures.
- **Chapter 5** is a bridging chapter for introducing the paper presented in Chapter 6 and provides a detailed background about nuclear grading, its importance, and challenges toward a reproducible nuclear grade.
- **Chapter 6** presents COMPASS (COMputer-assisted analysis combined with Pathologist's ASSessment), which is a personalized tool for reproducible nuclear atypia scoring and it is based on the pathologist's assessment of six

criteria related to the nuclear atypia along with computer-extracted features. This paper is currently being peer-reviewed for publication.

- **Chapter 7** serves as a bridging chapter for the study presented in chapter 8. It discusses the importance of the correct identification of carcinoma and benign subtypes and the necessity of the study presented in Chapter 8.
- **Chapter 8** introduces MuDeRN, a framework for classifying Hematoxylin-Eosin stained breast histopathological images either as benign or cancer; categorising cancer cases into four subclasses, namely ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma; and subdividing those cases classified as benign into four subcategories, namely adenosis, fibroadenoma, phyllodes a tumour, or tubular adenoma. This paper is currently being peer-reviewed for publication.
- **Chapter 9** provides an overview of the main findings of this thesis and their interpretation. It also discusses the findings in the context of literature and states the implications of the findings. Finally, it identifies the limitations of the studies and possibilities for future work.
- **Chapter 10** succinctly summarizes this thesis, its findings, and implications.

I used publicly available de-identified databases throughout the studies included in this thesis, therefore the studies were exempt from the ethical approval by the University of Sydney.

Abstract

Aims: This thesis aims at determining if computer-assisted analysis can be used to better understand pathologists' perception of mitotic figures on Hematoxylin-Eosin (HE) stained breast histopathological digital slides. It also explores the feasibility of reproducible histologic nuclear atypia scoring by incorporating computer-assisted analysis to cytological scores given by a pathologist. In addition, this thesis investigates the possibility of computer-assisted diagnosis for categorizing HE breast images into different subtypes of cancer or benign masses.

To achieve these aims, this thesis 1) examines the existing literature regarding computer-assisted analysis of HE images in breast pathology to identify knowledge deficiencies; 2) assesses the feasibility of relating image-processing features with disagreement in recognition of mitotic figures among pathologists; 3) proposes a tool for reproducible grading of nuclear atypia on HE breast images; and it 4) proposes a tool for automatic classification of HE breast images in multiple categories.

Materials and Methods: This thesis is comprised of three original research studies.

Study 1: A data set of 453 mitoses and 265 miscounted non-mitoses within breast cancer digital slides were considered. The MITOSIS-ATYPIA dataset, which is a publicly available dataset, was used in this experiment. In this dataset, two pathologists were asked to annotate the mitotic figures and label them either as a "true mitosis" or "probably a mitosis". In case of disagreement, the opinion of a third pathologist was requested. Based on the confidence level of three pathologists who annotated the mitoses, they were classified in three groups: those recognized by both two first pathologists (C1), those missed by one of the first two pathologists and recognized as a "true mitosis" by the third pathologist (C2), and those labelled as "probably a mitosis" by the majority of the readers (C3). The miscounted non-mitoses were annotated as a mitosis by only one of the pathologists, whereas the true mitosis were annotated by at least 2 of the pathologists. Shape-based, intensity-based and textural features were extracted from the objects in different channels of eight colour spaces. Two global descriptors representing the size of nuclei and chromatin density of each image were also extracted. Using Kruskal-Wallis H-test followed by the Tukey-Kramer test, the significantly different quantitative features among three categories of

mitotic figures and miscounted non-mitoses within the breast slides were identified. The study also extracted some rules, describing a trend which was observed from C1 (easily identifiable mitoses) to C3 (the most challenging mitoses).

Study 2: A new tool for reproducible nuclear atypia scoring in breast cancer histological images was proposed in this study. The new tool was tested on 600 images for which expert-consensus derived reference nuclear atypia scores were available. The images were acquired from 300 areas, once scanned by Aperio Scanscope XT scanner and once by a Hamamatsu Nanozoomer 2.0-HT scanner. The developed tool is called COMPASS (COMputer-assisted analysis combined with Pathologist's ASSessment) and it relied on two sets of features, where the first set is comprised of the scores given by the pathologists to six nuclear atypia-related cytological features, and the second set contained computer-extracted features. COMPASS was designed to assist junior pathologists to achieve an accuracy in nuclear grading comparable to that of senior pathologists. It was retrospectively tested for three junior pathologists who gave scores to six atypia-related criteria for each image.

Study 3: The third study proposed and tested MuDeRN (MUlti-category classification of breast histopathological image using DEep Residual Networks), which is a framework for classifying hematoxylin-eosin stained breast digital slides either as benign or cancer, and then categorizing cancer and benign cases into four different subtypes each. MuDeRN provided the diagnosis for each patient by combining outputs of deep residual networks' processed images in different magnification factors using a meta-decision tree. Images for each patient were heterogeneous and a meta-decision tree could potentially capture the nonlinearity for mapping the image-level diagnosis to the patient-level diagnosis. MuDeRN's performance was tested on a dataset of 7786 images from 81 patients where each patient had images at four magnification factors (x40, x100, x200, and x400) available. Images per each patient were heterogeneous, therefore for obtaining patient-level diagnosis from the image-level diagnoses instead of simple averaging or majority voting,

Results:

Study 1: It was found that the most challenging mitotic figures (C3) were smaller and rounder compared to other mitoses (C1 and C2). On the other hand, the sizes of the

miscounted non-mitoses were identical to those of easily to identify mitoses (C1 and C2) but miscounted non-mitoses were rounder than true mitoses. Compared to intensity-based features, textural features exhibited more differences between challenging mitotic figures (C3) and the easily identifiable mitoses (C1), while the intensity-based features from chromatin channels were the most discriminative features between the miscounted non-mitoses and the easily identifiable mitotic figures (C1). Among the texture features, features extracted using Gabor filter were the most discriminative features.

Study 2: The percentage agreement between the reference nuclear scores (consensus of pathologists) and COMPASS, if it had been adopted by the three junior pathologists, was 93.8%, 92.9%, and 93.1% respectively. The agreement rates were comparable to those of senior pathologists assessing the same dataset (i.e. 90.6% and 86.9%). The paired Mann-Whitney U test showed that the grades given by COMPASS when tested on Aperio images and the given grades for Hamamatsu images were not significantly different (junior pathologist 1: z=-1.1, P=0.29; junior pathologist 2: z=0.48, P=0.63; junior pathologist 3: z=0.86, P =0.39).

Study 3: For the malignant/benign classification of images, MuDeRN obtained correct classification rates (CCR) of 98.52%, 97.90%, 98.33%, and 97.66% in x40, x100, x200, and x400 magnification factors respectively. For eight-class categorization of images, CCRs were 95.40%, 94.90%, 95.70%, and 94.60% in x40, x100, x200, and x400 magnification factors respectively. For making patient-level diagnosis in eight-class categorization, MuDeRN obtained a CCR of 96.25%.

Conclusion: The findings from the first research study suggested that computer-aided image analysis can provide a better understanding of image-related features related to discrepancies among pathologists (the mitoses recognition task was evaluated). Two tasks done routinely by the pathologists are making diagnosis and grading the breast cancer. The second and third studies indicated that computer-assisted analysis can aid in both nuclear grading (COMPASS) and breast cancer diagnosis (MuDeRN). Therefore, three important tasks in breast pathology could benefit from the findings presented in this thesis. The results could be used to improve current status of breast cancer prognosis estimation through reducing the inter-pathologist disagreement in counting mitotic figures and reproducible nuclear grading. It can also improve

providing a second opinion to the pathologist for making a diagnosis and hence reduce diagnostic discrepancies among pathologists.

Acknowledgments

I would like to extend thanks to the many people who so generously contributed to the work presented in this thesis. Special mention goes to my enthusiastic supervisor, A/Professor Claudia Mello-Thoms. My PhD has been an amazing experience and I thank Claudia wholeheartedly, not only for her tremendous academic support, but also for giving me so many wonderful opportunities during my PhD journey. Similar, profound gratitude goes to Professor Patrick Brennan as my secondary supervisor, who has been a truly dedicated mentor, who shares his expertise so willingly

I am also hugely appreciative to the contributors of the Mitosis-Atypia challenge 2014 and BreakHis databases, for kindly providing us the access to their database. I also acknowledge the University of Sydney HPC Service at the University of Sydney for providing high performance computing resources that have contributed significantly to the research results reported within this thesis.

Special mention goes to the MIOPeG members for all their support and sharing their knowledge.

I also owe much gratitude to my lovely sister, Fariba, who never lost faith in me. Fariba, you always encouraged me and was the source of motivations! I also thank Hossein, my marvellous brother, and Nayyereh, my super wonderful sister-in-law, who sent me photos of my cute nephew, Heedar, whenever I was tired and needed energy to be able to keep going. A very special thank you to Muhammad, my amazing, awesome brother, who always patiently listened to my dreams and understood me. Finally, but by no means least, thanks go to Mom and Dad for almost unbelievable support. They are the most important people in my world. Mom and Dad, thank you so much for everything! I never would have been able to succeed without you!

This dissertation is dedicated to my husband, Navid, who has been a constant source of support and encouragement during the challenges of the PhD and life. I am truly thankful for having you in my life. Thank you, Navid, for your love and being my best friend. I owe everything to you. How else could I have made it through my PhD without having somebody to complain to every night? Thank you for making me have the most confidence I have ever had in my entire life.

Table of Contents

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Chapter 9

Parts of the work presented in this thesis have been published and/or presented in the following forums:

Journal papers

[1] Gandomkar, Ziba, Patrick C. Brennan, and Claudia Mello-Thoms. “Computer-based image analysis in breast pathology.” Journal of pathology informatics 7:43, 2016.

[2] Gandomkar, Ziba, Patrick C. Brennan, and Claudia Mello-Thoms. “Determining quantitative features describing appearance of challenging mitotic figures and miscounted non-mitotic objects.” Journal of pathology informatics 8:34, 2017.

[3] Gandomkar, Ziba, Patrick C. Brennan, and Claudia Mello-Thoms. “COMPASS: Nuclear Atypia Scoring of Breast Cancer by Computer-Assisted Analysis Combined with Pathologist’s Assessment.” Submitted to Journal of Digital Imaging, 2018. Under review.

[4] Gandomkar, Ziba, Patrick C. Brennan, and Claudia Mello-Thoms. “MuDeRN: Multi-category Classification of Breast Histopathological Image Using Deep Residual Networks.” Artificial Intelligence in Medicine Journal, 2018.

Conference presentations

[1] Z. Gandomkar, P. C. Brennan, and C. Mello-Thoms, “Identifying quantitative features describing challenging mitotic figures," Sydney Cancer Conference 2016, Sydney, Australia. (Abstract)

[2] Z. Gandomkar, P. C. Brennan, and C. Mello-Thoms, "Determining local and contextual features describing appearance of difficult to identify mitotic figures," in SPIE Medical Imaging, 2017, pp. 1014002-1014002-8. (Full paper)

[3] Z. Gandomkar, P. C. Brennan, and C. Mello-Thoms, “A Framework for Multi-class Categorization of Breast Histopathological Images Using Deep Residual Networks," 56th Australian Society for Medical Research National Scientific Conference 2017, Sydney, Australia. (Abstract)

[4] Z. Gandomkar, P. C. Brennan, and C. Mello-Thoms, “Nuclear Atypia Scoring by Combining Pathologist’s Assessment and Computer-Assisted Analysis," presented in BreastScreen conference 2018, Adelaide, Australia. (Abstract).

[5] Z. Gandomkar, P. C. Brennan, and C. Mello-Thoms, “A Framework for Detection of Malignant Cases Based on Breast Histopathological Images Using Deep Residual Networks," International Workshop on Breast Imaging 2018, Atlanta, Georgia, USA. (Full paper)

Chapter 1

Introduction

Breast cancer (BCa) is the most common non-skin cancer among women worldwide [1]. In spite of the increase in the incidence rate of BCa over the last few decades, the mortality rate from this disease in the developed countries has been decreasing due to improvements in treatment options [1] and early detection through screening mammography [2]. Mammography is the standard imaging examination for BCa screening and randomized clinical trials conducted between 1970 and 1990 supported its efficacy as a population-based organized BCa screening tool [3-5].

Radiologists might recall women attending screening for further imaging or a biopsy. In the USA, Elmore et. al (2015) estimated that 49% of women screened annually for a ten year period will experience at least one false-positive mammogram, and 19% will undergo a breast biopsy unnecessarily [6]. In another study, it was shown that in each round of a screening program, 10.6% of women with false-positive mammograms undergo fine needle aspiration or breast biopsy [7]. In the UK 4% of women attending screening mammography were called back for further examinations, and in total 1.76% of screened women undergo a biopsy [8]. Therefore, each year, pathologists evaluate a large number of breast histopathological slides, from which only one in four contains malignancy, and benign lesions and normal biopsies are far more prevalent [9]. Approximately 1.6 million women in the United States have breast biopsies each year [10]. In breast pathology, pathologists are responsible for different tasks such as determining whether a given lesion is benign or malignant, staging of BCa, determining cancer type, identifying the subtype of a benign lesion, grading, assessing surgical margins, and biomarker testing [11]. The tasks done by pathologists in routine clinical practice while interpreting the breast slides are illustrated in Figure 1. The pathologists' tasks which are related to this thesis are shown by check mark.

When a malignant mass is present, the pathologist should also stage the BCa. For staging BCa, pathologists evaluate the mass size and determine if the cancer is present in the lymph nodes (that is, determine whether the cancer has metastasized) [11]. Normally, cancer type is also stated in the pathology report. Pathologists also grade the cancer, as the cancer grade shows the tumour's aggressive potential [11]. Different grading methodologies have been proposed [11, 12]. Different cytological and histological components are taken into account in each one of these grading systems. Different grading systems, their contributing factors, and their reproducibility have

been discussed in Chapter 5. Among these grading systems, Scarff-Bloom-Richardson grading system (Nottingham grading system) is one of the most popular [12] as it is recommended in the US National Comprehensive Cancer Network (NCCN) guidelines[1]. This system considers three factors: the percentage of tumour area forming glandular/tubular structures, nuclear pleomorphism (changes in nuclear appearance), and number of mitoses per 10 high power fields [12].

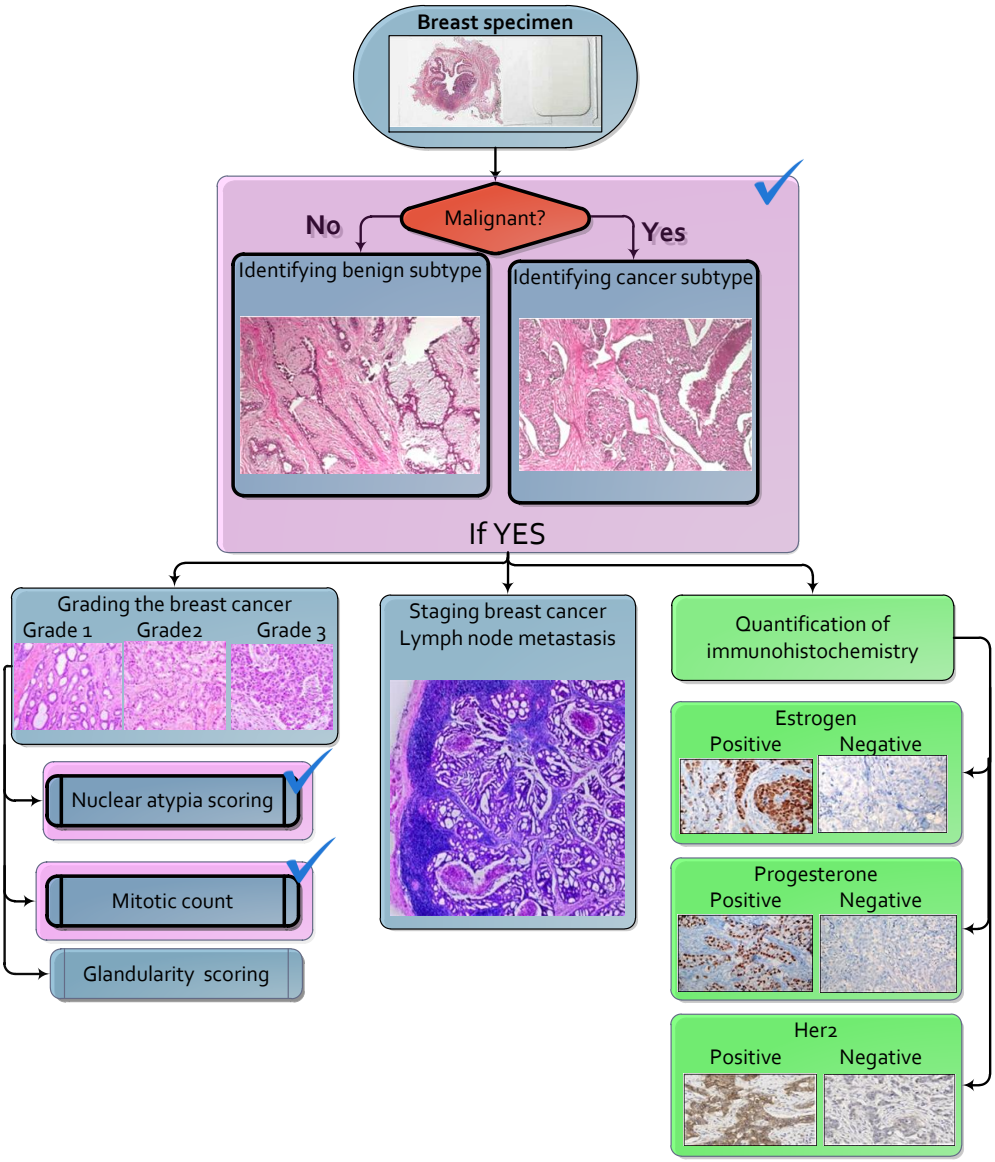


Figure 1- Tasks done by pathologists while interpreting the breast slide; the tasks covered in this thesis are shown by check mark.

[1] http://www.nccn.org/professionals/physician_gls/pdf/breast.pdf

Finally, the pathologist examines the expression of the biomarkers in BCa samples [11]. In BCa patients, three important biomarkers are evaluated: the estrogen receptor (ER), progesterone receptor (PR), and HER2 receptors. ER and PR statuses predict whether the patient can benefit from endocrine therapy or not [11]. An overexpression of HER2 shows a higher risk of cancer recurrence and predicts that patients can benefit from anthracycline and taxane-based chemotherapies but not endocrine therapy [13].

Usually, the pathological report of a breast biopsy is considered as the gold standard for further patient management and selection of the treatment options. However, recent studies have shown that there are disagreements among pathologists interpreting breast specimens. Table 1 summarizes findings of studies investigating discordance among pathologists in making a diagnosis about breast slides. Usually, an expert consensus review panel is considered as the gold standard and disagreement with the expert-driven consensus can lead to overinterpretation (overdiagnosis) or underinterpretation (undertreatment).

As shown in table 1, the expert breast pathologists have a high agreement rate in the diagnosis of invasive BCa, but the disagreement rates for the diagnosis of benign lesions and of atypical lesions can be high. The concordance level is usually measured using the agreement rate or the Cohen's kappa. The agreement rate is the number of concordance cases divided by total number of cases. Cohen's kappa measures inter-observer agreement for categorical items and is a more robust measure than the agreement rate, as it considers the possibility of the agreement happening by chance. It usually interpreted as follows: kappa$\leq$ 0 shows no agreement, 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as perfect agreement.

The high disagreement rates for the diagnosis of benign and atypical lesions is concerning as the treatment follow-up that each diagnosis would have received is different [26]. Considering the percentages from the studies presented in Table 1, if 100,000 core biopsies are performed per year, about 4,000-9,000 of them will lead to diagnosis of atypia [26]. Based on the numbers provided in [26] and [27], about half of these cases may lead to overdiagnosis and unnecessary invasive treatments.

Table 1-Findings of studies investigating discrepancies among pathologists making diagnosis of breast specimens.

Included categories	Study	N_C	N_P	Finding
Usual hyperplasia, atypical hyperplasia, or carcinoma in situ	[25]	24	6	Complete agreement among all six pathologists was seen in 58% of cases; five or more agreed for 71% of cases, and four or more arrived at the same diagnosis for 92% of cases.
Benign, benign with atypia, non-invasive malignant, and invasive malignant	[26]	30	26	Overall kappa of 0.71; kappa of 0.95 for malignant/benign classification; and kappa was nearly perfect for selection of benign versus malignant categories. There was less agreement for the categories of non-invasive malignant and benign with atypia (kappa coefficients of 0.59 and 0.22, respectively).
Benign with and without atypia; ductal carcinoma in situ, and invasive BCa	[22]	1970	12	A significant discrepancy of 3.35% for histologic classification
Benign with and without atypia; ductal carcinoma in situ, and invasive BCa	[27]	240	115	Overall agreement rate of 87% for benign without atypia; overall agreement rate of 48 %, with 17% overinterpretation and 35% underinterpretation for benign with atypia; overall agreement rate of 84 %, with 3% overinterpretation and 13% underinterpretation for ductal carcinoma in situ cases; overall agreement rate of 96% for invasive cases.

N_c represents number of cases while N_p represents number of cases.

Table 2 lists the findings of studies investigating the magnitude of disagreement among pathologists in grading BCa slides. As stated earlier, different grading systems have been proposed for grading BCa, however since the most popular is Nottingham grading system, I only included the inter-pathologists' studies using this grading system. As shown in the table, agreement is poorest for nuclear pleomorphism grade (fair to moderate) and it is strongest for tubular formations (substantial to perfect); and it is moderate to substantial for mitotic count. Studies presented in Table 2 used Cohen's kappa, however, it should be noted that as breast cancer grade is an ordinal variable, weighted Cohen's kappa represent the degree of disagreement better than Cohen's kappa. Disagreement among pathologists can also happen in biomarker reporting. Based on [22], biomarker profile was the second most common item with significant discrepancy (50 cases out 1970) between initial and second review pathology reports. The highest agreement was observed for HER2 grading, while ER-status of 19 patients and PR- status of 20 patients were changed.

Different reasons could cause disagreement among the pathologists. Allison et al. [23] divided underlying reasons into three categories, which were pathologist-related, diagnostic coding/study methodology-related, and specimen-related. Among pathologist-related factors, "professional differences of opinion on features meeting diagnostic criteria" was ranked first [23]. Diagnostic coding-related root causes were mostly miscategorizations of descriptive text diagnoses while specimen-related root causes included poor slide quality, artefacts, and limited diagnostic material [23]. Recent advances in digital scanners can potentially help in further understanding of the underlying reasons for discrepancies, and providing computerized tools for giving second opinions to the pathologists.

Table 2-Findings of studies investigating disagreement and agreement among pathologists for BCa grading

Study	N_C	N_P	Finding
[14]	50	5	Approximately 80% complete agreement was achieved for tubule formation, nuclear score, and mitotic count, with kappa values ranged from 0.46 to 0.69.
[15]	40	3	Pairwise kappa for agreement ranged from 0.68-0.83 (median 0.68) for overall grade. Kappa values were 0.54, 0.34 and 0.36 for tubule formation, nuclear pleomorphism and mitotic count respectively.
[16]	93	7*	Agreement rate of 31% in overall grade with kappa of 0.54; the agreement was best for tubular formations and poorest for nuclear grade.
[17]	35	13	Kappa ranged from 0.5 to 0.7, with the greatest agreement obtained in categorizing grade I (kappa=0.7), and grade III (kappa=0.7) tumours.
[18]	166	3	Overall agreement rate for grading of 72.3% of all cases
[19]	72	6	Pairwise kappa values from 0.43 to 0.74 for histologic grade. Generalized kappa values were 0.64, 0.52, and 0.4 for tubule formation, mitotic count, and nuclear pleomorphism.
[20]	10-23	5-7	Kappa ranged from 0.50 to 0.59 for overall grade. Pairwise kappa was the lowest for nuclear pleomorphism (kappa=0.37–0.50), highest for tubularity (kappa=0.57–0.83), and intermediate for mitotic count (kappa=0.45–0.64).
[21]	50	5	The polychoric correlations** among observers were 0.803, 0.712, 0.797 and 0.602 for the final grade, tubule formation, nuclear pleomorphism and mitotic figures, respectively. There were significant differences in thresholds and hence significant differences in classification of grades.

* seven pathology departments within the southern healthcare region of Sweden
** An estimate of the correlation between two normally distributed continuous variables

1-2- Whole slide imaging in breast pathology

1-2-1-What is whole slide imaging?

Whole slide imaging (WSI) refers to the scanning of an entire glass slide with high magnification and producing “digital slides” or “virtual slides” [24]. Pathologists use

monitors instead of light microscopes for assessing the digital slides and the WSI systems provide the pathologists with different options for annotating the slides, send them to a colleague for consultation, searching the database, and retrieving similar slides.

WSI involves software for data acquisition, archiving, slide viewing, and image processing and hardware for scanning the glass slide [28]. In late 1990s, the first virtual microscope was introduced. Although it was a major step towards digital pathology, the capabilities of the early virtual microscopes were limited, especially due to the length of time it took to scan a slide [28]. Today, most modern slide scanners are capable of producing high-resolution digital slides in a reasonable time, and many pathology labs are starting to undergo a transition from traditional workflow to fully-digital [29]. At first impression, the transition from glass slide to digital slides might seem to be like elimination of films in radiology. However, there are some fundamental differences between these two and transition toward digital pathology has its own considerations and benefits for the pathology department [28]. This section summarizes these advantages and considerations.

1-2-2- Advantages of WSI in breast pathology

WSI has the potential to be utilized in telepathology for primary diagnosis [15-41] and quality assurance (QA) [22, 42-48], clinical education [50-57], data management [60-63], and digital image analysis to aid pathologists [28, 64, 65]. In this section, the potential added advantages of WSI adoption in breast pathology are briefly discussed.

1-2-2-1- Telepathology for primary diagnosis

Telepathology can be used for diagnosis in remote areas and can eliminate cost and delays associated with posting a glass slide from areas without in-site pathologists as well as reduce pathologists' travels to remote area for reading slides. However, before broad adoption of digital pathology, some challenges should be tackled, such as improving the speed of the scanners, enhancing auto-focusing systems, and also developing appropriate software for data management in pathology labs. However, the most important challenge for adoption of WSI systems for primary diagnosis is proving that the performance of pathologists while using WSI is at least as good as their performance with light microscopy.

Recent studies showed that pathologists' performance in reading breast slides while using WSI platforms was comparable to conventional microscopy in BCa grading [17], benign/malignant determination [22], mitotic activity scoring [23],quantification of ER and PR [18,20], HER2 scoring [15,16,19,21,24] and also Ki-67 assessment [21,24]. However, some differences between conventional and virtual microscopy were reported in the studies as well; for example, in [30], it was reported that WSI had better sensitivity and lower specificity for HER2 scoring in comparison with light microscopy and Kondo et al [31] showed that the pathologists tended to assign higher HER2 scores with WSI than with glass slides [31]. Interestingly, Shaw et al [32] showed that performance with WSI is better for detecting tubule formation [32] and in [33], Cohen's kappa (κ) was used for measuring the agreement between the pathologists while assessing ER and PR expression in virtual slides. The κ between conventional and virtual microscopy ranged from 0.33 to 0.78 and the results showed that digitized evaluation of nuclear immunostaining was as precise as the manual evaluation of routine glass slides.

Recent advances in telepathology also made robotic microscopy possible. A robotic microscopy system is a fully-automated complex system which allows for the production of digital slides from tissue biopsies; it consists of a WSI system for producing and storing the slides, electromechanical part handling specimens, a pump for immersion oil (if necessary for producing slides), controls, and support accessories [34]. For example in [35], Singh et al investigated the feasibility of remote diagnosis of BCa slides using robotic microscopy. Although the agreement between diagnosis using light microscopy and robotic microscopy was promising, it has been stated that adoption of robotic microscopy was not feasible at the time of the study due to operational and technical problems.

Terpe et al used telepathology for an intraoperative frozen section service and showed that the error rate was similar to light microscopy [36]. However, in [37], it was reported that using telepathology for intraoperative analysis of the sentinel lymph node by frozen section caused decrease in sensitivity. As both studies had large number of patients (298 vs 628), lack of power is not the source of differences. Differences in the slide scanners' model or pathologists' level of expertise may be the reason of the observed behaviour.

Although the results obtained from comparison of diagnosis based on glass slides to digital slides were promising, certain limitations must be taken into account. First, in order to design a study for validation of WSI systems in breast histopathology, a test-set containing different types of cancers, benign, and normal tissues should be used. As suggested by [38], the magnitude of agreement between decisions made on glass slides and digital slides may differ for various biomarkers (or more generally tasks) and this suggests that a validation study for a specific task cannot be simply generalized to other pathology tasks. However, there are more than 10,000 possible diagnoses that a pathologist can render [39], and assessing all of them is not feasible. Rather than including all possible diagnoses in a test-set, I can make sure that the diagnostic features that comprise each diagnoses are properly represented in the WSI. Using pathologists from only one centre is the second limitation of the reviewed studies. Only Nassar et al [40] did a study in three different sites. They achieved comparable percentages of agreement between light microscopy and WSI across different sites [40]. Moreover, in a perfect study design, intra-observer variability should be considered as well. To do so, Campbell et al [41] asked the pathologists to examine the cases using conventional microscopy twice, with one reading considered as the original, and calculated intra-observer variability of the second reading using WSI [41]. Particularly when the pathology task is more subjective, such as detecting pleomorphism, stroma, the nature of the tumour border and lymphocytic infiltration, considering readers' variability while reading glass slides is crucial [32]. In summary, it can be concluded that WSI has the potential to be utilized in telepathology for primary diagnosis [15-41].

1-2-2-2- Telepathology in breast pathology quality assurance and consultation

Applications of telepathology for teleconsultation and QA were also assessed recently. The discrepancies between the interpretations of two different pathologists have been studied [22], and a high disagreement rate was observed, especially in classification of borderline cases, emphasizing the key role of QA in pathology labs. One of the major barriers to QA is problems associated with shipping glass slides between facilities, which raises risk of damage, time delays, and transporting costs. Telepathology allows fast inter-institutional QA.

Leong et al [42, 43] performed a clinical trial to investigate the accuracy of telepathology in the U.K. breast screening pathology QA program. The diagnostic accuracy, invasive tumour typing, tumour grading of telepathology compared to light microscopy was 98.8%, 91.3%, and 86.4% respectively. The study concluded that because of comparable results, telepathology can be used for QA purposes [42, 43]. In another study, Zito et al [44] used an internet-based platform to evaluate inter-observer reproducibility between pathologists using virtual slides (VS) of stereotactic core biopsy specimens of non-palpable breast lesions. The study showed similar results to those quality control studies using circulating glass slides and it concluded that telepathology can be used for quality control [44]. In [45], Terry et al presented the Canadian Immunohistochemistry Quality Control, which is a web-based program for QA developed in Canada. They showed that the telepathology can be used for inter-instantiation quality assurance and consultation.

Della Mea et al [46] performed one of the earliest studies on teleconsultation. It has been shown that digital pathology is effective for remote consultation [46]. However, they used only 48 cases, which is a limited number. In a similar study done in Germany, the impact of web-based service for teleconsultation in pathology was investigated [47]. The result indicated that the quality of telepathological diagnosis was as good as that of conventional diagnosis. In addition, it was reported that by using telepathology, the response was 1 to 2 days faster, as it avoided the delay of conventional post [47]. Later, the updates to the software for tackling some technical problems were presented in [48]. In another study done for evaluating performance of a WSI-based same-day second-opinion service, it was shown that in only 1.3% of the cases the original glass slide was requested to make a final decision [49]. In summary, the previous studies provided very strong evidences for supporting use of telepathology for quality assurance and consultation.

1-2-2-3- Virtual microscopy in breast pathology education

WSI systems have a lot of potential applications in clinical education because they provide the readers with different options for annotating, searching the database, and retrieving similar slides [50-52]. For example, Lundin et al [50] developed an educationally useful publicly available atlas of breast histopathology by using web based virtual microscopy technology. The user can see the virtual slides either with

supplementary diagnostic information or without it for self-assessment [50]. Khushi et al [51] also developed an open source tool for virtual slides management and archiving which can be used for educational purposes [51]. Bondi et al [52] used an e-learning platform (Docebo) to archive digital slides and showed that WSI systems are appropriate for proficiency tests and case sharing for consultation with more experienced colleagues [52].

One of the most important things in training the pathology residents is finding the best training methods for the residents to develop their diagnostic skills [53]. Recently a few studies analysed the gaze patterns of pathologists and residents while interpreting breast cancer virtual slides to understand the development of visual perception skills [53-57]. The results of this type of study can be used for optimizing clinical training of pathology residents.

1-2-2-4- Facilitation of Data Management

The advent of WSI systems could facilitate data management in breast pathology through faster and easier slide archiving, indexing, and retrieving. Recently, different online and offline tools were developed for storing, indexing, and content-based retrieval of breast virtual slides. For example, Schnorrenberg et al [58] developed the Biopsy Analysis Support System (BASS), which is software for indexing and content-based retrieval of breast digital slides [58]. As another example, in [59] Pathology Analytic Imaging Standards (PAIS), a data model was presented and a database was implemented based on it to manage data and retrieve slides relevant to the sent query. Zheng et al [60] developed a method for content-based slide retrieval from an archive of breast virtual slides. This software is capable of finding slides with the spatial texture property similar to the one queried [60]. INSPIRE is a web-based integrated informatics interface for aggregating annotation data of digital slides to perform and present statistical analyses [61]. Wright et al [62] developed RandomSpot which is a web-based tool for systematic random sampling of virtual slides [62].

One of the major issues of data management software for storing or transmission of pathology slides is the extremely large size of digitized slides. In [63], the usefulness of a visual discrimination model (VDM), as well as other distortion metrics for predicting the bit rates for visually lossless compression of breast digital slides, was investigated. It has been shown that VDM metrics could be utilized as a guide for

determining the compression rate of breast virtual slides and they reduced the data size 5–12 times of reversible compression methods [63].

1-2-2-5- Possibility of computer-assisted analysis

One of the major advantages of WSI systems in comparison with conventional microscopy is the possibility of analysing the digital slides using computer-based algorithms. Quantitative assessment of breast tissue for grading and quantifying biomarker status can be slow procedures whose accuracy may be affected by subjectiveness of the decisions made by pathologists. Recently many researchers and companies started working on computer-assisted systems for breast histopathology analysis. The primary purpose of the studies focusing on computer-assisted analysis of breast virtual slides can be classified in four categories: (i) segmentation of nuclei, tubule, or mitotic figures on BCa slides, (ii) classification of BCa slides as malignant or benign, (iii) BCa grading, and (iv) immunohistochemistry quantification. In chapter 2, computer-assisted image analysis in breast histopathology is reviewed in detail.

1-2-3- Considerations

At first impression, the transition from glass slide to digital slides might seem to be like elimination of films in radiology. However, there are some fundamental differences between these two [28, 64, 65]. First of all, in digital radiology, the images are acquired in digital format while in pathology, an additional step should be added to conventional microscopy system to scan the slides and produce digital images. Despite the rapid evolution in the technology associated with slide scanners, there are still some barriers in this area such as developing accurate and fast, fully automatic focusing systems and potential delay and extra expenses caused by adding the scanning step [28]. Secondly, the sizes of the digital slides are significantly larger than radiologic images. For example, the size of 100 slides at magnification of 40× (typical resolution of 0.25 micron/pixel) is approximately 80 Gigabytes. Therefore, huge storage devices and particular data management software are needed for archiving the virtual slides. Thirdly, workflow of radiologists is different from that of pathologists [65]. Studies done by McClintock et al [66] and Isaacs et al [67] showed that full adoption of WSI in the current workflow of a high-volume histology laboratory without making significant changes was not feasible [66, 67].

Moreover, the appropriate image standards, image compression protocol, guidelines for selecting suitable monitors for viewing the slides, and regulations are required to be established for adoption of whole slide imaging [64]. In April 2017, the Philips IntelliSite Pathology Solution WSI system, , manufactured by Philips Medical Systems Nederland BV, received clearance from US Food and Drug Administration (FDA) for primary diagnostic use [68]. This is the first and currently only WSI system which was allowed to be marketed for primary interpretation of surgical pathology slides in the United States, but the FDA gave 510(k) clearances to some manufacturers for manual and/or quantitative analysis of Immunohistochemistry [28]. To gain FDA approval and confidence from the pathologists, some investigations have been done to compare the performance of the pathologist under microscope and while assessing digital slides [28]. As discussed in the section 1-2-2-1 recent studies showed that pathologists' performance in reading breast slides while using WSI platforms was comparable to conventional microscopy in different tasks in breast pathology.

1-3- Knowledge deficiencies in the literature

After reviewing previous work on added benefits of computer-assisted analysis of Hematoxylin-Eosin stained breast histopathological digital slides, which is explained in the literature review (Chapter 2), the following shortcomings were identified:

- Lack of studies which link image processing features with disagreement among pathologists: previous studies investigated the quantitative image processing features related to disagreement among radiologists and expert consensus ground truth [69-71]. However, association of the image processing features with pathologists' decisions were not explored in breast pathology. Finding these associations can help in better understanding of underlaying reasons for disagreement among pathologists and potentially improve the diagnostic agreement.
- As stated earlier Nottingham grading system has three components. Previous studies achieved promising results for automatic mitotic figure detection [72-74] and also tubule segmentation [75-78], however, there is room for improving the results obtained for nuclear pleomorphism grading. Also as stated earlier the agreement is the

poorest for nuclear pleomorphism grade among three contributing factors in Nottingham grading system. As the breast cancer grade is related to the breast cancer prognosis [11], reproducible grading is highly required.

- Although many studies focused on automatic binary classification of breast histopathological slides as either malignant or benign [79-82], less attention was paid to multi-category classification of breast slides and differentiation of benign subtypes and identification of cancer subtypes. As different benign and cancer subtypes might require different patient management (especially in terms of how aggressive the treatment should be), accurate diagnosis is crucially important [11]

1-4- Aims and objectives

The aim of the studies included in this thesis was to explore the added benefits of computer-assisted analysis of Hematoxylin-Eosin stained breast histopathological digital slides. To realise this aim, the following objectives have been identified:

1) Conducting the literature review (a) to understand which tasks in breast pathology have been already addressed; (b) which image analysis techniques have been used; (c) to find the publicly available datasets, and (d) to determine the main deficiencies from the literature (Chapter 2).
2) (a) Determining the significantly different quantitative features among easily identifiable mitotic figures, challenging mitotic figures, and miscounted non-mitoses within breast slides and (b) identifying which colour spaces capture the difference among these groups better than others. The challenging mitotic figures are those mitoses for which the majority of the readers could not make a decision confidently while miscounted non-mitoses are false positive decisions. This study is an example of a framework for analysing the association of the pathologists' decisions (false-positive, true-positive, or false-negative) with image processing features (chapters 3 and 4).
3) Developing a tool for reproducible scoring of nuclear pleomorphism: The tool combines computer-extracted textural features with pathologists' assessment of cytological criteria. It considers each individual's unique perceptual pattern

and eliminates systematic over- or under-estimating of each grader (chapters 5 and 6).

4) Differentiation among benign and cancer subtypes in breast histopathological slides by aggregating outputs of multiple deep residual networks analysing images from different magnification levels (chapters 7 and 8).

1-5- Thesis structure

The reminder of the thesis is organised as follows:

Chapter 2 presents a review of the previous studies which used computer‑based image analysis in breast pathology. It aims at discussing the previous studies to find which features have been previously extracted from digital slides and which image processing tools have been used for stain normalization, segmentation, and classification of breast digital slides. The findings of these studies are summarized and compared while some key remaining challenges were identified. This chapter was published as the review paper "Computer-based image analysis in breast pathology" in the Journal of Pathology Informatics, 2016.

Chapter 3 is a bridging chapter that provides a detailed background about the mitotic figures, their importance and introduces the original article which is presented in Chapter 4.

Chapter 4 presents the published original article "Determining image processing features describing the appearance of challenging mitotic figures and miscounted nonmitotic objects." This was published in the Journal of Pathology Informatics in 2017. It sought to explore which image processing features differed significantly among easily identifiable mitoses, challenging mitoses (false negatives), and miscounted nonmitoses (false positives). It also compared the discriminative power of different colour spaces for distinguishing these three groups. I implemented the functions for extracting features that were used in this study using MATLAB and C++ (called as mex files in MATLAB).Where appropriate, MATLAB built-in functions were used.

Chapter 5 is a bridging chapter that provides a detailed background about the nuclear grading, its importance, and challenges toward a reproducible nuclear grade. It introduces the journal paper which is presented in Chapter 6.

Chapter 6 presents the original study "COMPASS: Pleomorphism Grading of Breast Cancer by Computer-Assisted Analysis Combined with Pathologist's Assessment." The paper introduces COMPASS (COMputer-assisted analysis combined with Pathologist's ASSessment), a tool proposed for reproducible nuclear pleomorphism scoring. This paper was submitted for publication to the Journal of the American Medical Informatics Association, 2017. It also presents the results for evaluating the performance of COMPASS for the three junior pathologists and discusses whether it could complement the senior pathologist's performance to some extent. I implemented the functions for extracting features that were used in this study using MATLAB. Where appropriate, MATLAB built-in functions were used.

Chapter 7 is a bridging chapter that provides a detailed background about the importance of identifying the carcinoma subtype as well as determining the subtype of the benign lesion. It introduces the journal paper which is presented in Chapter 8.

Chapter 8 presents the original article "Determining benign and cancer subtypes in breast histopathological slides by aggregating outputs of multiple deep residual networks analysing images from different magnification levels." This paper was submitted for publication to Artificial Intelligence in Medicine in 2017. The paper describes the proposed framework for classifying Hematoxylin-Eosin stained breast histopathological images either as benign or cancer, subdividing cancer cases into four subcategories, namely ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma, and classifying those cases classified as benign as adenosis, fibroadenoma, phyllodes a tumour, or tubular adenoma. This study has been implemented using Python (Keras library with TensorFlow backend).

Chapter 9 discusses the findings of the work, their implications, as well as limitations of the studies and possible avenues for improving this work and conducting future studies.

Chapter 10 concludes the thesis and summarizes the studies and their results.

References

[1] Y. C. Cheng and N. T. Ueno, "Improvement of survival and prospect of cure in patients with metastatic breast cancer," Breast cancer, vol. 19, pp. 191-199, 2012.

[2] K. C. Oeffinger, E. T. Fontham, R. Etzioni, A. Herzig, J. S. Michaelson, Y.-C. T. Shih, et al., "Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society," Jama, vol. 314, pp. 1599-1614, 2015.

[3] N. Bjurstam, L. Björneld, S. W. Duffy, T. C. Smith, E. Cahlin, O. Eriksson, et al., "The Gothenburg breast screening trial," Cancer, vol. 80, pp. 2091-2099, 1997.

[4] S. Shapiro, W. Venet, P. Strax, L. Venet, and R. Roeser, "Ten-to Fourteen-Year Effect of Screening on Breast Cancer Mortality 2," Journal of the National Cancer Institute, vol. 69, pp. 349-355, 1982.

[5] L. Tabar, A. Gad, L. Holmberg, U. Ljungquist, C. Fagerberg, L. Baldetorp, et al., "Reduction in mortality from breast cancer after mass screening with mammography: randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare," The Lancet, vol. 325, pp. 829-832, 1985.

[6] J. G. Elmore, M. B. Barton, V. M. Moceri, S. Polk, P. J. Arena, and S. W. Fletcher, "Ten-year risk of false positive screening mammograms and clinical breast examinations," New England Journal of Medicine, vol. 338, pp. 1089-1096, 1998.

[7] J. Chubak, D. M. Boudreau, P. A. Fishman, and J. G. Elmore, "Cost of breast-related care in the year following false positive screening mammograms," Medical care, vol. 48, p. 815, 2010.

[8] M. Bond, T. Pavey, K. Welch, C. Cooper, R. Garside, S. Dean, et al., "Psychological consequences of false-positive screening mammograms in the UK," Evidence-based medicine, vol. 18, pp. 54-61, 2013.

[9] D. L. Weaver, R. D. Rosenberg, W. E. Barlow, L. Ichikawa, P. A. Carney, K. Kerlikowske, et al., "Pathologic findings from the Breast Cancer Surveillance Consortium: population-based outcomes in women undergoing biopsy after screening mammography," Cancer, vol. 106, pp. 732-42, Feb 15 2006.

[10] M. Silverstein, "Where's the outrage?," J Am Coll Surg, vol. 208, pp. 78-9, Jan 2009.

[11] P. P. Rosen, Rosen's breast pathology: Lippincott Williams & Wilkins, 2001.

[12] L. P. Howell, R. Gandour-Edwards, and D. O'Sullivan, "Application of the Scarff-Bloom-Richardson tumor grading system to fine-needle aspirates of the breast," American journal of clinical pathology, vol. 101, pp. 262-265, 1994.

[13] J. D. Brenton, L. A. Carey, A. A. Ahmed, and C. Caldas, "Molecular classification and molecular forecasting of breast cancer: ready for clinical application?," Journal of clinical oncology, vol. 23, pp. 7350-7360, 2005.

[14] P. Robbins, S. Pinder, N. de Klerk, H. Dawkins, J. Harvey, G. Sterrett, et al., "Histological grading of breast carcinomas: a study of interobserver agreement," Hum Pathol, vol. 26, pp. 873-9, Aug 1995.

[15] M. Sikka, S. Agarwal, and A. Bhatia, "Interobserver agreement of the Nottingham histologic grading scheme for infiltrating duct carcinoma breast," Indian J Cancer, vol. 36, pp. 149-53, Jun-Dec 1999.

[16] P. Boiesen, P.-O. Bendahl, L. Anagnostaki, H. Domanski, E. Holm, I. Idvall, et al., "Histologic grading in breast cancer: reproducibility between seven pathologic departments," Acta oncologica, vol. 39, pp. 41-45, 2000.

[17] T. A. Longacre, M. Ennis, L. A. Quenneville, A. L. Bane, I. J. Bleiweiss, B. A. Carter, et al., "Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study," Modern pathology, vol. 19, p. 195, 2006.

[18] F. Theissig, K. D. Kunze, G. Haroske, and W. Meyer, "Histological grading of breast cancer. Interobserver, reproducibility and prognostic significance," Pathol Res Pract, vol. 186, pp. 732-6, Dec 1990.

[19] I. O. Ellis, D. Coleman, C. Wells, S. Kodikara, E. M. Paish, S. Moss, et al., "Impact of a national external quality assessment scheme for breast pathology in the UK," J Clin Pathol, vol. 59, pp. 138-45, Feb 2006.

[20] J. S. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, et al., "Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index," Modern pathology, vol. 18, p. 1067, 2005.

[21] N. Chowdhury, M. R. Pai, F. D. Lobo, H. Kini, and R. Varghese, "Interobserver variation in breast cancer grading: a statistical modeling approach," Anal Quant Cytol Histol, vol. 28, pp. 213-8, Aug 2006.

[22] L. Khazai, L. P. Middleton, N. Goktepe, B. T. Liu, and A. A. Sahin, "Breast pathology second review identifies clinically significant discrepancies in over 10% of patients," Journal of surgical oncology, vol. 111, pp. 192-197, 2015.

[23] K. H. Allison, L. M. Reisch, P. A. Carney, D. L. Weaver, S. J. Schnitt, F. P. O'malley, et al., "Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel," Histopathology, vol. 65, pp. 240-251, 2014.

[24] L. Pantanowitz, P. N. Valenstein, A. J. Evans, K. J. Kaplan, J. D. Pfeifer, D. C. Wilbur, et al., "Review of the current state of whole slide imaging in pathology," Journal of pathology informatics, vol. 2, 2011.

[25] S. J. Schnitt, J. L. Connolly, F. A. Tavassoli, R. E. Fechner, R. L. Kempson, R. Gelman, et al., "Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria," The American journal of surgical pathology, vol. 16, pp. 1133-1143, 1992.

[26] W. A. Wells, P. A. Carney, M. S. Eliassen, A. N. Tosteson, and E. R. Greenberg, "Statewide study of diagnostic agreement in breast pathology," JNCI: Journal of the National Cancer Institute, vol. 90, pp. 142-145, 1998.

[27] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. Tosteson, et al., "Diagnostic concordance among pathologists interpreting breast biopsy specimens," Jama, vol. 313, pp. 1122-1132, 2015.

[28] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman, "Digital imaging in pathology: whole-slide imaging and beyond," Annual Review of Pathology: Mechanisms of Disease, vol. 8, pp. 331-359, 2013.

[29] N. Stathonikos, M. Veta, A. Huisman, and P. J. van Diest, "Going fully digital: Perspective of a Dutch academic pathology lab," Journal of pathology informatics, vol. 4, 2013.

[30] C. B. Nunes, R. M. Rocha, M. A. Buzelin, D. Balabram, F. S. Foureaux, S. S. Porto, et al., "High agreement between whole slide imaging and optical microscopy for assessment of HER2 expression in breast cancer," Laboratory Investigation, vol. 94, pp. 71A-72A, 2014.

[31] Y. Kondo, T. Iijima, and M. Noguchi, "Evaluation of immunohistochemical staining using whole-slide imaging for HER2 scoring of breast cancer in comparison with real glass slides," Pathology International, vol. 62, pp. 592-599, 2012.

[32] E. C. Shaw, A. M. Hanby, K. Wheeler, A. M. Shaaban, D. Poller, S. Barton, et al., "Observer agreement comparing the use of virtual slides with glass slides in the pathology review component of the POSH breast cancer cohort study," Journal of Clinical Pathology, vol. 65, pp. 403-408, 2012.

[33] T. Micsik, E. Turanyi, Z. Sapi, L. Krecsak, G. Kiszler, T. Krenacs, et al., "Validation of digital immunohistochemical evaluation of hormonereceptors in breast cancer," Virchows Archiv, vol. 463, p. 228, 2013.

[34] S. M. Finkbeiner, "Robotic microscopy systems," ed: Google Patents, 2006.

[35] N. Singh, N. Akbar, C. Sowter, K. G. Lea, and C. A. Wells, "Telepathology in a routine clinical environment: Implementation and accuracy of diagnosis by robotic microscopy in a one-stop breast clinic," Journal of Pathology, vol. 196, pp. 351-355, 2002.

[36] H. J. Terpe, W. Müller, A. Liese, C. U. Vogel, and K. H. Broer, "Frozen section telepathology in the clinical routine of a breast cancer center," Pathologe, vol. 24, pp. 150-153, 2003.

[37] C. M. T. P. Francissen, R. F. D. Van La Parra, A. H. Mulder, A. M. Bosch, and W. K. De Roos, "Evaluation of the benefit of routine intraoperative frozen section analysis of sentinel lymph nodes in breast cancer," ISRN Oncology, vol. 1, 2013.

[38] M. A. Gavrielides, C. Conway, N. O'Flaherty, B. D. Gallas, and S. M. Hewitt, "Observer performance in the use of digital and optical microscopy for the interpretation of tissue-based biomarkers," Analytical Cellular Pathology, vol. 2014, 2014.

[39] S. R. Orell, G. F. Sterrett, M. Walters, and D. Whitaker, Manual and atlas of fine needle aspiration cytology: Churchill Livingstone, 1987.

[40] A. Nassar, C. Cohen, S. S. Agersborg, W. Zhou, K. A. Lynch, E. A. Barker, et al., "A multisite performance study comparing the reading of immunohistochemical slides on a computer monitor with conventional manual microscopy for estrogen and progesterone receptor analysis," American Journal of Clinical Pathology, vol. 135, pp. 461-467, 2011.

[41] W. S. Campbell, S. H. Hinrichs, S. M. Lele, J. J. Baker, A. J. Lazenby, G. A. Talmon, et al., "Whole slide imaging diagnostic concordance with light microscopy for breast needle biopsies," Human Pathology, vol. 45, pp. 1713-1721, 2014.

[42] F. J. W. M. Leong, A. K. Graham, P. Schwarezmann, and J. O. D. McGee, "Clinical trial of telepathology as an alternative modality in breast histopathology quality assurance," Telemedicine Journal and e-Health, vol. 6, pp. 373-377, 2000.

[43] F. J. W. M. Leong and J. O'D McGee, "Robotic interactive telepathology in proficiency testing/quality assurance schemes," Electronic Journal of Pathology and Histology, vol. 7, pp. 1-11, 2001.

[44] F. A. Zito, P. Verderio, G. Simone, V. Angione, P. Apicella, S. Bianchi, et al., "Reproducibility in the diagnosis of needle core biopsies of non-palpable breast lesions: An international study using virtual slides published on the world-wide web," Histopathology, vol. 56, pp. 720-726, 2010.

[45] J. Terry, E. E. Torlakovic, J. Garratt, D. Miller, M. Köbel, J. Cooper, et al., "Implementation of a Canadian external quality assurance program for breast cancer biomarkers: An initiative of Canadian Quality Control in Immunohistochemistry (cIQc) and Canadian Association of Pathologists (CAP) national standards committee/immunohistochemistry," Applied Immunohistochemistry and Molecular Morphology, vol. 17, pp. 375-382, 2009.

[46] V. Della Mea, F. Puglisi, M. Bonzanini, S. Forti, V. Amoroso, R. Visentin, et al., "Fine-needle aspiration cytology of the breast: A preliminary report on telepathology through internet multimedia electronic mail," Modern Pathology, vol. 10, pp. 636-641, 1997.

[47] T. Schrader, P. Hufnagl, W. Schlake, and M. Dietel, "Study of efficiancy of teleconsultation: the Telepathology Consultation Service of the Professional Assoziation of German Pathologists for the screening program of breast carcinoma," Verhandlungen der Deutschen Gesellschaft für Pathologie, vol. 89, pp. 211-218, 2005.

[48] S. Wienert, M. Beil, K. Saeger, P. Hufnagl, and T. Schrader, "Integration and acceleration of virtual microscopy as the key to successful implementation into the routine diagnostic process," Diagnostic Pathology, vol. 4, 2009.

[49] A. M. López, A. R. Graham, G. P. Barker, L. C. Richter, E. A. Krupinski, F. Lian, et al., "Virtual slide telepathology enables an innovative telehealth rapid breast care clinic," Seminars in Diagnostic Pathology, vol. 26, pp. 177-186, 2009.

[50] M. Lundin, J. Lundin, H. Helin, and J. Isola, "A digital atlas of breast histopathology: an application of web based virtual microscopy," Journal of Clinical Pathology, vol. 57, pp. 1288-1291, 2004.

[51] M. Khushi, G. Edwards, D. A. de Marcos, J. E. Carpenter, J. D. Graham, and C. L. Clarke, "Open source tools for management and archiving of digital microscopy data to allow integration with patient pathology and treatment information," Diagnostic Pathology, vol. 8, 2013.

[52] A. Bondi, S. Lega, P. Crucitti, P. Pierotti, R. Rapezzi, P. Sassoli De Bianchi, et al., "Quality assurance and automation," Cytopathology, vol. 23, p. 39, 2012.

[53] E. A. Krupinski, A. R. Graham, and R. S. Weinstein, "Characterizing the development of visual search expertise in pathology residents viewing whole slide images," Hum Pathol, vol. 44, pp. 357-64, Mar 2013.

[54] E. A. Krupinski, A. A. Tillack, L. Richter, J. T. Henderson, A. K. Bhattacharyya, K. M. Scott, et al., "Eye-movement study and human performance using telepathology virtual slides: implications for medical education and differences with experience," Hum Pathol, vol. 37, pp. 1543-56, Dec 2006.

[55] E. A. Krupinski and R. S. Weinstein, "Changes in visual search patterns of pathology residents as they gain experience," in Progress in Biomedical Optics and Imaging - Proceedings of SPIE, 2011.

[56] V. Raghunath, M. O. Braxton, S. A. Gagnon, T. T. Brunye, K. H. Allison, L. M. Reisch, et al., "Mouse cursor movement and eye tracking data as an indicator of pathologists' attention when viewing digital whole slide images," J Pathol Inform, vol. 3, p. 43, 2012.

[57] T. T. Brunyé, P. A. Carney, K. H. Allison, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Eye movements as an index of pathologist visual expertise: a pilot study," PloS one, vol. 9, p. e103447, 2014.

[58] F. Schnorrenberg, C. S. Pattichis, C. N. Schizas, and K. Kyriacou, "Content-based retrieval of breast cancer biopsy slides," Technology and Health Care, vol. 8, pp. 291-297, 2000.

[59] F. Wang, J. Kong, L. Cooper, T. Pan, T. Kurc, W. Chen, et al., "A data model and database for high-resolution pathology analytical image informatics," J Pathol Inform, vol. 2, p. 32, 2011.

[60] Y. Zheng, Z. Jiang, J. Shi, and Y. Ma, "Pathology image retrieval by block LBP based pLSA model with low-rank and sparse matrix decomposition," in Communications in Computer and Information Science vol. 437, ed, 2014, pp. 327-335.

[61] P. R. Quinlan, A. Ashfield, L. Jordan, C. Purdie, and A. M. Thompson, "An integrated informatics platform to facilitate transforming tissue into knowledge," Breast Cancer Research, vol. 12, p. S9, 2010.

[62] A. I. Wright, H. I. Grabsch, and D. E. Treanor, "RandomSpot: A web-based tool for systematic random sampling of virtual slides," J Pathol Inform, vol. 6, p. 8, 2015.

[63] J. P. Johnson, E. A. Krupinski, M. Yan, H. Roehrig, A. R. Graham, and R. S. Weinstein, "Using a Visual Discrimination Model for the Detection of Compression Artifacts in Virtual Pathology Images," Medical Imaging, IEEE Transactions on, vol. 30, pp. 306-314, 2011.

[64] L. Pantanowitz, P. N. Valenstein, A. J. Evans, K. J. Kaplan, J. D. Pfeifer, D. C. Wilbur, et al., "Review of the current state of whole slide imaging in pathology," Journal of pathology informatics, vol. 2, p. 36, 2011.

[65] J. D. Hipp, A. Fernandez, C. C. Compton, and U. J. Balis, "Why a pathology image should not be considered as a radiology image," Journal of pathology informatics, vol. 2, p. 26, 2011.

[66] D. S. McClintock, R. E. Lee, and J. R. Gilbertson, "Using computerized workflow simulations to assess the feasibility of whole slide imaging full

adoption in a high-volume histology laboratory," Analytical Cellular Pathology, vol. 35, pp. 57-64, 2012.

[67] M. Isaacs, J. K. Lennerz, S. Yates, W. Clermont, J. Rossi, and J. D. Pfeifer, "Implementation of whole slide imaging in surgical pathology: A value added approach," Journal of Pathology Informatics, vol. 2, 2011.

[68] B. Boyce, "An update on the validation of whole slide imaging systems following FDA approval of a system for a routine pathology diagnostic service in the United States," Biotechnic & Histochemistry, pp. 1-9, 2017.

[69] Z. Gandomkar, K. Tay, W. Ryder, P. C. Brennan, and C. Mello-Thoms, "iCAP: An Individualized Model Combining Gaze Parameters and Image-based Features to Predict Radiologists' Decisions While Reading Mammograms," IEEE transactions on medical imaging, vol. 36, pp. 1066-1075, 2017.

[70] M. A. Mazurowski, J. A. Baker, H. X. Barnhart, and G. D. Tourassi, "Individualized computer-aided education in mammography based on user modeling: Concept and preliminary experiments," Medical physics, vol. 37, pp. 1152-1160, 2010.

[71] S. Voisin, F. Pinto, G. Morin-Ducote, K. B. Hudson, and G. D. Tourassi, "Predicting diagnostic error in radiology via eye-tracking and image analytics: Preliminary investigation in mammography," Medical physics, vol. 40, 2013.

[72] M. Veta, P. J. van Diest, M. Jiwa, S. Al-Janabi, and J. P. Pluim, "Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method," PloS one, vol. 11, p. e0161286, 2016.

[73] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images," IEEE transactions on medical imaging, vol. 35, pp. 1313-1321, 2016.

[74] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in

International Conference on Medical Image Computing and Computer-assisted Intervention, 2013, pp. 411-418.

[75] K. Nguyen, M. Barnes, C. Srinivas, and C. Chefd'hotel, "Automatic glandular and tubule region segmentation in histological grading of breast cancer," in Proc. SPIE, 2015, p. 94200G.

[76] P. Maqlin, R. Thamburaj, J. J. Mammen, and A. K. Nagar, "Automatic detection of tubules in breast histopathological images," in Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012), 2013, pp. 311-321.

[77] A. Basavanhally, E. Yu, J. Xu, S. Ganesan, M. Feldman, J. Tomaszewski, et al., "Incorporating domain knowledge for tubule detection in breast histopathology using O'Callaghan neighborhoods," in SPIE Medical Imaging, 2011, p. 796310.

[78] M. Barnes, C. Chefd'hotel, S. Chukka, and K. Nguyen, "Automatic glandular and tubule detection in histological grading of breast cancer," ed: Google Patents, 2017.

[79] P. Filipczuk, M. Kowal, and A. Obuchowicz, "Multi-label fast marching and seeded watershed segmentation methods for diagnosis of breast cancer cytology," in Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, 2013, pp. 7368-7371.

[80] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in Pattern Recognition (ICPR), 2016 23rd International Conference on, 2016, pp. 2440-2445.

[81] M. A. Kahya, W. Al-Hayani, and Z. Y. Algamal, "Classification of breast cancer histopathology images based on adaptive sparse support vector machine," Journal of Applied Mathematics and Bioinformatics, vol. 7, p. 49, 2017.

[82] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," IEEE Transactions on Biomedical Engineering, vol. 63, pp. 1455-1462, 2016.

Chapter 2

Literature review

This chapter has been published as:

Gandomkar, Ziba, Patrick C. Brennan, and Claudia Mello-Thoms. “Computer-based image analysis in breast pathology.” Journal of pathology informatics 7:43, 2016.

J Pathol Inform

Editor-in-Chief:
Anil V. Parwani, Columbus, OH, USA
Liron Pantanowitz, Pittsburgh, PA, USA

OPEN ACCESS
HTML format

For entire Editorial Board visit : www.jpathinformatics.org/editorialboard.asp

Review Article

Computer-based image analysis in breast pathology

Ziba Gandomkar[1], Patrick C. Brennan[1], Claudia Mello-Thoms[1,2]

[1]Image Optimisation and Perception, Discipline of Medical Radiation Sciences, University of Sydney, Australia, [2]Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA

E-mail: *Mrs. Ziba Gandomkar - ziba.gandomkar@sydney.edu.au
*Corresponding author

Received: 23 May 2016 Accepted: 15 September 2016 Published: 21 October 2016

Abstract

Whole slide imaging (WSI) has the potential to be utilized in telepathology, teleconsultation, quality assurance, clinical education, and digital image analysis to aid pathologists. In this paper, the potential added benefits of computer-assisted image analysis in breast pathology are reviewed and discussed. One of the major advantages of WSI systems is the possibility of doing computer-based image analysis on the digital slides. The purpose of computer-assisted analysis of breast virtual slides can be (i) segmentation of desired regions or objects such as diagnostically relevant areas, epithelial nuclei, lymphocyte cells, tubules, and mitotic figures, (ii) classification of breast slides based on breast cancer (BCa) grades, the invasive potential of tumors, or cancer subtypes, (iii) prognosis of BCa, or (iv) immunohistochemical quantification. While encouraging results have been achieved in this area, further progress is still required to make computer-based image analysis of breast virtual slides acceptable for clinical practice.

Key words: Breast pathology, breast virtual slides, image analysis, whole slide imaging

Access this article online
Website: www.jpathinformatics.org
DOI: 10.4103/2153-3539.192814
Quick Response Code:

INTRODUCTION

Whole slide imaging (WSI) has the potential to be utilized in telepathology, clinical education, and digital image analysis to aid pathologists. As different types of specimen have different specifications, comprehensive studies should be carried out in each pathology subspecialty to assess the extent of added benefits of WSI in that field. A large proportion of the pathology slides are related to breast tissue; for example, in the United States, 1.6 million breast biopsies are assessed by the pathologists each year.[1] Recent studies suggested that WSI can be adopted in breast pathology as pathologists' performance in reading breast slides while using WSI platforms was comparable to conventional microscopy in breast pathology.[2]

One of the major advantages of WSI systems compared to conventional microscopy is the possibility of doing computer-based image analysis on the digital slides. Recently, many researchers and slide scanner vendors have started developing automated methods to facilitate pathologists' tasks in breast pathology. This review is restricted to the computer-based image analysis in breast pathology and aimed at discussing the previous studies, summarizing their results, and identifying the remaining challenges and areas where further studies are required.

This article may be cited as: Gandomkar Z, Brennan PC, Mello-Thoms C. Computer-based image analysis in breast pathology. J Pathol Inform 2016;7:43.

It should be noted that image analysis done in other pathology subspecialties or animals' tissue might be extendable to breast pathology; however, discussion about the potentials for extending these ideas to breast pathology is out of the scope of this review. For a broader review on digital pathology in general, please refer to the studies by Pantanowitz *et al.* and Ghaznavi *et al.*[3,4]

SEARCH STRATEGY

Three different databases, namely Scopus, PubMed, and IEEEXplore, were searched to find relevant studies published after 1995. Our overall search strategy included terms for digital slides (e.g., whole slide, digital pathology, virtual slide) and breast and was limited to English-language, original, human studies. We also searched references of the retrieved articles. The exact search statement for each database can be found in Appendix 1. For studies where the methodology evolved in two or more papers with considerable amount of overlap, only the most expanded version was included in the review.

Studies focusing on the application of WSI or computer-aided analysis to multispectral images or quantification of biomarkers other than four clinically important immunohistochemical (IHC) stains (i.e., estrogen receptor (ER), progesterone receptor (PR), Ki-67, and human epidermal growth factor receptor [HER2]) have been excluded as they are not currently widely used in the clinical practice.

CLASSIFICATION OF THE REVIEWED STUDIES

The primary purpose of the reviewed studies can be classified into four categories: (i) segmentation of desired regions or objects in the slide, (ii) classification of breast slides, (iii) prognosis of breast cancer (BCa), and (iv) IHC quantification. Most of the reviewed studies had a block diagram similar to the one shown in Figure 1. Some of the methods may not include one or more of the steps illustrated in Figure 1. Moreover, multiple steps may have been merged in some studies. As shown, features could be extracted from a segmented object or tissue texture. Each group of studies is discussed in this section.

Before processing the slides, preprocessing steps can be performed to eliminate the background,[5] segment diagnostically relevant area (DRA),[6,7] standardize the color,[6] or separate stain.[8] Color deconvolution is a commonly used preprocessing step to separate the H channel. Ruifrok and Johnston[9] proposed a formulation based on the Beer–Lambert law to map the red, green, and blue (RGB) color space to a set of three stains using color deconvolution. It should be noted that color deconvolution needs prior knowledge about the color vector of each stain (stain matrix). Standard stain matrix for a wide range of stain combinations is provided in a study by Ruifrok and Johnston.[9] However, use of image-specific stain matrix is more accurate. This motivated the development of an image-specific stain normalization algorithm to automatically estimate stain matrix for each slide, such as the one presented in a study by Khan *et al.*[10] As an alternative solution to overcome color variations due to dyeing, in a study by Ali *et al.*,[11] stain separation was done adaptively in cyan, magenta, and yellow color space rather than RGB. In addition, the RGB color space is not perceptually uniform. To overcome this problem, in the studies by Dundar *et al.*[12] and Basavanhally *et al.*,[13] lab color space which is a perceptually uniform color space was used. Finally, the color deconvolution assumes that the relation between spectral absorbance of a stain mixture and the concentrations of the pure stains is linear. This assumption is valid under monochromatic conditions; however, it introduces an error under nonmonochromatic conditions.[14]

Segmentation of Desired Regions or Objects in the Slide

A wide range of image processing methods has been used for segmenting objects in breast virtual slides. Accurate segmentation is important as it is an intermediate step of studies with various purposes. Because of its

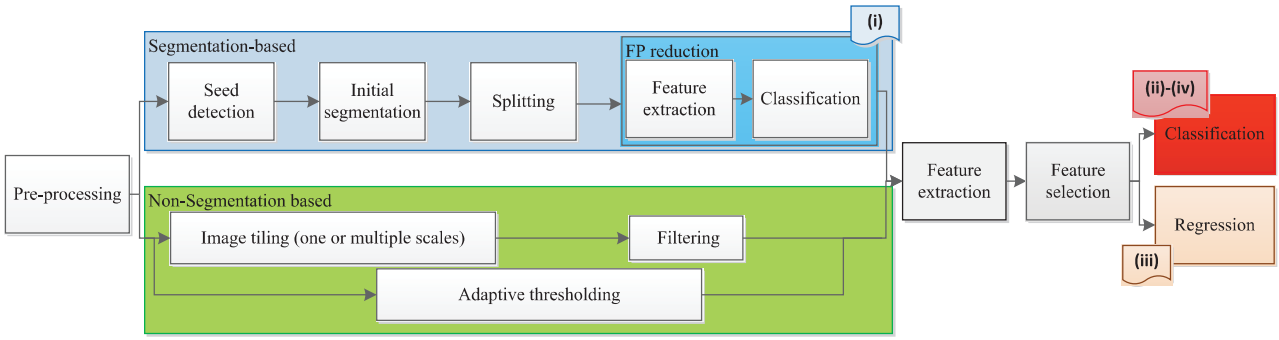


Figure 1: The common steps in the reviewed studies

importance, only proposing a segmentation method to handle difficulties in breast slides has been main subject of 25 reviewed studies. The studies are summarized in Table 1. In this section, methods proposed for segmenting DRAs,[31-34] epithelial nuclei,[5,15-19,35] lymphocyte cells,[7,13,20] tubule,[21-25] and mitotic figures[6,8,26-30] are discussed.

Table 1: Summary of the studies aimed at segmentation of structures in breast virtual slides

	Reference	Initialization/seed detection	Segmentation	Features; classifier	Result
Epithelial nuclei	[15]	Color deconvolution, morphological operation; fast radial symmetry transform	Multi-scale marker-controlled watershed	Morphology; rule-based discarding	TPR: 0.86 Specificity: 0.89
	[16]	Adaptive thresholding of sequentially filtered image; distance transform of overlapped cells	Gaussian mixture modeling of distance transform; cluster validation; occluded contour reconstruction	-	TPR: 0.97 (combined for cervical and BCa cells)
	[17]	Morphological reconstruction; adaptive thresholding	Marker extraction based on optimal H-minima transform; Marker-controlled watershed	-	ACC: 0.96 (combined for cervical and BCa cells)
	[5]	Background removal by graphcuts-based binarization; distance-constrained multiscale LoG filtering	Graphcuts-based method with combination of alpha expansion and graph coloring	-	ACC>0.94
	[18]	Color deconvolution, Calculating local features based on laws' texture	Probability map generation; ACM including shape priors	-	-
	[19]	Color deconvolution by singular value decomposition	Clustering; ACM	Intensity, morphology, texture; AdaBoost	ACC: 0.95
Lymphocyte	[20]	Constructing the shape priors	watershed; ACM combined with shape priors	Morphology; SVM	TPR: 0.86 PPV: 0.67
	[7]	Expectation maximization based segmentation of object classes	Geodesic ACM; Concavity detection; edge-path algorithm	Intensity, k-means clustering	TPR: 0.91 PPV: 0.78
	[13]	Thresholding luminance channel	Region growing	Size, luminance, proximity; Bayesian modeling and Markov random field	ACC: 0.9
Tubule	[21]	Color swatch definition by the user	Normalized cuts on weighted mean shift reduced color space; Color gradient based geodesic ACM	-	TPR: 0.86 PPV: 0.80
	[22]	Color deconvolution	Lumen detection by hierarchical normalized cut initialized color gradient based ACM	AF based on O'Callaghan neighborhood; RF	TPR: 0.86 PPV: 0.89
	[23]	K-means; identifying the nuclei nearer to each white region	Contour detection of the nuclei near-lumen using level set	Surrounding cell evenness; rule-based discarding	ACC: 0.9
	[24]	Nuclei detection by radial symmetry based method; classification of nuclei as normal/tumor	Lumen detection by thresholding	Morphology, texture, surrounding nuclei distribution	ACC: 0.91
	[25]	Super pixel generation; forming spatio-color-texture map using texton representation	Normalized graph cuts	-	TPR: 0.88 Specificity: 0.92

Contd...

Table 1: Contd...

	Reference	Initialization/seed detection	Segmentation	Features; classifier	Result
Mitotic cell	[26]	Color channel selection in RGB, HSV, lab, and luv spaces LoG on blue ratio channel; thresholding	Morphological operation; ACM	Intensity, texture, morphology; DT	(A* + DT) TPR: 0.74, PPV: 0.70 (H** + DT) TPR: 0.71, PPV: 0.56
	[8]	-	Chan-Vese level set method	Texture, morphology; SVM	TPR: 0.56 4.2 FP per high power field
	[27]	-	Multi-resolution graph based segmentation	Texture; clustering spatial refinement	TPR: 0.70
	[6]	Stain normalization; DRA segmentation	Gamma-Gaussian mixture modeling using expectation maximization	Texture; SVM	TPR: 0.72; PPV: 0.70
	[28]	-	Building an optimal training set	Deep neural network	TPR: 0.70; PPV: 0.88
	[29]	PCA analysis of RGB space	Adaptive thresholding and morphological operation	Intensity, texture, morphology, features extracted by convolutional neural networks; SVM	(A*) TPR: 0.59, PPV: 0.74 (H*) TPR: 0.44, PPV: 0.76
	[30]	LoG on blue ratio channel; thresholding	Morphological operation	Texture; DT, linear and nonlinear SVM	Selected features + SVM: TPR: 0.88, PPV: 0.60

*A: Aperio Scanscope CS (Aperio Technologies, Vista, California), **H: Hamamatsu NanoZoomer 2.0 HT (Hamamatsu Photonics, Bridgewater NJ). ACC: Accuracy, TPR: True positive rate, PPV: Positive predictive value, SVM: Support vector machine, PCA: Principal component analysis, DRA: Diagnostically relevant area, ACM: Active contour model, BCa: Breast cancer, DT: Decision tree, AF: Architectural features, RF: Random Forest, HSV: Hue-Saturation-Value color space

Diagnostically relevant areas

In pathology slides, large areas are empty. As stated, segmenting DRAs could be used as a preprocessing step to reduce the computational cost[6,7] or to avoid storage of non-DRAs with high magnification.[36] Due to its significance, there are studies aimed only at improving the accuracy of DRAs segmentation.

In the earliest method,[34] thresholding of gray-level image was used to segment DRAs. However, many important features of breast tissues are coded in the color and texture. Therefore, in a study by Mercan *et al.*,[32] a texton-based approach was proposed to distinguish between DRAs and irrelevant patches. In another study by Khan *et al.*,[31] Gabor-based texture features were used to differentiate hypocellular from hypercellular stroma. Gabor filters are extensively used in image analysis as they resemble the human visual system.

In a study by Peikari *et al.*,[33] the areas that attracted pathologists' attention were found using eye tracking data obtained from pathologists while assessing digital breast slides. The visual bag-of-words model with texture and color features was used to describe DRAs and train a logistic regression and a support vector machine (SVM) to predict DRAs in testing slides.

Epithelial cells

As shown in Figure 1, the segmentation procedure of epithelial cells included seed detection, initial segmentation, splitting, and false positive (FP) reduction. In one of the earliest studies by Dalle *et al.*,[37] gamma-corrected red channel was used to segment the epithelial cells. K-means clustering in RGB space was also utilized to detect cells.[38] However, in more recent studies, epithelial cells were mostly segmented from the H channel.[15,35,39,40] Thresholding the H channel followed by morphological operation is a low-computational cost approach to detect epithelial cells. Nonetheless, thresholding is sensitive to variations of stain and cannot handle overlapping cells. Hough transform as well as Laplacian of Gaussian (LoG) filtering[41] and its approximation, which is difference of Gaussian (DoG), are also popular tools to detect blob-like objects and are utilized to detect nuclei in breast tissue. They are more robust to the staining variations; however, the Hough transform is a computationally expensive approach and DoG should be deployed in a multi-resolution scheme to address cells with different sizes. The fast radial symmetry transform, which is a computationally efficient, noniterative procedure for localizing radial symmetry objects, has also been utilized for candidate nuclei locations detection.[15]

The initial seeds could be used for further fine segmentation using active contour models (ACMs),[18,19,37,39,40] Graphcuts,[5] or marker-controlled watershed.[15,35] Conventionally, the ACM relies on gray-level image, but breast slides are colored. To address this issue, in a study by Basavanhally *et al.*,[39] color gradient-based ACM was used. In addition, ACM cannot handle overlapping cells as they rely only on intensity information and do not incorporate knowledge about the nucleus shape. In a study by Veillard *et al.*,[18] a nucleic shape prior was included to deal with this issue. In a study by Al-Kofahi *et al.*,[5] Graphcuts could partially handle the segmentation the overlapping cells when combined with distance map constrained multi-scale LoG for initial seed detection. However, Graphcuts led to over-segmentation in enlarged highly textured nuclei and under-segmentation in partially broken or weakly stained touching cells. Marker-controlled watershed is a robust approach for separating the overlapping cells when the initial seeds are correctly localized. However, in case of severely overlapping cells, spurious initial seeds are inevitable. The adaptive H-minima transform[17] and Bayesian classification[16] scheme were proposed to handle severely touching cells.

In the studies by Veta *et al.*[35] and Vink *et al.*,[19] an extra FP reduction step was added to improve positive prediction value (PPV). In a study by Veta *et al.*,[35] morphological features were extracted from each segmented area and rule-based discarding was used to eliminate FPs. In a study by Vink *et al.*,[19] a more comprehensive feature set including intensity-based features, morphological and textural features was extracted from each segmented area and modified AdaBoost was used for classification of areas as true nuclei or FPs. Further investigations are still required to eliminate FPs and handle overlapping, enlarged, and broken cells.

Lymphocytes

Lymphocytic infiltration is a prognostic indicator; therefore, recently, researchers worked on the automatic segmentation of lymphocytes. In a study by Basavanhally *et al.*,[13] lymphocytes were initially segmented using region growing, which resulted in a large number of epithelial nuclei being detected as well. Bayesian modeling of size and luminance of lymphocytes and proximity modeling using Markov random field were used to eliminate nuclei. However, region growing cannot handle overlapping cells. To address this, a concavity detection scheme was proposed in a study by Fatakdawala.[7] The expectation maximization-based method was used to initialize the ACM and then overlapping cells were split. Despite achieving a true positive rate (TPR) of 86%, PPV was only 64%. In a study by Ali and Madabhushi,[20] shape priors were incorporated in ACM to handle overlapped cells and the watershed algorithm was used to initialize the ACM. Similarly, high TPR (86%) and low PPV (67%) were achieved. Therefore, it seems that adding an FP reduction module is required to eliminate epithelial nuclei.

Tubules

Tubules are characterized by a white region called lumen, surrounded by a single layer of nuclei in normal breast histopathology. Dalle *et al.*[42] utilized thresholding followed by morphological operations to segment lumen areas. However, thresholding was not robust to stain variation. Xu *et al.*[21] proposed a color gradient-based geodesic ACM which was initialized by weighted mean shift clustering and normalized cuts for lumen segmentation. Later, in a study by Basavanhally *et al.*,[22] domain knowledge was incorporated into method proposed in a study by Xu *et al.*[21] and each segmented area was classified either as true or false based on architectural features and an accuracy of 86% in detection of true lumen was achieved. Maqlin *et al.*[23] used heuristic rules based on evenness and closeness of strings of surrounding nuclei to eliminate lumen-like areas and achieved an accuracy of 90%.

The above-mentioned methods associated only the closest nuclei to the lumen. However, it could be surrounded by multiple layers of nuclei. Therefore, in a study by Nguyen *et al.*,[24] the global distribution of the nuclei and lumina were considered. Furthermore, a comprehensive feature set containing architectural, morphological, intensity-based, and textural features were used to distinguish true lumina from artifacts. Finally, the discussed methods focus on lumen detection; however, nuclear arrangement in tubules could be with a lumen or in solid islands without a lumen. Belsare *et al.*[25] proposed a novel integrated spatio-color-texture-based graph partitioning method to address this issue and achieved a correct classification rate (CCR) of 92% for segmentation.

Mitotic figures

Mitosis counting is tedious and subject to inter-observer variation. The automatic detection of mitosis could potentially address these problems. Roullier *et al.*[27] proposed a multi-resolution image analysis strategy for detection of mitotic figures based on Graph-based regularization. The method was analyzed WSI at different levels and segmented the relevant areas and detected the mitoses in the highest magnification, and it was completely unsupervised. However, the detection rate was 70% and no FP reduction step was adopted; hence, further improvement was required for deploying the method in a clinical setting.

In a study by Khan *et al.*,[6] the pixel intensities of mitotic and nonmitotic areas were modeled by a Gamma-Gaussian mixture. A set of textural and intensity-based features were extracted from each region labeled as mitosis by the first module. The features fed

into an SVM classifier which detects FP instances. The obtained TPR and PPV were 72% and 70%. Irshad *et al.*[30] detected the candidate region in the blue ratio image using thresholding followed by morphological operations and extracted a patch of size 80 pixel × 80 pixel from blue ratio and red and blue channels of RGB color space. A wide range of textural features including Haralick textural, gray-level run length, scale invariant feature transform, and Gabor-based features was extracted from each patch. The principal component analysis was used for dimension reduction. The features were fed into decision tree, linear SVM, and nonlinear SVM. It was shown that a decision tree achieved to the highest performance with a TPR of 76% and a PPV of 75%. Later, Irshad[26] investigated the added value of morphological features and features from other color spaces to the FP reduction step, but the result did not improve.

One of the difficulties in the detection of mitotic figure is its wide range of appearances. To handle this issue, in the studies by Cireşan *et al.* and Malon and Cosatto,[28,29] the learned features extracted by convolutional neural networks were utilized. Malon and Cosatto[29] achieved a TPR of 59% and a PPV of 75% while in a study by Cireşan *et al.*,[28] the PPV and TPR were improved to 80% and 70% using an optimized approach for sampling nonmitosis pixels in the training set. Further investigations are still required in mitotic detection filed to improve TPR and PVP and also deal with the wide range of variability in the appearance of mitotic figures.

Classification

Computer-based image analysis of breast slides may aim at classifying the virtual slide into different categories. The classification could be done based on the grade of BCa, the invasive potential of tumors, or cancer subtypes. Table 2 summarizes the purposes and methods of the reviewed studies aimed at classification of breast slides.

Cancer grading

Scarff-Bloom-Richardson grading system is a well-known grading system relying on magnitude of tubule formation, nuclear pleomorphism, and mitotic count. Segmenting the mitotic figures (which leads to the mitotic count) has been discussed in sections 1–5. The discussed studies in this section aimed at classifying the slides according to nuclear pleomorphism[37,43,44] or all three factors.[39,45]

The earliest reviewed study by Weyn *et al.*[44] applied wavelet transform in four levels on the segmented nuclei and the energy of filtered images in each scale was calculated. In addition to wavelet-based, Haralick, intensity-based, and morphological features were extracted and fed into a K-nearest neighbor classifier to separate individual nuclei and also each case in four categories (normal, nuclear atypia Grade I, II, III). A CCR of 64% for classification of individual nuclei and a CCR of 79% for case-based classification were observed. It was shown that textural features (wavelet-based and Haralick features) had a high additive value to intensity-based features. The dataset used in the study was highly imbalance (21 normal vs. eight Grade III cases) and the segmentation method was required further refinement.

In a study by Doyle *et al.*,[45] the centers of nuclei were manually segmented and a range of intensity-based and textural (Gabor-based and Haralick) features were extracted from each nucleus. The mean, standard deviation, minimum-to-maximum ratio, and mode of these features over all cells in each slide were calculated and formed the feature vector. The architectural features were also extracted from Voronoi diagram, Delaunay triangulation, minimum spanning tree, and nuclei density function. The architectural features resulted in the highest CCR for low- versus high-grade classification while the textural features resulted in a significantly lower CCR (73 vs. 93%). Despite the encouraging CCR, the fact that the segmentation was done manually limits generalizability of the study. In a study by Dalle *et al.*,[37] the nuclei segmentation was done automatically using polar transform, and area, compactness, and mean intensity were extracted from each segmented nuclei. A high value for CCR (92%) was achieved for scoring nuclear pleomorphism of 2396 region of interests (ROIs). However, the result could be biased as the dataset contained images from only six patients and did not include any patient with Grade I.

Pathologists implicitly integrate features from multiple field-of-views (FOVs) of different sizes when grading BCa. However, automatically selecting an optimal FOV size is not straightforward. In a study by Basavanhally *et al.*,[39] architectural and textural features were extracted from a multi-FOV of varying sizes and important features at different FOV sizes were identified to distinguish low/high-, low/intermediate-, and intermediate/high-grade patients. Unsurprisingly, the highest performance was obtained when distinguishing low from high-grade patients. Similar to results obtained in a study by Doyle *et al.*,[45] architectural features performed better than textural ones. It was also observed that the most discriminating architectural features were different in FOVs with various sizes while contrast played a dominant role among textural features. It was also shown that the multi-FOV classifier outperformed multi-scale classifier.

All of the above-mentioned methods extracted features from segmented nuclei only while pathologists rely on features from other structures such as tubules. To overcome this limitation, in a study by Petushi *et al.*,[43] features were also extracted from the tubule and showed that the density of tubule and number of Grade III would be useful parameters for BCa grading.

Table 2: Summary of the studies aimed at breast histopathology slides classification

References	Purpose	Processing/segmentation	Features; classifier	Result
[37]	Classification of BCa slides as score 2 or 3 based on nuclear Pleomorphism	Color deconvolution; thresholding and morphological operations; gradient in polar space	Morphology, texture; Gaussian modeling	CCR: 0.92
[43]	Classification of cells according to Nottingham histologic grade (3-Class)	Adaptive thresholding; morphological operation; Nuclei classification; Tubule detection	Texture, number of mitotic cells and tubules; linear, quadratic, ANN and DT	(The best: ANN) 3-Class CCR: 0.71
[44]	Classification of cells as benign or malignant (2-Class) Classification of cells based on nuclear pleomorphism (3-Class)	Thresholding; morphological operation; multiscale representation using wavelet	Morphology, intensity, texture; k-nearest neighbor	2-Class CCR: 0.89 3-Class CCR: 0.80
[39]	Classification of BCa slides as low (mBR 3-5), intermediate (mBR 6-7), and high (mBR 8-9) grade classes	Color deconvolution; morphological operation; color gradient-based ACM	Architecture, texture; boosted multi-FOV classifier	(Low/high) CCR: 0.93 (Low/intermediate) CCR: 0.72 (Low/high) CCR: 0.74
[45]	Classification of cancerous and noncancerous slides Classification of BCa slides as low or high grades	Extraction of 3400 features from manually segmented nuclei; dimension reduction by spectral clustering	(The best) Gabor-based texture; SVM (The best) architecture; SVM	CCR: 0.96 CCR: 0.93
[38]	Classification of BCa slides as benign or malignant	Adaptive thresholding; morphological operation; multi-label fast marching; watershed	Morphology, intensity, texture, architecture; k-nearest neighbor	Sensitivity: 0.97 specificity: 0.94
[12]	Classification of intraductal breast lesions as actionable (ADH and DCIS) or nonactionable (UDH)	Marker-controlled watershed	Morphology, intensity; multiple instances learning	CCR: 0.88
[46]	Classification of BCa slides as benign or malignant (2-Class) Classification of BCa slides into benign and two subtypes of cancer (3-Class classification)	Texton library construction by using four different filter banks; dimension reduction	Texton histogram; SVM, k-NN, DT, Bayesian, 4 boosting algorithms	(The best: Gentle AdaBoost) 2-Class CCR: 0.89 3-Class CCR: 0.80
[40]	Classification of nuclei and ROI in breast slides as benign or malignant	Color deconvolution in CMY; difference of Gaussian; Hough transform; ACM	Morphology, texture; SVM	(Nuclei) TPR: 0.81; FPR: 0.30 (ROI) TPR: 0.92; FPR: 0.20
[42]	Classification of BCa slides based on overall mBR grade (3-Class)	Cell localization and detection of tubular formations in low-resolution global image; classifying cells as epithelial/ tumor cells or candidate mitotic cells by Gaussian modeling	Tubules: Area of tubule/ area of slide; rule-based Nuclei: Color; Gaussian modeling Mitotic figures: Morphology, intensity; Gaussian modeling	-

mBR: Modified Scarff-Bloom-Richardson grading system, CCR: Correct classification rate, ROI: Region of interest, BCa: Breast cancer, ACM: Active contour model, UDH: Usual ductal hyperplasia; ADH: Atypical ductal hyperplasia, DCIS: Ductal carcinoma *in situ*, SVM: Support vector machine, TPR: True positive rate, FPR: False positive rate, ANN: Artificial neural network, k-NN: K-nearest neighbor, DT: Decision tree, FOV: Field-of-view

Benign versus malignant classification and distinguishing lesion subtypes

A pathologist usually inspects the breast tissue to determine if it is a benign or malignant lesion is present and also to identify the cancer type (if appropriate). Computer-aided detection (CAD) tools could help the pathologists in this task and make the results less susceptible to observer variation.

In the earliest CAD system,[44] it was shown that textural features (wavelet-based and Haralick) outperformed morphological and intensity-based features in differentiating benign from malignant cells. Later, in a study by Doyle *et al.*,[45] Gabor-based features, which are also a textural feature, achieved higher CCR compared to architectural features, and the diagnostic importance of nuclear texture in differentiating normal

from cancerous tissue has been shown. However, the wavelet-based, Haralick, and Gabor-based features are not easily interpretable to pathologists. Furthermore, the high dimension of feature vector when low number of training instances is available increases the chance of overfitting of the classifier to in hand data. Therefore, in a study by Cosatto *et al*.,[40] only the median nuclear area over an ROI and the number of large well-formed nuclei were utilized to train a linear SVM with a labeled dataset of 335 hand-picked ROIs. A sensitivity of 92% and a specificity of 80% were obtained. However, no information was provided about number of the patients from whom ROIs were picked.

In the real clinical practice, the ultimate goal is distinguishing patients with malignancy and not the individual ROIs. Pathologists judge each case based on multiple ROIs and label it accordingly. In a study by Filipczuk *et al*.,[38] a larger set from fifty patients (nine ROIs per patient) were classified as either benign or malignant using 84 features (morphological, intensity-based, and textural) extracted from isolated nuclei in each ROI. Sequential forward feature selection was used to reduce number of features, and a k-nearest neighbor was used as a classifier. The final diagnosis for each patient was obtained by a majority voting of the classification of all nine ROIs belonging to the same patient. A CCR of 100% was achieved. Considering the high CCR obtained in the study, a further investigation on this method on a larger data set is useful as no information about number of borderline cases was provided in the paper. Moreover, using majority voting could be questionable as pathologists consider a case malignant when at least one of the ROIs in the slide is positive. To address this issue, in a study by Dundar *et al*.,[12] learning with multiple instances was used to train an SVM classifier. According to the proposed classifier, a benign case was misclassified when at least one of the ROIs in a slide was classified as malignant, and a malignant case was misclassified when all of the ROIs in a slide were classified as benign. The method has been tested on a dataset of 20 well-defined ductal carcinoma *in situ* (DCIS), 12 borderline DCSI, 24 atypical ductal hyperplasia (ADH), and 39 usual ductal hyperplasia (UDH). DCSI and ADH cases were grouped as actionable (malignant) while UDH cases were considered nonactionable (benign). An overall accuracy of 87.9% was obtained while the accuracy on the borderline cases was 84.6%, comparable to that of nine pathologists on the same set (81.2% average). Despite encouraging result, for deploying such a system in clinical practice as an aid to pathologists, its additive value to a pathologist's diagnosis should also be assessed. Moreover, the proposed method (classification rule and features) is not easily interpretable to pathologists.

Unlike the above-mentioned studies, pathologists do not segment each individual nucleus within a slide; however, they analyze the scene holistically. In a study by Yang *et al*.,[46] textural features based on texton-based method were extracted without segmenting the structures in slides. CCRs of 89% and 80% were achieved in benign/malignant and multi-class (benign and two major cancer subtypes) classification.

Prognosis of Breast Cancer

The advent of WSI allows extracting quantitative features which could be helpful in predicting prognosis of BCa. In a study by Veta *et al*.,[47] an automatic nuclei segmentation algorithm[15] was utilized to extract size-related nuclear morphometric features and their prognostic value in male BCa was investigated. The results demonstrated that mean nuclear area has a significant prognostic value. In another study, Beck *et al*.[48] showed that quantitative stromal features are associated with survival. A comprehensive set of quantitative features from the BCa epithelium and stroma was extracted by utilizing a machine learning method called computational pathologist. The prognostic model was based on the extracted features and it was shown that the score from the model was strongly associated with overall survival. In addition, assessing significance of features revealed that survival was strongly related to three of the stromal features and the magnitude of association was stronger than the association of survival with epithelial features.

The presence of lymphocytic infiltration is also a prognostic indicator for in HER2 + BCa patients. Currently, pathologists do not routinely report the presence of LI as quantifying it is a tedious job. As discussed in 1–3, recently researchers worked on automatic detection of lymphocytes. However, a further step should be added to grade the extent of lymphocytic infiltration. In a study by Basavanhally *et al*.,[13] architectural features were extracted from Voronoi diagram, Delaunay triangulation, and minimum spanning tree using the centers of individually detected lymphocytes as vertices. A CCR of 90% was achieved in differentiating patients with high and low lymphocytic infiltration level. However, the dataset contained ROIs only from 12 patients. To assess the generalizability of the method, further analysis on a larger dataset is recommended.

Immunohistochemical Quantification

Currently, the standard procedure in pathology laboratories for assessment of IHC is visual examination of samples by a pathologist. The pathologist determines the status of receptors by counting positively stained cells. Hence, the procedure is tedious and prone to inter-observer variability due to subjectiveness. Computer-assisted methods in the field of IHC quantification aim at quantification of information extracted from IHC-stained samples to reduce inter-observer variability and assessment time.

HER2 receptors typically express on the cell membrane. Therefore, membrane segmentation is one of the main steps of automated methods for quantification of

HER2. A color-based approach,[49] water shedding,[50] and skeletonization[51] were used for membrane segmentation. After the segmentation stage, a group of features was extracted from the membrane and then utilized for prediction of HER2 score. The extracted features were based on membrane staining intensity,[49,51-53] membrane completeness,[49,53] or membrane color properties.[54,55] Instead of restricting the area for feature selection to the segmented membrane, Ali *et al.* utilized an algorithm which was previously used for analysis of astronomical images and extracted intensity-based features from the entire image without segmentation.[11,56] The agreement of the reviewed automated methods with the expert scoring is listed in Table 3.

In contrast to HER2 receptors, ER and PR overexpression typically results in nuclear immunoreactivity, and hence, nuclei segmentation is the first step of some of the reviewed methods listed in Table 3. The extracted features from segmented nuclei in the reviewed studies were based on nuclei staining intensity,[57,58] nuclei shape,[52,57] or nuclei color properties.[50,59] Rather than segmenting the nuclei, Amaral *et al.* proposed a method for predicting quick score values for receptor assessment based on color- and intensity-based features extracted from each pixel of test images without segmentation of nuclei.[60] As shown in Table 3, the reviewed automated methods for ER and PR assessment showed high agreement with the expert scoring.

Table 3: Automatic and semi-automatic methods for immunohistochemical quantification

	Method [reference]	Number of samples	Reference scoring	Result
ER and PR	Segmentation based			
	ImmunoRatio[50]	50 S	VC	*r*: 0.98 (combined with ki-67)
	NuclearQuant*[57]	195 C 53 ROI	Allred	κ: 0.859 ; κ (w): 0.986 κ (w): 0.981
	Definiens*	10 S	3-S	a: 100
	Aperio*[52]			*a*: 100
	Intensity analysis of segmented cell[58]	743 S	P/N	ρ: 0.74 (ER) and 0.62 (PR)
	Analyzing ratio of color components[59]	134 S	P/N	*a*: 85 (ER) and 81 (PR)
	Nonsegmentation based			
	Modified astronomical algorithms[11,56,60]	1769	Allred	*r*: 0.82
Her2	Segmentation based			
	MembraneQuant*[51] (based on membrane intensity)	309 ROI	4-S	κ: 0.872
	Definiens*	23 S	4-S	*a*: 100
	Aperio*[52] (based on membrane intensity)			*a*: 100
	Aperio* (based on membrane intensity)	77 S	4-S	*a*: 83.0 (H[a]); 73.4 (T[b]); 78.0 (CS[c])
	Using normalized color Histogram[55]			*a*: 94.6 (H[a]); 92.1 (T[b]); 92.5 (CS[c])
	Using membrane completeness and membrane intensity as features and minimum cluster distance classifier[49]	64 S	3-S	*a*: 80
	Using normalized color Histogram[54]	77 S	3-S	tb: 0.72
	Using membrane completeness and membrane intensity as features and minimum cluster distance classifier[53]	77 S	3-S	a: 81-83
	ImmunoMembrane[41]	144 S	3-S	κ (w): 0.80
	Nonsegmentation based			
	Intensity-based thresholding[61]	1648 S		*a*: 87; κ (w): 0.57
	Modified astronomical algorithms[11,56,60]	1653	4-S	*r*: 0.62
Ki-67	Segmentation based			
	ImmunoRatio[50]	50	VC	*r*: 0.98 (combined with ER and PR)
	Nonsegmentation based			
	Intensity-based thresholding[61]	1648 S	3-S	*a*: 87; κ (w): 0.57

[a]Hamamatsu NanoZoomer 2.0 HT (Hamamatsu Photonics, Bridgewater NJ), [b]Aperio ScanScope T2 (Aperio Technologies, Vista, California), [c]Aperio Scanscope CS (Aperio Technologies, Vista, California). *Commercial product; S: Slides, C: Cases, ROI: Regions of interest, VC: Visual counting, Allred: 0-8 grade system, 4-S: 4 grade system, 3-S: 3 grade system, P/N: Classification as positive or negative, ρ: Correlation coefficient *r*: R^2 for regression; *a*: Agreement percentage, κ: Cohen's kappa, κ (w): Weighted Cohen's kappa, tb: Kendall's coefficient of concordance, ER: Estrogen receptor, PR: Progesterone receptor

Similar to the methods for ER and PR assessment, quantitative assessment of Ki-67 could be done by extracting features based on either the segmented cell nuclei[50] or the percentage of stained area.[61] The agreement between the automated methods and the visual examination done by pathologist is reported in Table 3.

ImmunoRatio[50] and ImmunoMembrane[41] are two publicly available web-based applications. ImmunoRatio is a tool for quantitative assessment of ER, PR, and Ki-67 while ImmunoMembrane is an HER2 IHC analysis software. Both applications were tested and matched well with the pathologist's visual examination.[41,50]

DISCUSSION

The emphasis of this review was discussing the computer-based image analysis in breast pathology. In spite of encouraging results achieved by the reviewed studies, further progress is still required to make the CAD tools acceptable for clinical practice. For example, segmentation of severely overlapping and broken cells has not been fully addressed yet. Moreover, the low PPV of segmentation methods suggests that an FP reduction step should follow the initial segmentation. In addition, only a few studies focused on automatic segmentation and grading of tubule formation as well as distinguishing cancer subtypes; hence, further studies in these fields are required. Moreover, most of the studies attempted to extract features from the segmented objects. Further investigation of nonsegmentation-based methods in breast pathology is required as these methods avoid error propagation from the segmentation step and also mimic the human visual system which captures textural features.

Pathologists extract information from multiple ROIs and scales. Using multi-ROI and multi-scale approach to mimic the perception of pathologists could be a potential direction for future studies. Furthermore, clinicians prefer a CAD which provides physically interpretable features and classification rules; however, most of the current tools are "black box" systems.

One of the other major challenges of CAD is variability of breast tissue. Although standardization of slide preparation protocols, color normalization, noise reduction, and quality assurance programs will tackle the tissue variability problems to some extent, there is an inherent variability in the appearance of the objects within the breast tissue, which cannot be compensated. For example, the shape of epithelial cancerous nuclei may vary from almost normal-like round structure to highly irregularly shaped and enlarged nuclei with coarse and marginalized chromatin and prominent nucleoli. Moreover, the fact that different structures in breast histopathology slides may look similar decreases the specificity of CAD in detection of certain features. Another difficulty for segmentation-based CAD is separating clustered or overlapping cells. All these factors that affect adversely on the performance of CAD systems should be addressed to obtain a CAD which is robust enough to be used in the clinical practice of pathology.

In evaluation of CAD studies, inherent inter-pathologist variations should be considered. For example, in a study by Shaw *et al.*,[2] it was shown that intra- and inter-pathologist agreement for detection of pleomorphism is lower than that of IHC quantification. Similarly, automatic IHC quantification tools usually achieved higher agreement with pathologists' assessment in comparison with CADs aimed BCa grading [Tables 2 and 3].

Moreover, the additive value of CAD to pathologist's opinion should be investigated as CAD could be potentially used as "second reader." Finally, one of the major obstacles for researchers working on BCa digital slides is lack of publicly available data sets which enable them to evaluate the performance and robustness of their proposed algorithms. Having such reference databases whose ground truth was built based on a panel of expert pathologists would provide a unique opportunity for comparing different algorithms' performance against each other. Recently, two publicly available databases for mitosis detection have been introduced;[62] however, more databases containing virtual slides of different BCa types, different grades of BCa, and so on from different scanners are still required.

Financial Support and Sponsorship
Nil.

Conflicts of Interest
There are no conflicts of interest.

REFERENCES

1. Silverstein M. Where's the outrage? J Am Coll Surg 2009;208:78-9.
2. Shaw EC, Hanby AM, Wheeler K, Shaaban AM, Poller D, Barton S, *et al.* Observer agreement comparing the use of virtual slides with glass slides in the pathology review component of the POSH breast cancer cohort study. J Clin Pathol 2012;65:403-8.
3. Pantanowitz L, Valenstein PN, Evans AJ, Kaplan KJ, Pfeifer JD, Wilbur DC, *et al.* Review of the current state of whole slide imaging in pathology. J Pathol Inform 2011;2:36.
4. Ghaznavi F, Evans A, Madabhushi A, Feldman M. Digital imaging in pathology: Whole-slide imaging and beyond. Annu Rev Pathol 2013;8:331-59.
5. Al-Kofahi Y, Lassoued W, Lee W, Roysam B. Improved automatic detection and segmentation of cell nuclei in histopathology images. IEEE Trans Biomed Eng 2010;57:841-52.
6. Khan AM, Eldaly H, Rajpoot NM. A gamma-gaussian mixture model for detection of mitotic cells in breast cancer histopathology images. J Pathol Inform 2013;4:11.
7. Fatakdawala H, Xu J, Basavanhally A, Bhanot G, Ganesan S, Feldman M, *et al.* Expectation-maximization-driven geodesic active contour with overlap resolution (EMaGACOR): Application to lymphocyte segmentation on

breast cancer histopathology. IEEE Trans Biomed Eng 2010;57:1676-89.

8. Veta M, Van Diest PJ, Pluim JP. Detecting Mitotic Figures in Breast Cancer Histopathology Images. In Progress in Biomedical Optics and Imaging – Proceedings of SPIE; 2013.
9. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol 2001;23:291-9.
10. Khan AM, Rajpoot N, Treanor D, Magee D. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Trans Biomed Eng 2014;61:1729-38.
11. Ali HR, Irwin M, Morris L, Dawson SJ, Blows FM, Provenzano E, *et al.* Astronomical algorithms for automated analysis of tissue protein expression in breast cancer. Br J Cancer 2013;108:602-12.
12. Dundar MM, Badve S, Bilgin G, Raykar V, Jain R, Sertel O, *et al.* Computerized classification of intraductal breast lesions using histopathological images. IEEE Trans Biomed Eng 2011;58:1977-84.
13. Basavanhally AN, Ganesan S, Agner S, Monaco JP, Feldman MD, Tomaszewski JE, *et al.* Computerized image-based detection and grading of lymphocytic infiltration in HER2+breast cancer histopathology. IEEE Trans Biomed Eng 2010;57:642-53.
14. Haub P, Meckel T. A model based survey of colour deconvolution in diagnostic brightfield microscopy: Error estimation and spectral consideration. Sci Rep 2015;5:12096.
15. Veta M, van Diest PJ, Kornegoor R, Huisman A, Viergever MA, Pluim JP. Automatic nuclei segmentation in H & E stained breast cancer histopathology images. PLoS One 2013;8:e70221.
16. Jung C, Kim C, Chae SW, Oh S. Unsupervised segmentation of overlapped nuclei using Bayesian classification. IEEE Trans Biomed Eng 2010;57:2825-32.
17. Jung C, Kim C. Segmenting clustered nuclei using H-minima transform-based marker extraction and contour parameterization. IEEE Trans Biomed Eng 2010;57:2600-4.
18. Veillard A, Kulikova MS, Racoceanu D. Cell nuclei extraction from breast cancer histopathology images using colour, texture, scale and shape information. Diagn Pathol 2013;8:1.
19. Vink JP, Van Leeuwen MB, Van Deurzen CH, De Haan G. Efficient nucleus detector in histopathology images. J Microsc 2013;249:124-35.
20. Ali S, Madabhushi A. An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery. IEEE Trans Med Imaging 2012;31:1448-60.
21. Xu J, Janowczyk A, Chandran S, Madabhushi A. A Weighted Mean Shift, Normalized Cuts Initialized Color Gradient Based Geodesic Active Contour Model: Applications to Histopathology Image Segmentation. In SPIE Medical Imaging. International Society for Optics and Photonics; 2010.
22. Basavanhally A, Yu E, Xu J, Ganesan S, Feldman M, Tomaszewski J, *et al.* Incorporating Domain Knowledge for Tubule Detection in Breast Histopathology Using O'Callaghan Neighborhoods. In SPIE Medical Imaging. International Society for Optics and Photonics; 2011.
23. Maqlin P, Thamburaj R, Mammen JJ, Nagar AK. Automatic Detection of Tubules in Breast Histopathological Images. In Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012). Springer; 2013.
24. Nguyen K, Barnes M, Srinivas C, Chefd'hotel C. Automatic Glandular and Tubule Region Segmentation in Histological Grading of Breast Cancer. In SPIE Medical Imaging. International Society for Optics and Photonics; 2015.
25. Belsare AD, Mushrif MM, Pangarkar MA, Meshram N. Breast histopathology image segmentation using spatio-colour-texture based graph partition method. J Microsc 2016;262:260-73.
26. Irshad H. Automated mitosis detection in histopathology using morphological and multi-channel statistics features. J Pathol Inform 2013;4:10.
27. Roullier V, Lézoray O, Ta VT, Elmoataz A. Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization. Comput Med Imaging Graph 2011;35:603-15.
28. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In International Conference on Medical Image Computing and Computer-assisted Intervention. Springer; 2013.
29. Malon CD, Cosatto E. Classification of mitotic figures with convolutional neural networks and seeded blob features. J Pathol Inform 2013;4:9.
30. Irshad H, Jalali S, Roux L, Racoceanu D, Hwee LJ, Naour GL, Capron F. Automated mitosis detection using texture, SIFT features and HMAX biologically inspired approach. J Pathol Inform 2013;4:12.
31. Khan AM, El-Daly H, Simmons E, Rajpoot NM. HyMaP: A hybrid magnitude-phase approach to unsupervised segmentation of tumor areas in breast cancer histology images. J Pathol Inform 2013;4:1.
32. Mercan E, Aksoy S, Shapiro LG, Weaver DL, Brunye T, Elmore JG. Localization of Diagnostically Relevant Regions of Interest in Whole Slide Images. In Pattern Recognition (ICPR), 2014 22nd International Conference on; 2014.
33. Peikari M, Gangeh MJ, Zubovits J, Clarke G, Martel AL. Triaging diagnostically relevant regions from pathology whole slides of breast cancer: A texture based approach. IEEE Trans Med Imaging 2016;35:307-15.
34. Petushi S, Katsinis C, Coward C, Garcia F, Tozeren A. Automated Identification of Microstructures on Histology Slides. In Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on 2004. IEEE.
35. Veta M, Huisman A, Viergever MA, van Diest PJ, Pluim JP. Marker-Controlled Watershed Segmentation of Nuclei in H & E Stained Breast Cancer Biopsy Images. In Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on 2011. IEEE.
36. Racoceanu D, Capron F. Towards semantic-driven high-content image analysis: An operational instantiation for mitosis detection in digital histopathology. Comput Med Imaging Graph 2015;42:2-15.
37. Dalle JR, Li H, Huang CH, Leow WK, Racoceanu D, Putti TC. Nuclear pleomorphism scoring by selective cell nuclei detection. In WACV: IEEE; 2009.
38. Filipczuk P, Kowal M, Obuchowicz A. Multi-label fast marching and seeded watershed segmentation methods for diagnosis of breast cancer cytology. Conf Proc IEEE Eng Med Biol Soc 2013;2013:7368-71.
39. Basavanhally A, Ganesan S, Feldman M, Shih N, Mies C, Tomaszewski J, *et al.* Multi-field-of-view framework for distinguishing tumor grade in ER+breast cancer from entire histopathology slides. IEEE Trans Biomed Eng 2013;60:2089-99.
40. Cosatto E, Miller M, Graf HP, Meyer JS. Grading Nuclear Pleomorphism on Histological Micrographs. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on 2008. IEEE.
41. Tuominen VJ, Tolonen TT, Isola J. ImmunoMembrane: A publicly available web application for digital image analysis of HER2 immunohistochemistry. Histopathology 2012;60:758-67.
42. Dalle JR, Leow WK, Racoceanu D, Tutac AE, Putti TC. Automatic Breast Cancer Grading of Histopathological Images. In 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE; 2008.
43. Petushi S, Garcia FU, Haber MM, Katsinis C, Tozeren A. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. BMC Med Imaging 2006;6:14.
44. Weyn B, van de Wouwer G, van Daele A, Scheunders P, van Dyck D, van Marck E, *et al.* Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. Cytometry 1998;33:32-40.
45. Doyle S, Agner S, Madabhushi A, Feldman M, Tomaszewski J. Automated Grading of Breast Cancer Histopathology Using Spectral Clustering with Textural and Architectural Image Features. In Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on 2008. IEEE.
46. Yang L, Chen W, Meer P, Salaru G, Goodell LA, Berstis V, *et al.* Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens. IEEE Trans Inf Technol Biomed 2009;13:636-44.
47. Veta M, Kornegoor R, Huisman A, Verschuur-Maes AH, Viergever MA, Pluim JP, *et al.* Prognostic value of automatically extracted nuclear morphometric features in whole slide images of male breast cancer. Mod Pathol 2012;25:1559-65.
48. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, *et al.* Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci Transl Med 2011;3:108ra113.
49. Gavrielides MA, Masmoudi H, Petrick N, Myers KJ, Hewitt SM. Automated Evaluation of HER-2/neu Immunohistochemical Expression in Breast Cancer Using Digital Microscopy. In 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Proceedings, ISBI; 2008.
50. Tuominen VJ, Ruotoistenmäki S, Viitanen A, Jumppanen M, Isola J. ImmunoRatio: A publicly available web application for quantitative image analysis of estrogen receptor (ER), progesterone receptor (PR), and Ki-67.

Breast Cancer Res 2010;12:R56.
51. Micsik T, Kiszler G, Szabó D, Krecsák L, Hegedus C, Tibor K, *et al.* Computer aided semi-automated evaluation of HER2 immunodetection – A robust solution for supporting the accuracy of anti HER2 therapy. Pathol Oncol Res 2015;21:1005-11.
52. Lloyd MC, Allam-Nandyala P, Purohit CN, Burke N, Coppola D, Bui MM. Using image analysis as a tool for assessment of prognostic and predictive biomarkers for breast cancer: How reliable is it? J Pathol Inform 2010;1:29.
53. Masmoudi H, Hewitt SM, Petrick N, Myers KJ, Gavrielides MA. Automated quantitative assessment of HER-2/neu immunohistochemical expression in breast cancer. IEEE Trans Med Imaging 2009;28:916-25.
54. Keller B, Chen W, Gavrielides MA. Quantitative assessment and classification of tissue-based biomarker expression with color content analysis. Arch Pathol Lab Med 2012;136:539-50.
55. Keay T, Conway CM, O'Flaherty N, Hewitt SM, Shea K, Gavrielides MA. Reproducibility in the automated quantitative assessment of HER2/neu for breast cancer. J Pathol Inform 2013;4:19.
56. Gertych A, Mohan S, Maclary S, Mohanty S, Wawrowsky K, Mirocha J, *et al.* Effects of tissue decalcification on the quantification of breast cancer biomarkers by digital image analysis. Diagn Pathol 2014;9:213.
57. Krecsák L, Micsik T, Kiszler G, Krenács T, Szabó D, Jónás V, *et al.* Technical note on the validation of a semi-automated image analysis software application for estrogen and progesterone receptor detection in breast cancer. Diagn Pathol 2011;6:6.
58. Rexhepaj E, Brennan DJ, Holloway P, Kay EW, McCann AH, Landberg G, *et al.* Novel image analysis approach for quantifying expression of nuclear proteins assessed by immunohistochemistry: Application to measurement of oestrogen and progesterone receptor levels in breast cancer. Breast Cancer Res 2008;10:R89.
59. Sharangpani GM, Joshi AS, Porter K, Deshpande AS, Keyhani S, Naik GA, *et al.* Semi-automated imaging system to quantitate estrogen and progesterone receptor immunoreactivity in human breast cancer. J Microsc 2007;226(Pt 3):244-55.
60. Amaral T, McKenna SJ, Robertson K, Thompson A. Classification and immunohistochemical scoring of breast tissue microarray spots. IEEE Trans Biomed Eng 2013;60:2806-14.
61. Konsti J, Lundin M, Joensuu H, Lehtimäki T, Sihto H, Holli K, *et al.* Development and evaluation of a virtual microscopy application for automated assessment of Ki-67 expression in breast cancer. BMC Clin Pathol 2011;11:3.
62. Veta M, van Diest PJ, Willems SM, Wang H, Madabhushi A, Cruz-Roa A, *et al.* Assessment of algorithms for mitosis detection in breast cancer histopathology images. Med Image Anal 2015;20:237-48.

APPENDIX

Appendix 1

The advanced search option of the databases was used to find the articles. The search was limited to human studies. Endnote was used for reference management. After combining all references and omitting non-English references, duplicated studies were omitted by using a built-in function in Endnote. Nonoriginal studies (e.g., review paper, abstract paper or report), undetected duplicates, and nonrelevant studies were excluded based on scanning the article's title and abstracts. Then, included papers are downloaded and fully studied and a few of them were further excluded in case they were not original or relevant to the topic of the review.

The following statement was used to search Scopus:

TITLE-ABS-KEY ("breast") AND (TITLE-ABS-KEY ["virtual slide"] OR TITLE-ABS-KEY ["whole slide"] OR TITLE-ABS-KEY ["digital pathology"] OR TITLE-ABS-KEY ["digital histopathology"] OR TITLE-ABS-KEY ["whole-slide"] OR TITLE-ABS-KEY ["digitized histopathology"] OR TITLE-ABS-KEY ["digital slide"] OR TITLE-ABS-KEY ["digitized slide"] OR TITLE-ABS-KEY ["digitized cytology"] OR TITLE-ABS-KEY ["digital cytology"] OR TITLE-ABS-KEY ["digital cytopathology"] OR TITLE-ABS-KEY ["digitized cytopathology"] OR TITLE-ABS-KEY ["cell segmentation"] OR TITLE-ABS-KEY ["nuclei segmentation"] OR TITLE-ABS-KEY ["nucleus segmentation"]).

The following statement was used to search IEEEXplore

(QT breast QT) AND (["QT virtual slide QT"] OR ["QT whole slide"] OR ["QT digital pathology"] OR ["QT digital histopathology QT"] OR ["QT whole-slide"] OR ["digital slide"] OR ["QT digital cytology QT"] OR ["QT cell segmentation QT"] OR ["QT nucleus segmentation QT] OR ["QT nuclei segmentation QT"] OR ["QT histometry QT"] OR ["QT histology image QT"] OR ["QT histopathology image QT"] OR ["QT mitotic QT]).

The following statement was used to search PubMed

("Breast") AND (["virtual slide"] OR ["whole slide"] OR ["digital pathology"] OR ["digital histopathology"] OR ["whole-slide"] OR ["digital slide"] OR ["digital cytology"] OR ["cell segmentation"] OR ["nucleus segmentation"] OR ["nuclei segmentation"] OR ["histometry"] OR ["histology image"] OR ["histopathology image"]).

Chapter 3

Bridging Chapter for

"Determining quantitative features describing appearance of challenging mitotic figures and miscounted non-mitotic objects"

Study

The study was published in Journal of Pathology Informatics, 8(1), 2017.

3-1- Introduction

Mitosis is an essential stage of the cell cycle and represents division of the nucleus. The mitotic (M) phase of a cell cycle consists of mitosis and cytokinesis cytokinesis (division of the cytoplasm) and involves the division of the mother cell into two daughter cells [1]. Therefore, number of mitotic figures, or mitotic count, in a sample represent how active tissue is [1]. In the current practice of pathology, pathologists count the mitotic figures in a selected area of the most mitotically active part of the tumour on glass slides using light microscopy in ten high power fields or per unit area (2 mm^2) [2]. The active area of the tumour, selected for counting the mitotic figures, satisfies the following criteria: (1) areas with exclusively infiltrating breast cancer (BCa), such that any in situ component is avoided; (2) the periphery of the tumour section in which active growth is most likely to occur and in areas where there is no necrosis, inflammation, or calcifications; and (3) areas with high density of mitotic figures [2]. The mitotic count is a contributing factor in BCa grading since it indicates how much the tumour cells are dividing [1]. In addition, it was shown that the mitotic count has an independent prognostic value because it measures cellular proliferation, which is related to tumour aggressiveness [2].

Previous studies assessing the magnitude of agreement between pathologists for mitotic grading are summarized in Table 1. As shown the kappa values ranged from 0.38 (fair) to 0.70 (substantial), with the average of 0.51 (moderate) when considering studies in different countries. Among these studies, in [3] and [4] object-level recognition was studied. Malon et al [3] evaluated pathologists' agreement on the recognition of individual mitotic figures rather than mitotic count. They used a set of more than 4200 candidate mitotic figures taken from 2444 high power fields in 94 breast slides and asked three pathologists from USA and Japan to classify them either as mitotic figures or non-mitotic figures. The pairwise Cohen's kappa ranged from 0.13 to 0.44 with an average of 0.38. In [4], it was shown that pathologists, especially less experienced ones, often do not agree on recognition of mitotic figures and the disagreement rate is higher for smaller mitoses.

To aid pathologists for mitotic grading, recently different digital image analysis algorithms have been used for automatic mitotic figure detection. These studies were summarized in Chapter 2. In the study presented in chapter 4, the image processing

features previously used to automatically detect mitotic figures were utilized for determining the appearance of challenging mitotic figures and also the characteristics of non-mitotic figures that were miscounted as a mitosis by pathologists. This is the first time that such analysis has been done on mitotic figures.

Table 1- Magnitude of agreement for mitotic grading among pathologists in different studies. Np and Nc show number of pathologists and number of cases respectively.

Study	Site	Year	N_P	N_C	Kappa
[5]	Greece	1982	6	158	0.42
[6]	Australia	1992	2	76	0.64
[7]	USA	1995	6	75	0.52
[8]	Australia	1995	5	50	0.70
[9]	UK	1998	7	702*	0.39
[10]	Sweden	2000	7**	93	0.46
[11]	USA	2005	5-7	10-23	0.45
[12]	Italy	2005	10	20	0.57
[3]	USA, Japan***	2012	3	94	0.38
[13]	Netherlands	2013	2	100	0.64
[4]	Netherlands	2016	3	84	0.72

* 360 familial breast cancer subjects, 114 with BRCA1 mutation, 73 with BRCA2 mutation, 528 unselected for family history.
** Seven pathologists, each in a different pathology department in southern healthcare region of Sweden.
*** object-level recognition

Pathologists should distinguish true mitotic figures from other similar components in the tissue, such as apoptotic cells, tissue artefacts, and dark nuclei. Moreover, true mitoses have a broad range of appearances. Therefore, recognition of mitotic figures and exclusion of non-mitotic figures are challenging tasks. The features of challenging mitotic figures and the characteristics of non-mitotic objects misrecognised as mitotic figures can be discussed in the training sessions with the pathologists or pathology trainees. Previous studies suggested that the reproducibility of grading could improve to some extent after various teaching strategies [14-16]. For example, in [14], it was shown that, in a teaching session, use of a decision tree, which characterizes the diagnosis using the histological features, is to some extent effective in improving the proficiency of Gleason grading of prostatic cancer by general pathologists. As another example, in [16], it was shown that sharing the features of cases on which pathologists disagreed in a training session can improve the inter-pathologist agreement for breast cancer grading when agreement is measured approximately one hour after the teaching session.

In the study presented in the next chapter, I aim to identify the quantitative features extracted from the mitotic figures or their background contributing to the recognition difficulty of such figures, and to find rules to describe the appearance of miscounted non-mitotic objects. Identifying these features may be helpful in enhancing the training provided to registrars and pathologists in recognition of mitotic figures. By providing better training, the disagreement rate in the recognition of mitotic figures can be potentially reduced. In addition, predicting less easily identifiable mitotic figures may be useful for retrieving challenging slides during training procedures of pathologists. Moreover, the significantly different features among mitotic and non-mitotic objects can be used to automate the process of mitotic counting. Automatic mitotic counting can potentially reduce the pathologists' workload. The framework proposed for analysing mitotic figures and prediction of their difficulties could be extended to other tasks in breast pathology to construct rules for describing how the challenging cases in a specific task look like.

3-2- Materials and Methods

In chapter 4, I provide the methodological approach which enabled us to extract rules for describing the characteristics of challenging mitoses and non-mitoses. While the methods and results are explained in detail in the paper (Chapter 4), some further explanatory points about the dataset are discussed here.

3-2-1- Dataset

The images were obtained from the Mitosis Atypia 2014 data set [17]. In this dataset, images were available in three zoom levels (x10, x20, and x40). The locations of mitotic and non-mitotic figures were saved in .csv file. A sample hierarchy of files for a patient in the dataset is shown in figure 1. All slides were scanned by two different scanners, Aperio Scanscope XT and Hamamatsu Nanozoomer 2.0-HT. The specifications of these two scanners are shown in Table 2.

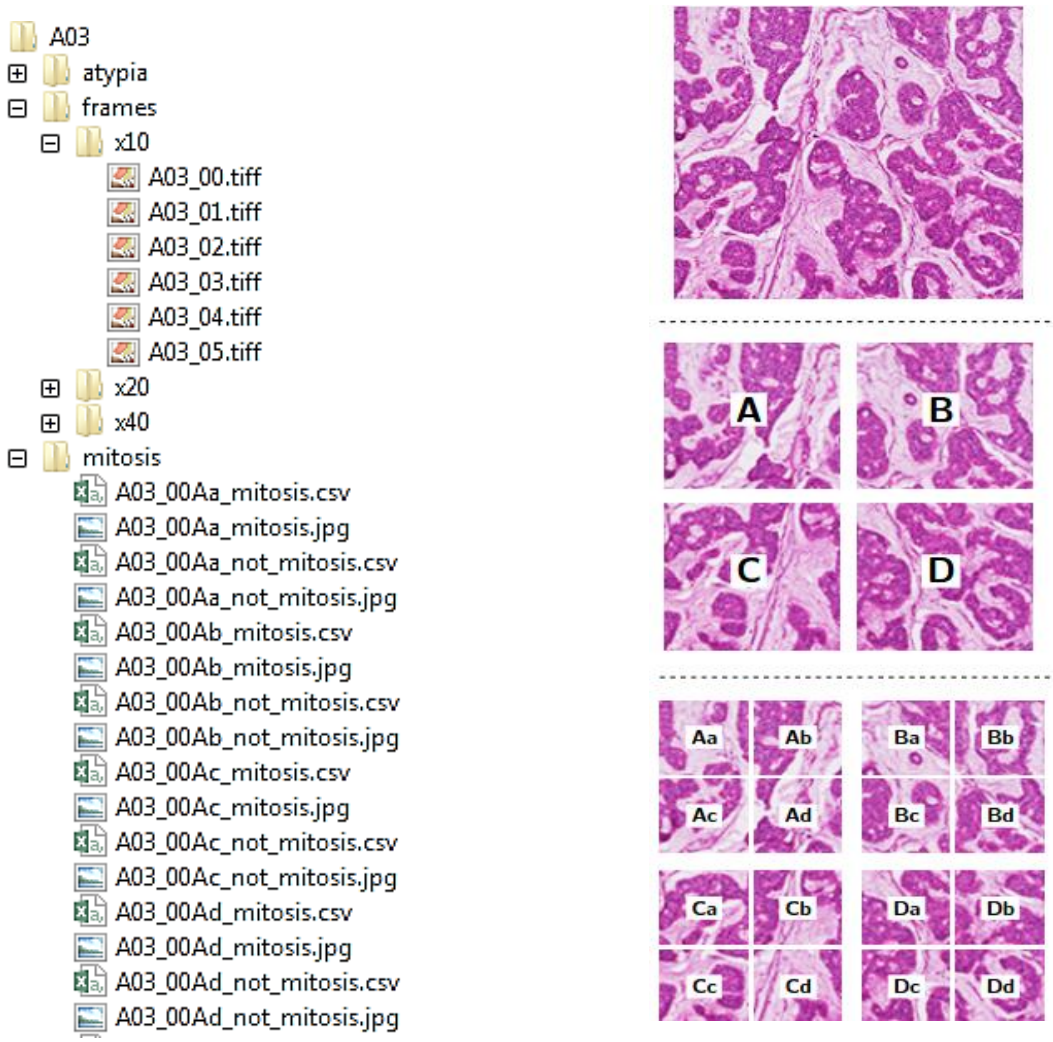


Figure 1- A sample hierarchy of files for a patient in the dataset. For each case, there are three folders, i.e. atypia, frames, and mitosis. The atypia folder contains the csv files for cytological criteria and nuclear grading and the mitosis folder contains the csv files for the location of mitotic figures. The mitoses were also overlaid on the original image and saved as a jpeg file. In the frames folder there were three subfolders, i.e. x10, x20, and x40. Each subfolder contained images in the corresponding magnification levels. For each image in the lowest magnification level (x10), four images in the intermediate magnification level (x20) and 16 images in the highest magnification level (x40) were provided. The naming protocol of images in x20 and x40 magnification factors is also shown. For naming the images in x20 A, B, C, and D were concatenated with the image name in x10 magnification factor. In naming the images in x40 a, b, c, and d were concatenated with the image name in x20 magnification factor.

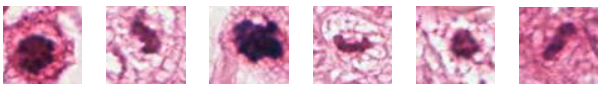
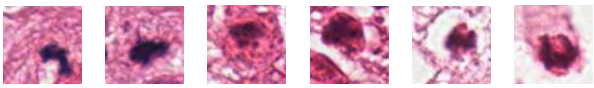
In this dataset, two pathologists have annotated objects within the images as true mitosis, probably a mitosis, or not a mitosis. In case of disagreement between these two pathologists, the opinion of a third experienced pathologist was requested, and the object has been labelled as mitosis or not mitosis according to the majority opinion.

Therefore, for some of the figures three opinions are available, while for others, only the opinions of two first pathologists were available. The dataset contains four different categories of objects with four confidence levels that show the probability of being a mitotic figure. Objects with a confidence level of 0.65 or above were considered as mitotic figures in the dataset.

Table 2- Specifications of two scanners

	Aperio Scanscope XT	**Hamamatsu Nanozoomer 2.0-HT**
Resolution at x40	0.2455 µm per pixel	0.227299 µm per pixel (horizontal) 0.227531 µm per pixel (vertical)
Dimensions of a x20 frame	1539 × 1376 pixels 755.649 × 675.616 μm^2	1663 × 1485 pixels 755.996474 × 675.76707 μm^2
Dimensions of a x40 frame	1539 × 1376 pixels 377.8245 × 337.808 μm^2	1663 × 1485 pixels 377.998237 × 337.883535 μm^2

Table 3- Sample figures from each category and the definitions of the categories

CL	**Images**	**Label**	**Definition**
1		C1	Mitoses that were recognized by both pathologists as a “true mitosis”
0.8		C2	Mitoses that were missed by one of the first two pathologists and recognized as a “true mitosis” by the third pathologist
0.65		C3	Mitoses that were labelled as “probably a mitosis” by the majority of the readers
0.2		C4	Non-mitotic objects that were recognized as a mitosis by only one reader and labelled as non-mitosis by other two pathologists

CL: confidence level in the dataset

Also in the Mitosis-Atypia dataset, three pathologists were asked to assess six criteria related to the nuclear atypia on each image. They gave a score from 1 to 3 to these criteria which are described in table 4. They were used as global descriptors of each image as one of the hypothesis that were tested in the paper (chapter 4) is whether the contextual features from the image affect the detectability of mitotic figures.

For segmentation of mitotic figures we used k-means clustering algorithm. It is a well-known clustering algorithm, which has been used in different data science applications such as load clustering [18, 19], radiologist' eye fixation clustering [, segmentation of structure within medical images, and etc.

Further information about the dataset is presented in the chapter 4, where the significantly different features among different categories are discussed thoroughly. It was found that the most challenging mitotic figures (C3) were smaller and rounder compared to other mitoses (C1 and C2). On the other hand, the sizes of the miscounted non-mitoses were identical to those of easily to identify mitoses (C1 and C2) but miscounted non-mitoses were rounder than true mitoses. Compared to intensity-based features, textural features exhibited more differences between challenging mitotic figures (C3) and the easily identifiable mitoses (C1), while the intensity-based features from chromatin channels were the most discriminative features between the miscounted non-mitoses and the easily identifiable mitotic figures (C1). Among the texture features, features extracted using Gabor filter were the most discriminative features. This is an interesting finding as the Gabor filter bank mimics human visual system and in perceptual tasks features based on Gabor filter banks exhibited high discriminative power.

Table 4- Six nuclear atypia criteria which were presented in the Mitosis-Atypia dataset

criterion	Score	Definition
Size of nuclei	1	0%-30% of tumour nuclei are bigger than normal epithelial nuclei.
	2	30%-60% of tumour nuclei are bigger than normal epithelial nuclei.
	3	More than 60% of tumour nuclei are bigger than normal epithelial nuclei.
Size of nucleoli	1	0%-30% of tumour cells have nucleoli size bigger than nucleoli of normal epithelial nuclei.
	2	30%-60% of tumour cells have nucleoli size bigger than nucleoli of normal epithelial nuclei.
	3	More than 60% of tumour cells have nucleoli size bigger than nucleoli of normal epithelial nuclei.
Density of chromatin	1	0%-30% of tumour cells have chromatin density higher than normal epithelial cells.
	2	30%-60% of tumour cells have chromatin density higher than normal epithelial cells.
	3	More than 60% of tumour cells have chromatin density higher than normal epithelial cells.
Thickness of nuclear membrane	1	0%-30% of tumour cells have nuclear membrane thickness higher than normal epithelial cells.
	2	30%-60% of tumour cells have nuclear membrane thickness higher than normal epithelial cells.
	3	More than 60% of tumour cells have nuclear membrane thickness higher than normal epithelial cells.
Regularity of nuclear contour	1	0%-30% of tumour cells have nuclear contour more irregular than normal epithelial cells.
	2	30%-60% of tumour cells have nuclear contour more irregular than normal epithelial cells.
	3	More than 60% of tumour cells have nuclear contour more irregular than normal epithelial cells.
Anisonucleosis*	1	Within the population of tumour cells, all nuclei are regular and/or nuclei size is not bigger than twice the size of normal epithelial cell nuclei.
	2	For cases that are not fitting neither with case 1 nor with case 3.
	3	Within the population of tumour cells, either nuclei size is irregular or nuclei size is bigger than 3 times the size of normal epithelial cell nuclei.

* size variation within a sample

References

[1] P. P. Rosen, Rosen's breast pathology: Lippincott Williams & Wilkins, 2001.

[2] L. Medri, A. Volpi, O. Nanni, A. M. Vecci, A. Mangia, F. Schittulli, et al., "Prognostic relevance of mitotic activity in patients with node-negative breast cancer," modern Pathology, vol. 16, p. 1067, 2003.

[3] C. Malon, E. Brachtel, E. Cosatto, H. P. Graf, A. Kurata, M. Kuroda, et al., "Mitotic figure recognition: Agreement among pathologists and computerized detector," Analytical Cellular Pathology, vol. 35, pp. 97-100, 2012.

[4] M. Veta, P. J. van Diest, M. Jiwa, S. Al-Janabi, and J. P. Pluim, "Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method," PloS one, vol. 11, p. e0161286, 2016.

[5] G. S. Delides, G. Garas, G. Georgouli, D. Jiortziotis, J. Lecca, T. Liva, et al., "Intralaboratory variations in the grading of breast carcinoma," Arch Pathol Lab Med, vol. 106, pp. 126-8, Mar 1982.

[6] J. M. Harvey, N. H. de Klerk, and G. F. Sterrett, "Histological grading in breast cancer: interobserver agreement, and relation to other prognostic factors including ploidy," Pathology, vol. 24, pp. 63-8, Apr 1992.

[7] H. F. Frierson, Jr., R. A. Wolber, K. W. Berean, D. W. Franquemont, M. J. Gaffey, J. C. Boyd, et al., "Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma," Am J Clin Pathol, vol. 103, pp. 195-8, Feb 1995.

[8] P. Robbins, S. Pinder, N. de Klerk, H. Dawkins, J. Harvey, G. Sterrett, et al., "Histological grading of breast carcinomas: a study of interobserver agreement," Hum Pathol, vol. 26, pp. 873-9, Aug 1995.

[9] S. R. Lakhani, J. Jacquemier, J. P. Sloane, B. A. Gusterson, T. J. Anderson, M. J. van de Vijver, et al., "Multifactorial analysis of differences between sporadic breast cancers and cancers involving BRCA1 and BRCA2 mutations," J Natl Cancer Inst, vol. 90, pp. 1138-45, Aug 05 1998.

[10] P. Boiesen, P. O. Bendahl, L. Anagnostaki, H. Domanski, E. Holm, I. Idvall, et al., "Histologic grading in breast cancer--reproducibility between seven

pathologic departments. South Sweden Breast Cancer Group," Acta Oncol, vol. 39, pp. 41-5, 2000.

[11] J. S. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, et al., "Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index," Mod Pathol, vol. 18, pp. 1067-78, Aug 2005.

[12] I. N. f. Q. A. o. T. B. Group, "Quality control for histological grading in breast cancer: an Italian experience," Pathologica, vol. 97, p. 1, 2005.

[13] S. Al-Janabi, H.-J. van Slooten, M. Visser, T. Van Der Ploeg, P. J. Van Diest, and M. Jiwa, "Evaluation of mitotic activity index in breast cancer using whole slide digital images," PloS one, vol. 8, p. e82576, 2013.

[14] D. Griffiths, J. Melia, L. McWilliam, R. Ball, K. Grigor, P. Harnden, et al., "A study of Gleason score interpretation in different groups of UK pathologists; techniques for improving reproducibility," Histopathology, vol. 48, pp. 655-662, 2006.

[15] P. Robbins, S. Pinder, N. de Klerk, H. Dawkins, J. Harvey, G. Sterrett, et al., "Histological grading of breast carcinomas: A study of interobserver agreement," Human Pathology, vol. 26, pp. 873-879, 1995/08/01/ 1995.

[16] A. Paradiso, I. Ellis, F. Zito, E. Marubini, S. Pizzamiglio, and P. Verderio, "Short-and long-term effects of a training session on pathologists' performance: the INQAT experience for histological grading in breast cancer," Journal of clinical pathology, vol. 62, pp. 279-281, 2009.

[17] L. Roux, D. Racoceanu, F. Capron, J. Calvo, E. Attieh, G. Le Naour, et al., "Mitos & atypia," Image Pervasive Access Lab (IPAL), Agency Sci., Technol. & Res. Inst. Infocom Res., Singapore, Tech. Rep, vol. 1, 2014.

[18] N. Haghdadi, B. Asaei, and Z. Gandomkar, "A clustering-based preprocessing on feeder power in presence of photovoltaic power plant," in Environment and Electrical Engineering (EEEIC), 2011 10th International Conference on, 2011, pp. 1-4: IEEE.

[19] N. Haghdadi, B. Asaei, and Z. Gandomkar, "Clustering-based optimal sizing and siting of photovoltaic power plant in distribution network," in Environment

and Electrical Engineering (EEEIC), 2012 11th International Conference on, 2012, pp. 266-271: IEEE.

[20] Z. Gandomkar, K. Tay, P. C. Brennan, and C. Mello-Thoms, "Recurrence quantification analysis of radiologists' scanpaths when interpreting mammograms," Medical physics, 2018.

[21] M. Delgermaa Demchig, M. Ziba Gandomkar, and C. Patrick, "Automatic Segmentation of the Dense Tissue in Digital Mammograms for BIRADS Density Categorization."

[22] Z. Gamdonkar, K. Tay, W. Ryder, P. C. Brennan, and C. Mello-Thoms, "iDensity: an automatic Gabor filter-based algorithm for breast density assessment," in Medical Imaging 2015: Image Perception, Observer Performance, and Technology Assessment, 2015, vol. 9416, p. 941607: International Society for Optics and Photonics.

Chapter 4

Determining quantitative features describing appearance of challenging mitotic figures and miscounted non-mitotic objects

This chapter has been published as:

Gandomkar, Ziba, Patrick C. Brennan, and Claudia Mello-Thoms. "Determining quantitative features describing appearance of challenging mitotic figures and miscounted non-mitotic objects." Journal of pathology informatics 8:34, 2017.

Original Article

Determining Image Processing Features Describing the Appearance of Challenging Mitotic Figures and Miscounted Nonmitotic Objects

Ziba Gandomkar[1], Patrick C. Brennan[1], Claudia Mello-Thoms[1,2]

[1]Medical Image Optimisation and Perception Research Group (MIOPeG), Discipline of Medical Radiation Sciences, Faculty of Health Sciences, University of Sydney, Australia, [2]Department of Biomedical Informatics, University of Pittsburgh School of Medicine, USA

Received: 04 March 2017 Accepted: 19 May 2017 Published: 07 September 2017

Abstract

Context: Previous studies showed that the agreement among pathologists in recognition of mitoses in breast slides is fairly modest. **Aims:** Determining the significantly different quantitative features among easily identifiable mitoses, challenging mitoses, and miscounted nonmitoses within breast slides and identifying which color spaces capture the difference among groups better than others. **Materials and Methods:** The dataset contained 453 mitoses and 265 miscounted objects in breast slides. The mitoses were grouped into three categories based on the confidence degree of three pathologists who annotated them. The mitoses annotated as "probably a mitosis" by the majority of pathologists were considered as the challenging category. The miscounted objects were recognized as a mitosis or probably a mitosis by only one of the pathologists. The mitoses were segmented using *k*-means clustering, followed by morphological operations. Morphological, intensity-based, and textural features were extracted from the segmented area and also the image patch of 63 × 63 pixels in different channels of eight color spaces. Holistic features describing the mitoses' surrounding cells of each image were also extracted. **Statistical Analysis Used:** The Kruskal–Wallis H-test followed by the Tukey-Kramer test was used to identify significantly different features. **Results:** The results indicated that challenging mitoses were smaller and rounder compared to other mitoses. Among different features, the Gabor textural features differed more than others between challenging mitoses and the easily identifiable ones. Sizes of the non-mitoses were similar to easily identifiable mitoses, but nonmitoses were rounder. The intensity-based features from chromatin channels were the most discriminative features between the easily identifiable mitoses and the miscounted objects. **Conclusions:** Quantitative features can be used to describe the characteristics of challenging mitoses and miscounted nonmitotic objects.

Keywords: Breast cancer, breast histopathology, intensity-based features, mitotic figures, textural features

Introduction

Mitotic count is one the contributing factors in Bloom–Richardson–Elston grading system[1] and also has an independent prognostic value in breast cancer.[2] However, previous studies showed that the agreement among pathologists in counting mitoses is fairly modest.[1,3-7] For example, Meyer *et al*.[1] asked a group of five to seven pathologists to examine 10–23 patients' slides and found that the Cohen's kappa for pairwise agreement on mitotic grade ranged from 0.45 to 0.67 (moderate agreement) and the average Cohen's kappa for the object-level agreement was 0.38 (fair agreement). Malon *et al*.[5] performed larger object-level agreement on 4204 figures and found that the Cohen's kappa ranged from 0.13 to 0.44 (slight to fair agreement). Furthermore, a quality control program[8] performed among 13 Italian pathologists showed that the Cohen's kappa for agreement of pathologists to the reference value ranged from 0.11 to 0.86 (slight to good agreement).

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_22_17

Address for correspondence: Ms. Ziba Gandomkar, Brain and Mind Centre, 94 Mallett Street, Camperdown NSW 2050, Australia. E-mail: ziba.gandomkar@sydney.edu.au

How to cite this article: Gandomkar Z, Brennan PC, Mello-Thoms C. Determining image processing features describing the appearance of challenging mitotic figures and miscounted nonmitotic objects. J Pathol Inform 2017;8:34.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2017/8/1/34/214164

Investigating causal agents for discrepancies in recognition of mitotic figures will be helpful in avoiding overcounting and undercounting of these figures. The wide range of appearances of mitotic figures and their similarity to other objects such as apoptoses within histopathology slides could be two potential reasons for misrecognition.

With the advent of whole-slide imaging, computer-based image analysis on the digital slides has become possible. Previously, different image processing features were used to automatically detect the mitoses on breast slides.[9] In this study, we aimed at determining image processing features differed significantly among easily identifiable mitoses, challenging mitoses, and miscounted nonmitoses. This study also seeks to compare different color spaces to determine which color spaces capture the difference among groups better than others.

Materials and Methods

Data

The images were obtained from the Mitosis-Atypia grand challenge 2014 dataset,[10] which is publicly available and contained the location of mitotic figures. The slides were from six patients and were scanned using Aperio ScanScope XT slide scanner (Aperio Technologies, Vista, CA). Each image, with the size of 1376 × 1539 pixels, covered 0.1276 mm^2 of tissue.

The dataset contained 453 mitoses and 265 nonmitoses. Two senior pathologists were asked to mark the center of mitoses and label them either as a "true mitosis" or "probably a mitosis." In case of disagreement, the opinion of a third expert pathologist was requested. Based on the annotation provided by the readers, the marked objects were classified into four categories, those recognized by both pathologists (C1), those missed by one of the first two pathologists and recognized as a "true mitosis" by the third pathologist (C2), those labeled as "probably a mitosis" by majority of the readers (C3), and objects recognized as a mitosis by only one reader and labeled as nonmitosis by other two pathologist (C4). As stated in the dataset, the first three groups were considered as true mitoses while C4 was considered as false positive markings, i.e. miscounted nonmitoses. In the original dataset, the confidence level of being mitoses for these four categories was 1, 0.8, 0.6, and 0.2, respectively. Here, we considered C1 as the easily identifiable group while C3 was considered the most challenging category because the majority of the readers could not make a decision about it confidently.

Feature extraction

The center of each mitotic figure was provided in the dataset. Hence, an image patch of 63 × 63 pixels in the neighborhood of each annotation was used and the mitotic figures were segmented using *k*-means clustering.[11,12] Mitotic figures have different appearances in different stages of mitosis. For example, during telophase, where two daughter cells are being created, two separated connected components could be detected in the image patch, and in late metaphase, a hole could exist within the extracted component. In addition, in anaphase, segmentation might result in a large connected component and a few nearby smaller regions. To address these issues, morphological closing followed by filling holes was used. In addition, the convexity (area over convex area) of the largest connected component centered at the image patch was calculated. Only the largest connected component was considered as a mitotic figure unless the convexity was low and the area of the second largest connected component was comparable to that of the largest connected component. In this case, both connected components were included in the final segmentation.

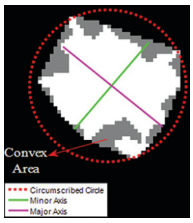
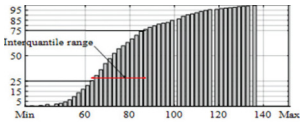
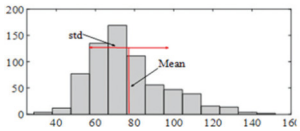
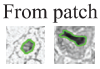
From each segmented mitotic figure, 13 shape-based features were extracted. The features are listed in Table 1. A brief intuitive description about the features along with a few sample images with high and low values of the particular feature is also shown in the table. When more than two components were segmented, major and minor axis length, and features of the second group [Table 1] were extracted from the larger component. In addition, the intensity-based features were also extracted from the largest segmented area. These features were extracted from each of the channels of red, blue, green (RGB) color space. Moreover, the image patches were converted to YDbDr, YUV, Lab, hue, saturation, and value (HSV), hue, saturation, and lightness (HSL), XYZ, and CAT02 long, medium, and short (LMS), and the features were also extracted from channels of these color spaces. A sample mitotic figure in different channels is shown in Figure 1. The intensity-based features were listed in Table 1 as well.

In addition, three groups of textural features, namely Haralick and Shanmugam texture,[13] neighborhood grey tone difference matrix,[14] and grey level run length matrix[15] were extracted from both image patches and the segmented areas. The features extracted from the segmented area describe the second and higher order statistics of the intensity value distribution within the area, while those extracted from the patch describe the local context of the objects. In addition, Gabor textural features[10] were extracted from the patches and the segmented areas' border. The textural features are also listed in Table 1. Finally, we extracted three measures for describing the contrast between the segmented area and its surrounding. The first and second contrast measures are the difference and the ratio of mean intensities of the mitotic figure and its surrounding respectively while the third measure is the difference in mean intensities normalized to the sum of the standard deviation of intensities of the mitotic figure and its surrounding.

In the original dataset, three pathologists also assessed six criteria related to the nuclear atypia in each image and gave a score from 1 to 3. These criteria were nuclei size, nucleoli size, anisonucleosis, chromatin density, and membrane thickness. Here, we used these criteria as global descriptors of each

image. We hypothesized that in addition to the appearance of the mitotic figure and its local surrounding area, the contextual features from the image could also affect the detectability of mitotic figures. As there was variability among the pathologists'

Table 1: Extracted features from each object

Feature	Sample		Description
	High	Low	
Shape-based (13) Group 1: Size-related (6) Area, perimeter, major axis length, minor axis length, convex area, circumscribed circle diameter Group 2: Other (7) Euler number, Solidity, Eccentricity, Compactness, circularity, mean distance to the nearest border point (MBD), shape factor	 Area  Solidity  Eccentricity  Circularity		 Solidity is defined as area/convex area and is lower when the area is concave Eccentricity is 0 for circle, 1 for the line, in between for ellipse Euler number shows whether there are any holes in the mask generated after segmentation Compactness, circularity, MBD, and shape factor are different measures for quantifying the compactness and circularity
Intensity-based features (21) First order statistical (6) Mean, SD, kurtosis, skewness, Maximum, min Percentiles (11) 5th, 10th, 15th, 20th, 25th, 50th, 75th, 80th, 85th, 90th, 95th percentiles Ranges (4) Minimum-Maximum, Interquartile, 5th-95th percentile, 1st-95th percentile	 Mean  Different interquartile while mean is similar		  These describe the distribution of intensity values in the segmented area. These features could be calculated from the probability density function (top) or cumulative distribution function (bottom) of intensity values
2nd order textural (40) Haralick texture (area[20] and patch [20]) Energy, entropy; contrast, dissimilarity, homogeneity; correlation, autocorrelation, information measures of correlation 1 and 2 (infh1 and infh2); cluster prominence, cluster shad, maximum probability; sum of squares, sum of average; sum of variance, difference of variance; sum of entropy, difference of entropy; inverse difference normalized, inverse difference moment normalized	 Dissimilarity From area  Homogeneity From area  Correlation From patch  infh1 From patch		Features are extracted from grey level co-occurrence matrix and capture textural information For example, energy measures the image homogeneity. Entropy indicates the randomness of the image texture. Contrast measures the local variations. Dissimilarity shows local variations but considers absolute differences. Homogeneity is inversely correlated to contrast and shows the local homogeneity. Correlation indicates the linear dependency among grey levels of neighboring pixels. Sum of average indicates the presence of punctate regions of high intensity. infh1 and infh2 measure grey level linear dependency with respect to directional entropy and randomness of spatial dependency
Higher order textural (14) GLRM (area[7] and patch [7]) SRE, LRE, GLN, RP, RLN, LGRE, HGRE	 LRE From area  RLN From area		A run is a string of consecutive pixels which have the same grey level. It gives clues about the relative distribution of intensity levels. SRE (LRE) shows the prevalence of short (long) runs, high in the fine (coarse) texture. GLN (RLN) is low when intensity levels (run lengths) are equally distributed among runs (grey levels). RP is lowest in linear texture. LGRE (HGRE) is high when there are long runs in the low (high) grey levels
Higher order textural (10) NGTDM (area[5] and patch [5]) Coarseness, NGTDM contrast, busyness, complexity, strength	 Coarseness  Complexity		Coarseness: The perceived granularity NGTDM contrast: The local fluctuations of intensity levels Busyness: The rate of changing intensity Complexity: Sum of normalized differences between intensity values taken in pairs Strength: Combines the concepts of coarseness and busyness
Gabor textural features (24) Energy features (border[12] and patch [12]) Energy contented of filtered images in 2 scales (2 and 4 pixel/cycle) and 6 orientations (S1O1, S1O2,…S1O6, S2O1, S2O2,…, S2O6)	 From patch		Border features: High when the sharp border exists in the particular orientation Patch features: High when there are linear structures in the patch and sharp borders

Contd...

Table 1: Contd...

Feature	Sample		Description
	High	Low	
Contrast (3) Three measures based on difference in mean intensity of mitotic figure and its surrounding (Co1, Co2, Co3)	 C3	Co_1=Mean intensity$_I$-Mean intensity$_S$ $Co_2 = \frac{\text{Mean intensity}_I}{\text{Mean intensity}_S}$ $Co_3 = \frac{C1}{(std_I + std_S)}$  I (Image) S (surround)

SD: Standard deviation, SRE: Short run emphasis, LRE: Long run emphasis, GLN: Grey level nonuniformity, RP: Run percentage, RLN: Run length nonuniformity, LGRE: Low grey level run emphasis, HGRE: High grey level run emphasis, NGTDM: Neighborhood grey tone difference matrix, MBD: Mean Border Distance

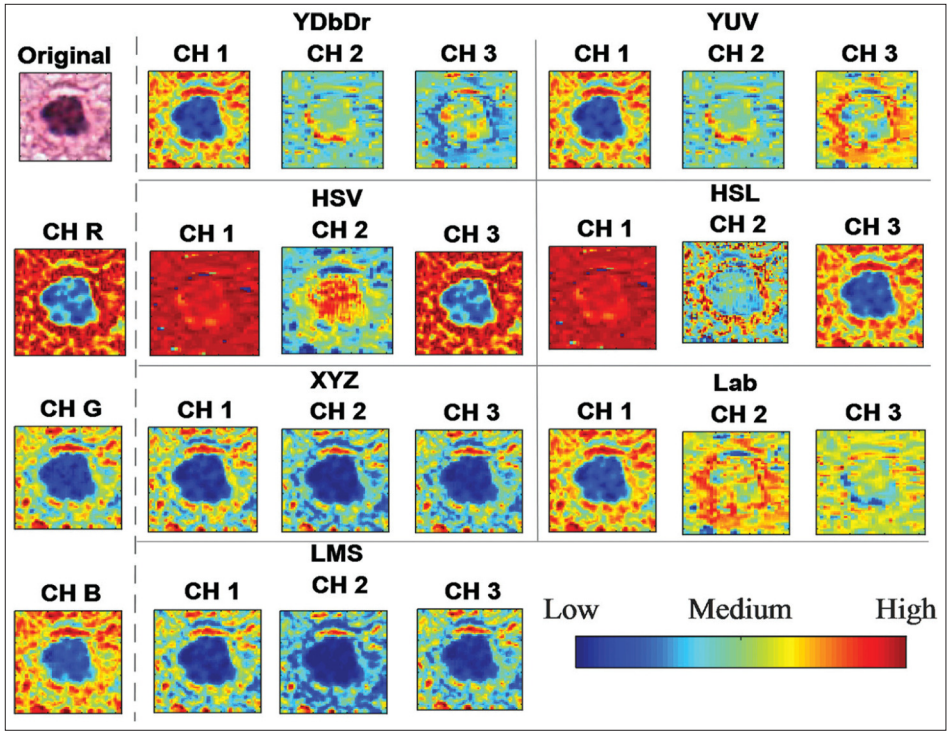


Figure 1: A mitotic figure in different color spaces

grades, the majority of the different scores was considered as a grade of each image.

Statistical analysis

To find features that differed significantly among different categories, the Kruskal–Wallis H-test was utilized and a $P < 0.05$ was considered as statistically significant. For those features differed significantly based on the result of Kruskal–Wallis H-test, pairwise comparisons were done using the rank-based Tukey-Kramer test to identify the pairs that differed significantly. Here six pairs, namely, C1 versus C2, C1 versus C3, C1 versus C4, C2 versus C3, C2 versus C4, and C4 versus C3 were possible. Moreover, we sought to determine whether there was a trend in any of the features from C4 to C1 and also examined features for trend from C1 to C3 (from the most challenging ones to easily identifiable). For this purpose, Spearman correlation was used. The statistical tests were performed using MATLAB 2016 b (Mathwork Inc., Natick, MA).

Results

Significantly different features

As stated in section 2–4, 14 sets of local features were extracted from either the segmented area or the patch for each channel. From the shape-based features, all features except solidity (χ^2 (3, 718) =7.7992, P = 0.0504) and eccentricity (χ^2 (3, 718) =3.9022, P = 0.2722) were significantly different among the groups.

A *post hoc* test showed that the miscounted objects (C4) were different from the challenging mitotic figures (C3) in terms of

circularity and mean distance to the nearest border point. None of the shape-based features resulted in a significant difference between C4 and C2, while the circularity of miscounted objects (mean rank = 382.41, standard deviation [SD] = 12.74) was significantly larger than that of easily identifiable mitotic figures (mean rank = 307.78, SD = 15.54). The mean distance to the border was also significantly different between C4 (mean rank = 389.03, SD = 12.74) and C1 (mean rank = 332.75, SD = 15.54). All size-related and shape factor features of challenging mitotic figures were significantly different from those of both C2 and C1. Among the size-related features, the magnitude of differences between the challenging mitotic figures and others were highest for the length of major axis. Shape factor was significantly different between C3 (mean rank = 314.02, SD = 15.55) and C1 (mean rank M = 378.13, SD = 15.54) as well as C3 (mean rank = 314.02, SD = 15.55) and C2 (mean rank = 406.87, SD = 21.05). Compactness was significantly different between C3 (mean rank = 321.98, SD = 15.54) and C2 (mean rank = 407.50, SD = 21.05) while circularity was significantly different between C3 (mean rank = 390.33, SD = 15.54) and C1 (mean rank = 307.78, SD = 15.54).

For different types of intensity-based features, the percentage of significantly different features is shown in Figure 2a in each color space. Overall, intensity-based features are more highlighted in Lab color space. Similarly, for eight sets of textural features, the percentage of significantly different features per each color space was shown in Figure 2b and c. As shown, overall the discriminative ability of the textural features from the image patch was the highest. Furthermore, XYZ and LMS captured the differences in the texture better than other color spaces.

Figure 3a-c shows the percentage of features that differed significantly per each feature set in the pairwise comparisons of C4 (miscounted non-mitoses) with C1, C2, and C3. As shown, textural features from patches generally captured the difference between different categories of mitotic figures and the miscounted objects better than other features. The features were ranked based on the obtained *P* value for each comparison. After considering the 50 features with the lowest *P* value, it was found that the percentile and contrast features from Db (YDbDr), U (YUV), and b (Lab) led to the lowest *P* values (that is, highest differences) for C1 versus C4 and also C2 versus C4. The Haralick and Gabor textural features extracted from the patches from all LMS and XYZ along with saturation channel led to the highest difference between C1 and C3.

Figure 4a and b presents the percentage of features that differed significantly per each feature set in the pairwise comparisons of C3 versus C1 and C3 versus C2, respectively. Ranking the features showed that the major axis length, perimeter, convex area, contrast, and Gabor features extracted from the patches led to the lowest *P* values for both C3 versus C1 and C3 versus C2 and the hue channel captured the differences better than others.

The analysis of the global features showed that all global scene descriptors except nuclei contour were significantly different among the groups. The *post hoc* test revealed the nuclei size of surrounding cells was significantly smaller for the miscounted objects (mean rank = 329.52, SD = 12.11) compared to C1 (mean rank = 408.69, SD = 20.02) and C2 (mean rank = 410.42, SD = 14.78). It was also shown that the nuclei size was significantly smaller in the challenging mitotic figures (mean rank = 326.41, SD = 14.78) in comparison with the easily identifiable ones (C1 and C2). On the other hand, the nucleoli size was significantly larger for C3 (mean rank = 410.66, SD = 13.94) compared to C1 (mean rank = 352.43, SD = 13.94). The value was lowest for C4 and the difference was significant for C4 versus C2 and C4 versus C3.

The mean anisonucleosis score of the mitoses' surrounding cells was lowest for the miscounted objects and differed significantly from all three categories of mitotic figures. The density of chromatin was significantly different for C4 (mean rank = 341.36, SD = 6.71) versus C3 (mean rank = 371.47, SD = 8.19). Finally, the membrane thickness was significantly different for C4 (mean rank = 336.02,

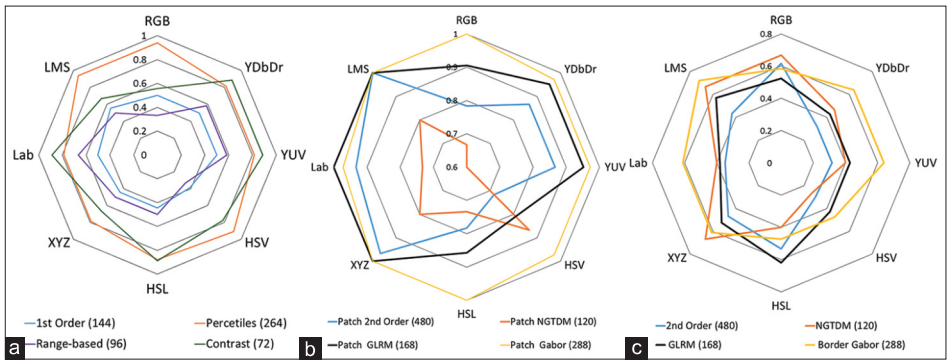


Figure 2: Percentage of features that differ significantly in each color space; total number of features in each set is shown in the parenthesis. (a) Intensity-based features; (b) textural features from patches; (c) textural features from the segmented area

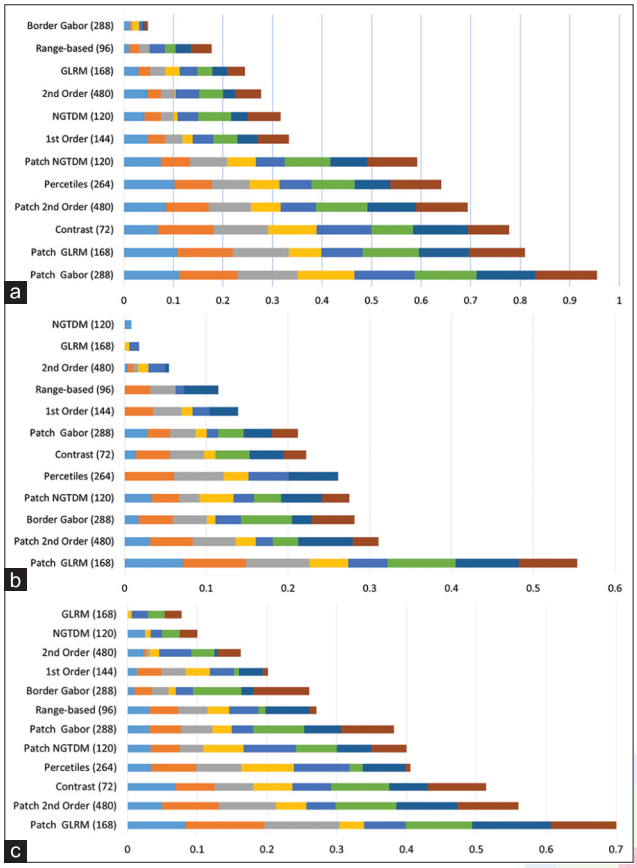


Figure 3: Percentage of each type of features that differed significantly in the pairwise comparisons of C4 (miscounted nonmitoses) with C3 (a), C2 (b), C1 (c)

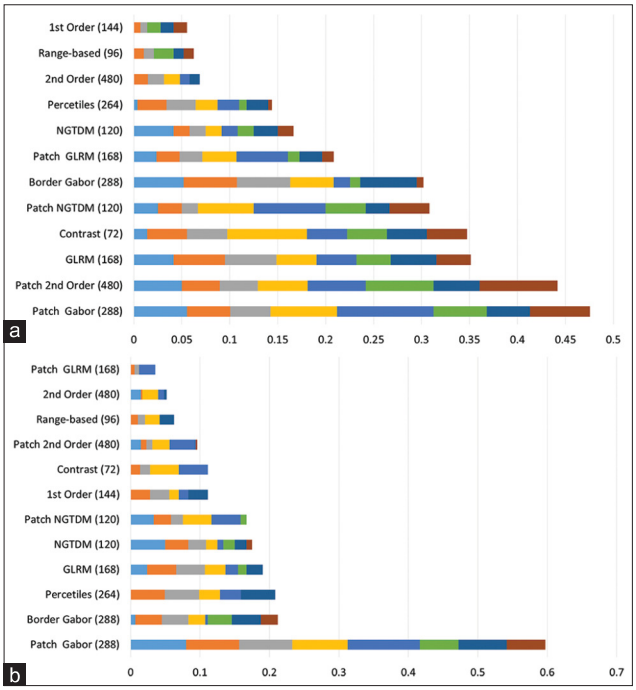


Figure 4: Percentage of features in each set that differed significantly in the pairwise comparisons of C3 with C1 (a) and C2 (b)

SD = 7.69) versus C3 (mean rank = 382.43, SD = 9.39) and C1 (mean rank = 369.36, SD = 9.39). Again none of the global descriptors led to a significant difference for C1 versus C2.

Seventeen features were significantly different among all pairwise comparisons except C1 versus C2. All these features were percentile-based, from which fifteen were from Db (YDbDr), U (YUV), b (Lab) channels (five features per channel) and two from saturation channel of HSV and HSL.

For 1241 features, the *P* value for Spearman correlation between features' value and the ordinal variable represent the groups (C1 to C4) was <0.05. The distribution of 100 features with the highest correlation across different feature types is shown in Figure 5. The corresponding Spearman correlation ranged from 0.21 to 0.32 with $P < 0.0001$. For 1059 features, there was a trend from the easily identifiable category (C1) to the challenging category (C3). Figure 5 also shows the distribution of 100 highest correlated features across different feature sets. The range of Spearman correlation for these features was 0.18–0.25 with $P < 0.0001$.

Table 2 shows a few examples of the features for which a trend was observed from C1 to C3. The examples are shown in a format of the if-then condition, for example, the first rule in the table could be interpreted as "recognition of mitotic figures with larger

Table 2: Examples of features for which a trend was observed within mitoses categories

	Increase in variable resulted in a significant level of correlation → increase or decrease in the probability of being challenging
1	↑ Major axis length → ↓challenging
2	↑ Compactness → ↑ challenging
3	↑ 95th percentile in CH b (Lab); roughly means ↑ greenness → ↓challenging
4	↑ 5th percentile of CH. Db (YDbDr); roughly means ↑ blueness → ↑challenging
5	↑ Haralick contrast CH R (RGB); roughly means ↑ local contrast in CH R → ↓challenging
6	↑ Infh1 CH R (RGB); roughly means ↑ linear dependency with respect to directional entropy → ↓challenging
7	↑ Dissimilarity in the segmented area CH R (RGB); roughly means ↑ local intensity variations → ↓challenging
8	↑ Coarseness of the segmented area CH B (RGB); roughly means ↑ granularity within the area → ↑challenging
9	↑ Business CH H (HSV); roughly means↑patches with high rate of changes in hue → ↓challenging
10	↑ Patch Gabor feature (+30°, −30° 1st Scale) CH R (RGB); roughly means↑thin linear structures → ↓challenging
11	↑ Patch Gabor feature (0° 1st Scale) CH L (Lab); roughly means↑thin linear structures→ ↓challenging
12	↑ Patch correlation CH V (HSV); roughly means ↑ linear dependency to neighboring pixels → ↑challenging

HSV: Hue, saturation, and values, RGB: Red, green, and blue. An upward arrow, ↑, and downward arrow, ↓, represent an increase and decrease in the value of the parameters (either features or probability of belonging to a certain group) respectively. An arrow to the right, →, indicates a conditional, an "if…then…" rule

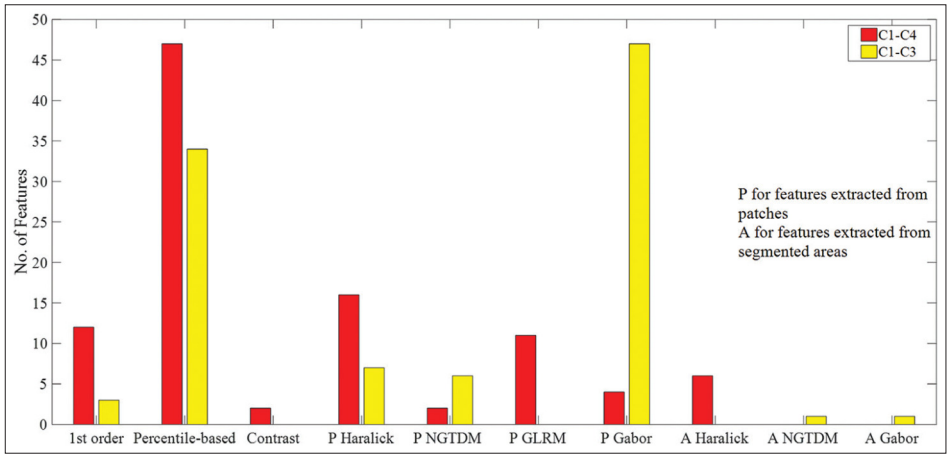


Figure 5: Distribution of features ranked the highest in terms of showing a trend from C1 to C4 (red) and C1 to C3 (yellow)

Table 3: Examples of features for which a trend was observed from easily identifiable mitoses to nonmitoses

	Increase in variable resulted in a significant level of correlation → increase or decrease in the probability of being nonmitoses
1	↑ Mean intensity CH U (YUV); roughly means ↑ blueness → ↑nonmitoses
2	↑ Range (95th-5th) CH Dr (YDbDr); roughly means ↑ range of redness → ↑nonmitoses
3	↑ Contrast CH Z (XYZ); roughly means ↑ difference in blueness of the area and its surrounding → ↑nonmitoses
4	↑ Sosvh CH S (HSV); roughly means ↑ heterogeneity of saturation values of the area → ↑nonmitoses
5	↑ Patch strength CH M (LMS); roughly means ↑ combined granularity and change rate → ↑nonmitoses
6	↑ Patch coarseness CH V (YUV); roughly means ↑ granularity of patch → ↑nonmitoses
7	↑ Patch RP CH G (RGB); roughly means ↓ linear structures → ↓nonmitoses
8	↑ Patch GLN CH B (RGB); roughly means ↑ evenness of blue values distribution across runs → ↓nonmitoses
9	↑ Patch LRE CH S (HSV); roughly means ↑ similar saturation value for a long pixel sequence → ↑nonmitoses
10	↑ HGRE CH V (YUV); roughly means↑granularity within the area → ↓nonmitoses
11	↑ Patch Gabor feature (120 2nd scale) CH L (LMS); roughly means↑thick linear structures→ ↓nonmitoses
12	↑ Patch homogeneity CH G (RGB); roughly means ↑ granularity within the area → ↑nonmitoses

LRE: Long run emphasis, RGB: Red, green, and blue, HSV: Hue, saturation, and values, LMS: Long, medium and short. An upward arrow, ↑, and downward arrow, ↓, represent an increase and decrease in the value of the parameters (either features or probability of belonging to a certain group) respectively. An arrow to the right, →, indicates a conditional, an "if…then…" rule

major axis length is less challenging." Similarly, Table 3 shows examples of observed C1–C4, which represent the 100%–20% confidence level of being mitoses in the original dataset.

Discussion

This study focused on a systematic analysis of different types of features and various color spaces to find quantitative features that differed significantly among easily identifiable mitotic figures, challenging ones, and the objects that were misrecognized as being mitotic figures by a pathologist.

Overall, the data demonstrated that quantitative features can capture the differences between appearances of the challenging mitotic figures (C3) and the easily identifiable ones (C1 and C2) and also miscounted objects (C4), and those of all three categories of mitotic figures (C1, C2, and C3). However, none of the features were significantly different between the mitotic figures missed by only one pathologist (C2) and those found by all pathologists (C1). The results further suggested that the missing the mitotic figures in category C2 could not be because of their appearance (image-based features). Furthermore, the mitotic figures were not equally distributed among the categories; C1 and C3 contained 178 mitoses and C2 contained 97 ones. The lower number of significantly different features in comparisons between C2 with other categories could be due to a lack of observations and thus lower statistical power.

It was also found that the distribution of significantly different features varies among color spaces. We can group the color spaces into four categories: RGB, which is an additive color space and often used for display technology, the Lab family, which separates luminance from chromatin and includes Lab, YUV, and YDbDr, the HSV family, which includes HSV and HSL and present color as HSVs, and color spaces such as XYZ and LMS, which are based on responsivity of three types of cones in the human eye. As shown in Figure 6, these two perceptually motivated color spaces capture the difference better than other color spaces. RBG followed in terms of capturing such differences. It should be noted that there is L, M, and S channels, named after cone cells responsive to LMS wavelength roughly correspond to RGB channels in RGB space. However, the better performance of LMS compared to RGB could be as result of being more robust to illumination conditions.

As indicated in Figure 7, the discriminative ability of various color spaces was not the same for different comparison pairs. As expected, the different color spaces in the same family perform almost similarly. Among different mitoses categories,

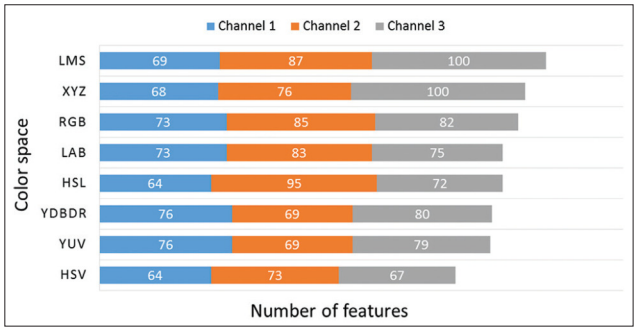


Figure 6: Distribution of significantly different features among various color spaces

the miscounted objects' appearances and less easily identifiable mitoses resulted to the highest number of significantly different features. This could be because less easily identifiable mitoses deviate from the typical appearances of mitotic figures that pathologists have in their mind, while the miscounted objects share some similarity with this typical appearance and thus they are mistaken as being true mitoses because of this similarity. Across all color spaces, the numbers of significantly different features for the three comparisons that involved the miscounted objects were higher than those of comparisons among different categories of mitoses. Therefore, in spite of differences between the challenging mitoses and the easily identifiable ones, the magnitude of similarity between them still exceeds that of mitoses versus nonmitoses. In addition, features from HSL and HSV resulted into a higher number of significantly different features for C3 versus C1 and C3 versus C2, while LMS and XYZ captured the differences among the miscounted objects and mitoses better than other color spaces.

The analysis of the shape-based features indicated that the less easily identifiable mitotic figures are rounder and smaller (based on all of the size-related measures) than other mitotic figures, while the average miscounted object's size was similar to that of an average easily identifiable mitotic figure. However, the miscounted objects were rounder (based on circularity measure). The median of circularity of the different mitoses categories is lowest for C1 and highest for C3, and the circularity of C3 was approximately similar to that of the miscounted objects. In the early metaphase, the circularity measure is high, and it could be hypothesized that most of the miscounted objects were mistaken by cells in their early metaphase. It should be noted that performing morphological closing on the binary masks corresponding to the mitotic figures slightly affects the shape-based features. Morphological closing is defined as a dilation followed by erosion, using a structuring element. The size of the structuring element determines the smoothness of edge, the size of the filled gaps, and the size of the filled holes in the output of the operator. Here, we used a disk-shaped structuring element with a radius of two pixels. The morphological closing operator resulted in a slightly larger area (on average <1.4% increase in the size of the area when only one connected component existed); however, this operation was helpful because it included a few

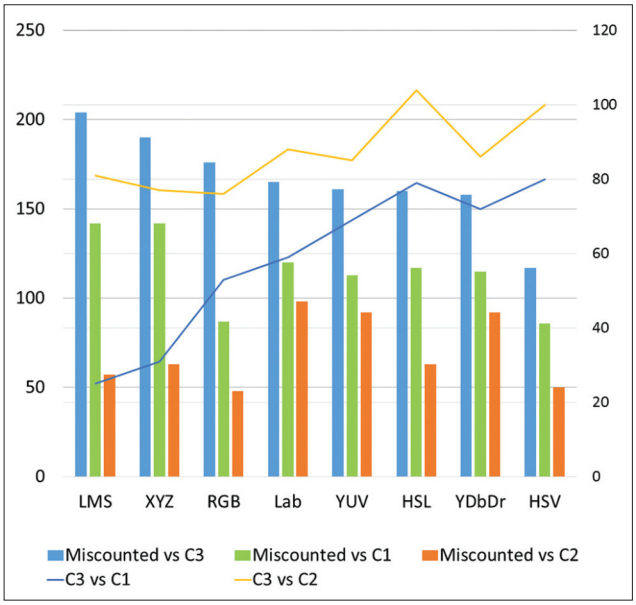


Figure 7: Number of significantly different features in various color spaces for pairwise comparisons

nearby smaller regions and thus it generated a better estimate for the mask of the mitoses. As a two-pixel disk-shaped structuring element was used here, smaller regions that are one or two pixels apart from the main connected component were included as a result of applying the morphological closing.

Also, in the *k*-means clustering algorithm, the final segmentation depends on the initial seeds. This could result in a noisy border in the final segmented region. The use of morphological opening could be beneficial in smoothing the edges and compensating the segmentation error. Among all features, the perimeter was affected the most (about 6%), however, the smoothed borders in the output of the morphological closing could lead to better estimates for the perimeter of mitotic figures. Among other size-related shape-based features, i.e. major axis length, minor axis length, convex area, and circumscribed circle diameter the observed changes were <1%. The effect of the operation on other shape-based features was negligible as they are proportional values. In addition, it should be mentioned that the size of the structuring element was small enough (two pixels) to make sure that the outer-borders of the segmented areas did not change dramatically. Moreover, changes due to the closing operation were significantly smaller than the differences between categories. For example, the changes in area due to the closing operation were about 7.5 pixels, which was significantly smaller than the differences between miscounted objects and mitotic figures (on average 90 pixels). Therefore, the closing operation did not affect the final comparisons.

As shown in Figures 3 and 4, some of the textural and intensity features differed significantly between different pairs. Ranking the obtained *P* value of significantly different features for comparison between the challenging mitoses (C3) and easily identifiable ones (C1) revealed that Gabor features from the

patches are the most discriminative features, while a high proportion of intensity-based features showed no significant differences between the two groups. Therefore, challenging mitoses are mostly different from the easily identifiable ones in their texture and shape rather than the intensity levels. Gabor filter bank mimics the human visual system, which responds selectively to orientations and scales. Utilizing the energy content of a filtered image in a particular scale and orientation as a feature resembles the same procedure. This could be the reason why these features captured information about a mitotic figure being perceptually challenging for the readers. The significantly different intensity based features were either extracted from chromatin channels of Lab family or hue channel. Compared to other color spaces, Lab family is more perceptually uniform, which means that a change in the value of their channels produces a change of the same amount in perceived color. On the other hand, for comparing nonmitoses and easily identifiable mitoses, intensity-based features were the most discriminative ones.

All global scene descriptors except nuclei contour led into significant *P* values. When the context of mitotic figures contained smaller nuclei or larger nucleoli, recognizing mitotic figures was more difficult for pathologists. It could be hypothesized that these two factors made the scene more complex. Also, it was shown that the miscounted objects were often marked in images with smaller nuclei size, smaller nucleoli size, lower anisonucleosis score, and lower membrane thickness score.

The Kruskal–Wallis H-test is a nonparametric equivalent of ANOVA and dealt the groups as categorical variables. The group variable could be also treated as an ordinal variable, as the confidence level decrease from C1 to C3. We used the Spearman correlation to find whether there is a trend in change of the feature values from the less easily identifiable category to the challenging one. A similar analysis was also utilized to explore trends from C1 to C4, which represent the 100%–20% confidence level of being mitoses in the original dataset. Identifying these rules could be beneficial in improving the training provided to pathology residents and general pathologists. Paradiso *et al*. showed that extracting the methodological skills required to enhance performance from a quality control study[8] and reviewing these skills in a training course can increase the pathologists' performance in a short-term.[16]

Our study has a number of limitations. First, we used the opinion of pathologists while assessing hematoxylin and eosin (H and E) images as the ground truth. A definition of mitosis phase could be better accomplished with a combination of Ki-67 label and H and E image data. Particularly for C3, which was labeled as "probably a mitosis" by the majority of the readers, the Ki-67 label could help in establishing the ground truth. Moreover, the pathologists who annotated the images for this data set were expert readers (as stated in the data description). The quantitative features describing the appearances of challenging mitoses for pathology trainees or less experienced observers could be different from those features extracted here. Using the methodology described here for pathology residents could help us in quantifying the perceptual difficulty in recognition of mitotic figures for them and improve their education. In this study we used all cases and readers of a publicly available dataset (Mitosis-atypia challenge 2014), which included only data from senior pathologists, and hence we did not have access to data from less experienced readers. However, this study provides a methodology to systematically analyze the quantitative data and presents preliminary results from such methodology, and we hypothesize that a similar methodology could be used to analyze data from pathology residents and study their behavior. The data suggested that none of the features differed significantly between C1 and C2. Even in the simplest visual tasks, missing a few targets are inevitable. However, as other categories outnumbered C2, the lack of statistical power could be the reason of insignificant *P* values of the pairwise comparisons involved C2. One potential venue for further work is the expansion of this study to a larger dataset to quantify the differences between C1 and C2. In addition, intrapathologist variations could exist in recognizing mitotic figures. Therefore, another potential venue for future work could be investigating whether a reader would miss mitotic figures belonging to C2 (i.e., the ones which were missed one time) in the second reading.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Meyer JS, Alvarez C, Milikowski C, Olson N, Russo I, Russo J, *et al.* Breast carcinoma malignancy grading by Bloom-Richardson system vs. proliferation index: Reproducibility of grade and advantages of proliferation index. Mod Pathol 2005;18:1067-78.
2. Medri L, Volpi A, Nanni O, Vecci AM, Mangia A, Schittulli F, *et al.* Prognostic relevance of mitotic activity in patients with node-negative breast cancer. Mod Pathol 2003;16:1067-75.
3. Veta M, van Diest PJ, Jiwa M, Al-Janabi S, Pluim JP. Mitosis counting in breast cancer: Object-Level interobserver agreement and comparison to an automatic method. PLoS One 2016;11:e0161286.
4. Tsuda H, Akiyama F, Kurosumi M, Sakamoto G, Yamashiro K, Oyama T, *et al*. Evaluation of the interobserver agreement in the number of mitotic figures breast carcinoma as simulation of quality monitoring in the Japan National Surgical Adjuvant Study of Breast Cancer (NSAS-BC) Protocol. Cancer Sci 2000;91:451-7.
5. Malon C, Brachtel E, Cosatto E, Graf HP, Kurata A, Kuroda M, *et al.* Mitotic figure recognition: Agreement among pathologists and computerized detector. Anal Cell Pathol (Amst) 2012;35:97-100.
6. Longacre TA, Ennis M, Quenneville LA, Bane AL, Bleiweiss IJ, Carter BA, *et al.* Interobserver agreement and reproducibility in classification of invasive breast carcinoma: An NCI breast cancer family registry study. Mod Pathol 2006;19:195-207.
7. Dalton LW, Page DL, Dupont WD. Histologic grading of breast carcinoma. Cancer 1994;73:2765-70.
8. INfQAoTB Group. Quality control for histological grading in breast cancer: An Italian experience. Pathologica 2005;97:1.
9. Gandomkar Z, Brennan PC, Mello-Thoms C. Computer-based image analysis in breast pathology. J Pathol Inform 2016;7:43.
10. Roux L, Racoceanu D, Capron F, Calvo J, Attieh E, Le Naour *G, et al*. MITOS and ATYPIA. 2014. Available at: http://mitos- atypia-14.grand-challenge. [Last accessed on 2017 May 10].
11. Ray S, Turi RH. Determination of number of clusters in k-means

clustering and application in colour image segmentation. In: Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques. 1999. p. 137-43.
12. Filipczuk P, Kowal M, Obuchowicz A. Multi-label fast marching and seeded watershed segmentation methods for diagnosis of breast cancer cytology. In: Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE. 2013. p. 7368-71.
13. Haralick RM, Shanmugam K. Textural features for image classification. IEEE Trans Syst Man Cybern 1973;3:610-21.
14. Amadasun M, King R. Textural features corresponding to textural properties. IEEE Trans Syst Man Cybern 1989;19:1264-74.
15. Galloway MM. Texture analysis using gray level run lengths. Comput Graph Image Process 1975;4:172-9.
16. Paradiso A, Ellis IO, Zito FA, Marubini E, Pizzamiglio S, Verderio P. Short- and long-term effects of a training session on pathologists' performance: The INQAT experience for histological grading in breast cancer. J Clin Pathol 2009;62:279-81.

Chapter 5

Bridging Chapter for

"COMPASS: Nuclear Atypia Scoring of Breast Cancer by Computer-Assisted Analysis Combined with Pathologist's Assessment"

Study

The study is submitted for publication to IEEE Journal of Biomedical and Health Informatics, 2017.

5-1- Background

The aggressive potential of a tumour determines the prognosis of breast cancer (BCa) [1] and is an estimate of the expected outcome of a BCa such as the recurrence probability (the risk of the cancer coming back after treatment) and the patient's life expectancy [1-4]. Prognostic factors are helpful in the selection of a treatment regimen; for example, expensive treatments, which may cause serious side effects, such as hormonal treatments and adjuvant chemotherapy [4], are only advisable for patients with a poor prognosis [1].

It has been shown that the Nottingham modification of the Scarff-Bloom-Richardson (NSBR) BCa grading system [2] could be used to estimate BCa prognosis [3]. However, in the routine BCa patient management, NSBR score is not being used due to the considerable inter-pathologist variability in the score which could affect patients' risk assessment for hormonal treatment and adjuvant chemotherapy [4]. The NSBR grade is the average of scores given by a pathologist to three contributing components, namely, the degree of gland formation, the magnitude of nuclear atypia, and the number of mitotic figures [2].

Table 1 summarizes the results of the previous studies [5-12] investigating the magnitude of inter-observer variability in NSBR grading and its components. As shown, the agreement on the nuclear atypia score was the weakest [5-11] with Cohen's kappa ranging from 0.19 to 0.64, with the average of 0.39 across all eight studies (fair agreement) [5-12]. Nuclear atypia score represents the cytological features of tumour cells relative to normal cells [1]. Different characteristics should be taken into account for nuclear atypia scoring, such as nuclei shape, size, margin and chromatin, and nucleoli size and appearance [1], and with the lack of quantitative or semi-quantitative measurements, it is very difficult to ascertain nuclear atypia scores [5].

The NSBR grading system was originally developed for patients with invasive ductal carcinoma, which accounts for about 80% of all invasive BCa. Invasive lobular carcinoma, which is the second most common type of BCa, has different histological and cytological characteristics compared to the invasive ductal carcinoma and the NSBR grade of lobular carcinoma lack the prognostic implications it has for ductal carcinoma [13, 14].

For invasive lobular carcinoma, nuclear atypia score is the most useful prognostic tool in comparison with the NSBR grade [15, 16]. Tubule formation and mitotic count differ slightly among patients with invasive lobular carcinoma and hence nuclear atypia score could be a more informative prognostic factor compared to the NSBR grade [15].

Nowadays, fine-needle aspiration is being used for pre-operative diagnosis of BCa. Assessing tubule formation and counting mitotic figures in ten different tumour areas is not feasible on fine-needle aspiration of the breast [17, 18]. Hence, much of the emphasis for assessment of fine-needle aspiration of the breast should be placed on nuclear atypia scoring [17, 18].

Table 1- The results of studies investigating Cohen's kappa values for inter-pathologist agreement in Nottingham histologic grading and also scoring its components. N_p and N_c show number of pathologists and number of cases respectively. For each study, the weakest agreement is shown in bold.

Study	Site	Year	N_P	N_C	Tubularity (k)	Nuclear (k)	Mitotic (k)	Overall (k)
[5]	Greece	1982	6	158	0.45	**0.19**	0.42	0.30
[6]	Australia	1992	2	76	0.65	**0.46**	0.64	0.60
[7]	USA	1995	6	75	0.64	**0.40**	0.52	0.55
[8]	Australia	1995	5	50	0.67	**0.64**	0.70	0.70
[9]	UK	1998	7	702*	0.51	**0.23**	0.39	-
[10]	Sweden	2000	7**	93	0.61	**0.44**	0.46	0.54
[11]	USA	2005	5-7	10-23	0.73	**0.27**	0.45	0.58
[12]	Italy	2005	10	20	0.74	0.47	0.57	**0.45**

* 360 familial BCa subjects, 114 with BRCA1 mutation, 73 with BRCA2 mutation, 528 unselected for family history.
** Seven pathologists, each in a different pathology department in southern healthcare region of Sweden

Owing to the importance of the nuclear atypia grade and the poor agreement among pathologists in scoring it, recently a few studies aimed at devising automatic algorithms for nuclear pleomorphism scoring [19, 20]. In the paper presented in chapter 6, COMPASS (COMputer-assisted analysis combined with Pathologist's ASSessment), a novel tool for reproducible scoring of nuclear atypia, is explained. Unlike previous nuclear atypia grading algorithms which aimed at providing an independent second opinion to the pathologists [19,20], COMPASS combines the pathologist's assessment of six criteria related to the nuclear atypia with computer-extracted features and assigns a nuclear pleomorphism score to the image based on

both subjective scores and objective features. In the previous automatic nuclear grading methods, the features were either extracted from the segmented nuclei (segmentation-based methods [19]) or from the entire tissue [20]. However, COMPASS is a hybrid segmentation-based and texture-based approach to extract the computer-related features from the digitized slides. It involves a coarse segmentation to restrict further analysis to a few regions of interest, followed by textural feature extraction from these areas.

COMPASS combines the pathologists' assessment of cytological criteria with computer-extracted features and outputs nuclear atypia score. It uses cytological features as previous studies showed that cytological grades are correlated to both the histologic nuclear score and NSBR grade. The cytological features considered in six cytological grading systems and their relationship with histological grading are shown in Table 2. Cytological criteria that were taken into account by COMPASS are shown in bold in the table. As shown, an inter-pathologist variation exists in cytological grading. Computer-extracted textural features were added in COMPASS to mitigate the inter-pathologist variation in assessment of cytological features, describe the overall appearance of nuclei, and complement the cytological characteristics for outputting nuclear atypia score.

An additional uniqueness of COMPASS is that, being a personalized model, it considers each individual's unique perceptual pattern, and eliminates systematic over- or under- estimating of each grader. Previous studies in the filed of medical image perception showed that readers have their own error making pattern [31-33] and their own visual search strategies [34, 35].

5-2- Materials and Methods

In the paper presented in chapter 6, I discuss the steps of COMPASS and the obtained results in detail. Some further explanatory points about the dataset and a few supplementary points which were not covered in the paper (chapter 6) are discussed below.

5-2-1- Dataset

The images were obtained from Mitosis Atypia 2014 data set [36]. As explained in chapter 3, in this dataset, images were available in three different magnification factors (x10, x20, and x40). All slides were scanned by two different scanners, Aperio Scanscope XT and Hamamatsu Nanozoomer 2.0-HT.

Two experienced senior pathologists were asked to provide a nuclear atypia score for 300 images of 11 patients at x20 magnification. All images were selected based on opinion of a pathologists so that they included the tumour. In case of disagreement, the opinion of a third pathologist was requested. As nuclear atypia scoring needs the evaluation of size and shape of a large population of nuclei, a wide area should be considered. Therefore, scoring was done at x20 magnification. The majority voting was used for assigning the final score for image. The output files were saved in two .csv files with the format of XXX_cna_score all.csv (scores) and XXX_cna_score decision.csv (vote of the majority) where XXX is image ID.

The nuclear atypia score could be 0-3. A score of 0 was given to an image when it does not contain enough tumour epithelial cells. Only three images were scored as 0 by all three pathologists. These images were excluded from analysis. A score of 1 is assigned to an image where the tumour epithelial nuclei are small with little increase in size in comparison with cells, have regular margins and uniform nuclear chromatin. A score of 2 is given to an image where the tumour epithelial cells are larger than normal, have open, vesicular nuclei with visible nucleoli, and both size and shape vary moderately. Finally when noticeable variation is size and shape are seen in the image, especially when large and bizarre nuclei are present, and nuclei are vesicular with prominent, often multiple nucleoli, a score of 3 is given. Four sample images per each grade are shown in Figure 1.

As stated in chapter 3, in addition to nuclear atypia scores, three junior pathologists were asked to assess six criteria related to the nuclear atypia on each image at x40 magnification factor. The criteria were nuclei size, nucleoli size, anisonucleosis (size variation within a population of nuclei), chromatin density, regularity of nuclear contour, and membrane thickness. The detailed definition of these parameters can be found in chapter 3 in section 3-2-1.

Table 2- Comparative table of cytological grading systems for BCa.

Year [Ref]	Considered features	Relation with histological grading	Kappa*
1980 **[21]**	Cell dissociation: isolated or in cluster Nuclei: **enlarged or not, anisokaryosis,** naked, budding, **level of hyperchromasia** Nucleoli: **enlarged or not** Mitosis: count	[22] AR: 66.66% [23]AR: 68% [23]SC: 0.613 [23]k=0.42 [24]AR: 77.19% [24]SC: 0.715 [24]k=0.57	[23]0.561 [24]0.65
1980 **[25, 26]**	Nuclei: **size** (I: normal, II: twice, III: three-fold), **membrane contour** (I: round & smooth, II: smooth, III: irregular), **anisonucleosis** (I: absent, II: moderate, III: marked), **level of hyperchromasia** Nucleoli: **presence of macro-nucleoli**	[23]AR:76.3% [23]SC: 0.654 [23]k=0.53 [24]AR: 70.18% [24]SC: 0.535 [24]k= 0.48	[23]0.616 [24]0.63
1994 **[27]**	Cell: **size** (I: 1-2 rbc, II: 3-4 rbc, III: ≥5rbc), pleomorphism Nuclei: **membrane contour** (I: smooth, II: Fold, III: buds or clefts), **chromatic** (I: vesicular, II: granular, II: clumped and cleared) Nucleoli: **appearance** (I: indistinct, II: noticeable, III: Prominent) Cell dissociation: isolated or in cluster	[28]AR: 78.57% [28]SC: 0.804 [22] AR: 83.33% [24]AR: 77.19% [24]SC: 0.799 [24]k= 0.62	[23]0.708 [24] 0.71
2000 **[29]**	Cell: **Size** (I: 1-2 rbc, II: 3-4 rbc, III: ≥5rbc), nuclear/cytoplasmic ratio (I: 50, II: 50-80, III: >80) Nuclei: pleomorphism, **chromatic** (I: fine, II: granular, II: coarse), **level of hyperchromasia** Nucleoli: **appearance** (I: indistinct, II: noticeable, III: Prominent) Necrosis: presence	[23]AR: 66.60% [23]SC:0.615 [23]k=0.401 [24]AR: 75.44% [24]SC: 0.686 [24]k= 0.53	[23]0.618 [24]0.59
2003 **[30]**	Cell: cellularity (I: scanty, II: moderate, II: abundant) Nuclei: **size** (I: 1-2 rbc, II: 3-5 rbc, III: >5rbc), **membrane contour** (I: smooth, II: Fold, III: buds or clefts) Nucleoli: **appearance** (I: indistinct, II: noticeable, III: Prominent and macro) Cell dissociation: isolated or in cluster Lymphatic response: degree of presence Mitosis: count	[23]AR: 72.20% [23]SC: 0.696 [23]k=0.515 [24]AR: 66.67% [24]SC: 0.744 [24]k= 0.46	[23]0.615 [24] 0.68

rbc = red blood cell

AR: agreement rate (percentage of concordance cases) between histologic and cytological grades, SC: Spearman's correlation between histologic and cytological grades, k= Cohen's kappa for measuring the agreement level between histologic and cytological grades

* kappa values for inter-pathologist agreement in cytological grading

Further information about the dataset is presented in the chapter 6.

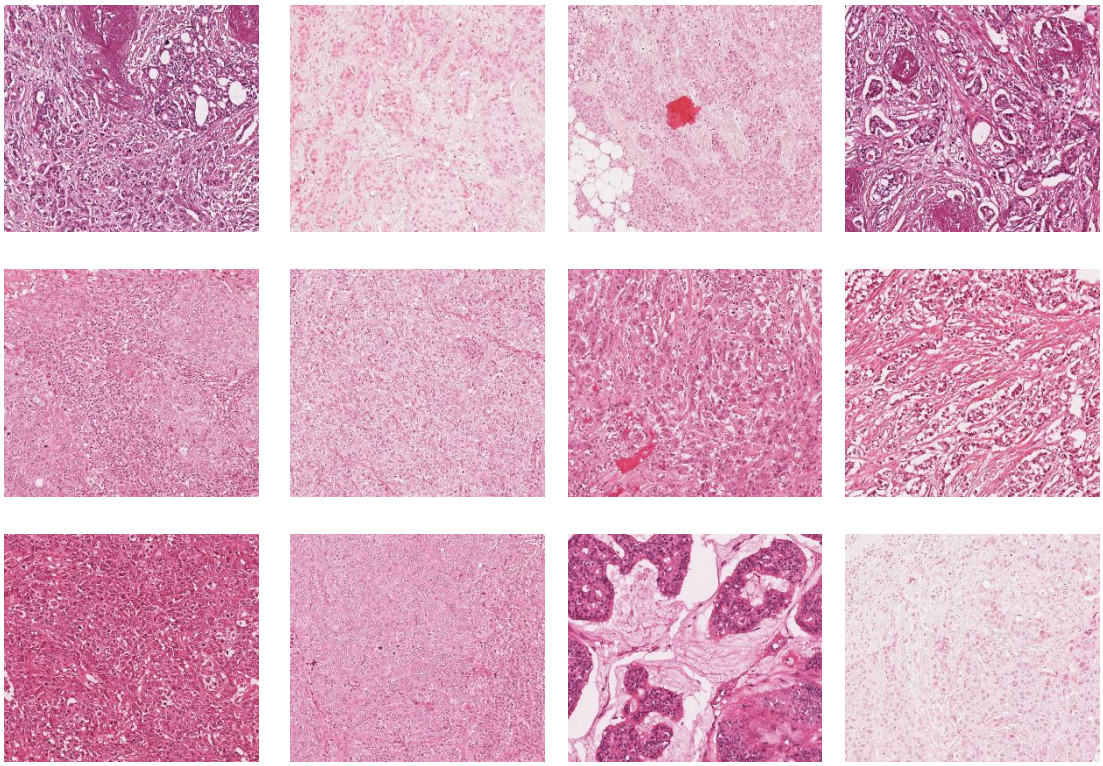


Figure 1- Sample images in x20 magnification level with the nuclear atypia score of 1(top), 2(middle), 3(bottom).

5-2-2- Supplementary points for implementation

COMPASS is a personalized tool which combines the pathologist's assessment of six atypia-related criteria with computer-extracted features and assigns a nuclear atypia score to the image based on both subjective scores and objective features. Figure 2 shows the framework for COMPASS. As the model is personalized, first the parameters of the model are estimated for each pathologist by asking the readers to assign scores to six nuclear atypia criteria on the images in the database for which the expert-consensus derived reference nuclear atypia scores are available. After this training stage, COMPASS can be used to score new images. To do so, the pathologist is asked to assign the scores to six atypia criteria and then COMPASS, whose parameters are now adjusted for the reader, gives a nuclear pleomorphism score to the image by combining the scores given by the pathologist with computer-extracted features.

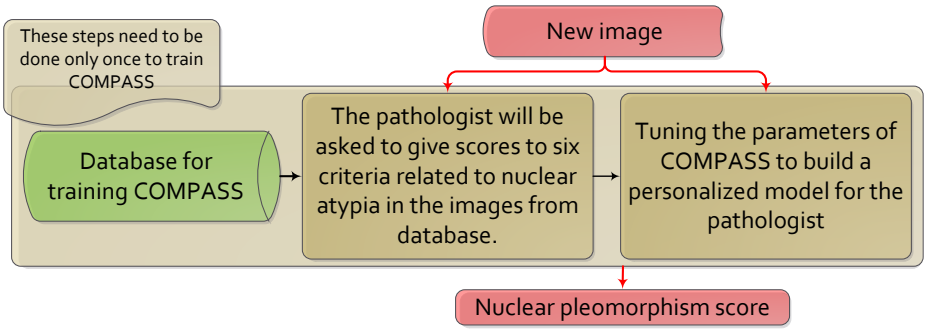


Figure 2- The framework for COMPASS.

COMPASS is comprised of two independent modules. The first module relies on the pathologist's assessment on six criteria related to nuclear atypia. These criteria were fed into a non-linear regression model, RM1, which assigns the atypia score to the input image. The second module involves another non-linear regression model, RM2, which assigns atypia scores to patches from the input image based on the computer-extracted textural features. From each input image, ten patches containing epithelial cells were automatically selected by COMPASS in a way that the patches were cantered at locations with high density of cancerous epithelial cells. For each image, RM2 assigns ten scores corresponding to ten patches selected from that image. Finally, the score given by RM1, along with minimum, median, and maximum of ten scores given by RM2 were fed into the third non-linear regression model, RM3, which produces a three-scale nuclear atypia score.

Different steps of COMPASS are further explained in the paper (chapter 6). In this section, two supplementary points about the process for automatic selection of patches and the procedure for training and testing COMPASS using the dataset are covered.

5-2-2-1- Automatic selection of patches

For automatic selection of the patches, first the stain normalization method suggested in [32, 33] was utilized to minimize inconsistencies in staining of different images and differences between images from two scanners. The output of the stain normalization step for a sample image from both scanners is shown in Figure 3.

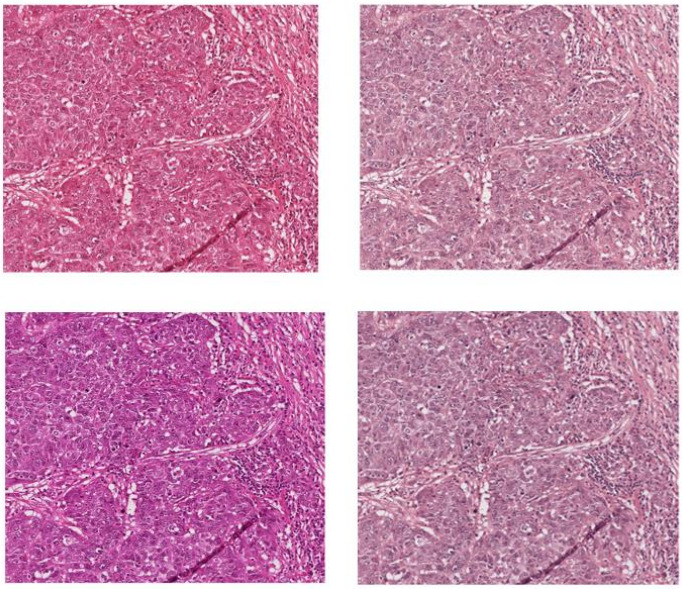


Figure 3- Original images (left column) scanned by the two scanners and the respective stain normalized images (right column).

Colour deconvolution was utilized to separate H and E channels of the stained-normalized image [37, 38]. The complement of the H channel was then processed with morphological closing (a dilation followed by an erosion) using a disk of radius 2. This was followed by filling holes within the image to generate H_P (the processed image). In the context of greyscale images, holes are areas of dark pixels surrounded by lighter pixels. Finally, the candidate locations for epithelial cells are detected by thresholding H_P and removing the connected components whose areas are less than 30 pixels. The threshold value was found empirically and set to 80. The thresholded image is called H_{Th1}.

In order to extract appropriate image patches, I needed to make sure that the imperfect areas (e.g. folded tissues) and areas with normal epithelial and lymphocyte cells were excluded from H_{Th1}. To eliminate these areas, three different masks were generated and subtracted from H_{Th1}. The first mask was obtained by thresholding the complement of the E image followed by removing all connected components whose areas were smaller than 5000 pixels. Next, the holes were filled to generate $Mask_1$. To generate $Mask_2$, the complement of the H channel was filtered by a Gabor filter bank with the wavelength of 20 pixels/cycle and 8 equally-spaced orientations. Next, the maximum filter response was recorded for each pixel. Finally, the maximum response image was thresholded to generate $Mask_2$. $Mask_3$ includes areas with normal epithelial tissue and lymphocytes which are darker, smaller in size, rounder, and without irregularities or

broken areas in the membrane. Filtering the H_P with a Laplacian of Gaussian (LoG) followed by thresholding of the filtered image was used to detect normal epithelial tissue and lymphocytes. Previously, LoG was utilized to detect epithelial cells [39] and mitotic figures [40]. The standard deviation of the filter determines the size of the structure which is detected by the LoG. Here I found the appropriate size empirically and set it to 20 pixels. The output of LoG filter was then thresholded and the connected components with an area smaller than 2000 pixels were eliminated from $Mask_3$. All three masks were subtracted from H_{Th1} to generate H_{Th2} was generated. As stated previously, I wanted to find hypercellular areas. To do so, H_{Th2} was convolved with a Gaussian filter to generate H_F. Therefore, when multiple cells are present in a neighbourhood of a pixel, it will have a high value in H_F. Finally, H_F was normalized and ten points whose intensity is at least 0.75 were randomly selected from H_F. The pairwise distance among all selected points should be greater than 100 pixels.

A sample image, along with separated H and E channels are shown in Figure 4 (a)-(c). The complement of the H channel was then processed with morphological closing (a dilation followed by an erosion) using a disk of radius 2. This was followed by filling holes within the image to generate HP (the processed image). In the context of greyscale images, holes are areas of dark pixels surrounded by lighter pixels. Finally, the candidate locations for epithelial cells are then detected by thresholding HP and removing the connected components whose areas are less than 30 pixels. The threshold value was found empirically and set to 80. The HP corresponding to the image shown in Figure 4 (a) is shown in Figure 4 (d) and the thresholded image (HTh1) overlaid on the original image using green colour is indicated in Figure 4 (e).

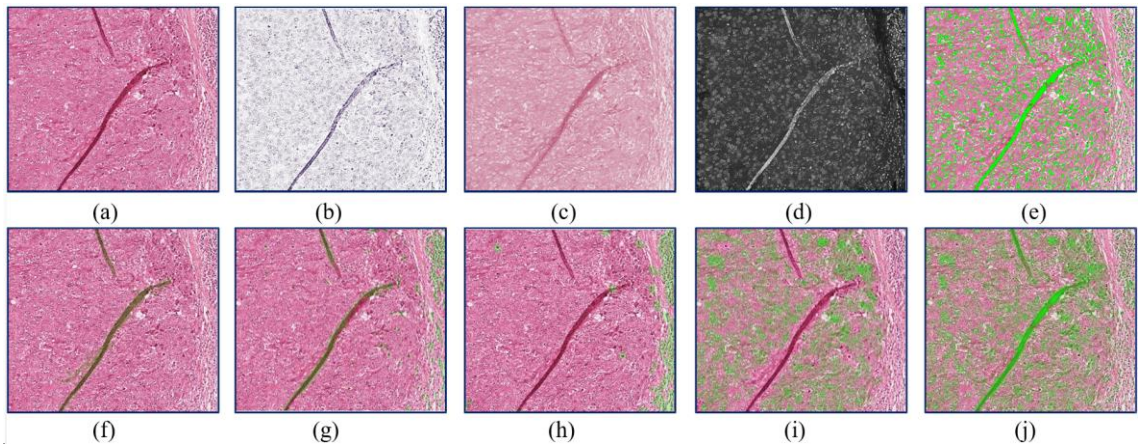


Figure 4- (a) Original image; (b) and (c) outputs of colour deconvolution separated H and E channels respectively; (d) the H channel image after being processed; (e) the thresholded image in the first step; (f-h) three masks; (i) HF (j) HF if the masks were not subtracted from the thresholded image.

5-2-3- Evaluation procedure of the COMPASS

To evaluate COMPASS' performance, I had to first train and then test the trained model on unseen instances. To do so, I used leave-one-image-out cross validation. Each time one of the images served as the test data, while the rest of the images (the training data) were utilized for estimating the parameters of COMPASS. Thus, the test data was not used for training the model and the trained model is completely blind to the test data. In the training process, the parameters of RM1, RM2, and RM3 should be estimated. As RM3 combined the output of RM1 and RM2, before training RM3, the parameters of RM1 and RM2 should be estimated. Therefore, in each iteration of leave-one-out cross validation, the training procedure had two steps: (1) training RM1 and RM2; (2) training RM3. Therefore, the training set should be divided so that part of the training set was used in the first step and part of the training set was used in the second step.

Each time (in each iteration of leave-one-out cross validation), the training data was partitioned into five subsets with roughly identical sizes and roughly the same class proportions as in the original dataset. Four subsets were utilized to estimate the parameters of RM1 and RM2. Next, the images in the remaining subset were inputted to RM1 and RM2. As stated earlier, four features were extracted from the scores given by RM1 and RM2 to each instance in this subset and used to train RM3. Finally, a score was given to the test data by the trained model. Figure 5 shows the procedure for training COMPASS. This procedure was repeated five times; each time one of the subsets was used to estimate the parameters of RM3, the rest of them were used to estimate the parameters of RM1 and RM2. Therefore, five scores were given to each test data. The median value of all these scores was assigned to each image. For training RM3, the instances from grade 1 and 3 were up-sampled by applying the Synthetic Minority Oversampling TEchnique (SMOTE) [41]. SMOTE is a common approach to oversample a dataset and it is typically used when the dataset is imbalanced. The SMOTE algorithm oversamples the minority class, which has a smaller number of samples. In oversampling, a sample from the dataset is taken, and its k nearest

neighbours are taken into account. In order to make a synthetic data point, one of these neighbours are selected randomly and a vector between this neighbour and the current data point is considered. The vector is then multiplied by a random number between 0 and 1 and added to the current point to generate the synthetic data point [36]. The numbers of nearest neighbours to use were set to 3 and 5 for grade 1 and 3 respectively and the percentage of SMOTE instances to create was set to 200% and 400%. The hyper-parameters of RM3 were also set by using Bayesian optimization.

To set the hyperparameters of RM1, RM2, RM3, in each of the five repetitions, ten-fold-cross-validation was used for the Bayesian optimization. In the training process of each one of the regression models, first, their hyper-parameters were set by using Bayesian optimization, and then the internal parameters of RM3 with the optimal hyper-parameter was estimated. The Bayesian optimization uses a 10-fold-cross validation. I did so to automate the entire parameter estimation (both hyper-parameters and parameters). Usually there is a bias because of manually selecting hyper-parameters by trial and error. However, here we have a fully automatic method. This is really important in this experiment as our sample size was really small and the manual selection of parameters may lead to overfitted models to the utilized data. In the implementation of Bayesian optimization, to achieve a result robust to partitioning noise, at every iteration, the cross-validation was repartitioned. I used MATLAB 2017a (Mathworks Inc, Natick, MA) for training and testing the regression tree ensembles (RM1, RM2, and RM3) and also for optimizing their hyperparameters.

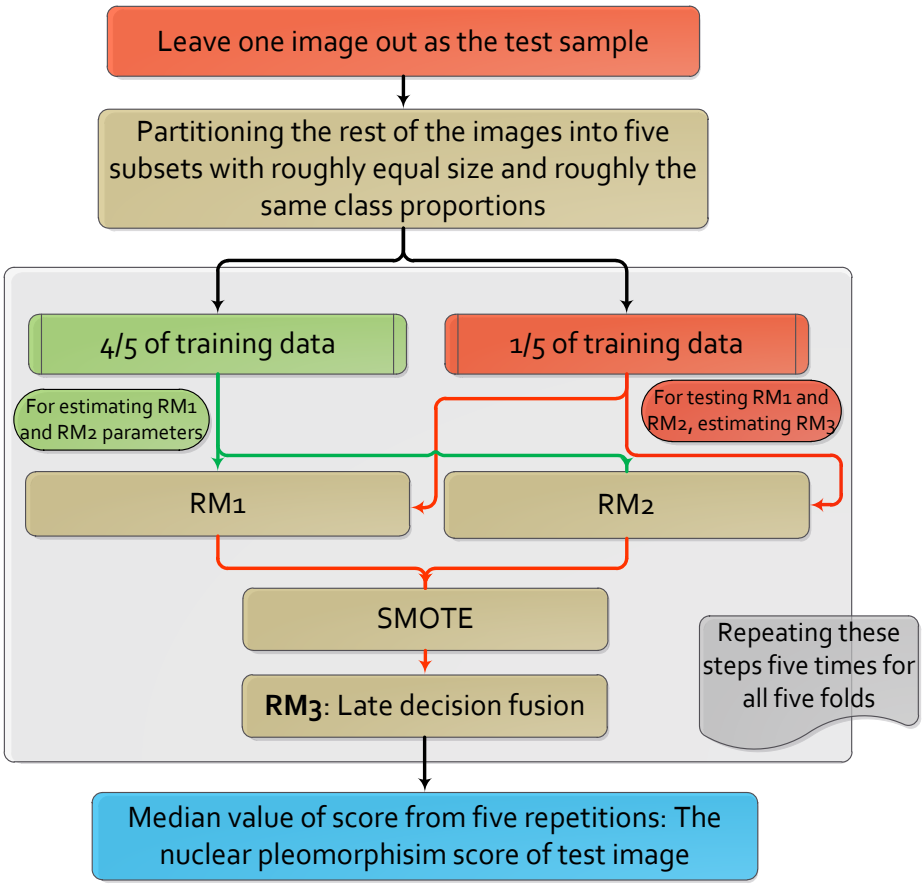


Figure 4- The procedure for training and testing COMPASS using the dataset. Each time one of the images served as the test data, while the rest of the images (the training data) were utilized for estimating the parameters of COMPASS. In each iteration of leave-one-out cross validation, training procedure had two steps: (1) training RM1 and RM2; (2) training RM3. In each iteration of leave-one-out cross validation, the training data was partitioned into five subsets with roughly identical sizes and roughly the same class proportions as in the original dataset. Four subsets were utilized to estimate the parameters of RM1 and RM2. Next, the images in the remaining subset were inputted to RM1 and RM2 and their output was used to train RM3. Finally, a score was given to the test data by the trained model. This procedure was repeated five times and each time after training all three models, the test instance was inputted and therefore, five scores were given to each test data. The median value of all these scores was assigned to each image.

References

[1] S. Mook, M. K. Schmidt, E. J. Rutgers, A. O. van de Velde, O. Visser, S. M. Rutgers, et al., "Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online Adjuvant! program: a hospital-based retrospective cohort study," The lancet oncology, vol. 10, pp. 1070-1076, 2009.

[2] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up," Histopathology, vol. 19, pp. 403-410, 1991.

[3] N. E. Roberti, "The role of histologic grading in the prognosis of patients with carcinoma of the breast," Cancer, vol. 80, pp. 1708-1716, 1997.

[4] J. M. Bueno-de-Mesquita, D. Nuyten, J. Wesseling, H. van Tinteren, S. Linn, and M. van De Vijver, "The impact of inter-observer variation in pathological assessment of node-negative breast cancer on clinical risk assessment and patient selection for adjuvant systemic treatment," Annals of oncology, vol. 21, pp. 40-47, 2009.

[5] G. S. Delides, G. Garas, G. Georgouli, D. Jiortziotis, J. Lecca, T. Liva, et al., "Intralaboratory variations in the grading of breast carcinoma," Arch Pathol Lab Med, vol. 106, pp. 126-8, Mar 1982.

[6] J. M. Harvey, N. H. de Klerk, and G. F. Sterrett, "Histological grading in breast cancer: interobserver agreement, and relation to other prognostic factors including ploidy," Pathology, vol. 24, pp. 63-8, Apr 1992.

[7] H. F. Frierson, Jr., R. A. Wolber, K. W. Berean, D. W. Franquemont, M. J. Gaffey, J. C. Boyd, et al., "Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma," Am J Clin Pathol, vol. 103, pp. 195-8, Feb 1995.

[8] P. Robbins, S. Pinder, N. de Klerk, H. Dawkins, J. Harvey, G. Sterrett, et al., "Histological grading of breast carcinomas: a study of interobserver agreement," Hum Pathol, vol. 26, pp. 873-9, Aug 1995.

[9] S. R. Lakhani, J. Jacquemier, J. P. Sloane, B. A. Gusterson, T. J. Anderson, M. J. van de Vijver, et al., "Multifactorial analysis of differences between sporadic breast cancers and cancers involving BRCA1 and BRCA2 mutations," J Natl Cancer Inst, vol. 90, pp. 1138-45, Aug 05 1998.

[10] P. Boiesen, P. O. Bendahl, L. Anagnostaki, H. Domanski, E. Holm, I. Idvall, et al., "Histologic grading in breast cancer--reproducibility between seven pathologic departments. South Sweden Breast Cancer Group," Acta Oncol, vol. 39, pp. 41-5, 2000.

[11] J. S. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, et al., "Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index," Mod Pathol, vol. 18, pp. 1067-78, Aug 2005.

[12] I. N. f. Q. A. o. T. B. Group, "Quality control for histological grading in breast cancer: an Italian experience," Pathologica, vol. 97, p. 1, 2005.

[13] M. S. Wachtel, A. Halldorsson, and S. Dissanaike, "Nottingham Grades of Lobular Carcinoma Lack the Prognostic Implications They Bear for Ductal Carcinoma1," Journal of Surgical Research, vol. 166, pp. 19-27, 2011/03/01/ 2011.

[14] P. Sinha, S. Bendall, and T. Bates, "Does routine grading of invasive lobular cancer of the breast have the same prognostic significance as for ductal cancers?," European Journal of Surgical Oncology (EJSO), vol. 26, pp. 733-737, 2000.

[15] A. L. Adams, D. C. Chhieng, W. C. Bell, T. Winokur, and O. Hameed, "Histologic grading of invasive lobular carcinoma: does use of a 2-tiered nuclear grading system improve interobserver variability?," Annals of diagnostic pathology, vol. 13, pp. 223-225, 2009.

[16] A. L. Adams, Y. Li, J. D. Pfeifer, and O. Hameed, "Nuclear Grade and Survival in Invasive Lobular Carcinoma: A Case Series with Long-term Follow-up," The breast journal, vol. 16, pp. 445-447, 2010.

[17] J. P. Phukan, A. Sinha, and J. P. Deka, "Cytological grading of breast carcinoma on fine needle aspirates and its relation with histological grading," South Asian Journal of Cancer, vol. 4, pp. 32-34, Jan-Mar 2015.

[18] C. Bansal, M. Pujani, K. Sharma, A. Srivastava, and U. Singh, "Grading systems in the cytological diagnosis of breast cancer: A review," Journal of Cancer Research and Therapeutics, vol. 10, pp. 839-845, October 1, 2014 2014.

[19] E. Cosatto, M. Miller, H. P. Graf, and J. S. Meyer, "Grading nuclear pleomorphism on histological micrographs," in Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, 2008, pp. 1-4.

[20] A. M. Khan, K. Sirinukunwattana, and N. Rajpoot, "A global covariance descriptor for nuclear atypia scoring in breast histopathology images," IEEE journal of biomedical and health informatics, vol. 19, pp. 1637-1647, 2015.

[21] J. Mouriquand and D. Pasquier, "Fine needle aspiration of breast carcinoma: a preliminary cytoprognostic study," Acta cytologica, vol. 24, pp. 153-159, 1980.

[22] P. Pandey, A. Dixit, S. Chandra, and S. Kaur, "A comparative and evaluative study of two cytological grading systems in breast carcinoma with histological grading: an important prognostic factor," Analytical Cellular Pathology, vol. 2014, 2014.

[23] D. Einstien, B. Omprakash, H. Ganapathy, and S. Rahman, "Comparison of 3-tier cytological grading systems for breast carcinoma," ISRN oncology, vol. 2014, 2014.

[24] K. Saha, G. Raychaudhuri, B. K. Chattopadhyay, and I. Das, "Comparative evaluation of six cytological grading systems in breast carcinoma," Journal of Cytology / Indian Academy of Cytologists, vol. 30, pp. 87-93, Apr-Jun 2013.

[25] E. R. Fisher, C. Redmond, and B. Fisher, "Histologic grading of breast cancer," Pathol Annu, vol. 15, pp. 239-51, 1980.

[26] P. Arul and S. Masilamani, "Comparative evaluation of various cytomorphological grading systems in breast carcinoma," Indian journal of

medical and paediatric oncology: official journal of Indian Society of Medical & Paediatric Oncology, vol. 37, p. 79, 2016.

[27] I. A. Robinson, G. McKee, A. Nicholson, P. A. Jackson, M. G. Cook, J. D'Arcy, et al., "Prognostic value of cytological grading of fine-needle aspirates from breast carcinomas," The Lancet, vol. 343, pp. 947-949, 1994/04/16/ 1994.

[28] S. Pal and M. Gupta, "Correlation between cytological and histological grading of breast cancer and its role in prognosis," Journal of Cytology, vol. 33, pp. 182-186, October 1, 2016 2016.

[29] E. Taniguchi, Q. Yang, W. Tang, Y. Nakamura, L. Shan, M. Nakamura, et al., "Cytologic grading of invasive breast carcinoma. Correlation with clinicopathologic variables and predictive value of nodal metastasis," Acta Cytol, vol. 44, pp. 587-91, Jul-Aug 2000.

[30] M. Khan, A. Haleem, H. Al Hassani, and H. Kfoury, "Cytopathological grading, as a predictor of histopathological grade, in ductal carcinoma (NOS) of breast, on air-dried Diff-Quik smears," Diagnostic cytopathology, vol. 29, pp. 185-193, 2003.

[31] Z. Gandomkar, K. Tay, P. C. Brennan, and C. Mello-Thoms, "A model based on temporal dynamics of fixations for distinguishing expert radiologists' scanpaths," in Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment, 2017, vol. 10136, p. 1013606: International Society for Optics and Photonics.

[32] Z. Gandomkar, K. Tay, W. Ryder, P. C. Brennan, and C. Mello-Thoms, "Predicting radiologists' true and false positive decisions in reading mammograms by using gaze parameters and image-based features," in Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment, 2016, vol. 9787, p. 978715: International Society for Optics and Photonics.

[33] Z. Gandomkar, K. Tay, P. C. Brennan, E. Kozuch, and C. Mello-Thoms, "Can eye-tracking metrics be used to better pair radiologists in a mammogram reading task?," Medical physics, 2018.

[34] Z. Gandomkar, P. C. Brennan, and C. Mello-Thoms, "A cognitive approach to determine the benefits of pairing radiologists in mammogram reading," in Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment, 2018, vol. 10577, p. 1057704: International Society for Optics and Photonics.

[35] Z. Gandomkar, K. Tay, W. Ryder, P. C. Brennan, and C. Mello-Thoms, "iCAP: An Individualized Model Combining Gaze Parameters and Image-based

Features to Predict Radiologists' Decisions While Reading Mammograms," IEEE transactions on medical imaging, vol. 36, no. 5, pp. 1066-1075, 2017.

[36] L. Roux, D. Racoceanu, F. Capron, J. Calvo, E. Attieh, G. Le Naour, et al., "Mitos & atypia," Image Pervasive Access Lab (IPAL), Agency Sci., Technol. & Res. Inst. Infocom Res., Singapore, Tech. Rep, vol. 1, 2014.

[37] M. Macenko, M. Niethammer, J. Marron, D. Borland, J. T. Woosley, X. Guan, et al., "A method for normalizing histology slides for quantitative analysis," in Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on, 2009, pp. 1107-1110.

[38] M. Niethammer, D. Borland, J. Marron, J. T. Woosley, and N. E. Thomas, "Appearance Normalization of Histology Slides," in MLMI, 2010, pp. 58-66.

[39] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," IEEE Transactions on Biomedical Engineering, vol. 57, pp. 841-852, 2010.

[40] H. Irshad, "Automated mitosis detection in histopathology using morphological and multi-channel statistics features," Journal of pathology informatics, vol. 4, 2013.

[41] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

Chapter 6

COMPASS: Nuclear Atypia Scoring of Breast Cancer by Computer-Assisted Analysis Combined with Pathologist's Assessment

The work covered in this chapter has been submitted as:

Gandomkar, Ziba, Patrick C. Brennan, and Claudia Mello-Thoms. "COMPASS: Nuclear Atypia Scoring of Breast Cancer by Computer-Assisted Analysis Combined with Pathologist's Assessment." Journal of Digital Imaging, 2018.

COMPASS: Computer-Assisted Analysis Combined with Pathologist's Assessment for Nuclear Atypia Scoring of Breast Cancer

Ziba Gandomkar, Patrick C. Brennan, Claudia Mello-Thoms

***Abstract*— The inter-pathologist agreement for nuclear atypia grading of breast cancer is poor due to the non-quantitative nature of the scoring. In this paper, we proposed COMPASS (COMputer-assisted analysis combined with Pathologist's ASSessment), a tool for reproducible nuclear atypia scoring. COMPASS relies on two sets of features, where the first set includes the scores given by the pathologists to six cytological characteristics related to nuclear atypia, while the second set includes textural features extracted by computer. The COMPASS's performance was evaluated using 300 images for which expert-consensus derived reference nuclear pleomorphism scores were available and were scanned by two scanners from different vendors. A personalized model was built for three junior pathologists who gave scores to six atypia-related criteria for each image. Leave-One-Out cross validation was used and COMPASS was trained and tested for each junior pathologist separately. Percentage agreement between COMPASS and the reference nuclear scores was 93.8%, 92.9%, and 93.1% for three junior pathologists. The COMPASS's performance in nuclear grading was almost identical for both scanners, with Cohen's kappa ranging from 0.80-0.86 for different pathologists and different scanners. Cohen's kappa of COMPASS were comparable to the Cohen's kappa for two senior pathologists (0.79 and 0.68) assessing the same dataset.**

***Index Terms*— Breast, Breast Cancer, Microscopy, Nuclear atypia grading, Nuclear pleomorphism grading, Pattern recognition.**

I. Introduction

BREAST cancer is a heterogeneous disease and different treatment options are available for the women diagnosed with it. Prognostic factors, which represent the aggressive potential of the tumor, could provide valuable information for the selection of a treatment regimen. For example, hormonal treatment and adjuvant chemotherapy, which are used to increase patient survival, are expensive and could cause serious side effects, and hence are only advisable for high risk patients [1, 2].

*Z. Gandomkar (email: ziba.gandomkar@sydney.edu.au) , P. C. Brennan, and C. Mello-Thoms are with Image Optimisation and Perception, Discipline of Medical Imaging and Radiation Sciences, Faculty of Health Sciences, University of Sydney, Sydney, NSW, Australia.

C. Mello-Thoms is also affiliated with the department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA.

Previous studies have shown that the Nottingham modification of the Scarff-Bloom-Richardson (NSBR) breast cancer grading system [3] provides useful prognostic information [4]. However, application of the NSBR score is still limited in routine patient management due to various reasons. Among them, the considerable inter-pathologist variability and subjectiveness are major hindrances. In [5], it was shown that inter-reader variations impact on a patient's risk assessment for hormonal treatment and adjuvant chemotherapy.

The NSBR grading system has three contributing components, namely, the degree of gland formation, the magnitude of nuclear pleomorphism, and the number of mitotic figures [3]. The overall NSBR score is an average of scores of these three components. Previous studies investigated the magnitude of inter-observer variability in NSBR grading and its components. The percentage agreement among pathologists in previous studies ranged from 43% to 74%, with Cohen's kappa ranging from 0.19 to 0.74 [6-11]. It was also shown that among these three components, the agreement on the nuclear pleomorphism score was the weakest, with percentage agreement of 55-68% and Cohen's kappa ranging from 0.27 to 0.5 [6, 9].

Nuclear pleomorphism (or atypia) score represents the variations in size, shape, and appearance of tumor cells relative to normal cells. In the clinical practice, with the lack of quantitative or semi-quantitative measurements, the pathologist must decide how to categorize a nucleus with mixed features (for example, small but with an irregular shape) and that might explain why the agreement among readers is very poor.

In addition to being a contributing factor of the NSBR grade, the nuclear atypia score might be a more useful prognostic tool compared to the overall NSBR grade for patients with invasive lobular carcinoma, as mitotic activity and tubule formation vary little in these patients[12]. Due to the importance of the nuclear atypia grade and the lack of agreement among pathologists for grading it, recently a few studies aimed at devising automatic algorithms for nuclear pleomorphism scoring [13-16].

In this paper, we propose a method for reproducible nuclear pleomorphism scoring called COMPASS (COMputer-assisted analysis combined with Pathologist's ASSessment). Unlike previous algorithms which aimed at providing an independent

second opinion to the pathologists, COMPASS combines the pathologist's assessment of six criteria related to the nuclear atypia with computer-extracted features and assigns a nuclear pleomorphism score to the image based on both subjective scores and objective features. Another novelty of COMPASS is being a hybrid segmentation-based and texture-based approach to extract the computer-related features from the digitized slides. In the previous automatic nuclear grading methods, the features were either extracted from the segmented nuclei (segmentation-based methods [13, 14]) or from the entire tissue[15]). However, COMPASS involves a coarse segmentation to restrict further analysis to a few regions of interest followed by textural feature extraction from these areas. Another uniqueness of COMPASS is that, being a personalized model, it considers each individual's unique perceptual pattern, and eliminates systematic over- or under- estimating of each grader. In [17], it was shown that some pathologists are prone to under-grading while others systematically over-grade the cases. Junior pathologists are target users for COMPASS as in general less experienced pathologists have lower agreement levels with a consensus of expert readers [11, 17] and could benefit significantly from such an algorithm. Unfortunately, due to lack of expert pathologists with subspecialty training in reading breast biopsies, many specimens are currently interpreted by less experienced or general pathologists. This paper aims at investigating possibility of improving junior pathologists' performances to a level comparable to the expert readers' performance by using computer-extracted features combined with a systematic evaluation of cytological features by the pathologists.

II. Materials and Methods

A. Dataset

Three-hundred images were obtained from the Mitosis Atypia challenge 2014 data set [18], which is publicly available. Three of the images were excluded as there was no tumor region present in them and hence no atypia grade was associated with them.

Nuclear pleomorphism scores were given by two experienced senior pathologists. In case of disagreement, a third pathologist scored the image and the final score was obtained based on a vote of the majority. Based on NSBR [3], a score of 1 is given to an image when there is little increase in the size of nuclei in comparison with normal breast epithelial cells, the outlines of nuclei are regular, and the nuclear chromatin is uniform. When the cells are larger than normal with visible nucleoli and have open vesicular nuclei, with moderate variations in size and shape among cells, a score of 2 is assigned. A score of 3 is appropriate when nuclei are vesicular with prominent, often multiple nucleoli, have noticeable variations in shape and size, and large and bizarre nuclei are present in the sample [3]. All images were scanned by two different scanners, namely, Aperio Scanscope XT and Hamamatsu Nanozoomer 2.0-HT. The pathologists graded at ×20 magnification, which covered approximately 0.511 mm^2 of tissue.

In addition, three junior pathologists were asked to evaluate six criteria related to nuclear atypia and give a score from one to three for each criterion. These criteria were nuclei size, nucleoli size, anisonucleosis (size variation within a population of nuclei), chromatin density, regularity of nuclear contour, and membrane thickness. Some of these criteria (nuclei size, nucleoli size, and regularity of nuclear contour) are explicitly mentioned in NSBR grading [3]. These criteria are also components of some other nuclear grading systems [19, 20] which tried to quantify other factors that contribute the pathologists' judgments about nuclear atypia grading. For example, in Fisher's modification of Black's nuclear grading, anisonucleosis, nuclear membrane, chromatin density, and nucleoli size are taken into account [21], while in Robinson's nuclear grading system, which showed high level of concordance with NSBR grade [20], nuclei size, nucleoli size, cell uniformity, regularity of nuclear contour, and membrane thickness were taken into account [22].

For each image at ×20 magnification, the criteria were evaluated on four images at ×40 magnification (resolution of 0.2455 µm/pixel for Aperio and horizontal resolution of 0.2273 µm/pixel and vertical resolution of 0.2275 µm/pixel for Hamamatsu) to make sure that the detailed nuclei features were visible to the junior pathologists. Hence, for an image at ×20 magnification, each pathologist gave 24 (6 criteria×4

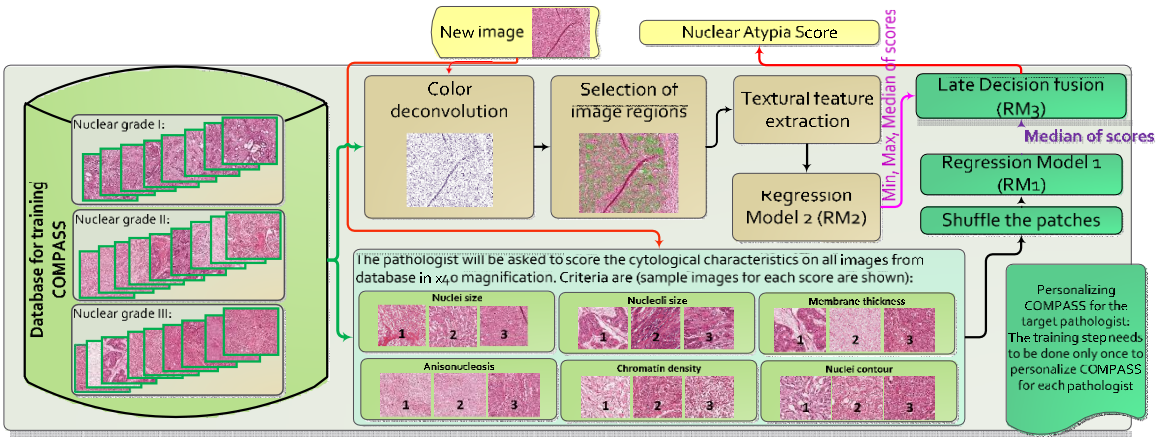


Fig. 1. The steps of COMPASS.

images) scores.

B. CAMPASS

1) Overview

The steps of COMPASS are depicted in Fig. 1. As shown, COMPASS consists of two independent modules, where the first module generates a score based on the pathologist's assessment of the four images at ×40 magnification, and the second module, which generates ten scores corresponding to ten image patches, is based on textural features from the image at ×20 magnification. In the last stage of COMPASS, the scores corresponding to each image from both modules are combined by using an ensemble of trees for regression and a single score is given to the image.

2) Computer-aided feature extraction

In this step, the patches containing epithelial cells were selected from each image and the textural features were extracted from these image patches, whose size was 251×251 pixels and were centered at locations with high density of cancerous epithelial cells. To find the centers of the patches, the stain normalization method suggested in [23] was utilized to minimize inconsistencies in staining of different images.

Color deconvolution was utilized to separate H and E channels of the stained-normalized image [23]. A sample image, along with separated H and E channels are shown in Fig. 2 (a)-(c). The complement of the H channel was then processed with morphological closing (a dilation followed by an erosion) using a disk of radius 2. This was followed by filling holes within the image to generate HP (the processed image). In the context of greyscale images, holes are areas of dark pixels surrounded by lighter pixels. Finally, the candidate locations for epithelial cells are then detected by thresholding HP and removing the connected components whose areas are less than 30 pixels. The threshold value was found empirically and set to 80. The HP corresponding to the image shown in Fig. 2 (a) is shown Fig. 2 (d) and the thresholded image (HTh1) overlaid on the original image using green color is indicated in Fig. 2 (e).

In order to extract appropriate image patches, we needed to make sure that the imperfect areas (e.g. folded tissues) and areas with normal epithelial and lymphocyte cells were excluded from HTh1. To eliminate these areas, three different masks were generated and subtracted from HTh1. The first mask was obtained by thresholding the complement of the E image followed by removing all connected components whose areas were fewer than 5000 pixels. Next, the holes were filled to generate Mask1. In order to generate Mask2, the complement of the H channel was filtered by a Gabor filter bank with the wavelength of 20 pixel/cycle and 8 equally- spaced orientations. Next, the maximum filter response was recorded for each pixel. Finally, the maximum response image was thresholded to Mask2. Mask3 includes areas with normal epithelial tissue and lymphocytes which are darker, smaller in size, rounder, and without irregularities or broken areas in their membrane. Therefore filtering the HP with a Laplacian of Gaussian (LoG) followed by thresholding of the filtered image was used. Previously, LoG was utilized to detect epithelial cells [24] and mitotic figures [25]. The standard

TABLE I
EXTRACTED FEATURES FROM EACH IMAGE PATCH

Feature Type (feature name)
First order statistics features (AVE, STD, 1st, 5th, 25th, 50th, 75th, 95th, 99th percentile of intensity)
Haralick texture features averaged over four direction for d= 3 pixels (Contrast, Correlation, Cluster Prominence, Cluster Shade, Dissimilarity, Energy, Entropy, Homogeneity, Sum of squares, Sum average, Sum entropy, Difference variance, Difference entropy, Information measure of correlation 1 and 2, Inverse difference normalized, Inverse difference moment normalized)
Local binary patterns (uniform local binary patterns with number of number of neighbours = 8)
Features from grey level run length matrix (Short run emphasis, Long run emphasis, Grey level non-uniformity, Run percentage, Run length non-uniformity, Low grey level run emphasis, High grey level run emphasis)
Gabor-based features (AVE energy of filtered image using Gabor filter bank in one scale and six orientations)
Features based on Maximum response filters (AVE energy of in eight filtered images)

AVE and STD are average and standard deviation respectively.

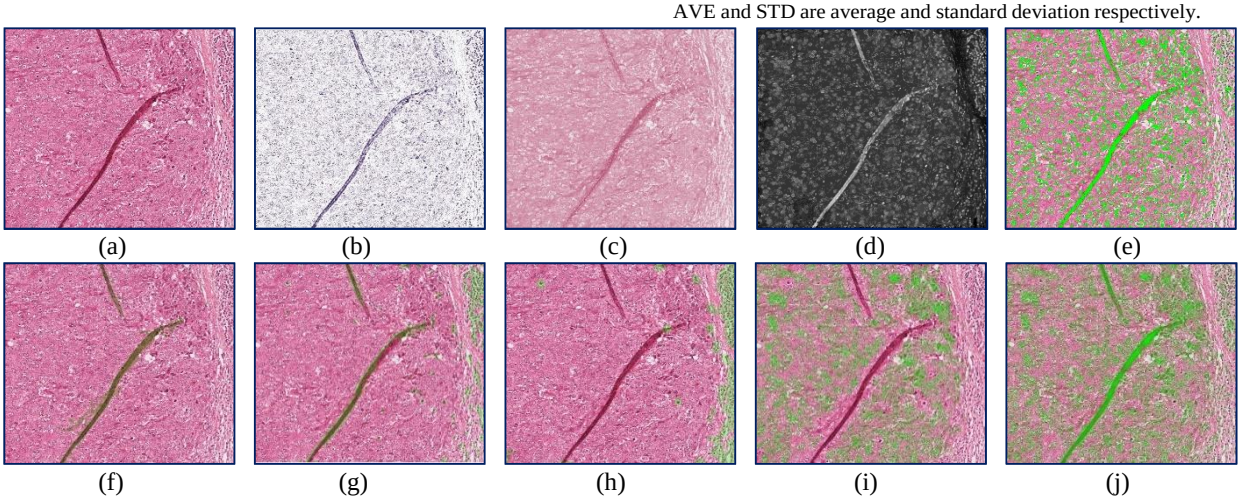


Fig. 2. (a) Original image; (b) and (c) outputs of colour deconvolution separated H and E channels respectively; (d) the H channel image after being processed; (e) the thresholded image in the first step; (f-h) three masks; (i) HF (j) HF if the masks were not subtracted from the thresholded image.

deviation of the filter determines the size of the structure which is detected by the LoG. Here we found the appropriate size empirically and set it to 20 pixels. The output of LoG filter was then thresholded and the connected components with an area less than 2000 pixels were eliminated from Mask3. All three masks were subtracted from HTh1 and HTh2 was generated. As stated previously, we want to find hypercellular areas. To do so, HTh2 was convolved with a Gaussian filter to generate HF. Therefore, when multiple cells are present in a neighborhood of a pixel, it will have a high value in HF. Three masks and HF are shown in Fig 2 (f)-(i). Figure 2 (j) depicts HF if the masks were not subtracted from HTh2. As shown, the subtraction is essential to restrict the analysis to the tumor areas. Finally, HF was normalized and ten pixels whose intensity is at least 0.75 were randomly selected from HF. The distance of the selected points should be more than 100 pixels.

Next, the textural features listed in Table I were extracted from each patch. The textural features were extracted from H- channel, blue-ratio channel [26], and each one of three RGB channels. The images were also converted to Lab, YUV, HSL, and LMS color spaces and the features were extracted from each channel of these color spaces.

3) Regression models

As shown in Fig. 1, in the intermediate steps of COMPASS there are two regression models, namely, Regression model 1 (RM1) and Regression model 2 (RM2). The input of RM1 were the scores given by the pathologists while RM2 relied on the textural features. Both RM1 and RM2 were ensembles of trees for regression which comprised of a weighted combination of multiple regression trees.

Junior pathologists scored six atypia-related criteria on four images at ×40 magnification for each image in the dataset. This resulted in a 24-dimensional feature vector for each image. If bizarre nuclei were present in one of the four images at ×40 magnification, the grade of the image is 3, and this does not depend on the arrangement of the four images. Therefore, for an input image, all 24 possible combinations of the shuffling of these four ×40 images were generated. Then RM1 assigns a score to each of these 24 possible permutations, and the final score of the image is the median of these values. For each test image, ten patches were selected as suggested in 2-2-
2. Each one of these patches was inputted to RM2.

For training RM1 all 24 possible combinations of four ×40 images were generated for each ×20 image in the training set. This increases the size of the training set by 24 times and makes RM1 invariant to the spatial layout of structures within the image. For training RM2, each patch was considered as an instance, and the grade of the image (from which the patch was selected) was considered as the grade of the patch. Hence, the size of the training set for RM2 was ten times larger than the number of the images.

One of the main challenges in using ensemble models is setting the hyper-parameters of the model because they could affect the performance of the model. We used Bayesian optimization for hyperparameter tuning [27]. Here the optimization searched over the ensemble method, namely, either Bag (bootstrap aggregation) or LSboost (least squares boosting), over the number of weak learners, over the learning rate for shrinkage of the LSBoost method, over the minimum number of leaf node observations in the template tree, and over the number of features to select at random for each split in the tree.

4) Late decision *fusion*

As shown in Fig. 1, the median of 24 values given by RM1 to 24 possible permutations of four ×40 images, along with minimum, median, and maximum scores given by RM2 to ten patches of each image built the feature vector for RM3. RM3 was an ensemble of trees for regression as well. In order to find the cut-off values to threshold the scores from regression models and produce three-scale atypia grades, two receiver operating characteristic (ROC) curves were generated, one for detecting high grade images (grade 3 against combined grade 1 and 2) and one for low grade images (grade 1 against combined grade 2 and 3) and their optimal operating points were found.

For training RM3, the instances from grade 1 and 3 were upsampled by applying the Synthetic Minority Oversampling TEchnique (SMOTE) [28]. The numbers of nearest neighbors to use were set to 3 and 5 for grade 1 and 3 respectively and the percentage of SMOTE instances to create was set to 200% and 400%. The hyper-parameters of RM3 were also set by using Bayesian optimization [27].

C. Evaluation of COMPASS

As COMPASS is a personalized tool, first the parameters of the model should be estimated for each pathologist by asking the readers to assign scores to six nuclear atypia criteria on the images for which the expert-consensus derived reference nuclear pleomorphism scores are available. After this training stage, COMPASS can be used to score new images. Therefore for evaluating COMPASS' performance, we need to first train the model and then test the trained model on unseen data. To do so, leave-one-image-out cross validation (LOOCV) was used. Hence, each time one of the images served as the test data, the rest of the images (training data) were utilized for estimating the parameters of COMPASS. The percentage agreement and Cohen's kappa [29] were calculated for each junior pathologist.

In each iteration of LOOCV, the training data was

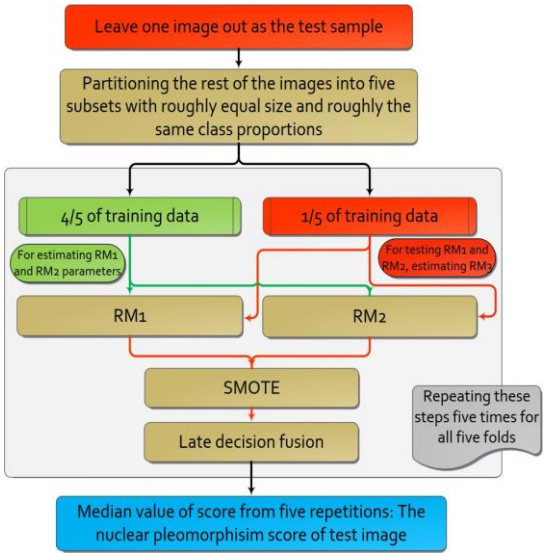


Fig. 3. The evaluation procedure of COMPASS.

TABLE II

CONFUSION MATRICES; COLUMNS ARE TRUE LABELS WHILE ROWS ARE LABELS FROM COMPASS. COMPASS' PERFORMANCE FOR IMAGES SCANNED BY (A)-(C) APERIO SCANNER (D)-(E) HAMAMATSU SCANNER. JP STANDS FOR JUNIOR PATHOLOGIST.

(a) JP1

	G1	G2	G3
G1	18	1	0
G2	5	215	5
G3	0	6	47

(b) JP2

G1	G2	G3
20	3	0
3	213	6
0	6	46

(c) JP3

G1	G2	G3
16	3	0
7	214	8
0	5	44

(d) JP1

	G1	G2	G3
G1	16	0	0
G2	7	212	5
G3	0	10	47

(e) JP2

G1	G2	G3
19	3	0
4	207	5
0	12	47

(f) JP3

G1	G2	G3
16	1	0
7	217	5
0	4	47

partitioned into five subsets with roughly identical size and roughly the same class proportions as in the original dataset. Four subsets were utilized to estimate the parameters of RM1 and RM2. Next, the images in the remaining subset were inputted to RM1 and RM2. As stated earlier, four features were extracted from the scores given by RM1 and RM2 to each instance in this subset and used to train RM3. Finally, a score was given to the test data by the trained model. Fig. 3 shows the procedure for training COMPASS. This procedure was repeated five times; each time one of the subsets was used to estimate the parameters of RM3, the rest of them were used to estimate the parameters of RM1 and RM2. Therefore, five scores were given to each test data. The median value of all these scores was assigned to each image. In order to set the hyper-parameters of regression models, in each of the five repetitions, ten-fold-cross-validation was used for Bayesian optimization. To achieve a result robust to partitioning noise, at every iteration, the cross-validation was repartitioned.

TABLE III

COHEN'S KAPPA AND PERCENTAGE AGREEMENT OF COMPASS AND SENIOR PATHOLOGISTS. THE HIGHEST ACCURACY IS SHOWN IN BOLD.

Senior pathologists			***COMPASS***		
	1	**2**		**JP1**	**JP2**	**JP3**
I	0.79	0.68	**SA**	0.86	0.85	0.80
E	0.85	0.73	**SH**	0.81	0.81	0.85
G1	78.3%	82.6%		73.9%	**84.8%**	69.6%
G2	90.1%	91.4%		96.2%	94.6%	**97.1%**
G3	**98.1%**	69.2%		90.4%	89.4%	87.5%
T	90.6%	86.9%		**93.4%**	92.9%	93.3%

JP stands for junior pathologist; I (E): The cases to which the readers could not assign a grade were included (excluded); SA: Aperio. SH: Hamamatsu Nanozoomer 2.0-HT Scanner. G and T stand for grade and total.

TABLE IV

AUC VALUES FOR DETECTION OF GRADES 3 AND 1.

	Detection of grade 3				Detection of grade 1			
JP	**Sum**	**RM1**	**SA**	**SH**	**Sum**	**RM1**	**SA**	**SH**
1	0.631	0.784*	0.977*	0.941*	0.591	0.638	0.934*	0.888*
2	0.708	0.786	0.975*	0.959*	0.756	0.746	0.948*	0.907*
3	0.635	0.744*	0.963*	0.926*	0.754	0.779	0.935*	0.850

JP represents the junior pathologist for whom COMPASS was trained and tested. SA and SH represent the performance of COMPASS for Aperio and Hamamatsu scanners.* Shows that AUC value for cumulative score (Sum column) is significantly lower than the compared AUC. The P-values were calculated based on [30].

III. RESULTS

A. Performance of COMPASS

As described in section II-C leave-one-image-out cross validation was used to evaluate the performance of COMPASS for each scanner. COMPASS is personalized hence it should be trained and tested for each reader separately. Table II shows the confusion matrices of COMPASS for each scanner and each junior pathologist. In the table, the upper triangular part of the matrix represents "under-graded instances" and the lower part represents "over- graded instances" based on COMPASS.

The paired Mann-Whitney U test was used to compare the grades given by COMPASS when tested on Aperio images with the given grades for Hamamatsu images. The differences among grades from different pathologists were not significantly different (junior pathologist 1: z=-1.1, P=0.29; junior pathologist 2: z=0.48, P=0.63; junior pathologist 3: z=0.86, P =0.39). Also, Spearman's rank-order correlation coefficients between the scores (before thresholding it to produce three scale grades) given to the images from two scanners were 0.74, 0.77, and 0.75 for three junior pathologists.

B. Comparison of COMPASS with senior pathologists

We simulated the adoption of COMPASS by the junior pathologists and compared the performance of COMPASS to that of senior pathologists. The average CCR per each grade is shown in Table III for all junior pathologists. The values are an average of the two scanners. Similarly, on the right side of the table, CCRs are shown for the senior pathologists. As shown, the overall performance of COMPASS was comparable to that of the senior pathologists.

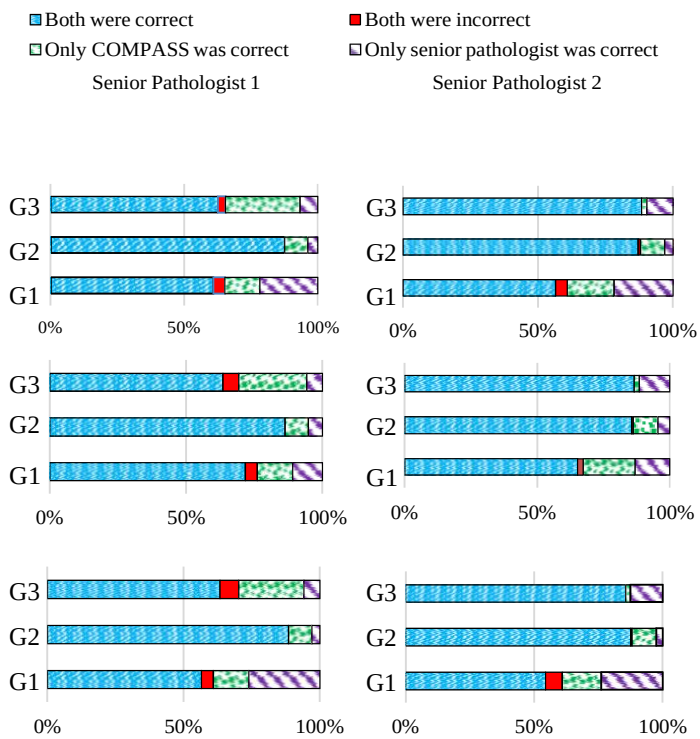


Fig. 4. The percentage of concordant and discordant cases for each atypia category based on scores given by COMPASS and the senior pathologists. The values are the average of two scanners. G1, G2, and G3 indicate grade 1, 2, and 3 respectively. Each row represents one of the junior pathologists.

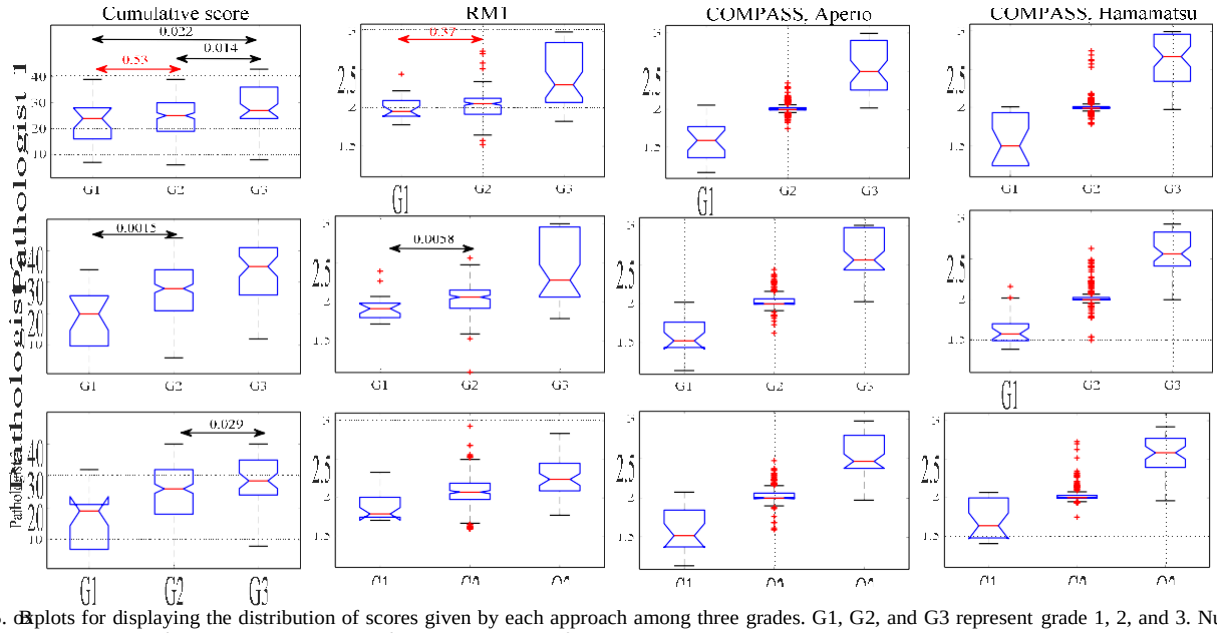


Fig. 5. Boxplots for displaying the distribution of scores given by each approach among three grades. G1, G2, and G3 represent grade 1, 2, and 3. Numbers above double arrows in the figure show the P-values of Tukey-Kramer test for each pair. When the P-values were not shown between a pair, it means P- value< 0.001. The red numbers show insignificant differences between a pair.

Cohen's kappa was also calculated to measure the magnitude of agreement of COMPASS with the ground truth. The value is calculated for each pathologist and each scanner separately and shown in Table III. Similarly, Cohen's kappa was calculated for the senior pathologists. In the databases, there were nine images to which the senior pathologists did not assign once with and once without these images. As shown the agreement level is perfect except when COMPASS was adopted for the third pathologist and images from Aperio scanner were used. The Cohen's kappa value is substantial for this arrangement. For the senior readers, the agreement level was substantial to perfect.

The joint distribution of grades from COMPASS and each of the senior pathologists was investigated to find the percentage of the images that both graded correctly, only one of them graded correctly, or both graded incorrectly. The result is shown in Fig 4. Results for both scanners were combined to generate the plots. Each row in the plot shows COMPASS as adopted by one of the junior pathologists, and each column represent one of the senior pathologists. Among misclassified images, the percentage of images which were graded incorrectly by both COMPASS and senior pathologists (orange areas in the plots) were lower than those which were graded correctly by one of them. Hence, to some extent, COMPASS could complement the senior pathologist's performance.

C. *Added benefits of textural features*

The added benefit of textural features was investigated by comparing the performance of COMPASS against that of two baseline approaches that only used the scores given by the pathologists to the cytological characteristics of images. The first one was grading based on the total cumulative score given to all criteria. Most of the nuclear grading systems produce the final nuclear grade of each sample by summing up all scores given to the considered criteria [19, 20]. Hence, we also used this approach for comparison. The second approach was based on the score assigned by RM1. One possible benefit of COMPASS is to use a complex non-linear regression model to associate the scores given by the pathologists to the nuclear grade. This part is done by RM1 in COMPASS. Therefore we compared the performance of COMPASS with that of RM1 to investigate the importance of the added textural features. The boxplots, which display the distribution of scores among different grades, are shown in Fig. 5. The plots were generated from two baseline approaches as well as COMPASS for both scanners. The Kruskal-Wallis test resulted to P-values<0.0001 for all approaches and all pathologists, except for the first approach (sum) when adopted for the first pathologist, which led to a P-value of 0.007 ($\chi2(2,297)=5.03$). The Rank-based version of Tukey's HSD (Tukey-Kramer) test showed that differences between all possible multiple pairs for all approaches were significant (P<0.05) except for grade 1 against grade 2 for the first and second approach when adopted by the first reader. The results of the rest of the comparisons are indicated in the figure.

Also, the AUC for detecting high grade images (i.e. grade 3) and low grade images (i.e. grade 1) is reported in Table IV. The AUC for detecting high grade images (i.e. grade 3) shows the binary classification for categorizing images as high grade (grade 3) and low/intermediate grade (grade 1 and 2). The AUC for low grade images (i.e. grade 1) shows the performance of binary classification for categorizing images as high/intermediate grade (i.e. grade 2 and 3) and low grade (i.e. grade 1). As shown, the cumulative score led to the poorest results for detecting grade 3 for all pathologists, however, the differences between AUC of the cumulative score and RM1 were not significant for the second pathologist. For detecting low grade (i.e. grade 1) RM1 outperformed the cumulative score for pathologist 1 and 3 while the two approaches resulted in an almost similar AUC values for pathologist 2.

IV. Discussion

In this paper, COMPASS, a personalized algorithm for reproducible nuclear pleomorphism grading, was introduced. The Leave-one-out cross validation was used and a percentage agreement of 93.4%, 92.9%, and 93.3% was achieved between COMPASS and the reference nuclear grade for three pathologists. Therefore, the percentage agreement was almost

identical for three junior pathologists. The results also suggested that the performance of COMPASS was approximately similar for both scanners. However, the CCRs of COMPASS varied among different nuclear grades. As shown in tables II and III, CCRs were the highest in grade two and the lowest in grade one for all three junior pathologists. Similarly, on average, senior pathologists achieved the highest CCR in grade two.

Most of the previous algorithms aiming at automatic nuclear grading segmented the cells within an image and then extracted features from the segmented areas. COMPASS also detects the nuclei, however, it does not improve the coarse segmentation and extracts the textural features from neighborhoods with a high density of nuclei. Therefore, the impact of segmentation errors on the features extracted by COMPASS has been compensated to some extent. A recent fully-automatic algorithm based on textural features was proposed in [15]. It achieved CCRs of 65.22%, 90.09%, and 69.23% for the three nuclear grades and a Cohen's kappa of 0.6123 (substantial agreement) on the same images that we used here. By taking advantage of scores from pathologists and restricting the analysis to the areas with high nuclear density, COMPASS obtained average (across three junior pathologists) CCRs of 76.1%, 96.0%, and 89.1% for the three nuclear grades and a Cohen's kappa of 0.83 (perfect agreement). As shown in Table 4, the scores given by RM1 (which relies on features from pathologist's assessment) achieved an average AUC of 77% for detecting high grade cases. This shows the higher discriminative ability of scores given by the pathologists in detecting high grade images compared to low grade images (average AUC of 72%).

Cohen's kappa ranged from 0.80-0.86 for different pathologists and different scanners. The values were comparable to the Cohen's kappa for senior pathologists while assessing the same dataset. Fig. 5 shows the percentage of images on which COMPASS agreed with each one of the senior pathologists. It should be noted that even two senior pathologists did not agree with each other on all images. Their agreement rate was 74%, 84%, and 69% for the three nuclear grades. Estimating the internal parameters of COMPASS was computationally expensive and this procedure should be done for each pathologist. Specifically, this was due to the fact that we set the hyper-parameters of COMPASS by using Bayesian optimization and repeated the late decision fusion step five times to avoid partitioning noise. However, the training step should be done only once and after that COMPASS can be used to assess new images.

The study has a number of limitations; first, the prevalence of different grades in the dataset was different from real clinical practice. Therefore, the agreement rate reported here and Cohen's kappa will change if the class proportions change. However, having a balanced data set may improve the CCRs of the grade 1 and 3, as more data will be available for training the model. Here, we dealt with the class imbalance problem by adopting the SMOTE [28] algorithm for upsampling of minority classes (i.e. grade 1 and 3); nonetheless COMPASS could benefit from larger sample size. Secondly, intra-pathologist variability in scoring six atypia- related criteria should be investigated. Thirdly, the reported results were based on 300 images from eleven patients. Here we used leave-one-image-out cross validation to evaluate the performance of COMPASS. However, the results would be more realistic if the test images were from different patients. In the publicly available challenge dataset, 124 test images from different patients were provided, however the junior pathologists did not asses those images. Hence, COMPASS could not be used for grading them.

Fourthly, as the ultimate goal of breast cancer grading is utilizing it as a prognostic factor in patient management, investigating the association between the nuclear grade outputted by COMPASS and patient survival would strengthen the study. Relating COMPASS' output to patients' prognosis could be a future step of this study. Also, the internal parameters of COMPASS are estimated based on the current performance of the junior pathologist. However, the scores given by the junior pathologists could change as they gain experience. Therefore, the parameters of the model should be updated on a regular basis. Investigating paradigm for updating the parameters and algorithmic considerations (e.g. whether the hyper-parameters should be updated or not) could be a possible avenue for future work. Finally, COMPASS was tested retrospectively and we assumed that the junior pathologists would accept the nuclear grade given by COMPASS. However, in a more realistic set-up, the junior pathologists would score six atypia-related criteria and then COMPASS would combine these scores with the computer-extracted textural features using previously trained non-linear regression models and output the nuclear grade to the junior pathologists, who would assign the final nuclear grade to the image.

In summary, COMPASS, which is a personalized tool, can assist junior pathologists in nuclear grading of breast cancer and achieved a performance that was comparable to that of the senior pathologists. This study has also demonstrated that COMPASS, if it had been adopted by the junior pathologists, could play the role of the second reader, and it could also complement the senior pathologist's performance to some extent. The findings also underscore the importance of textural computer-extracted features to supplement the junior pathologist's assessment of the case.

ACKNOWLEDGMENT

We would like to thank Mitosis-Atypia challenge organizers who collected the data utilized in the study and kindly provided us the access to their dataset after the challenge time. We also acknowledge the University of Sydney HPC Service at the University of Sydney for providing high performance computing resources that have contributed to the research results reported within this paper.

REFERENCES

[1] S. Mook, M. K. Schmidt, E. J. Rutgers, A. O. van de Velde, O. Visser, S. M. Rutgers, *et al.*, "Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online Adjuvant! program: a hospital-based retrospective cohort study," *The lancet oncology*, vol. 10, pp. 1070-1076, 2009.

[2] E. A. Rakha, M. E. El-Sayed, D. G. Powe, A. R. Green, H. Habashy, M. J. Grainge, *et al.*, "Invasive lobular carcinoma of the breast: response to hormonal therapy and outcomes," *European Journal of Cancer*, vol. 44, pp. 73-83, 2008.

[3] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up," *Histopathology,* vol. 19, pp. 403-410, 1991.

[4] [4] N. E. Roberti, "The role of histologic grading in the prognosis of patients with carcinoma of the breast," *Cancer,* vol. 80, pp. 1708-1716, 1997.

[5] J. Bueno-de-Mesquita, D. Nuyten, J. Wesseling, H. van Tinteren, S. Linn, and M. van De Vijver, "The impact of inter-observer variation in pathological assessment of node-negative breast cancer on clinical risk assessment and patient selection for adjuvant systemic treatment," *Annals of oncology,* vol. 21, pp. 40-47, 2010.

[6] H. F. Frierson, R. A. Wolber, K. W. Berean, D. W. Franquemont, M. J. Gaffey, J. C. Boyd, *et al.*, "Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma," *American journal of clinical pathology,* vol. 103, pp. 195-198, 1995.

[7] J. M. Harvey, N. H. de Klerk, and G. F. Sterrett, "Histological grading in breast cancer: interobserver agreement, and relation to other prognostic factors including ploidy," *Pathology,* vol. 24, pp. 63-68, 1992.

[8] T. A. Longacre, M. Ennis, L. A. Quenneville, A. L. Bane, I. J. Bleiweiss, B. A. Carter, *et al.*, "Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study," *Modern pathology,* vol. 19, pp. 195-207, 2006.

[9] J. S. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, *et al.*, "Breast carcinoma malignancy grading by Bloom–Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index," *Modern pathology,* vol. 18, pp. 1067-1078, 2005.

[10] A. Paradiso, I. Ellis, F. Zito, E. Marubini, S. Pizzamiglio, and P. Verderio, "Short-and long-term effects of a training session on pathologists' performance: the INQAT experience for histological grading in breast cancer," *Journal of clinical pathology,* vol. 62, pp. 279-281, 2009.

[11] R. Zhang, H.-j. Chen, B. Wei, H.-y. Zhang, Z.-g. Pang, H. Zhu, *et al.*, "Reproducibility of the Nottingham modification of the Scarff-Bloom-Richardson histological grading system and the complementary value of Ki-67 to this system," 2010.

[12] A. L. Adams, D. C. Chhieng, W. C. Bell, T. Winokur, and O. Hameed, "Histologic grading of invasive lobular carcinoma: does use of a 2-tiered nuclear grading system improve interobserver variability?," *Annals of diagnostic pathology,* vol. 13, pp. 223-225, 2009.

[13] E. Cosatto, M. Miller, H. P. Graf, and J. S. Meyer, "Grading nuclear pleomorphism on histological micrographs," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008, pp. 1-4.

[14] J.-R. Dalle, H. Li, C.-H. Huang, W. K. Leow, D. Racoceanu, and T. C. Putti, "Nuclear pleomorphism scoring by selective cell nuclei detection."

[15] A. M. Khan, K. Sirinukunwattana, and N. Rajpoot, "A Global Covariance Descriptor for Nuclear Atypia Scoring in Breast Histopathology Images," *IEEE Journal of Biomedical and Health Informatics,* vol. 19, pp. 1637-1647, 2015.

[16] C. Lu, M. Ji, Z. Ma, and M. Mandal, "Automated image analysis of nuclear atypia in high-power field histopathological image," *Journal of microscopy,* vol. 258, pp. 233-240, 2015.

[17] B. Dunne and J. Going, "Scoring nuclear pleomorphism in breast cancer," *Histopathology,* vol. 39, pp. 259-265, 2001.

[18] L. Roux, D. Racoceanu, F. Capron, J. Calvo, E. Attieh, G. Le Naour, *et al.*, "MITOS & ATYPIA," 2014.

[19] C. Bansal, M. Pujani, K. L. Sharma, A. Srivastava, and U. Singh, "Grading systems in the cytological diagnosis of breast cancer: A review," *Journal of Cancer Research & Therapeutics,* vol. 10, 2014.

[20] K. Saha, G. Raychaudhuri, B. K. Chattopadhyay, and I. Das, "Comparative evaluation of six cytological grading systems in breast carcinoma," *Journal of Cytology,* vol. 30, p. 87, 2013.

[21] A. Abati and G. McKee, "Grading of breast carcinoma in fine-needle aspiration cytology," *Diagnostic cytopathology,* vol. 19, pp. 153-154, 1998.

[22] I. Robinson, G. McKee, and M. Kissin, "Typing and grading breast carcinoma on fine-needle aspiration: Is this clinically useful information?," *Diagnostic cytopathology,* vol. 13, pp. 260-265, 1995.

[23] M. Macenko, M. Niethammer, J. Marron, D. Borland, J. T. Woosley, X. Guan, *et al.*, "A method for normalizing histology slides for quantitative analysis," in *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, 2009, pp. 1107-1110.

[24] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Transactions on Biomedical Engineering,* vol. 57, pp. 841-852, 2010.

[25] H. Irshad, "Automated mitosis detection in histopathology using morphological and multi-channel statistics features," *Journal of pathology informatics,* vol. 4, p. 10, 2013.

[26] H. Chang, L. A. Loss, and B. Parvin, "Nuclear segmentation in H&E sections via multi-reference graph cut (MRGC)," in *International symposium biomedical imaging*, 2012.

[27] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of bayesian optimization," *Proceedings of the IEEE,* vol. 104, pp. 148-175, 2016.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[29] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: the kappa statistic," *Fam Med,* vol. 37, pp. 360-363, 2005.

[30] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology,* vol. 143, pp. 29-36, 1982.

Chapter 7

Bridging Chapter for

"MuDeRN: Multi-category Classification of Breast Histopathological Image Using Deep Residual Networks"

Study

The study is published in the Artificial Intelligence in Medicine Journal, 2018.

7-1- Introduction

A recent study in the USA showed that an estimated 19% of women screened annually for a ten year period will undergo a breast biopsy [1]. Therefore, each year, large numbers of breast histopathological slides are interpreted in pathology labs. Benign breast lesions are far more prevalent, and only one in four cases depicts malignancy [2]. Interpretation of all these slides is very time-consuming and subject to inter-pathologist variations. As the pathologists' diagnoses on the cases are considered the gold standard for further treatment of the patients, incorrect diagnoses can lead to inappropriate patient management [3-5]. Misdiagnosing invasive or pre-invasive breast cancer as being a benign lesion could result in worse outcomes as the cancer may advance [6]. On the other hand, overinterpretation of benign cases is also harmful as the patients may undergo unnecessary treatments [7].

Correct recognition of the subtype of benign breast lesions and breast cancers is very important as further patient management is dependent on the pathological diagnosis [3-7]. For example, the risk of developing invasive cancer in the future varies among different benign subtypes [11] and misclassifying benign lesions may result in overestimation or underestimation of the subsequent breast cancer risk, which could lead to inappropriate patient management [7]. On the other hand, determining the malignancy subtype can be helpful in predicting the patient's response to therapy, for example, appropriate treatment options for invasive lobular cancer are different from those for invasive ductal cancer [3-6, 10].

The magnitude of disagreement among pathologists for generating a diagnosis in breast pathology has been previously studied [3-5, 8, 9]. One of the earliest studies was carried out in 1992 [8], where six pathologists were asked to classify 24 cases into three categories, namely, usual hyperplasia, atypical hyperplasia, or carcinoma in situ. Complete agreement among all six pathologists was seen only in 58% of cases. Later, Wells et al. (1998) included 30 cases with benign, benign with atypia, non-invasive malignant, and invasive malignant and asked 26 pathologists to diagnose the cases [9]. The overall kappa for agreement of pathologists with expert-consensus was 0.71.

Elmore et al. (2015) [4] investigated the degree of disagreement of 115 pathologists with expert consensus-derived reference in diagnosing benign without atypia, benign

with atypia, ductal carcinoma in situ, and invasive cancer. Their results suggested that 3% of benign cases without atypia and 17% of benign cases with atypia were overinterpreted as ductal carcinoma in situ or invasive carcinoma and 10% of ductal carcinoma in situ or invasive carcinoma cases were underinterpreted as benign cases with or without atypia.

Lawton et al. (2014) investigated the magnitude of the discrepancies among ten pathologists for distinguishing fibroadenomas from phyllodes tumours, which are two subtypes of benign lesions without atypia. Their results suggested that all pathologists agreed about the final diagnoses in only 53% of cases [10]. In [11], the agreement among pathologists in classifying benign lesions into three categories, namely fibroadenoma, phyllodes tumour, and other benign subtype was explored. It was shown that the overall Cohen's kappa for inter-pathologist agreement only 0.48, which suggested moderate agreement [11]. Longacre et al. [12] showed that inter-pathologist disagreement exist for identifying the subtypes of invasive breast carcinoma as well [12]. They showed that the agreement of diagnoses with expert consensus-derived reference diagnoses were 75.0%, 62.5%, 95.8%, 56.3%, 41.7%, 90.0%, and 92.0% for tubular, papillary, mucinous, medullary, metaplastic, lobular, and ductal carcinoma (other than medullary and tubular) respectively [12].

Beside the inter-pathologist agreement, Jackson et al. (2017) showed that diagnostic disagreement is observed even when the same pathologist interprets the same case at two time points [13]. The reported intra-pathologist disagreement rates were 8% for invasive breast cancer, 16% for ductal carcinoma in situ, 47% for benign with atypia, and 16% for benign without atypia.

Studying the underlying reasons for disagreement could be helpful in improving agreement. In [3], the reasons for disagreement were divided into three groups, namely pathologist-related, diagnostic coding/study methodology-related, and specimen-related. "Professional differences of opinion on features meeting diagnostic criteria" was ranked first among pathologist-related reasons [3]. Computer-aided analysis which provides an objective categorization can help pathologists reduce the discrepancies in diagnostic disagreement.

As discussed in chapter 2 (literature review paper), textural, intensity-based, and morphological features were previously used for binary [14-16] or multi-class classification [16] of breast histopathological images. Recently deep learning has been also used for categorizing breast histopathological slides [17-19]. However, these studies [17-19] mainly focused on benign/malignant classification of images. Moreover, usually majority voting is used for making patient-level diagnosis based on the image-level diagnosis and no non-linear trainable method was proposed. Fianlly, none of the previous studies provided an algorithm for combining the diagnoses for different magnification factors.

In the paper presented in chapter 8, MuDeRN (MUlti-category classification of breast histopathological images using DEep Residual Networks), a novel a framework for eight-class categorization of hematoxylin-eosin stained breast digital slides is explained. MuDeRN classifies the cases either as benign or cancer, and then categorizes cancer and benign cases into four different classes each. More specifically, MuDeRN aims at:

- Classification of breast histopathological images as benign or malignant
- Categorization of malignant images as ductal carcinoma, lobular carcinoma, mucinous carcinoma, or papillary carcinoma
- Classification of benign images as adenosis, fibroadenoma, phyllodes tumour, or tubular adenoma

Some of the subtypes differentiated by MuDeRN share several similar features but require different patient management strategies. Among benign subtypes, fibroadenoma and benign phyllodes are both benign fibro-epithelial tumours. They share many similar features, and the benign phyllodes tumour looks similar to a giant fibroadenoma [20]. However, treatment is different for these two subtypes. The benign phyllodes tumours should be surgically removed as their growth is very fast and if they remained untreated, eventually they create a visible lump and may cause pain; however fibroadenoma does not necessarily have to be surgically removed. Tubular adenoma and adenosis are both benign epithelial proliferations without atypia. Tubular adenoma is completely benign and does not result in an increased risk of subsequent breast cancer [21], while adenosis is associated with a higher risk of the development of invasive cancer compared to the general population [22-24]. Among malignant

subtypes, in both mucinous carcinoma and invasive lobular carcinoma, mucin production might increase, but prognosis from mucinous carcinoma is better than that of invasive lobular carcinoma [25, 26], and hence more aggressive treatment options are only advisable for invasive lobular carcinoma [27].

The performance of MuDeRN was evaluated using the BreakHis database [28]. It is a publicly available dataset of hematoxylin-eosin stained breast histopathological slides, where the images were acquired in four visual magnification factors. BreakHis was previously used for malignant/benign classification [17-19, 28] and eight-class categorization [19]. The major contributions of this work compared to [17-19, 28] are:

- Enhancing the correct classification rate in both image-level (i.e. making a diagnosis for each image independently without considering the patient information) and patient-level (i.e. assigning a label to each patient by combining the class labels appointed to all images of that patient)
- proposing a framework for combining classification results of a patient's images from different magnification factors to make the ultimate patient-level diagnosis

7-2- Materials and Methods

In chapter 8, I provide the detailed description about the steps of MuDeRN, but some further explanatory points about the dataset are discussed here.

7-2-1- Dataset

The BreakHis database [28] is a publicly available dataset of hematoxylin-eosin stained breast histopathological slides in four visual magnification factors, namely x40, x100, x200, and x400. The effective pixel size and objective lens for each magnification factor are shown in Table 1. The database contains images from four BCa subtypes and four benign subtypes. It contained 7786 images, which were acquired from 81 patients, but the number of images per patient differed from patient to patient, however on average 96 images were provided per patient. The areas covered in the images were selected by an experienced pathologist in each magnification level in a way that the image contains diagnostically relevant features of the disease. So, depending on the opinion of the pathologist, different number of images was available

for different magnification level. Detailed number of images per case per magnification factor is provided in [28], where the dataset has been introduced.

In this section, microscopic characteristics of each subtype included in the BreakHis, are briefly explained. It should be noted that in real clinical practice scanning a test sample four times using four different microscopic lenses will not be required for using MuDeRN. Slides can be scanned at the highest magnification level and down sampled to produce the lower magnification factors.

Table 3- effective pixel size and objective lens for each magnification factor

Visual magnification	Objective lens	Effective pixel size (µm)
x40	x4	0.49
x100	x10	0.20
x200	x20	0.10
x400	x40	0.05

7-2-1-1- Benign subtypes

7-2-1-1-1- Adenosis

Adenosis is a benign breast lesion without atypia which involves epithelial proliferations of small acini and terminal ducts. It accounts for 12 - 28% of all benign lesions [24-27]. Usually in adenosis the number of glands are greater than usual and lobules are enlarged and collagen is often present. Diagnosing adenosis is important as it is associated with 1.5 to 2-fold increased risk of developing subsequent invasive breast cancer [22]. Figure 1 depicts an adenosis from the BreakHis database.

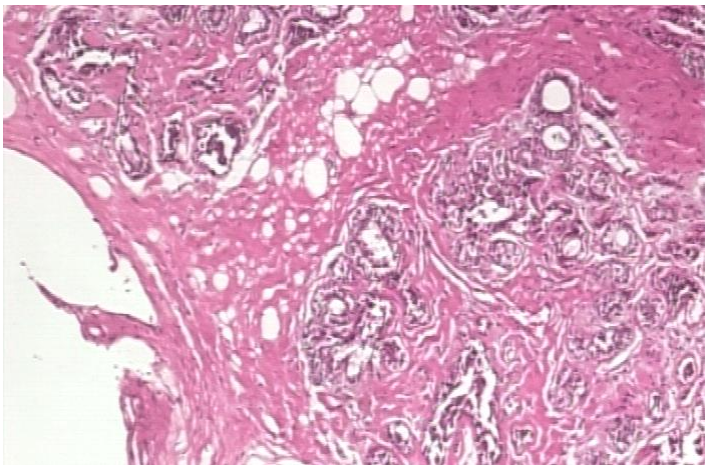


Figure 1- A sample image in adenosis class (patient ID=22549G) at x40 magnification factor (image ID: SOB_B_A-14-22549G-40-026).

7-2-1-1-2- Fibroadenoma

Fibroadenoma is a common benign lesion (accounting for 18.5% of all benign biopsies [29]) which mostly affect women in their 20s and 30s. Tumours are mostly firm and well circumscribed. Fibroadenoma is comprised of epithelial and stromal components. Figure 2 shows a fibroadenoma from the BreakHis database.

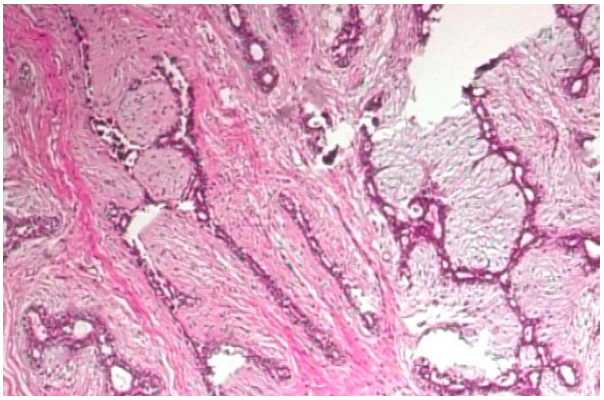


Figure 2- A sample image in fibroadenoma class (patient ID=14134E) at x40 magnification factor (image ID: SOB_B_F-14-14134E-40-002).

7-2-1-1-3- Phyllodes tumour

Phyllodes tumours are usually observed in middle age women and make up 0.20% of all benign breast lesions [30]. They are mostly firm, round and well circumscribed. Their main feature under the microscope is stromal hypercellularity and overgrowth, normally no nuclear atypia is observed. Similar to fibroadenoma, they arise from interlobular stroma, however they are much larger and more cellular than

fibroadenoma. Phyllodes in Greek means leaf-like. The phyllodes tumour gets its name as in these tumours, projection of stroma into ducts create leaf-like pattern. In the BreakHis database only benign phyllodes tumours were included, however, in general, phyllodes tumours could be benign, borderline or malignant. In benign phyllodes tumours, less than 2 mitotic figures are observed in ten high power fields, while in borderline tumours and malignant ones 2-5 and more than 5 mitosis are observed respectively[31, 32]. Figure 3 presents a phyllodes tumour from the BreakHis database.

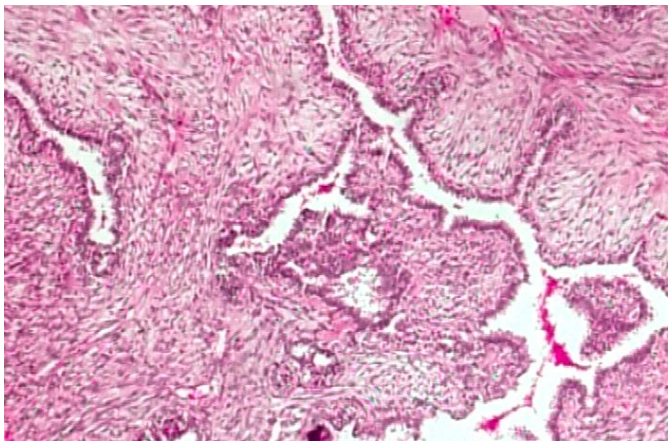


Figure 3- A sample image in phyllodes tumour class (patient ID=21998AB) at x40 magnification factor (image ID: SOB_B_PT-14-21998AB-40-004). The leaf-like architecture can be seen in the figure.

7-2-1-1-4- Tubular adenoma

Tubular adenoma accounts for up to 10% of benign lesions and is usually observed in younger women. It is characterized by being well circumscribed, as well as absence of atypical nuclei and presence of densely packed tubules. No or scant stroma is often observed [31, 32]. Figure 4 indicates a tubular adenoma from the BreakHis database.

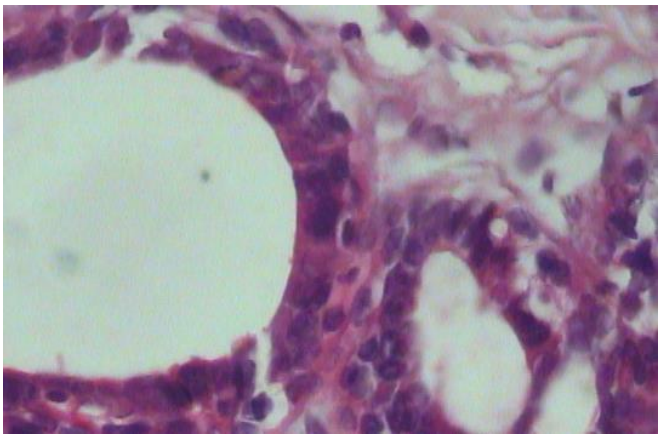


Figure 4- A sample image in tubular adenoma class (patient ID=3411F) at x400 magnification factor (image ID: SOB_B_TA-14-3411F-400-012). As shown no atypical cells are present and the tubule is densely packed.

7-2-1-2- Malignant subtypes

7-2-1-2-1- Invasive ductal carcinoma

Invasive ductal carcinoma is the most common type of BCa, accounting for 75% to 80% of all BCa. Most of the time when the lesion lacks diagnostic criteria of any other BCa subtypes, it is diagnosed as invasive ductal carcinoma (i.e. a diagnosis of exclusion) [24-27]. Tumours in this category are typically firm with ill-defined borders and noted chalky streaks are observed on cut sections, however some tumours may show a clear border. The tumour cells could grow in sheets, nests, cords or individual cells. Degree of tubule formation and nuclear atypia differ among patients and mitotic figures are obvious [31]. Figure 5 depicts an invasive ductal carcinoma from the BreakHis database.

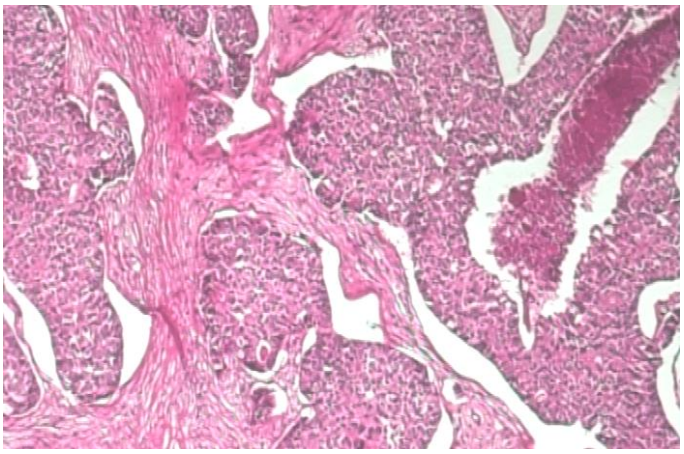


Figure 5- A sample image in invasive ductal carcinoma class (patient ID=2523) at x40 magnification factor (image ID: SOB_M_DC-14-2523-40-010).

7-2-1-2-2- Invasive lobular carcinoma

Invasive lobular carcinoma is the second most common type of BCa, accounting for 10%-15% of BCa [33]. Mostly the tumour nuclear grade is 1 or 2 and tumour cells differ slightly from normal cells in size, shape, and appearance; the chromatin is evenly dispersed and typically no nucleoli is observed. Cells are often encircling normal ducts [31].

Compared to ductal carcinoma, cells of lobular carcinoma are smaller, more uniform and discohesive which grow in Indian file or singly. Discohesion of tumour cells is a consequence of lack of E-cadherin, which is an epithelial calcium-dependent protein for cell adhesion. Most of the time, tubularity grade is 3, as cells do not form tubules [31]. Figure 6 indicates an invasive lobular carcinoma from the BreakHis database.

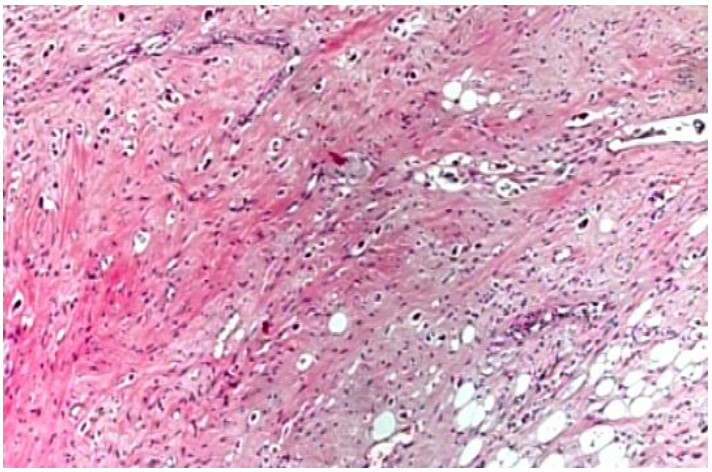


Figure 6- A sample image in invasive lobular carcinoma class (patient ID=15570C) at x40 magnification factor (image ID: SOB_M_LC-14-15570C-40-021). As shown cells are small, uniform and discohesive.

7-2-1-2-3- Mucinous carcinoma

Mucinous carcinoma is a subtype of BCa, accounting for 0.5% to 3% of all BCa cases [34]. It is characterized by low grade tumour cells floating in a lightly stained amorphous mucin and cells may be solid, acinar, or detached. Mitotic figures and in situ epithelial component are usually not present and tumours are often enclosed by connective tissue bands. The prognosis of mucinous carcinoma is usually better than invasive ductal carcinoma [35]. Figure 7 shows a mucinous carcinoma from the BreakHis database.

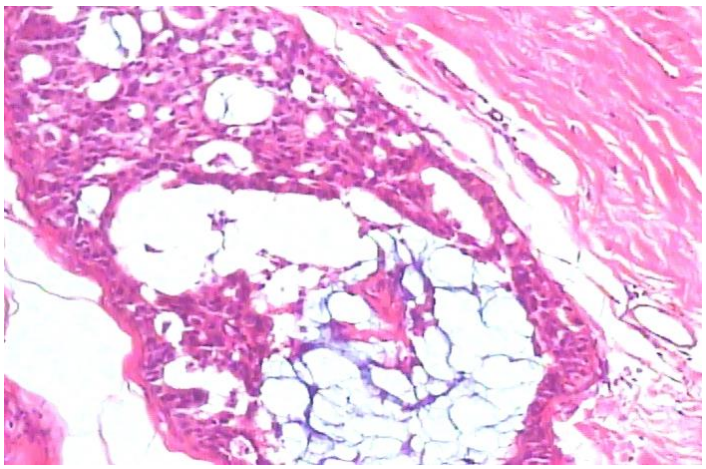


Figure 7- A sample image in mucinous carcinoma class (patient ID=18842D) at x100 magnification factor (image ID: SOB_M_MC-14-18842D-100-004). Prominent amount of mucin in spaces can be seen.

7-2-1-2-4- Papillary carcinoma

Another subtype of BCa included in the BreakHis database is papillary carcinoma. It is typically observed in the central part of breast and is very uncommon subtype of BCa, accounting for 1%-2% of this disease [31, 36]. It is usually characterized by being well circumscribed and the presence of a delicate network of fibrovascular stroma in an arborizing pattern. Either papillary or solid foci formed by ducts almost filled by a solid neoplastic proliferation is observed under the microscope while myoepithelial cells are often absent. Usually the histologic grade of the tumour in this category is 1 or 2 [32]. Figure 8 depicts a papillary carcinoma from the BreakHis database.

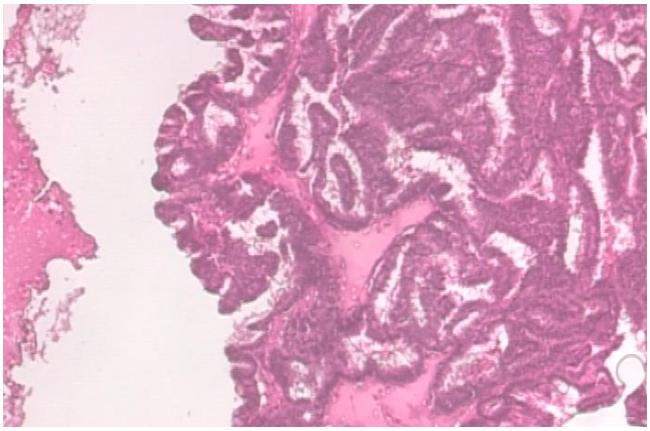


Figure 8- A sample image in papillary carcinoma class (patient ID=9146) at x40 magnification factor (image ID: SOB_M_PC-14-9146-40-004).

7-3- Evaluation of MuDeRN

For evaluation of MuDeRN, 27-fold cross validation was used. The main reason that I used cross validation instead of splitting the data into training, validation, and test sets, was that there was not enough data available for training the ResNets and Metadata without losing significant testing capability. Hence, eighty-one patients were randomly divided into 27 subsets, from which 24 contained one benign patient and two cancer patients and 3 contained three cancer patients. Also, I made sure that all subsets contained at least one patient with ductal carcinoma. This was done because for some categories I had only a few patients and I wanted to make sure that all these patients were not grouped into one subset. Each time one of the sets served as the test set and the rest of the patients were split into the training set with 70 patients and the

validation set with 8 patients. The parameters of the ResNets were estimated based on the training data while the validation data was used for training the MDT for the patient-level diagnosis. As the number of benign images were approximately half of the malignant images, I upsampled the benign class by extracting twice as many patches from the training and validation sets.

References

[1] J. G. Elmore, M. B. Barton, V. M. Moceri, S. Polk, P. J. Arena, and S. W. Fletcher, "Ten-year risk of false positive screening mammograms and clinical breast examinations," New England Journal of Medicine, vol. 338, pp. 1089-1096, 1998.

[2] D. L. Weaver, R. D. Rosenberg, W. E. Barlow, L. Ichikawa, P. A. Carney, K. Kerlikowske, et al., "Pathologic findings from the Breast Cancer Surveillance Consortium: population-based outcomes in women undergoing biopsy after screening mammography," Cancer, vol. 106, pp. 732-42, Feb 15 2006.

[3] K. H. Allison, L. M. Reisch, P. A. Carney, D. L. Weaver, S. J. Schnitt, F. P. O'malley, et al., "Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel," Histopathology, vol. 65, pp. 240-251, 2014.

[4] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. Tosteson, et al., "Diagnostic concordance among pathologists interpreting breast biopsy specimens," Jama, vol. 313, pp. 1122-1132, 2015.

[5] L. Khazai, L. P. Middleton, N. Goktepe, B. T. Liu, and A. A. Sahin, "Breast pathology second review identifies clinically significant discrepancies in over 10% of patients," Journal of surgical oncology, vol. 111, pp. 192-197, 2015.

[6] M. B. Bedell, M. E. Wood, D. C. Lezotte, S. M. Sedlacek, and M. M. Orleans, "Delay in diagnosis and treatment of breast cancer: implications for education," Journal of Cancer Education, vol. 10, pp. 223-228, 1995.

[7] K. McPherson, "Screening for breast cancer-balancing the debate," BMJ: British Medical Journal, vol. 340, 2010.

[8] S. J. Schnitt, J. L. Connolly, F. A. Tavassoli, R. E. Fechner, R. L. Kempson, R. Gelman, et al., "Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria," The American journal of surgical pathology, vol. 16, pp. 1133-1143, 1992.

[9] W. A. Wells, P. A. Carney, M. S. Eliassen, A. N. Tosteson, and E. R. Greenberg, "Statewide study of diagnostic agreement in breast pathology," JNCI: Journal of the National Cancer Institute, vol. 90, pp. 142-145, 1998.

[10] T. J. Lawton, G. Acs, P. Argani, G. Farshid, M. Gilcrease, N. Goldstein, et al., "Interobserver Variability by Pathologists in the Distinction Between Cellular Fibroadenomas and Phyllodes Tumors," International journal of surgical pathology, vol. 22, pp. 695-698, 08/26 2014.

[11] G. Cserni, Z. Orosz, J. Kulka, Z. Sápi, E. Kálmán, and R. Bori, "Divergences in diagnosing nodular breast lesions of noncarcinomatous nature," Pathology & Oncology Research, vol. 12, pp. 216-221, 2006.

[12] T. A. Longacre, M. Ennis, L. A. Quenneville, A. L. Bane, I. J. Bleiweiss, B. A. Carter, et al., "Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study," Modern pathology, vol. 19, p. 195, 2006.

[13] S. L. Jackson, P. D. Frederick, M. S. Pepe, H. D. Nelson, D. L. Weaver, K. H. Allison, et al., "Diagnostic Reproducibility: What

Happens When the Same Pathologist Interprets the Same Breast Biopsy Specimen at Two Points in Time?," Annals of surgical oncology, vol. 24, pp. 1234-1241, 12/02 2017.

[14] B. Weyn, G. van de Wouwer, A. van Daele, P. Scheunders, D. van Dyck, E. van Marck, et al., "Automated breast tumor diagnosis and grading based on wavelet chromatin texture description," Cytometry, vol. 33, pp. 32-40, 1998.

[15] P. Filipczuk, M. Kowal, and A. Obuchowicz, "Multi-label fast marching and seeded watershed segmentation methods for diagnosis of breast cancer cytology," in Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, 2013, pp. 7368-7371.

[16] L. Yang, W. Chen, P. Meer, G. Salaru, L. A. Goodell, V. Berstis, et al., "Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens," IEEE Transactions on Information Technology in Biomedicine, vol. 13, pp. 636-644, 2009.

[17] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in Neural Networks (IJCNN), 2016 International Joint Conference on, 2016, pp. 2560-2567.

[18] F. A. Spanhol, P. R. Cavalin, L. S. Oliveira, C. Petitjean, and L. Heutte, "Deep Features for Breast Cancer Histopathological Image Classification."

[19] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model," Scientific Reports, vol. 7, 2017.

[20] R. H. El Khouli and A. Louie, "Case of the Season: A Giant Fibroadenoma in the Guise of a Phyllodes Tumor; Characterization Role of MRI," Seminars in roentgenology, vol. 44, pp. 64-66, 2009.

[21] F. A. Tavassoli, Pathology of the Breast: McGraw Hill Professional, 1999.

[22] R. A. Jensen, D. L. Page, W. D. Dupont, and L. W. Rogers, "Invasive breast cancer risk in women with sclerosing adenosis," Cancer, vol. 64, pp. 1977-1983, 1989.

[23] C. Wells, I. McGregor, C. Makunura, P. Yeomans, and J. Davies, "Apocrine adenosis: a precursor of aggressive breast cancer?," Journal of clinical pathology, vol. 48, pp. 737-742, 1995.

[24] M. C. Bois, Z. Al-Hilli, D. W. Visscher, T. L. Hoskin, M. H. Frost, D. C. Radisky, et al., "Microglandular adenosis and risk of breast cancer: a Mayo benign breast disease cohort study," in LABORATORY INVESTIGATION, 2016, pp. 32A-32A.

[25] A. R. Frost, S. Terahata, I. T. Yeh, R. S. Siegel, B. Overmoyer, and S. G. Silverberg, "An analysis of prognostic features in infiltrating lobular carcinoma of the breast," Mod Pathol, vol. 8, 1995.

[26] A. Dumitru, A. Procop, A. Iliesiu, M. Tampa, L. Mitrache, M. Costache, et al., "Mucinous Breast Cancer: a Review Study of 5 Year Experience from a Hospital-Based Series of Cases," Mædica, vol. 10, pp. 14-18, 2015.

[27] K. L. Maughan, M. A. Lutterbie, and P. S. Ham, "Treatment of breast cancer," Chemotherapy, vol. 51, p. 53, 2010.

[28] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," IEEE

Transactions on Biomedical Engineering, vol. 63, pp. 1455-1462, 2016.

[29] D. L. Weaver, R. D. Rosenberg, W. E. Barlow, L. Ichikawa, P. A. Carney, K. Kerlikowske, et al., "Pathologic findings from the breast cancer surveillance consortium," Cancer, vol. 106, pp. 732-742, 2006.

[30] J. A. Tice, E. S. O'Meara, D. L. Weaver, C. Vachon, R. Ballard-Barbash, and K. Kerlikowske, "Benign Breast Disease, Mammographic Breast Density, and the Risk of Breast Cancer," JNCI Journal of the National Cancer Institute, vol. 105, pp. 1043-1049, 2013.

[31] A. Talei, M. Akrami, M. Mokhtari, and S. Tahmasebi, "Surgical and Clinical Pathology of Breast Diseases," in Histopathology-Reviews and Recent Advances, ed: InTech, 2012.

[32] P. P. Rosen, Rosen's breast pathology: Lippincott Williams & Wilkins, 2001.

[33] C. I. Li, B. O. Anderson, J. R. Daling, and R. E. Moe, "Trends in incidence rates of invasive lobular and ductal breast carcinoma," Jama, vol. 289, pp. 1421-1424, 2003.

[34] I. Chtourou, S. K. Makni, I. Bahri, K. Abbes, A. Sellami, I. Fakhfakh, et al., "[Pure colloid carcinoma of the breast: anatomoclinical study of seven cases]," Cancer Radiother, vol. 13, pp. 37-41, Jan 2009.

[35] I. K. Komenaka, M. B. El-Tamer, A. Troxel, D. Hamele-Bena, K. A. Joseph, E. Horowitz, et al., "Pure mucinous carcinoma of the breast," Am J Surg, vol. 187, pp. 528-32, Apr 2004.

[36] S. J. Bhosale, A. Y. Kshirsagar, S. R. Sulhyan, S. V. Jagtap, and Y. P. Nikam, "Invasive Papillary Breast Carcinoma," Case Reports in Oncology, vol. 3, pp. 410-415, Sep-Dec 11/13 2010.

Chapter 8

MuDeRN: Multi-category Classification of Breast Histopathological Image Using Deep Residual Networks

This chapter has been published as:

Gandomkar, Ziba, Patrick C. Brennan, and Claudia Mello-Thoms. "MuDeRN: Multi-category Classification of Breast Histopathological Image Using Deep Residual Networks." Artificial Intelligence in Medicine Journal, 2018.

Title of the article:

MuDeRN: Multi-category Classification of Breast Histopathological Image Using Deep Residual Networks

Authors:

Ziba Gandomkar[1], Patrick C. Brennan[1], Claudia Mello-Thoms[1, 2]

[1] Image Optimisation and Perception, Discipline of Medical Imaging and Radiation Sciences, Faculty of Health Sciences, University of Sydney, Sydney, NSW, Australia.

[2] Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA.

The corresponding author:

Ziba Gandomkar

Postal address: Room 726, Level 7, Brain and Mind Centre, 94 Mallett Street, Camperdown NSW 2050, Australia.

E-mail: ziba.gandomkar@sydney.edu.au.

Telephone number: +61291144300

Keywords:

Benign breast lesion, breast cancer, breast cancer subtypes, deep learning, deep residual networks.

Abstract

Motivation: Identifying carcinoma subtype can help to select appropriate treatment options, and determining the subtype of benign lesions can be beneficial to estimate the patients' risk of developing cancer in the future. Pathologists' assessment of lesion subtypes is considered as the gold standard, however, sometimes strong disagreements among pathologists for distinction among lesion subtypes have been previously reported in the literature.

Objective: To propose a framework for classifying hematoxylin-eosin stained breast digital slides either as benign or cancer, and then categorizing cancer and benign cases into four different subtypes each.

Materials and Methods: We used data from a publicly available database (BreakHis) of 81 patients where each patient had images at four magnification factors (x40, x100, x200, and x400) available, for a total of 7786 images. The proposed frame work, called MuDeRN (MUlti-category classification of breast histopathological image using DEep Residual Networks) consisted of two stages. In the first stage, for each magnification factor, a deep residual network (ResNet) with 152 layers has been trained for classifying patches from the images as benign or malignant. In the next stage, the images classified as malignant were subdivided into four cancer subcategories and those categorized as benign were classified into four subtypes. Finally, the diagnosis for each patient was made by combining outputs of ResNets' processed images in different magnification factors using a meta-decision tree.

Results: For the malignant/benign classification of images, MuDeRN's first stage achieved correct classification rates (CCR) of 98.52%, 97.90%, 98.33%, and 97.66% in x40, x100, x200, and x400 magnification factors respectively. For eight-class categorization of images based on the output of MuDeRN's both stages, CCRs in four magnification factors were 95.40%, 94.90%, 95.70%, and 94.60%. Finally, for making patient-level diagnosis, MuDeRN achieved a CCR of 96.25% for eight-class categorization.

Conclusions: MuDeRN can be helpful in the categorization of breast lesions.

Keywords: benign breast lesion, Breast cancer, breast cancer subtypes, deep learning, deep residual networks.

1- Introduction

Breast cancer (BCa) is the most common non-skin cancer among women worldwide. In spite of the increase in incidence rate of BCa over last few decades, the mortality rate from BCa in the developed countries has been decreased due to improvements in treatment options and early detection of BCa through screening mammography [1]. For every 1000 women who have participated in screening mammography, 15.6 to 17.5 need a needle biopsy [2] but only one in four are diagnosed with BCa [3]. Therefore, each year, pathologists evaluate a large number of breast histopathological slides, from which only about 25% contains malignancy, and benign lesions are far more prevalent.

The diagnoses made by pathologists on the cases are usually considered as the gold standard for further treatment of the patients. However, recent studies have shown that the pathologists might disagree with an expert consensus-derived reference diagnoses in distinguishing benign cases from cancer [4-6]. In [5], 6900 individual case diagnoses made by 115 pathologists were compared with an expert consensus-derived ground truth and 17% of benign cases with atypia and 3% of benign cases without atypia were misdiagnosed as ductal carcinoma in situ or invasive carcinoma, while 10% of invasive carcinoma or ductal carcinoma in situ were misdiagnosed as benign cases with or without atypia. Also, it was shown that pathologists who interpret a smaller number of cases per week and those working as a general pathologists make more diagnostic errors than experts [3-5]. Allison et al. [4] divided underlying reasons for disagreement among the pathologists into three categories, which were pathologist-related, diagnostic coding/study methodology-related, and specimen-related. Among pathologist-related factors, “professional differences of opinion on features meeting diagnostic criteria” was ranked first. Computer-assisted analysis can be helpful in reducing the discrepancies in benign/malignant classification by providing an objective classification.

Recently, with the advent of whole slide imaging and production of digital histopathology slides, many researchers have started developing computer-aided detection tools for classification of breast slides as benign or malignant [7]. For example, Weyn et al. [8] used wavelet‑based, Haralick, intensity‑based, and morphological features extracted from segmented nuclei and their surrounding for classification of breast histopathological slides as benign or malignant, and achieved a correct classification rate (CCR) of 79% for case-based classification. In [9], 84 features (morphological, intensity‑based, and textural) extracted from isolated nuclei were utilized for classifying images as either benign or malignant. A sensitivity of 97% and a specificity of 94% has been achieved. However, both methods are computationally expensive as the epithelial nuclei were segmented first. Unlike these methods, Yang et al., [10] extracted textural features using a texton-based approach without segmenting the structures in slides. Using this method, 89% of images were classified correctly.

Pathologists are responsible for not only identifying whether a lesion is malignant or benign but also determining the benign or cancer subtypes, as both benign and malignant breast lesions encompass different subcategories with heterogeneous categories. Different treatment options are available for BCa patients and determining the BCa subtype could be helpful in predicting the patient’s response to therapy, for example, invasive lobular cancer gains a clear benefit from systemic therapy when compared to

invasive ductal cancer [11]. The correct recognition of benign lesion type is also important because the patient's risk of developing subsequent BCa varies among different types of benign lesions [12].

Cserni et al. [13] showed that there are discrepancies among pathologists for determining benign lesion subtypes. The study asked six pathologists to classify benign lesions into three categories, namely fibroadenoma, phyllodes tumor, and anything other these subtypes. The overall Cohen's kappa for categorizing was 0.48, which suggests a moderate agreement [14]. Lawton et al. [15] investigated the agreement among ten pathologists for distinguishing fibroadenomas from phyllodes tumors and found that there was 100% agreement only in 53% of cases. In [16], the interobserver agreement for classification of invasive breast carcinoma was studied and the highest agreement rates among 13 pathologists were achieved for mucinous, lobular, and tubular subtypes, with agreement rates of 96.0%, 78.7%, and 78.0% respectively.

Similar to the malignant/benign classification, computer-assisted analysis could help pathologists to increase diagnostic agreement in multi-class categorization of lesions. In spite of the importance of determining the lesion subtype, only a few previous studies aimed at automatic classification of breast lesions into different subtypes. In [10], six subtypes of BCa were divided into two subgroups: cancer class I that contains ductal carcinoma in situ and lobular carcinoma in situ, and cancer class II containing invasive ductal carcinoma, invasive lobular carcinoma, lymph-node-negative metastasis, and soft tissue metastasis. Using a texton-based approach, images were classified into three classes, i.e. benign, cancer type I, and cancer II and a CCR of 80% was achieved.

Recently there has been growing research on the application of deep learning in medical image segmentation and classification. A few studies have applied deep learning for analyzing breast histopathology slides. In [17], it was used to detect mitotic figures within breast slides. Wang et al. used deep learning for identifying metastatic BCa and obtained an area under the receiver operating curve of 0.925 [18]. Spanhol et al. used AlexNet for classifying breast histopathological images as benign and malignant [19]. In [20], a context-aware stacked convolutional neural network architecture was used for classifying whole slide images as benign, ductal carcinoma in situ, or invasive ductal carcinoma.

This paper focused on three tasks which are: (i) classification of breast histopathological images as benign or malignant, (ii) categorization of malignant images as ductal carcinoma, lobular carcinoma, mucinous carcinoma, or papillary carcinoma; and (iii) classification of benign images as adenosis, fibroadenoma, phyllodes tumour, or tubular adenoma. Previously Han et al. [21] used GoogLeNet [22] for classifying breast histopathological images into similar eight categories and used majority voting for patient classification. Although this aimed at addressing an almost similar problem, we improved both the image-level (i.e. considering each image individually without incorporating the patient information for decision making) and the patient-level (i.e. appointing a single label to each patient by aggregating the class labels assigned to all images of that patient) classification CCRs. This was achieved by first carrying out the stain normalization as a pre-processing step. Secondly, we used a deeper network and a two-stage classifier and thirdly, we utilized a meta-decision tree (MDT) [23] for making the patient-level diagnosis based on the four magnification factors.

In this study, we proposed a frame work, called MuDeRN (MUlti-category classification of breast histopathological image using DEep Residual Networks) for classifying patients based on hematoxylin-eosin stained breast digital slides either as benign or cancer, and then categorizing cancer and benign cases into four different subtypes each. MuDeRN used a very deep residual neural network [24], i.e. a deep residual network with 152 layers (ResNet-152), for classification of breast histopathological images as benign or malignant. Images were acquired in four different magnification factors and for each factor, a separate network has been trained. Malignant images were then subdivided into four subcategories while benign images were classified as four benign subtypes. Eventually, the final diagnosis for a patient was made by combining outputs of networks for different magnification factors using an MDT [23]. It considers the confidence level of the label given by the networks for each magnification factor and also the CCR of the assigned labels to select the best magnification factor for making a patient-level diagnosis. The major contributions of this work are using ResNet for the first time for differentiation of benign and malignant subtypes and also proposing a framework for combining outputs based on different magnification factors to make the ultimate diagnosis for a patient.

2- Materials and Methods

In this section, the dataset we used and MuDeRN's steps are discussed. This study was exempt from the requirement for approval by the Human Research Ethics Committees at the University of Sydney because the data was obtained from a publicly available dataset and all images were de-identified.

The steps of MuDeRN for categorization of lesion subtypes is shown in Figure 1. Briefly, for each patient a set of images from four magnification factors (x40, x100, x200, and x400) were available. To mitigate the color variations, images were normalized by two different methods, which are explained in 2-3. From each normalized image, square image patches were extracted and fed into a ResNet for classification. In each magnification factor, a separated network was trained. The classification was done in two stages. In the first stage (S1) patches were classified either as benign or malignant and an image-level decision was made by using weighted majority voting. On average 24 images were available per patient per each magnification factor. For making patient-level diagnosis, an MDT [23] was used to combine the probabilities of being malignant given to the different images of a patient. The second stage consists of two modules, M and B. The images which were classified as malignant by the first stage were fed into module M where they were subdivided into four cancer subtypes, while those classified as benign were inputted to module B, where they classified into four categories. The architectures of the module in the second stage was almost identical to that of the first stage but for the four classes.

2-1- Dataset

To evaluate the performance of MuDeRN, we used the BreakHis database [25], which is a publicly available dataset of hematoxylin-eosin (HE) stained breast histopathological slide. The images were acquired in four visual magnification factors, namely x40, x100, x200, and x400 with the effective pixel size of 0.49 µm, 0.20 µm, 0.10 µm, and 0.05 µm respectively. The images were stored in a format of three-channel red–green–blue (RGB) TrueColor (24-bit color depth, 8 bits per color channel) color space. For each patient, the pathologist identifies a region of interest (ROI). The undesired areas, such as text annotations or black

border, were removed and the images were cropped to a dimension of 700 ×460 pixels. Finally, out-of-focus images were also discarded.

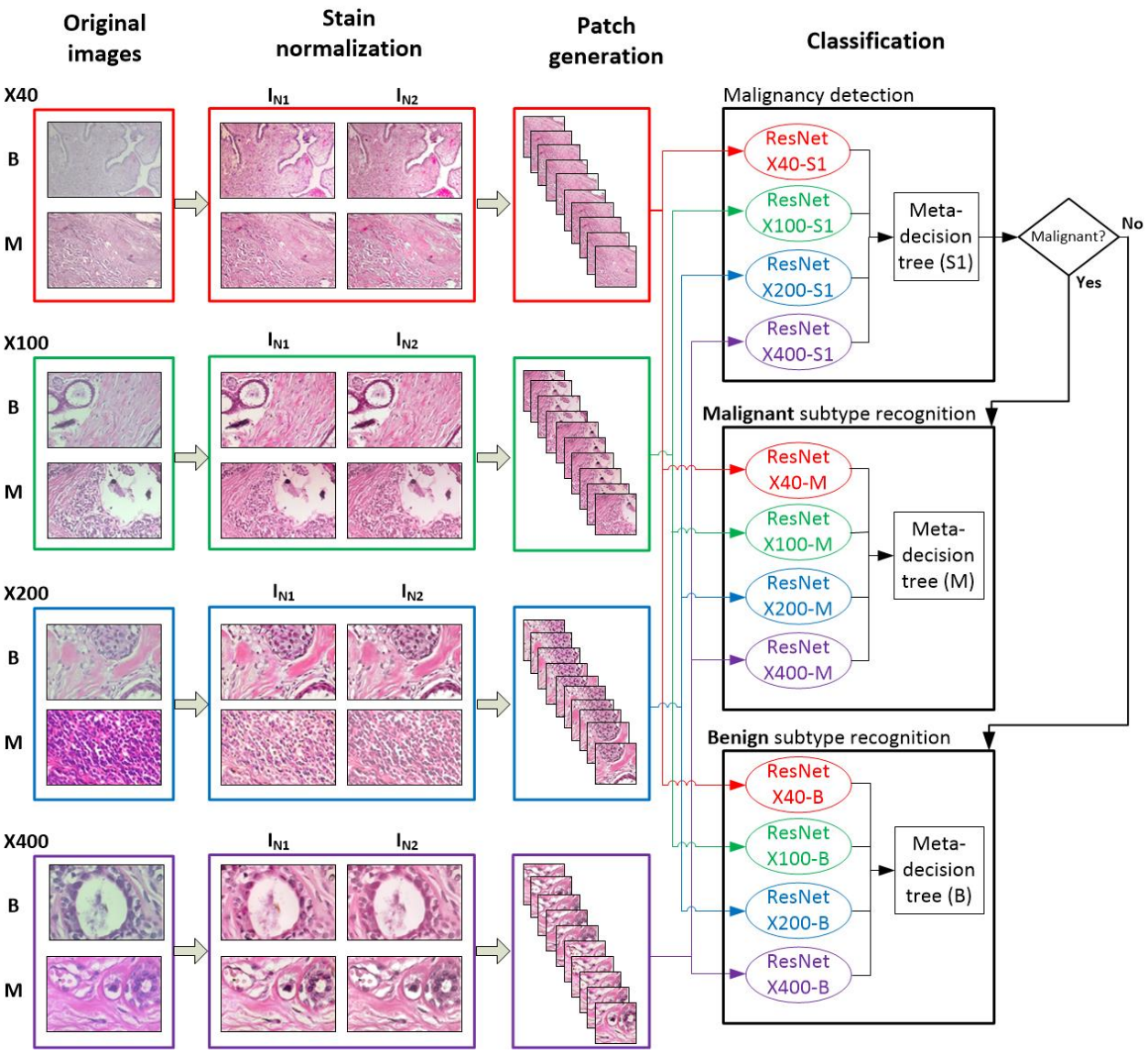


Figure 1- The steps of MuDeRN

On average 24.23, 25.28, 24.46, and 22.15 images were available per patient in x40, x100, x200, and x400 respectively. The BreakHis database comprises of 82 folders corresponding to 82 patients, however, one of the patients (Patient ID: 13412) was a borderline case (has features of both ductal and lobular carcinoma) and hence was placed in both ductal and lobular groups. This patient was included in benign/malignant classification but excluded for tumor sub-type recognition. Figure 2 shows the distribution of images over different sub-types.

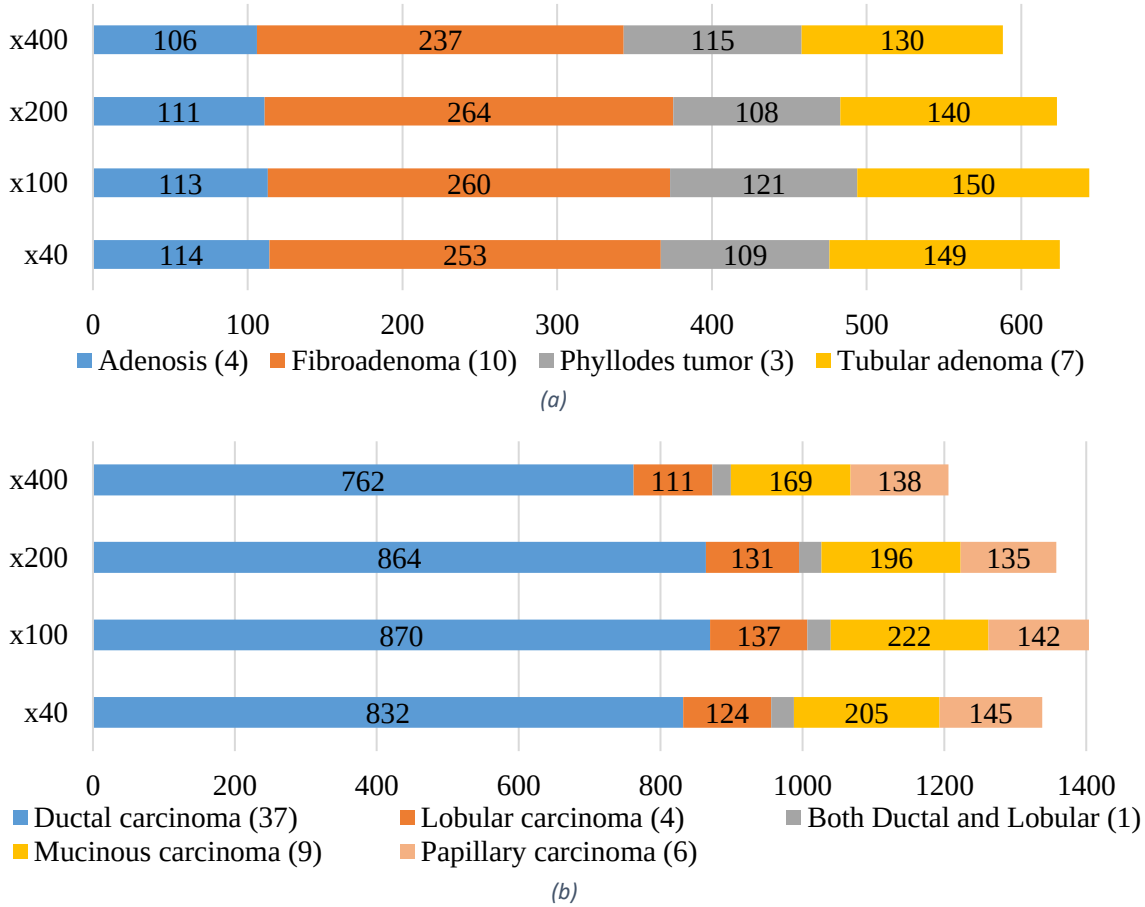


Figure 2- Distribution of (a) benign (b) malignant images by magnification factor and class, number of patients in each category is shown in parentheses. Numbers in each row represent number of images in BreakHis for all patients per each subtype.

2-2- Deep Residual Network

Deep neural networks are cascades of layers of nonlinear processing units that form a hierarchy corresponding to multiple levels of the data representation, starting by learning low-level features (such as edges and lines) to higher-level features (which combine the low-level features to elements of tissue). They are increasingly popular models for automatic classification and segmentation in medical image analysis when large-scale labeled data is available. Although the deep neural networks originated from previously existing artificial neural networks, training their deep architectures have recently become practical due to emergence of high-performance GPU computing, which makes it feasible to train networks with many hidden layers in a reasonable time.

The AlexNet [26] is one of the earliest deep neural networks which contains five convolutional layers followed by fully connected layers. It won the ImageNet Large Scale Visual Recognition Competition-2012. The AlexNet aimed at classification of images into 1000 object categories. Unlike the conventional neural networks which use hyperbolic tangent as the activation function, the AlexNet uses rectified linear units (ReLU) as they are several times faster.

Recent evidence indicated that deeper networks (a network with more layers), such as VGG19 (19 layers) [27] and GoogLeNet (22 layers) [22] achieved better results on the ImageNet dataset. However, simply

stacking more convolutional layers will not lead to a lower classification error and "overly deep plain networks" have higher training error compared to their shallower counterpart [24]. This phenomenon, which is called the degradation problem, could be due to the optimization difficulty of finding the weights of all hidden layers in a feasible time when the network is overly deep. To tackle this problem, ResNet was proposed in [24]. In Figure 3, the building block of the ResNet is compared to the plain network. In the plain networks (Figure 3(a)) the mapping from input to output can be represented by the nonlinear H(x) function. Assume that instead of H(x), F(x)= H(x)-x is used. As shown in Figure 3(b), at the output of the second weight layer x was added to the F(x) and then their sum passes to the ReLU. He et al. [24] showed that adding this shortcut from input to the output of the stacked layers could tackle the optimization difficulties of the deeper networks as the gradient can flow directly from later layers to the earlier layers. As in ResNet, the shortcut connections simply perform identity mapping, extra parameters (and hence extra computational complexity) is not added to the optimization task.

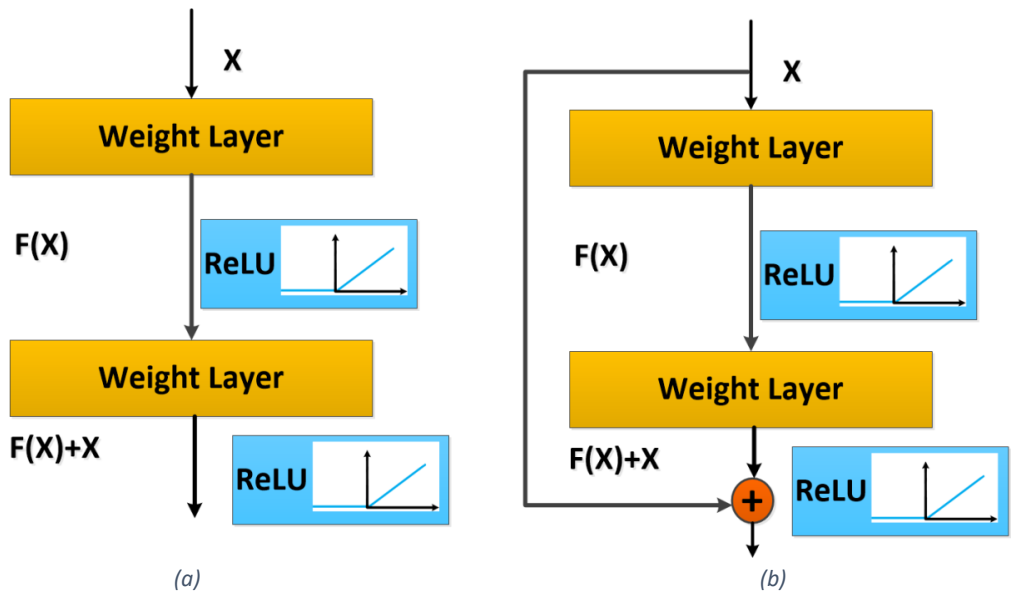


Figure 3- Building block of (a) a plain net (b) a ResNet. ReLU is a rectified linear unit.

ResNet models with 50, 101, and 152 layers were trained and tested on the ImageNet dataset [24]. As expected the error was the least for the ResNet with 152 layers. Previously ResNet-152 was used for analysis of histopathological slide in [28]. It achieved an overall accuracy of 93.0% for classification of colorectal whole-slide images into six classes (five types of colorectal polyps and the normal class) and performed better than ResNet with 50 and 101 layers. Therefore, in this study, we used the ResNet-152 for classification.

2-3- Stain normalization

Inconsistencies in color are major issues in analysis of histopathological slides. The inconsistencies could be due to different reasons such as the use of different chemicals for staining, variations in color concentrations, or differences in scanners from different vendors. Different algorithms have been suggested for stain normalization. Each algorithm has its own advantages and limitations and works better for a group of images but has some flaw when applied to other images. Hence, here we used two different stain normalization methods and produce two stain normalized images, I_{N1} and I_{N2}, for each image.

I_{N1} was produced using a stain normalization algorithm based on histogram specification [29], where the images from a patient were transformed to a set of new images so that the histograms of the output images in different color channels approximately match the target image histogram in the corresponding channel. Figure 4 indicates the target image that used in this study. This image was selected from the Mitosis-Atypia database[1] based on the opinion of a pathologist who was asked to select an image with an appropriate staining which includes lumen, stroma and epithelial cells and does not have any artifact or tissue folding. We did the histogram matching for the stack of images of a given patient in each magnification factor rather than doing the normalization image by image. This was done to mitigate the visual artifacts in the images due to the assumption of the histogram specification that the proportion of pixels in each color is almost identical in the source and target image. If each time only a single image had been considered, this assumption might have been violated as only a very small tissue area with limited tissue elements and hence limited number of colors would have been taken into account. The second approach used for stain normalization was first proposed in [30], where the mean and standard deviation of each channel of the images were matched to that of the reference image by using a set of linear transformation in La*b* color space.

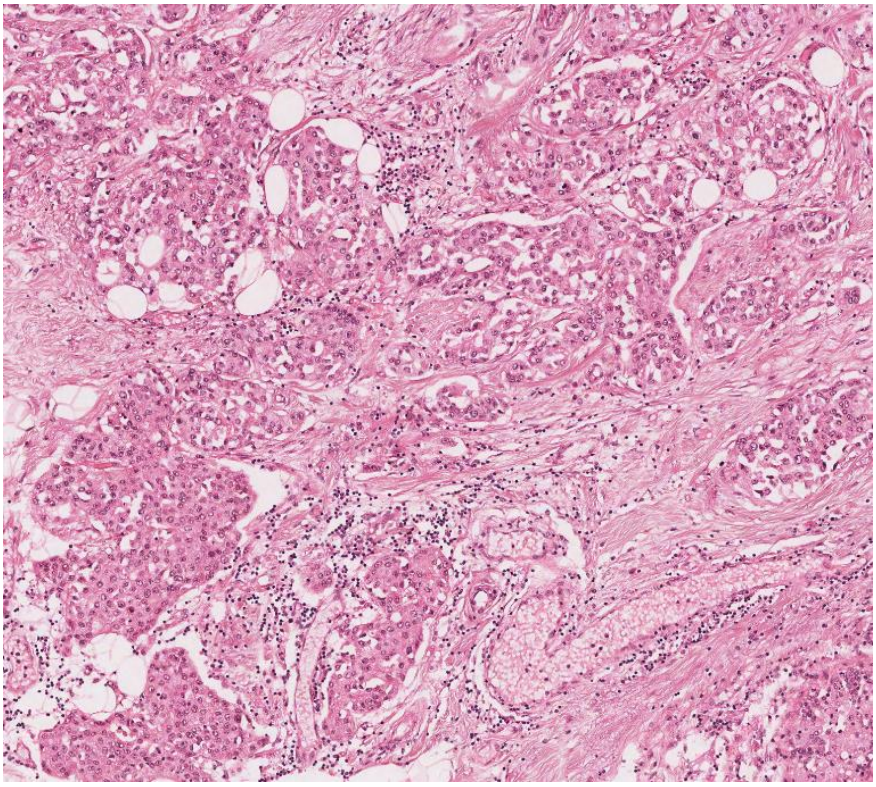


Figure 4- This target image was used as a reference image to which all the images were mapped.

2-4- Stage 1: Benign/Malignant classification

As shown in figure 1, in the first stage binary classification was performed for detecting the malignant and benign cases. In this section, first we explain the steps of MuDeRN for classifying images of a patient as

[1] http://mitos-atypia-14.grand-challenge.org/

malignant or benign. Then we explain in details how MuDeRN was trained and tested by using BreakHis database.

The first stage was composed of four ResNets, where each was trained for classifying breast histopathological images of a specific magnification factor. The input size of the ResNets was 224 x 224 while the images in BreakHis database were 700 ×460 pixels. Therefore, both I_{N1} and I_{N2}, were resized to 341×224. Then five overlapping patches with size of 224 x 224 were extracted from each one of the stained normalized images by using a sliding window. Therefore for each image, ten patches (five from I_{N1} and five from I_{N2}) were extracted. As stated in section 2.1, for each patient, images in four magnification factors were available. N_{x40}, N_{x100}, N_{x200}, and N_{x400} indicate the number of images in x40, x100, x200, and x400 magnification factors. Assume the ith image in the xth magnification factor of a patient was inputted to the ResNet for the xth magnification factor, then the probability that the image is belonging to the jth class, was found using (1).

$$CL_j^{i,x} = \frac{\sum_{p=1}^{10} cl_{j,p}^{i,x}}{10} \quad ,j \in \{M, B\}, p \in \{1, .., 10\} \tag{1}$$

Where $(i, x) \in \{(1, x40), \dots, (N_{x40}, x40), (1, x100), \dots, (N_{x100}, x100), \dots, (N_{x200}, x200), \dots, (N_{x400}, x400)\}$.

M and B represent malignant and benign classes and $cl_{j,p}^{i}$ shows the probability of the pth patch of the ith image belonging to the jth class. To find the class label for each image, the weighted majority voting was used. Therefore, the label assigned to the ith image was $J_{x,i} = \underset{j}{argmax}\, CL_j^{x,i}$. The image-level CCR was calculated as the number of images in each magnification factor which were classified correctly by the total number of images in that magnification factor.

In order to make the final diagnosis for a patient, image-level diagnoses for four magnification factors have been combined using a MDT. As stated earlier, on average we have approximately 24 images per patient per magnification factor. First, in each magnification factor, the average probability of a patient belonging to jth class was calculated using (2) where N_x shows number of images in the xth magnification factor.

$$CL_j^x = \frac{\sum_{i=1}^{N_x} CL_j^{i,x}}{N_x} \tag{2}$$

In each magnification factor, the class maximizing $CL_j^x, J_x = \underset{j}{argmax}\, CL_j^x$, was found and for each patient, we had $\{J_{x40}, \underset{j}{max}\, CL_j^{x40}, J_{x100}, \underset{j}{max}\, CL_j^{x100}, J_{x200}, \underset{j}{max}\, CL_j^{x200}, J_{x400}, \underset{j}{max}\, CL_j^{x400}\}$. This was then fed into the MDT (S1) to make the final diagnosis for a patient. We used MDTs as suggested in [23] for combining multiple classifiers. The output of the MDT identify ResNet from which magnification factor should be considered to assign the label for a test data by considering the confidence level of the label assigned in different magnification factors and the performances of each magnification factor when used for classifying the validation set. So, the MDT specifies the best magnification factor for a specific test patient. For example if the confidence of the ResNet in the magnification factor of x200 was 100% (the highest one compared to all other magnification factors) for a test data and the CCR of the ResNet in x200 for the

validation set was also 100%, then the MDT would assign the label of test data as outputted by the ResNet in the magnification factor of x200.

For evaluation of MuDeRN, 27-fold cross validation was used. The main reason that we used cross-validation instead of splitting the data into training, validation, and test sets, was that there was not enough data available for training the ResNets and Metadata without losing significant testing capability.

Eighty-one patients were randomly divided into 27 subsets, from which 24 contained one benign patient and two cancer patients and 3 contained three cancer patients. Also, we made sure that all subsets contained at least one patient with ductal carcinoma. This was done because for some categories we had only a few patients and we wanted to make sure that all these patients were not grouped into one subset. Each time, one of the sets served as the test set and the rest of the patients were split into the training set with 70 patients and the validation set with 8 patients. The parameters of ResNets were estimated based on the training data while the validation data was used for training the MDT for the patient-level diagnosis.

As the number of benign images were approximately half of the malignant images, we upsampled the benign class by extracting twice as many patches from the training and validation sets. Therefore, for each benign image, 20 patches were extracted from I_{N1} and I_{N2} while ten patches were extracted from each malignant image.

Training the ResNet from the scratch (i.e. the random initialization of the network's weight and training the model) requires a very large scale dataset as the network has a large number of parameters. Therefore, we fine-tuned the ResNet, previously trained on the ImageNet (1.2M labeled images) by continuing to train it on the training set using stochastic gradient descent (SGD) with back-propagation with a small learning rate (0.0001) for 50 epochs. Although the classification task done on the ImageNet data set is completely different to breast histopathological image classification, due to lack of training data, making a model from the scratch was not feasible. In [21], it was shown that the accuracy of GoogLeNet for eight-class classification of breast histopathological slides was higher for fine-tuning in comparison with training from the scratch. Before start training, the last classification layer of the pre-trained ResNet model was removed and replaced with a classification layer with only two classes as ResNet had been trained for classifying image into 1000 categories.

Image augmentation artificially creates training images through different ways of processing or combination of multiple processing and is usually required to improve the performance of deep networks and avoid overfitting of the network to the training set. Here the training data was augmented by random combination of image rotation by 90°, 180°, or 270°, flipping about horizontal or vertical axes, and random horizontal and vertical shifting between ±10 pixels. In each epoch, one pass over the training patches was completed and in each pass, the image patches were randomly augmented.

The validation set has only eight members, which makes training the MDT difficult. Therefore, an up-sampling strategy was required. Assume $P_j | \, j \in \{M, B\}$ indicates total number of patches extracted from an image. Here we set P_M and P_B to 10 and 20 respectively as number of benign images were approximately half of that of the malignant images. For upsampling, we randomly grouped patches from

different images of a single patient together in a way that each group contains only one patch from a particular image; hence for each patient, P_j samples were generated. For example, when the validation set contains 2 benign and 6 malignant patients, in total 100 samples, i.e. 2 (number of benign patients) × 20 (P_B)+ 6 (number of malignant patients) × 10 (P_M). Therefore, for the confidence level of each group of patches,$\widehat{CL}_{j,p}^{x}$, we will have

$$\widehat{CL}_{j,p}^{x} = \frac{\sum_{i=1}^{N_x} cl_{j,p}^{i,x}}{N_x} \tag{3}$$

Similar to (2), for each magnification factor, the class with the maximum value, $\hat{J}_{x,p} = \underset{j}{argmax}\, \widehat{CL}_{j,p}^{x}$, was found. Therefore we have P_j samples for each patient and the data used for training the MDT had the format of $\left\{ \hat{J}_{x,p}, \underset{j}{max}\, \widehat{CL}_{j,p}^{x} \mid x \in \{x40, x100, x200, x400\}, j \in \{M, B\}, p \in \{1, \dots, P_j\} \right\}$.

After estimating the parameters of the ResNets in four magnification factors and also the parameters of the MDT, for each test image, ten patches were randomly selected from I_{N1} and I_{N2}. Patches in each magnification factor were fed into the corresponding ResNet, which outputted $cl_{j,p}^{i,x}$. Using (1) and (2), the value of CL_j^x was calculated for each image. Finally for each patient in the test subset $\{ J_{x40}, \max_j CL_j^{x40}$, J_{x100} , $\max_j CL_j^{x100}$, J_{x200} , $\max_j CL_j^{x200}$, J_{x400} , $\max_j CL_j^{x400}\}$ was generated and inputted to the trained MDT to make the final diagnosis for each patient.

2-5- Stage 2: Differentiation of lesion sub-types

Based on the decision made in the first stage, malignant images were fed into the module M as shown in Figure 1, and categorized into four cancer subtypes, while benign images were inputted to the module B and classified into four categories.

The architecture of both modules were almost identical to what was used in the first stage. However, there were two differences between them. First, in S1 we used the model pre-trained on the ImageNet data set as a starting point for the fine-tuning. Here, we used the ResNet trained at the S1 as the starting point. We hypothesized that the first few layers of the ResNet already learned the low-level features for describing the breast histopathological images and hence it could be a better starting point.

Second, the number of classes in this stage is four per each module, therefore we have $j \in \{1,2,3,4\}$. In the module for processing benign images, the total number of patches extracted from images in the training and validation sets for each class, P_j, were 24 for adenosis and phyllodes tumor, 10 for Fibroadenoma, and 16 for tubular adenoma. Similarly the minority classes were upsampled in the module M, resulting to 10, 70, 40, and 60 patches per image for ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma respectively. Each time half of the patches were extracted from I_{N1} while the other half were extracted from I_{N2}. In S1, a sliding window was used to extract overlapping image patches, here we used sliding window but we also randomly rotated (0°, 90°, or 180°) and flipped (none, horizontal, or vertical) the image patches as well. This was done because for the minority classes

we extracted 20-35 patches per normalized image and different patches are almost identical with a very slight shift.

3- Results

The values of CCRs for the ResNets processing images of different magnification factors in the first stage are shown in figure 3. For all magnification factors, CCRs for the benign category were lower than those for the malignant one. The differences between CCR of benign and malignant categories are significant for all magnification factors (x40: z=-2.32, p-value=0.020; x200: z=- 2.48, p=0.013; x200: z=- 2.88, p-value=0.004; x400: z=- 3.07, p-value=0.002). The overall CCR varied across different magnification factors and ranged from 97.9% to 98.3%, but the differences among overall CCRs of different magnification factors were not statistically significant.

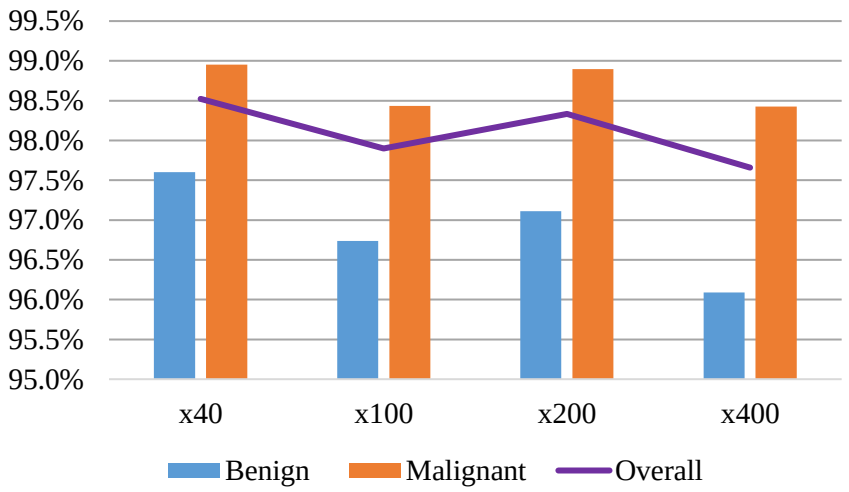


Figure 5- Accuracy of ResNets in the first stage for malignant/benign classification of images in different magnification factors

The CCR values for the recognition of different benign and malignant subtypes in various magnification factors are listed in Tables 1 and 2 respectively. In table 1, we included all benign images, no matter whether they were detected correctly by the malignant/benign classification module. Similarly, all malignant images regardless of their labels from the first stage were included in table 2. Table 3 indicates the overall CCRs when outputs from both stages were combined. Hence, the image was considered correctly classified when it was assigned the appropriate label by the first stage and then in the next stage, the subtype was correctly identified. Here we presented results for both four-class categorization (Tables 1 and 2) and eight-class categorization (Table 3) separately because we wanted to show the performances of stand-alone modules for distinction among benign subtypes and distinguishing among cancer subtypes, as some pathologists might prefer to classify images into benign and malignant themselves and use MuDeRN to aid in the distinction among subtypes, so that the error from the first stage does not propagate in the classification done by the second stage.

As shown in Table 1, overall CCR value of x200 magnification factor was the highest, however, the differences among the CCR values for different magnification factors were not significantly different. For adenosis and Fibroadenoma, the x200 magnification factor achieved the highest CCR value while for the

Phyllodes tumor class, the highest CCRs was achieved when images from the lowest magnification factor were classified. For tubular adenoma CCR value of x100 magnification factor was the highest.

Table 1- Accuracy of the MuDeRN's module for the recognition of different benign subtypes in various magnification factors. The highest CCR for each class is shown in bold.

	x40	x100	x200	x400
Adenosis	89.47%	87.61%	**89.19%**	86.79%
Fibroadenoma	98.02%	96.15%	**99.24%**	97.47%
Phyllodes tumor	**94.50%**	92.56%	93.52%	94.78%
Tubular adenoma	94.63%	**96.67%**	95.00%	95.38%
Overall	95.04%	94.10%	**95.51%**	94.56%

As shown in Table 2, the overall CCR value for x200 magnification factor was the highest, which is similar to the results presented in Table 1 for the recognition of different benign subtypes. Here for each class, a different magnification factor resulted in the highest CCR for each class.

Table 2- Accuracy of the MuDeRN's module for the recognition of different malignant subtypes in various magnification factors. The highest CCR for each class is shown in bold.

	x40	x100	x200	x400
Ductal carcinoma	**98.44%**	97.82%	98.26%	98.29%
Lobular carcinoma	97.58%	97.81%	97.71%	**98.20%**
Mucinous carcinoma	96.59%	96.85%	**97.45%**	96.45%
Papillary carcinoma	95.86%	**96.48%**	96.30%	96.38%
Overall	97.78%	97.52%	**97.89%**	97.80%

Table 3 shows the MuDeRN's CCR at the image-level in each magnification factor and for each class. As shown, for two benign subtypes and two cancer subtypes, x100 magnification factor performed the best and for the rest of the subtypes, x200 magnification factor outperformed others.

Table 3- Accuracy of MuDeRN in different magnification factors. The highest CCR for each class is shown in bold.

	x40	x100	x200	x400
Adenosis	82.46%	85.84%	**89.19%**	86.79%
Fibroadenoma	96.44%	92.31%	**95.83%**	93.25%
Phyllodes tumor	**93.58%**	93.39%	92.59%	89.57%
Tubular adenoma	**93.29%**	92.67%	91.43%	92.31%
Ductal carcinoma	**97.96%**	97.13%	97.57%	97.38%
Lobular carcinoma	95.16%	95.62%	**96.18%**	94.59%
Mucinous carcinoma	**95.61%**	94.59%	95.41%	93.49%
Papillary carcinoma	95.17%	96.48%	95.56%	**95.65%**
Overall	95.60%	94.89%	**95.69%**	94.63%

Finally, as explained in section 2.5, MDT were used to make a patient-level diagnosis through combining the outputs from different magnification factors. A CCR of 98.77% was achieved in the first stage for classification of patients either as benign or malignant while the overall CCR for the patient-level diagnosis was 96.25%.

Training the MDT added an extra computational burden to the algorithm, so one might question the advantage of MDT over a non-trainable aggregation strategy. The most common nontrainable way to aggregate image-level classification and produce patient-level diagnose is majority voting of the image-level results. We compared the performance of MDT with that of majority voting to explore the added benefit of using MDT for aggregating the image-level results. Figure 6 shows the comparison of two aggregation methods. As shown overall the MDT method performed about 4% better than nontrainable method. As it can be seen, the differences between two methods varied for different diseases. The advantage of MDT was more prominent for different types of benign diseases.

A few recent studies used AlexNet and GoogleNet [22] for binary-class classification using data from the same database. In table 4, the image-level CCRs in different magnification factors of our algorithm were compared with those of studies that used the same database. Z-test for proportions showed that the CCR of MuDeRN was significantly higher than other methods in all magnification factors ($p<0.05$). As shown in table 4, for eight-class categorization of images, the CCRs were also improved significantly compared to GoogLeNet exept for x100 magnification.

In patient-level, the GoogLeNet achieved the best CCR for x40 magnification factor which was 97.1% while the best CCR for AlexNet was 90% at the same magnification factor. MuDeRN outputted a single patient-level diagnosis for all magnification factors which differed significantly from that of AlexNet ($z= 2.42$, p-value=0.015) but was not significantly different from that of GoogLeNet ($z= 0.75$, p-value=0.453). For eight-class categorization in patient-level, the CCR of MuDeRN was about 1.55% higher that than that of the GoogLeNet (the best result was obtained for x200) but the difference was not significant ($z=-0.45$, p-value= 0.624).

Table 4- Comparison of the MuDeRN with the state-of-the-art accuracy for malignant/benign classification and malignant and benign subtype identification of images. The second highest CCR for the patient-level classification and also the second highest CCR for image-level classification per magnification factor have been underlined.

	Image-level				**Patient-level**			
Two-class classification	**x40**	**x100**	**x200**	**x400**	**x40**	**x100**	**x200**	**x400**
Hand-crafted features **[25]**	82.8%	80.7%	84.2%	81.2%	83.8%	82.1%	85.1%	82.3%
AlexNet (32*32 patches) **[19]**	89.6%	85.0%	84.0%	80.8%	88.6%	84.5%	85.3%	81.7%
Combination of AlexNets **[19]**	85.6%	83.5%	83.1%	80.8%	90.0%	88.4%	84.6%	86.1%
Deep features**[31]**	84.6%	84.8%	84.2%	81.6%	84.0%	83.9%	86.3%	82.1%
AlexNet17 **[21]**	85.6%	83.5%	83.1%	80.8%	90.0%	88.4%	84.6%	86.1%
GoogLeNet **[21]**	<u>95.8%</u>	<u>96.9%</u>	<u>96.7%</u>	<u>94.9%</u>	<u>97.1%</u>	95.7%	96.5%	95.7%
MuDeRN	**98.5%**	**97.9%**	**98.3%**	**97.7%**	**98.77%**			
Eight-class classification								
AlexNet17 **[21]**	86.4%	75.8%	72.6%	84.6%	74.6%	73.8 %	76.4%	79.2%

GoogLeNet **[21]**	92.8%	93.9%	93.7 %	92.9%	94.1%	93.2%	94.7%	93.5%
MuDeRN	**95.6%**	**94.9%**	**95.7%**	**94.6%**	**96.25%**			

4- Discussion

In this paper, MuDeRN, a framework was proposed to classify patients based on HE stained breast histopathological images either as benign or malignant, and also categorize them into eight classes, representing different subtypes of benign lesions and carcinomas. MuDeRN consisted of two stages. The first stage had a single module composed of four ResNets, where each one dealt with a specific magnification factor and a MDT for combining image-level predictions to classify patients either as benign or malignant. The second stage was comprised of two modules, one for categorizing malignant images into four subtypes and one for classifying benign images into four subcategories. MuDeRN was tested on a database containing 7786 images in four magnification factors from 81 patients. It achieved an average CCR of 98.10% over all magnification factors for classifying the images as benign or malignant while an average CCR of 95.15% for classifying images into eight classes. At the patient-level, MuDeRN achieved a CCR of 98.77% for malignant/benign classification, and 96.25% for the eight-category classification.

As shown in table 3, the CCR values varied among different subtypes. This could be due to the fact that numbers of patients in different subtypes were not similar. For example, ductal carcinoma which was achieved the highest CCR had the highest number of patients as well. By providing larger number of cases, the ResNets learn the characteristics of lesion better. For the adenosis subtype, the CCR was the lowest. That could be due to the reason that the adenosis has different subtypes (i.e. sclerosing, tubular, apocrine, microglandular) and larger number of cases is required so that the network learns features of different variations of the disease.

For binary classification, the ResNet processing the images from the lowest magnification factor, i.e. x40, achieved the highest overall CCR. This is in line with the results obtained in [25] where the conventional classifiers and the textural features were used for the binary classification the BreakHis database. The pathologists also start by evaluating the slide in the lowest magnification factor and then zoom in to a few areas at the higher magnification factors for making the final diagnosis. This behavior could explain the fact that the images in the lowest magnification factor of the database (i.e. x40 magnification factor) had slightly more discriminative power compared to other magnification factors. For eight-category classification, the ResNet analyzing images of x200 magnification factor achieved the highest CCR value. This could imply that the x40 magnification factor is more informative for making decision about existence of malignancy however further information, especially cytological features, from the higher magnification factors is required for identification of lesion subtypes.

As shown in table 4, ResNet performed better than GoogLeNet which itself had a higher CCR compared to the AlexNet in image-level binary classification. This could be due to the fact that ResNet is deeper than GoogLeNet which is deeper than AlexNet. Another contributing factor for achieving a higher CCR in this study could be stain normalization; as in [21], the images were not stain normalized. Also, the images were downsized to 256 × 256 in [21]. As the aspect ratio of the original images in BreakHis was about 1.52, resizing them could change the aspect ratio of the structures within the tissue and result in altering some

informative features of the image. Here we extracted square patches from the images and then used the weighted majority voting for image-level classification. Similarly, for eight-class classification of images, the ResNet outperformed the GoogLeNet and the differences were significant in all magnification factors. In the patient-level, the differences between the performance of GoogLeNet (at the best magnification factor) and MuDeRN was not significantly for both binary and eight-category classification. This could be due to small sample size and also absence of borderline cases. By including in situ cases to the database, which are more challenging, the differences between the performances could become significant.

This study has a number of limitations. First, cases with non-invasive BCa (ductal carcinoma in situ and lobular carcinoma in situ) were not included in the BreakHis database. These types of BCa are pre-invasive and demonstrate features between benign and invasive cancer and making diagnoses for these cases are more difficult. Although the results obtained in this study are promising, that could be to some extent because of lack of the borderline cases in the database. Therefore, including in situ cases could be a possible avenue for future work. Secondly, the regions of interest were manually selected by the pathologists in the BreakHis database, which makes MuDeRN semi-automatic. Therefore, one potential future work could be adding a preprocessing stage which automatically selects the diagnostically relevant areas of the whole slide images. Also, the target image for stain normalization was selected manually based on opinion of a pathologist. Selecting a different image as the target image could affect the appearance of stain normalized images and some variability exists in this selection, which will propagate through the framework. Thirdly, in the BreakHis database only four benign subtypes and four cancer subtypes were considered, however, both benign lesions and invasive cancer have other subtypes which should be included. In addition, for some subtypes, only a few cases were included and the performance of MuDeRN should be investigated on a larger database including more patients from theses subtypes. Also, the performance of MuDeRN as the second reader should be evaluated. Providing an independent second opinion could be particularly helpful when the slides were evaluated by general or less experienced pathologists.

Acknowledgment

We would like to thank contributors of BreakHis database content who kindly provided us the access to their database. We also acknowledge the University of Sydney HPC Service at the University of Sydney for providing high performance computing resources that have contributed significantly to the research results reported within this paper.

References

[1] J. Erlay, M. Ervik, R. Dikshit, S. Eser, and C. Mathers, "Cancer incidence and mortality worldwide: IARC CancerBase No. 11," in *GLOBOCAN 2012 v1. 0*, ed, 2012.

[2] N. Calonge, D. B. Petitti, T. G. DeWitt, A. J. Dietrich, K. D. Gregory, D. Grossman*, et al.*, "Screening for breast cancer: US Preventive Services Task Force recommendation statement," *Annals of internal medicine,* vol. 151, pp. 716-726, 2009.

[3] D. L. Weaver, R. D. Rosenberg, W. E. Barlow, L. Ichikawa, P. A. Carney, K. Kerlikowske*, et al.*, "Pathologic findings from the Breast Cancer Surveillance Consortium: population-based outcomes in women undergoing biopsy after screening mammography," *Cancer,* vol. 106, pp. 732-42, Feb 15 2006.

[4] K. H. Allison, L. M. Reisch, P. A. Carney, D. L. Weaver, S. J. Schnitt, F. P. O'malley*, et al.*, "Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel," *Histopathology,* vol. 65, pp. 240-251, 2014.

[5] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. Tosteson*, et al.*, "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *Jama,* vol. 313, pp. 1122-1132, 2015.

[6] L. Khazai, L. P. Middleton, N. Goktepe, B. T. Liu, and A. A. Sahin, "Breast pathology second review identifies clinically significant discrepancies in over 10% of patients," *Journal of surgical oncology,* vol. 111, pp. 192-197, 2015.

[7] Z. Gandomkar, P. C. Brennan, and C. Mello-Thoms, "Computer-based image analysis in breast pathology," *Journal of pathology informatics,* vol. 7, 2016.

[8] B. Weyn, G. van de Wouwer, A. van Daele, P. Scheunders, D. van Dyck, E. van Marck*, et al.*, "Automated breast tumor diagnosis and grading based on wavelet chromatin texture description," *Cytometry,* vol. 33, pp. 32-40, 1998.

[9] P. Filipczuk, M. Kowal, and A. Obuchowicz, "Multi-label fast marching and seeded watershed segmentation methods for diagnosis of breast cancer cytology," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, 2013, pp. 7368-7371.

[10] L. Yang, W. Chen, P. Meer, G. Salaru, L. A. Goodell, V. Berstis*, et al.*, "Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens," *IEEE Transactions on Information Technology in Biomedicine,* vol. 13, pp. 636-644, 2009.

[11] R. Barroso-Sousa and O. Metzger-Filho, "Differences between invasive lobular and invasive ductal carcinoma of the breast: results and therapeutic implications," *Therapeutic advances in medical oncology,* vol. 8, pp. 261-266, 2016.

[12] M. Guray and A. A. Sahin, "Benign breast diseases: classification, diagnosis, and management," *The oncologist,* vol. 11, pp. 435-449, 2006.

[13] G. Cserni, Z. Orosz, J. Kulka, Z. Sápi, E. Kálmán, and R. Bori, "Divergences in diagnosing nodular breast lesions of noncarcinomatous nature," *Pathology & Oncology Research,* vol. 12, pp. 216-221, 2006.

[14] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics,* pp. 159-174, 1977.

[15] T. J. Lawton, G. Acs, P. Argani, G. Farshid, M. Gilcrease, N. Goldstein*, et al.*, "Interobserver Variability by Pathologists in the Distinction Between Cellular Fibroadenomas and Phyllodes Tumors," *International journal of surgical pathology,* vol. 22, pp. 695-698, 08/26 2014.

[16] T. A. Longacre, M. Ennis, L. A. Quenneville, A. L. Bane, I. J. Bleiweiss, B. A. Carter*, et al.*, "Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study," *Modern pathology,* vol. 19, p. 195, 2006.

[17] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2013, pp. 411-418.

[18] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718,* 2016.

[19] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *Neural Networks (IJCNN), 2016 International Joint Conference on*, 2016, pp. 2560-2567.

[20] B. E. Bejnordi, G. Zuidhof, M. Balkenhol, M. Hermsen, P. Bult, B. van Ginneken*, et al.*, "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," *arXiv preprint arXiv:1705.03678,* 2017.

[21] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model," *Scientific Reports,* vol. 7, 2017.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov*, et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[23] L. Todorovski and S. Džeroski, "Combining classifiers with meta decision trees," *Machine learning,* vol. 50, pp. 223-249, 2003.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[25] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering,* vol. 63, pp. 1455-1462, 2016.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[28] B. Korbar, A. M. Olofson, A. P. Miraflor, K. M. Nicka, M. A. Suriawinata, L. Torresani*, et al.*, "Deep-Learning for Classification of Colorectal Polyps on Whole-Slide Images," *arXiv preprint arXiv:1703.01550,* 2017.

[29] S. Kothari, J. H. Phan, R. A. Moffitt, T. H. Stokes, S. E. Hassberger, Q. Chaudry*, et al.*, "Automatic batch-invariant color segmentation of histological cancer images," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, 2011, pp. 657-660.

[30] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications,* vol. 21, pp. 34-41, 2001.

[31] F. A. Spanhol, P. R. Cavalin, L. S. Oliveira, C. Petitjean, and L. Heutte, "Deep Features for Breast Cancer Histopathological Image Classification."

Chapter 9

Discussion

9-1- Thesis Overview and Major Contributions

Each year pathologists evaluate a large volume of breast specimens from which only 25% are diagnosed with breast cancer and the rest of them are benign [1]. The pathologists' workflow when assessing a breast biopsy is shown in figure 1. As shown, pathologists determine whether the biopsy is malignant or benign and then identify the subtype of benign or malignant lesions [2]. When a malignant mass is present, they also determine the stage and grade of breast cancer [2]. For staging breast cancer, pathologists evaluate the mass size and determine whether the cancer has metastasized (that is, whether the cancer has spread to the lymph nodes) [3]. As shown in figure 1 (under grading the breast cancer), for grading a breast cancer three factors are considered: the magnitude of glandularity (glandular/tubular structures formation in tumour area), magnitude of nuclear atypia (changes in nuclear appearance), and number of mitoses per 10 high power fields [4]. Making diagnosis, grading and staging are done based on assessing Hematoxylin-Eosin stained slides, but pathologists are also responsible for the quantification of three immunohistochemical stains (i.e., estrogen receptor, progesterone receptor, and human epidermal growth factor 2) [2].

Pathologists prepare a report including the diagnosis as well as the grade and stage of a breast cancer (if present) for other clinicians involved in patient care, such as a surgeon or a breast physician. The pathology reports are considered as the gold standard and used for selecting the appropriate treatment for the patient. However, recent studies have shown that there are discrepancies among pathologists in making diagnosis, staging and grading breast cancer [5-16]. Some of these discordances could lead to delay in treatment, unnecessary treatments [7, 8], inappropriate

undertreatment [17], differences in patient management [7, 8], and also impact patients' risk for having a subsequent breast cancer [18].

With the advent of whole slide imaging, computer-assisted analysis of breast histopathological slides became possible [19, 20]. Computer-assisted analysis can aid pathologists in making diagnoses, staging and grading the breast cancer, and quantification of immunohistochemical stains [21]. It also can be used to enhance educational schemes, to reduce disagreement and provide better understanding about image-related features that lead to discordance among pathologists [22-24].

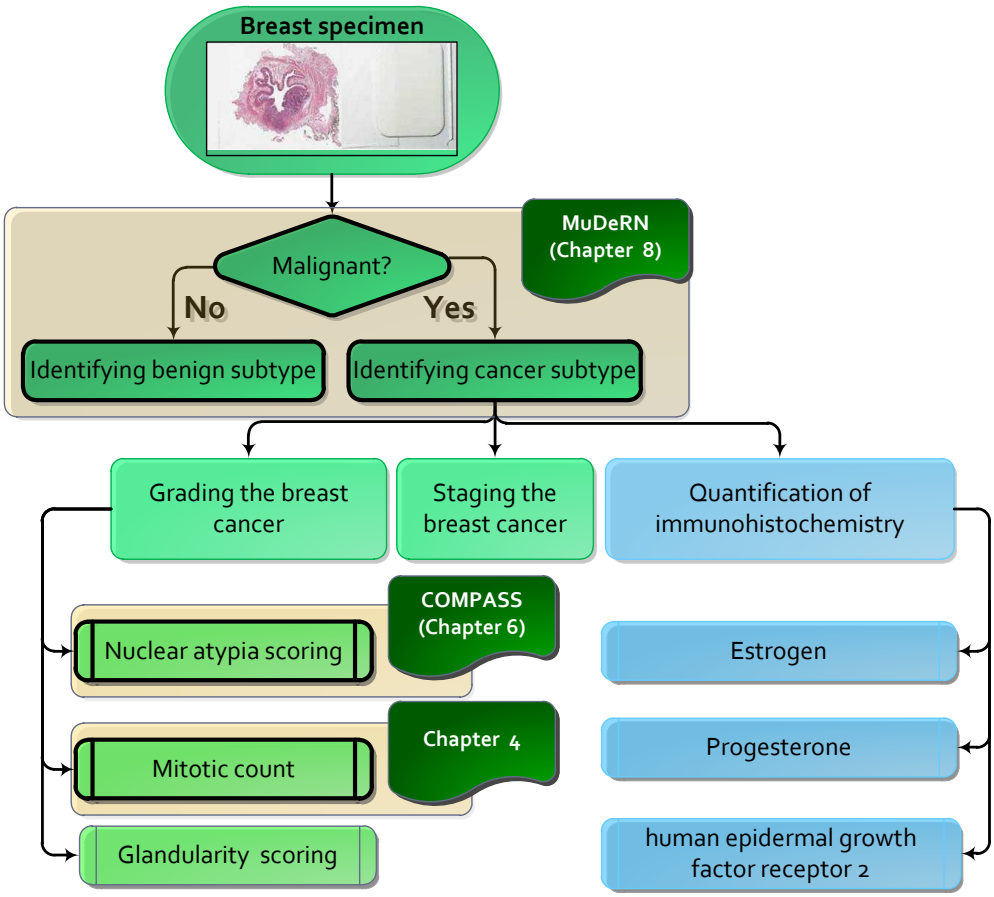


Figure 1- Pathologists' workflow while interpreting a breast biopsy; tasks shown in green boxes are done based on Hematoxylin-Eosin stained breast histopathological digital slides while tasks shown in blue boxes require immunohistochemical staining. The focus of this thesis is on computer-assisted analysis of Hematoxylin-Eosin stained breast images. The emphasis of different chapters is also indicated. As shown, the studies

involved in this thesis covered five components of pathologists' workflow for interpreting Hematoxylin-Eosin stained breast slides.

This thesis aimed at addressing some of the main deficiencies of the existing literature in computer-assisted analysis of breast histopathological digital slides. Based on thoroughly reviewing the existing studies, it was found that previous studies showed high agreement between computer-assisted methods and the expert scoring in the quantification of three clinically important estrogen receptor, progesterone receptor, and human epidermal growth factor 2. Therefore, I narrowed down the focus of this thesis to computer-assisted analysis of Hematoxylin-Eosin stained digital breast slides. As shown in figure 1, the studies done in this thesis cover all components of pathologists' workflow for the interpretation of Hematoxylin-Eosin stained breast slides except glandularity scoring and staging the breast cancer.

The breast cancer grade is the average of scores assigned by a pathologist to three contributing components, namely, glandularity, the level of nuclear atypia, and the miotic count [4]. Previous studies indicated that scoring tubule formation achieved substantial to perfect inter-pathologist agreement [9-16] and the agreement level for this component is the strongest among the three contributing factors in breast cancer grading. Also, although only a few studies focused on automatic segmentation of tubules in the tumour area, the agreement with the expert segmentation was quite high [25, 26]. Therefore, among the three factors evaluated for breast cancer grading, this thesis focused on the mitoses count and the nuclear atypia score. Among the three contributing components, most of the previous studies were devoted to the automatic detection of mitotic figures [27-35], and the state-of-the-art methods achieved high accuracy in detection of mitotic figures [36]. Hence, this thesis focused on the

mitotic counting task from the medical image perception angle and aimed at determining the computer-extracted features related to the disagreement among pathologists in recognition of mitotic figures (chapters 3 and 4). The framework presented here for linking quantitative image processing features with inter-pathologists variations in recognition of mitotic figure could be extended to other tasks in breast pathology. There were previous studies that focused on nuclear atypia scoring, however, most of them aimed at classifying each individual cell into different atypia grades [37, 38] or only addressed the segmentation of epithelial nuclei [39-45] and did not extend their methodology to provide the nuclear atypia score based on the features of segmented nuclei. Given the fact that the inter-pathologist variation in nuclear atypia scoring was the highest among three components of breast cancer grading system, COMPASS, a method for reproducible nuclear atypia scoring was proposed in this thesis (chapters 5 and 6).

As shown in figure 1, pathologists classify the slides into benign or malignant and also determine the exact subtype of the lesion. Many previous studies focused on the benign/malignant classification of breast histopathological images [37, 38, 46-49], however, less attention was paid to determining the cancer/benign subtypes. Also, most of the previous studies worked on classifying a region of interest (ROI) in an image while in the clinical practice pathologists assess different ROIs in a slide at single or multiple zoom levels and make the final diagnosis for a patient rather than for each ROI. In this thesis, MuDeRN was proposed to address these two shortcomings for computer-aided diagnosis; MuDeRN aimed at identifying the cancer subtypes and differentiating benign subtypes, and it involves a framework for patient-level diagnoses which incorporates information from different ROIs and different magnification levels.

Staging breast cancer involves the segmentation of the tumour area and the assessment of metastasis. The state-art-of-the art methods for segmenting the tumour area achieved a level of agreement with expert segmentation [50, 51]. As part of staging a breast cancer, pathologists evaluates whether the tumour had spread to the lymph nodes, chest wall, or skin. Therefore, assessing tissues other than breast might be required. Therefore, this thesis did not cover the application of computer-assisted methods in staging of breast cancer.

In summary, MuDeRN could assist pathologists in the first step of their workflow for interpreting Hematoxylin-Eosin stained breast slides where the patients are classified as benign or cancer, and the subtype of cancer or benign mass is identified. In the second phase of pathologist' workflow, in case of presence of breast cancer, COMPASS could assist in achieving a reproducible nuclear atypia score. The study presented in chapter 4 could be helpful in reducing the discrepancies among the pathologists for recognition of mitotic figures. In the rest of this section, the findings of this thesis are discussed in the context of existing literature. In particular, the objectives, which were listed earlier in the introduction chapter (chapter 1) and met in this thesis, are discussed in each sub-section.

Reviewing the image processing techniques which worked successfully in analysing Hematoxylin-Eosin stained digital slides and their main showed that:

- The reviewed automated methods for stain assessment showed high agreement with the expert scoring in the quantification of four clinically important immunohistochemical stains (i.e., estrogen receptor,

progesterone receptor, Ki‑67, and human epidermal growth factor).

- The automatic methods for segmentation of tumour area achieved a high agreement with the expert's segmentation.
- Many studies focused on benign/malignant classification but there is lack of studies aimed at **identification of the cancer subtypes or differentiation of benign subtypes from each other**.
- Among three components of Scarff‑Bloom‑Richardson grading system, the mitotic figure detection got the most attention in the previous studies. Although only a few studies focused on tubule formation, the agreement with the expert segmentation was quite high. There were studies that focused on nuclear atypia scoring, however, most of them were in cell-level, i.e., classifying each individual cell. Given the fact that the agreement on the nuclear atypia score was the weakest among three components of Scarff‑Bloom‑Richardson grading system, further developments on **reproducible nuclear atypia scoring** are required.
- I identified a lack of studies which **link quantitative image processing features with disagreement among pathologists.**
- Pathologists evaluate different regions of interest (ROI) in a histopathological slide and assess the slide at different zoom levels. Most of the reviewed papers worked on ROIs and there is still room for improving the framework for **patient-**

level diagnoses which incorporate information from different ROIs and different magnification levels.

- There is also a lack of studies testing the proposed algorithms in a **prospective** manner, where the computer-assisted algorithm provides feedback to the pathologist and then the pathologist makes the final decision taking into account this feedback.
- There is also lack of **personalized computer-assisted analysis tool**, which consider each pathologist's unique error making patterns.

The literature review paper presented in chapter 2 was limited to the added benefits of computer-assisted analysis in breast pathology. Although some studies conducted in other pathology subspecialties or animal tissues could be extended to breast pathology, I only focused on breast pathology in the review. Previously, in [19] and [20], broader reviews on the current status of digital pathology in general, its benefits and potential challenges were carried out. Gurcan et al. (2009) reviewed the computer assisted analysis in histopathology in general [21] and Irshad et al (2014) reviewed computer aided segmentation of nuclei in histopathology in general [52]. Finally, our study can complement the review done by Veta et. al (2014) [53] on breast cancer histopathology image analysis. Our review covered more recent studies and included more studies in breast slide classification, while [53] focused on studies aimed at segmentation of elements within the breast digital slides and less emphasis was placed on breast histopathological image classification.

9-1-1- Feasibility of relating quantitative features with discrepancies among pathologists

The first study of this thesis aimed to **explore the feasibility of relating quantitative image processing features with disagreement among pathologists**. The study focused on the mitotic recognition task[1].

In summary, the results of this study showed that quantitative image-processing features can capture the differences between appearances of challenging mitotic figures and the easily identifiable mitoses, and also between miscounted objects (false positives) and mitotic figures. It was also found that the discriminative power of different colour spaces for distinguishing miscounted objects from mitotic figures varies and two perceptually motivated colour spaces (LMS and XYZ) capture the difference better than other colour spaces.

Examination of the shape-based features showed that challenging mitotic figures are rounder and smaller than other mitotic figures. This could be due to the fact that the recognition of mitoses specifications, such as hairy outline, is more difficult for smaller objects. Previous studies suggested that providing precise constraints in mitoses counting protocol can lead to standardization of the mitotic index and eventually decrease discrepancies in grading among the pathologists [54-56]. Our results showed that the miscounted non-mitotic objects are rounder than the true mitoses, thus suggesting that considering a quantitative roundness measure (such as compactness) as a constraint in mitoses counting may decrease disagreement among pathologists for counting mitotic figures.

The circularity measure showed that the miscounted objects were rounder than mitotic figures. As in the early metaphase, the circularity measure is high, it could be hypothesized that most of miscounted objects were

[1] "Determining image processing features describing the appearance of challenging mitotic figures and miscounted non-mitotic objects", Journal of Pathology Informatics, 2017

mistaken by cells in their early metaphase. The results suggested that by restricting counted circular objects to those with very evident features of mitoses in early metaphase (round with clotted visible hairy extensions of nuclear material [54-56]), the number of miscounted non-mitoses can be reduced.

The challenging mitoses and the easily identifiable mitoses exhibit an almost similar level of hyperchromicity, as most of the intensity-based features did not differ significantly between them, however texture-based features significantly differed between these two groups. On the other hand, for comparing non-mitoses and easily identifiable mitoses, intensity-based features were the most discriminative ones. So, constraints on cells based on the level of hyperchromicity can reduce the number of miscounted non-mitoses objects.

Among the scene descriptors, the results suggested that the higher density of chromatin and smaller size of surrounding nuclei led to higher magnitude of difficulty in the identification of mitotic figures. This could be due to the fact that the above-mentioned factors led to a higher probability of finding similar objects in the slide and made the scene more complex for pathologists. The results also showed that the miscounted non-mitoses were often annotated in images with smaller cell size. These findings could be used to notify the pathologists that counting in a particular slide with the above-mentioned features leads to error so that they pay extra attention.

9-1-2- Computer-assisted nuclear atypia grading

The second original study presented in this thesis was aimed at developing a tool for **reproducible nuclear atypia grading.** A personalized tool,

called COMPASS (COMputer-assisted analysis combined with Pathologist's ASSessment), was proposed and its performance was evaluated.

In summary, COMPASS relies on two modules processing two different sets of features. The first set includes the scores given by the pathologists to six cytological characteristics related to nuclear atypia, while the second set comprises textural computer-extracted features. COMPASS' performance was evaluated using the Mitosis-Atypia database, which includes 600 images with expert-consensus derived reference nuclear atypia scores. Half of the images were produced by Aperio Scanscope XT scanner and half of them were acquired by Hamamatsu Nanozoomer 2.0-HT scanner. COMPASS was retrospectively personalized for three junior pathologists who gave scores to six atypia-related criteria for images in the database.

The magnitude of COMPASS's agreement (for all junior pathologists and both scanners) with expert-consensus reference nuclear grade was comparable to that of senior pathologists while assessing the same dataset. Owing to stain normalization [57] in the pre-processing step, as expected COMPASS' performance did not differ between two scanners. Also, it was shown that COMPASS could supplement the senior pathologist's performance, as only a few images were mis-graded by both senior pathologists and COMPASS, and for most of the images at least one of them provided the correct grade.

COMPASS is a hybrid method in a sense that it involves a coarse nuclei segmentation as well as textural analysis. Previous automatic nuclear grading methods either extracted the features from the segmented nuclei (segmentation-based methods [37, 58]) or from the entire tissue (textural

analysis [58]). However, COMPASS uses the initial segmentation for restricting the analysis to diagnostically relevant ROIs.

Previously, in the field breast radiology, personalized computer-aided analysis tools were developed[59]. COMPASS is the first personalized computer-assisted analysis tool in breast pathology. Dunne and Going (2001) showed that some pathologists are prone to under-grading while others systematically over-grade the cases [60]. COMPASS is a personalized model, trained based on scores given by each individual; hence it can reduce systematic over-grading and under-grading for each individual.

The results showed that COMPASS outperformed a recent fully-automatic algorithm based on textural features [58] tested on the same dataset for the three nuclear grades. This was achieved by using cytological scores given by junior pathologists, and also limiting the analysis to the ROIs with high nuclear density.

Also, the added-benefits of computer-extracted features to the cytological features assessed by the pathologist were evaluated. To do so, the performance of COMPASS was compared with two baseline approaches that only consider the scores given by the junior pathologists to the cytological features. The grade of the first approach was simply the total cumulative score given to all criteria while the second approach build a non-linear model to link six scores given by the junior pathologists to a reference nuclear atypia grade. The fact that COMPASS achieved the highest accuracy (compared to two baseline approaches) suggested that textural features provide complementary information to cytological scores.

Although the added benefit of textural features was observed both for distinguishing grade I from higher grade tumours (grade II and III), and also grade III from lower grade tumours (grade I and II), the magnitude of this added benefit (increment in the performance of the junior pathologists) was higher for distinguishing grade I tumours. This suggests that the included cytological features have higher discriminative power for distinguishing grade III tumours from borderline ones (grade II) while more information about the characteristics of image is required for differentiating grade I from grade II.

In recent years use of fine needle aspiration (FNA) for the pre-operative diagnosis of breast cancer increased. In the National Cancer Institute of USA workshop on FNA, it was recommended that breast cancer grade should be stated in the pathology report for the management of a given patient based on prognostic information [61, 62]. Due to the limited sample size in FNA, histological grading is not completely feasible based on FNA and previous studies emphasized that the cytological grade on FNA should correspond to the histologic grade [61, 62]. Various cytological grading systems based on different cytological features have been proposed [63-68] but their agreement rate with the histologic grade ranged from 66.6% to 78.57%, and none of them reached a substantial agreement (that is, agreement rate above 80% which is equivalent to Cohen's kappa above 0.61) [69-71]. Our results suggest that computer-extracted features can supplement cytological features for predicting histologic nuclear grade and hence can be used for grading FNA breast specimens.

9-1-3- Computer-assisted diagnosis

The last study of this thesis is presented in chapter 8[2] and aimed at addressing the lack of studies on **identification of the cancer subtypes or differentiation of benign subtypes**. It also aimed at proposing a framework for providing **patient-level diagnoses** through aggregating information from different ROIs and different magnification levels.

MuDeRN (MUlti-category classification of breast histopathological image using DEep Residual Networks) has been proposed and evaluated in the study. MuDeRN comprised of two stages; in its first stage, images at each magnification factor were classified as benign or malignant while in its second stage, the images classified as malignant were subdivided into four cancer subcategories, i.e., ductal carcinoma, lobular carcinoma, mucinous carcinoma, or papillary carcinoma, and those labelled as benign were classified into four subtypes, i.e., adenosis, fibroadenoma, phyllodes tumour, or tubular adenoma. Finally, the diagnosis for each patient was made by combining outputs of images in different magnification factors using a meta-decision tree. MuDeRN was tested on a publicly available dataset, called BreakHis which includes 7786 images in four magnification factors from 81 patients.

In the binary classification task, MuDeRN achieved higher accuracy for the malignant group. That could be due the fact that number of malignant patients were approximately 2.5 times higher than that of benign cases, and hence more data were provided for MuDeRN to learn the variant appearances of malignant cases.

---

[2] "MuDeRN: Multi-category Classification of Breast Histopathological Image Using Deep Residual Networks," Journal of Artificial Intelligence in Medicine, 2018.

In the eight-category classification task, on average the accuracies were higher for malignant subtypes across all magnification factors. Among different subtypes, invasive ductal carcinoma achieved the highest correct classification rate (CCR), and this may be owing to the fact that this category has the highest number of patients in BreakHis, and hence MuDeRN learnt its various characteristics well. The lowest and the second lowest CCRs correspond to the adenosis and phyllodes tumour subtypes, for which BreakHis included only four and three patients, respectively. No patients with mixed adenosis and phyllodes tumour were included in the database. Invasive lobular carcinoma subtypes had also only four patients, however, the differences between included lobular carcinoma cases and other malignant cases were very clear. No patients with mixed ductal and lobular carcinoma were included. Another possible reason for explaining the low CCR of adenosis could be fact that the adenosis has four sub-categories (i.e. sclerosing, tubular, apocrine, and microglandular) and thus a large number of cases is required so that the network can learn the features of the different variants of the disease.

Similar to [72] where hand-crafted features (manually designed features such as Haralick texture features, intensity based-features, etc used in traditional machine learning frameworks) were used for binary-classification of BreakHis, MuDeRN achieved the highest CCR in the lowest magnification factor for the binary-classification task, while for the eight-category classification task, the highest CCR was obtained for x200 magnification factor. This suggests that the lowest magnification factor is more informative for differentiating malignant from benign, however, further information, especially cytological features, from higher magnification factors is essential for the accurate determination of lesion subtypes. This is similar to the pathologists' behaviour in the clinical

practice where they start exploring the image in the lowest available magnification level and then select the final diagnosis from a list of potential diagnosis after zooming in to a few areas at a higher magnification.

Cserni et al. (2006) showed that there is only moderate agreement among pathologists for classifying benign lesions into fibroadenoma, phyllodes tumour, and other benign subtypes other than these two [18], and Lawton et al. (2104) indicated that only in 53% of cases pathologists reached agreement for distinguishing fibroadenomas from phyllodes tumours [73]. As stated, MuDeRN is able to differentiate these two subtypes and hence it can provide a second opinion and assist the pathologist. The other benign subtypes which are handled by MuDeRN are adenosis and tubular adenoma. They are both due to hyperplasia of lobuli and exhibit several similar features [74], therefore MuDeRN can assist pathologists for differentiating these two subtypes by providing an objective diagnosis. Previous studies also showed that diagnostic discrepancies exist for categorising the subtypes of invasive breast carcinoma, and the magnitude of disagreement ranged from 38.5% for papillary carcinoma to 8% for ductal carcinoma [75]. Hence, a second opinion from MuDeRN could be helpful for pathologists.

MuDeRN achieved the highest CCR compared to the previous automatic methods tested on the BreakHis database for benign/malignant classification [72, 76-78]. The second best algorithm [78] also used a deep neural network called GoogLeNet [79], however, the residual neural networks used in MuDeRN are deeper and also include shortcuts from input to the output of the stacked layers [80]. Only one previous algorithm [78] addressed the eight-category classification task of BreakHis database

using GoogLeNet [79]. The results showed that MuDeRN outperformed the previous study [78]. This better performance was achieved through applying stain normalization as a pre-processing step, using a deeper network and a two-stage classifier and utilizing a meta-decision tree [81] for making the patient-level diagnosis by aggregating outputs of multiple deep residual networks analysing images at different magnification factors.

9-2- Limitations

In this thesis, I used the Mitosis-Atypia[3] and BreakHis[4] databases, which are both publicly available. Although using publicly available databases makes our study comparable to the previous studies in the literature and facilitates replicating our work in future studies, it should be noted that the availability of certain data types was limited, and this may have hindered certain aspects of the three studies described here.

I used the Mitosis-Atypia database in the first and second original research studies. The number of included patients in Mitosis-Atypia database is low (only eleven patients). Also, the prevalence of different nuclear grades in the dataset was different from real clinical practice and images in grade II considerably outnumber those in grades I and III. In addition, in both mitotic recognition and COMPASS studies intra-pathologist variability in scoring was not investigated as the data for multiple assessments from a single pathologist was not available in the Mitosis-Atypia database.

The BreakHis database, which was used for evaluating MuDeRN, only contained benign patients without atypia and invasive carcinoma, that is,

[3] https://mitos-atypia-14.grand-challenge.org/
[4] https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

no in situ breast cancer (ductal carcinoma in situ and lobular carcinoma in situ) or benign lesions with atypia (atypical ductal hyperplasia and typical lobular hyperplasia) were included. Previous studies showed that the degree of diagnostic discrepancies among pathologists are higher for in situ breast cancer and benign lesions with atypia compared to benign lesions without atypia and invasive carcinoma, as they exhibit borderline diagnostic features [5, 7, 8]. Also, the numbers of patients in three subtypes (adenosis, invasive lobular carcinoma, and phyllodes tumours) were smaller than five, and this may limit the generalizability of results obtained from MuDeRN, particularly for these categories.

In spite of above-mentioned limitations of the utilized data, using the publicly available datasets enabled us to compare our results with those of others and also made it possible for other researchers to further extend our work.

In both Mitosis-Atypia 2014 and BreakHis databases, the initial ROIs were manually selected by expert breast pathologists, which makes both COMPASS and MuDeRN semi-automatic. This limits the capacity of COMPASS and MuDeRN for being used in a completely automated clinical practice.

Also, the images were stain normalized before feeding to COMPASS and MuDeRN. Various stain normalization methods are available [82], and selecting a different stain normalization could affect the appearance of stain normalized images and therefore change the obtained results. All of the current stain normalization method have their own advantages and disadvantages[82] and further research in this filed is ongoing [83-85].

The studies presented in this thesis required use of high performance computing facilities and cannot be finished in a reasonable time using an ordinary desktop computer. Extracting various image processing features from different colour spaces in the mitotic figure recognition study, training and testing COMPASS using leave-one-out cross validation and setting its hyper-parameters using Bayesian optimization, and finally, evaluation of MuDeRN using 27-fold-cross validation were computationally expensive tasks which were accomplished by means of high performance computing service at the University of Sydney. However, this is not a major limitation for the studies, as nowadays high performance computing services are available with reasonable costs. More importantly, extracting rules, training COMPASS, or MuDeRN should be done only once, as after that they can be used easily without any extensive computation.

Moreover, studies in this thesis focused on grading and making diagnoses based on Hematoxylin-Eosin stained images. Another important task done by the pathologists based on Hematoxylin-Eosin stained slides is staging the breast cancer, which involves determining size of tumour, assessing whether the cancer spread to nearby lymph nodes and other parts of body, but this was not examined in this thesis.

Finally, one major limitation of using the tools presented in this thesis and using whole slide imaging in general, is the extra time required for scanning (i.e. digitizing the slides). Although current commercial slide scanners are quite fast compared to their previous generation and the current speed is about 4 minutes per 2'x2' slide at the highest magnification level, further improvement is still required in terms of their speed. Also, the current auto-focusing algorithms can be improved. A set of slides can be loaded to most of the new scanners at the same time and the slides can

be scanned overnight. Moreover, the images can be down sampled digitally, so usually scanning at the highest magnification level will be sufficient for algorithm like MuDeRN, which may require images from multiple magnification factors.

9-3- Future directions

As stated in the previous section, lack of data, particularly for certain categories, may limit generalisability of findings presented in this thesis. One possible avenue for future work will be including more patients in the databases. More specifically, COMPASS will benefit from including more patients with nuclear atypia grade of I and III and MuDeRN will benefit from increasing number of patients in adenosis, invasive lobular carcinoma, and phyllodes tumours subtypes as wells as adding carcinoma in situ and benign with atypia subtypes. The mitotic recognition study will benefit from having a dataset whose ground truth was obtained based on both Ki-67 label and Hematoxylin-Eosin stained images data, as a combination of these images leads to more reliable labelling of mitotic figures. Specifically for those objects labelled as “probably a mitosis” by the majority of senior pathologists, the Ki-67 label could be beneficial in establishing the ground truth. Furthermore, in the Mitosis-Atypia database, senior pathologists annotated the mitotic figures on images. One potential avenue for future work could be asking pathology trainees or less experienced pathologists to annotate the images and repeat the framework proposed in chapter 4 for them as the quantitative features describing the appearances of challenging mitoses for less experienced observers could be different from those features extracted for the experienced ones.

A future step for improving COMPASS and MuDeRN could be adding a pre-processing block for analysing the entire whole slide image and automatically selecting the tumour areas which will be then fed into COMPASS and MuDeRN. In a recent study, GoogLeNet was used for segmenting the tumour area and an accuracy about 98% was achieved [51], hence adding this pre-processing step is feasible. In addition, three studies presented here only focused on images from a particular magnification levels, which were available in the utilized databases. A nice future work will be studying other magnification levels and investigating the performance of computerized tools presented here in other magnification levels.

Dealing with the borderline cases presenting categories of various diseases are crucially important. Especially as these tools will ultimately play the role of "surrogate second reader", investigating the performances of tools presented in this study and future computerized tool on the borderline cases is really important. Such as investigation will help us to optimize how computerized tools should be implemented in clinical practice of pathologists.

Also, as discussed in the Introduction, pathologists make the final diagnosis based on both Hematoxylin-Eosin stained and immunohistochemical images. As shown Chapter 2, the current computerized methods perform really well in quantification of immunohistochemical results. These tools can be incorporated with the computerized histological assessment tools to provide the final diagnosis for the patient.

Another potential future work could be investigating the pathologists' performances in recognition of mitotic figures after a training session

where it was discussed the obtained rules which described the appearances of easily identifiable mitoses, challenging mitoses (false negative), and miscounted non-mitoses. Previously Paradiso et al. showed that extracting the methodological skills required to enhance performance from a quality control study [16] and reviewing these skills in a training course can increase the pathologists' performance in a short-term [86]. Also, the rules can be used to develop a training software for breast pathologists which retrieves the mitotic figures and non-mitoses objects with different levels of difficulty at different stages of their learning curves. Another future step could be extending the framework presented in chapter 4 for recognition of mitotic figure to other tasks in breast pathology. Exploring the association between quantitative image processing features with disagreement among pathologists in different pathology tasks could help in training pathology registrars and improving the understanding about underlying reasons for diagnostic errors.

Moreover, nowadays by using whole slide imaging, it is possible to record pathologists' navigation pattern. In future, we can relate the image features to pathologists' navigation patterns and diagnostic errors. From the perceptual studies done on radiologists, it is well-known that radiologists' first impression of the image, often referred as the gist response, could predict the ultimate diagnoses of the case [87, 88]. Considering the size virtual slides and the hierarchical search pattern (from low magnification factor to high magnification factor), understanding the importance of peripheral processing of pathological images might help in understanding about underlying reasons for diagnostic errors.

Another potential avenue for future work is testing COMPASS and MuDeRN in a prospective scenario. In this thesis COMPASS and

MuDeRN were tested retrospectively where it was assumed that the pathologists would accept the decision of COMPASS and MuDeRN. However, in a more realistic set-up, they will provide a feedback to pathologists who would make the ultimate decision. In medical imaging, both in radiology and pathology, the human interaction with computerized programs is poorly understood. One possible direction for the future projects can be testing different conditions and determining the best implementation of using computerized tools in practice. At least the following conditions should be explored:

1. Condition A: pathologists will read cases as a blinded second reader to the computerized tool. They will be aware that their decision will be compared with the computerized tool's decision afterwards.

2. Condition B: pathologists will read the cases as an informed second reader to the computerized tool readings.

3. Condition C: pathologists will be asked read the cases as a third reader, having both the first pathologist's decision and the computerized tool decision.

The added benefit of these conditions should be compared with pathologists while reading on their own.

In chapter 6, it was shown that COMPASS, if it had been adopted by the junior pathologists, could be used as a second opinion for senior pathologists as it complements the senior pathologist's performance to some extent. A potential future step could be performing similar study for MuDeRN to investigate its usefulness as a second opinion for senior pathologists. Studying the performances of both MuDeRN and COMPASS in double reading scenarios when the first reader is a less

experienced pathologist could be also done in the future. This will be very useful in practice as providing an independent second opinion could be particularly helpful when the slides were evaluated by general or less experienced pathologists [1, 5, 7].

In the future, COMPASS can be extended by adding a module which relates the nuclear grade outputted by COMPASS (or the feature sets used by COMPASS) to the patient survival. As stated earlier, nuclear grading is utilized as a prognostic factor in patient management. Therefore, outputting an index showing patient survival will strengthen COMPASS.

In practice, both benign and malignant entities might be present on one slide. In the current study, the utilized databases did not include such cases. One potential future work will be examining the performance of MuDeRN on a database containing cases with benign and malignant entities. Moreover, 10 to 50 whole slide images might be available for a patient. Current version of MuDeRN combines image-level diagnoses for four magnification factors using a meta-decision tree in order to make the final diagnosis for a patient. A potential extension of this work can be extending MuDeRN so that it can handle multiple slides per case.

Another possible future work could be investigating MuDeRN when tested on tissue where none of these eight categories are encountered. It can be hypothesize that if the new benign sample from a category other than the ones included in the BreakHis database share some features with other benign categories in terms of cells' approaches, tissue texture, etc and these characteristics have been captured by the trained network, then MuDeRN might be able to categorize it as a benign (although this new

disease category was not included in the BreakHis). However, in future this hypothesis should be investigated.

Finally, specific microscopic details (eg, mitotic figures) were reported to be difficult to identify [87] as under the microscope some 3-dimensional information can be captured by optically zooming in and out. As an avenue for future studies, the performance of tools with whole slide images that are acquired with an additional z-axis in their entirety can be explored to take advantage of 3-dimensional information.

9-4- Summary

In summary, the findings presented in this thesis contribute to the existing literature by:

i. Showing that quantitative image processing features are associated with disagreement among pathologists in recognition of mitotic figures, the results suggested that precise constraints based on quantitative features can be placed in mitotic figure counting protocol to decrease discrepancies in mitotic figure recognition among the pathologists.
ii. Indicating that holistic features, which explain the characteristics of the entire image rather than just the mitotic figures themselves, are related to disagreement among pathologists in recognition of mitotic figures. For example, smaller size of surrounding nuclei led to higher degree of difficulty in the identification of mitotic figures.
iii. Proposing COMPASS, an individualized tool for reproducible nuclear atypia grading, and indicating that:

- COMPASS outperformed the previous automatic algorithms for nuclear atypia grading tested on the same database;
- COMPASS has a performance comparable to senior breast pathologists in nuclear atypia grading.

iv. Demonstrating the value of combining pathologist's assessment on cytological features and computer-extracted textural features for reproducible nuclear atypia scoring.

v. Confirming the value of a hybrid approach for nuclear atypia grading which involves an initial coarse nuclei detection followed by textural analysis in areas with high density of epithelial cells.

vi. Illustrating that included cytological features to some extent contain information for predicting histologic nuclear atypia grade and computer-extracted features can complement this information.

vii. Proposing MuDeRN, a framework based on deep residual networks for multi-category classification, and indicating that MuDeRN outperformed previous automatic classification algorithms tested on the same database.

viii. Proposing a framework based on meta-decision tree for providing patient-level diagnoses through aggregating the decisions made at the different magnification levels.

References

[1] D. L. Weaver, R. D. Rosenberg, W. E. Barlow, L. Ichikawa, P. A. Carney, K. Kerlikowske, et al., "Pathologic findings from the Breast Cancer Surveillance Consortium: population-based outcomes in women undergoing biopsy after screening mammography," Cancer, vol. 106, pp. 732-42, Feb 15 2006.

[2] F. A. Tavassoli, Pathology of the Breast: McGraw Hill Professional, 1999.

[3] P. P. Rosen, Rosen's breast pathology: Lippincott Williams & Wilkins, 2001.

[4] L. P. Howell, R. Gandour-Edwards, and D. O'Sullivan, "Application of the Scarff-Bloom-Richardson tumor grading system to fine-needle aspirates of the breast," American journal of clinical pathology, vol. 101, pp. 262-265, 1994.

[5] K. H. Allison, L. M. Reisch, P. A. Carney, D. L. Weaver, S. J. Schnitt, F. P. O'malley, et al., "Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel," Histopathology, vol. 65, pp. 240-251, 2014.

[6] I. O. Ellis, D. Coleman, C. Wells, S. Kodikara, E. M. Paish, S. Moss, et al., "Impact of a national external quality assessment scheme for breast pathology in the UK," J Clin Pathol, vol. 59, pp. 138-45, Feb 2006.

[7] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. Tosteson, et al., "Diagnostic concordance among pathologists interpreting breast biopsy specimens," Jama, vol. 313, pp. 1122-1132, 2015.

[8] L. Khazai, L. P. Middleton, N. Goktepe, B. T. Liu, and A. A. Sahin, "Breast pathology second review identifies clinically significant discrepancies in over 10% of patients," Journal of surgical oncology, vol. 111, pp. 192-197, 2015.

[9] G. S. Delides, G. Garas, G. Georgouli, D. Jiortziotis, J. Lecca, T. Liva, et al., "Intralaboratory variations in the grading of breast carcinoma," Arch Pathol Lab Med, vol. 106, pp. 126-8, Mar 1982.

[10] J. M. Harvey, N. H. de Klerk, and G. F. Sterrett, "Histological grading in breast cancer: interobserver agreement, and relation to other prognostic factors including ploidy," Pathology, vol. 24, pp. 63-8, Apr 1992.

[11] H. F. Frierson, Jr., R. A. Wolber, K. W. Berean, D. W. Franquemont, M. J. Gaffey, J. C. Boyd, et al., "Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma," Am J Clin Pathol, vol. 103, pp. 195-8, Feb 1995.

[12] P. Robbins, S. Pinder, N. de Klerk, H. Dawkins, J. Harvey, G. Sterrett, et al., "Histological grading of breast carcinomas: a study of interobserver agreement," Hum Pathol, vol. 26, pp. 873-9, Aug 1995.

[13] S. R. Lakhani, J. Jacquemier, J. P. Sloane, B. A. Gusterson, T. J. Anderson, M. J. van de Vijver, et al., "Multifactorial analysis of differences between sporadic breast cancers and cancers involving BRCA1 and BRCA2 mutations," J Natl Cancer Inst, vol. 90, pp. 1138-45, Aug 05 1998.

[14] P. Boiesen, P. O. Bendahl, L. Anagnostaki, H. Domanski, E. Holm, I. Idvall, et al., "Histologic grading in breast cancer--reproducibility between seven pathologic departments. South Sweden Breast Cancer Group," Acta Oncol, vol. 39, pp. 41-5, 2000.

[15] J. S. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, et al., "Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index," Mod Pathol, vol. 18, pp. 1067-78, Aug 2005.

[16] I. N. f. Q. A. o. T. B. Group, "Quality control for histological grading in breast cancer: an Italian experience," Pathologica, vol. 97, p. 1, 2005.

[17] J. M. Bueno-de-Mesquita, D. Nuyten, J. Wesseling, H. van Tinteren, S. Linn, and M. van De Vijver, "The impact of inter-observer variation in pathological assessment of node-negative breast cancer on clinical risk assessment and patient selection for adjuvant systemic treatment," Annals of oncology, vol. 21, pp. 40-47, 2009.

[18] G. Cserni, Z. Orosz, J. Kulka, Z. Sápi, E. Kálmán, and R. Bori, "Divergences in diagnosing nodular breast lesions of noncarcinomatous nature," Pathology & Oncology Research, vol. 12, pp. 216-221, 2006.

[19] L. Pantanowitz, P. N. Valenstein, A. J. Evans, K. J. Kaplan, J. D. Pfeifer, D. C. Wilbur, et al., "Review of the current state of whole slide imaging in pathology," Journal of pathology informatics, vol. 2, p. 36, 2011.

[20] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman, "Digital imaging in pathology: whole-slide imaging and beyond," Annual Review of Pathology: Mechanisms of Disease, vol. 8, pp. 331-359, 2013.

[21] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," IEEE reviews in biomedical engineering, vol. 2, pp. 147-171, 2009.

[22] M. Lundin, J. Lundin, H. Helin, and J. Isola, "A digital atlas of breast histopathology: an application of web based virtual microscopy," Journal of Clinical Pathology, vol. 57, pp. 1288-1291, 2004.

[23] M. Khushi, G. Edwards, D. A. de Marcos, J. E. Carpenter, J. D. Graham, and C. L. Clarke, "Open source tools for management and archiving of digital microscopy data to allow integration with patient pathology and treatment information," Diagnostic Pathology, vol. 8, 2013.

[24] A. Bondi, S. Lega, P. Crucitti, P. Pierotti, R. Rapezzi, P. Sassoli De Bianchi, et al., "Quality assurance and automation," Cytopathology, vol. 23, p. 39, 2012.

[25] K. Nguyen, M. Barnes, C. Srinivas, and C. Chefd'hotel, "Automatic glandular and tubule region segmentation in histological grading of breast cancer," in Medical Imaging 2015: Digital Pathology, 2015, p. 94200G.

[26] P. Maqlin, R. Thamburaj, J. J. Mammen, and A. K. Nagar, "Automatic detection of tubules in breast histopathological

images," in Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012), 2013, pp. 311-321.

[27] H. Irshad, L. Roux, and D. Racoceanu, "Multi-channels statistical and morphological features based mitosis detection in breast cancer histopathology," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2013, pp. 6091-6094.

[28] M. Veta, P. J. Van Diest, and J. P. W. Pluim, "Detecting mitotic figures in breast cancer histopathology images," in Progress in Biomedical Optics and Imaging - Proceedings of SPIE, 2013.

[29] V. Roullier, O. Lézoray, V. T. Ta, and A. Elmoataz, "Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization," Computerized Medical Imaging and Graphics, vol. 35, pp. 603-615, 2011.

[30] V. Roullier, O. Lézoray, V. T. Ta, and A. Elmoataz, "Mitosis extraction in breast-cancer histopathological whole slide images," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) vol. 6453 LNCS, ed, 2010, pp. 539-548.

[31] V. Roullier, V. T. Ta, O. Lézoray, and A. Elmoataz, "Graph-based multi-resolution segmentation of histological whole slide images," in 2010 7th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2010 - Proceedings, 2010, pp. 153-156.

[32] A. M. Khan, H. Eldaly, and N. M. Rajpoot, "A gamma-gaussian mixture model for detection of mitotic cells in breast cancer histopathology images," J Pathol Inform, vol. 4, p. 11, 2013.

[33] C. D. Malon and E. Cosatto, "Classification of mitotic figures with convolutional neural networks and seeded blob features," Journal of pathology informatics, vol. 4, 2013.

[34] H. Irshad, "Automated mitosis detection in histopathology using morphological and multi-channel statistics features," Journal of pathology informatics, vol. 4, 2013.

[35] H. Irshad, S. Jalali, L. Roux, D. Racoceanu, L. J. Hwee, G. Le Naour, et al., "Automated mitosis detection using texture, SIFT features and HMAX biologically inspired approach," Journal of Pathology informatics, vol. 4, 2013.

[36] K. Paeng, S. Hwang, S. Park, and M. Kim, "A unified framework for tumor proliferation score prediction in breast histopathology," in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, ed: Springer, 2017, pp. 231-239.

[37] E. Cosatto, M. Miller, H. P. Graf, and J. S. Meyer, "Grading nuclear pleomorphism on histological micrographs," in Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, 2008, pp. 1-4.

[38] B. Weyn, G. van de Wouwer, A. van Daele, P. Scheunders, D. van Dyck, E. van Marck, et al., "Automated breast tumor diagnosis and grading based on wavelet chromatin texture description," Cytometry, vol. 33, pp. 32-40, 1998.

[39] C. Jung and C. Kim, "Segmenting clustered nuclei using H-minima transform-based marker extraction and contour parameterization," Biomedical Engineering, IEEE Transactions on, vol. 57, pp. 2600-2604, 2010.

[40] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," Biomedical Engineering, IEEE Transactions on, vol. 57, pp. 841-852, 2010.

[41] M. Veta, A. Huisman, M. A. Viergever, P. J. van Diest, and J. P. Pluim, "Marker-controlled watershed segmentation of nuclei in H&E stained breast cancer biopsy images," in Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, 2011, pp. 618-621.

[42] M. Veta, P. J. van Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. Pluim, "Automatic nuclei segmentation in H&E stained breast cancer histopathology images," PLoS One, vol. 8, p. e70221, 2013.

[43] C. Jung, C. Kim, S. W. Chae, and S. Oh, "Unsupervised segmentation of overlapped nuclei using Bayesian classification," Biomedical Engineering, IEEE Transactions on, vol. 57, pp. 2825-2832, 2010.

[44] A. Veillard, M. S. Kulikova, and D. Racoceanu, "Cell nuclei extraction from breast cancer histopathologyimages using colour, texture, scale and shape information," Diagnostic Pathology, vol. 8, pp. S5-S5, 09/30 2013.

[45] J. P. Vink, M. Van Leeuwen, C. Van Deurzen, and G. De Haan, "Efficient nucleus detector in histopathology images," Journal of microscopy, vol. 249, pp. 124-135, 2013.

[46] L. Yang, W. Chen, P. Meer, G. Salaru, L. A. Goodell, V. Berstis, et al., "Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens," IEEE Transactions on Information Technology in Biomedicine, vol. 13, pp. 636-644, 2009.

[47] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on, 2008, pp. 496-499.

[48] S. Petushi, F. U. Garcia, M. M. Haber, C. Katsinis, and A. Tozeren, "Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer," BMC Medical Imaging, vol. 6, p. 14, 2006.

[49] P. Filipczuk, M. Kowal, and A. Obuchowicz, "Multi-label fast marching and seeded watershed segmentation methods for diagnosis of breast cancer cytology," Conf Proc IEEE Eng Med Biol Soc, vol. 2013, pp. 7368-71, 2013.

[50] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. Shih, et al., "Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent," Scientific Reports, vol. 7, p. 46450, 2017.

[51] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," arXiv preprint arXiv:1606.05718, 2016.

[52] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review-current status and future potential," IEEE Reviews in Biomedical Engineering, vol. 7, pp. 97-114, 2014.

[53] M. Veta, J. P. W. Pluim, P. J. Van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: A review," IEEE Transactions on Biomedical Engineering, vol. 61, pp. 1400-1411, 2014.

[54] P. J. van Diest, J. P. Baak, P. Matze-Cok, E. C. Wisse-Brekelmans, C. M. van Galen, P. H. Kurver, et al., "Reproducibility of mitosis counting in 2,469 breast cancer specimens: results from the Multicenter Morphometric Mammary Carcinoma Project," Hum Pathol, vol. 23, pp. 603-7, Jun 1992.

[55] J. P. Baak, P. J. van Diest, T. Benraadt, E. Matze-Cok, J. Brugghe, L. T. Schuurmans, et al., "The Multi-Center Morphometric Mammary Carcinoma Project (MMMCP) in The Netherlands: value of morphometrically assessed proliferation and differentiation," J Cell Biochem Suppl, vol. 17g, pp. 220-5, 1993.

[56] M. Veta, P. J. van Diest, M. Jiwa, S. Al-Janabi, and J. P. Pluim, "Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method," PloS one, vol. 11, p. e0161286, 2016.

[57] M. Macenko, M. Niethammer, J. Marron, D. Borland, J. T. Woosley, X. Guan, et al., "A method for normalizing histology

slides for quantitative analysis," in Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on, 2009, pp. 1107-1110.

[58] J.-R. Dalle, H. Li, C.-H. Huang, W. K. Leow, D. Racoceanu, and T. C. Putti, "Nuclear pleomorphism scoring by selective cell nuclei detection," in WACV, 2009.

[59] Z. Gandomkar, K. Tay, W. Ryder, P. C. Brennan, and C. Mello-Thoms, "iCAP: An Individualized Model Combining Gaze Parameters and Image-based Features to Predict Radiologists' Decisions While Reading Mammograms," IEEE transactions on medical imaging, vol. 36, pp. 1066-1075, 2017.

[60] B. Dunne and J. Going, "Scoring nuclear pleomorphism in breast cancer," Histopathology, vol. 39, pp. 259-265, 2001.

[61] S. Sinha, N. Sinha, R. Bandyopadhyay, and S. K. Mondal, "Robinson's cytological grading on aspirates of breast carcinoma: Correlation with Bloom Richardson's histological grading," Journal of Cytology/Indian Academy of Cytologists, vol. 26, p. 140, 2009.

[62] M. Bibbo, A. Abati, A. Ferenczy, J. Robitaille, E. Franco, J. Arseneau, et al., "The uniform approach to breast fine needle aspiration biopsy," Acta Cytologica, vol. 40, pp. 1120-1126, 1996.

[63] J. Mouriquand and D. Pasquier, "Fine needle aspiration of breast carcinoma: a preliminary cytoprognostic study," Acta cytologica, vol. 24, pp. 153-159, 1980.

[64] E. R. Fisher, C. Redmond, and B. Fisher, "Histologic grading of breast cancer," Pathol Annu, vol. 15, pp. 239-51, 1980.

[65] P. Arul and S. Masilamani, "Comparative evaluation of various cytomorphological grading systems in breast carcinoma," Indian journal of medical and paediatric oncology: official journal of Indian Society of Medical & Paediatric Oncology, vol. 37, p. 79, 2016.

[66] I. A. Robinson, G. McKee, A. Nicholson, P. A. Jackson, M. G. Cook, J. D'Arcy, et al., "Prognostic value of cytological grading of fine-needle aspirates from breast carcinomas," The Lancet, vol. 343, pp. 947-949, 1994/04/16/ 1994.

[67] E. Taniguchi, Q. Yang, W. Tang, Y. Nakamura, L. Shan, M. Nakamura, et al., "Cytologic grading of invasive breast carcinoma. Correlation with clinicopathologic variables and predictive value of nodal metastasis," Acta Cytol, vol. 44, pp. 587-91, Jul-Aug 2000.

[68] M. Khan, A. Haleem, H. Al Hassani, and H. Kfoury, "Cytopathological grading, as a predictor of histopathological grade, in ductal carcinoma (NOS) of breast, on air-dried Diff-Quik smears," Diagnostic cytopathology, vol. 29, pp. 185-193, 2003.

[69] D. Einstien, B. Omprakash, H. Ganapathy, and S. Rahman, "Comparison of 3-tier cytological grading systems for breast carcinoma," ISRN oncology, vol. 2014, 2014.

[70] K. Saha, G. Raychaudhuri, B. K. Chattopadhyay, and I. Das, "Comparative evaluation of six cytological grading systems in breast carcinoma," Journal of Cytology / Indian Academy of Cytologists, vol. 30, pp. 87-93, Apr-Jun 2013.

[71] S. Pal and M. Gupta, "Correlation between cytological and histological grading of breast cancer and its role in prognosis," Journal of Cytology, vol. 33, pp. 182-186, October 1, 2016 2016.

[72] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," IEEE Transactions on Biomedical Engineering, vol. 63, pp. 1455-1462, 2016.

[73] T. J. Lawton, G. Acs, P. Argani, G. Farshid, M. Gilcrease, N. Goldstein, et al., "Interobserver Variability by Pathologists in the Distinction Between Cellular Fibroadenomas and Phyllodes Tumors," International journal of surgical pathology, vol. 22, pp. 695-698, 08/26 2014.

[74] P. Spieler and M. Rössle, Nongynecologic Cytopathology: A Practical Guide: Springer Science & Business Media, 2012.

[75] T. A. Longacre, M. Ennis, L. A. Quenneville, A. L. Bane, I. J. Bleiweiss, B. A. Carter, et al., "Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study," Modern pathology, vol. 19, p. 195, 2006.

[76] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in Neural Networks (IJCNN), 2016 International Joint Conference on, 2016, pp. 2560-2567.

[77] F. A. Spanhol, P. R. Cavalin, L. S. Oliveira, C. Petitjean, and L. Heutte, "Deep Features for Breast Cancer Histopathological Image Classification."

[78] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model," Scientific Reports, vol. 7, 2017.

[79] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1-9.

[80] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[81] L. Todorovski and S. Džeroski, "Combining classifiers with meta decision trees," Machine learning, vol. 50, pp. 223-249, 2003.

[82] D. Onder, S. Zengin, and S. Sarioglu, "A review on color normalization and color deconvolution methods in histopathology," Applied Immunohistochemistry & Molecular Morphology, vol. 22, pp. 713-719, 2014.

[83] A. Janowczyk, A. Basavanhally, and A. Madabhushi, "Stain normalization using sparse autoencoders (StaNoSA): Application to digital pathology," Computerized Medical Imaging and Graphics, vol. 57, pp. 50-61, 2017.

[84] M. D. Zarella, C. Yeoh, D. E. Breen, and F. U. Garcia, "An alternative reference space for H&E color normalization," PloS one, vol. 12, p. e0174489, 2017.

[85] N. Alsubaie, N. Trahearn, S. E. A. Raza, D. Snead, and N. M. Rajpoot, "Stain Deconvolution Using Statistical Analysis of Multi-Resolution Stain Colour Representation," PloS one, vol. 12, p. e0169875, 2017.

[86] A. Paradiso, I. Ellis, F. Zito, E. Marubini, S. Pizzamiglio, and P. Verderio, "Short-and long-term effects of a training session on

pathologists' performance: the INQAT experience for histological grading in breast cancer," Journal of clinical pathology, vol. 62, pp. 279-281, 2009.

[87] Z. Gandomkar et al., "Detection of the abnormal gist in the prior mammograms even with no overt sign of breast cancer," in 14th International Workshop on Breast Imaging (IWBI 2018), 2018, vol. 10718, p. 1071804: International Society for Optics and Photonics.

[88] P. C. Brennan et al., "Radiologists can detect the 'gist'of breast cancer before any overt signs of cancer appear," Scientific reports, vol. 8, no. 1, p. 8717, 2018.

Chapter 10

Conclusion

Breast cancer is the most commonly diagnosed non-melanoma cancer among women worldwide [1] and hence large numbers of breast biopsies are examined by pathologists each year. For example, annually in the USA, pathologists evaluate 1.6 million breast specimens [2, 3]. Therefore, even slightly improving current practice in breast pathology through computer-assisted analysis can benefit many women. This thesis aimed at determining if computer-aided analysis of Hematoxylin-Eosin (HE) stained breast histopathological digital slides can be used to better understand pathologists' perception of mitotic figures. It also investigated the possibility of reproducible nuclear atypia scoring by combining computer-assisted analysis with cytological scores given by a pathologist. Furthermore, this thesis examined the feasibility of computer-assisted analysis for classification of HE breast images into various subcategories of benign or cancer masses.

This thesis involved a literature review on the current status of computer-assisted analysis in breast pathology along with three original research studies addressing deficiencies from the existing literature of computer-assisted analysis of Hematoxylin-Eosin stained images in breast pathology. The first study explored the feasibility of relating quantitative image processing features with disagreement among pathologists in the mitotic recognition task.

The results of the first study suggested that there are quantitative image-based features that differ significantly among easily identifiable mitotic figures, challenging mitotic figures, and miscounted non-mitoses within breast slides. The rules, which were extracted in this study and described the appearances of easily identifiable mitoses, challenging mitoses (false negatives), and miscounted non-mitoses (false positives), could be helpful

in providing detailed and more objective constraints in mitoses counting protocols to mitigate the discrepancies among pathologists in the recognition of mitotic figures. These rules can be also discussed with the pathologists in training sessions. Moreover, these rules can be utilised to develop an educational software for pathology trainees which starts instruction with retrieving easy mitoses and non-mitoses and retrieves more difficult objects as training continues.

The second original research study proposed a personalized tool for reproducible nuclear atypia scoring. The tool, called COMPASS (COMputer-assisted analysis combined with Pathologist's ASSessment), relies on both computer-extracted features and scores given by the pathologists to cytological characteristics. COMPASS was designed to assist junior pathologists. It achieved a performance comparable to a senior pathologist in nuclear grading and hence could be potentially used to reduce inter-pathologist variations in nuclear grading [4-11], which was observed more among less experienced pathologists [12, 13]. It was also shown that textural features provide complementary information to cytological scores for histologic nuclear grading. Currently, fine needle aspiration (FNA) is being increasingly utilized and providing breast cancer grade based on FNA in the pathology report is recommend, as it could be beneficial for the management of a given patient [14, 15]. Because of limited sample size in FNA, histological grading is not completely practical based on FNA, and several attempts have been made to produce a cytological grade on FNA which corresponds well with histological grade. Our findings suggested that computer-extracted features complement the cytological scores for predicting histologic nuclear grade and therefore can potentially be used in grading FNA breast biopsies.

In the last study, MuDeRN (MUlti-category classification of breast histopathological image using DEep Residual Networks) has been proposed and tested. It distinguishes malignant patients from benign ones based on Hematoxylin-Eosin breast images and then categorizes cancer and benign cases into four different subtypes each. MuDeRN achieved a correct classification rate of 96.25% for eight-class categorization of patients. This high accuracy suggests that MuDeRN could be helpful in the categorization of breast lesions to reduce discrepancies among pathologists' diagnoses.

In conclusion, the findings presented in the first research study of this thesis demonstrated that computer-aided image analysis can help in better understanding of image-related features related to discrepancies among pathologists (the mitoses recognition task was evaluated). The second and third research studies indicated the feasibility of computer-assisted nuclear grading (COMPASS) and computer-assisted diagnosis (MuDeRN). Therefore, three important tasks in breast pathology could benefit from the findings presented in this thesis. The results could be used to improve current status of breast cancer prognosis estimation through reducing the inter-pathologists disagreement in counting mitotic figures and reproducible nuclear grading. It can also improve provide a second opinion to the pathologist for making diagnosis and hence reduce diagnostic discrepancies among pathologists.

References

[1] J. Ferlay, C. Héry, P. Autier, and R. Sankaranarayanan, "Global burden of breast cancer," in Breast cancer epidemiology, ed: Springer, 2010, pp. 1-19.

[2] M. Silverstein, "Where's the outrage?," J Am Coll Surg, vol. 208, pp. 78-9, Jan 2009.

[3] M. J. Silverstein, A. Recht, M. D. Lagios, I. J. Bleiweiss, P. W. Blumencranz, T. Gizienski, et al., "Special report: Consensus conference III. Image-detected breast cancer: state-of-the-art diagnosis and treatment," J Am Coll Surg, vol. 209, pp. 504-20, Oct 2009.

[4] G. S. Delides, G. Garas, G. Georgouli, D. Jiortziotis, J. Lecca, T. Liva, et al., "Intralaboratory variations in the grading of breast carcinoma," Arch Pathol Lab Med, vol. 106, pp. 126-8, Mar 1982.

[5] J. M. Harvey, N. H. de Klerk, and G. F. Sterrett, "Histological grading in breast cancer: interobserver agreement, and relation to other prognostic factors including ploidy," Pathology, vol. 24, pp. 63-8, Apr 1992.

[6] H. F. Frierson, Jr., R. A. Wolber, K. W. Berean, D. W. Franquemont, M. J. Gaffey, J. C. Boyd, et al., "Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma," Am J Clin Pathol, vol. 103, pp. 195-8, Feb 1995.

[7] P. Robbins, S. Pinder, N. de Klerk, H. Dawkins, J. Harvey, G. Sterrett, et al., "Histological grading of breast carcinomas: a study

of interobserver agreement," Hum Pathol, vol. 26, pp. 873-9, Aug 1995.

[8] S. R. Lakhani, J. Jacquemier, J. P. Sloane, B. A. Gusterson, T. J. Anderson, M. J. van de Vijver, et al., "Multifactorial analysis of differences between sporadic breast cancers and cancers involving BRCA1 and BRCA2 mutations," J Natl Cancer Inst, vol. 90, pp. 1138-45, Aug 05 1998.

[9] P. Boiesen, P. O. Bendahl, L. Anagnostaki, H. Domanski, E. Holm, I. Idvall, et al., "Histologic grading in breast cancer--reproducibility between seven pathologic departments. South Sweden Breast Cancer Group," Acta Oncol, vol. 39, pp. 41-5, 2000.

[10] J. S. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, et al., "Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index," Mod Pathol, vol. 18, pp. 1067-78, Aug 2005.

[11] I. N. f. Q. A. o. T. B. Group, "Quality control for histological grading in breast cancer: an Italian experience," Pathologica, vol. 97, p. 1, 2005.

[12] B. Dunne and J. Going, "Scoring nuclear pleomorphism in breast cancer," Histopathology, vol. 39, pp. 259-265, 2001.

[13] R. Zhang, H.-j. Chen, B. Wei, H.-y. Zhang, Z.-g. Pang, H. Zhu, et al., "Reproducibility of the Nottingham modification of the Scarff-Bloom-Richardson histological grading system and the complementary value of Ki-67 to this system," 2010.

[14] S. Sinha, N. Sinha, R. Bandyopadhyay, and S. K. Mondal, "Robinson's cytological grading on aspirates of breast carcinoma: Correlation with Bloom Richardson's histological grading," Journal of Cytology/Indian Academy of Cytologists, vol. 26, p. 140, 2009.

[15] M. Bibbo, A. Abati, A. Ferenczy, J. Robitaille, E. Franco, J. Arseneau, et al., "The uniform approach to breast fine needle aspiration biopsy," Acta Cytologica, vol. 40, pp. 1120-1126, 1996.

Bibliography

Adams, A. L., Chhieng, D. C., Bell, W. C., Winokur, T., & Hameed, O. (2009). Histologic grading of invasive lobular carcinoma: does use of a 2-tiered nuclear grading system improve interobserver variability? *Annals of diagnostic pathology, 13*(4), 223-225.

Adams, A. L., Li, Y., Pfeifer, J. D., & Hameed, O. (2010). Nuclear Grade and Survival in Invasive Lobular Carcinoma: A Case Series with Long-term Follow-up. *The breast journal, 16*(4), 445-447.

Al-Janabi, S., van Slooten, H.-J., Visser, M., Van Der Ploeg, T., Van Diest, P. J., & Jiwa, M. (2013). Evaluation of mitotic activity index in breast cancer using whole slide digital images. *PloS one, 8*(12), e82576.

Al-Kofahi, Y., Lassoued, W., Lee, W., & Roysam, B. (2010). Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering, 57*(4), 841-852.

Allison, K. H., Reisch, L. M., Carney, P. A., Weaver, D. L., Schnitt, S. J., O'malley, F. P., . . . Elmore, J. G. (2014). Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. *Histopathology, 65*(2), 240-251.

Alsubaie, N., Trahearn, N., Raza, S. E. A., Snead, D., & Rajpoot, N. M. (2017). Stain Deconvolution Using Statistical Analysis of Multi-Resolution Stain Colour Representation. *PloS one, 12*(1), e0169875.

Arul, P., & Masilamani, S. (2016). Comparative evaluation of various cytomorphological grading systems in breast carcinoma. *Indian journal of medical and paediatric oncology: official journal of Indian Society of Medical & Paediatric Oncology, 37*(2), 79.

Baak, J. P., van Diest, P. J., Benraadt, T., Matze-Cok, E., Brugghe, J., Schuurmans, L. T., & Littooy, J. J. (1993). The Multi-Center Morphometric Mammary Carcinoma Project (MMMCP) in The Netherlands: value of morphometrically assessed proliferation and differentiation. *J Cell Biochem Suppl, 17g*, 220-225.

Bansal, C., Pujani, M., Sharma, K., Srivastava, A., & Singh, U. (2014). Grading systems in the cytological diagnosis of breast cancer: A review. *Journal of Cancer Research and Therapeutics, 10*(4), 839-845. doi:10.4103/0973-1482.140979

Bedell, M. B., Wood, M. E., Lezotte, D. C., Sedlacek, S. M., & Orleans, M. M. (1995). Delay in diagnosis and treatment of breast cancer: implications for education. *Journal of Cancer Education, 10*(4), 223-228.

Bhosale, S. J., Kshirsagar, A. Y., Sulhyan, S. R., Jagtap, S. V., & Nikam, Y. P. (2010). Invasive Papillary Breast Carcinoma. *Case Reports in Oncology, 3*(3), 410-415. doi:10.1159/000321270

Bibbo, M., Abati, A., Ferenczy, A., Robitaille, J., Franco, E., Arseneau, J., . . . Bernacki, E. G. (1996). The uniform approach to breast fine needle aspiration biopsy. *Acta cytologica, 40*(6), 1120-1126.

Boiesen, P., Bendahl, P. O., Anagnostaki, L., Domanski, H., Holm, E., Idvall, I., . . . Ferno, M. (2000). Histologic grading in breast cancer--reproducibility between seven pathologic departments. South Sweden Breast Cancer Group. *Acta Oncol, 39*(1), 41-45.

Bois, M. C., Al-Hilli, Z., Visscher, D. W., Hoskin, T. L., Frost, M. H., Radisky, D. C., . . . Carter, J. M. (2016). *Microglandular adenosis and risk of breast cancer: a Mayo benign breast disease cohort study.* Paper presented at the LABORATORY INVESTIGATION.

Bondi, A., Lega, S., Crucitti, P., Pierotti, P., Rapezzi, R., Sassoli De Bianchi, P., & Naldoni, C. (2012). Quality assurance and automation. *Cytopathology, 23*, 39.

Brennan, P. C., Gandomkar, Z., Ekpo, E. U., Tapia, K., Trieu, P. D., Lewis, S. J., . . . Evans, K. K. (2018). Radiologists can detect the 'gist'of breast cancer before any overt signs of cancer appear. *Scientific Reports, 8*(1), 8717.

Bueno-de-Mesquita, J. M., Nuyten, D., Wesseling, J., van Tinteren, H., Linn, S., & van De Vijver, M. (2009). The impact of inter-observer variation in pathological assessment of node-negative breast cancer on clinical risk assessment and patient selection for adjuvant systemic treatment. *Annals of oncology, 21*(1), 40-47.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Chtourou, I., Makni, S. K., Bahri, I., Abbes, K., Sellami, A., Fakhfakh, I., . . . Boudawara, T. S. (2009). [Pure colloid carcinoma of the breast: anatomoclinical study of seven cases]. *Cancer Radiother, 13*(1), 37-41. doi:10.1016/j.canrad.2008.06.004

Cosatto, E., Miller, M., Graf, H. P., & Meyer, J. S. (2008). *Grading nuclear pleomorphism on histological micrographs.* Paper presented at the Pattern Recognition, 2008. ICPR 2008. 19th International Conference on.

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N. N., . . . Madabhushi, A. (2017). Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Scientific Reports, 7*, 46450.

Cserni, G., Orosz, Z., Kulka, J., Sápi, Z., Kálmán, E., & Bori, R. (2006). Divergences in diagnosing nodular breast lesions of noncarcinomatous nature. *Pathology & Oncology Research, 12*(4), 216-221.

Dalle, J.-R., Li, H., Huang, C.-H., Leow, W. K., Racoceanu, D., & Putti, T. C. (2009). *Nuclear pleomorphism scoring by selective cell nuclei detection.* Paper presented at the WACV.

Delgermaa Demchig, M., Ziba Gandomkar, M., & Patrick, C. Automatic Segmentation of the Dense Tissue in Digital Mammograms for BIRADS Density Categorization.

Delides, G. S., Garas, G., Georgouli, G., Jiortziotis, D., Lecca, J., Liva, T., & Elemenoglou, J. (1982). Intralaboratory variations in the grading of breast carcinoma. *Arch Pathol Lab Med, 106*(3), 126-128.

Doyle, S., Agner, S., Madabhushi, A., Feldman, M., & Tomaszewski, J. (2008). *Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features.* Paper presented at the Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on.

Dumitru, A., Procop, A., Iliesiu, A., Tampa, M., Mitrache, L., Costache, M., . . . Cirstoiu, M. (2015). Mucinous Breast Cancer: a Review Study of 5 Year Experience from a Hospital-Based Series of Cases. *Mædica, 10*(1), 14-18.

Dunne, B., & Going, J. (2001). Scoring nuclear pleomorphism in breast cancer. *Histopathology, 39*(3), 259-265.

Einstien, D., Omprakash, B., Ganapathy, H., & Rahman, S. (2014). Comparison of 3-tier cytological grading systems for breast carcinoma. *ISRN oncology, 2014*.

El Khouli, R. H., & Louie, A. (2009). Case of the Season: A Giant Fibroadenoma in the Guise of a Phyllodes Tumor; Characterization Role of MRI. *Seminars in roentgenology, 44*(2), 64-66. doi:10.1053/j.ro.2008.12.003

Ellis, I. O., Coleman, D., Wells, C., Kodikara, S., Paish, E. M., Moss, S., . . . Winder, R. (2006). Impact of a national external quality assessment scheme for breast pathology in the UK. *J Clin Pathol, 59*(2), 138-145. doi:10.1136/jcp.2004.025551

Elmore, J. G., Barton, M. B., Moceri, V. M., Polk, S., Arena, P. J., & Fletcher, S. W. (1998). Ten-year risk of false positive screening mammograms and clinical breast examinations. *New England Journal of Medicine, 338*(16), 1089-1096.

Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A. N., . . . Schnitt, S. J. (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama, 313*(11), 1122-1132.

Elston, C. W., & Ellis, I. O. (1991). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology, 19*(5), 403-410.

Ferlay, J., Héry, C., Autier, P., & Sankaranarayanan, R. (2010). Global burden of breast cancer *Breast cancer epidemiology* (pp. 1-19): Springer.

Filipczuk, P., Kowal, M., & Obuchowicz, A. (2013). *Multi-label fast marching and seeded watershed segmentation methods for diagnosis of breast cancer cytology.* Paper presented at the Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE.

Filipczuk, P., Kowal, M., & Obuchowicz, A. (2013). Multi-label fast marching and seeded watershed segmentation methods for diagnosis of breast cancer cytology. *Conf Proc IEEE Eng Med Biol Soc, 2013*, 7368-7371. doi:10.1109/embc.2013.6611260

Fisher, E. R., Redmond, C., & Fisher, B. (1980). Histologic grading of breast cancer. *Pathol Annu, 15*(Pt 1), 239-251.

Frierson, H. F., Jr., Wolber, R. A., Berean, K. W., Franquemont, D. W., Gaffey, M. J., Boyd, J. C., & Wilbur, D. C. (1995). Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. *Am J Clin Pathol, 103*(2), 195-198.

Frost, A. R., Terahata, S., Yeh, I. T., Siegel, R. S., Overmoyer, B., & Silverberg, S. G. (1995). An analysis of prognostic features in infiltrating lobular carcinoma of the breast. *Mod Pathol, 8*.

Gamdonkar, Z., Tay, K., Ryder, W., Brennan, P. C., & Mello-Thoms, C. (2015). *iDensity: an automatic Gabor filter-based algorithm for breast density assessment.* Paper presented at the Medical Imaging 2015: Image Perception, Observer Performance, and Technology Assessment.

Gandomkar, Z., Brennan, P., & Mello-Thoms, C. (2018). *Nuclear Atypia Scoring by Combining Pathologist's Assessment and Computer-Assisted Analysis.* Paper presented at the BreastScreen Australia Conference 2018.

Gandomkar, Z., Brennan, P. C., & Mello-Thoms, C. (2016). Computer-based image analysis in breast pathology. *Journal of pathology informatics, 7*.

Gandomkar, Z., Brennan, P. C., & Mello-Thoms, C. (2017a). Determining image processing features describing the appearance of challenging mitotic figures and miscounted nonmitotic objects. *Journal of pathology informatics, 8*.

Gandomkar, Z., Brennan, P. C., & Mello-Thoms, C. (2017b). *Determining local and contextual features describing appearance of difficult to identify mitotic figures.* Paper presented at the Medical Imaging 2017: Digital Pathology.

Gandomkar, Z., Brennan, P. C., & Mello-Thoms, C. (2018a). *A cognitive approach to determine the benefits of pairing radiologists in mammogram reading.* Paper presented at the Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment.

Gandomkar, Z., Brennan, P. C., & Mello-Thoms, C. (2018b). *A framework for distinguishing benign from malignant breast histopathological images using deep residual networks.* Paper presented at the 14th International Workshop on Breast Imaging (IWBI 2018).

Gandomkar, Z., Brennan, P. C., & Mello-Thoms, C. (2018c). MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. *Artificial intelligence in medicine*.

Gandomkar, Z., Ekpo, E. U., Lewis, S. J., Evans, K. K., Tapia, K., Trieu, P.-D., . . . Brennan, P. C. (2018). *Detection of the abnormal gist in the prior mammograms even with no overt sign of breast cancer.* Paper presented at the 14th International Workshop on Breast Imaging (IWBI 2018).

Gandomkar, Z., Tay, K., Brennan, P. C., Kozuch, E., & Mello-Thoms, C. (2018). Can eye-tracking metrics be used to better pair radiologists in a mammogram reading task? *Medical physics*.

Gandomkar, Z., Tay, K., Brennan, P. C., & Mello-Thoms, C. (2017). *A model based on temporal dynamics of fixations for distinguishing expert radiologists' scanpaths.* Paper presented at the Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment.

Gandomkar, Z., Tay, K., Brennan, P. C., & Mello-Thoms, C. (2018). Recurrence quantification analysis of radiologists' scanpaths when interpreting mammograms. *Medical physics*.

Gandomkar, Z., Tay, K., Ryder, W., Brennan, P. C., & Mello-Thoms, C. (2016). *Predicting radiologists' true and false positive decisions in reading mammograms by using gaze parameters and image-based features.* Paper presented at the Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment.

Gandomkar, Z., Tay, K., Ryder, W., Brennan, P. C., & Mello-Thoms, C. (2017). iCAP: An Individualized Model Combining Gaze Parameters and Image-based Features to Predict Radiologists' Decisions While Reading Mammograms. *IEEE transactions on medical imaging, 36*(5), 1066-1075.

Ghaznavi, F., Evans, A., Madabhushi, A., & Feldman, M. (2013). Digital imaging in pathology: whole-slide imaging and beyond. *Annual Review of Pathology: Mechanisms of Disease, 8*, 331-359.

Griffiths, D., Melia, J., McWilliam, L., Ball, R., Grigor, K., Harnden, P., . . . Waller, M. (2006). A study of Gleason score interpretation in different groups of UK pathologists; techniques for improving reproducibility. *Histopathology, 48*(6), 655-662.

Group, I. N. f. Q. A. o. T. B. (2005). Quality control for histological grading in breast cancer: an Italian experience. *Pathologica, 97*(1), 1.

Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., & Yener, B. (2009). Histopathological image analysis: A review. *IEEE reviews in biomedical engineering, 2*, 147-171.

Haghdadi, N., Asaei, B., & Gandomkar, Z. (2011). *A clustering-based preprocessing on feeder power in presence of photovoltaic power plant.* Paper presented at the Environment and Electrical Engineering (EEEIC), 2011 10th International Conference on.

Haghdadi, N., Asaei, B., & Gandomkar, Z. (2012). *Clustering-based optimal sizing and siting of photovoltaic power plant in distribution network.* Paper presented at the Environment and Electrical Engineering (EEEIC), 2012 11th International Conference on.

Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., & Li, S. (2017). Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model. *Scientific Reports, 7*.

Harvey, J. M., de Klerk, N. H., & Sterrett, G. F. (1992). Histological grading in breast cancer: interobserver agreement, and relation to other prognostic factors including ploidy. *Pathology, 24*(2), 63-68.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Howell, L. P., Gandour-Edwards, R., & O'Sullivan, D. (1994). Application of the Scarff-Bloom-Richardson tumor grading system to fine-needle aspirates of the breast. *Am J Clin Pathol, 101*(3), 262-265.

Irshad, H. (2013). Automated mitosis detection in histopathology using morphological and multi-channel statistics features. *Journal of pathology informatics, 4*.

Irshad, H., Jalali, S., Roux, L., Racoceanu, D., Hwee, L. J., Le Naour, G., & Capron, F. (2013). Automated mitosis detection using texture, SIFT features and HMAX biologically inspired approach. *Journal of pathology informatics, 4*(Suppl).

Irshad, H., Roux, L., & Racoceanu, D. (2013). *Multi-channels statistical and morphological features based mitosis detection in breast cancer histopathology.* Paper presented at the Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS.

Irshad, H., Veillard, A., Roux, L., & Racoceanu, D. (2014). Methods for nuclei detection, segmentation, and classification in digital histopathology: A review-current status and future potential. *IEEE reviews in biomedical engineering, 7*, 97-114. doi:10.1109/RBME.2013.2295804

Jackson, S. L., Frederick, P. D., Pepe, M. S., Nelson, H. D., Weaver, D. L., Allison, K. H., . . . Elmore, J. G. (2017). Diagnostic Reproducibility: What Happens When the Same Pathologist Interprets the Same Breast Biopsy Specimen at Two Points in Time? *Annals of surgical oncology, 24*(5), 1234-1241. doi:10.1245/s10434-016-5695-0

Janowczyk, A., Basavanhally, A., & Madabhushi, A. (2017). Stain normalization using sparse autoencoders (StaNoSA): Application to digital pathology. *Computerized Medical Imaging and Graphics, 57*, 50-61.

Jensen, R. A., Page, D. L., Dupont, W. D., & Rogers, L. W. (1989). Invasive breast cancer risk in women with sclerosing adenosis. *Cancer, 64*(10), 1977-1983.

Jung, C., & Kim, C. (2010). Segmenting clustered nuclei using H-minima transform-based marker extraction and contour parameterization. *Biomedical Engineering, IEEE Transactions on, 57*(10), 2600-2604.

Jung, C., Kim, C., Chae, S. W., & Oh, S. (2010). Unsupervised segmentation of overlapped nuclei using Bayesian classification. *Biomedical Engineering, IEEE Transactions on, 57*(12), 2825-2832.

Khan, A. M., Eldaly, H., & Rajpoot, N. M. (2013). A gamma-gaussian mixture model for detection of mitotic cells in breast cancer histopathology images. *J Pathol Inform, 4*, 11. doi:10.4103/2153-3539.112696

Khan, A. M., Sirinukunwattana, K., & Rajpoot, N. (2015). A global covariance descriptor for nuclear atypia scoring in breast histopathology images. *IEEE journal of biomedical and health informatics, 19*(5), 1637-1647.

Khan, M., Haleem, A., Al Hassani, H., & Kfoury, H. (2003). Cytopathological grading, as a predictor of histopathological grade, in ductal carcinoma (NOS) of breast, on air-dried Diff-Quik smears. *Diagnostic cytopathology, 29*(4), 185-193.

Khazai, L., Middleton, L. P., Goktepe, N., Liu, B. T., & Sahin, A. A. (2015). Breast pathology second review identifies clinically significant discrepancies in over 10% of patients. *Journal of surgical oncology, 111*(2), 192-197.

Khushi, M., Edwards, G., de Marcos, D. A., Carpenter, J. E., Graham, J. D., & Clarke, C. L. (2013). Open source tools for management and archiving of digital microscopy data to allow integration with patient pathology and treatment information. *Diagnostic Pathology, 8*(1). doi:10.1186/1746-1596-8-22

Komenaka, I. K., El-Tamer, M. B., Troxel, A., Hamele-Bena, D., Joseph, K. A., Horowitz, E., . . . Schnabel, F. R. (2004). Pure mucinous carcinoma of the breast. *Am J Surg, 187*(4), 528-532. doi:10.1016/j.amjsurg.2003.12.039

Lakhani, S. R., Jacquemier, J., Sloane, J. P., Gusterson, B. A., Anderson, T. J., van de Vijver, M. J., . . . Easton, D. F. (1998). Multifactorial analysis of differences between sporadic breast cancers and cancers involving BRCA1 and BRCA2 mutations. *J Natl Cancer Inst, 90*(15), 1138-1145.

Lawton, T. J., Acs, G., Argani, P., Farshid, G., Gilcrease, M., Goldstein, N., . . . Reynolds, C. (2014). Interobserver Variability by Pathologists in the Distinction Between Cellular Fibroadenomas and Phyllodes Tumors. *International journal of surgical pathology, 22*(8), 695-698. doi:10.1177/1066896914548763

Li, C. I., Anderson, B. O., Daling, J. R., & Moe, R. E. (2003). Trends in incidence rates of invasive lobular and ductal breast carcinoma. *Jama, 289*(11), 1421-1424.

Longacre, T. A., Ennis, M., Quenneville, L. A., Bane, A. L., Bleiweiss, I. J., Carter, B. A., . . . Layfield, L. J. (2006). Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study. *modern Pathology, 19*(2), 195.

Lundin, M., Lundin, J., Helin, H., & Isola, J. (2004). A digital atlas of breast histopathology: an application of web based virtual microscopy. *Journal of clinical pathology, 57*(12), 1288-1291.

Macenko, M., Niethammer, M., Marron, J., Borland, D., Woosley, J. T., Guan, X., . . . Thomas, N. E. (2009). *A method for normalizing histology slides for quantitative analysis.* Paper presented at the Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on.

Malon, C., Brachtel, E., Cosatto, E., Graf, H. P., Kurata, A., Kuroda, M., . . . Yagi, Y. (2012). Mitotic figure recognition: Agreement among pathologists and computerized detector. *Analytical Cellular Pathology, 35*(2), 97-100.

Malon, C. D., & Cosatto, E. (2013). Classification of mitotic figures with convolutional neural networks and seeded blob features. *Journal of pathology informatics, 4*.

Maqlin, P., Thamburaj, R., Mammen, J. J., & Nagar, A. K. (2013). *Automatic detection of tubules in breast histopathological images.* Paper presented at the Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012).

Maughan, K. L., Lutterbie, M. A., & Ham, P. S. (2010). Treatment of breast cancer. *Chemotherapy, 51*, 53.

McPherson, K. (2010). Screening for breast cancer-balancing the debate. *BMJ: British Medical Journal, 340*.

Medri, L., Volpi, A., Nanni, O., Vecci, A. M., Mangia, A., Schittulli, F., . . . Amadori, D. (2003). Prognostic relevance of mitotic activity in patients with node-negative breast cancer. *modern Pathology, 16*(11), 1067.

Meyer, J. S., Alvarez, C., Milikowski, C., Olson, N., Russo, I., Russo, J., . . . Parwaresch, R. (2005). Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. *Mod Pathol, 18*(8), 1067-1078. doi:10.1038/modpathol.3800388

Mook, S., Schmidt, M. K., Rutgers, E. J., van de Velde, A. O., Visser, O., Rutgers, S. M., . . . Ravdin, P. M. (2009). Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online Adjuvant! program: a hospital-based retrospective cohort study. *The lancet oncology, 10*(11), 1070-1076.

Mouriquand, J., & Pasquier, D. (1980). Fine needle aspiration of breast carcinoma: a preliminary cytoprognostic study. *Acta cytologica, 24*(2), 153-159.

Nguyen, K., Barnes, M., Srinivas, C., & Chefd'hotel, C. (2015). *Automatic glandular and tubule region segmentation in histological grading of breast cancer.* Paper presented at the Medical Imaging 2015: Digital Pathology.

Niethammer, M., Borland, D., Marron, J., Woosley, J. T., & Thomas, N. E. (2010). *Appearance Normalization of Histology Slides.* Paper presented at the MLMI.

Onder, D., Zengin, S., & Sarioglu, S. (2014). A review on color normalization and color deconvolution methods in histopathology. *Applied Immunohistochemistry & Molecular Morphology, 22*(10), 713-719.

Paeng, K., Hwang, S., Park, S., & Kim, M. (2017). A unified framework for tumor proliferation score prediction in breast histopathology *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 231-239): Springer.

Pal, S., & Gupta, M. (2016). Correlation between cytological and histological grading of breast cancer and its role in prognosis. *Journal of Cytology, 33*(4), 182-186. doi:10.4103/0970-9371.190449

Pandey, P., Dixit, A., Chandra, S., & Kaur, S. (2014). A comparative and evaluative study of two cytological grading systems in breast carcinoma with histological grading: an important prognostic factor. *Analytical Cellular Pathology, 2014*.

Pantanowitz, L., Valenstein, P. N., Evans, A. J., Kaplan, K. J., Pfeifer, J. D., Wilbur, D. C., . . . Colgan, T. J. (2011). Review of the current state of whole slide imaging in pathology. *Journal of pathology informatics, 2*(1), 36.

Paradiso, A., Ellis, I., Zito, F., Marubini, E., Pizzamiglio, S., & Verderio, P. (2009). Short-and long-term effects of a training session on pathologists' performance:

the INQAT experience for histological grading in breast cancer. *Journal of clinical pathology, 62*(3), 279-281.

Petushi, S., Garcia, F. U., Haber, M. M., Katsinis, C., & Tozeren, A. (2006). Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. *BMC Medical Imaging, 6*(1), 14.

Phukan, J. P., Sinha, A., & Deka, J. P. (2015). Cytological grading of breast carcinoma on fine needle aspirates and its relation with histological grading. *South Asian Journal of Cancer, 4*(1), 32-34. doi:10.4103/2278-330X.149948

Robbins, P., Pinder, S., de Klerk, N., Dawkins, H., Harvey, J., Sterrett, G., . . . Elston, C. (1995). Histological grading of breast carcinomas: a study of interobserver agreement. *Hum Pathol, 26*(8), 873-879.

Roberti, N. E. (1997). The role of histologic grading in the prognosis of patients with carcinoma of the breast. *Cancer, 80*(9), 1708-1716.

Robinson, I. A., McKee, G., Nicholson, A., Jackson, P. A., Cook, M. G., D'Arcy, J., & Kissin, M. W. (1994). Prognostic value of cytological grading of fine-needle aspirates from breast carcinomas. *The Lancet, 343*(8903), 947-949. doi:https://doi.org/10.1016/S0140-6736(94)90066-3

Rosen, P. P. (2001). *Rosen's breast pathology*: Lippincott Williams & Wilkins.

Roullier, V., Lézoray, O., Ta, V. T., & Elmoataz, A. (2010) Mitosis extraction in breast-cancer histopathological whole slide images. *Vol. 6453 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 539-548).

Roullier, V., Lézoray, O., Ta, V. T., & Elmoataz, A. (2011). Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization. *Computerized Medical Imaging and Graphics, 35*(7-8), 603-615. doi:10.1016/j.compmedimag.2011.02.005

Roullier, V., Ta, V. T., Lézoray, O., & Elmoataz, A. (2010). *Graph-based multi-resolution segmentation of histological whole slide images.* Paper presented at the 2010 7th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2010 - Proceedings.

Roux, L., Racoceanu, D., Capron, F., Calvo, J., Attieh, E., Le Naour, G., & Gloaguen, A. (2014). Mitos & atypia. *Image Pervasive Access Lab (IPAL), Agency Sci., Technol. & Res. Inst. Infocom Res., Singapore, Tech. Rep, 1*.

Saha, K., Raychaudhuri, G., Chattopadhyay, B. K., & Das, I. (2013). Comparative evaluation of six cytological grading systems in breast carcinoma. *Journal of Cytology / Indian Academy of Cytologists, 30*(2), 87-93. doi:10.4103/0970-9371.112647

Schnitt, S. J., Connolly, J. L., Tavassoli, F. A., Fechner, R. E., Kempson, R. L., Gelman, R., & Page, D. L. (1992). Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria. *The American journal of surgical pathology, 16*(12), 1133-1143.

Silverstein, M. (2009). Where's the outrage? *J Am Coll Surg, 208*(1), 78-79. doi:10.1016/j.jamcollsurg.2008.09.022

Silverstein, M. J., Recht, A., Lagios, M. D., Bleiweiss, I. J., Blumencranz, P. W., Gizienski, T., . . . Willey, S. C. (2009). Special report: Consensus conference III. Image-detected breast cancer: state-of-the-art diagnosis and treatment. *J Am Coll Surg, 209*(4), 504-520. doi:10.1016/j.jamcollsurg.2009.07.006

Sinha, P., Bendall, S., & Bates, T. (2000). Does routine grading of invasive lobular cancer of the breast have the same prognostic significance as for ductal cancers? *European Journal of Surgical Oncology (EJSO), 26*(8), 733-737.

Sinha, S., Sinha, N., Bandyopadhyay, R., & Mondal, S. K. (2009). Robinson's cytological grading on aspirates of breast carcinoma: Correlation with Bloom Richardson's histological grading. *Journal of Cytology/Indian Academy of Cytologists, 26*(4), 140.

Spanhol, F. A., Cavalin, P. R., Oliveira, L. S., Petitjean, C., & Heutte, L. Deep Features for Breast Cancer Histopathological Image Classification.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). *Breast cancer histopathological image classification using convolutional neural networks.* Paper presented at the Neural Networks (IJCNN), 2016 International Joint Conference on.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering, 63*(7), 1455-1462.

Spieler, P., & Rössle, M. (2012). *Nongynecologic Cytopathology: A Practical Guide*: Springer Science & Business Media.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). *Going deeper with convolutions.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Talei, A., Akrami, M., Mokhtari, M., & Tahmasebi, S. (2012). Surgical and Clinical Pathology of Breast Diseases *Histopathology-Reviews and Recent Advances*: InTech.

Taniguchi, E., Yang, Q., Tang, W., Nakamura, Y., Shan, L., Nakamura, M., . . . Kakudo, K. (2000). Cytologic grading of invasive breast carcinoma. Correlation with clinicopathologic variables and predictive value of nodal metastasis. *Acta Cytol, 44*(4), 587-591.

Tavassoli, F. A. (1999). *Pathology of the Breast*: McGraw Hill Professional.

Tice, J. A., O'Meara, E. S., Weaver, D. L., Vachon, C., Ballard-Barbash, R., & Kerlikowske, K. (2013). Benign Breast Disease, Mammographic Breast Density, and the Risk of Breast Cancer. *JNCI Journal of the National Cancer Institute, 105*(14), 1043-1049. doi:10.1093/jnci/djt124

Todorovski, L., & Džeroski, S. (2003). Combining classifiers with meta decision trees. *Machine learning, 50*(3), 223-249.

van Diest, P. J., Baak, J. P., Matze-Cok, P., Wisse-Brekelmans, E. C., van Galen, C. M., Kurver, P. H., . . . et al. (1992). Reproducibility of mitosis counting in 2,469 breast cancer specimens: results from the Multicenter Morphometric Mammary Carcinoma Project. *Hum Pathol, 23*(6), 603-607.

Veillard, A., Kulikova, M. S., & Racoceanu, D. (2013). Cell nuclei extraction from breast cancer histopathologyimages using colour, texture, scale and shape information. *Diagnostic Pathology, 8*(Suppl 1), S5-S5. doi:10.1186/1746-1596-8-S1-S5

Veta, M., Huisman, A., Viergever, M. A., van Diest, P. J., & Pluim, J. P. (2011). *Marker-controlled watershed segmentation of nuclei in H&E stained breast cancer biopsy images.* Paper presented at the Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on.

Veta, M., Pluim, J. P. W., Van Diest, P. J., & Viergever, M. A. (2014). Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering, 61*(5), 1400-1411. doi:10.1109/TBME.2014.2303852

Veta, M., van Diest, P. J., Jiwa, M., Al-Janabi, S., & Pluim, J. P. (2016). Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method. *PloS one, 11*(8), e0161286.

Veta, M., van Diest, P. J., Kornegoor, R., Huisman, A., Viergever, M. A., & Pluim, J. P. (2013). Automatic nuclei segmentation in H&E stained breast cancer histopathology images. *PloS one, 8*(7), e70221. doi:10.1371/journal.pone.0070221

Veta, M., Van Diest, P. J., & Pluim, J. P. W. (2013). *Detecting mitotic figures in breast cancer histopathology images.* Paper presented at the Progress in Biomedical Optics and Imaging - Proceedings of SPIE.

Vink, J. P., Van Leeuwen, M., Van Deurzen, C., & De Haan, G. (2013). Efficient nucleus detector in histopathology images. *Journal of microscopy, 249*(2), 124-135.

Wachtel, M. S., Halldorsson, A., & Dissanaike, S. (2011). Nottingham Grades of Lobular Carcinoma Lack the Prognostic Implications They Bear for Ductal Carcinoma1. *Journal of Surgical Research, 166*(1), 19-27. doi:https://doi.org/10.1016/j.jss.2010.05.016

Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

Weaver, D. L., Rosenberg, R. D., Barlow, W. E., Ichikawa, L., Carney, P. A., Kerlikowske, K., . . . Maygarden, S. J. (2006). Pathologic findings from the breast cancer surveillance consortium. *Cancer, 106*(4), 732-742.

Weaver, D. L., Rosenberg, R. D., Barlow, W. E., Ichikawa, L., Carney, P. A., Kerlikowske, K., . . . Ballard-Barbash, R. (2006). Pathologic findings from the Breast Cancer Surveillance Consortium: population-based outcomes in women undergoing biopsy after screening mammography. *Cancer, 106*(4), 732-742. doi:10.1002/cncr.21652

Wells, C., McGregor, I., Makunura, C., Yeomans, P., & Davies, J. (1995). Apocrine adenosis: a precursor of aggressive breast cancer? *Journal of clinical pathology, 48*(8), 737-742.

Wells, W. A., Carney, P. A., Eliassen, M. S., Tosteson, A. N., & Greenberg, E. R. (1998). Statewide study of diagnostic agreement in breast pathology. *JNCI: Journal of the National Cancer Institute, 90*(2), 142-145.

Weyn, B., van de Wouwer, G., van Daele, A., Scheunders, P., van Dyck, D., van Marck, E., & Jacob, W. (1998). Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. *Cytometry, 33*(1), 32-40.

Yang, L., Chen, W., Meer, P., Salaru, G., Goodell, L. A., Berstis, V., & Foran, D. J. (2009). Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens. *IEEE Transactions on Information Technology in Biomedicine, 13*(4), 636-644.

Zarella, M. D., Yeoh, C., Breen, D. E., & Garcia, F. U. (2017). An alternative reference space for H&E color normalization. *PloS one, 12*(3), e0174489.

Zhang, R., Chen, H.-j., Wei, B., Zhang, H.-y., Pang, Z.-g., Zhu, H., . . . Bu, H. (2010). Reproducibility of the Nottingham modification of the Scarff-Bloom-Richardson histological grading system and the complementary value of Ki-67 to this system.