UNIVERSITY OF DEUSTO

DEEP LEARNING FOR BREAST CANCER DIAGNOSIS

Doctoral Thesis

by

Zabit Hameed

Advisor: Prof. Begoña Garcia-Zapirain
Co-advisor: Dr. Ibon Oleagordea Ruiz

Faculty of Engineering
University of Deusto, Spain.

Bilbao, June 2023

UNIVERSITY OF DEUSTO

DEEP LEARNING FOR BREAST CANCER DIAGNOSIS

by

Zabit Hameed

A thesis presented to the Faculty of Engineering for the degree of
Doctor of Philosophy *with* International Mention

Advisor: Prof. Begoña Garcia-Zapirain
Co-advisor: Dr. Ibon Oleagordea Ruiz

Student sign:

Advisors sign:

Bilbao, June 2023

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this research are original and have not been submitted in whole or in part for consideration for any other degree or qualification at the University of Deusto or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgment.

Zabit Hameed

June 2023
Bilbao, Spain.

Dedication

*This thesis is dedicated to my beloved parents **Aye** and **Dada**.*

Abstract

Breast cancer is a common malignancy and a leading cause of cancer-related deaths in women worldwide. Its early diagnosis can significantly reduce morbidity and mortality rates in women. To this end, histopathological diagnosis is usually followed as a gold standard approach. However, this process is tedious, labor-intensive, and may be subject to inter-reader variability. Accordingly, an automatic diagnostic system can assist in improving the quality of diagnosis. The key intent of this thesis is to design, optimize, and validate end-to-end systems based on deep learning for the effective and efficient diagnosis of breast malignancy using histopathology images.

In our first contribution, we conducted a systematic review of state-of-the-art supervised machine and deep learning approaches in the detection, segmentation, and classification of breast lesions using widely used medical imaging modalities including mammography, sonography, magnetic resonance imaging, and histopathology during the years 2016 and 2022. It is inferred that convolutional neural networks are progressively being exploited in computer-aided diagnosis of breast cancer. Furthermore, it is deduced that mammography and magnetic resonance imaging were mostly utilized in detection and segmentation tasks, followed by sonography. Whereas, mammography and histopathology were predominantly used in classification tasks.

In our second contribution, we accomplished the first case study related to the binary classification of breast cancer. In this study, we presented an ensemble deep learning approach for the definite classification of non-carcinoma and carcinoma histopathology images using our collected dataset. We trained four different models based on pre-trained VGG16 and VGG19 architectures. Initially, we followed 5-fold cross-validation operations on all the individual models, namely, fully-trained VGG16, fine-tuned VGG16, fully-trained VGG19, and fine-tuned VGG19 models. Then, we followed an ensemble strategy by taking the average of predicted probabilities and found that the ensemble of fine-tuned VGG16 and fine-tuned VGG19 performed competitive classification performance, especially on the carcinoma class. The ensemble of fine-tuned VGG16 and VGG19 models offered sensitivity of 97.73% for the carcinoma

class. Moreover, it offered accuracy of 95.29% and F1-score of 95.29%. These experimental results demonstrated that our proposed deep learning approach is effective for the automatic classification of complex-natured histopathology images of breast cancer, more specifically for carcinoma images.

In our third and final contribution, we accomplished the second case study related to the multiclass classification of breast cancer. In this study, we presented a deep learning approach to automatically classify hematoxylin-eosin-stained microscopy images into normal tissues, benign lesions, in situ carcinoma, and invasive carcinoma using our collected dataset. The proposed model exploited six intermediate layers of the Xception network to retrieve robust and abstract features from input images. First, we optimized the proposed model on the original (unnormalized) dataset using 5-fold cross-validation. Then, we investigated its performance on four normalized datasets resulting from Reinhard, Ruifrok, Macenko, and Vahadane stain normalization. For original images, the proposed framework yielded accuracy of 98.00% along with Cohen's kappa score of 0.969. Furthermore, it achieved an average AUC-ROC score of 0.998 as well as a mean AUC-PR value of 0.995. Specifically, for in situ carcinoma and invasive carcinoma, it offered sensitivity of 96.00% and 99.00%, respectively. For normalized images, the proposed architecture performed better for Macenko normalization compared to the other three techniques. In this case, the proposed model achieved accuracy of 97.79% together with Cohen's kappa score of 0.965. In addition, it attained an average AUC-ROC score of 0.997 and a mean AUC-PR value of 0.991. Especially, for in situ carcinoma and invasive carcinoma, it offered sensitivity of 96.00% and 99.00%, respectively. These results demonstrate that our proposed model outperformed the baseline AlexNet as well as state-of-the-art VGG16, VGG19, Inception-v3, and Xception models with their default settings. Furthermore, it can be inferred that although stain normalization techniques offered competitive performance, they could not surpass the results of the original dataset.

This thesis is accomplished with three journal articles whereas the fourth one is under review. Similarly, two papers are presented and published at the international conferences. Moreover, an international research stay was successfully performed at the Université Laval in Canada.

Keywords: Artificial intelligence, biomedical engineering, biomedical image processing, breast cancer diagnosis, deep learning, histopathology.

Acknowledgement

The work presented in this thesis could never have been achieved without the support and guidance of many people. I would first like to express my unfeigned gratitude to the Spanish Ministry of Science for giving me the opportunity as a Personal Investigador to complete my doctoral thesis in engineering at the University of Deusto.

It is with a heartfelt appreciation that express my thanks to my thesis director Professor Begoña García-Zapirain for her continuous support, motivation, and enthusiasm. Her advisory has enabled me to grow as a passionate independent thinker, always aiming to learn and discover new knowledge. Also, I thank to my co-advisor Dr. Ibon Oleagordia Ruiz for his continuous guidance during this journey. I am grateful to the entire team of the eVida Research Group for their moral support during this period.

I extend my heartfelt gratitude to Professor Simon Duchesne for giving me the exceptional opportunity for an invaluable international research stay at the Université Laval in Canada.

A special thanks to my family who have shared the Ph.D. journey with me, and always showered me with their love and kindness. I would also like to thank all my friends in Bilbao who made this journey more enjoyable.

Finally and most importantly, I am eternally grateful to my parents and wife for their unwavering love, support, and prayers at every phase of my life. I will always strive to be the best version of myself.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

According to Global Cancer Statistics 2020, breast cancer is the most common malignancy and the leading cause of cancer mortality in women worldwide [17]. It was diagnosed in 20.26 million women, accounting for 11.7% of all cancer cases in 2020. Moreover, it caused 0.96 million deaths in women, representing 6.9% of all cancer deaths in 2020 [17]. These statistics indicate that it is one of the deadliest cancers affecting the women's population worldwide. Therefore, the premature understanding of the underlying pathophysiology of breast cancer is crucial to reduce morbidity and mortality among women.

To that end, the pathological study is followed as a benchmark to comprehend the pathophysiology of breast tumors. In this method, tissue samples are collected and mounted on glass slides, and subsequently stained these slides for a better portrayal of tumoral characteristics. Afterward, pathologists proceed with the microscopic examination of these slides to conclude a possible diagnosis of breast tissues. Nevertheless, the manual interpretation of histopathology images can be a tedious and time-consuming process. Moreover, morphological criteria followed during the manual analysis depend mainly on the domain experience of engaged pathologists. For instance, a study revealed that the overall concordance rate of diagnostic interpretation among participating pathologists was around 75% [18]. Therefore, computer-assisted diagnostic systems could help in improving the overall diagnostic process of breast malignancy.

During the last decade, deep learning (DL) models have made remarkable progress in computer vision, specifically in medical image processing, due to their abilities to automatically learn complicated and advanced features from images [19, 20]. It encouraged various researchers to exploit DL models for breast cancer diagnosis using

medical images. The current thesis designed, optimized, and validated novel end-to-end systems based on DL approaches to effectively and efficiently diagnose breast lesions using microscopy images. The research presented in this thesis could help clinicians diagnose breast tissues as non-carcinoma and carcinoma. Furthermore, it could assist clinicians in diagnosing breast lesions as normal, benign, in situ carcinoma, and invasive carcinoma.

1.1 Research hypothesis and objectives

The analysis of the problem outlined in the previous section leads to the following hypothesis.

- *Optimized deep learning frameworks can effectively and efficiently diagnose breast cancer as* ***non-carcinoma and carcinoma*** *as well as* ***normal, benign, in situ carcinoma, and invasive carcinoma*** *using histopathology images.*

Considering the aforementioned hypothesis, the current thesis leverages supervised DL architectures to effectively diagnose breast malignancy using microscopy images. To achieve this, the four objectives listed below need to be fulfilled.

- **Objective 1:** To collect and annotate a dataset containing whole slide images retrieved from eighty female patients suffering from breast cancer.
- **Objective 2:** To define the state-of-the-art developments in the supervised machine and deep learning for breast cancer diagnosis using widely followed medical imaging modalities.
- **Objective 3:** To design, optimize, and validate a supervised DL framework aimed at effective ***binary classification*** of breast cancer using our collected dataset. This objective encompasses the first case study in this thesis.
- **Objective 4:** To design, optimize, and validate a supervised DL framework intended at effective ***multiclass classification*** of breast cancer as normal, benign, in situ carcinoma, and invasive carcinoma using our collected dataset. This objective comprehends the second case study in this thesis.

1.2 Scientific and social impact

This section aims to highlight the scientific as well as social impact of the research works presented in the thesis.

The contributions that emerged from this thesis have started showing its scientific significance. In terms of scientific impact, publication in upper-quartile peer-reviewed journals is an important measure of the quality of any scientific work. The present thesis is accomplished with three journal articles whereas the fourth one is under review. Similarly, two papers have been presented and published at the international conferences. Furthermore, an international research stay was successfully performed at the Université Laval in Canada. After successful publications, the citation is one of the standard indicators for evaluating the quality of research works. The research articles published during the development of this thesis accumulated more than 250 citations until June 2023. Moreover, one of our published articles was recently highlighted by a prestigious Biomedical Engineering Community of the Nature Portfolio which can be accessed here: https://bioengineeringcommunity.nature.com/posts/connecting-ai-pathology-using-tiatoolbox. These statistics indicate that scientific contributions achieved with this thesis are being validated by the research community worldwide.

In terms of social impact, our propounded models could assist clinicians during the diagnostic process of breast tissues. It is worth mentioning that the manual diagnostic decisions during the microscopic analysis sometimes might be difficult due to intra-class variation and inter-class consistency within the histopathology images of breast cancer. To that end, the propounded end-to-end systems are effective and may assist in reducing the bias caused by human errors. Furthermore, the proposed systems are comparably efficient and may decrease the time required in the diagnostic process. Finally, it could help to reduce cancer-related mortality rates among female patients worldwide.

1.3 Research methodology

This section explains the research methodology followed in each of the two studies presented in this thesis, as displayed in Figure 1.1. It should be noted that although both studies are related to breast cancer diagnosis, each study case has a comparably different background, as well as different materials and methods.

- **State-of-the-art:** The main objective of this step is to analyze and understand the current state-of-the-art machine and deep learning techniques involved in

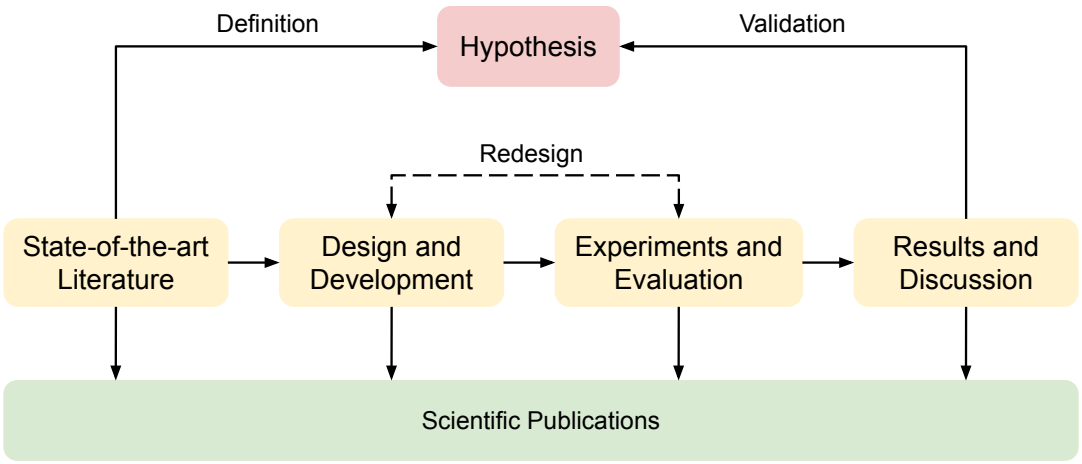


Figure 1.1 The research methodology followed in each of the two studies presented in this thesis.

breast cancer diagnosis. The knowledge procured during this stage will lead to the formulation of the hypothesis and, subsequently, to the comparison with the obtained results.

- **Design and development:** After literature analysis and processing of knowledge secured in the previous phase, this stage involves the design and development of different elements of the proposed system.
- **Experiment and evaluation:** At this stage, numerous metrics are defined for the evaluation of the propounded system. The designed system is redesigned based on its performance on the validation data.
- **Results and discussion:** This step aims to compare as well as discuss the acquired results with state-of-the-art studies, which leads to the final assessment of the established hypothesis.

1.4 Thesis organization

This section describes the complete structure of this thesis. Chapter 1 introduces the significance of the thesis as well as the methodology followed for its accomplishment. Chapters 2 presents relevant literature on breast cancer diagnosis. Chapter 3 and 4 comprehend the main contributions and each contains a separate case study with its own

materials, methods, results, discussions, and conclusions. Finally, Chapter 5 highlights conclusive remarks together with future recommendations. A brief description of each chapter is presented below.

- **Chapter 1 - Introduction:** Chapter 1 introduces the significance of this thesis and the methodology followed for its accomplishment. It starts by introducing the research hypothesis together with the objectives to be fulfilled during the development of this thesis. It then explains the research methodology followed to fulfill the stated objectives. Finally, it highlights the structure of this thesis.

- **Chapter 2 - Literature review:** Chapter 2 discusses relevant approaches used in the computed-assisted diagnosis of breast malignancy. This chapter analyzes 142 supervised machine and deep learning studies in the detection, segmentation, and classification of breast cancer during the years 2016 to 2022.

- **Chapter 3 - Study I: Binary classification of Breast Cancer:** Chapter 3 presents the first case study of the thesis in which we designed, optimized, and validated an end-to-end system based on an ensemble of deep CNN models to classify breast lesions into non-carcinoma and carcinoma.

- **Chapter 4 - Study II: Multiclass classification of Breast Cancer:** Chapter 4 introduces the second case study of the thesis in which we designed, optimized, and validated an end-to-end system based on multilevel features of CNN models to classify breast lesions into normal, benign, in situ carcinoma, and invasive carcinoma.

- **Chapter 5 - Conclusion:** Chapter 5 outlines conclusions extracted from the final evaluation of the research work. It also discusses how the objectives set in Chapter 1 were successfully accomplished. Furthermore, it summarizes the scientific contributions and international collaboration carried out during the development of this thesis. Finally, it highlights the limitations and future recommendations of the thesis.

Chapter 2

Literature Review

2.1 Introduction

The anatomy of a female breast is composed of milk-producing lobules, milk-carrying ducts, adipose tissues, fibrous tissues, blood vessels, and lymph vessels [21]. The lobules are also known as glandular tissues and adipose tissues are often called fatty tissues. The complete structure of a female breast is illustrated in Figure 2.1 [1]. Breast malignancy usually occurs due to abnormalities in epithelial tissues of the breast and may infect the nearby stroma, ducts, and lobules [21, 22]. The emanated tumors can be benign or malignant. Benign tumors are non-cancerous and arise from small structural changes in the breast, whereas malignant tumors are cancerous and are divided into in situ and invasive carcinomas [21, 22]. In situ carcinoma predominates within ducts and does not spread to adjacent tissues [21, 22]. However, invasive cancer can invade nearby areas through the immune system or systemic circulation [21, 22]. Its early diagnosis can reduce morbidity and mortality in women worldwide.

To that end, the breast cancer diagnosis usually begins with radiology imaging exams including mammography [23], sonography [23], magnetic resonance imaging (MRI) [23], followed by histopathology imaging [22]. Expert radiologists and pathologists analyze these images for possible diagnosis of breast malignancy. However, the traditional manual approach may lead to biased results due to domain expertise. For radiology imaging like mammography, sonography, and MRI, the American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) aimed to standardize the clinical assessment of breast malignancy [24]. The BI-RADS lexicon guidelines help radiologists grade breast tissues according to predefined categories and thus improve consistency. However, this process could be subjective due

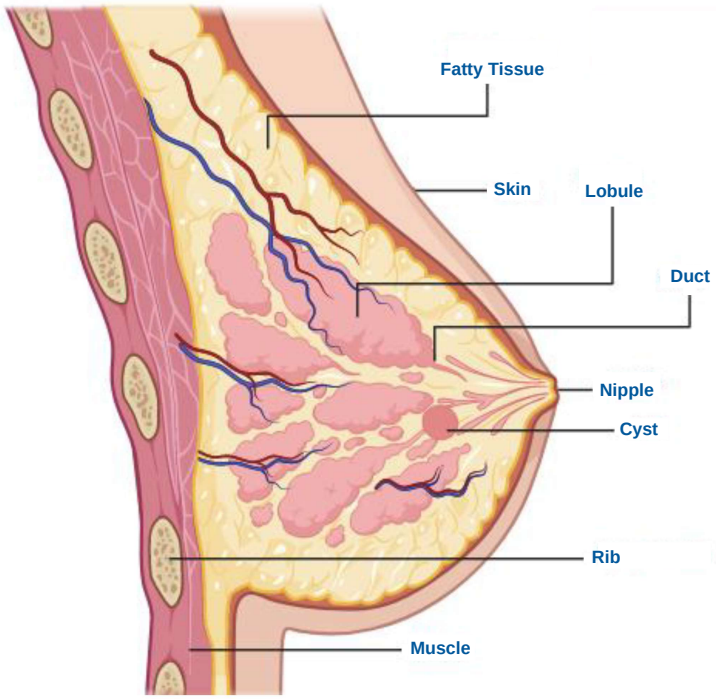


Figure 2.1 The anatomical structure of a female breast [1].

to inter-observer variability. Similarly, for histopathology images, pathologists follow numerous morphological criteria to distinguish breast lesions. Nevertheless, the criteria used in the manual interpretation of histology images could be subjective and may be prone to error. This led to the diagnostic agreement among pathologists being around 75% [18]. In addition, there still exists a shortage of human experts to provide timely diagnosis and refer patients to the appropriate clinical care.

To this end, computer-aided detection and diagnosis (CAD) systems are usually followed as a second opinion with an aim to assist in the diagnostic process. The CAD systems can be classified as computer-aided detection (CADe) and computer-aided diagnosis (CADx) systems [25]. The CADe systems help to detect and locate breast lesions in medical images. Whereas the CADx systems assist to categorize and diagnose breast tumors, for instance, benign and malignant [25]. The CADe and CADx systems may vary from traditional machine learning (ML) techniques to novel deep learning (ML) models. Nevertheless, ML CAD systems mainly rely on feature engineering which may lead to biased results. Therefore, DL models have been increasingly used in radiology and histopathology images to effectively improve the diagnostic process of breast malignancy. A generalized CAD framework is portrayed in Figure 2.2 [2]. It should be noted that we considered CADe systems from input images to region of in-

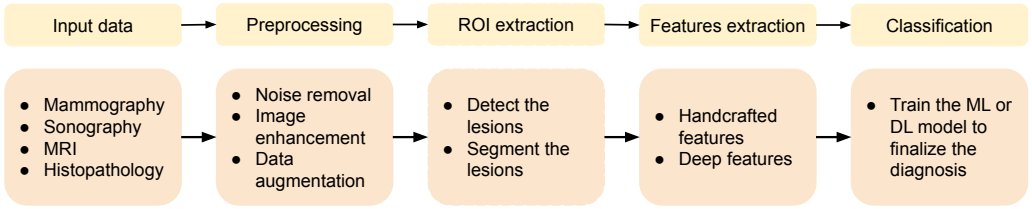


Figure 2.2 The complete process of computer-aided breast cancer diagnosis [2].

terest (ROI) extraction using ML and DL detection and segmentation models. Wheres, we considered CADx systems from input images to lesion diagnosis using ML and DL classification frameworks [25].

The rationale of the current systematic review is to address the following research questions (RQs). It should be noted that we selected published studies from January 2016 to December 2022 due to the rise of DL frameworks during the last decade [26].

- *RQ1: Which imaging modalities have been followed the most in the detection, segmentation, and classification of breast cancer?*
- *RQ2: How many datasets have been publicly released related to breast cancer diagnosis using mammography, sonography, MRI, and histopathology?*
- *RQ3: Which ML and DL models (CADe systems) are mostly followed in the detection and segmentation of breast cancer?*
- *RQ4: Which ML and DL models (CADx systems) are mostly followed in the classification of breast cancer?*

The remaining sections of this systematic review article are organized as follows. Section 2.2 highlights the radiology imaging modalities followed by histopathology imaging. Section 2.3 explained the databases searched using specific terms, inclusion criteria, and exclusion criteria of studies analyzed within the current systematic review. Section 2.4 outlined ML and DL methodologies in the detection, segmentation, and classification of breast malignancy. Section 2.5 discussed the summary of research carried out during the years 2016 and 2022 using supervised ML and DL models in the detection, segmentation, and classification of breast cancer. Finally, section 2.6 outlined the conclusion, consideration, and future direction.

2.2 Medical imaging in breast cancer diagnosis

In this section, we provided answers to the research questions, *RQ1* and *RQ2*. To that end, we discussed widely used medical imaging modalities including mammography, sonography, magnetic resonance imaging, and histopathology imaging in the detection, segmentation, and diagnosis of breast cancer, as well as their associated datasets available from January 2016 to December 2022, in the succeeding subsections.

2.2.1 Mammography

Mammography is specialized medical imaging that uses low-dose (20-32 peak kilovoltage) X-rays to create radiographic images (mammograms) of the breast and is conventionally compressed in two different planes, called mediolateral oblique (MLO) and craniocaudal (CC) views [27]. Being the simplest form, screen-film mammography is used as a gold standard for screening and diagnosis of breast malignancy [27]. However, due to its technical and practical advantages, full-field digital mammography is increasingly adopted for breast cancer screening and diagnosis worldwide [28]. Screening mammograms aim to detect tumors at an earlier stage before the symptoms appear, whereas diagnostic mammograms are utilized in case of irregular symptoms [29]. The main focus during mammogram analysis involves the presence of smaller white spots called calcifications and larger abnormal areas known as masses [30]. However, the standard two-dimensional (2D) mammogram may lead to tissue overlap due to the projection of three-dimensional (3D) breast structures on 2D images [27, 31]. Consequently, advanced digital mammography approaches such as contrast-enhanced spectral mammography (CESM) and digital breast tomosynthesis (DBT) are increasingly used in breast imaging [27, 31]. The CESM leverages a dual-energy technique for image acquisition following the injection of an iodinated contrast medium. Specifically, it reveals the angiogenesis patterns of tumors using contrast media and provides a 2D contrast enhancement map of breast tissues [27, 31]. Whereas, the DBT is a 3D mammogram that acquires multiple images of the compressed breast at different narrow angles, which are then reconstructed into a stack of thin slices. This allows a comprehensive assessment of suspicious lesions that may be difficult to visualize during routine mammography [27, 31]. These novel methods have better sensitivity than conventional mammography for screening and diagnostic purposes. Further details about these procedures can be found in [27, 31].

To sum up, each category of this modality is widely used for the screening and detection of breast cancer. Nevertheless, it has limited detecting capabilities in some

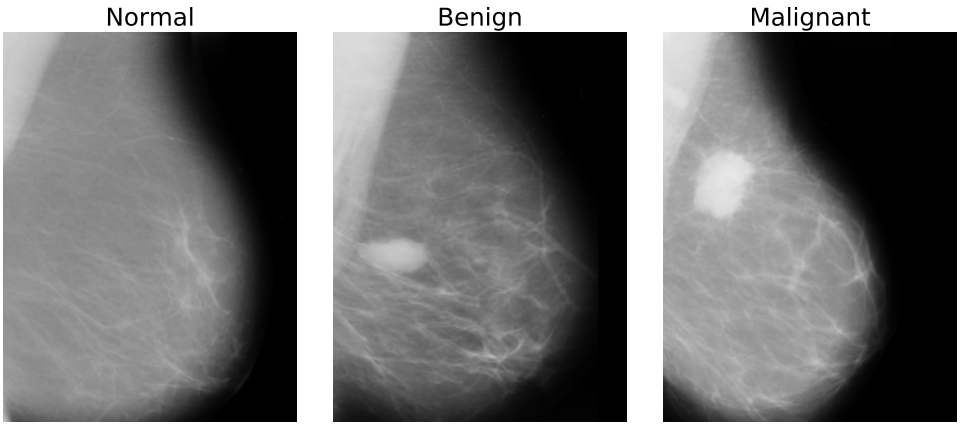


Figure 2.3 An example of normal, benign, and malignant mammograms [3].

cases, for instance, dense glandular breasts of young women [27, 31]. An example of normal, benign, and malignant mammograms are depicted in Figure 2.3. Similarly, the strengths and limitations of mammography are given in Table 2.1. Whereas, its publicly available datasets are provided in Table 2.2.

2.2.2 Sonography

Sonography is an ultrasound imaging that utilizes high-frequency (3-12 megahertz) sound waves to acquire images (sonograms) of breast tissues and is often used as a complementary screening tool along with mammography [27]. It can assess the morphology, orientation, internal structures, and margins of lesions from multiple planes. Evaluation of these features helps to differentiate benign and malignant tumors, especially in dense fibroglandular breasts where mammography may not be effective [27, 32]. Ultrasound is available in classic 2D grayscale as well as advanced color formats such as elastography. In clinical practices, malignant tumors are considered rigid compared to benign lesions [31]. Elastography uses this principle to measure tissue elasticity and can be conducted in two ways: strain elastography (SE) and shear wave elastography (SWE) [32]. The SE evaluates variations in tissue texture due to a freehand transducer pressure on the region of interest. This creates a colored map called elastogram which is then superimposed on the grayscale 2D ultrasound [27, 32]. Whereas, SWE leverages shear waves (10-200 hertz) which are induced by the acoustic radiation force and propagate transversely through tissues. These waves can pass through stiff lesions faster than soft tissues and are calculated as Young's modulus in kilopascals. This produces a colored image depicting elasticity in kilopascals, which is

Figure 2.4 An example of normal, benign, and malignant sonograms [4].

directly proportional to the speed of shear waves. Similar to SE, the resultant colored map of tissue elasticity is then overlaid on the grayscale 2D ultrasound in the region of interest [27, 32]. It should be noted that SWE depicts quantitative measurements of tissue elasticity, unlike SE, which shows relative elasticity before and after compression. Other than SE and SWE, other popular techniques include Nakagami imaging [33], color doppler [27, 32], power doppler [27, 32], automatic breast ultrasound (ABUS) [27, 32], contrast-enhanced ultrasound (CEUS) [27, 32], and 3D ultrasound [27, 32]. Further details of these approaches can be found in [27, 31, 32]. These novel sonography approaches have increased the diagnostic accuracy of breast malignancy compared to grayscale ultrasound imaging [31, 33, 34].

To conclude, it is an interactive, dynamic, safe, painless, and real-time scanning modality. Furthermore, it does not use radiation compared to mammograms and can be used for pregnant women. Nonetheless, it yielded comparably complex-natured and may also suffer from inter-observer variability. Also, there is no strong evidence to use it for routine screening and thus cannot be employed in the early detection of breast cancer [27, 31]. An example of normal, benign, and malignant sonograms are portrayed in Figure 2.4. Likewise, the advantages and disadvantages of sonography are given in Table 2.1. Whereas, its publicly available datasets are provided in Table 2.2.

2.2.3 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) uses a strong homogeneous magnetic field of at least 1.5 Tesla and radiofrequency (RF) waves generated by dedicated breast coils to acquire highly detailed 2D images of breast tissues [35]. These cross-sectional 2D images or slices can be stacked to generate a comprehensive 3D model of the area of interest. The MRI is usually followed as a supplemental tool with mammography and sonography in the diagnosis of breast lesions [35]. It can be utilized when the results

of mammography or sonography are not clear, especially in the case of dense breasts [35]. Moreover, it can be used as a screening tool for women at higher hereditary risk for the development of breast cancer [35]. The MRI leverages the magnetic properties of hydrogen atoms (protons) of a human body to produce diagnostics images. In this process, three main components are involved: a primary magnet, gradient coils, and RF coils [36, 37]. The primary magnet is used to create a strong static magnetic field. Similarly, gradient coils can produce three secondary fields over the primary field which in turn allows selection of slices in the axial, sagittal, and coronal planes. Moreover, RF coils are responsible for sending RF signals into breast tissues as well as receiving echo signals from breast tissues as MRI images [36, 37]. When a human body is placed in the primary magnet, hydrogen protons align either parallel or antiparallel to the applied field, resulting in a magnetic vector (longitudinal magnetization). The RF pulses are then transmitted to disturb the alignment of spinning protons or precession, deflecting the magnetization vector (decreasing the longitudinal magnetization). At this stage, hydrogen protons become synchronized and result in a net magnetization vector at the right angle of the main field, called transverse magnetization. When the source of radio waves is turned off, it brings back the magnetizing vector to its original position, and thus an MRI image is retrieved [36, 37]. There are two ways to measure the time required for protons to relax completely. The first T1-relaxation is the time it takes for a magnetic vector to return to its resting state. Whereas, the second is T2-relaxation, which is the time it takes for the axial spin to return to its resting state [36, 37]. To enhance the quality of conventional grayscale MRI, dynamic contrast-enhanced MRI (DCE-MRI) is followed which leverages gadolinium-based contrast media to detect breast cancers via tumor angiogenesis [35, 38]. Its basic protocol consists of a single precontrast T1-weighted acquisition and multiple postcontrast T1-weighted acquisitions to record the kinetic behavior of the contrast media accumulated in a tumor [35, 38]. Nonetheless, DCE-MRI offers limited specificity due to overlapping morphological and kinetic features of benign and malignant lesions, leading to unnecessary breast biopsies [39]. To improve specificity, other MRI techniques such as diffusion weighted imaging (DWI) [27, 40], magnetic resonance spectroscopy (MRS) [27, 40] and magnetic resonance elastography (MRE) [27, 40] have been explored, with DWI being the most robust method [41]. Furthermore, multiparametric breast MRI protocols are increasingly being performed, in which non-contrast T2-weighted and DWI acquisitions are also performed besides the native T1-weighted acquisition, as discussed in [35, 40, 41].

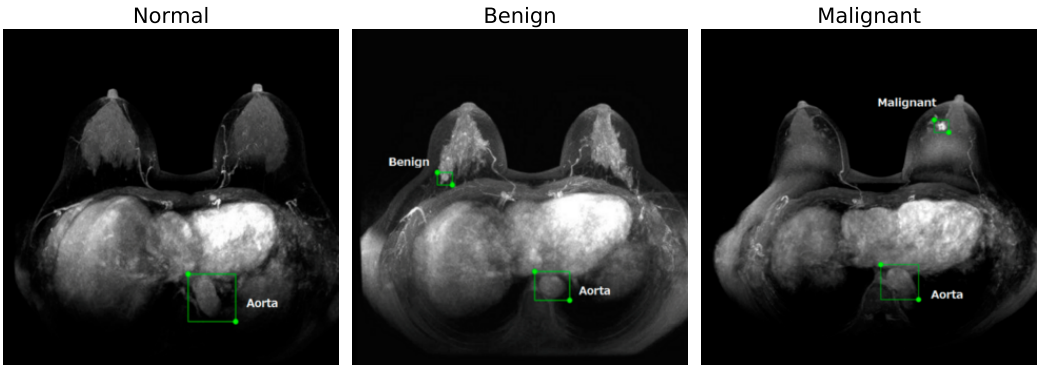


Figure 2.5 An example of normal, benign, and malignant MRI [5].

In summary, the MRI offers the highest sensitivity in the detection of breast lesions compared to mammography and sonography. Also, it does not use radiation and thus can be used in women at higher risk. However, MRI is comparably expensive and also takes a considerable time to perform. Therefore, it may not be feasible for breast screening in a large population. Furthermore, it may lead to "false positive" results which in turn require unnecessary biopsies [27, 31]. An example of normal, benign, and malignant MRI images are illustrated in Figure 2.5. Similarly, the pros and cons of MRI are given in Table 2.1. Whereas, its publicly available datasets are provided in Table 2.2.

2.2.4 Histopathology

Histopathology is a procedure in which tissue samples are collected and mounted on glass slides, and subsequently stained these slides for a better portrayal of morphological and immunophenotypical characteristics of breast tumors [42]. After that, pathologists proceed with the microscopic examination of these slides to conclude a possible diagnosis of breast cancer [42]. The complete steps of the histopathological procedure have been discussed in [8, 43]. This process is also called biopsy in medical terminology. The digitized image generated from each slide is called a whole slide image (WSI) whereas an image extracted from a specific region within WSI is called a microscopy image [8, 43]. Pathologists evaluate and annotate regions of interest within a stained WSI at different zoom levels under the microscope. It should be noted that hematoxylin and eosin (H&E) stain usually helps pathologists to visualize breast tissues in an effective way. In histopathology image analysis, pathologists mainly examine cellular findings compared to radiology image analysis where radiologists typically focus on structural elements of a breast [44].

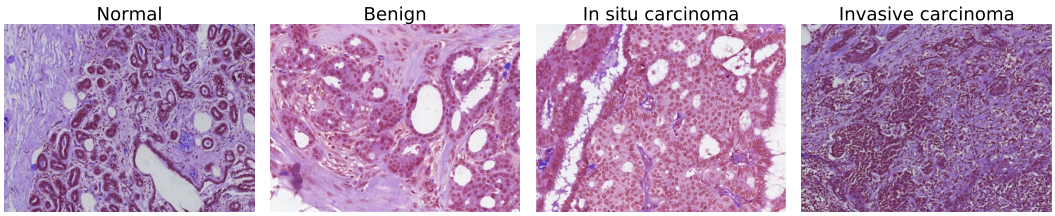


Figure 2.6 An example of normal, benign, in situ carcinoma, and invasive carcinoma microscopy images [6].

Table 2.1 The advantages and disadvantages of mammography, sonography, magnetic resonance imaging, and histopathology imaging in breast cancer.

Imaging modality	Strengths	Limitations
Mammography	Non-invasive, efficient, and cost-effective method [45]	Exposure to potentially harmful radiation [46]
	Routinely used for breast cancer screening [47]	Limited application in dense breasts [45]
	Can detect breast cancer at an early stage [46]	May require additional diagnostic tests [47]
Sonography	Non-invasive and real-time approach [45]	Inherent artifacts and speckle noise [47]
	Recommended for routine checkups during pregnancy [45]	Unclear tumor contour due to shadowing effect [45, 47]
	No potentially harmful ionizing radiation [46]	Comparatively poor image quality [45]
MRI	Non-invasive method for high risk patients [45]	Comparatively expensive [46]
	Identify suspicious areas more accurately [46]	Not recommended during pregnancy [45]
	No potentially harmful ionizing radiation [45]	May create allergy due to contrast media [45]
Histopathology	Can diagnose various types of cancers [45]	An invasive approach [45]
	Highly accurate methodology [45]	High expertise is required to analyze microscopy images [45]
	Provide comprehensive analysis of tissues [45]	Extensive care is needed during a biopsy procedure [45, 47]

To summarize, it is considered a gold-standard approach for the diagnosis of breast cancer. It has the potential to characterize different types of breast tumors that might not be possible with radiology imaging such as mammography, sonography, and MRI. However, morphological criteria used by pathologists to delineate different breast lesions could be subjective and may lead to biased results [8, 43]. An example of normal, benign, in situ carcinoma, and invasive carcinoma microscopy images are outlined in Figure 2.6. Likewise, the strengths and limitations of histopathology are given in Table 2.1. Whereas, its publicly available datasets are provided in Table 2.2.

Table 2.2 The list of published datasets related to mammography, sonography, magnetic resonance imaging, and histopathology imaging in breast cancer.

Imaging modality	Dataset	Number of images/cases	Year	Link
Mammography	MIAS [3]	322	1994	https://www.repository.cam.ac.uk/handle/1810/250394
	Mini-MIAS [3]	322	1994	http://peipa.essex.ac.uk/info/mias.html
	DDSM [48]	2620	2001	http://www.eng.usf.edu/cvprg/mammography/database.html
	Mammographic mass [49]	961	2007	http://archive.ics.uci.edu/ml/datasets/Mammographic+Mass
	BCDR [50]	FM: 1010, DM: 724	2012	https://www.bcdr.eu/
	INBreast [51]	410	2012	http://medicalresearch.inescporto.pt/breastresearch/GetINbreastDatabase.html
	CBIS-DDSM [52]	6671	2017	https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=22516629
	OPTIMAM [53]	172282	2020	https://medphys.royalsurrey.nhs.uk/omidb/
	CDD-CESM [54]	2006	2022	https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=109379611
Sonography	OASBUD [55]	100	2017	http://bluebox.ippt.gov.pl/~hpiotrzk/index.html
	UDIAT [56]	163	2017	http://www2.docm.mmu.ac.uk/STAFF/m.yap/dataset.php
	BUSI [4]	780	2020	https://scholar.cu.edu.eg/?q=afahmy/pages/dataset
	BUSIS [57]	562	2022	http://cvprip.cs.usu.edu/busbench
MRI	RIDER [58]	10	2008	https://wiki.cancerimagingarchive.net/display/Public/RIDER+Breast+MRI
	TCGA-BRCA [59]	84	2016	https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=19039112
	Duke MRI [60]	922	2018	https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70226903
Histopathology	WDBC [61]	569	1995	https://archive-beta.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
	TCGA-BRCA [62]	1062	2013	https://portal.gdc.cancer.gov/projects/TCGA-BRCA
	MGH-BIDMC [63]	167	2014	https://datadryad.org/stash/dataset/doi:10.5061/dryad.pv85m
	BCBH [64]	249	2015	https://rdm.inesctec.pt/dataset/nis-2017-003
	BreakHis [65]	7909	2015	https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis
	CAMELYON17 (train) [66]	500	2017	https://camelyon17.grand-challenge.org
	HASHI [67]	389	2018	https://datadryad.org/stash/dataset/doi:10.5061/dryad.1g2nt41
	PUIH [68]	3771	2018	http://ear.ict.ac.cn/?page_id=1616
	BACH (train) [69]	Part A: 400, Part B: 30	2019	https://iciar2018-challenge.grand-challenge.org
	BRACS [70]	WSI: 547, ROI: 4539	2022	https://www.bracs.icar.cnr.it/

2.3 Data collection

We conducted the current systematic review in accordance with the PRISMA statement [7], as illustrated in Figure 2.7 and explained in the following subsections.

2.3.1 Searched databases

We performed a literature search using the web of science by considering widely used publishers including ACM Digital Library, Elsevier, Hindawi, IEEE, Oxford University Press, MDPI, Nature Portfolio, Public Library of Science, Springer Nature, and Wiley Online Library.

2.3.2 Search terms

We used the search terms such as "breast cancer machine learning", "breast cancer deep learning", "breast carcinoma machine learning", "breast carcinoma deep learning", "breast malignancy machine learning", and "breast malignancy deep learning". We further refined our search using different filters as follows: We first selected publication years from 2016-2022 together with document type as an article or proceeding paper. We then chose topics of breast cancer scanning, computer vision and graphics, artificial intelligence, and machine learning along with different categories including engineering electrical electronic, engineering biomedical, computer science information systems, computer science interdisciplinary applications, computer science artificial intelligence, computer science theory methods, telecommunications, radiology nuclear medicine medical imaging, imaging science photographic technology, medical informatics, and mathematical computational biology. This search yielded a total of 1945 papers. The duplicated and retracted publications were then eliminated using the Zotero software [71], retaining 1470 papers.

2.3.3 Inclusion criteria

Titles and abstracts contain crucial information to be considered to select papers meeting desired criteria. To that end, we searched for supervised ML and DL methodologies using mammography, sonography, MRI, and histopathology imaging for the detection, segmentation, and classification of breast cancer. We organized the bibliography using the Zotero software [71]. These criteria reduced the number of relevant papers to 552 for full-text reading. Whereas, the finalized 142 papers analyzed in this systematic review are portrayed in Figure 2.8. Additionally, these studies are depicted in Figure 2.9 and illustrated in Figure 2.10.

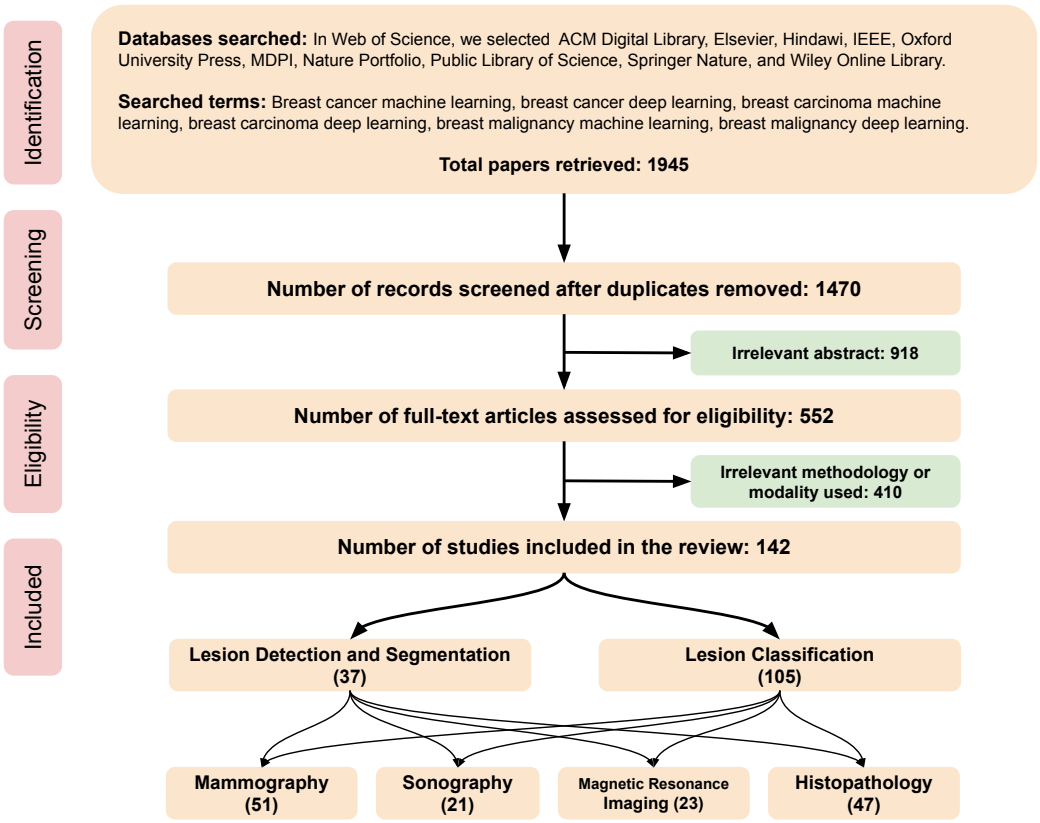


Figure 2.7 The completed flow diagram of systematic review based on PRISMA statement [7].

2.3.4 Exclusion criteria

We excluded papers not relevant to the detection, segmentation, and diagnosis of breast cancer using mammography, sonography, MRI, and histopathology. Moreover, we eliminated articles that generalized datasets; for instance, papers that used datasets relating to breast cancer as well as lung cancer, prostate cancer, and colorectal cancer. Similarly, we removed articles that were ambiguous about datasets, for instance, works that did not cite the dataset. Furthermore, we precluded articles that utilized artificial intelligence approaches other than the supervised ML and DL models. Additionally, we excluded those focused on BI-RADS classification. We also excluded those focused on the molecular classification of breast cancer. Lastly, we excluded the ones used for the proliferation and prognosis of breast malignancy.

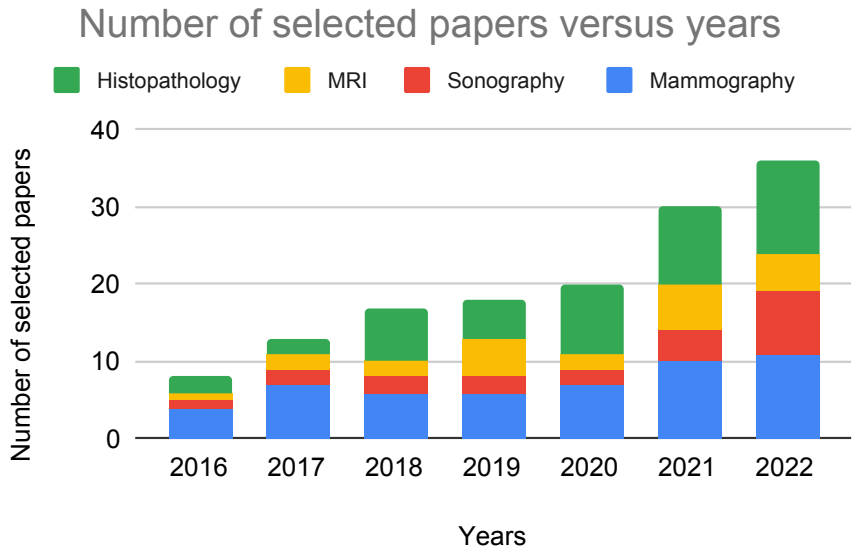


Figure 2.8 The distribution of articles analyzed per year.

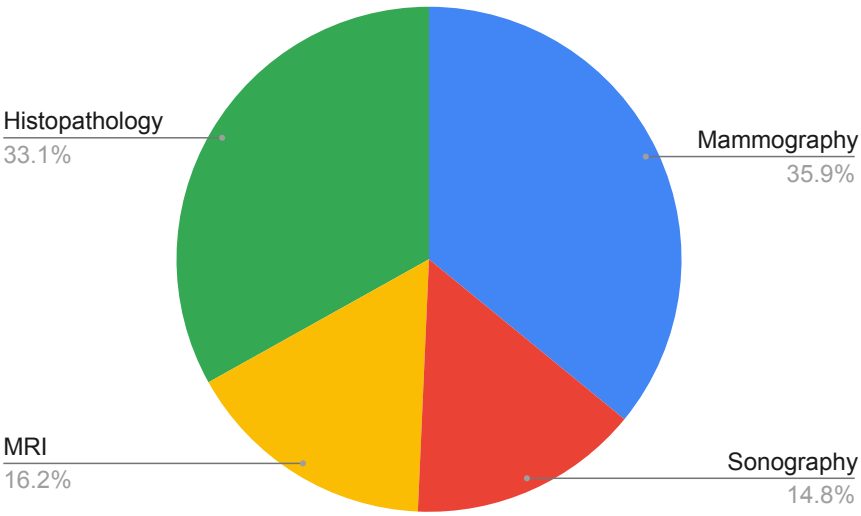


Figure 2.9 The pie chart of articles analyzed per modality.

2.4 Machine and deep learning approaches

This section briefly discussed supervised ML and DL models used in radiology and histopathology imaging for computer-aided breast cancer diagnosis.

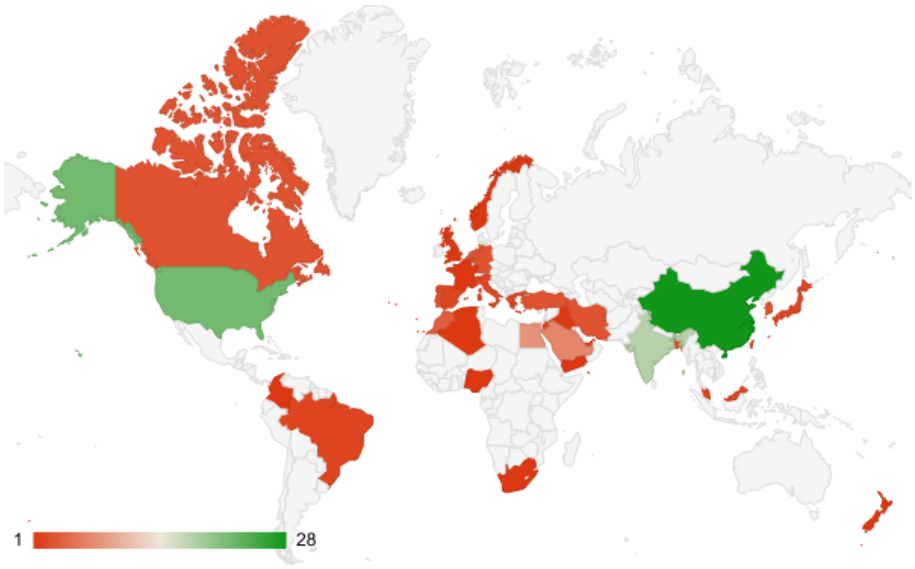


Figure 2.10 The map of articles analyzed per country.

2.4.1 Machine learning techniques

The term ML was defined as a sub-field of artificial intelligence by Arthur Samuel in 1959 [72]. A typical supervised ML approach includes preprocessing, lesion detection and segmentation, feature extraction, and lesion classification, as shown in Figure 2.2. A variety of ML methods have been proposed to detect, segment, and classify breast malignancy. The detection task is performed usually by scanning the whole image using a multiscale sliding window strategy [73]. Similarly, the classical segmentation task is typically performed using threshold-based segmentation, region-based segmentation, and edge-based segmentation methods [74]. Numerous features such as geometric (shape and margin), intensity or density (first-order radiomics), and texture (second-order radiomics) characteristics are then extracted from segmented areas. The geometric features include shape features, Zernike moments, and Fourier descriptors [75]. Similarly, intensity or density features include mean, variance, median, standard deviation, skewness, kurtosis, entropy, correlation, energy, uniformity, homogeneity, and smoothness among others [75]. Likewise, texture features include gray-level co-occurrence matrix (GLCM), gray-level run length matrix (GLRLM), gray-level size zone matrix (GLSZM), neighborhood gray-tone difference matrix (NGTDM), local binary pattern (LBP), scale-invariant feature extraction (SIFT), local quantization, threshold adjacent statistics (TAS), and parameter-free threshold adjacent statistics (PFTAS) among others [65, 75]. Finally, these features are used to train supervised ML classifiers. Some examples of widely followed super-

vised ML classifiers include decision tree (DT) [76], random forest (RF) [77], k-nearest neighbor (KNN) [78], Naive Bayes (NB) [79], and support vector machine (SVM) [80]. Moreover, the ensemble ML techniques include adaptive boosting (AdaBoost) classifier [81] and gradient boosting machine (GBM) [82] along with its effective variants like extreme gradient boosting (XGBoost), categorical boosting (CatBoost), and light GBM (LightGBM) classifiers. Further details on the above-mentioned traditional ML methodologies can be found in their respective papers.

However, traditional ML approaches mainly rely on feature selection which might lead to biased results. To that end, the DL models are progressively being followed in the automatic diagnosis of breast cancer.

2.4.2 Deep learning models

Recently, DL models have made remarkable advances in computer vision, particularly in biomedical image processing, because of their capability to automatically learn advanced characteristics from input images [19]. Particularly, convolutional neural networks (CNNs) are widely used in image-related tasks due to their ability to effectively share parameters across different layers within a DL model [83]. Consequently, various CNN-based architectures have been proposed during the past few years for the detection, segmentation, and classification of breast lesions. For lesion detection, the state-of-the-art DL frameworks mainly exploited region-based CNN (R-CNN) [84], Fast R-CNN [85], Faster R-CNN with region proposal network (RPN) [86], and you look once (YOLO) [87] architectures. Similarly, for lesion segmentation, the novel DL frameworks mostly employed fully convolutional network (FCN) [88], U-Net [89], SegNet [90], Mask R-CNN [91], DeepLab [92], and V-Net [93] architectures. Finally, for lesion classification, the cutting-edge DL frameworks commonly leveraged AlexNet [14], VGGNet [9], Inception [94], ResNet [95], Xception [16], DenseNet [96], and EfficientNet [97] architectures. Further details on the aforementioned models using deep CNN can be found in their respective papers.

2.4.3 Performance evaluation

The classification performance of state-of-the-art ML and DL models is mainly evaluated on the elements of the confusion matrix, also called contingency table [43, 6]. The table is comprised of four terms, namely, True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). In the context of breast cancer, the TP refers to images correctly predicted as cancerous and the FP represents non-cancerous

images mistakenly predicted as cancerous. Whereas the FN represents cancerous images mistakenly predicted as non-cancerous, and the TN refers to images correctly predicted as non-cancerous. These four terms lead us to sensitivity or recall, specificity, positive predictive value (PPV) or precision, negative predictive value (NPV), accuracy, and F1-measure [98]. Furthermore, the area under the receiver operating characteristic (ROC) curve is usually utilized to evaluate the performance of classification models [99]. Similarly, the mean average precision (mAP) is typically used for detection tasks. Whereas, the Jaccard index or intersection over union (IoU) and the Dice similarity coefficient (DSC) are normally followed for segmentation tasks [99]. The above-mentioned terminologies are mathematically defined as follows.

- Sensitivity: The sensitivity, also called recall or true positive rate, evaluates the correctly predicted proportion of actual positive images, as given in equation 4.3.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2.1}$$

- Specificity: The specificity, also known as true negative rate, assesses the correctly predicted proportion of actual negative images, as provided in equation 2.2.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2.2}$$

- PPV: The PPV, also called precision, analyzes the correctly predicted proportion of total predicted positive images, as given in equation 2.3.

$$\text{PPV} = \frac{TP}{TP + FP} \tag{2.3}$$

- NPV: The NPV evaluates the correctly predicted proportion of the total predicted negative images, as given in equation 2.4.

$$\text{PPV} = \frac{TN}{TN + FN} \tag{2.4}$$

- Accuracy: The accuracy examines a proportion of correctly predicted images among the total actual images, as stated in equation 4.4.

$$\text{Accuracy} = \frac{TP + FN}{TP + TN + FP + FN} \tag{2.5}$$

- F1-measure: The F1-measure, also called F1-score, estimates a harmonic average between recall and precision, as given in equation 2.6.

$$\text{F1-measure} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.6}$$

- ROC Curve: The ROC curve illustrates a relationship between true positive rate (TPR) and false positive rate (FPR) at different threshold values. The TPR is also known as sensitivity or recall, whereas FPR is equal to 1-specificity. The ROC curve shows that increasing TPR results in rising the FPR and vice versa. The TPR and FPR can be mathematically calculated using equations 4.3 and 4.6, respectively. The AUC is the area under the ROC curve and is a single-value metric.

$$\text{FPR} = \frac{FP}{FP + TN} \tag{2.7}$$

- PR Curve: The PR curve depicts an inverse relationship between precision and recall at different thresholds. A PR curve illustrates that a higher precision value results in a lower recall score and vice versa. The precision and recall can be mathematically computed using equations 4.3 and 2.3, respectively. The AP is the area under the PR curve, and mAP is the average of AP over all the classes. The mAP can be mathematically computed using the equation 2.8, where N is the total number of classes.

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{2.8}$$

- IoU: The IoU, also known as the Jaccard index, is an F-measure based metric that assesses the degree to which the predicted mask matches the ground truth, as given in equation 2.9.

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{2.9}$$

- DSC: The DSC is also an F-measure based metric that estimates the extent to which a predicted mask matches the ground truth, as stated in equation 2.10. The difference between IoU and DSC is that the former penalizes more than the latter for both under- and over-segmentation [99].

$$\text{DSC} = \frac{2TP}{2TP + FP + FN} \tag{2.10}$$

2.5 Discussion

In this section, we provided answers to the research questions, *RQ3* and *RQ4*. To that end, we analyzed supervised ML and DL models for the detection, segmentation, and classification of breast cancer from January 2016 to December 2022, as discussed in the succeeding subsections.

2.5.1 ML and DL in breast cancer detection and segmentation

In this section, we analyzed supervised ML and DL CADe systems used to detect and/or segment breast lesions using mammography, sonography, MRI, and histopathology images.

ML and DL in breast cancer detection and segmentation using mammography

Numerous studies proposed supervised ML and DL models to detect and/or segment breast lesions using mammography, as given in Table 2.3. For instance, Hu et al. [100] designed a new system for detecting masses in digital mammograms using a visual saliency model and CNN features. First, a visual saliency model was applied to mammograms to highlight masses. Then, the image was segmented into sub-images using a sliding window. Next, features were extracted from the sub-images using AlexNet as a feature extractor followed by the SVM classifier. Finally, a sliding window fusion algorithm was applied to complete the detection. The proposed detection system achieved a mean sensitivity of 0.940 and an average false positive per image (FPI) of 3.7 on the DDSM dataset. Taheri et al. [101] presented a new method to classify mammograms into normal and abnormal classes using SVM classifier. This process included denoising using an adaptive median filter, removing unwanted objects using Harris corner detection (HCD), feature extraction, and classification using the SVM. Three features were used to train the SVM: an intensity value, an auto□correlation matrix value of detected corners, and the energy. This approach resulted in a recall of 92.5% and a precision of 96.8% using the DDSM dataset. Kooi et al. [102] compared a deep CNN with a reference CAD system based on manually created features for detecting malignant lesions. It was found that CNN outperformed the CAD system at low sensitivity, whereas at high sensitivity its performance was comparable. Also, the CNN model along with manually designed features achieved the highest AUC of 0.941 on a larger private dataset of 63,262 FFDM images. In addition, the performance of CNN was equivalent to that of experienced radiologists. Mordang et

Table 2.3 ML and DL in detection and segmentation of breast cancer using mammography during 2016-2022

Author	Year	Country	Type	Dataset	Task	Technique	Optimal Results
Hu et al. [100]	2016	China	Conference	DDSM	Detection	AlexNet and SVM	Sensitivity: 0.940
Taheri et al. [101]	2016	USA	Conference	DDSM	Detection	SVM	Recall: 92.50%, Precision: 96.80%
Kooi et al. [102]	2017	Netherlands	Journal	Private FFDM	Detection	CNN	AUC score: 0.941
Mordang et al. [103]	2017	Netherlands	Journal	Private FFDM	Detection	LDA, RF, and SVM	AUC score: 0.937
Al-masni et al. [104]	2018	Republic of Korea	Journal	DDSM	Detection	YOLO	Accuracy: 99.70%
Ting et al. [105]	2019	Malaysia	Journal	Mini-MIAS	Detection	CNNI-BCC based on CNN	Accuracy: 90.50%, Sensitivity: 89.47%, Specificity: 90.71%, AUC Score: 0.901
Oliveira et al. [106]	2019	Portugal	Conference	INbreast	Detection and Segmentation	VGG16 and ResNet50	Sensitivity: 0.850, DSC: 0.830
Agarwal et al. [107]	2020	Spain	Journal	OMI-DB and INbreast	Detection	Faster R-CNN	Sensitivity: 0.760, Specificity: 0.880, AUC score: 0.870
Al-Antari et al. [108]	2020	Republic of Korea	Journal	DDSM and INbreast	Detection	YOLO, ResNet50, and InceptionResNet-v2	Accuracy: 99.17%
Hassan et a. [109]	2020	Egypt	Journal	Private FFDM, MIAS, INbreast, CBIS-DDSM	Detection and Classification	ALexNet and GoogLeNet	Accuracy: 95.50%
Pérez-Benito et al. [110]	2020	Spain	Journal	Private FFDM	Segmentation	ECNN based on CNN	DSC FGT: 0.770
Viegas et al. [111]	2021	Portugal	Journal	INbreast	Detection and Segmentation	Mask R-CNN	Sensitivity: 0.909, DSC tumor: 0.891
Yan et al. [112]	2021	France	Journal	CBIS-DDSM and INbreast	Detection	YOLO-v3 with VGG16	Sensitivity: 0.960
Hamed Aly [113]	2021	Egypt	Journal	INbreast	Detection	YOLO-v3	Accuracy: 89.40%
Kulkarni and Rabidas [114]	2022	India	Journal	DDSM	Detection	SqueezeU-Net based on U-Net	Accuracy: 90.81%, Sensitivity: 89.83%
Ramesh et al. [115]	2022	India	Journal	MIAS	Segmentation	Modified GoogLeNet, SVM, DT, RF, and NB	Mean IoU: 89.11%, DSC tumor: 82.15%
Dap and Jha [116]	2022	India	Conference	INbreast	Segmentation	U-Net and BCDU-Net	Mean IoU: 0.810, DSC tumor: 0.872

al. [103] proposed a CADe framework aimed to improve the detection of microcalcification lesions by reducing obvious false positives (OFP) using mammograms. The proposed architecture incorporated three independent-level classifiers to determine the final suspicious score. The microcalcification classifier is used to calculate the calci-

fication score at the most basic level. The breast arterial calcification (BAC) classifier is used to compute the groups of microcalcification at the middle level. The detection error classifier is used to compute the examination characteristics at the highest level. Finally, suspicious scores were calculated using the output of these three classifiers. To categorize malignant and OPFs, the SVM with linear kernel provided the best classification performance. It yielded an AUC of 0.937 on 80 examinations having 158 groups of malignant microcalcification. Al-masni et al. [104] designed a CAD system for the simultaneous detection and classification of breast masses in mammograms using the YOLO network. First, a set of randomly selected 600 images was rotated three times to create an augmented dataset of 2400 images with an equal proportion of benign and malignant cases. Next, the preprocessing phase included peripheral density correction and dataset normalization to 0 and 1. After that, the YOLO network was employed for feature selection, and masses were subsequently detected using a confidence score. Lastly, detected masses were classified as benign or malignant using two fully connected layers. The proposed model YOLO-based CAD system offered a detection accuracy of 99.70%. In addition, it attained a classification accuracy of 97.00%. Ting et al. [105] presented a novel framework called CNN improvement for breast cancer classification (CNNI-BCC) to detect and classify mammographic lesions into healthy, benign, and malignant. It was composed of feature-wise data augmentation (FWDA), CNN-based classification (CNNBS), and interactive detection-based lesion locator (IDBLL). Initially, the original images were labeled and ROI patches were extracted. During FWDA, each patch was rotated and flipped to generate eight 128×128 pixels image patches. Next, CNNBS was trained on the augmented ROI patches to classify them into their respective categories. Finally, the IDBLL was able to detect and classify breast lesions using the predicted bounding boxes. The proposed CNNI-BCC network offered accuracy of 90.50%, sensitivity of 89.47%, specificity of 90.71%, and an AUC of 0.901 on the mini-MIAS dataset. Oliveira et al [106] designed a lightweight CNN-based framework to detect and segment breast lesions in mammograms. During preprocessing, data augmentation was performed using affine transformations and cropping. The ResNet50 model was then fine-tuned to generate the region proposals. Moreover, the VGG16 network was utilized to reduce the false positive (FP) by classifying region proposals as background and mass. Finally, a graph-based approach was followed for the contour refinement of lesions. The propounded system was able to detect masses with sensitivity of 0.850 and to segment with a DSC value of 0.830 using the INbreast dataset. Agarwal et al. [107] proposed an automatic framework for detecting benign and malignant masses in FFDM images using the Faster R-CNN

object detection model. It utilized mammograms obtained from different scanners in the OMI-DB and INbreast datasets. The OMI-DB containing FFDM images acquired with Hologic and General Electric scanners were utilized in this study called OMI-H and OMI-G datasets, respectively. During preprocessing, images were downsampled for computational limitations and normalized for intensity standardization. Initially, the Faster R-CNN pretrained on natural images was fine-tuned on the OMI-H dataset to detect masses in whole mammograms. Subsequently, transfer learning was adopted to fine-tune the Faster R-CNN pretrained on the large OMI-H dataset to detect masses in smaller OMI-G and INbreast datasets. It yielded sensitivity of 0.760, specificity of 0.880, and an AUC of 0.870 on the OMI-G dataset. Whereas, it obtained an AUC of 0.900 on the INbreast dataset. Al-Antari et al. [108] proposed a CAD system based on DL to detect (and classify) breast lesions using mammograms. During preprocessing, images were treated for peripheral density correction and contrast enhancement. Also, data augmentation was applied by rotating, flipping, and scaling the mammograms. In the first step, the YOLO detector was practiced to detect breast lesions from whole mammograms. In the second step, regular CNN, ResNet50, and InceptionResNet-v2 networks were employed to classify FFDM images into benign and malignant. On the one hand, the YOLO model achieved a detection accuracy of 99.17% and 97.27% using DDSM and INbreast datasets, respectively. On the other hand, InceptionResNet-v2 outperformed regular CNN and ResNet50 models in the classification task. It yielded accuracy of 97.50% on the DDSM dataset. Whereas, it attained accuracy of 95.32% on the INBrest dataset. Hassan et al. [109] designed an automatic CAD system to detect (and classify) breast masses using mammograms. Initially, original images were processed to produce their enhanced version. Then, the maximally stable extremal regions (MSER) detector was used to detect breast masses in both original and enhanced images. Next, a feature-matching process was applied between these regions to detect the mass areas. Following that, patches of detected masses with a size of 200×200 pixels were extracted. The MSER achieved an average detection accuracy of 95.50% using mammograms of four datasets, including a private dataset, DDSM, MIAS, and INbreast datasets. Finally, mass images were resized to train the pretrained AlexNet and GoogLeNet models to categorize them as benign or malignant. It was discovered that the fine-tuned AlexNet achieved a maximum accuracy of 97.89% with an AUC of 98.32% on a private dataset having 95 mammograms. Whereas, it attained accuracy of 98.53% with an AUC of 98.95% on the MIAS dataset. Pérez-Benito et al. [110] designed an Entirely CNN (ECNN) model to segment fibroglandular tissue (FGT) in mammograms. During preprocessing, breast regions were detected, and pectoral

muscles were excluded. A histogram-based approach was then adopted to normalize mammograms acquired with eleven distinct devices. To generate the ground truth, two experienced radiologists segmented breast tissues and FGT in FFDM images using semiautomatic tools. Lastly, FGT were segmented using the proposed nine-layered ECNN model. It was discovered that normalizing gray-level values from different acquisition devices improved the performance of the model. This method procured a mean DSC score of 0.770 to segment FGT using a private dataset of 6680 mammograms obtained from 1785 subjects. Viegas et al. [111] utilized Mask R-CNN to detect and segment masses in mammography images by considering random and case-wise partitioning of the dataset. During preprocessing, images were cropped, normalized, and converted from grayscale to pseudo-color format. The Mask R-CNN was then trained on pseudo-color mammograms to detect and segment the masses. It was found that case-wise portioning produced more reliable results than random partitioning, with sensitivity of 0.909 for mass detection and a DSC value of 0.891 for mass segmentation on the INbreast dataset. Yan et al. [112] proposed a multi-task framework based on YOLO-v3 that considered both the CC and MLO views of mammograms to detect (and classify) breast masses. It was deduced that the proposed dual-view mass matching methodology outperformed conventional single-view mammograms. It offered a detection sensitivity of 0.960 and a classification accuracy of 0.879 on the INbreast dataset. Hamed Aly [113] leveraged YOLO-v3 to detect (and classify) breast masses using FFDM images. During preprocessing, images were scaled to the range of 0 to 255, followed by their normalization between 0 and 1. Furthermore, images were resized and augmented during the training process. The proposed model obtained a detection accuracy of 89.40% and a classification accuracy of 95.50% on the INbreast dataset. Kulkarni and Rabidas [114] proposed a modified U-Net called SqueezeU-Net to detect (and classify) benign and malignant calcification lesions using FFDM images. During preprocessing, images were resized to 128×128 pixels, normalized, and augmented. The U-Net and SqueezeU-Net were then utilized to detect as well as classify the calcification lesions. It was discovered that the SqueezeU-Net surpassed the U-Net network. Specifically, for the detection task, it procured accuracy of 90.81% and sensitivity of 89.83% on the DDSM dataset. Whereas, for the classification task, it offered accuracy of 97.30% and sensitivity of 97.37% on the same dataset. Ramesh et al. [115] proposed a DL architecture based on DCNN to segment and classify benign and malignant lesions using FFDM images. During processing, images were enhanced using the CLAHE method. The propounded model based on modified GoogLeNet was then trained to segment the suspected regions. After that, twenty-four features related

Table 2.4 ML and DL in detection and segmentation of breast cancer using sonography during 2016-2022

Author	Year	Country	Type	Dataset	Task	Technique	Optimal Results
Almajalid et al. [117]	2018	USA	Conference	Private sonograms	Segmentation	Modified U-Net	DSC tumore: 0.825
Xu et al. [118]	2019	China	Journal	Private sonograms	Segmentation	CNN	mean IoU: 85.10%
Zhou et al. [119]	2021	China	Journal	Private ABUS	Detection	Faster R-CNN	Recall: 95.06%, Precision: 74.05%
Daoud et al. [120]	2022	Jordan	Journal	Private sonograms and BUSI	Detection	DexiNed edge-detection model	Recall: 0.900, Precision: 0.910, F1-score: 0.900
Micheal et al. [121]	2022	China	Journal	Private sonograms	Segmentation	U-Net	DSC tumor: 99.62%, Accuracy: 98.15%
Podda et al. [122]	2022	Italy	Journal	BUSI and OASBUD	Segmentation	U-Net, VGG19, ResNet50, Inception-v3, and Xception	Mean IoU: 76.23%, DSC tumor: 82.60%

to shape and texture were extracted to train SVM, DT, RF, and NB classifiers. For the segmentation task, it procured a mean IoU of 89.11% and DSC of 82.15% on the MIAS dataset. Whereas, for classification, it attained accuracy of 99.12% and sensitivity of 99.89% on the same dataset. Dap and Jha [116] exploited the U-Net and BCDU-Net models to segment breast lesions in FFDM images. During preprocessing, the ROI images were extracted and resized to 224×224 pixels. Furthermore, the images were augmented using geometric transformations. The U-Net and BCDU-Net were then employed to segment the breast lesions. It was found that the BCDU-Net outran the U-Net, yielding a mean IoU of 0.810 and a DSC value of 0.872 on the INbreat dataset.

ML and DL in breast cancer detection and segmentation using sonography

Various studies presented supervised ML and DL frameworks to detect and/or segment breast malignancy using sonography, as provided in Table 2.4. For example, Almajalid et al. [117] introduced a novel framework based on the U-Net model to segment benign and malignant lesions using sonograms. During preprocessing, noise speckles were reduced using the speckle reducing anisotropic diffusion (SRAD) method, followed by contrast enhancement using histogram equalization. Moreover, data augmentation was performed by applying geometric transformations. The modified U-Net obtained a DSC of 0.825 on a private dataset with 221 sonogram images. Xu et al. [118] designed a new CNN-based model to segment the FGT, mass, skin, and fatty tissues in 3D breast ultrasound images. Initially, manual segmentation and labeling were performed by experienced radiologists to generate the ground truth for training purposes. The designed model was composed of two modules called CNN-I and CNN-II. The CNN-I

took normalized slices of 128×128 pixels to perform pixel labeling in three orthogonal ultrasound image planes. For each plan, it generated four outputs that corresponded to the aforementioned four functional tissues. Next, CNN-II took these twelve values from three plans to segment the FGT, mass, skin, and fatty tissues. It yielded a mean IoU score of 85.50% on a private dataset of 3D sonograms from 21 subjects. Zhou et al. [119] proposed a 3D multi-view approach based on improved Faster R-CNN to detect tumors in ABUS volumes. Initially, an improved Faster R-CNN was empoyed to get 2D bounding boxes in each of axial, coronal and sagittal planes. A multi-view position analysis scheme was then introduced to final 3D bounding boxes. This framework achieved recall of 95.06% and precision of 74.05% on a private dataset having 75 ABUS volumes from 75 patients. Daoud et al. [120] proposed a DL framework to detect benign and malignant tumors in sonograms. During preprocessing, grayscale sonograms were first transformed to RGB images. Four different object-detection models were then employed to localize the ROI areas having tumors. Also, images were processed using the DexiNed edge-detection model to generate the edge map. The ROI areas together with edge maps were combined to select the ROI in sonograms. The proposed strategy achieved a recall of 0.900, precision of 0.910, and F1-score of 0.900 on a private dataset having 380 sonograms. Whereas, it offered a recall of 0.890, precision of 0.900, and F1-score of 0.88 on the BUSI dataset. Micheal et al. [121] utilized a modified U-Net model to segment benign and malignant lesions in sonograms. During preprocessing, ground truth ROI images were obtained using a binarization method based on marks drawn by an experienced radiologist. The proposed U-Net architecture was then trained on these images. It obtained a DSC of 99.62% and accuracy of 98.15% on a private dataset containing 620 sonograms. Podda et al. [122] designed a novel DL pipeline based on an ensemble of DCNN models to segment (and delineate) normal, benign, and malignant tissues using sonograms. An ensemble of U-Net with a backbone incorporating customized CNN, Inception modules with traditional CNN, VGG19, ResNet50, and DenseNet121 was used for segmentation. Whereas, an ensemble of ResNet50, Inception-v3, Xception, InceptionResNet-v2, and DenseNet201 was adopted for classification. For the segmentation task, it attained a mean IoU of 76.23% and DSC of 82.60% on the BUSI dataset. Whereas, for the classification task, it yielded accuracy of 91.14% on the same dataset.

ML and DL in breast cancer detection and segmentation using MRI imaging

Several studies introduced supervised ML and DL architectures to detect and/or segment using MRI, as given in Table 2.5. Dalmis et al. [123] suggested a new method to

Table 2.5 ML and DL in detection and segmentation of breast cancer using MRI during 2016-2022

Author	Year	Country	Type	Dataset	Task	Technique	Optimal Results
Dalmis et al. [123]	2017	Netherlands	Journal	Private DCE-MRI	Segmentation	U-Net	DSC breast: 0.944, DSC FGT: 0.850
Benjelloun et al. [124]	2018	Belgium	Conference	Private DCE-MRI	Segmentation	U-Net	Mean IoU: 76.14%
Zhang et al. [125]	2019	USA	Journal	Private precontrast MRI	Segmentation	U-Net	DSC breast: 0.860, DSC FGT: 0.830
El Adoui et al. [126]	2019	Belgium	Journal	Private DCE-MRI	Segmentation	U-Net and SegNet	Mean IoU: 74.14%
Jiao et al. [127]	2020	China	Journal	Private DCE-MRI	Segmentation and detection	U-Net++ for breast segmentation and Faster R-CNN for mass detection	DSC breast: 0.951, Sensitivity: 0.874
Ayatollahi et al. [128]	2021	Iran	Journal	Private Ultrafast DCE-MRI	Detection	Modified RetinaNet	Sensitivity: 0.950
Galli et al. [129]	2021	Italy	Journal	Private DCE-MRI	Segmentation	U-Net	DSC tumor: 70.37%
Zhang et al. [130]	2021	USA	Journal	Private DCE-MRI	Segmentation	U-Net	DSC breast: 0.970, DSC FGT: 0.950
Huo et al. [131]	2021	China	Journal	Private DCE-MRI	Segmentation	nnU-Net	DSC breast: 0.968, DSC FGT: 0.877
Guo et al. [132]	2022	China	Journal	Private MRI	Segmentation	CNN with SVM	DSC tumor: 0.930

automatically segment FGT of the breast in DCE-MRI images using the U-Net model. Two different approaches were followed to obtain three-class labels in DCE-MRI images, namely, nonbreast tissues, fat tissue inside the breast, and FGT inside the breast. In the first approach, two successive 2-class U-Nets were applied, the first was responsible for breast segmentation, and the second was for FGT segmentation within the obtained breast mask. In the second approach, the same task was performed using a single 3-class U-Net framework. For breast segmentation, the average DSC values resulting from a 3-class U-Net and 2-class U-Nets were 0.933 and 0.944, respectively. Furthermore, for FGT segmentation, the average DSC values resulting from 3-class and 2-class U-Nets were 0.850 and 0.811, respectively. These results were achieved on a private dataset of 66 DCE-MRI images. Benjelloun et al. [124] leveraged the U-Net architecture to automatically segment breast cancer tissues in DCE-MRI images. During preprocessing, images were first pretreated for bias field correction. Subsequently, ground truth segmentation was obtained for each MRI image with the help of an experienced radiologist by considering only the ROI of having a tumor. The proposed network achieved a mean IoU of 76.14% using a private dataset of 96 volumes of DCE-MRI images with 5452 slices.Whereas, for the classification task, it offered

accuracy of 88.60%, sensitivity of 95.30%, and AUC of 93.60%. Zhang et al. [125] leveraged U-Net to segment the FGT in precontrast breast MRI images. The goal of this work was to obtain three-class labels, which included nonbreast tissues, fat tissue within the breast, and FGT within the breast. The ground truth was initially generated using a template-based segmentation method. The first U-Net was then employed to separate the breast tissues from the entire image. The second U-Net, on the other hand, was utilized to segment the fat tissues and FGT within the obtained breast mask. The presented architecture yielded a mean DSC of 0.86 ($\pm$0.05) for breast segmentation and 0.83 ($\pm$0.06) for FGT segmentation on a private dataset having precontrast MRI images of 314 patients. El Adoui et al. [126] proposed two approaches for segmenting breast tumors in DCE-MRI images based on SegNet and U-Net. During preprocessing, images were first pretreated for bias field correction. Following that, ground truth segmentation was obtained for each MRI image with the help of an experienced radiologist by considering only the ROI of having a tumor. Furthermore, data augmentation was performed by scaling, flipping, and rotating the data. Next, two architectures were trained to segment the breast tumors based on SegNet and U-Net. This methodology yielded mean IoU scores of 68.88% and 76.14% using SegNet and U-Net, respectively, on a private dataset having 86 DCE-MRI volumes from 43 patients. Jiao et al. [127] proposed a framework based on deep CNNs to automatically segment breast regions and detect breast masses using DCE-MRI images. For the segmentation task, breast region labels were created which were then used to train the U-Net++ model. It assisted in removing interference from organs outside the breast area. Next, for the detection task, preprocessed images from the segmentation task were employed to train the Faster R-CNN to identify the locations of breast masses. For breast segmentation, the U-Net++ yielded a DSC of 0.951 and IoU of 0.908. Whereas, for mass detection, the Faster R-CNN offered sensitivity of 0.874 on a private dataset containing DCE-MRI images of 75 patients. Ayatollahi et al. [128] introduced a CADe model based on modified RetinaNet to detect benign and malignant lesions in ultrafast DCE-MRI images. During preprocessing, T1 weighted images were subjected to motion correction, temporal normalization, and cropping of the breast tissues. The proposed CADe framework yielded sensitivity of 0.950 on a private dataset having 572 lesions from 462 patients. Galli et al. [129] developed a DL pipeline based on the U-Net model to segment lesions in breast MRI images. The complete pipeline was composed of breast masking, motion correction, slice extraction, and lesion segmentation. Initially, a fully automated algorithm based on multiplanar 2D U-net was employed to extract the breast masks. Then, a 3D nonrigid intensity-based registration was used to reduce misalignment be-

Table 2.6 ML and DL in detection and segmentation of breast cancer using histopathology during 2016-2022

Author	Year	Country	Type	Dataset	Task	Technique	Optimal Results
Ni et al. [133]	2019	China	Conference	Private WSI	Segmentation	WSI-Net based on CNN and RF	Mean IoU: 71.84%
Patil et al. [134]	2020	USA	Conference	HASHI	Segmentation	RUBIC-Net based on U-Net	Mean IoU: 87.60%, DSC tumor: 89.45%
Li and Lu [135]	2021	China	Conference	Camelyon17	Segmentation	U-Net	DSC tumor: 0.846
Lu et al.[136]	2022	China	Journal	Camelyon17	Detection and Segmentation	YOLO-v4 and GCPANet	Recall: 0.680, F1-score: 0.787, DSC tumor: 0.852

tween a slice acquired at different times. Next, slice images were extracted using three temporal acquisitions: pre-contrast, two minutes after contrast agent injection, and six minutes following contrast contrast administration. Finally, the U-Net model was utilized to segment breast lesions. The propounded pipeline procured a DSC value of 70.37% on a private containing bilateral DCE-MRI images from 33 patients. Zhang et al. [130] exploited U-Net architecture with and without transfer learning to segment FGT in DCE-MRI images. Initially, the first U-Net was applied to segment breast tissues from the entire image. The second U-Net was then used to delineate FGT from fat tissues within the obtained breast mask. It was found that the U-Net benefited from transfer learning procured means DSC values of 0.970 and 0.950 for breast and FGT, respectively, on a private dataset containing 166 fat-sat and 286 non-fat-sat DCE-MRI studies. Huo et al. [131] employed nnU-Net to segment FGT in DCE-MRI images. During preprocessing, the pre-contrast T1-weighted images were normalized between 0 and 255. Two separate nnU-Net models were used to segment breast and FGT tissues. The proposed framework based on nnU-Net attained an average DSC of 0.968 for breast segmentation and a mean DSC of 0.877 for FGT segmentation on a private dataset having 100 DCE-MRI cases. Guo et al. [132] suggested a CNN-SVM model to segment tumors in MRI images. During preprocessing, two experienced radiologists manually annotated the tumor and nontumor regions. The CNN and SVM were trained in parallel for accurate segmentation. The proposed CNN-SVM acquired a DSC value of 0.930 on a private dataset with 272 MRI cases.

ML and DL in breast cancer detection and segmentation using histopathology

Many studies developed supervised ML and DL models to detect and/or segment breast lesions using histopathology, as provided in Table 2.6. Ni et al. [133] proposed WSI-Net, a novel network for automatically segmenting (and classifying) breast WSI im-

ages. The WSI-Net is composed of a baseline semantic segmentation model called DeepLab and an extra classification branch. During preprocessing, each WSI was binarized using the Otsu thresholding to remove non-tissue areas. Next, every WSI was divided into overlapped patches, which were fed into WSI-Net to perform pixel-wise prediction. The lower layer of DeepLab was used to identify and discard non-malignant patches, whereas the remaining potentially cancerous ones were forwarded to its higher layer to perform pixel-wise segmentation. The DeepLab model produced segmentation outputs of one-eighth of the input patches, which were then aggregated using majority voting to get a segmentation map for each WSI. Finally, eighteen morphological features were extracted from the segmented map of each WSI to train the RF classifier. The WSI-Net achieved a mean IoU of 71.84% and accuracy of 87.00% on a private dataset of 300 breast WSIs. Patil et al. [134] proposed a modified version of U-Net called RUBIC-Net to segment breast lesions using downsampled WSIs of breast cancer. It leveraged the pretrained ResNet152 network as a counterpart to the encoder and decoder in the U-Net model. Furthermore, dilated convolutions were used to effectively capture ROI features. During preprocessing, WSIs were downsampled to 320×320 pixels and subsequently normalized between 0 and 1. Likewise, the binary masks were scaled and normalized. The designed network was then trained on low-resolution WSIs to produce binary masks for breast lesions. The proposed RUBIC-Net outperformed U-Net and achieved an IoU of 87.60% and DSC of 89.45% on the HASHI dataset. Li and Lu [135] integrated the MobileNet-v2, ResNet101, and U-Net models to segment lesions in breast WSIs. During preprocessing, the Otsu algorithm was adopted to remove background areas. The MobileNet-v2 was then employed to filter out non-cancerous tissues. Next, the ResNet101 network was used to classify normal and malignant lesions, leading to a heatmap of cancerous tissues. Finally, U-Net was employed to segment cancerous regions within WSIs. The proposed approach using the U-Net model attained a DSC of 0.846 to segment tumor areas. Lu et al.[136] proposed an innovative DL framework to detect and segment breast lesions in WSI images. The recognition module based on the YOLO-v4 was used to quickly recognize the main lesions. Whereas the segmentation module based on the global context-aware progressive aggregation network (GCPANet) was employed to perform fine segmentation of the lesion regions. For the detection, the propounded approach attained sensitivity of 0.680 and an F1-score of 0.787 on the Camelyon17 dataset. Whereas, for the segmentation, it yielded a DSC value of 0.852 on the Camelyon17 dataset.

2.5.2 ML and DL in breast cancer classification

In this section, we analyzed ML and DL CADx systems used to classify breast lesions using mammography, sonography, MRI, and histopathology imaging.

ML and DL in breast cancer classification using mammography

Table 2.7 ML and DL in breast cancer classification using mammography during 2016-2022

Author	Year	Country	Type	Dataset	Task	Technique	Optimal Results
Arevalo et al. [137]	2016	Colombia	Journal	BCDR-FM	Classification	CNN	AUC score: 0.826
Jiao et al. [138]	2016	China	Journal	DDSM	Classification	CNN and SVM	Accuracy: 96.7%
Kashyap et al. [139]	2017	India	Conference	mini-MIAS	Classification	SVM	Accuracy: 94.21%
Guan et al. [140]	2017	USA	Conference	MIAS and DDSM	Classification	VGG16 with transfer learning	Accuracy: 0.950, AUC score: 0.971
Sonar et al. [141]	2017	India	Conference	MIAS and DDSM	Classification	Hybrid SVM-KNN	Accuracy: 100%
Hepsağ et al. [142]	2017	Turkey	Conference	mini-MIAS and BCDR	Classification	CNN	Accuracy: 0.8800, Precision: 0.8600, Recall: 0.9000, F1-score: 0.8800
Antropava et al. [143]	2017	USA	Journal	Private FFDM	Classification	VGG19 with SVM	AUC score: 0.862
Yeh and Chan [144]	2018	Taiwan	Conference	Private DBT	Classification	LeNet	Accuracy: 87.12%, Precision: 87.27%, Recall: 87.09%
Chougrad et al. [145]	2018	Morocco	Journal	DDSM, BCDR, INbreast, MIAS	Classification	VGG16, ResNet50, Inception-v3 transfer learning	Accuracy: 98.23%, AUC score: 0.990
Gao et al. [146]	2018	USA	Journal	Private CESM and IN-Brest	Classification	ResNet50 with GBT	Accuracy: 0.890, AUC score: 0.910
Hagos et al. [147]	2018	Netherlands	Conference	Private FFDM	Classification	VGGNet	AUC score: 0.933
Mohamed et al. [148]	2018	USA	Journal	Private FFDM	Classification	AlexNet	AUC score: 0.970
Li et al. [149]	2019	China	Journal	Private FFDM	Classification	DenseNet and DenseNetII	Accuracy: 94.55%, Sensitivity: 95.60%, Specificity: 95.36%
Laghmati et al. [150]	2019	Morocco	Conference	Mammographic mass	Classification	ANN, KNN, DT, and SVM	Accuracy: 0.840, Sensitivity: 0.860, Specificity: 0.820
Tsochatzidis et al. [151]	2019	Greece	Journal	CBIS-DDSM	Classification	AlexNet, VGG, GoogLeNet, and ResNet	Accuracy: 0.804

Table 2.7 continued from previous page

Author	Year	Country	Type	Dataset	Task	Technique	Optimal Results
Benzebouchi et al. [152]	2019	Algeria	Conference	DDSM	Classification	CNN	Accuracy: 97.89%, Sensitivity: 98.90%, Specificity: 96.90%, AUC score: 98.20%
Li et al. [153]	2020	China	Journal	Private FFDM and DBT	Classification	VGG16	Accuracy: 95.10%, Sensitivity: 70.80%, Specificity: 98.90%, AUC score: 0.910
Wessels and Haar [154]	2020	South Africa	Conference	MIAS and DDSM	Classification	CNN mini VGGNet and mini GoogLeNet	Accuracy: 66.16%, Precision: 67.63%, Recall: 66.84%, AUC score: 0.850
Saranyaraj et al. [155]	2020	India	Journal	DDSM	Classification	CNN based on LeNet	Accuracy: 96.23%, AUC score: 0.9842
Ragab et al. [156]	2021	Egypt	Journal	CBIS-DDSM and MIAS	Classification	ALexNet, GoogLeNet, ResNet18, ResNet50 and ResNet101	Accuracy: 97.90%, Sensitivity: 0.980, Specificity: 0.980, AUC score: 1.000
Heidari et al. [157]	2021	USA	Journal	Private FFDM	Classification	SVM with RPA	Accuracy: 75.2%, AUC score: 0.840
Malebary et al. [158]	2021	KSA	Journal	DDSM and MIAS	Classification	CNN and LSTM	Accuracy: 0.960, Sensitivity: 0.970, Specificity: 0.980, F1 score: 0.970
Lee et al. [159]	2021	USA	Journal	CBIS-DDSM	Classification	CNN	AUC score: 0.860
Tsochatzidis et al. [160]	2021	Greece	Journal	CBIS-DDSM	Classification	ResNet50 with U-Net	Accuracy: 0.776, AUC score: 0.862
El Houby et al. [161]	2021	Egypt	Journal	MIAS, DDSM, and INbreast	Classification	CNN	Accuracy: 96.52%, Sensitivity: 96.55%, Specificity: 96.49%, AUC score: 0.980
Song et al. [162]	2021	China	Journal	Private CESM	Classification	Res2Net50	Accuracy: 96.60%, Sensitivity: 96.40%, Specificity: 96.40%, AUC score: 0.966
Al-Fahaidy et al. [163]	2022	Yemen	Journal	MIAS	Classification	SVM	Accuracy: 87.10%, Sensitivity: 90.00%
Samee et al. [164]	2022	KSA	Journal	INbreast and Mini-MIAS	Classification	SVM, KNN, NB, and ensemble classifiers	Accuracy: 98.62%, Sensitivity: 98.28%
Marathe et al. [165]	2022	USA	Journal	Private FFDM	Classification	LightGBM	Accuracy: 0.530, Sensitivity: 1.000, AUC score: 0.730
Singh et al. [166]	2022	India	Journal	INbreast	Classification	KNN, SVM, DT, NB, RF, and ET	Accuracy: 90.40%, Sensitivity: 92.00%
Karthiga et al. [167]	2022	India	Journal	INbreast, DDSM, and MIAS	Classification	AlexNet, VGG16, and VGG19	Accuracy: 96.53%

Table 2.7 continued from previous page

Author	Year	Country	Type	Dataset	Task	Technique	Optimal Results
Hekal et al. [168]	2022	Egypt	Journal	CBIS-DDSM	Classification	AlexNet, ResNet50, ResNet101, and DenseNet201	Accuracy: 94.00%, Sensitivity: 92.00%
Alshammari et al. [169]	2022	KSA	Journal	Private FFDM	Classification	SVM, NB, DT, DA, and KNN	Accuracy: 100.00%
Song et al. [170]	2022	China	Journal	Private CESM and CDD-CESM	Classification	MDIB model based on ResNet18 and MLP	Accuracy: 97.18%, Sensitivity: 94.86%, AUC score: 0.973

Numerous studies proposed supervised ML and DL models to classify breast lesions using mammography, as given in Table 2.7. Arevalo et al. [137] designed a CNN-based architecture for the classification of film mammograms. The proposed CNN model outperformed hand□crafted features given by radiologists. Besides, the standalone model gained area under the ROC curve (AUC) of 0.822 and outperforms the histogram of oriented gradients (HOG) and histogram of gradient divergence (HGD) descriptors. Also, the combination of handcrafted and learned descriptors obtained an AUC of 0.826 on the BCDR-FM dataset. Jiao et al. [138] employed CNN to extract middle- and high-level features from mammograms, followed by two SVM classifiers. The outputs of two SVM models trained on different features were analyzed to find the consistency and inconsistency between their results. All consistent outcomes were considered correct. In inconsistent cases, gray information were used to calculate their benign and malignant classes. This method achieved accuracy of 96.7% on the DDSM dataset. Kashyap et al. [139] proposed a novel approach to classify mammograms by considering breast density. The images were enhanced using a fractional order differential-based filter, followed by their segmentation with the c-means clustering algorithm. Subsequently, the LBP and dominant rotated local binary pattern (DRLBP) features were extracted to train SVM with different functions. The SVM with DRLBP and RBF kernel functions offered the highest accuracy of 94.21% on the mini-MIAS dataset. Guan et al. [140] leveraged the VGG16 with feature extraction and fine-tuning strategies to classify mammograms. Following the preprocessing, one ROI was extracted from each mass-containing image. The grayscale mass images were subsequently converted to RGB images to train a newly developed CNN network, VGG16 as a feature extractor, and VGG16 as a fine-tuned model. It was inferred that VGG16 with feature extraction yielded accuracy of 0.906 on the MIAS dataset, whereas it achieved accuracy of 0.950 on the DDSM dataset. Sonar et al. [141] proposed a modified hybrid SVM-KNN model to classify breast mammograms into be-

nign and malignant. The preprocessing included noise removal, artifact elimination, and image enhancement. Next, the ROI images were extracted using c-means clustering and active contour techniques. Following that, the GLCM texture features were extracted and utilized to train the suggested hybrid SVM-KNN supervised classifier. This framework reached accuracy of 100% on the DDSM dataset. Hepsağ [142] developed a CNN model for the classification of mammography masses and calcification lesions separately. The preprocessing methods involved in this study were cropping, augmentation, and balancing the datasets. The designed model yielded accuracy of 0.87, precision of 0.78, recall of 0.90, and F1-score of 0.84 on the mass images of the mini-MIAS dataset. Similarly, it attained accuracy of 0.88, precision of 0.86, recall of 0.90, and F1-score of 0.88 using the mass images of the BCDR dataset. Antropava et al. [143] combined CNN features with handcrafted features, followed by SVM, to classify breast lesions of breast FFDM images. The grayscale ROI images obtained from the FFDM dataset were initially converted to RGB images by duplicating them across three channels. These ROI images were then fed into the VGG19 model, which generated five feature vectors that were then concatenated into a single feature vector. The fusion of CNN-based features and handcrafted features offered an AUC of 0.862 on 739 ROI images extracted from a private FFDM dataset of 245 breast lesions. Yeh and Chan [144] compared CNN-based CAD with feature-based CAD for breast cancer classification using DBT images. For both systems, the image processing step included noise removal, intensity conversion, morphological treatment, and ROI extraction. The CNN-based CAD achieved accuracy of 87.12% ($\pm$0.035) compared to a feature-based CAD with accuracy of 74.85% ($\pm$0.122) on a private dataset having DBT images of 20 practical cases. Chougrad et al. [145] developed a CAD system by exploring transfer learning of CNN models to classify mammography mass lesions. During preprocessing, fixed-size ROI patches were extracted followed by their global contrast normalization. Three state-of-the-art CNN models including VGG16, ResNet50, and Inception-v3 were fine-tuned to achieve optimal classification performance. The fine-tuned Inception-v3 outperformed the other two networks on all of the DDSM, BCDR, and INbreast datasets. Moreover, the inception-v3 model was fine-tuned on a larger dataset comprised of the above-mentioned three datasets and achieved accuracy of 98.23% with an AUC of 0.99 on the independent MIAS dataset. Gao et al. [146] suggested a shallow-deep CNN model using ResNet50 along with a gradient boosting tree to classify breast tumors into benign and malignant using CESM images. The shallow-deep CNN was employed to render the recombined images from low-energy images of the CESM. Whereas, the deep-CNN was utilized to extract novel

features from the low-energy and recombined images. During preprocessing, lesions were manually identified and tumor images were normalized between 0 and 1. These images were then resized to 224×224 pixels to train the designed framework. The proposed approach achieved accuracy of 0.85 with an AUC of 0.84 using only the low-energy images of the private CESM dataset. Whereas it provided accuracy of 0.89 with an AUC of 0.91 using both the low energy and recombined images of the private dataset of 49 CESM cases. Hagos et al. [147] proposed a novel CNN-based model using symmetry information to classify breast mammographic masses. The suspicious mass candidates were first detected automatically by analyzing local lines and gradient orientation features. Following that, images were resized to 300×300 pixels, and data augmentation was performed. The suggested symmetry model outperformed a baseline model similar to the VGG network. The designed system attained an AUC of 0.933 on a private dataset of 28,294 FFDM images. Mohamed et al. [148] employed an improved AlexNet network to categorize breast density using FFDM images. During preprocessing, histogram equalization was used to calibrate the contrast intensity of images, followed by normalization of the dataset. The modified AlexNet framework consisted of five convolutional layers, three maximum pooling layers, and three fully connected layers with a final two-way softmax function. In the first model, the two outputs corresponded to the categories of scattered density and heterogeneously density. Whereas, the two outputs of the second model denoted non-dense (fatty and scattered density) and dense (heterogeneous and extreme density) classes. The second model outperformed the first one, achieving an AUC of 0.97 for the MLO view on a private dataset of 15,415 FFDM images. Li et al. [149] designed an improved DenseNet model, known as DenseNet-II, to classify benign and malignant mammography images in an effective way. During preprocessing, images were normalized with zero-mean and unit standard deviation. Besides, images were enhanced using affine transformations and random cropping. The DenseNet-II model has an Inception module, three dense blocks, and transitional layers. It was found that the DenseNet-II framework with a few parameters outperformed AlexNet, VGG19, GoogLeNet, and DenseNet networks by achieving accuracy of 94.55%, sensitivity of 95.60%, and specificity of 95.36% on a private dataset having 2042 FFDM cases. Laghmati et al. [150] leveraged numerous ML techniques including ANN, KNN, DT, and SVM to classify breast mammographic masses as benign or malignant. Five attributes were considered at the input of each classifier to predict the severity of each breast mass. It was found that the ANN outperformed the others with accuracy of 0.84, sensitivity of 0.86, and specificity of 0.82 on the mammographic mass dataset. Tsochatzidis et al. [151] investigated the

effect of weight initialization on the performance of deep CNN models in classifying mammographic mass lesions. During preprocessing, mass images were extracted from the original mammograms. A total of eight deep CNN models were trained from scratch with random weights and fine-tuned with pretrained weights. It was found that models initialed with pretrained weights outperformed the networks initialized with random weights. Among the pretrained models, ResNet50 architecture yielded a highest accuracy of 0.804 on the CBIS-DDSM dataset. The proposed network achieved a segmentation DSC of 0.830 and a detection sensitivity of 0.850 on the INbreast dataset. Benzebouchi et al. [152] proposed a CNN-based architecture to classify mammography mass images as benign or malignant. The segmentation was manually performed to generate mass images which were employed to train the model. The proposed system yielded accuracy of 97.89%, sensitivity of 98.90%, specificity of 96.90%, and an AUC of 98.20% on the DDSM dataset. Li et al. [153] evaluated numerous deep CNN models to distinguish normal, benign, and malignant breast tissues using FFDM and DBT. During preprocessing, two experienced radiologists extracted 2D ROI patches from FFDMs as well as 2D and 3D ROI patches from DBT images. The data was augmented by rotating, flipping, and scaling the images. Four different studies were assessed with and without transfer learning using the VGG16 network. In addition, the combined effect of FFDM and DBT was explored in these four studies. It was concluded that transfer learning benefited mass classification for both the FFDM and DBT. Furthermore, it was found that combining FFDM and DBT improved classification accuracy. For malignant masses, the best model achieved accuracy of 95.10%, sensitivity of 70.80%, specificity of 98.90%, and an AUC of 0.910 using a private dataset containing 1854 2D and 3D ROI patches from FFDM and DBT images. Wessels and Haar [154] designed a CAD system based on deep CNN to categorize normal, benign, and malignant breast tissues using FFDM. During preprocessing, images were downsampled to 128×128 pixels, followed by their normalization. Two deep CNN architectures including mini VGGNet and mini GoogLeNet were leveraged to classify breast tissues into normal, benign, and malignant classes. It was noticed that the mini GoogLeNet provided a higher accuracy of 66.16%, a precision of 67.63%, a recall of 66.84%, and an AUC of 0.850 on the DDSM dataset. Saranyaraj et al. [155] developed a DCNN to classify mammograms as normal, benign, or malignant. During preprocessing, images were cropped to 200×200 pixels and de-noised to preserve the edge information. The proposed five-layered DCNN was trained by employing different sets of hyperparameters. It was found that the performance of the DCNN was sensitive to the choice of hyperparameters. The optimized DCNN acquired accuracy of 96.23% along with an AUC

of 0.9842 on the DDSM dataset. Ragab et al. [156] leveraged DCNN to classify mammograms into benign and malignant. During preprocessing, images were enhanced using the CLAHE method, and ROI patches were extracted. Subsequently, features were extracted using the AlexNet, GoogleNet, ResNet18, ResNet50, and ResNet101 architectures. These features were then fused to train the SVM classifier with different kernels. It was found that the SVM with quadratic kernel achieved maximum classification performance. It attained accuracy of 97.90%, sensitivity of 0.980, specificity of 0.980, and an AUC of 1.00 on the CBIS-DDSM dataset. Similarly, it procured accuracy of 97.40%, sensitivity of 0.990, specificity of 0.952, and an AUC of 1.00 on the MIAS dataset. Heidari et al. [157] investigated the effect of feature dimension reduction in optimizing ML models to classify mammographic lesions into benign and malignant. During preprocessing, breast masses were segmented and 181 features were extracted. Numerous feature reduction techniques were employed to train the SVM model. It was found that the random projection algorithm (RPA) outperformed the others by achieving accuracy of 75.2% and an AUC of 0.840 on a private dataset of 1487 FFDM images. Malebary et al. [158] proposed a CAD system based on CNN and LSTM to classify mammograms into normal, benign, and malignant. During preprocessing, lesions were segmented using k-means clustering to generate ROI images. Thereafter, features extracted with ResNet50 and LSTM were concatenated to train a CNN model. Lastly, the RF and XGBoost classifiers were followed to finalize the predictions. The proposed framework acquired accuracy of 0.960, sensitivity of 0.970, specificity of 0.980, and an F1-score of 0.970 on the DDSM dataset. Similarly, it offered accuracy of 0.950, sensitivity of 0.970, specificity of 0.970, and an F1-score of 0.980 on the MIAS dataset. Lee et al. [159] compared segmentation-free and segmentation-dependent CAD systems to classify mammograms as benign or malignant. It was noticed that the segmentation-free CNN outperformed segmentation-dependent models by procuring an AUC of 0.86 on the CBIS-DDSM dataset. Tsochatzidis et al. [160] integrated segmentation information into CNN to effectively classify breast lesions into benign and malignant. During preprocessing, ROI images of 1024×1024 pixels were extracted from original mammograms. Segmentation maps were acquired from either ground-truth or a modified U-Net model to train a modified ResNet50 model. The proposed strategy obtained accuracy of 0.776 and AUC of 0.862 using ground-truth segmentation maps, whereas it attained a maximum accuracy of 0.768 and AUC of 0.857 using a U-Net-based automatic segmentation on the CBIS-DDSM dataset. El Houby et al. [161] proposed a novel CNN model to classify mammography images into benign and malignant. During preprocessing, mammograms were denoised with a nonlinear filter,

enhanced with CLAHE, and ROI images were extracted. Moreover, ROI patches were resized to 208×208 pixels, and data augmentation was performed. The designed CNN model yielded accuracy of 96.52% and 95.30%, sensitivity of 96.55% and 98.00%, specificity of 96.49% and 92.6%, and AUC of 0.980 and 0.974 on the INbreast and MIAS datasets, respectively. Song et al. [162] proposed a new multiview multimodal network based on Res2Net50 to classify CESM images as benign or malignant. During preprocessing, images were denoised, followed by data augmentation and contract normalization. The proposed model was then trained simultaneously on low-energy and dual-energy subtracted images. It procured accuracy of 96.60%, sensitivity of 96.40%, specificity of 96.40%, and AUC of 0.966 on a private dataset having 760 CESM images from 95 patients. Al-Fahaidy et al. [163] proposed an ML model to classify breast masses as benign or malignant using FFDM images. The preprocessing step involved noise removal with median filtering, artifact suppression, and background separation. The ROI images of 64×64 pixels were then extracted using the seeded region growing technique. After that, shape-based, first-order, second-order, fractal dimension and wavelet features were extracted, and 307 effective features were identified using the sequential forward selection method. The suggested SVM model was trained on these optimal features which yielded accuracy of 87.10% and sensitivity of 90.00% on the MIAS dataset. Samee et al. [164] exploited DCCN together with LR and PCA to categorize FFDM images into benign and malignant. The preprocessing step included ROI extraction, data augmentation, and generating pseudo-color RGB images. These mammograms were then fed into the AlexNet, VGG16, and GoogLeNet to extract features. Following that, significant features were retrieved with PCA and LR models to train SVM, KNN, NB, and ensemble classifiers. This methodology attained accuracy of 98.62% and sensitivity of 98.28% on the INbreast and accuracy of 98.80% and sensitivity of 99.62% on the mini-MIAS dataset. Marathe et al. [165] leveraged an ML approach to distinguish amorphous calcification lesions as benign or malignant. During preprocessing, three different binary masks were generated from each input ROI image by employing a multiscale segmentation method. A set of local and global features were then extracted from these masks. The local features were then grouped together using the k-means clustering, and concatenated with global features to train the LightGBM classifier. This methodology yielded accuracy of 0.530, sensitivity of 1.00, and AUC of 0.730 on a private dataset with 276 ROI patches from 248 FFDM images. Singh et al. [166] developed an improved ML framework to delineate breast masses into benign and malignant using FFDM images. In the beginning, ROI areas with mass lesions were extracted. Then, a total of 125 geometric and textural features were re-

trieved from each ROI image. Next, the twenty most discriminatory features were used to train six classifiers, which included KNN, SVM, DT, RF, NB, and ET models. The results revealed that the KNN outperformed other classifiers with accuracy of 90.40%, sensitivity of 92.00%, and specificity of 88.00% on the INbreast dataset. Karthiga et al. [167] presented an innovative DCCN model to demarcate benign and malignant lesions using FFDM images. The preprocessing step involved contrast enhancement with CLAHE, data augmentation, segmentation with polynomial curve fitting, and ROI selection. It was found that the propounded framework surpassed fine-tuned AlexNet, VGG16, and VG19 networks, and yielded accuracy of 95.95% on the MIAS dataset, 99.39% on the DDSM, and 96.53% on the INbreast dataset. Hekal et al. [168] introduced an ensemble system based on DCCN models to classify benign and malignant lesions using FFDM images. During reprocessing, the extracted ROI images were converted to the suspected module region (SNR) using the Otsu thresholding method. The ensemble model comprising AlexNet, ResNet50, ResNet101, and DenseNet201 was then trained on SNR images. After that, the extracted features were fused to train the SVM classifier. The propound system yielded accuracy of 94.00%, sensitivity of 92.00%, and specificity of 93.00% on the CBIS-DDSM dataset. Alshammari et al. [169] employed various ML techniques to categorize benign and malignant tumors using FFDM images. During preprocessing, an experienced radiologist manually labeled tumor lesions to extract ROI images. Twelve features pertaining to shape, density, and texture were then retrieved from each ROI region. The SVM, NB, DT, DA, and KNN classifiers were then trained using two optimal features and an optimization algorithm It was noticed that optimized SVM and NB surpassed other classifiers, achieving accuracy of 100.0% on a private dataset with 42 FFDM images. Song et al. [170] designed multi-feature deep information bottleneck (MDIB) network to delineate CESM images to benign and malignant. During preprocessing, images were normalized and augmented. Initially, the ResNet18 network was trained on multi-view multi-modal images of the CESM to extract multimodal features. These features were then used to train MLP to calculate the information bottleneck and mutual information among the extracted multimodal features. Finally, these features were concatenated to effectively calculate the breeast malignancy. The proposed MDIB framework attained accuracy of 97.18%, sensitivity of 94.86%, and AUC of 0.973 on a private dataset containing 760 CESM images from 95 patients.

Table 2.8 ML and DL in breast cancer classification using sonography during 2016-2022

Author	Year	Country	Type	Dataset	Task	Technique	Optimal Results
Shan et al. [171]	2016	USA	Journal	Private sonograms	Classification	DT, RF, SVM, and ANN	Accuracy: 78.50%, AUC score: 0.830
Singh et al. [172]	2017	India	Journal	Private sonograms	Classification	Fusion of BPANN, SVM and expert opinion	Accuracy: 98.96%, Sensitivity: 99.28%, Specificity: 98.66%, AUC score: 98.97%
Antropava et al. [143]	2017	USA	Journal	Private sonograms	Classification	VGG19 with SVM	AUC score: 0.902
Sultan et al. [173]	2018	USA	Journal	Private sonograms	Classification	RF classifier	Sensitivity: 0.920, Specificity: 0.950, AUC score: 0.960
Byra et al. [174]	2019	USA	Journal	Private sonograms, UDIAT, OASBUD	Classification	CNN	Accuracy: 0.887, Sensitivity: 0.848, Specificity: 0.897, AUC score: 0.936
Wang et al. [175]	2020	Canada	Journal	Private ABUS	Classification	Inception-v3	Accuracy: 0.880, Sensitivity: 0.886, AUC score: 0.947
Daoud et al. [176]	2020	Jordan	Journal	Private sonograms and UDIAT	Classification	VGG19 with SVM	Accuracy: 96.1%, Sensitivity: 95.7%, Specificity: 96.3%, AUC score: 0.981
Mishra et al. [177]	2021	India	Journal	BUSI	Classification	LR, DT, SVM, RF, AdaBoost, and gradient boosting	Accuracy: 0.974, Sensitivity: 0.960, AUC score: 0.970
Eroğlu et al. [178]	2021	Turkey	Journal	BUSI	Classification	AlexNet, ResNet50, and MobileNet-v2	Accuracy: 95.60%, Sensitivity: 95.60%
Shia and Chen [179]	2021	Taiwan	Journal	Private sonograms	Classification	ResNet101 with linear SVM	Sensitivity: 94.34%, Specificity of 93.22%, AUC score: 0.938
Misra et al. [180]	2022	Republic of Korea	Journal	Private grayscale and SE	Classification	AlexNet and ResNet	Accuracy: 90.00%, Sensitivity: 88.89%, F1-score: 89.79%
Hoffmann et al. [181]	2022	Germany	Journal	Private grayscale and SWE	Classification	DCNN	accuracy: 93.53%, Sensitivity: 94.42%, AUC score: 96.55%
Mishra et al. [182]	2022	India	Journal	BUSI and UDIAT	Classification	RF, AdaBoost, gradient boosting, and SVM	Accuracy: 0.974, Sensitivity: 0.977, F1-score: 0.970, AUC score: 0.991
Liu et al. [183]	2022	China	Journal	Private sonograms	Classification	ResNet101 with PCA, NB, and SVM	Accuracy: 89.17%, Recall: 86.49%, AUC score: 0.950
Wang et al. [184]	2022	China	Journal	Private ABUS	Classification	ResNet101-v2	Accuracy: 77.50%, Sensitivity: 85.00%, AUC score: 0.850

ML and DL in breast cancer classification using sonography

Various studies presented supervised ML and DL frameworks to classify breast lesions using sonography, as provided in Table 2.8. Shan et al. [171] proposed a methodology to translate descriptive BI-RADS features of ultrasound imaging into digital features. The optimum feature sets were then selected from digitized features using a bottom-up

searching method. Different combinations of digitized features were used to analyze the classification performance of DT, RF, SVM, and ANN classifiers. The RF model provided the highest accuracy of 78.50% with an AUC of 0.83 on a private dataset of 283 sonograms. Singh et al. [172] intended to improve the clinical use of CAD systems in breast lesion classification using sonography. At first, noise in sonograms was reduced using wavelet-based filtering, and ROI patches were obtained. Second, 457 features were extracted, from which only the 19 most relevant were selected using multiple feature selection methods. Optimal results were obtained using a multi-criteria feature selection method. Finally, classification was performed by integrating BPANN, SVM, and expert opinion. The proposed hybrid method using a multi-criterion feature selection technique yielded accuracy of 98.96%, sensitivity of 99.28%, specificity of 98.66%, and an AUC of 98.97% on a private dataset of 178 sonograms. Antropava et al. [143] joined CNN features with handcrafted features, followed by SVM, to classify breast lesions of breast sonograms. The grayscale ROI images obtained from sonograms were first converted to RGB images by duplicating them across three channels. Then, all the original varying-sized ROI patches were converted to fixed sizes before feature extraction. The fixed-sized ROI images were then fed into the VGG19 model, which produced five feature vectors that were then concatenated to get a single feature vector. The fusion of CNN-based features and handcrafted features yielded an AUC of 0.902 on 2393 ROI images extracted from a private dataset having 1125 sonogram lesions. Sultan et al. [173] followed an ML approach to classify breast lesions as benign or malignant using multimodal sonography images. During preprocessing, lesions were manually traced by an experienced clinician. The grayscale and Doppler features were then extracted automatically to train the RF classifier. It was concluded that combining Doppler and grayscale morphologic features improved diagnostic performance. This approach procured the highest sensitivity of 0.920, specificity of 0.950, and AUC of 0.960 on a private dataset containing grayscale and Doppler images from 160 breast lesions. Byra et al. [174] developed a framework based on the fine-tuning of VGG19 to classify breast sonography masses into benign and malignant. In the beginning, ROI patches were extracted by an experienced radiologist. During preprocessing, ROI images were median filtered, cropped, and resized to the default VGG19 image size of 22×224 pixels. Also, data augmentation was applied to improve the training process. The concept of the matching layer was introduced at the start of the pretrained model to convert the input grayscale ultrasound images to RGB images. This layer reinforced the discriminative power of the pretrained VGG19 network and yielded accuracy of 0.887 ($\pm$0.028), sensitivity of 0.848 ($\pm$0.039), specificity of 0.897 ($\pm$0.035), and an

AUC of 0.936 ($\pm$0.019) on a private dataset containing 882 ultrasound masses with one mass per patient. Wang et al. [175] proposed a multiview CNN model based on Inception-v3 to classify ABUS images into benign and malignant. The ground truth of each lesion was annotated by a physician using a bounding box which was further verified by an experienced radiologist. The lesion patches were cropped from different slices of each lesion to train the model. As the ABUS images can be visualized in coronal and transverse plans, the proposed multiview Inception-v3 was able to effectively extract multiview features from both views. This approach outperformed the traditional ML models as well as single-view CNNs and achieved accuracy of 0.880, sensitivity of 0.886, specificity of 0.876, and an AUC of 0.947 ($\pm$0.016) on a private dataset of 316 breast lesions from 263 patients. Daoud et al. [176] combined deep features with handcrafted morphological features to effectively classify breast sonograms into benign and malignant. During preprocessing, ROI images were obtained from the original images using bounding boxes. The VGG19 was then trained on ROI images to extract deep features at six different levels. These deep features were combined with handcrafted texture and morphological features to train the SVM classifier. Furthermore, a feature selection algorithm was used to choose an optimal combination of deep and handcrafted features. It was found that the combination of deep features and handcrafted morphological features offered promising results with accuracy of 96.1% ($\pm$2.2), sensitivity of 95.7% ($\pm$4.2), specificity of 96.3% ($\pm$3.6), and an AUC of 0.981 on a private dataset containing 380 sonograms. Mishra et al. [177] followed the ML approach to classify breast tumors as benign or malignant using sonograms. Initially, original sonograms were fused with ground truth to produce masked RGB images. Next, 3810 handcrafted features were extracted, including HOG, Hu moments, shape features, and texture features. Following that, the recursive feature elimination technique was utilized to select ten prominent features. Lastly, oversampling was performed using the synthetic minority oversampling technique (SMOTE). It was found that the AdaBoost classifier outperformed others with accuracy of 0.974, sensitivity of 0.960, and AUC of 0.970 on the BUSI dataset. Eroğlu et al. [178] proposed a hybrid model based on DCCN to classify sonograms into normal, benign, and malignant. During preprocessing, data augmentation was applied to normal and malignant images to balance the dataset. The proposed hybrid model comprised of AlexNet, ResNet50, and MobileNet-v2 was employed to extract salient features. The most prominent features were then selected using the minimum redundancy maximum relevance method. This approach yielded accuracy of 95.60% and sensitivity of 95.60% on the BUSI dataset. Shia and Chen [179] exploited pretrained ResNet-101 as a feature extractor together with lin-

ear SVM to classify breast sonograms as benign or malignant. During preprocessing, images were resized 224×224 pixels after the removal of the non-related contents. In addition, data augmentation was used by applying geometric transformations. The pretrained ResNet-101 was then utilized to extract prominent features, followed by the linear SVM to predict the breast malignancy. It was concluded that the proposed approach offered comparable performances to those reported by three experienced physicians. Specifically, it procured sensitivity of 94.34%, specificity of 93.22%, and AUC of 0.938 on malignant images of a private dataset containing 2099 breast sonograms retrieved from 543 patients. Misra et al. [180] developed an ensemble DL architecture to classify benign and malignant tumors using grayscale B-mode and SE sonograms. During preprocessing, images were augmented by applying geometric transformations. The AlexNet and ResNet were then fine-tuned on augmented B-mode and SE images, and the resulting features were combined to predict the level of malignancy. This approach yielded an image-level accuracy of 90.00%, sensitivity of 88.89%, specificity of 91.10%, and F1-score of 89.79% on a private dataset containing 261 B-mode images and 261 SE images. Hoffmann et al. [181] introduced a DCNN framework to categorize benign and malignant lesions using sonography images with each having grayscale and SWE sonograms. During preprocessing, each original sonogram was cropped into four images, two grayscale and two SWE sonograms. In addition, these images were resized to $224 \times 224 \times 3$ pixels. Different combinations of sonograms were used to evaluate the performance of the proposed DCNN model. It was found that combining grayscale and SWE images with the surrounding B-mode image using two ensembled DCNN models achieved the best results, with accuracy of 93.53%, sensitivity of 94.42%, specificity of 90.75% and an AUC of 96.55% on a private dataset of 746 sonography images. Mishra et al. [182] fused CNN features with radiomic features to distinguish between benign and malignant tumors using sonograms. A total of 3812 radiomic features were extracted including shape, HOG, GLCM, and LBP features. After that, relevant features were selected using the recursive feature elimination technique. The SVM, RF, AdaBoost, and gradient boosting were used to finalize the prediction. It was concluded that the gradient boosting classifier outperformed others and acquired accuracy of 0.978, sensitivity of 0.977, F1-score of 0.970, and AUC of 0.991 on the BUSI dataset. Whereas, it attained accuracy of 0.974, sensitivity of 0.979, F1-score of 0.972, and AUC of 0.987 on the UDIAT dataset. Liu et al. [183] proposed a new framework using morphological, textural, and deep features to classify sonograms into benign and malignant. During preprocessing, two experienced pathologists extracted ROI images, which were then resized to 256×256 pixels. For morphological features,

k-means clustering was used to segment regions, and the least correlated features were then used to train the NB classifier. For textural features, the LBP, HOG, and GLCM features were extracted, followed by PCA to find the least correlated features. For deep features, the ResNet101 was used as a feature extractor, followed by PCA to select the least correlated features. The textural and deep features were then used to train the SVM classifier. Finally, the results obtained from NB and SVM were averaged together to calculate the breast malignancy. This methodology procured accuracy of 89.17%, recall of 86.49%, F1-score of 88.28%, and AUC of 0.950 on a private dataset with 791 FFDM images. Wang et al. [184] propounded an innovative DCNN model to classify benign and malignant tumors using ABUS volumes. During preprocessing, lesion volumes were manually marked by three experienced radiologists. The ResNet-v2 model was then provided with the original ABUS and morphological information to accurately predict the level of malignancy. It was observed that the ResNet101-v2 achieved a maximum accuracy of 0.775, sensitivity of 0.850, F1-score of 0.820, and AUC of 0.850 on a private dataset containing 769 ABUS volumes from 743 patients.

ML and DL in breast cancer classification using MRI imaging

Several studies introduced supervised ML and DL architectures to classify breast lesions using MRI, as given in Table 2.9. Razavi et al. [185] proposed a new criterion for selecting four sets of vital morphological features. These include three shape descriptors generated by the sphere packing method, namely volume□radius histogram, packing fraction of enclosing sphere, and graph topological features, whereas the fourth is the Zernike descriptor. These four methods resulted in a total number of 142 features. Also, mean decrease in accuracy (MDA) and mean decrease gini (MDG) were applied to select the top 30 features. The RF classifier outperformed the NB, AdaBoost, and SVM classifiers, achieving accuracy of 90.56%, a precision of 90.30%, and an AUC of 0.94 on a private dataset consisting of 106 non-mass-like lesions of DCE-MRI images. Antropava et al. [143] fused CNN features with handcrafted features, followed by SVM, to classify breast lesions of breast DCE-MRI images. The grayscale ROI images obtained from DCE-MRI images were first converted to RGB images by duplicating them across three channels. In this case, one precontrast and two post-contrast slices of every ROI were combined to get the corresponding RGB image, before feature extraction. The fixed-size ROI images were then fed into the VGG19 model, which produced five feature vectors that were then concatenated to get a single feature vector. The fusion of CNN-based features and handcrafted features yielded an AUC of 0.892 on 690 ROI patches extracted from a private dataset of 690 DCE-MRI images. Zheng et al.

Table 2.9 ML and DL in breast cancer classification using MRI during 2016-2022

Author	Year	Country	Type	Dataset	Task	Technique	Optimal Results
Razavi et al. [185]	2016	Germany	Conference	Private DCE-MRI	Classification	NB, RF, AdaBoost, and SVM	Accuracy: 90.56%, Precision: 90.30%, AUC score: 0.940
Antropava et al. [143]	2017	USA	Journal	Private DCE-MRI	Classification	VGG19 with SVM	AUC score: 0.892
Zheng et al. [186]	2018	China	Conference	Private DCE-MRI	Classification	Dense Convolutional LSTM with ResNet50	Accuracy: 0.847, Recall: 0.782, Precision: 0.815
Luo et al. [187]	2019	Hong Kong	Conference	Private DCE-MRI	Classification	3D ResNet34	Accuracy: 0.855, Sensitivity: 0.857, Specificity: 0.852, AUC score: 0.902
Ji et al. [188]	2019	USA	Journal	Private DCE-MRI	Classification	SVM	Sensitivity: 99.50%, AUC score: 0.890
Haarburger et al. [189]	2019	Germany	Conference	Private DCE-MRI	Classification	3D ResNet18 with multi-scale curriculum learning strategy	Accuracy: 0.810, AUC score: 0.890%
Feng et al. [190]	2020	China	Journal	Private multi-sequence MRI	Classification	KFLI model based on ResNet50 and LSTM	Accuracy: 85.00%, Sensitivity: 84.60%, Specificity: 85.70%, AUC score:0.890
Thakran et al. [191]	2021	India	Journal	Private multiparametric MRI	Classification	NB, RF, and SVM	Accuracy: 0.940, Sensitivity: 0.960, Specificity: 0.920
Fujioka et al. [192]	2021	Japan	Journal	Private DCE-MRI	Classification	Xception, InceptionResNet-v2, Inception-v3, DenseNet	Sensitivity: 74.50%, Specificity: 96.00%, AUC score: 0.895
Amin et al. [193]	2022	UAE	Journal	Private MRI	Classification	SVM and ANN	Accuracy: 98.52%, Sensitivity: 97.00%
Rashid et al. [194]	2022	Turkey	Journal	Private MRI	Classification	CNN with SVM	Accuracy: 95.28%, AUC score: 0.974
Gui et al. [195]	2022	China	Journal	Private MRI	Classification	AlexNet, VGG16, GoogLeNet, ResNet50, and DenseNet	Accuracy: 0.925, Sensitivity: 0.950, AUC score: 0.958
Tsuchiya et al. [196]	2022	Japan	Journal	Private MRI	Classification	SVM, RF,	Accuracy: 93.00%, Sensitivity: 92.00%, AUC score: 0.970

[186] presented a novel structure termed dense convolutional LSTM to classify smaller breast lesions using DCE-MRI and DWI images. During preprocessing, two experienced radiologists performed pixel-level labeling task. The DCE-MRI images were normalized and cropped into $40 \times 40 \times 40 \times t$, where t is the time sequence. Whereas, DWI-MRI images were normalized and cropped into $40 \times 40 \times 40$ at the same location. Initially, the LSTM states were instantiated using Apparent Diffusion Coefficient (ADC) maps generated from DWI images. The proposed dense convolutional LSTM framework was then exploited to extract time-intensity information from DCE-MRI, followed by the ResNet50 to predict the malignancy. The proposed approach outran

the default ResNet50 network and acquired accuracy of 0.847, recall of 0.782, and precision of 0.815 on a private dataset having 72 DCE-MRI images from 72 patients. Luo et al. [187] introduced a novel framework based on 3D ResNet34 to classify breast lesions followed by their localization in DCE-MRI images. During preprocessing, the 3D ResNet34 model was first trained using the proposed Cosine Margin Sigmoid Loss (CMSL). During the classification task, deep features learned by the 3D ResNet34 were embedded into a hyper-sphere to discriminate between benign and malignant classes in an angular feature space. Whereas, during the localization task, tumors were localized by leveraging correlations among the learned deep features using the COrrelation Attention Map (COAM). The suggested model guided by CMSL attained accuracy of 0.855, sensitivity of 0.857, specificity of 0.852, and an AUC of 0.902 on a private dataset containing 10,290 DCE-MRI images from 1715 subjects. Ji et al. [188] investigated the effectiveness of radiomic features along with SVM to classify mass and non-mass lesions of DCE-MRI studies into benign and malignant. Breast lesions were first manually located with the assistance of a skilled radiologist. A computerized approach was then used to perform three-dimensional tumor segmentation. Following that, numerous radiomic features were extracted, including size, shape, morphology, enhancement texture, and enhancement variance. Finally, these features were combined to train the SVM to classify mass and non-mass lesions as benign or malignant. This technique attained an AUC of 0.890 along with sensitivity of 99.50% on a private dataset containing 1979 DCE-MRI studies. Haarburger et al. [189] proposed a new framework using 3D ResNet18 in conjunction with a multiscale curriculum learning strategy to classify whole breast DCE-MRI images into benign and malignant categories at a patient level. It should be noted that the 3D ResNet18 does not depend on lesion segmentation. Also, the multiscale curriculum learning strategy can effectively train the network on the patch level as well as on the whole breast level. This methodology yielded an AUC of 0.89 ($\pm$0.01) and accuracy of 0.810 ($\pm$0.02) on a private dataset containing 408 DCE-MRI studies. Feng et al. [190] developed a Knowledge-driven Feature Learning and Integration (KFLI) framework to classify benign and malignant breast lesions using multi-sequence MRI images containing DCE-MRI and DWI sequences. The two main components of the KFLI included a sequence division module based on domain knowledge and an adaptive weighting module for feature integration. The proposed framework leveraged the properties of DCE-MRI and DWI sequences. The DCE-MRI sequence was divided into three sub-sequences known as over appearance (OA), peripheral tissue (PT), and all phases of DCE-MRI (AP-DCE). The OA and PT sequences contained morphological features, whereas the AP-DCE em-

bodied hemodynamic information. Similarly, the DWI sequence comprehended data about water molecule diffusion. The OA, PT, and DWI patches were used to train the ResNet50 model. On the other hand, the AP-DCE patch with several sequential images was used to train the ResNet50 followed by the LSTM network. Features learned from the above-mentioned four types of patches were then integrated using an adaptive weighting module. The designed KFLI model offered accuracy of 85.00%, sensitivity of 84.60%, specificity of 85.70%, and an AUC of 0.908 on a private dataset of 100 multi-sequence MRI images from 100 patients. Thakran et al. [191] utilized the ML approach to categorize breast tumors as benign or malignant in multiparametric MRI images. During preprocessing, motion correction and automatic segmentation of tumors were performed. Then, from each multiparametric MRI image, 128 features were extracted, including 8 quantitative and 120 texture features. Following that, an optimized feature vector with fifteen quantitative and texture features was computed to train NB, RF, and SVM classifiers. It was found that SVM outperformed others with accuracy of 0.940, sensitivity of 0.960, and specificity of 0.920 on a private dataset containing multiparametric MRI images from 60 patients. Fujioka et al. [192] leveraged DCNN frameworks to classify breast lesions into benign and malignant in DCE-MRI images. During processing, each maximum intensity projection of the DCE-MRI image was converted into an image of 512×512 pixels. These images were then used to train various DL architectures. It was inferred that the InceptionResNet-v2 surpassed others by yielding sensitivity of 74.50%, specificity of 96.00%, and AUC of 0.895 on a private dataset containing 358 DCE-MRI cases. Amin et al. [193] exploited SVM and ANN together with optimal features to categorize benign and malignant tumors using MRI images. During preprocessing, a median filter was used to ensure uniformity of images, followed by the watershed algorithm to segment tumor areas. After that, five features were extracted to train SVM and ANN models. It was found that the SVM with linear kernel surpassed the ANN by securing accuracy of 98.52% and sensitivity of 97% on a private dataset with 56 MRI studies. Rashid et al. [194] explored both ML and DL models to classify benign and malignant tumors using MRI images. For the ML approach, seventy-five features were extracted from the selected ROI patches, including shape and textural radiomic features. Furthermore, for the DL approach, the proposed CNN-SVM and fine-tuned Inception-v3 models were trained. It was found that the SVM with a linear kernel trained on 75 radiomic features attained accuracy of 97.06%, whereas the CNN-SVM attained accuracy of 95.28% and AUC of 0.974 on a private dataset containing 35 MRI cases. Gui et al. [195] leveraged five DCNN models to classify benign and malignant lesions in MRI images. During preprocessing, an

experienced radiologist manually performed lesion-level annotation. The slice-level annotation was then carried out using the region-growing algorithm. These slice-level images were used to train AlexNet, VGG16, GoogLeNet, ResNet50, and DenseNet networks. It was concluded that the DenseNet model surpassed others and obtained accuracy of 0.925, sensitivity of 0.950, and AUC of 0.958 on a private dataset containing 487 lesions from 337 patients. Tsuchiya et al. [196] proposed a hybrid ML model by leveraging radiomic and radiological features to classify benign and malignant tumors in DCE-MRI images. During preprocessing, ROI areas of tumors were extracted. Next, a total of 1070 first-order and second-order radiomic features were retrieved among which 35 were selected using the LASSO technique. Three ML classifiers including SVM, RF, and XGBoost were trained on these optimal features. The hybrid model attained a maximum accuracy of 93.00%, sensitivity of 92.00%, and AUC of 0.970 on a private dataset with 88 MRI studies.

ML and DL in breast cancer classification using histopathology

Table 2.10 ML and DL in breast cancer classification using histopathology during 2016-2022

Author	Year	Country	Type	Dataset	Task	Technique	Optimal Results
Modi et al. [197]	2016	India	Conference	WDBC	Classification	BN, NB, and RF	Accuracy: 96.84%
Spanhol et al. [198]	2016	Brazil	Conference	BreakHis: Binary	Classification	CNN	Accuracy: 85.60%
Zhi et al. [199]	2017	New Zealand	Conference	BreakHis: Binary	Classification	VGG16 transfer learning	Accuracy: 94.80%
Araújo et al. [64]	2017	Portugal	Journal	BCBH	Classification	CNN with SVM	Accuracy: 77.80%
Adeshina et al. [200]	2018	Nigeria	Conference	BreakHis: Multiclass	Classification	Deep CNN with AdaBoost	Accuracy: 91.54%, Precision: 63.36%, Recall: 76.67%
Jonnalagedda et al. [201]	2018	USA	Conference	BreakHis: Binary	Classification	MVPNet based on CNN	Accuracy: 92.20%, Sensitivity: 94.20%, Specificity: 92.30%
Liu [202]	2018	China	Conference	WDBC	Classification	LR classifier	Accuracy: 96.50%
Huang and Chung [203]	2018	China	Conference	BCBH and BACH	Classification	ResNet with MLP	Accuracy: 95.00%
Awan et al. [204]	2018	UK	Conference	BACH	Classification	ResNet50 with SVM	Accuracy: 90.00%
Chennamsetty et al. [205]	2018	India	Conference	BACH	Classification	ResNet101 and DenseNet161	Accuracy: 97.50%
Bardou et al. [206]	2018	China	Journal	BreakHis: Multiclass	Classification	CNN and SVM	Accuracy: 88.23%
Yang et al. [207]	2019	China	Journal	BACH	Classification	EMS-Net based on ResNet152, DenseNet161, and ResNet101	Accuracy: 91.75%

Table 2.10 continued from previous page

Author	Year	Country	Type	Dataset	Task	Technique	Optimal Results
Li et al. [208]	2019	China	Journal	BCBH	Classification	ResNet50 with SVM	Accuracy: 88.89%
De Matos et al. [209]	2019	Canada	Conference	BreakHis: Binary	Classification	Inception-v3 with PCA and SVM	Accuracy: 91.00%
Abdar et al. [210]	2020	Canada	Journal	WDBC	Classification	Nested Ensemble classifier based on NB	Accuracy: 98.07%, Recall: 98.10%, F1-score: 98.10%
Hajiabadi et al. [211]	2020	Iran	Journal	WDBC	Classification	ANN	Accuracy: 0.960
El-Shair et al. [212]	2020	USA	Conference	WDBC	Classification	LR, NB, and MLP	Accuracy: 0.970, Recall: 0.920, F1-score: 0.960
Ray et al. [213]	2020	USA	Conference	WDBC	Classification	LR, NB, and MLP	Accuracy: 0.975, Precision: 0.980, Recall: 0.950, F1-score: 0.970
Dhahri et al. [214]	2020	KSA	Journal	MGH-BIDMC and WDBC	Classification	LR, KNN, Gaussian NB, extremely randomized tree (ET), and AdaBoost	Accuracy: 96.07%, Precision: 96.00%, Recall: 96.00%, F1-score: 96.00%
Mewada et al. [215]	2020	KSA	Journal	BCBH and BreakHis	Classification	CNN with spectral-spatial features	Accuracy: 97.45%, Sensitivity: 96.59%, Specificity: 97.73%, AUC score: 99.03%
Assiri et al. [216]	2020	KSA	Journal	WDBC	Classification	LR, SVM, and MLP	Accuracy: 99.42%, Precision: 99.40%, Recall: 994.0%, F1-score: 99.40%
Saxena et al. [217]	2020	India	Journal	BreakHis	Classification	AlexNet and ResNet50 with SVM	Accuracy: 93.92%
Ibraheem et al. [218]	2021	Egypt	Journal	BreakHis: Binary	Classification	CNN with residual blocks	Accuracy: 97.04%, Sensitivity: 97.14%
Inan et al. [219]	2021	Bangladesh	Conference	WDBC	Classification	Ensemble model using LR, KNN, and SVM	Accuracy: 98.25%, Recall: 100%, Precision: 97.30%
Boumaraf et al. [220]	2021	China	Journal	BreakHis: Multiclass	Classification	ResNet18 with transfer learning	Accuracy: 92.03%, Recall: 90.28%, F1 measure: 90.77%
Ghosh et al. [221]	2021	Bangladesh	Conference	WDBC	Classification	MLP, CNN, RNN, LSTM, and GRU	Accuracy: 99.10%, Sensitivity: 98.00%
Yari et al. [222]	2021	Norway	Conference	BreakHis: Multiclass	Classification	ResNet50	Accuracy: 94.33%
Rashmi et al. [223]	2021	India	Journal	Private and BreakHis: Binary	Classification	BCHisto-Net based on CNN	Accuracy: 95.00%, Cohen's kappa: 0.890

Table 2.10 continued from previous page

Author	Year	Country	Type	Dataset	Task	Technique	Optimal Results
Hu et al. [224]	2021	China	Journal	BreakHis: Binary	Classification	Modified ResNet34	Accuracy: 94.03%
Mashudi et al. [225]	2021	Malaysia	Conference	WDBC	Classification	KNN, SVM, RF, Bagging, and AdaBoost	Accuracy: 98.77%
AlKassar et al. [226]	2021	Iraq	Journal	BreakHis: Multiclass	Classification	Xception and DenseNet	Accuracy: 92.20%
Saleh et al. [227]	2022	Egypt	Journal	WDBC	Classification	An optimized RNN	Accuracy: 96.74%, Recall: 96.74%, and F1-score: 96.80%
Zaalouk [228]	2022	Egypt	Journal	BreakHis: Multiclass	Classification	VGG19, Xception, DenseNet201, and ResNet152	Accuracy: 93.32%, Recall: 92.36%, F1-score: 92.44%
Madhulika and Sampath [229]	2022	India	Conference	BreakHis: Binary	Classification	ResNet50 with linear SVM	Accuracy: 96.92%
Sharma et al. [230]	2022	India	Journal	BreakHis: Multiclass	Classification	ResNet50 with linear SVM	Accuracy: 0.980, Recall: 0.980, F1-score: 0.910
Labrada and Barkana [231]	2022	USA	Conference	BreakHis: Binary	Classification	DT, SVM, KNN, and NN	Accuracy: 96.90%, Recall: 97.40%, F1-score: 97.70%, AUC score: 98.80%
Abbasniya et al. [232]	2022	Iran	Journal	BreakHis: Binary	Classification	Inception-ResNet-v2 with XGBoost, CatBoost, LightBoost	Accuracy: 96.46%, F1-score score: 97.44%
Aljuaid et al. [233]	2022	KSA	Journal	BreakHis: Multiclass	Classification	ResNet18, Inception-v3, and ShuffleNet	Accuracy: 97.81%, Sensitivity: 97.65%, SpecificitY: 97.31%
Alaoui et al. [234]	2022	Morocco	Conference	BreakHis: Binary	Classification	Ensemble of seven pre-trained DCNN models	Accuracy: 93.80%, Recall: 95.70%, F1-score score: 93.80%
Nakach et al. [235]	2022	Morocco	Journal	BreakHis: Binary	Classification	Inception-v3 with XGBoost	Accuracy: 92.52%, Recall: 95.36%, F1-score score: 92.93%
Khan et al. [236]	2022	Bangladesh	Journal	BACH and BreakHis: Binary	Classification	MultiNet based on VGG16, DenseNet201, and NasNetMobile	Accuracy: 0.980, Recall: 0.980, F1-score: 0.980
Xu et al. [237]	2022	China	Journal	BreakHis: Binary	Classification	DenseNet201 with RBF SVM	Accuracy: 94.88%, Recall: 93.97%, F1-score: 93.38%
Silva and Cortes [238]	2022	Brazil	Journal	BreakHis: Binary	Classification	ResNet18, ResNet152, and GoogLeNet	Accuracy: 0.840, Recall: 0.860, F1-score: 0.880

Many studies developed supervised ML and DL models to classify breast lesions using histopathology, as provided in Table 2.10. Modi and Ghanchi [197] leveraged feature selection techniques such as information gain (IG), relief (R), correlation attributes (CA), and correlation-based feature selection (CFS) to compare the classification performance of Bayes network (BN), Naïve Bayes (NB) and random forest (RF) algorithms. It was found that the RF model together with CA achieved the highest accuracy of 96.84% on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Spanhol et al. [198] proposed a CNN architecture with three convolutional layers and two fully connected layers for the classification of breast histopathology images into benign and malignant. During preprocessing, patches of 32×32 and 64×64 pixels were extracted from the input images using sliding windows and random extraction methods. Furthermore, these patches were normalized by subtracting the mean image from each patch before training the proposed CNN model. The output was calculated by combining patch probabilities with the sum, product, and maximum rules. It was found that the maximum rule outperformed the sum and product rules by achieving image level accuracy of 85.60 (± 4.8) for 40X on the binary task of the BreakHis dataset. Zhi et al. [199] designed a custom model comprising six bottom convolutional layers of the VGG16 network to classify breast histopathology images as benign or malignant. First, five random patches were extracted from each image of all resolutions in the BreakHis dataset. The proposed model was then trained three times based on different patches which resulted in three classifiers. The ensemble of these three classifiers offered accuracy of 93.3% for 40X, 94.6% for 100X, 94.8% for 200X, and 88.40% for 400X on the binary task of the BreakHis dataset. Araújo et al. [64] suggested a novel CNN framework composed of five convolutional layers and three fully connected layers to classify breast histopathology images. During preprocessing, each image was divided into 35 patches of 512×512 pixels with 50% overlap, followed by channel-wise patch normalization. The proposed CNN was trained on augmented patches to perform patch-wise classification. Also, deep CNN features were then used to train the SVM classifier. The image-wise classification was calculated by combining patch probabilities with the sum, maximum, and majority voting rules. This novel architecture attained the highest image accuracy of 85.00% using the majority voting rule. Adeshina et al. [200] presented a deep CNN architecture together with AdaBoost to classify breast cancer histopathology images. During preprocessing, images were resized into $400 \times 400 \times 3$ pixels, followed by their normalization. The proposed CNN-based model as a feature extractor in conjunction with the AdaBoost classifier achieved accuracy of 91.54%, a precision of 63.36%, and a recall of 76.67% on the multiclass task of the

BreakHis dataset. Jonnalagedda et al. [201] designed a CNN-based framework called multi-viewing path DL neural network (MVPNet), as well as A new data augmentation method called NuView, for magnification independent classification of breast cancer histopathology images. The MVPNet simultaneously analyzed local and global features of a tissue image using multiple kernels. Whereas the NuView leveraged the location of tumor areas to generate images of random magnification. This approach significantly reduced the number of parameters compared to standard transfer learning models such as the ResNet50 network. The proposed MVPNet together with NuView augmentation offered accuracy of 92.20% ($\pm$1.6), sensitivity of 94.20% ($\pm$2.2), and specificity of 92.30% ($\pm$2.4) on the binary task of the BreakHis dataset. Liu [202] evaluated the effectiveness of features selection on the classification performance of logistic regression using WDBC dataset. Among the ten features, maximum texture and maximum perimeter were found to be the most effective. The combination of these two features achieved a maximum accuracy of 96.5% on the WDBC dataset. Huang and Chung [203] utilized a proposed deep spatial fusion network composed of an adapted ResNet and MLP to compute spatial correlations among patches. Every input image was divided into 12 non-overlapping patches following data augmentation and normalization. A modified ResNet model was initially trained on these patches to obtain the corresponding discriminatory features in the form of probability maps. The MLP was then trained on spatially distributed probability maps to calculate class probability for each input image. The proposed strategy achieved accuracy of 86.1% and 95.00% on BCBH and BACH datasets, respectively. Awan et al. [204] proposed a context-aware methodology to classify histopathology images using ResNet50 model together with the SVM classifier. Each input image was first divided into 12 non-overlapping patches. The ResNet50 model was then employed to generate an 8192-dimensional feature vector for each patch. The dimensionality of these 12 features was further reduced using the PCA technique. For context-based image classification, flattened features of 2×2 overlapping blocks of patches were used to train the SVM classifier. The proposed approach achieved accuracy of 90.00% on the BACH dataset. Chennamsetty et al. [205] used an ensemble of a ResNet101 and two DenseNet161 networks to classify breast histopathology images. During preprocessing, images were downsampled using bilinear interpolation. Moreover, the images were normalized to zero mean and unit standard deviation, computed from the pretrained ImageNet as well as from the entire training histology dataset. The ResNet101 and DenseNet161 models were fine-tuned on the ImageNet normalized dataset. Whereas the other DenseNet161 network was fine-tuned with a normalized dataset from the training histology dataset. An

ensemble of the above-mentioned three networks on two different normalized datasets offered accuracy of 97.50% on the BACH dataset. Bardou et al. [206] compared the performance of an ML model based on handcrafted features with that of a newly designed DL model to automatically classify breast cancer histopathology images. In the ML approach, handcrafted features were extracted using two coding models called bag-of-words and locality-constrained linear coding. These features were then used to train the SVM classifier. Whereas in the DL approach, a CNN model was designed with five convolutional layers followed by two fully connected layers. The newly designed CNN model outperformed the other, achieving accuracy of 88.23% for 40X on the multiclass task of the BreakHis dataset. Yang et al. [207] proposed a novel framework for classifying breast histopathology images called Ensemble of MultiScale convolutional neural Network (EMS-Net). The EMS-Net leveraged an ensemble of fine-tuned DenseNet161, ResNet152, and ResNet101 models. During preprocessing, each image was converted to multiple scales, and patches of 224×224 pixels were extracted at each scale. Also, patches were normalized to a zero mean and a unit variance, followed by data augmentation. The proposed EMS-Net was trained on the augmented patches of multiple scales and achieved accuracy of 91.75% ($\pm$2.32) using five-fold cross-validation on the BACH dataset. Li et al. [208] utilized cell-level and tissue-level information to train the ResNet50 model as a feature extractor for classifying breast histopathology images. During preprocessing, H&E stain inconsistency was normalized using Reinhard normalization. Two different patches of sized 128×128 pixels and 512×512 pixels were extracted from each input image to preserve the cell-level and tissue-level features, respectively. Small patches were extracted without overlap and then clustered by class using the k-means algorithm to select discriminative patches. Whereas, large patches were extracted with 50% overlap. Both types of patches were resized to 224×224 pixels before training the ResNet50 model. The extracted features were used to train the SVM model. The proposed strategy achieved accuracy of 88.89% on the BCBH dataset. De Matos et al. [209] proposed a new framework using transfer learning with inception-v3 along with PCA and SVM to classify breast histopathology images. Each input image was first converted into overlapped patches of size 150×150 pixels. Patch features were then extracted with two different strategies. In the first type, patch-wise deep features were extracted using Inception-v3 as a feature extractor followed by PCA. In the second type, patch-wise handcrafted features were extracted using the PFTAS method. These features were independently used to train the SVM model to filter out irrelevant patches by employing tissue structures from another histopathology image dataset. The remaining relevant patches were used to train

the SVM classifier. It was found that transfer learning with Inception-v3, followed by PCA, attained a maximum accuracy of 91.00% ($\pm$3.00) for 100X on the binary task of the BreakHis dataset. Abdar et al. [210] developed a novel Nested Ensemble (NE) framework to classify benign and malignant breast tumors. It employed stacking and voting (SV) techniques to combine classifiers for ensemble learning. The NE architecture has two major parts including the Classifier and MetaClassifier. The Classifier incorporated various ML models. Whereas, the MetaClassifier embodied two or three different classification algorithms. Four two-layered NE frameworks were designed with two having two classifiers and two carrying three classifiers in their MetaClassifiers. It was found that the SV-Naïve Bayes-3-MetaClassifier outperformed others and yielded accuracy of 98.07%, a precision of 98.10%, a recall of 98.10%, and an F1-score 98.10% of on the WDBC dataset. Hajiabadi et al. [211] introduced an ensemble loss function to categorize benign and malignant breast tumors. The new robust and generalized objective function resulted from the linear combination of Hing, Correntropy, and Cross-entropy loss functions. The dataset was first normalized between -1 and 1. The proposed ensemble loss function was then employed to train a simpler ANN with three hidden layers, KNN, RF, and SVM classifiers. It was found that the ANN outperformed KNN, RF, and SVM models and attained accuracy of 0.960 on noisy labels of the WDBC dataset. El-Shair et al. [212] compared the performance of several ML algorithms, including LR, NB, and MLP, in classifying benign and malignant breast tumors. The dataset was predominantly normalized between 0 and 1. The features were then ranked using mutual information, and ten optimal features were selected to train classifiers. It was found that MLP and LR outperformed the NB and yielded accuracy of 0.970, a precision of 1.00, a recall of 0.920, and an F1-score of 0.960 on the WDBC dataset. Ray et al. [213] proposed a cloud-based platform that utilized DT and RF to classify breast tumors into benign and malignant. During preprocessing, the dataset was cleaned by removing missing values, followed by data normalization between 0 and 1. After that, PCA was employed to select the ten most crucial features. It was found that RF surpassed DT with accuracy of 0.975, a precision of 0.980, a recall of 0.950, an F1-score of 0.970, and an AUC of 0.990 on the WDBC dataset. Dhahri et al. [214] employed five different ML algorithms based on reduced features to classify benign and malignant breast tumors. The most significant features were first selected using the tabu search algorithm. Then, the KNN, Gaussian NB (GNB), LR, extremely randomized tree (ET), and AdaBoost classifiers were trained using the reduced features. It was discovered that significant features selected using the tabu search algorithm improved the performance of classifiers. On

the one hand, the AdaBoost outperformed others by achieving accuracy of 96.07%, a precision of 96.00%, sensitivity of 96.00%, an F1-score of 96.00%, and an AUC of 95.00% on the MGH-BIDMC dataset. On the other hand, the LR outperformed others by obtaining accuracy of 98.00%, a precision of 99.00%, sensitivity of 99.00%, an F1-score of 99.00%, and an AUC of 98.00% on the WDBC dataset. Mewada et al. [215] integrated spectral features of multi-resolution wavelet transform with spatial features of CNN to classify breast histopathology images. During preprocessing, images were partitioned to patches of 512×512 pixels with 50% overlapping. Also, data augmentation was performed by applying random rotation and mirroring. The proposed wavelet-CNN-based framework was then trained on augmented patches having similar labels as the original images. It yielded a maximum accuracy of 97.58% for 40X on the binary task of the BreakHis dataset. Also, it offered accuracy of 97.45%, sensitivity of 96.59%, specificity of 97.73%, and an AUC of 99.03% on the BCBH dataset. Assiri et al. [216] presented an ensemble model based on the majority voting mechanism to classify breast tumors. Initially, the performance of eight cutting-edge ML classifiers was evaluated using the WDBC dataset. Subsequently, the three best classifiers including LR, SVM with SGD, and MLP were chosen based on their F-scores. The classification results of these three classifiers were forwarded to the ensemble classifier to calculate the final outputs. This framework offered accuracy of 99.42%, a precision of 99.40%, a recall of 994.0%, and an F1-measure of 99.40% on the WDBC dataset. Saxena et al. [217] investigated ten state-of-the-art pretrained CNN architectures as feature extractors for classifying breast histopathology images. During preprocessing, each input image was divided into six non-overlapped patches of 224×224 pixels. These patches were then fed into a pretrained deep CNN model under the consideration of extracting patch features. Following that, the extracted patch features were concatenated to generate a feature vector representing an image. Finally, feature vectors were employed to train the linear SVM classifier to finalize the classification. The ResNet50 with linear SVM yielded a maximum image-level accuracy of 90.12% for 200X on the binary task of the BreakHis dataset. Whereas, AlexNet with linear SVM attained a top patient-level accuracy of 93.92% for 200X on the same dataset. Ibraheem et al. [218] designed a framework having three parallel CNN with residual blocks to classify breast histopathology images into benign and malignant. During preprocessing, original images were resized to 224×224 pixels, followed by data augmentation. This model yielded accuracy of 97.04%, sensitivity of 97.14%, and specificity of 95.23% for 200X on the binary task of BreakHis dataset. Inan et al. [219] designed a hybrid model by integrating the probabilities of three ML classifiers

to classify breast tumors into benign and malignant. During preprocessing, relevant features were selected using the Kendall rank correlation method. Furthermore, the dataset was augmented with SMOTE and normalized using robust scaling. The proposed ensemble model incorporated probabilities of individual LR, KNN, and SVM classifiers, and achieved accuracy of 98.25%, recall of 100%, and precision of 97.30% on the WDBC dataset. Boumaraf et al. [220] leveraged DCNN with transfer learning for magnification-independent classification of breast histopathology images. During preprocessing, images were resized to 224×224 pixels using bilinear interpolation and normalized with global contrast normalization. In addition, the dataset was augmented by applying geometric transformations. The proposed framework based on block-wise fine-tuning of ResNet18 achieved accuracy of 92.03%, recall of 90.28%, precision of 91.38%, and F1 measure of 90.77% on the multiclass task of the BreakHis dataset. Ghosh et al. [221] performed a comparative analysis of different DL techniques to classify breast tumors as benign or malignant. During preprocessing, the dataset was converted into numerical values. Then, seven different DL models including MLP, CNN, RNN, LSTM, and GRU were trained and compared based on their classification performance. It was concluded that LSTM and GRU outperformed others and attained accuracy of 99.10% and sensitivity of 98.00% on the WDBC dataset. Yari et al. [222] exploited DCNN with transfer learning for magnification-independent classification of breast histopathology images. During preprocessing, images were augmented by applying geometric transformations. The ResNet50 network was then used as a feature extractor which achieved accuracy of 94.33% on the multiclass task of the BreakHis dataset. Rashmi et al. [223] proposed a new BCHisto-Net framework to classify breast histopathology images as benign or malignant. It encompassed two parallel CNN branches with one for extracting regional features and the other for retrieving global features. The preprocessing step involved stain normalization and data augmentation. The local and global features extracted from the proposed BCHisto-Net were then aggregated to predict the malignancy. The proposed model attained accuracy of 95.00% and Cohen's kappa of 0.89 on a private KMC dataset having 1043 images of 100X resolutions. Moreover, it offered accuracy of 89.00% and Cohen's kappa of 0.77 for 100X on the binary task of the BreakHis dataset. Hu et al. [224] presented a modified ResNet34 to classify breast histopathology images into benign and malignant. The preprocessing stage involved stain normalization and data augmentation. The modified ResNet34 network was then trained to estimate the malignancy. It offered a maximum accuracy of 94.03% for 40X on the binary task of the BreakHis dataset. Mashudi et al. [225] evaluated several ML techniques to classify breast tumors

as benign or malignant. During preprocessing, twenty-three prominent features were selected using the DT algorithm, followed by their normalization between 0 and 1. These features were then used to train KNN, SVM, and ensemble models such as Bagging, RF, and AdaBoost classifiers. It was found that the AdaBoost ensemble model outperformed others by obtaining accuracy of 98.77% on the WDBC dataset. AlKassar et al. [226] employed the Xception and DenseNet networks for efficient classification of breast histopathology images. During preprocessing, stain normalization was performed. Then, shallow features were extracted using the Xception model whereas deep features were retrieved with the DenseNet network. Lastly, these features were used to train an ensemble classifier to maximize the classification performance. This approach offered accuracy of 92.20% for 40X on the multiclass task of the BreakHis dataset. Saleh et al. [227] proposed an optimized DL framework based on RNN to classify breast tumors into benign and malignant. During preprocessing, three feature selection methodologies including the correlation matrix, univariate feature selection, and recursive feature elimination were applied to select salient features. Numerous ML classifiers including DT, RF, NB, KNN, and SVM were then exploited to predict breast malignancy. It was concluded that the optimized deep RNN using the univariate feature selection technique outperformed the aforementioned ML classifiers, attaining accuracy of 96.74%, precision of 96.39%, recall of 96.74%, and F1 value of 96.80% on the WBDC dataset. Zaalouk et al. [228] leveraged numerous pretrained DCNN models with feature extraction and fine-tuning strategies for magnification-independent classification of breast cancer histopathology images. During preprocessing, images were downsampled to 200×200 pixels, followed by data augmentation and normalization. After that, the VGG19, Xception, DenseNet201, InceptionResNet-v2, and ResNet152 networks were investigated using feature extraction and fine-tuning approaches. It was concluded that the fine-tuned Xception network outperformed others, obtaining accuracy of 93.32%, recall of 92.36%, and F1-score of 92.44% on the multiclass task of the BreakHis dataset. Madhulika and Sampath [229] exploited the ResNet50 network accompanied with linear SVM to classify breast cancer histopathology images as benign or malignant. During preprocessing, images were downsampled to 224×2224 pixels, followed by normalization and data augmentation. The ResNet50 was utilized to extract vital features, which were then used to train the linear SVM to predict tumor severity. The proposed strategy achieved accuracy of 96.92% on the binary task of the BreakHis dataset. Sharma et al. et al. [230] proposed a DL framework composed of ResNet50 as a feature extractor together with SVM to categorize breast cancer histopathology images. During preprocessing, images were downsampled to 175×115

pixels. The ResNet50 network as a feature extractor was then amalgamated with linear SVM to predict the tumor class. This approach acquired an average accuracy of 0.980, recall of 0.980, and F1-score of 0.910 on the multiclass task of the BreakHis dataset. Labrada and Barkana [231] utilized the ML approach based on nuclei characteristics to distinguish histopathology images into benign and malignant. During preprocessing, cell nuclei were segmented using the Otsu method. A total of thirty-three geometric, directional, and intensity features were then extracted from segmented nuclei to train the DT, SVM, KNN, and NN classifiers. It was found that the NN exceeded others, achieving accuracy of 96.90%, recall of 97.40%, AUC of 98.80%, and F1-score of 97.70% for 400X on the binary task of the BreakHis dataset. Abbasniya et al. [232] employed pretrained DCNN models as feature extractors along with gradient boosting algorithms to delineate breast cancer histopathology images as benign or malignant. It should be noted that no comprehensive preprocessing methods were used besides image normalization. The pretrained Inception-ResNet-v2 was then employed to extract prominent features from images, followed by an ensemble of XGBoost, CatBoost, and LightGBM classifiers. This methodology attained an average accuracy of 96.46% and an F1-score of 97.44% on the binary task of the BreakHis dataset. Aljuaid et al. [233] leveraged pretrained DCNN models as feature extractors to demarcate breast cancer histopathology images. During preprocessing, images were enhanced using Median and Gaussian filters. Furthermore, images were resized according to the input of pretrained networks, followed by data augmentation. Three models including ResNet18, Inception-v3, and ShuffleNet were then utilized to predict the level of malignancy. It was concluded that the ResNet18 surpassed Inception-v3 and ShuffleNet, yielding accuracy of 97.81%, sensitivity of 97.65%, and specificity of 97.31% on the multiclass task of the BreakHis dataset. Alaoui et al. [234] introduced an ensemble framework based on seven pretrained DCNN models to classify breast histopathology images into benign and malignant. During preprocessing, images were enhanced using the CLAHE method, normalized, and augmented. The VGG16, VGG19, ResNet50, Inception-v3, Inception-ResNet-v2, Xception, and MobileNet models were then stacked together, followed by the LR classifier. This methodology attained accuracy of 93.80%, recall of 95.70%, and F1-score of 93.80% for 40X on the binary task of the BreakHis dataset. Nakach et al. [235] utilized pretrained DCNN models as feature extractors together with boosting classifiers to categorize breast histopathology images as benign or malignant. During preprocessing, images were enhanced using the CLAHE method, normalized, and augmented using geometric transformations. The DenseNet201, MobileNet-v2, and Inception-v3 were employed as feature extractors, followed by the AdaBoost,

GBM, XGBoost, and LightGBM to predict breast malignancy. It was found that the Inception-v3 in conjunction with the XGBoost carrying 200 trees outran the other two networks, yielding accuracy of 92.52%, recall of 95.36%, and F1-score of 92.93% for 40X on the binary task of the BreakHis dataset. Khan et al. [236] presented a framework called MultiNet framework based on pretrained VGG16, DenseNet201, and NasNetMobile as feature extractors to classify breast cancer histopathology images. During preprocessing, images were resized to 224×224 pixels, normalized, and augmented. The features extracted from the aforementioned networks were then integrated to effectively predict breast malignancy. The proposed MultiNet procured accuracy of 0.980, a recall of 0.980, and an F1-score of 0.980 on the BACH dataset. Moreover, it offered accuracy of 0.990, recall of 0.990, and F1-score of 0.990 on the binary task of the BreakHis dataset. Xu et al. [237] leveraged pretrained DenseNet201 as a feature extractor together with RBF SVM to categorize breast cancer histopathology images as benign or malignant. During preprocessing, images were resized and normalized according to the input of the pretrained model. The DenseNet201 was then employed as a feature extractor, followed by the RBF SVM to predict breast malignancy. It obtained accuracy of 94.88%, recall of 93.97%, and F1-score of 93.38% on the binary task of the BreakHis dataset. Silva and Cortes [238] exploited three pretrained DCCN frameworks to classify breast cancer histopathology images as benign or malignant. During preprocessing, images were resized to 224×224 pixels, followed by normalization and augmentation. The pretrained ResNet18, ResNet152, and GoogLeNet networks were then used as feature extractors to predict breast malignancy. It was noted that the ResNet152 achieved a maximum performance with accuracy of 0.84, recall of 0.86, precision of 0.90, and an F1-score of 0.880 on the binary task of the BreakHis dataset.

2.6 Chapter summary

The present systematic review focused on leveraging supervised ML and DL models in the detection, segmentation, and classification of breast lesions using widely used medical imaging modalities including mammography, sonography, MRI, and histopathology from January 2016 to December 2022. To that end, we analyzed 142 studies using supervised ML and DL approaches. It is concluded that mammography and MRI are mostly followed for the detection and segmentation tasks, followed by sonography. Whereas, mammography and histopathology are commonly exploited for the classification task, followed by sonography and MRI. Similarly, it is inferred that mammogra-

phy and histopathology have more publicly available datasets compared to sonography and MRI datasets. It is further deduced that the CNN is progressively being followed in computer-aided diagnosis of breast malignancy. However, it is still supported by traditional ML techniques. Besides, it is noticed that most of the articles investigated the diagnosis of benign and malignant lesions or their sub-classes, whereas, only a few studies have focused on patient-level diagnosis.

However, several limitations should be considered before interpreting the findings of the current systematic review. Firstly, it did not consider thermography and tomography techniques such as computerized tomography and positron emission tomography due to a comparably low number of published studies or lack of publicly released datasets. Secondly, it did not include semi-supervised, unsupervised, and self-supervised ML and DL frameworks. Thirdly, it did not count on attention-based models, transformer-related architectures, graph-based models, and generative adversarial networks. Lastly, it did not address breast cancer prognosis, malignancy recurrence in breast patients treated with neoadjuvant chemotherapy, response to chemotherapy, and survivability prediction.

In the future, it is recommended to publish more publicly available annotated datasets, especially related to MRI and sonography. Similarly, the patient-level approaches should be endorsed to further improve the diagnostic procedure [220, 233]. Likewise, we recommend exploring semi-supervised, unsupervised, and self-supervised frameworks based on the CNN model. Furthermore, the proposed architectures should be cross-validated to reduce underfitting and overfitting problems, which in turn could make models more generalized toward unseen test data. Finally, there is a critical need to build robust and efficient DL models to assist clinicians in diagnosing breast cancer at an early stage, which could save billions of dollars in healthcare costs [239].

Chapter 3

Study I: Binary Classification of Breast Cancer

This chapter elucidates numerous concepts related to the binary classification of breast cancer using a DL approach. The research findings of this study have been published in a peer-reviewed journal entitled "Breast cancer histopathology image classification using an ensemble of deep learning models" [43], as explained in the succeeding sections.

3.1 Introduction

Cancer is one of the critical public health issues around the world. According to the Global Burden of Disease (GBD) study, there have been 24.5 million cancer incidence and 9.6 million cancer deaths worldwide in 2017 [240]. These statistics indicate that cancer incidence expanded by 33% between 2007 and 2017 worldwide [240]. Specifically, breast cancer is the most common malignancy and the leading cause of cancer-related mortalities among women worldwide [240, 241]. Thus, premature diagnosis of this pathology is crucial to preclude its progression and reduce its morbidity rates in women.

Breast cancer is a heterogeneous disease, composed of numerous entities with distinctive biological, histological and clinical characteristics [242]. This malignancy erupts from the growth of abnormal breast cells and might invade the adjacent healthy tissues [242]. Its clinical screening is initially performed by utilizing radiology images, for instance, mammography, ultrasound imaging, and Magnetic Resonance Imaging (MRI) [23, 243]. However, these non-invasive imaging approaches may not be capa-

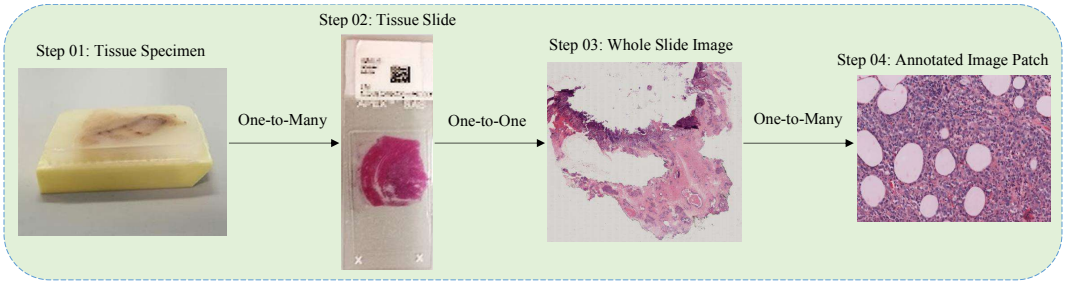


Figure 3.1 The complete process of biopsy is depicted in Figure. Steps 01 and 02 are taken from [8] whereas steps 03 and 04 are retrieved from our own dataset.

ble of determining the cancerous areas efficiently. To this end, the biopsy technique is usually used to analyze the malignancy in breast cancer tissues more comprehensively. The process of biopsy includes the collection of tissue samples, mounting them on microscopic glass slides, and staining these slides for better visualization of nuclei and cytoplasm [42]. Pathologists then carry out the microscopic analysis of these slides in order to finalize the diagnosis of breast cancer [42]. The complete process of biopsy technique is depicted in Figure 3.1, and is comprehensively described in [8].

However, the manual analysis of complex-natured histopathological images is fairly a time-consuming and tedious process, and could be prone to errors. Also, the morphological criteria used in the classification of these images are somehow subjective, which leads to the result that an average diagnostic concordance among the pathologists is approximately 75% [18]. Therefore, the computer-assisted diagnosis [23, 42, 244] plays a significant role to assist pathologists in analyzing the histopathology images. Specifically, it improves the diagnostic accuracy of breast cancer by reducing the inter-pathologist variations in diagnostic decisions [42]. However, the conventional computerized diagnostic approaches, ranging from rule-based systems to machine learning techniques, may not effectively challenge the intra-class variation and inter-class consistency within the histopathology images of breast cancer [245]. Also, these methodologies mainly rely on feature extraction methods like scale-invariant feature transform [246], speed robust features [247] and local binary patterns [248] which all are based on supervised information and can be prone to biased results during the classification of breast cancer histopathology images [245]. Therefore, the need for efficient diagnosis leads to an advanced set of computational models based on multiple layers of nonlinear processing units, called deep learning [19].

Recently, deep learning models [8, 19, 249–251] have made remarkable progress in computer vision, specifically in biomedical image processing, due to their abilities to automatically learn complicated and advanced features from images, which

inspired various researchers to leverage these models in the classification of breast cancer histopathology images [8]. Especially convolutional neural networks (CNNs) [83] are widely used in image-related tasks due to their abilities to effectively share parameters across various layers within a deep learning model. Numerous CNN-based architectures have been proposed during the past few years; however, AlexNet [14] is considered as of the first deep CNNs to achieve considerable accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) during 2012. Thereafter, VGG architecture [9] introduced the idea of leveraging deeper networks with smaller convolutional filters, and achieved second place at ILSVRC 2014. The intuition of multiple stacked smaller convolutional filters can provide an effective receptive field and is also used in recently proposed pretrained models, including Inception Network [94] and residual neural network (ResNet) [95]. In this work, we employed two different approaches of VGG architecture for an efficient classification of breast cancer histopathology by utilizing our own created dataset. The main contributions of this study are: First, we created a private dataset of whole slide images (WSI) from breast cancer patients with the help of experienced pathologists. Then, image patches were extracted from the WSI images, composed of non-carcinoma and carcinoma classes. Next, we selected and trained different combinations of pretrained VGG16 and VGG19 [9] deep learning architectures (discussed in Section 3.3). Specifically, we evaluated an individual as well as ensemble performances of fully-trained and fine-tuned VGG16 and VGG19 frameworks [9]. Of note, our main objective is the correct classification of the carcinoma class on a priority basis and we found that the ensemble of fine-tuned VGG16 and VGG19 approach [9] provided superior performance in the classification of non-carcinoma and carcinoma histopathology images of breast cancer.

The remaining sections of this chapter are provided as follows. Section 3.2 presents related work. Section 3.3 demonstrates the materials and methods used to conduct this research. Section 3.4 shows experimental setup and Section 3.5 illustrates the results along with discussion. Finally, Section 3.6 highlights the conclusion and future direction of this study.

3.2 Related work

With the evolution of machine learning in biomedical engineering, numerous studies leveraged handcrafted features-based approaches for the classification of histopathology images related to breast cancer. For instance, Kowal et al. [252] focused on the nuclei segmentation and extracted forty-two morphological, topological and texture

features from the segmented nuclei of 500 fine-needle biopsy images of breast cancer. Then, these features were utilized to train three different classifiers in order to classify these images into benign and malignant classes. Similarly, Filipczuk et al. [253] also showed interest in the segmentation of nuclei and extracted twenty-five shape-based and texture-based features from the segmented nuclei of 737 cytology images of breast cancer. Based on these features, four different machine learning classifiers, namely, KNN (K-nearest neighbor), NB (Naive Bayes), DT (decision tree), and SVM (support vector machine), were trained for the classification of these cytological images into benign and malignant cases. Apart from nuclei segmentation [252, 253], other studies focused on the extraction of global features from the whole images. For instance, Zhang et al. [254] combined local binary patterns, statistics from the gray level co-occurrence matrix and the curvelet transform, and designed a cascade random space ensemble scheme (with rejection options) for an efficient classification of the microscopic biopsy images of breast cancer. Although the traditional machine learning approaches have made satisfactory performances in analyzing the histological images of breast cancer, their performances mainly rely on the selection of features on which they are trained. Furthermore, they might not be capable of effectively extracting and organizing the discriminative information from data [255].

In contrast to the traditional machine learning approaches based on handcrafted features, deep learning models have the ability to yield complicated and high-level features from images automatically [255]. Consequently, numerous recent studies employed deep learning approaches, with and without leveraging the pre-trained models, for the classification of breast cancer histopathology images. Of note, most of these studies employed BreakHis dataset [65] for the classification task. For instance, Spanhol et al. [198] employed CNN for the classification of breast cancer histopathology images and achieved 4 to 6 percentage points higher accuracy on BreakHis dataset [65] when using a variation of AlexNet [14]. Similarly, Bayramoglu et al. [256] utilized CNN in order to classify the histopathology images breast cancer irrespectively of their resolution using BreakHis dataset [65]. Specifically, the authors proposed single-task and multi-task CNN architectures; whereas the former was capable of predicting malignancy only and the latter was able to predict malignancy and magnification intensity of images simultaneously. These studies leveraging BreakHis dataset provided various state-of-the-art performances; however, they are relying on the same dataset. In this study, we followed the recent approaches of Araújo et al. [64] and Yan et al. [68] and presented a dataset for the classification of breast cancer histology images using deep learning models. However, our dataset contains only non-carcinoma and carcinoma

classes, unlike [64] and [68] which have four classes in their classification problem. The explanation of our dataset and proposed methodologies are comprehensively discussed in the next section 3.3.

3.3 Materials and methods

In this section, we introduced our dataset, followed by its preprocessing methodology and training, validation, and testing criteria along with the augmentation process. Then, we discussed the layout of the VGG model and finally, we described the architecture of our proposed ensemble architecture.

3.3.1 Data collection

We collected overall 544 whole slides images (WSI) from 80 patients suffering from breast cancer in the pathology department of Colsanitas Colombia University, Bogotá, Colombia. The tumor tissue fragments were fixed in formalin and embedded in paraffin. Subsequently, 4 mm cuts were made that were stained with hematoxylin and eosin ($H\&E$). For the Immunohistochemistry studies, the paraffin-embedded tissue sections were treated with xylene to render them diaphanous (the paraffin being removed later by passing it through decreasing alcohol concentrations until 100% water was reached). Rehydrated sections were rinsed in phosphate buffered saline (PBS) containing 1% Tween-20. For the detection of proteins, sections were heated in a high pH Envision FLEX target retrieval solution at 65 C for 20 min and then incubated for 20 min at room temperature in the same solution. Endogenous peroxidase activity (3% H_2O_2) and non-specific binding (33% foetal calf serum) were blocked and the sections were incubated overnight at 4 C with the following primary antibodies: anti-ER (estrogen receptor), anti-PR (progesterone receptor), anti-HER-2 (human epidermal growth factor receptor-2), anti-myosin, anti-Ki-67 (proliferation-associated biomarker). Next, an Ultra View universal DAB kit was used following the manufacturer's recommendations in conjunction with an automated staining procedure.

The tissue sections were then scanned at high resolution ($400\times$) using a Roche iScan HT scanner (https://diagnostics.roche.com/global/en/products/instruments/ventana-iscan-ht.html). These WSI images representing multiple cases from every patient were analyzed using H & E, hormone receptors, including ER, PR, $HER2$, myosin, and Ki-67. Next, two pathologists examined the digital whole slides of tissue stained with H & E and extracted 845 areas from WSI, among which 408 are non-carcinoma and 437 are carcinoma im-

Table 3.1 Characteristics of our proposed dataset.

Images	Quantity	Color Model	Staining
Carcinoma	437	RGB	H & E
Non-carcinoma	408	RGB	H & E
Total	845	RGB	H & E

ages.The carcinoma class has images of malignant tumors whereas the non-carcinoma class contains images of normal tisues as well as benign images of non-tumor glandular tissues. These areas were photographed at 200× (50 micrometers of resolution) and exported to png format using Qupath 0.1.2 software [257]. The dimensions of these images were noted as 1278×760 pixels. This dataset is considered to be balanced and its statistics are represented in Table 3.1. The main objective related to this dataset is the automatic classification of breast cancer histopathology images, most importantly the carcinoma images.

3.3.2 Preprocessing

The dataset used in this study contains histopathology images of breast cancer stained with H & E, which is widely used to assist pathologists during the microscopic assessment of tissue slides. However, it is difficult to maintain the same staining concentration through all the slides, which results in color differences among the acquired images. These contrast differences may adversely affect the training process of the CNN model and thus the color normalization is usually applied. In this study, we followed the recent studies [64, 68] and employed the approach proposed by [12] for colour normalization. In this method, images are first converted into optical density (OD) by using a logarithmic transformation. Next, singular value decomposition (SVD) is applied to OD tuples to obtain two-dimensional projections with higher variance. Then, the resulting color space transform is applied to the original images. Finally, the histogram of images is stretched in order to cover the lower 90% of data. However, the classification performance of our proposed model deteriorated upon using the normalized images, which is also comprehensively explained in [258]. Eventually, we omitted the stain normalization process and thus used the original images in this study. The example of original and normalized carcinoma images are shown in Figure 3.2.

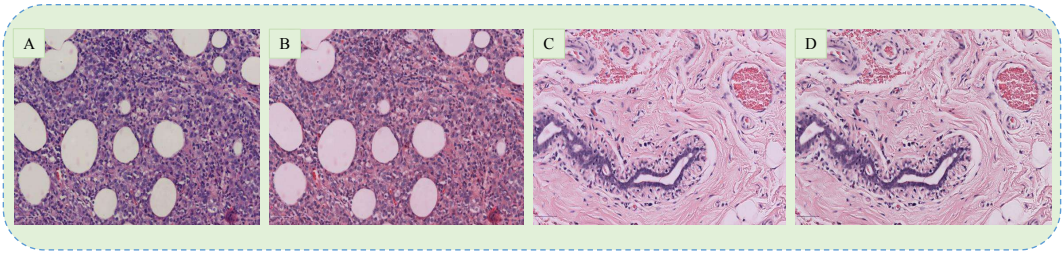


Figure 3.2 The examples of original (A,C) and normalized (B,D) images of carcinoma and non-carcinoma cases.

Table 3.2 Criteria for the selection of training, validation, and test images.

	No. of Images	Percentage
Training	540	64%
Validation	135	16%
Test	170	20%
Total	845	100%

3.3.3 Training criteria

For the individual and ensemble models, we selected 80% of images for training and the remaining 20% for testing purposes with the same percentage of carcinoma and non-carcinoma images. In this way, 675 images were used for training whereas the remaining 170 images were kept for testing the model. Following [259], we used 5-fold cross-validation on training images which means that 540 images were used for training and 135 images for validation purpose. Again, we have an equal percentage of non-carcinoma and carcinoma images in training and validation. These statistics about training, validation, and testing the models are depicted in Table 3.2.

3.3.4 Data augmentation

Image data augmentation is a technique used to expand the dataset by generating modified images during the training process. By employing the *ImageDataGenerator* provided by Keras deep learning library [260], we generate batches of tensor image data with real-time data augmentation. With this type of data augmentation, we want to ensure that our network, when trained, sees new variations of our data at each and every epoch. Firstly, an input batch of images is presented to the *ImageDataGenerator*, which then transforms each image in the batch by a series of random translations, rotations, etc. The rotation which we specified "rotation range = 40" corresponds to a

Table 3.3 Parameters of data augmentation.

Parameters of Image Augmentation	values
Zoom range	0.2
Rotation range	40
Width shift range	0.2
Height shift range	0.2
Horizontal flip	True
Fill mode	Reflect

random rotation angle between [−40, 40] degrees. We also set the "width and height shift range = 0.2" which specifies the upper bound of the fraction of the total width by which the image is to be randomly shifted, either towards the left or right for width or up or down for height. Of note, the rotation operation may rotate some pixels out of the image frame and leave behind empty pixels within the frame which must be filled. We used the "reflect mode" in order to fill these empty pixels. Finally, the randomly transformed batch is then returned to the calling function. All these parameters along with their values are shown in Table 3.3.

3.3.5 VGG architecture

Pretrained models usually help in a better initialization and convergence when the dataset is comparably small as compared to natural image datasets, and this result has been extensively used in other areas of medical imaging too [68]. To this end, we employed deep CNN-based pretrained model proposed by Visual Geometry Group (VGG) of Oxford University [9]. The VGG model is in fact one of the most influential contributions since it reinforced the notion that CNNs have to have a deep network of layers in order for this hierarchical representation of visual data to work. Although numerous follow-up works made improvements in VGG architecture; however, we used its early layouts in this study, called VGG16 and VGG19 architectures. These names are given because of the fact that VGG16 contains sixteen weight layers whereas VGG19 carries nineteen weight layers in their basic structures [9].

The complete framework of the VGG16 model is portrayed in Figure 3.3. It is composed of five convolutional blocks and every block has multiple convolution layers (with relu activation), together with a max-pooling layer. It strictly uses 3×3 filters with stride and pad of 1, along with 2×2 maxpooling layers with stride 2. The basic architecture of VGG19 is the same as that of VGG16, except three extra convolutional

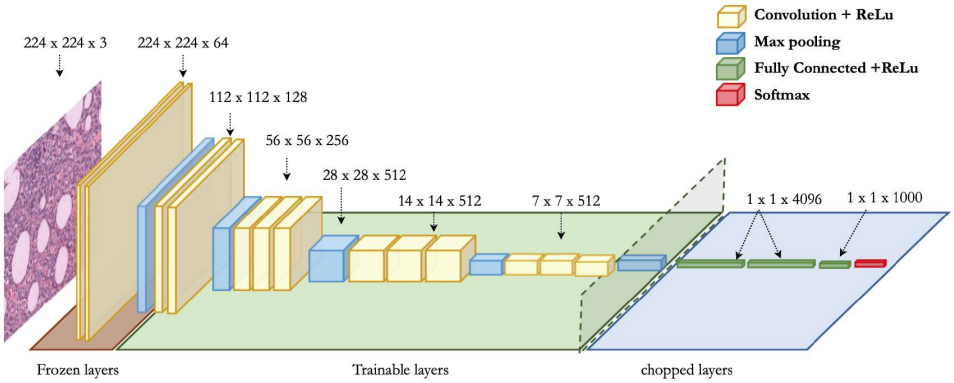


Figure 3.3 Representation of fine-tuned VGG16 architecture [9]. In fine-tuned VGG16 and VGG19 models, the first block (comprising two convolutional layers and one max-pooling layer) is frozen whereas the rest of layers are trainable. However, in fully-trained VGG16 and VGG19 models, all the five blocks are trainable.

layers. We tried four different approaches by using these two pretrained architectures. For fully-trained VGG16, we employed all the five blocks and replaced the last three layers by a single dense layer with 256 nodes, as shown in Figure 3.3. The final output layer is composed of binary cross-entropy loss function which is mathematically shown in Equation (3.1). Also, for fine-tuned VGG16, we froze the first block (with two convolutional layers and one max-pooling layer) and used the remaining four blocks for the training purpose. Again, we used one dense layer of 256 nodes along with the same loss function of binary cross-entropy. Similarly, for fully-trained VGG19, we trained all the blocks along with one dense layer of 128 nodes. Also, we froze the first block and trained the remaining blocks in the fine-tuned VGG19 model along with a single dense layer of 128 nodes. The final layer in case of VGG19 is also composed of binary cross-entropy loss function, as shown in Equation (3.1).

$$Binary\ cross\ entropy = -\frac{1}{m}\sum_{i}^{m}(y_i * log\,(p(y_i)) + (1-y_i) * log(1-p(y_i))) \tag{3.1}$$

3.3.6 Proposed ensemble approach

The architecture of our proposed ensemble approach is illustrated in Figure 3.4. It is composed of an ensemble of fine-tuned VGG16 and fine-tuned VGG19 models. First, for both models, training images (80%) are arranged in 5-folds, out of which four are used for training and one is used for model validation or evaluation. Of note,

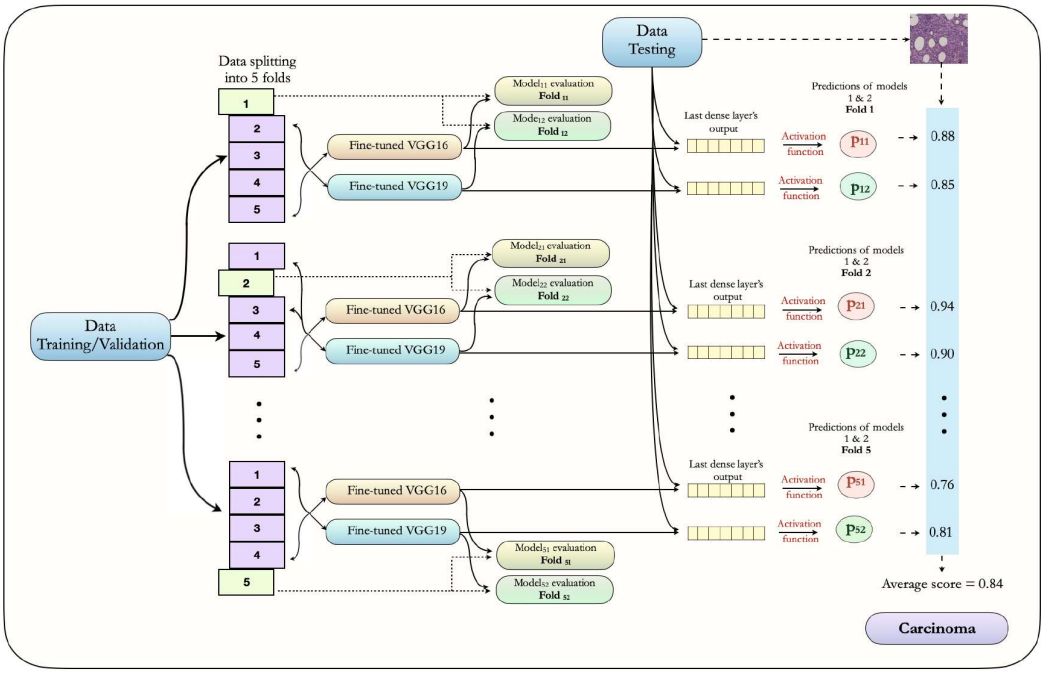


Figure 3.4 The proposed ensemble architecture using the fine-tuned VGG16 and VGG19 models along with 5-fold cross-validation approach.

these folds are mutually exclusive and have equal percentages of non-carcinoma and carcinoma images. Also, we used image augmentation during the training process, as described in Table 3.3. In every fold, we trained each model for 200 epochs; however, we saved weights of the best model only, based on a minimum value of loss function. In this way, we saved the weight for 5 folds for both models. Then, the test images (20%) are utilized in order to make the final prediction in the form of probabilities. The average probability for every class (non-carcinoma and carcinoma) is derived by taking the mean of ten probability values, obtained from 5-fold VGG16 and 5-fold VGG19 models (10 folds in total). In this way, we considered the average probability of both the models in order to classify images into non-carcinoma or carcinoma classes. The final results of our proposed ensemble deep learning approach are discussed in section 3.5.

3.4 Experimental setup

In this section, we explained the experimental environment, followed by the interpretation of evaluation metrics in our proposed model, and finally, we elucidated the tuning of hyperparameters.

3.4.1 Implementation

We implemented all the experiments related to this article by using *Python 3.7.6* along with *TensorFlow 2.1.0* and *Keras 2.2.4* installed on a standard PC with dual Nvidia GeForce GTX 2070 graphical processing unit (GPU) support. Moreover, this PC has a RAM capacity of 32.0 GB and holds a 3.60 GHz Intel® Core™ i9-9900K processor with 16 logical threads as well as 16 MB of cache memory.

3.4.2 Evaluation metrics

The overall performance of our proposed model relies on elements of confusion matrix, also called error matrix or contingency table. This evaluation matrix contains four terms, namely, True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). In our problem, TP refers to those images that were correctly classified as carcinoma and the FP represents the non-carcinoma images mistakenly classified as carcinoma. Whereas, the FN represents the images belonging to carcinoma class that were classified as non-carcinoma, and the TN refers to the non-carcinoma images correctly classified. The classification performance of our proposed model was evaluated on the testing set using four performance measures based on confusion matrix, namely, precision, sensitivity (recall), overall accuracy, and F1-score, using python scikit-learn module. These performance measures can be calculated as follow:

- Precision: It quantifies exactness of a model, and represents the ratio of carcinoma images accurately classified out of the union of predicted same-class images.

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

- Sensitivity: Sensitivity, also called "recall" computes completeness of a model. It represents the ratio of images accurately classified as carcinoma out of the total number of carcinoma images.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.3}$$

- Accuracy: It evaluates correctness of a model, and is the ratio of the number of images accurately classified out of the total number of testing images.

$$Accuracy = \frac{TP + FN}{TP + TN + FP + FN} \tag{3.4}$$

- F1-score: It represents the harmonic average of precision and recall, and is usually used for the optimization of a model towards either precision or recall.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.5}$$

3.4.3 Hyperparameter tuning

Neural networks have a powerful property of learning sophisticated connections between their inputs and outputs automatically [20]. However, some of these connections might be the result of sampling noise, so they can prevail during the training process but could not exist within the real test dataset. This issue may lead to overfitting problems and thus may degrade the prediction performance of a deep learning model [20]. For this very reason, we followed the tuning process of hyperparameters in order to get the generalized performance of our proposed model. The methodology used for the selection of optimal hyperparameters is as follows: First, we selected binary cross-entropy as a loss function for our binary classification problem. Then, Adam (adaptive moment estimation) algorithm [261, 262] was used during the training process in order to perform optimization through 200 epochs. At this stage, we tried three different learning rates (0.001, 0.0001, and 0.00001) and three different batch sizes (16, 32, and 64) while keeping in mind the values used in the recently published study [206, 262]. During the model training, our primary aim was to minimize the generalization gap between training loss and validation loss, and found that the batch size of 32 worked well together with the learning rate of 0.0001. Also, we used a dropout of 0.3 in order to prevent the model from overfitting during the training process [263]. Next, we saved the weights of five best models based on their minimal validation loss by using a 5-fold cross validation approach. Finally, we employed these weights for the class prediction on the test dataset. Of note, we used the convolutional filters, pooling filters, strides, and padding with their default values mentioned in the original VGG16 and VGG19 architectures [9]. All the optimal values of hyperparameters used in this study are provided in Table 3.4.

3.5 Results and discussion

In this section, we evaluated the performances of our proposed deep learning models by taking into consideration the average predicted probabilities. First, we highlighted the performance metrics of individual models and then we discussed the competitiveness of

Table 3.4 Hyperparameters used in the individual and an ensemble models.

Hyperparameters	VGG16 with Data Augmentation	VGG19 with Data Augmentation
Train approach	5-fold cross-validation	5-fold cross-validation
Optimizer	Adam	Adam
Loss function	Binary cross-entropy	Binary cross-entropy
Learning rate	0.0001	0.0001
Batch size	32	32
Convolution	3×3 with stride 1	3×3 with stride 1
Padding	Same	Same
Pooling	2×2 max-pooling with stride 2	2×2 max-pooling with stride 2
Epochs	200	200
Drop out	0.3	0.3
Regularizer	N/A	N/A
Architecture	Fully-trained and Fine-tuned	Fully-trained and Fine-tuned

our proposed models with recently published studies, especially in terms of carcinoma classification.

3.5.1 Results of VGG16 architecture

The performance metrics of fully-trained VGG16 architecture on our dataset are shown in Table 3.5. It can be noticed that these metrics vary across different folds although using the same test samples. Interestingly, the average recall value (sensitivity) of carcinoma class is noted as 94.55% (± 2.59). Also, the highest accuracy and F1 score are noted during Fold 1, in contrast to their lowest values during Fold 2. The overall accuracy of the fully-trained VGG16 model is 91.41 (± 3.40) along with the average F1 score of 91.38 (± 3.42). The accuracy curves of this model are depicted in Figure 3.5, whereas its loss curves are displayed in Figure 3.6.

Similar to fully-trained VGG16 architecture, the performance metrics of fine-tuned VGG16 framework are also presented in Table 3.5. Again, we used the same test set across all the folds. In this case, the average recall value of carcinoma class can be noticed as 94.09% (± 3.35). Moreover, the highest accuracy and F1 score are found during Fold 5, whereas their respective lowest values can be seen during Fold 1. Overall, the fine-tuned VGG16 models provided an average accuracy of 91.67% (± 3.69) as well as an average F1 score of 91.63% (± 3.69). The accuracy curves of this model are also illustrated in Figure 3.5, whereas its loss curves are presented in Figure 3.6. Lastly, the training and prediction times of fully-trained and fine-tuned VGG16 models are provided in Table 3.6.

Table 3.5 Performance metrics of VGG16 architecture on our dataset.

Architecture	Folds	Confusion Matrices			Performance Evaluation (%)				Average (%)	
		Predict → Actual ↓	NC	C	Precision	Recall	F1	Test	Acc.	F1
Fully-Trained VGG16	Fold 1	Non-carcinoma	75	7	97.40	91.46	94.34	82	**94.71**	**94.70**
		Carcinoma	2	86	92.47	97.73	95.03	88		
	Fold 2	Non-carcinoma	65	17	90.28	79.27	84.42	82	**85.88**	**85.80**
		Carcinoma	7	81	82.65	92.05	87.10	88		
	Fold 3	Non-carcinoma	73	9	93.59	89.02	91.25	82	**91.76**	**91.75**
		Carcinoma	5	83	90.22	94.32	92.22	88		
	Fold 4	Non-carcinoma	70	12	95.89	85.37	90.32	82	**91.18**	**91.13**
		Carcinoma	3	85	87.63	96.59	91.89	88		
	Fold 5	Non-carcinoma	78	4	91.76	95.12	93.41	82	**93.53**	**93.53**
		Carcinoma	7	81	95.29	92.05	93.64	88		
	Avg.	Non-carcinoma	–	–	93.78	88.05	90.75	82	**91.41**	**91.38**
		Carcinoma	–	–	89.65	94.55	91.98	88		
Fine-Tuned VGG16	Fold 1	Non-carcinoma	67	15	87.01	81.71	84.28	82	**85.29**	**85.27**
		Carcinoma	10	78	83.87	88.64	86.19	88		
	Fold 2	Non-carcinoma	74	8	92.50	90.24	91.36	82	**91.76**	**91.76**
		Carcinoma	6	82	91.11	93.18	92.13	88		
	Fold 3	Non-carcinoma	76	6	95.00	92.68	93.83	82	**94.12**	**94.11**
		Carcinoma	4	84	93.33	95.45	94.38	88		
	Fold 4	Non-carcinoma	73	9	96.05	89.02	92.41	82	**92.94**	**92.92**
		Carcinoma	3	85	90.43	96.59	93.41	88		
	Fold 5	Non-carcinoma	75	7	96.15	91.46	93.75	82	**94.12**	**94.11**
		Carcinoma	3	85	92.39	96.59	94.44	88		
	Avg.	Non-carcinoma	–	–	93.34	89.02	91.13	82	**91.67**	**91.63**
		Carcinoma	–	–	90.23	94.09	92.11	88		

Table 3.6 The training and prediction times of fully-trained and fine-tuned models.

Model	Single Training Time	5-Fold Training Time	Prediction Time
Fully-trained VGG16	17 min. 50 sec.	89 min	30 sec.
Fine-tuned VGG16	17 min. 25 sec.	87 min	31 sec.
Fully-trained VGG19	20 min. 40 sec.	103 min	35 sec.
Fine-tuned VGG19	19 min. 55 sec.	99 min	36 sec.

3.5.2 Results of VGG19 architecture

The performance metrics of fully-trained VGG19 architecture on our dataset are presented in Table 3.7. In this case, the average recall value (sensitivity) for carcinoma images is 95.45% (± 3.41) which is 0.9 percentage points higher than that of the fully-trained VGG16 model. Also, the maximum values of the accuracy and an F1 scores occurred during Fold 3, whereas their minimum values found during Fold 4. Finally,

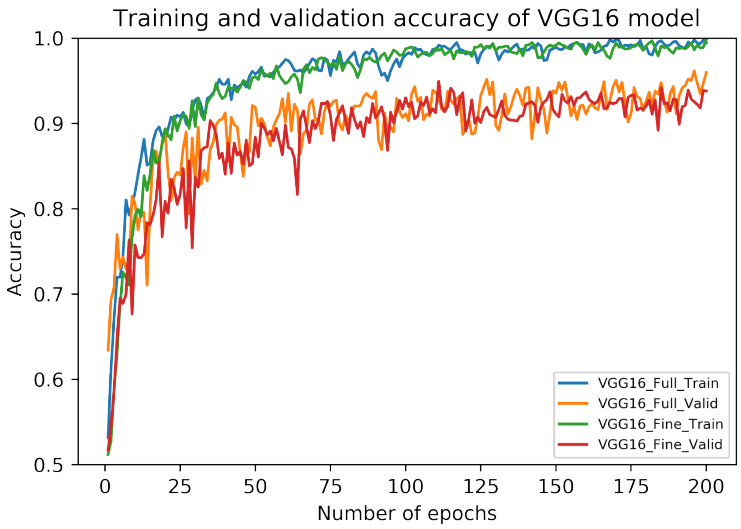


Figure 3.5 The training and validation accuracy curves of fully-trained and fine-tuned VGG16 models.

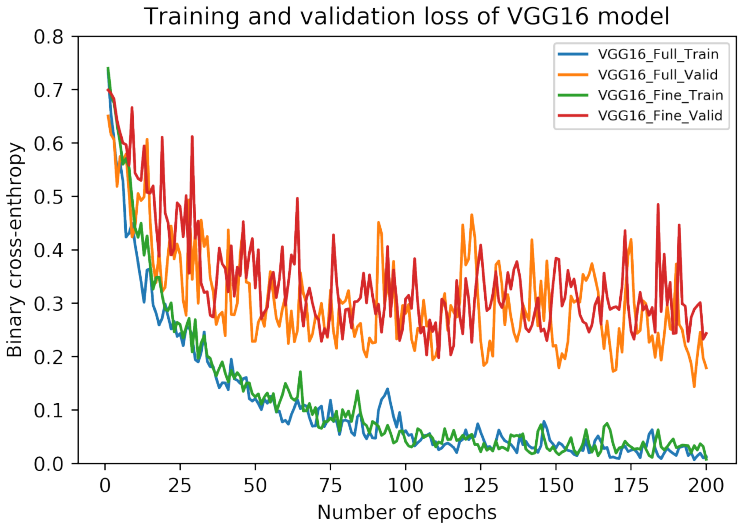


Figure 3.6 The training and validation loss curves of fully-trained and fine-tuned VGG16 models.

the overall accuracy of the fully-trained VGG19 model is 90.35% (± 1.35) in together with the average F1 score of 90.31% (± 1.35). The accuracy curves of this model are illustrated in Figure 3.7, whereas its loss curves are portrayed in Figure 3.8.

Similar to the fully-trained VGG19 model, the performance metrics of fine-tuned VGG19 architecture are portrayed in Table 3.7. The average recall value for carcinoma cases is 95.68% (± 3.15) which reflects 1.59 percentage points higher than that

Table 3.7 Performance metrics of VGG19 architecture on our dataset.

Architecture	Folds	Confusion Matrices			Performance Evaluation (%)				Average (%)	
		Predict → Actual ↓	NC	C	Precision	Recall	F1	Test	Acc.	F1
Fully-Trained VGG19	Fold 1	Non-carcinoma	66	16	98.51	80.49	88.59	82	**90.00**	**89.89**
		Carcinoma	1	87	84.47	98.86	91.10	88		
	Fold 2	Non-carcinoma	71	11	94.67	86.59	90.45	82	**91.18**	**91.15**
		Carcinoma	4	84	88.42	95.45	91.80	88		
	Fold 3	Non-carcinoma	69	13	98.57	84.15	90.79	82	**91.76**	**91.70**
		Carcinoma	1	87	87.00	98.86	92.55	88		
	Fold 4	Non-carcinoma	69	13	90.79	84.15	87.34	82	**88.24**	**88.21**
		Carcinoma	7	81	86.17	92.05	89.01	88		
	Fold 5	Non-carcinoma	73	9	91.25	89.02	90.12	82	**90.59**	**90.58**
		Carcinoma	7	81	90.00	92.05	91.01	88		
	Avg.	Non-carcinoma	–	–	94.76	84.88	89.46	82	**90.35**	**90.31**
		Carcinoma	–	–	87.21	95.45	91.09	88		
Fine-Tuned VGG19	Fold 1	Non-carcinoma	64	18	98.46	78.05	87.07	82	**88.82**	**88.67**
		Carcinoma	1	87	82.86	98.86	90.16	88		
	Fold 2	Non-carcinoma	75	7	93.75	91.46	92.59	82	**92.94**	**92.94**
		Carcinoma	5	83	92.22	94.32	93.26	88		
	Fold 3	Non-carcinoma	75	7	97.40	91.46	94.34	82	**94.71**	**94.70**
		Carcinoma	2	86	92.47	97.73	95.03	88		
	Fold 4	Non-carcinoma	76	6	96.20	92.68	94.41	82	**94.71**	**94.70**
		Carcinoma	3	85	93.41	96.59	94.97	88		
	Fold 5	Non-carcinoma	71	11	89.87	86.59	88.20	82	**88.82**	**88.81**
		Carcinoma	8	80	87.91	90.91	89.39	88		
	Avg.	Non-carcinoma	–	–	95.14	88.05	91.32	82	**92.00**	**91.96**
		Carcinoma	–	–	89.77	95.68	92.56	88		

of the fine-tuned VGG16 model. In this case, the highest values of accuracy and an F1 score are noted for Fold 3 and 4, whereas their low values occurred during Fold 1. The average accuracy and F1 score in this case are 91.67% (± 2.99) and 91.63% (± 3.03), respectively. The accuracy curves of this model are also presented in Figure 3.7, whereas its loss curves are shown in Figure 3.8. Finally, like the VGG16 models, the training and prediction times of fully-trained and fine-tuned VGG19 frameworks are also given in Table 3.6.

3.5.3 Results of ensemble VGG16 and VGG19

The performance metrics of the ensemble VGG16 and VGG19 framework are shown in Table 3.8. In this approach, we ensemble the fully-trained VGG16 and VGG19 architectures and the fine-tuned VGG16 and VGG19 frameworks by taking the average of output probabilities among all the folds in the aforementioned architectures. Inter-

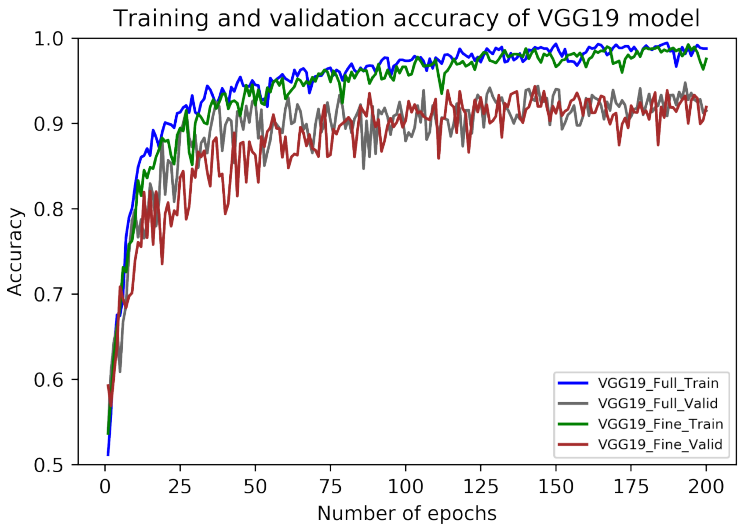


Figure 3.7 The training and validation accuracy curves of fully-trained and fine-tuned VGG19 models.

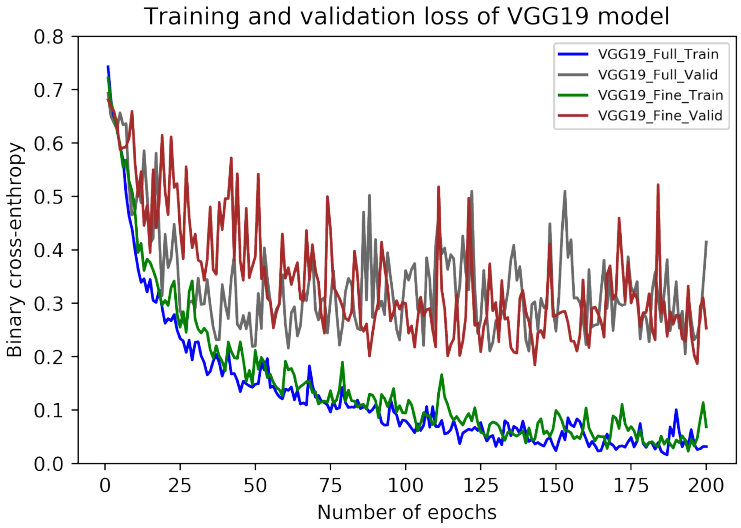


Figure 3.8 The training and validation loss curves of fully-trained and fine-tuned VGG19 models.

estingly, the recall value for the carcinoma class is noted as the same (97.73%) in both fully-trained and fine-tuned ensemble approaches. However, the fine-tuned approach offered high accuracy and F1 score (overall) compared to the fully-trained approach, as shown in Table 3.8.

Table 3.8 Performance metrics of ensemble VGG16 and VGG19 architectures.

Ensemble Method	Confusion Matrices			Performance Evaluation (%)				Average (%)	
	Predict → Actual ↓	NC	C	Precision	Recall	F1	Test	Accuracy	F1
Full-Trained VGG16+VGG19	Non-carcinoma	73	9	97.33	89.02	92.99	82	**93.53**	**93.51**
	Carcinoma	2	86	90.53	97.73	93.99	88		
Fine-Tuned VGG16+VGG19	Non-carcinoma	76	6	97.44	92.68	95.00	82	**95.29**	**95.29**
	Carcinoma	2	86	93.48	97.73	95.56	88		

3.5.4 Discussion

The effectiveness of our proposed ensembling approach can be compared with various state-of-the-art studies used for the classification of breast cancer histopathology images. Most of these novel deep learning approaches are based on BreakHis dataset [65]. For instance, Spanhol et al. [198] employed a variant of AlexNet [14] for the classification of benign and malignant images of BreakHis dataset [65]. The authors used sum, product and maximum fusions rules along with different patch sizes and reported an image level accuracy of 84.0% (± 3.2) for $200\times$ image magnification. In the following year, Bayramoglu et al. [256] proposed a magnification independent approach for BreakHis dataset. Specifically, the authors presented "single task CNN" and "multi-task CNN" frameworks, where the former predicts malignancy and the latter predicts malignancy as well as the magnification level in the benign and malignant images. For $200\times$ magnification, the authors reported an accuracy of 84.63% (± 2.72) and 82.56% (± 3.49) for single task CNN and multi-task CNN, respectively. Both of these studies [198, 256] reported better classification performance than the traditional handcrafted machine learning approaches. In comparison with Spanhol et al. [198] and Bayramoglu et al. [256], our approach shows better classification performance despite using a comparatively small dataset. Recently, Han et al. [264] proposed a structured deep learning model called class structure-based deep CNN (CSDCNN) for the classification of benign and malignant histopathology images of breast cancer, and reported an accuracy of 96.7% (± 2.0) on BreakHis dataset for $200\times$ magnification factor. Similarly, Nahid et al. [265] first used clustering algorithm in order to retrieve the statistical and geometrical clusters hidden in the histopathology images. The authors then evaluated the effect of deep CNN in together with short-term memory (LSTM) network for the efficient classification of benign and malignant images, and thus achieved an accuracy of 91.0% on BreakHis dataset for $200\times$ magnification. Lastly, Daniz et al. [266] employed fine-tuned AlexNet [14] and VGG16 [9] models for the classification of breast cancer histopathology images. The authors followed 5-fold cross-validation

approach and reported a maximum accuracy of 91.37% (± 1.72) when using fine-tuned AlexNet [14] on BreakHis dataset for $200\times$ magnification. These state-of-the-art studies [198, 256, 264–266] along with other novel frameworks are comprehensively reviewed in [267]. Although having a small dataset, our results are still competitive with the novel deep learning frameworks [198, 256, 264–267]. In summary, the results demonstrated that our proposed ensemble deep learning model can retrieve various multi-level and multi-scale features from histopathology images of breast cancer. It also became clear from the comparison process that the results of our proposed architecture is competitive with numerous state-of-the-art studies using comparably bigger datasets.

3.6 Chapter summary

In this study, we presented an ensemble deep learning approach for the classification of breast cancer histopathology images using our collected dataset. The main objective of this work was to effectively classify carcinoma images. We found that it could be better to use the average predicted probabilities of two individual models. To this end, we employed an ensemble of fine-tuned VGG16 and VGG19 models and achieved a relatively more robust model. The proposed ensemble approach provides competitive performance on the classification of complex-natured histopathology images of breast cancer. However, our collected dataset is comparatively small in contrast to the datasets used in numerous state-of-the-art studies. Also, our dataset contains merely two-class images. The future indications of this study include the extension of our dataset and the inclusion of images for multi-class classification problems. Also, other pretrained models need to be included in the future work. Finally, it will be interesting to apply similar ensemble criteria to histopathology images of different cancers, such as lung cancer.

Chapter 4

Study II: Multiclass Classification of Breast Cancer

This chapter elucidates numerous concepts related to the multiclass classification of breast cancer using a DL approach. The research findings of this study have been published in a peer-reviewed journal entitled "Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network" [6], as explained in the succeeding sections.

4.1 Introduction

According to Global Cancer Statistics 2020, breast cancer is the most common malignancy and the primary cause of cancer-related mortalities in the female population worldwide [17]. Specifically, 2.26 million (11.7% of the total cancer incidence) women were diagnosed, with a mortality of 0.69 million (6.9% of the total cancer deaths) during 2020 [17]. Therefore, the premature understanding of breast tumor pathophysiology is crucial, which may help in reducing the morbidity and mortality rates in women worldwide. This malignancy is considered a heterogeneous collection of diseases with distinct biological, clinical, and treatment response behaviors [268]. It mainly occurs due to abnormalities in the epithelial tissues of the breast and may invade the adjacent stroma, mammary duct, or lobes [22]. Although, the routine clinical analysis of breast cancer can be carried out by exploiting numerous radiology images, including ultrasound, mammography, and Magnetic Resonance Imaging (MRI) [23, 243]. Nevertheless, these non-invasive methodologies might not characterize the heterogeneous behaviors of breast tumors effectively. Therefore, the patho-

logical study is followed as a benchmark to comprehend the pathophysiology of breast tumors. In this method, tissue samples are collected and mounted on glass slides, and subsequently stained these slides for a better portrayal of tumoral morphological and inmunophenotypical characteristics [42]. After that, pathologists proceed with the microscopic examination of these slides to conclude a possible diagnosis of breast cancer [42]. The complete steps of the histopathological procedure have been discussed in [269] and [43].

However, the manual interpretation of histopathology images can be a tedious as well as a time-consuming process, and may lead to biased results. Moreover, the morphological criteria used during the manual analysis depend on the domain experience of the pathologists involved. For instance, one study revealed that the overall concordance rate of diagnostic interpretation among participating pathologists was around 75% [18]. To that end, the computer-aided diagnosis (CAD) [42, 23, 270] can help pathologists to improve diagnostic accuracy by reducing inter-pathologist variations during the diagnostic process of breast cancer. Nonetheless, traditional computerized diagnostic approaches, ranging from rule-based systems to machine learning methods, may not be sufficient to deal with the inter-class consistency and intra-class variability of complex-natured histopathology images of breast cancer. Furthermore, these conventional methodologies usually leverage feature extraction techniques such as scale-invariant feature transform [246], speed robust features [247] and local binary patterns [248], all of which are dependent on supervised information and hence may cause biased results when classifying these images. Therefore, the demand for an efficient and effective diagnosis yielded an advanced set of computational models based on numerous layers of nonlinear processing units, known as deep learning [19, 20].

In vision-related tasks, the convolutional neural network (CNN) [83] is considered superior to traditional multilayer perceptron due to having translational equivariance and translational invariance properties, the former resulting from parameter sharing and the latter from pooling operations [19, 20]. Especially, deep CNN architectures have made significant progress over the last decade among which AlexNet [14] is considered as the earliest deep CNN model to achieve decent accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) during 2012. Subsequently, VGG network [9] was presented with a novel idea of utilizing a deep network with small-sized convolutional filters, and it secured second position at the ILSVRC during 2014. At this point, Szegedy et al. [94] introduced the Inception architecture by staking multiple smaller convolutional filters to obtain an effective receptive field, and attained first place at the ILSVRC in 2014. The following year, He at al. [95] pointed out that

increasing the network depth after certain level may degrade its performance and they employed residual connections to overcome this problem, and earned first position at the ILSVRC in 2015. Consequently, numerous state-of-the-art studies leveraged the aforementioned architectures, pre-trained on ImageNet, to accurately classify breast cancer histopathology images using publicly available datasets, including BreakHis [65] and BACH [69] datasets. For instance, Jiang et al. [271] proposed a modified ResNet model [95] and achieved state-of-the-art accuracy for multiclass classification on BreaKHis dataset [65]. Similarly, the top studies of the BACH challenge [69] exploited either a single pre-trained network or an ensemble of pre-trained architectures for multiclass classification of breast microscopy images. Recently, Elmannai et al. [272] acknowledged the effectiveness of Inception modules and residual connections as feature extractors, and achieved state-of-the-art performance on the BACH dataset [69]. To this end, we leveraged the Xception model [16], stands for extreme inception, which is based on the efficient utilization of Inception and residual connections (see 4.2.5). As a feature extractor, it can provide consistent results in the classification of histopathology images of different magnification levels [273]. Our approach effectively utilizes the concepts introduced in [68, 273, 274, 16, 275] to extract salient features from histopathology images using the pre-trained Xception model [16] as a feature extractor.

The rationale and significance of this study are as follows: 1) To annotate and prepare a private dataset aimed to classify breast cancer histopathology images into normal tissue, benign lesion, in situ carcinoma, and invasive carcinoma [69]. Of note, the dataset prepared in this study is an extension of our previously published work on binary classification [43]. 2) To evaluate the performance of four widely used stain normalization methods [274]. 3) To propose a deep learning model based on multilevel features extracted from intermediate layers of the pre-trained Xception model [16]. 4) To optimize the proposed model for the accurate classification of breast cancer histopathology images on the original and normalized images, especially for carcinoma classes. To our knowledge, this is the first study that annotated a new private dataset, proposed a generalized as well as a computationally efficient model based on the Xception network [16] as a feature extractor, and evaluated the results of four widely used stain normalization approaches [274]. In summary, our proposed model provided consistent results for the definite classification of breast cancer histopathology images into four classes and also outperformed state-of-the-art results.

The remaining sections of this chapter are organized as follows. Section 4.2 describes materials and methods along with the proposed model. Section 4.3 explains

the findings, and section 4.4 compares the results of our proposed framework to state-of-the-art research. Finally, section 4.5 summarizes the conclusion as well as the future prospects of this work.

4.2 Materials and methods

In this section, we presented the dataset used in this study, followed by the analysis of four stain normalization techniques. Then, we elucidated the training criteria and in-place data augmentation used in this work. Next, we explained the proposed model and its implementation setup. Lastly, we described the model evaluation and the hyperparameter optimization of our proposed model.

4.2.1 Colsanitas dataset

In this study, we used the same dataset as presented in [43] which contains 544 whole slide images (WSIs), retrieved from 80 breast cancer patients at the pathology department of Colsanitas clinic with a dependence of the Sanitas University, Bogotá, Colombia. The protocols followed to convert histology samples into their corresponding digital images are discussed in [43], including collection and fixation, dehydration and clearing, paraffin embedding, staining and mounting, and digitalization [276]. The tissues were scanned at high magnification (40X) using a Roche iScan HT scanner (https://diagnostics.roche.com/global/en/products/instruments/ventana-iscan-ht.html). The WSI images are stained with hematoxylin and eosin (H&E) and illustrate multiple cases from each patient, as explained in [43]. It is worthy to mention that the dataset annotated for our previously published work [43] contained merely 845 images aimed at binary classification. Whereas the dataset annotated for the current study includes 2250 images formulated for multiclass classification [69]. Two experienced pathologists examined the H&E-stained WSI images and extracted 2250 images, including 600 normal tissues, 250 benign lesions, 250 in situ carcinoma, and 1150 invasive carcinoma. These images were exported as original pixels in .tiff format using Qupath 0.2.3 software [257]. The dimensions of these images are same as that of the BACH dataset [69] (2048×1536 pixels), with a pixel size of $0.46\mu m \times 0.46\mu m$. The complete characteristics of our created dataset is provided in Table 4.1. Also, the examples of normal tissue, benign lesion, in situ carcinoma, and invasive carcinoma images from the Colsanitas dataset are illustrated in Figure 4.1.

Table 4.1 Characteristics of our collected Colsanitas dataset.

Image	Quantity	Size ($w \times h \times c$)	Pixel size	Colour	Staining
Normal	600	$2048 \times 1536 \times 3$	$0.46\mu m \times 0.46\mu m$	RGB	H&E
Benign	250	$2048 \times 1536 \times 3$	$0.46\mu m \times 0.46\mu m$	RGB	H&E
In situ	250	$2048 \times 1536 \times 3$	$0.46\mu m \times 0.46\mu m$	RGB	H&E
Invasive	1150	$2048 \times 1536 \times 3$	$0.46\mu m \times 0.46\mu m$	RGB	H&E
Total	2250	$2048 \times 1536 \times 3$	$0.46\mu m \times 0.46\mu m$	RGB	H&E

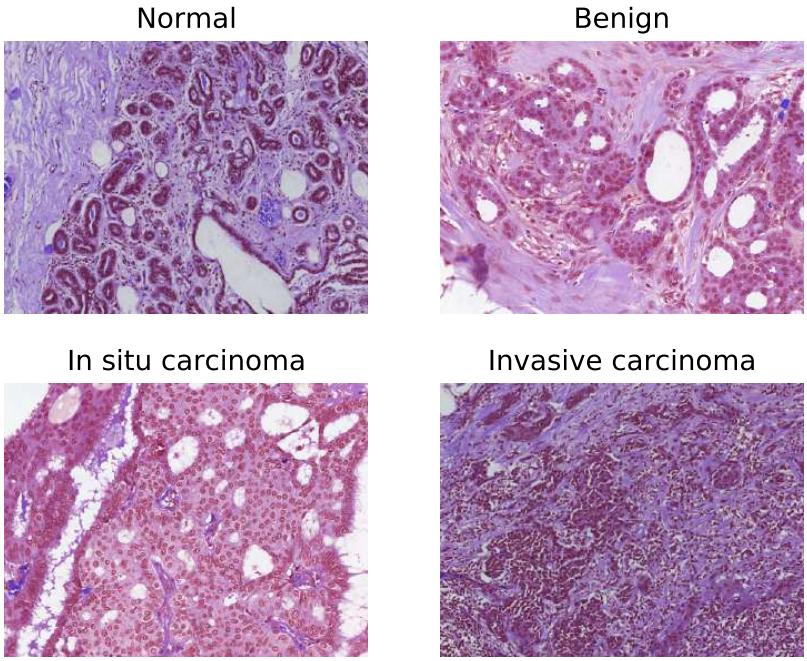


Figure 4.1 An example of H&E stained normal tissue, benign lesion, in situ carcinoma, and invasive carcinoma from our collected dataset.

4.2.2 Preprocessing

The datasets used in this work contain breast cancer histopathology images retrieved from H&E-stained whole-slide images. However, the stain concentration cannot be maintained in all the slides which may result in contrast differences among the exported images. These colour variations in acquired images may affect the performance of computer-aided diagnostic systems [276]. Lyon et al. [277] highlighted the need for the normalization of reagents and procedures in histopathological practice. Therefore, various colour preprocessing techniques, including colour-transfer and colour-decovolution, are introduced in the literature to standardize the stain appearance. For

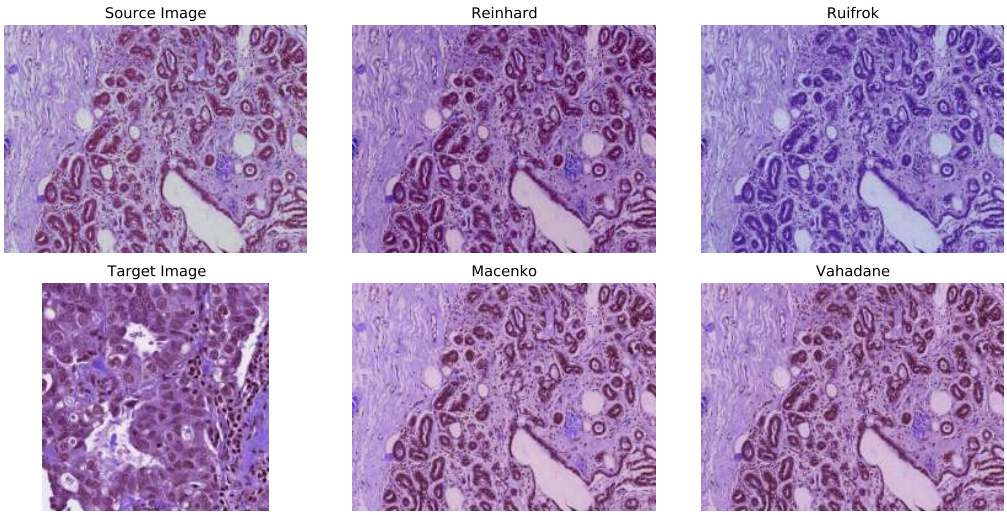


Figure 4.2 An example of H&E-stained source image, target image, and four pre-processed images resulting from Reinhard [10], Ruifrok [11], Macenko [12], and Vahadane [13] stain normalization.

instance, Reinhard et al. [10] developed a colour-transfer methodology in which RGB-format images are mapped to the colour distribution of a target image. In this method, a natural RGB image is first converted to a perceptual colour space with decorrelated axes, called $l\alpha\beta$. The mean values and standard deviations of each channel are then adjusted in both the images (source and target) in the colour space. Finally, the $l\alpha\beta$ colour space is converted to get a normalized RGB image. However, this type of global normalization is based on the unimodal distribution of pixels in each channel of colour space, which may not be appropriate when using multiple coloured stains. Therefore, numerous studies have concluded that stain separation prior to stain normalization has a relatively significant impact on the experimental results. For instance, Ruifrok et al. [11] introduced a colour-deconvolution method to separate the stains. For each stain in a histopathology image, the individual RGB channels are first transformed to their respective optical density (OD) values using Lambert-Beer's law. Then, the orthogonal transformation of OD values is carried out to get independent information regarding individual stains. Next, the OD vectors are normalized to achieve an unbiased absorption factor for each stain. After that, the normalized OD vectors are combined to form a normalized OD matrix. Lastly, a normalized image is created by leveraging the normalized OD matrix. In the following years, Macenko et al. [12] also followed a colour-deconvolution approach and concluded that H&E stains can be separated linearly in an OD colour space. First, a histology image is converted to its OD values using the loga-

rithmic transformation. Then, singular value decomposition (SVD) is applied to OD tuples to obtain a two-dimensional plane corresponding to the two largest singular values. Next, these OD-transformed pixels are projected onto the plan and normalized to unit length. After that, an angle is calculated at each point with respect to the first SVD direction, yielding a histogram that depicts the intensity of each stain. At this point, all of the intensity histograms are scaled to the same pseudo-maximum and compared to each other. Lastly, the concentration of each stain is determined by using the H&E matrix of the OD values and stain normalization is performed. Ultimately, using the H&E matrix with the normalized stain concentration, a normalized image is created. Recently, Vahadane et al. [13] developed a stain separation framework, called structure-preserving colour normalization (SPCN), which aimed to preserve the structure information of the source image. First, an RGB image is converted to OD values using Lambert-Beer's law. Then, for stain separation, a sparseness constraint (λ) is added to the optimization problem to reduce the solution space of the non-negative matrix factorization (NMF), called Sparse NMF (SNMF). In other words, a sparse constraint (λ) is added to the NMF to effectively separate the stains. Next, the proposed SNMF is used to estimate the color appearances and stain density maps of source and target images. Finally, a normalized image is generated by combining the scaled density map of a source image with the color appearance of a target image. Further theoretical and mathematical details of the aforementioned normalization techniques can be found in their respective original works [10–13] as well as in the review paper [278]. For the implementation, we utilized Warwick's Stain Normalization Toolbox (https://github.com/TissueImageAnalytics/tiatoolbox). Figure 4.2 depicts an example of a source image, a target image, and four normalized images using the above-mentioned practices.

4.2.3 Training procedure

We selected 80 percent of the images for training and the remaining 20 percent for testing, with an equal percentage of images from each of the four classes. Next, following [43, 279] we applied 5-fold cross-validation on the training dataset, which means that the training dataset (80%) is split into five equal subsets. Among these, four parts (64%) were used for training and one part (16%) was used for validating(evaluating) the model. Once finalizing the model, we included the validation part into the training dataset and retrained the model with all 80% of the images. Of note, the test subset is always the same for all the models. All these details are given in Table 4.2 and illustrated in Figure 4.3.

Table 4.2 Selection criteria for training, validation, and test images.

	Colsanitas dataset				Extended Colsanitas dataset				
	Nor.	Ben.	Ins.	Inv.	Nor.	Ben.	Ins.	Inv.	Percentage
Train	384	160	160	736	384	640	640	736	64%
Valid	96	40	40	184	96	160	160	184	16%
Test	120	50	50	230	120	50	50	230	20%
Total	600	250	250	1150	600	850	850	1150	100%

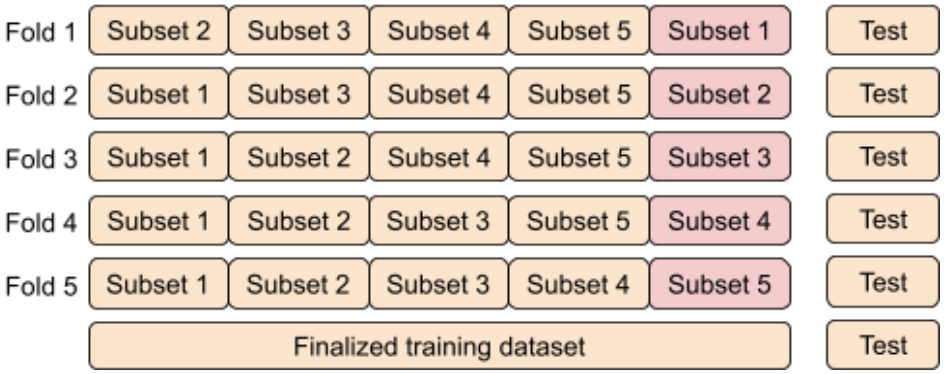


Figure 4.3 An illustration of the training process based on 5-fold cross-validation.

4.2.4 Data augmentation

In-place data augmentation or on-the-fly data augmentation is a technique in which a batch of original images is transformed into its new variation during each and every epoch of the training process. By employing this approach, we want to ensure that the model experiences new variations of input images at each epoch during the training process. To achieve this, we employed ImageDataGenerator provided by Tensorflow deep learning library [280]. The whole process of in-place data augmentation is as follows: 1) First, ImageDataGenerator takes a batch of input images. 2) Then, it transforms every image in the input batch by applying a series of random translations and rotations. In this work, we set "rotation range = 0.2" which corresponds to a random rotation between [-20, 20] degrees. However, it usually rotates some pixels out of the image frame, leaving empty pixels within the image, which we filled using "fill mode = reflect mode". Similarly, we specified " width and height shift range = 0.2" which indicates the percentage of width or height of the image to be shifted randomly, either towards left/right for width or up/down for the height. Also, we selected "zoom range = 0.2" which specifies random zoom-in operation. Of note, we did not apply horizon-

Table 4.3 Parameters and their values used in in-place data augmentation.

Parameters of ImageDataGenerator	Selected values
Zoom range	0.2
Rotation range	0.2
Width shift range	0.2
Height shift range	0.2
Horizontal flip	False
Vertical flip	False
Fill mode	Reflect

tal or vertical shifts operation because we already did these shifts when expanding the Colsanitas dataset. 3) Finally, it returns the randomly transformed batch of images. All the parameters and their selected values are provided in Table 4.3.

4.2.5 Proposed model

A straightforward way to increase the performance of a neural network is to increase the number of layers (length) and the number of units at each layer (width). However, the downsides of uniformly increasing network size include a larger number of parameters and computational resources [94]. Therefore, to address the issues of computational efficiency and the number of parameters, Szegedy et al. [94] introduced the concept of Inception in 2015. The inception module leveraged the idea of "network-in-network" [281] for dimensionality reduction. Also, it convolves an input with different sized filters and concatenates the output. Specifically, the Inception-v1 or GoogleNet, based on inception modules, utilized 12 times fewer parameters than AlexNet [14] and won the ILSVRC in 2014. In the following years, Inception-v2 or Batch Normalization [282], Inception-v3 [15], and Inception-v4 [283] were introduced which are considered to be the improved versions of Inception-v1 [94]. In addition to the Inception-v4 architecture, the Inception-ResNet-v1 and Inception-ResNet-v2 models were introduced, which utilized residual connections together with Inception modules [283]. Leveraging inception modules in conjunction with residual connections led to the development of an efficient architecture, called Xception network, which stands for "Extreme Inception" [16]. The Xception is an efficient network mainly depends on two crucial things: 1) depthwise separable convolution and 2) shortcuts between convolution blocks as in ResNet architecture [95]. Overall, the Xception model has 36 convolutional layers structured together into 14 modules, with each module having a

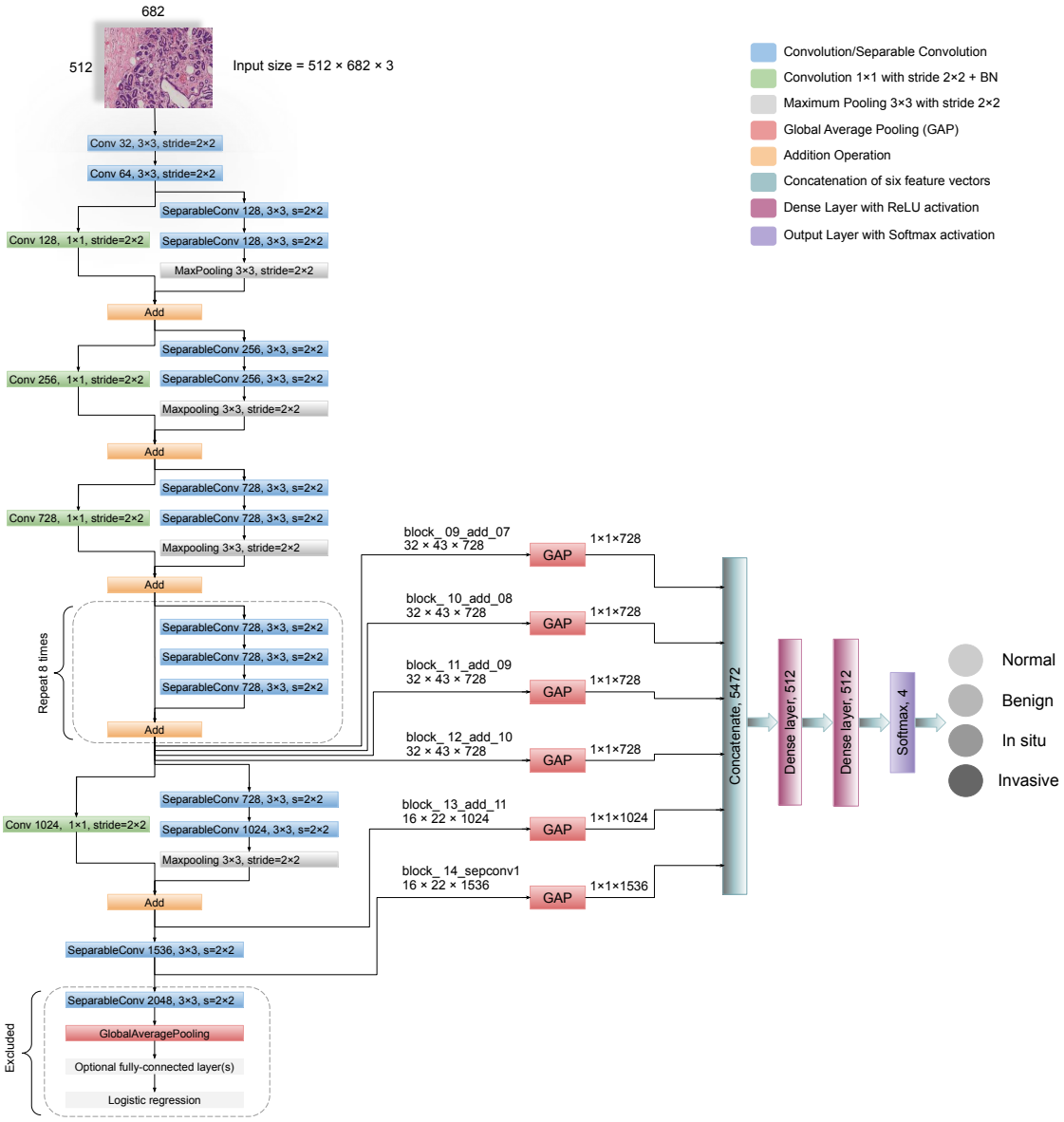


Figure 4.4 The complete framework of our proposed model is illustrated along with all the layers. For every input image, six different features are extracted followed by the global average pooling. These multilevel features are then concatenated (merged) horizontally to form a single vector of $1 \times 1 \times 5472$ which is used for classification.

linear residual connection around it, except the first and last one, as shown in Figure 4.4.

Our proposed model leveraged the Xception network [16] to retrieve robust and abstract features from the intermediate layers, as shown in Figure 4.4. First, the model takes an RGB image of height 512 and width 682 at its input layer. Of note, we re-

duced the dimension of original images in such a way that the ratio of height and width remained the same. In this way, we preserve the original structure of images, unlike [275] that used the dimension of 512×512. Then, following [16, 275, 68], we utilized global average pooling (GAP) on six different layers to obtain the corresponding feature vectors. GAP layers help to decrease the number of parameters and to reduce the overfitting [273]. It is worth mentioning that before finalizing these six layers, we checked the results of different layers from the last seven blocks of the Xception network on the original dataset using k-fold cross-validation. We found that these six layers offered consistent performance in classifying each class with minimal variation. After that, we concatenated (merged) these vectors horizontally to acquire the finalized vector of the dimension 5472 pixels for each image. After the images are converted to their corresponding feature vectors, we trained two dense layers with each having the dimension of 512 nodes and Rectified Linear Unit (ReLU) activation. Lastly, the output layer is comprised of four nodes with Softmax activation and is used for the classification of the given images into four categories. The Softmax function transforms a vector k real-valued numbers into a vector of k probabilities that sum to 1, as explained in [20]. In our case, the input to the Softmax function is a real-valued vector with $k = 4$, whereas its output is a vector of $k = 4$ probabilities that sum to 1. The mathematical explanation of softmax function is given in equation 4.1 and is described in [20].

$$Softmax(\boldsymbol{z})_i = \frac{exp(z_i)}{\sum_{j=1}^{k} exp(z_j)} \tag{4.1}$$

Where $\boldsymbol{z} = (z_1, z_2, z_3, z_4)$ is the input vector to the Softmax function and k is the number of classes. Moreover, $exp(z_i)$ shows the exponential of the i^{th} real-valued number in the input vector and its value is always positive. Laslty, the normalization term $\sum_{j=1}^{k} exp(z_j)$ depicts the sum of exponential of all the input real-valued numbers and its value is also always positive. In this way, we get a vector of probabilities that sums to 1.

4.2.6 Implementation setup

We implemented all the experiments using Python version 3.8.5 and TensorFlow 2.4.1 [280], installed on a standard computer machine with two Nvidia GeForce GTX 2070 graphical processing units (GPUs) support. Furthermore, the machine has a RAM of 32.0 GB and holds a 3.60 GHz Intel® Core™ i9-9900K processor with 16 logical

threads and 16 MB of cache memory. We followed the distributed training approach of TensorFlow [280] by using both the GPUs using the "tf.distribute.MirroredStrategy (devices=['/gpu:0','/gpu:1'])" strategy.

4.2.7 Model evaluation

The classification performance of the proposed framework leverages the elements of confusion matrix, also known as contingency table [43, 284]. For multiclass classification problem, we defined the elements of the confusion matrix in terms of the target class and non-target class, which can be applied to every individual class [284]. For instance, the target class could be invasive and non-target class could be non-invasive. True Positive (TP) refers to the images that are correctly classified as the target class (invasive), and False Positive (FP) shows the non-target images (non-invasives) that are falsely classified as the target class (invasive). Whereas, False Negative (FN) indicates the images of target class (invasive) classified as non-target class (non-invasive), and True Negative (TN) denotes the correctly classified non-target images (non-invasive). Of note, FP is also called *type I error* and FN is also called *type II error* in the literature. Furthermore, following [285], we assessed the performance of our proposed model using receiver operating characteristic (ROC) curves and precision-recall (PR) curves along with their area under the curve (AUC) values for every class (one-vs-rest method) for the original and normalized datasets. Lastly, we computed the Cohen's kappa statistic for the original as well as normalized datasets.

- Precision: It calculates the exactness of a model and defines the ratio of images correctly classified as the target class (invasive) out of all predicted same-class images.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

- Sensitivity: Sensitivity, also known as recall, evaluates the completeness of a model. It determines the ratio of images accurately classified as the target class (invasive) out of all actual same-class images.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4.3}$$

- Accuracy: It computes the correctness of a model and is defined as the proportion of the number of accurately classified images out of total actual test images.

$$\text{Accuracy} = \frac{TP + FN}{TP + TN + FP + FN} \tag{4.4}$$

- F1-score: It indicates the harmonic average of precision and recall and is commonly employed to optimize a model for either precision or recall.

$$\text{F1-score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.5}$$

- ROC Curve: The ROC curve shows a relationship between true positive rate (TPR) and false positive rate (FPR) at different thresholds. TPR is also called sensitivity or recall, whereas FPR is equivalent to 1-specificity. An ROC curve depicts that increasing TPR results in also increasing FPR and vice versa. The mathematical formula of TPR is shown in equation 4.3 whereas that of FPR is provided in equation 4.6.

$$\text{FPR} = \frac{FP}{FP + TN} \tag{4.6}$$

- PR Curve: The PR curve shows an inverse relationship between precision and recall at different thresholds. A PR curve illustrates that increasing precision value results in decreasing recall score and vice versa. The mathematical formula of precision is given in equation 4.2 whereas that of recall is given in equation 4.3.

- Cohen's kappa: It calculates the degree of agreement between the true values and predicted values. It is widely used in to handle multiclass and imbalanced dataset problems. Its mathematical formula is provided in equation 4.7 where p_o and p_e represent observed and expected agreements, respectively.

$$\text{k} = \frac{p_o - p_e}{1 - p_e} \tag{4.7}$$

4.2.8 Hyperparameter optimization

Neural networks can learn complicated patterns between their inputs and outputs automatically [19, 20]. Many of these input-output connections, however, may be the result of sampling noise that prevailed during training but may not exist in the test dataset. This can result in the overfitting problem and thus reduce the prediction ability of a deep learning model. To that end, it is crucial to follow the process of hyperparameter tuning to obtain the generalized predictive performance of the proposed network. In this study, we followed the 5-fold cross-validation approach (see 4.2.3) to get the

Table 4.4 The optimal hyperparameters of our proposed model.

Hyperparameters	Optimal values
Train approach	5-fold cross-validation
Loss function	Categorical cross-entropy
Optimizer	Adam
Learning rate	0.00001
Batch size	64
Convolution	$1 \times 1, 3 \times 3, 5 \times 5$
Maxpooling	2×2 with stride 2
Epochs	1000
Dropout	0.1
Regularizer	$L2$

best set of hyperparameters. The procedure followed for obtaining the optimum hyperparameters values is as follows: For our multiclass classification task, we first selected categorical cross-entropy as an objective function. Then, we employed Adam (adaptive moment estimation) algorithm [261, 43] during the training to optimize the model through 1000 epochs. At this point, we checked three variants of learning rates $(0.001, 0.0001, 0.00001)$ and three distinct batch sizes $(16, 32, 64)$ based on recently published studies [275, 43]. We found that the learning rate of 0.00001 together with a batch size of 64 worked well in reducing the generalization gap between training and validation loss. Next, we saved the weights of five models resulted from the lowest validation loss, and evaluate the predictive performance of each model on the unseen test dataset. Importantly, we aimed to maximize the mean value of test accuracy while minimizing the standard deviation after checking the predictive abilities of five individual models. For the final model, we trained the proposed framework with all the training images (training and validation) and saved the weights of the optimum model based on the minimum validation loss. Lastly, we employed these weights to predict the classes of the test images. Importantly, we used the default parameters specified in the original architecture of the Xception paper for the convolutional filters, pooling filters, strides, and padding [16]. All the hyperparameters and their optimal values used in this study are presented in Table 4.4.

4.3 Results

In this section, we explained and compared the classification performance of our proposed framework by considering the original (unnormalized) and normalized images.

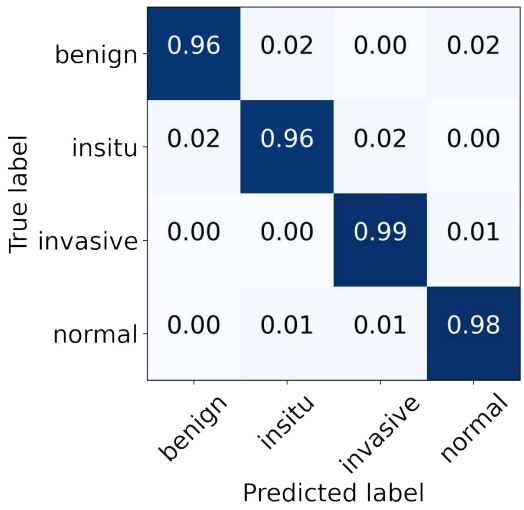


Figure 4.5 The final normalized confusion matrix of original dataset.

4.3.1 Results without normalization

For the original (unnormalized) dataset, the performance metrics of our proposed model are provided in Table 4.5. During the cross-validation, we reported the highest accuracy of 96.88% during folds 1, 2, and 4, whereas the lowest accuracy of 95.33% during fold 5, which led to a mean accuracy of 96.22% (± 0.66). The finalized model offered an accuracy value of 98.00%, as shown in Table 4.5. Specifically, for in situ and invasive carcinomas, we reported sensitivity values of 96.00% and 99.00%, respectively. Similarly, for benign lesions, we found a sensitivity score of 96.00% which is similar to that of in situ carcinoma. The finalized results of all the four classes using the original dataset are shown in Figure 4.5. Furthermore, the ROC and PR curves for every class of the original dataset along with their AUC scores are depicted in Figure 4.6. The AUC-ROC values vary from 0.998 to 0.999 whereas the AUC-PR values range from 0.990 to 0.999, as displayed in Figure 4.6. Of note, the accuracy and loss curves of the original dataset are provided with every normalized dataset for better visualization and comparison, which are discussed within the next subsections.

4.3.2 Results of Reinhard normalization

For the Reinhard normalization, the performance metrics of our proposed architecture are given in Table 4.6. During the cross-validation, we noted higher accuracy of 97.11% at fold 4 and lower accuracy of 95.33% at fold 5, yielding a mean accuracy of 96.44% (± 0.68). The finalized model attained an accuracy of 97.33%, as stated in

Table 4.5 Evaluation metrics of our proposed model using the original dataset.

Folds	Confusion Matrices					Performance Evaluation					
	Predict → Actual ↓	Ben.	Ins.	Inv.	Nor.	Prec.	Rec.	F1	Test	Accuracy	Kappa
Fold 1	Benign	43	3	2	2	1.00	0.86	0.92	50	96.88%	0.951
	In situ	0	49	1	0	0.91	0.98	0.94	50		
	Invasive	0	1	226	3	0.98	0.98	0.98	230		
	Normal	0	1	1	118	0.96	0.98	0.97	120		
Fold 2	Benign	48	1	1	0	0.96	0.96	0.96	50	96.88%	0.952
	In situ	2	48	0	0	0.89	0.96	0.92	50		
	Invasive	0	4	222	4	0.99	0.97	0.98	230		
	Normal	0	1	1	118	0.97	0.98	0.98	120		
Fold 3	Benign	47	0	2	1	0.98	0.94	0.96	50	96.00%	0.937
	In situ	1	48	1	0	0.96	0.96	0.96	50		
	Invasive	0	1	226	3	0.95	0.98	0.97	230		
	Normal	0	1	8	111	0.97	0.93	0.94	120		
Fold 4	Benign	47	1	0	2	0.94	0.94	0.94	50	96.88%	0.952
	In situ	2	47	1	0	0.94	0.94	0.94	50		
	Invasive	0	1	226	3	0.99	0.98	0.98	230		
	Normal	1	1	2	116	0.96	0.97	0.96	120		
Fold 5	Benign	46	2	1	1	0.85	0.92	0.88	50	95.33%	0.928
	In situ	3	47	0	0	0.90	0.94	0.92	50		
	Invasive	0	2	224	4	0.99	0.97	0.98	230		
	Normal	5	1	2	112	0.96	0.93	0.95	120		
Final	Benign	48	1	0	1	0.98	0.96	0.97	50	**98.00%**	**0.969**
	In situ	1	48	1	0	0.96	0.96	0.96	50		
	Invasive	0	0	227	3	0.99	0.99	0.99	230		
	Normal	0	1	1	118	0.97	0.98	0.98	120		

Table 4.6. Especially for in situ carcinoma, we observed a sensitivity of 96.00% which is equivalent to that of the original dataset. Whereas for invasive carcinoma, we noted a sensitivity of 98.00% which is 1.00% lower than the original dataset. These finalized results of all the four classes using the Reinhard-based normalized dataset are portrayed in Figure 4.7. In addition, the ROC and PR curves for each class of the Reinhard normalization together with their AUC values are illustrated in Figure 4.8. In this case, the AUC-ROC values range from 0.997 to 0.999 whereas AUC-PR scores vary from

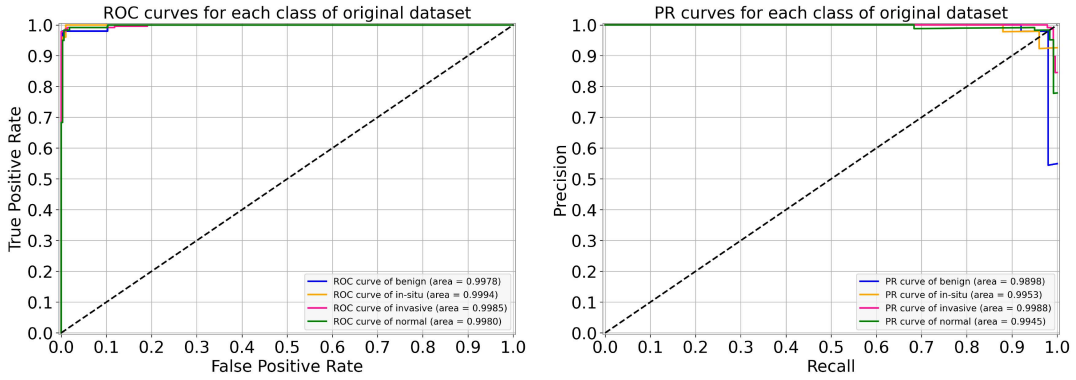


Figure 4.6 For the original dataset, the left side shows ROC curves for each class with an average AUC-ROC of 0.998. Whereas the right side depicts its PR curves for every class with a mean AUC-PR of 0.995.

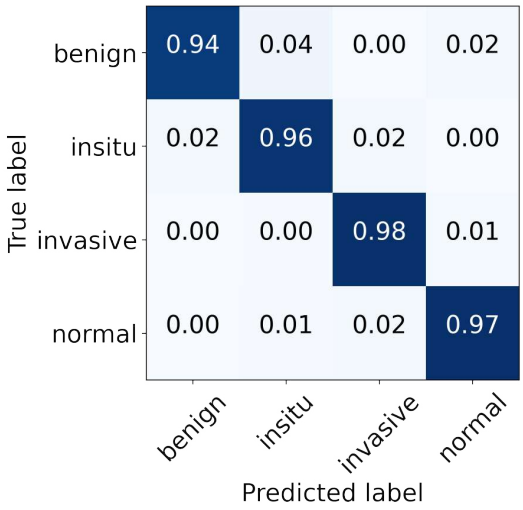


Figure 4.7 The final normalized confusion matrix of Reinhard dataset.

0.989 to 0.998, as shown in Figure 4.8. The accuracy curves of Reinhard normalization along with the original ones are shown on the left side of Figure 4.9, whereas their corresponding loss curves are depicted on the right side of Figure 4.9. It can be seen that there is no significant difference in these curves. Based on these results, we concluded that although the Reinhard normalization achieved a competitive classification performance, it could not outperform results of the original (unnormalized) dataset.

4.3.3 Results of Ruifrok normalization

For the Ruifrok normalization, the performance metrics of our proposed framework are presented in Table 4.7. During the cross-validation, we observed a highest accuracy of

Table 4.6 Evaluation metrics of our proposed model using Reinhard normalization.

Folds	Confusion Matrices					Performance Evaluation					
	Predict → Actual ↓	Ben.	Ins.	Inv.	Nor.	Prec.	Rec.	F1	Test	Accuracy	Kappa
Fold 1	Benign	43	4	1	2	1.00	0.86	0.92	50	96.44%	0.945
	In situ	0	49	1	0	0.89	0.98	0.93	50		
	Invasive	0	1	225	4	0.98	0.98	0.98	230		
	Normal	0	1	2	117	0.95	0.97	0.96	120		
Fold 2	Benign	46	2	0	2	0.98	0.92	0.95	50	96.88%	0.952
	In situ	1	49	0	0	0.89	0.98	0.93	50		
	Invasive	0	3	223	4	1.00	0.97	0.98	230		
	Normal	0	1	1	118	0.95	0.98	0.97	120		
Fold 3	Benign	47	2	1	0	0.98	0.94	0.96	50	96.44%	0.944
	In situ	1	48	1	0	0.92	0.96	0.94	50		
	Invasive	0	1	226	3	0.97	0.98	0.97	230		
	Normal	0	1	6	113	0.97	0.94	0.96	120		
Fold 4	Benign	47	2	0	1	0.96	0.94	0.95	50	97.11%	0.955
	In situ	1	47	1	1	0.94	0.94	0.94	50		
	Invasive	0	0	227	3	0.99	0.99	0.99	230		
	Normal	1	1	2	116	0.96	0.97	0.96	120		
Fold 5	Benign	47	3	0	0	0.87	0.94	0.90	50	95.33%	0.928
	In situ	2	47	0	1	0.87	0.94	0.90	50		
	Invasive	2	2	223	3	0.99	0.97	0.98	230		
	Normal	3	2	3	112	0.97	0.93	0.95	120		
Final	Benign	47	2	0	1	0.98	0.94	0.96	50	**97.33%**	**0.959**
	In situ	1	48	1	0	0.92	0.96	0.94	50		
	Invasive	0	1	226	3	0.99	0.98	0.98	230		
	Normal	0	1	2	117	0.97	0.97	0.97	120		

96.88% during fold 2 and a lowest accuracy of 96.00% during fold 5, which resulted in a mean accuracy of 96.31% ($\pm$0.37). The finalized model yielded an accuracy of 97.33%, as mentioned in Table 4.7. Particularly, the sensitivity for in situ class is 96.00%, which is equal to both the original and the Reinhard normalization. Likewise, the sensitivity for invasive class is 99.00%, which is the same as that of the original but 1.00% higher than the Reinhard normalization. These optimal results of all the classes for the Ruifrok-based normalized dataset are depicted in Figure 4.10. Moreover, the

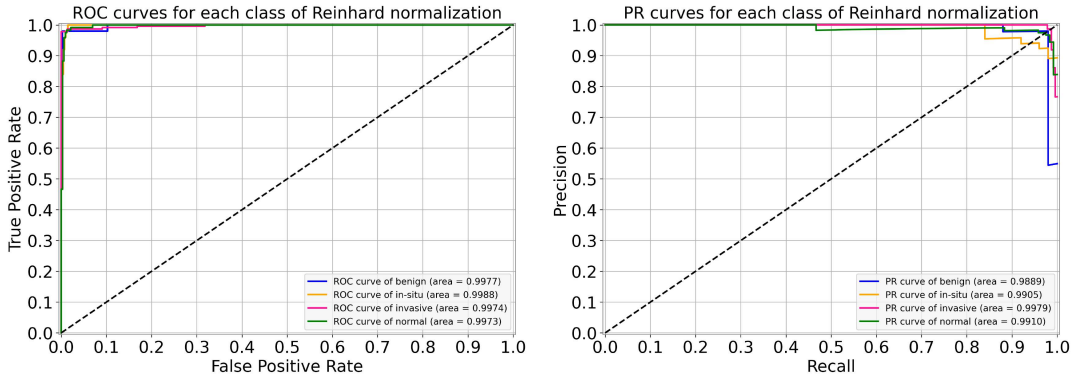


Figure 4.8 For Reinhard normalization, the left-hand side represents ROC curves for each class with an average AUC-ROC of 0.998. Whereas the right-hand side depicts its PR curves for every class with a mean AUC-PR of 0.992.

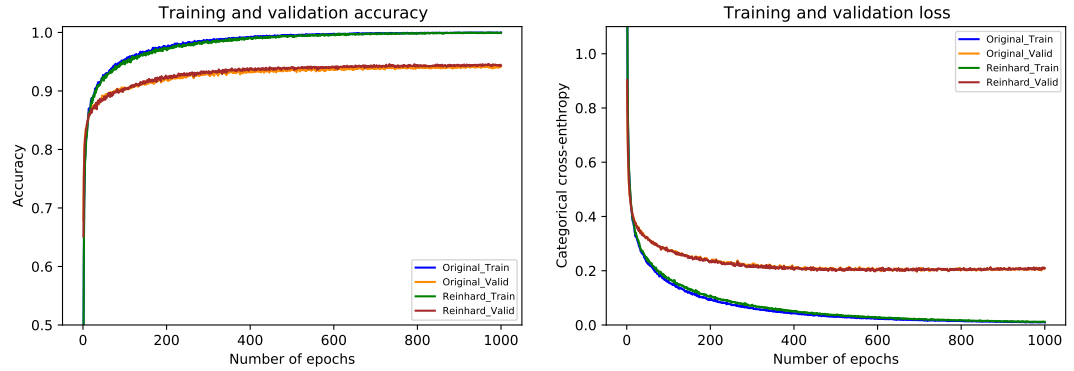


Figure 4.9 The left-hand side shows a comparison of training and validation accuracy curves of the original dataset and Reinhard normalization. Whereas the right-hand side depicts a comparison of training and validation loss curves of the original dataset and Reinhard normalization.

ROC and PR curves for an individual class of the Ruifrok normalization in conjunction with their AUC scores are provided in Figure 4.11. In this case, the AUC-ROC values range from 0.997 to 0.999 whereas the AUC-PR scores range from 0.980 to 0.999, as demonstrated in Figure 4.11. The comparison of accuracy curves, in this case, is shown on the left block of Figure 4.12, whereas their corresponding loss curves are illustrated on the right block of Figure 4.12. Like the Reinhard normalization, it can be seen that there is no significant difference in these curves. Thus, it can be concluded that the classification performance using the Ruifrok normalization is the same as Reinhard normalization in terms of accuracy.

Table 4.7 Evaluation metrics of our proposed model using Ruifrok normalization.

Folds	Confusion Matrices					Performance Evaluation					
	Predict → Actual ↓	Ben.	Ins.	Inv.	Nor.	Prec.	Rec.	F1	Test	Accuracy	Kappa
Fold 1	Benign	42	6	0	2	0.98	0.84	0.90	50	96.22%	0.941
	In situ	1	48	1	0	0.84	0.96	0.90	50		
	Invasive	0	2	225	3	0.99	0.98	0.98	230		
	Normal	0	1	1	118	0.96	0.98	0.97	120		
Fold 2	Benign	48	2	0	0	0.98	0.96	0.97	50	96.88%	0.952
	In situ	1	48	1	0	0.89	0.96	0.92	50		
	Invasive	0	3	222	5	0.99	0.97	0.98	230		
	Normal	0	1	1	118	0.96	0.98	0.97	120		
Fold 3	Benign	44	3	0	3	0.96	0.88	0.92	50	96.00%	0.937
	In situ	2	45	3	0	0.94	0.90	0.92	50		
	Invasive	0	0	226	4	0.97	0.98	0.98	230		
	Normal	0	0	3	117	0.94	0.97	0.96	120		
Fold 4	Benign	46	3	0	1	0.94	0.92	0.93	50	96.44%	0.945
	In situ	1	47	2	0	0.94	0.94	0.94	50		
	Invasive	1	0	224	5	0.98	0.97	0.98	230		
	Normal	1	0	2	117	0.95	0.97	0.96	120		
Fold 5	Benign	47	3	0	0	0.90	0.94	0.92	50	96.00%	0.938
	In situ	1	47	2	0	0.92	0.94	0.93	50		
	Invasive	1	1	224	4	0.98	0.97	0.98	230		
	Normal	3	0	3	114	0.97	0.95	0.96	120		
Final	Benign	45	3	1	1	0.98	0.90	0.94	50	**97.33%**	**0.958**
	In situ	1	48	1	0	0.94	0.96	0.95	50		
	Invasive	0	0	227	3	0.98	0.99	0.98	230		
	Normal	0	0	2	118	0.97	0.98	0.98	120		

4.3.4 Results of Macenko normalization

For the Macenko normalization, the performance metrics of our proposed system are provided in Table 4.8. During the cross-validation, we observed the uppermost accuracy of 97.11% in fold 4 as well as the lowermost accuracy of 96.00% in fold 3, resulting in a mean accuracy of 96.10% (± 0.88). The finalized model got an accuracy of 97.78%, as given in Table 4.8. In particular, the sensitivity values for in situ

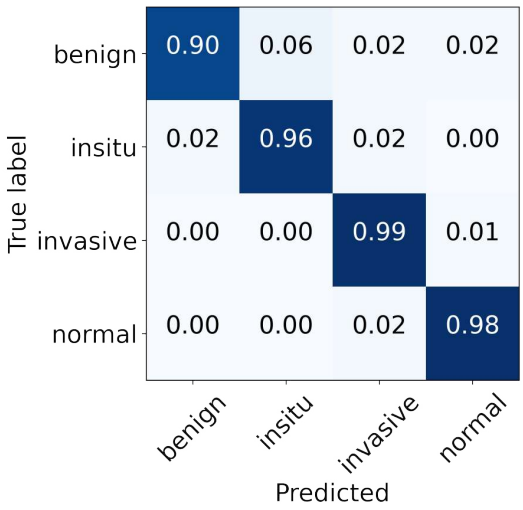


Figure 4.10 The final normalized confusion matrix of Ruifrok dataset.

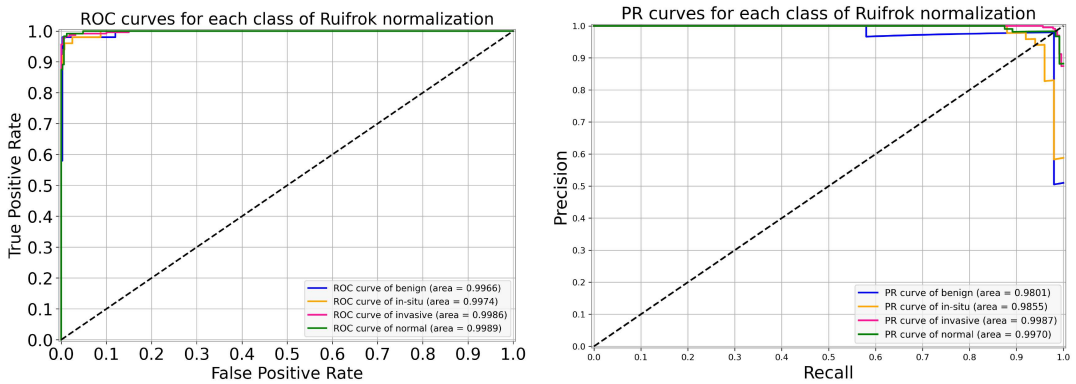


Figure 4.11 For Ruifrok normalization, the left side represents ROC curves for an individual class with an average AUC-ROC of 0.998. Whereas the right side depicts its PR curves for every class with a mean AUC-PR of 0.990.

and invasive carcinomas, in this case, are 96% and 99%, which are equal to that of the original dataset. These optimum results for all four classes are illustrated in Figure 4.13. Besides, the ROC and PR curves for each class of Macenko normalization with their corresponding AUC scores are shown in Figure 4.14. Here, AUC-ROC scores vary between 0.995 and 0.999 whereas AUC-PR values range from 0.981 to 0.998, as indicated in Figure 4.14. The relationship between accuracy curves is shown on the left portion of Figure 4.15, whereas their relative loss curves are depicted on the right portion of Figure 4.15. Interestingly, the validation loss improved as compared to the original dataset; however, no considerable changes occurred in validation accuracy. These statistics pointed out that the Macenko-based normalization has slightly outper-

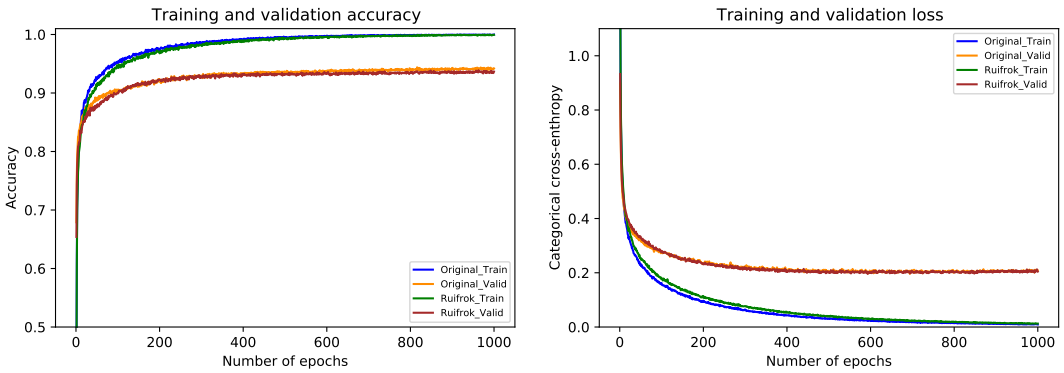


Figure 4.12 The left side demonstrates a comparison of training and validation accuracy curves of the original dataset and Ruifrok normalization. Whereas the right side illustrates a comparison of training and validation loss curves of the original dataset and Ruifrok normalization.

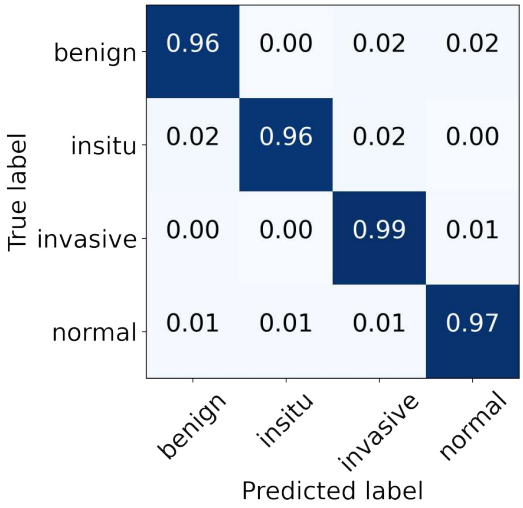


Figure 4.13 The final normalized confusion matrix of Macenko dataset.

formed the Reinhard and Ruifrok approaches in terms of accuracy. Also, it offered the same potential as the original dataset in terms of sensitivity for the in situ and invasive carcinomas.

4.3.5 Results of Vahadane normalization

Lastly, the performance metrics of our suggested model for Vahadane normalization are given in Table 4.9. During the cross-validation, we found a maximum accuracy of 97.77% during fold 4 and a minimum accuracy of 95.77% during fold 3, with a mean accuracy of 96.57% (± 0.75). The accuracy of the finalized model is noted as

Table 4.8 Evaluation metrics of our proposed model using Macenko normalization.

Folds	Confusion Matrices					Performance Evaluation					
	Predict → Actual ↓	Ben.	Ins.	Inv.	Nor.	Prec.	Rec.	F1	Test	Accuracy	Kappa
Fold 1	Benign	42	3	2	3	0.98	0.84	0.90	50	96.88%	0.951
	In situ	0	49	1	0	0.91	0.98	0.94	50		
	Invasive	1	1	227	1	0.98	0.99	0.98	230		
	Normal	0	1	1	118	0.97	0.98	0.98	120		
Fold 2	Benign	46	2	1	1	0.96	0.92	0.94	50	96.44%	0.945
	In situ	1	48	0	1	0.91	0.96	0.93	50		
	Invasive	0	2	223	5	0.99	0.97	0.98	230		
	Normal	1	1	1	117	0.94	0.97	0.96	120		
Fold 3	Benign	48	0	0	2	0.96	0.96	0.96	50	96.00%	0.937
	In situ	1	48	1	0	0.98	0.96	0.97	50		
	Invasive	0	0	227	3	0.96	0.99	0.97	230		
	Normal	1	1	9	109	0.96	0.91	0.93	120		
Fold 4	Benign	47	1	0	2	0.96	0.94	0.95	50	97.11%	0.955
	In situ	1	48	1	0	0.96	0.96	0.96	50		
	Invasive	0	0	227	3	0.98	0.99	0.98	230		
	Normal	1	1	3	115	0.96	0.96	0.96	120		
Fold 5	Benign	48	1	1	0	0.84	0.96	0.90	50	96.22%	0.941
	In situ	3	46	1	0	0.96	0.92	0.94	50		
	Invasive	1	0	226	3	0.99	0.98	0.98	230		
	Normal	5	1	1	113	0.97	0.94	0.96	120		
Final	Benign	48	0	1	1	0.94	0.96	0.95	50	**97.78%**	**0.965**
	In situ	1	48	1	0	0.99	0.96	0.97	50		
	Invasive	1	0	227	2	0.99	0.99	0.99	230		
	Normal	1	1	1	117	0.97	0.97	0.97	120		

97.33%, as indicated in Table 4.9. Specifically, the sensitivity for in situ carcinoma is 94% which is 2.00% lower than the original dataset. Likewise, the sensitivity for invasive carcinoma is 98% which is 1.00% percent lower than the original dataset. These concluded results of all the four classes are illustrated in Figure 4.16. Also, the ROC curves and PR curves for every class of Vahadane normalization along with their AUC values are shown in Figure 4.17. In this scenario, the AUC-ROC values vary 0.997 and 0.999 whereas the AUC-PR scores range from 0.986 to 0.996, as mentioned in Figure

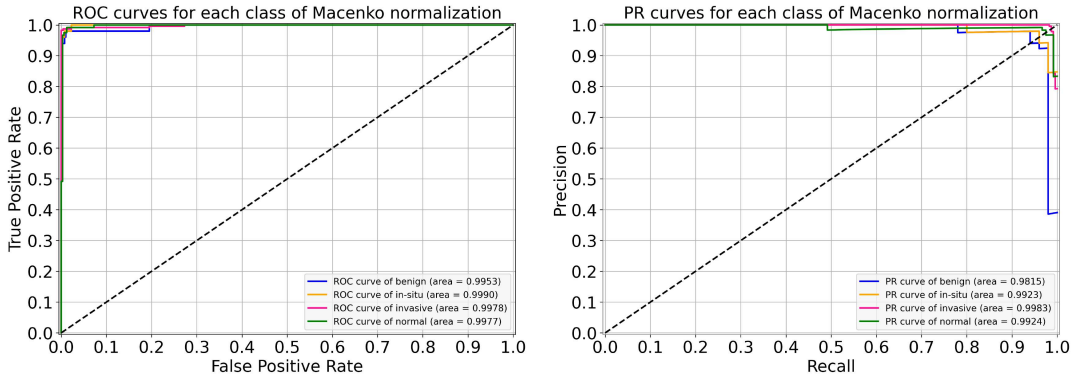


Figure 4.14 For Macenko normalization, the left block illustrates ROC curves for each class with an average AUC-ROC of 0.997. Whereas the right block depicts its PR curves for the individual class with a mean AUC-PR of 0.991.

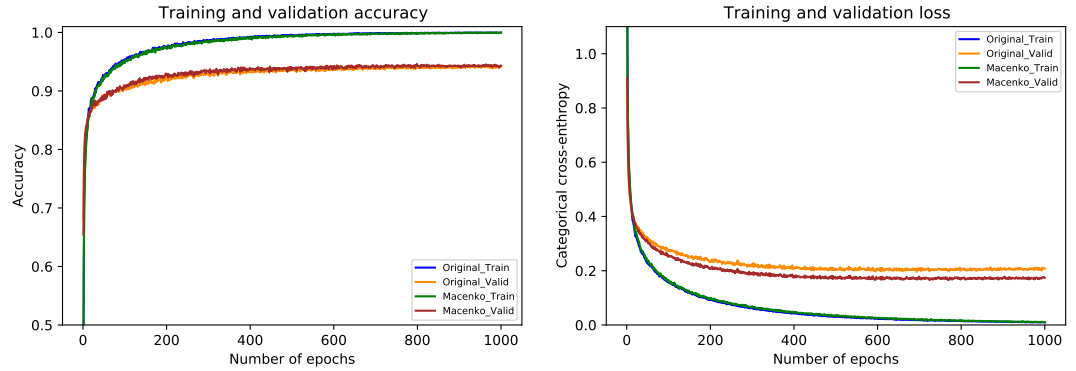


Figure 4.15 The left graph represents a comparison of training and validation accuracy curves of the original dataset and Macenko normalization. Whereas the right graph portrays a comparison of training and validation loss curves of the original dataset and Macenko normalization.

4.17. The correlation between accuracy curves is shown on the left side of Figure 4.18, whereas their corresponding loss curves are displayed on the right side of Figure 4.18. Similar to the Macenko normalization, a slight improvement in validation loss can be seen; however, no such improvement occurred in validation accuracy. These statistical analysis show that Vahadane normalization has the same performance as Reinhard and Ruifrok normalization, but is slightly lower than the original and Macenko normalization in terms of accuracy.

Finally, the sensitivity values of normal tissue, benign lesion, in situ carcinomas, and invasive carcinomas are collectively illustrated in Figure 4.19. Specifically, for in situ carcinomas, the sensitivity of original dataset is equivalent to Reinhard [10], Ruifrok [11], and Macenko [12]; however, it is 2% higher than the Vahadane [13]

Table 4.9 Evaluation metrics of our proposed model using Vahadane normalization.

Folds	Confusion Matrices					Performance Evaluation					
	Predict → Actual ↓	Ben.	Ins.	Inv.	Nor.	Prec.	Rec.	F1	Test	Accuracy	Kappa
Fold 1	Benign	42	4	2	2	1.00	0.84	0.91	50	96.66%	0.948
	In situ	0	49	1	0	0.91	0.98	0.94	50		
	Invasive	0	0	226	4	0.98	0.98	0.98	230		
	Normal	0	1	1	118	0.95	0.98	0.97	120		
Fold 2	Benign	45	2	0	3	0.98	0.90	0.94	50	96.44%	0.945
	In situ	1	48	0	1	0.91	0.96	0.93	50		
	Invasive	0	2	222	6	1.00	0.97	0.98	230		
	Normal	0	1	0	119	0.92	0.99	0.96	120		
Fold 3	Benign	46	1	1	2	0.98	0.92	0.95	50	95.77%	0.934
	In situ	1	48	1	0	0.96	0.96	0.96	50		
	Invasive	0	0	227	3	0.95	0.99	0.97	230		
	Normal	0	1	9	110	0.96	0.92	0.94	120		
Fold 4	Benign	47	1	0	2	0.98	0.94	0.96	50	97.77%	0.965
	In situ	1	48	1	0	0.96	0.96	0.96	50		
	Invasive	0	0	227	3	0.99	0.99	0.99	230		
	Normal	0	1	1	118	0.96	0.98	0.97	120		
Fold 5	Benign	48	1	1	0	0.89	0.96	0.92	50	96.22%	0.942
	In situ	2	48	0	0	0.94	0.96	0.95	50		
	Invasive	0	1	223	6	0.99	0.97	0.98	230		
	Normal	4	1	1	114	0.95	0.95	0.95	120		
Final	Benign	46	1	1	2	0.98	0.92	0.95	50	**97.33%**	**0.958**
	In situ	1	47	2	0	0.96	0.94	0.95	50		
	Invasive	0	0	226	4	0.99	0.98	0.98	230		
	Normal	0	1	0	119	0.95	0.99	0.97	120		

dataset and this small difference is equivalent to one sample in case of in situ carcinoma. Moreover, for invasive carcinoma, the proposed model offered higher sensitivity of 99% for original dataset, which is equivalent to Ruifrok [11] and Macenko [12] but 1% lower than Reinhard [10] and Vahadane [13]. In summary, our proposed model achieved generalized performance for the original as well as normalized datasets.

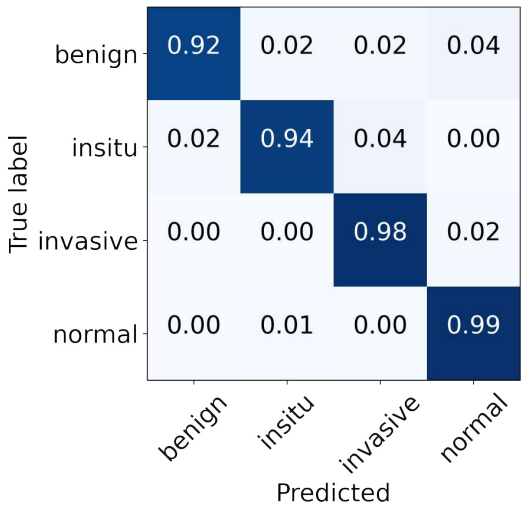


Figure 4.16 The final normalized confusion matrix of Vahadane dataset.

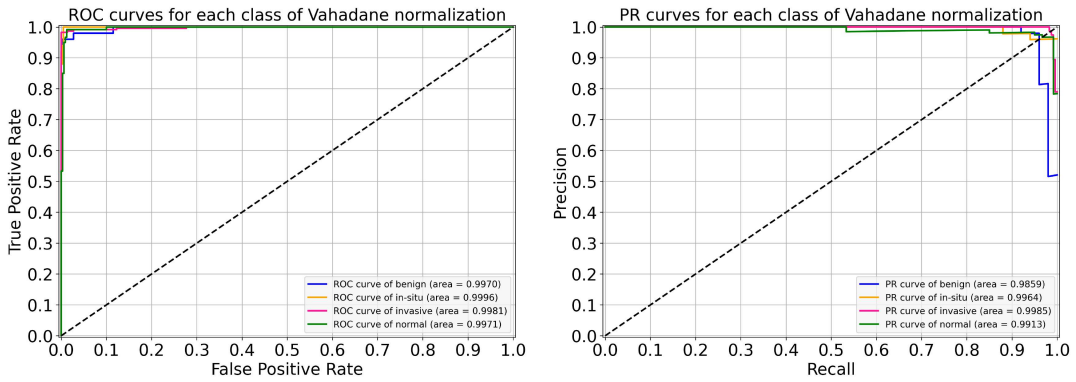


Figure 4.17 For Vahadane normalization, the left side shows ROC curves for each class with an average AUC-ROC of 0.998. Whereas the right side portrays its PR curves for the individual class with a mean AUC-PR of 0.993.

4.4 Discussion

The effectiveness of our proposed approach based on multilevel features can be compared with the baseline model (AlexNet [14]) and state-of-the-art deep learning architectures including VGG16 [9], VGG19 [9], Inception-v3 [15], and Xception [16] networks as feature extractors with their default settings. To that end, we leveraged the same optimal hyperparameters that we selected in our optimized framework, as discussed in the subsection 4.2.8. Furthermore, we used the same input image size as our proposed model to effectively compare the results, unlike Hao et al. [286], where the authors selected input image dimensions based on an individual pre-trained CNN model. We trained all of the aforementioned models on 80% of the images, whereas

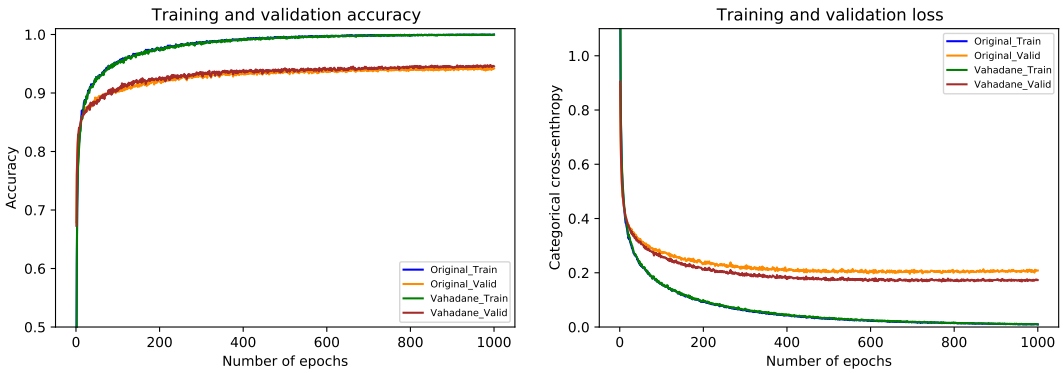


Figure 4.18 The left side shows a comparison of training and validation accuracy curves of the original dataset and Vahadane normalization. Whereas the right side depicts a comparison of training and validation loss curves of the original dataset and Vahadane normalization.

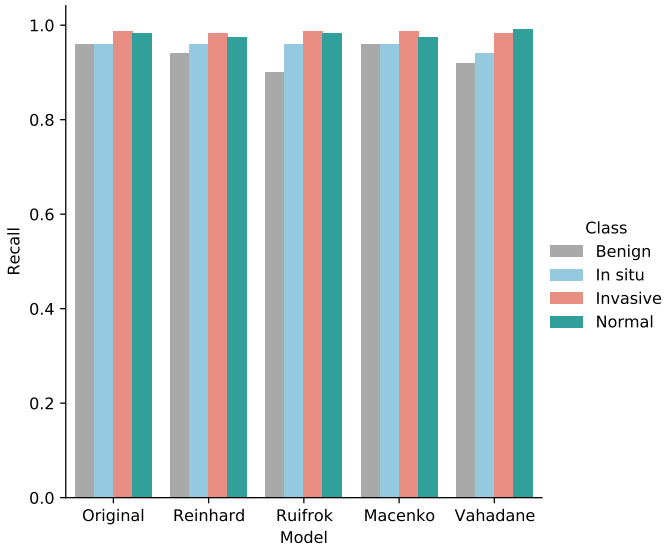


Figure 4.19 The sensitivity (recall) values of normal, benign, in situ carcinoma, and invasive carcinoma for the original, Reinhard [10], Ruifrok [11], Macenko [12], and Vahadane [13] datasets.

the remaining 20% of the images were used for the test purpose, as explained in the subsection 4.2.3. Of note, we chose AlexNet [14] as a baseline model because it was the first deep CNN model to achieve promising accuracy on the ILSVRC in 2012. Similarly, we considered VGG16 [9] and VGG19 [9] because our recently published study [43] employed these models to perform binary classification on a dataset that was generated from the same WSI images as used in the current study, as explained in subsection 4.2.1. Furthermore, the reason for selecting the Inception-v3 [15] lies in the

Table 4.10 Comparison of the proposed model based on multilevel features of Xception network with default versions of AlexNet [14] (baseline), VGG16 [9], VGG19 [9], Inception-v3 [15], and Xception [16] models as feature extractors.

Model	Dataset	Accuracy	F1-score	Kappa	Parameters
AlexNet [14]	Original	82.44%	77.25%	0.720	40.72 million
	Reinhard	76.66%	69.75%	0.633	
	Ruifrok	81.55%	75.75%	0.708	
	Macenko	81.33%	75.75%	0.702	
	Vahadane	78.89%	72.25%	0.667	
VGG16 [9]	Original	90.44%	86.50%	0.852	35.95 million
	Reinhard	88.00%	82.50%	0.814	
	Ruifrok	87.11%	82.50%	0.800	
	Macenko	89.55%	86.00%	0.839	
	Vahadane	89.55%	86.25%	0.838	
VGG19 [9]	Original	87.33%	81.75%	0.805	41.26 million
	Reinhard	88.89%	82.75%	0.824	
	Ruifrok	88.00%	81.75%	0.814	
	Macenko	89.11%	84.25%	0.832	
	Vahadane	89.11%	83.00%	0.832	
Inception-v3 [15]	Original	94.66%	91.25%	0.917	23.08 million
	Reinhard	94.44%	91.25%	0.914	
	Ruifrok	94.44%	91.50%	0.914	
	Macenko	93.55%	90.25%	0.900	
	Vahadane	93.77%	90.00%	0.904	
Xception [16]	Original	96.44%	95.00%	0.945	22.12 million
	Reinhard	96.66%	94.75%	0.948	
	Ruifrok	96.66%	94.75%	0.948	
	Macenko	96.00%	94.25%	0.938	
	Vahadane	95.56%	93.75%	0.931	
Proposed	Original	98.00%	97.50%	0.969	20.71 million
	Reinhard	97.33%	96.25%	0.959	
	Ruifrok	97.33%	96.25%	0.958	
	Macenko	97.78%	97.00%	0.965	
	Vahadane	97.33%	96.25%	0.958	

simplicity and robustness of its architecture, as discussed in the subsection 4.4. Finally, the motive behind choosing the plain Xception [16] is that it that it could be crucial to

evaluate its classification performance along with its modified architecture. Overall, the evaluation metrics of all the models are summarized in Table 4.10. Further details of these results can be found in the Supplementary Information (SI) file. The detailed comparison of our proposed architecture with each of the aforementioned models is as follows:

The performance metrics of the default AlexNet [14] model (baseline) as a feature extractor are given in Table 4.10 (further details can be found in Supplementary Table A.1). For the original dataset, it offered an accuracy of 82.44%, F1-score of 77.25%, and Cohen's kappa score of 0.720. Among the four normalized datasets, it yielded the highest accuracy of 81.55%, F1-measure of 75.75%, and Cohen's kappa of 0.708 for Ruifrok normalization. However, the baseline model shows overfitting as portrayed in the loss curves of Supplementary Figure A.1. Furthermore, it is a computationally expensive model with 40.7 million of training parameters, as mentioned in Table 4.10. In contrast, our proposed approach leveraged 20.01 million fewer parameters and achieved 15.56 percentage points higher accuracy along with a 24.9 percentage points gain in Cohen's kappa value for the original dataset.

Similarly, the performance measurements of the default VGG16 [9] model as a feature extractor are also compiled in Table 4.10 (more details are available in Supplementary Table A.2). For the original dataset, it gained an accuracy of 90.44%, F1-score of 86.50%, and Cohen's kappa statistic of 0.852. It acquired the highest accuracy of 89.55%, F1-measure of 86.25%, and Cohen's kappa of 0.838 for Vahadane normalization among the four normalized datasets. It can be noticed that VGG16 [9] outperformed the baseline model. Nevertheless, it shows overfitting as illustrated in the loss curves of Supplementary Figure A.2. Moreover, like the baseline AlexNet [14], it is a computationally expensive model with a total number of 35.95 million training parameters, as stated in Table 4.10. Conversely, our proposed model utilized 15.24 million lower parameters and achieved 7.56 percentage points higher accuracy along with 11.7 percentage points increase in Cohen's kappa score for the original dataset.

Likewise, the performance metrics of the default VGG19 [9] model as a feature extractor are provided in Table 4.10 (additional details are given in Supplementary Table A.3). It attained an accuracy of 87.33%, F1-measure of 81.75%, and Cohen's kappa value of 0.805 For the original dataset among the normalized datasets, it reached a maximum accuracy of 89.11%, F1-score of 84.25%, and Cohen's kappa of 0.832 for Macenko normalization. It can be observed that VGG19 [9] also outperformed the baseline model similar to the VGG16 [9] model. Nonetheless, it exhibits overfitting as portrayed in the loss curves of Supplementary Figure A.3. Furthermore, like the

baseline AlexNet [14] and VGG16 [9], it is a computationally expensive model with a total number of 41.26 million training parameters, as stated in Table 4.10. Contrary to VGG19 [9], our proposed framework utilized 20.55 million fewer parameters and achieved 10.67 percentage points higher accuracy together with 16.4 percentage points increase in Cohen's kappa score for the original dataset.

Moreover, the performance measurements of the default Inception-v3 [15] model as a feature extractor are also outlined in Table 4.10 (further details are provided in Supplementary Table A.4). For the original dataset, it attained an accuracy of 94.66%, F1-measure of 91.25%, and Cohen's kappa score of 0.917. Among the normalized datasets, it gained a top accuracy of 94.44%, F1-score of 91.50%, and Cohen's kappa of 0.914 for Ruifrok normalization. Interestingly, the default Inception-v3 [15] using 23.03 million training parameters offered promising results compared to the baseline AlexNet [14], and state-of-the-art VGG16 [9] and VGG19 [9] models. However, it shows overfitting as illustrated in the loss curves of Supplementary Figure A.4. In contrast, our proposed strategy leveraged 2.37 million lower training parameters and yielded 3.34 percentage points higher accuracy in conjunction with 5.5 percentage points increase in Cohen's kappa value for the original dataset.

Lastly, the performance metrics of the default Xception [16] model as a feature extractor are presented in Table 4.10 (more details can be found in Supplementary Table A.5). For the original dataset, it obtained an accuracy of 96.44%, F1-measure of 95.00%, and Cohen's kappa statistic of 0.945. Among the normalized datasets, it attained the highest accuracy of 96.66%, F1-score of 94.75%, and Cohen's kappa of 0.948 for both the Reinhard and Ruifrok normalization. It employed 22.21 million of training parameters and outperformed the baseline AlexNetNet [14] and state-of-the-art VGG16 [9], VGG19 [9], and Inception-v3 [15] models. These results demonstrate that the default Xception model as a feature extractor also offered promising results due to its robust performance in classifying histopathology images [273]. However, the default Xception model started overfitting which can be noticed in the loss curves of Supplementary Figure A.5. This can be due to using merely one GAP layer in its default framework. In comparison, our proposed approach used 1.41 million fewer parameters and yielded 1.56 percentage points high accuracy together with a 2.4 percentage points improvement in Cohen's kappa score for the original dataset.

In summary, these results demonstrate that the baseline AlexNet [14], as well as the state-of-the-art VGG16 [9] and VGG19 [9], are computationally expensive models. Furthermore, Inception-v3 [15] and Xception [16] networks offered promising performance but suffered from the overfitting problem. In contrast, our proposed model

based on multilevel features of the Xception [16] network outperformed all the default state-of-the-art frameworks with a fewer number of training parameters. Also, our proposed model offered resistance to overfitting due to the usage of multiple GAP layers [273]. Thus, it can be concluded that when used as a feature extractor, it is better to first check the Xception model with its default setting and then use multiple GAP layers to decrease the overfitting problem [273]. Overall, our proposed model using multilevel features from the intermediate layers of the Xception [16] network outperformed the baseline as well as state-of-the-art models with their default settings in classifying the breast cancer histopathology images. Interestingly, it provided minimal variance among the results on original and normalized datasets, and thus acts as a generalized deep learning model.

4.5 Chapter summary

The purpose of this study is to leverage deep learning to classify the hematoxylin-eosin-stained breast cancer microscopy images of our collected dataset into normal tissue, benign lesion, in situ carcinoma, and invasive carcinoma. To achieve this, we utilized six intermediate layers of the pre-trained Xception model to extract salient features from input images. We first optimized the proposed architecture on the unnormalized dataset, and then evaluated its performance on normalized datasets resulting from Reinhard, Ruifrok, Macenko, and Vahadane stain normalization procedures. Overall it is concluded that the proposed approach provides a generalized state-of-the-art classification performance towards the original and normalized datasets. Also, it can be deduced that even though the aforementioned stain normalization methods offered competitive results, they did not outperform the results of the original dataset. In future, we recommend to use the stain normalization techniques based on generative adversarial networks. Similarly, we suggest exploiting other recently developed pre-trained models by adopting feature extraction and fine-tuning strategies. Furthermore, it would be interesting take to the advantage of semi-supervised, unsupervised and self-supervised learning. Lastly, the concepts introduced in this study can be applied to histopathology image classification of different cancers, such as colorectal and lung cancers.

Chapter 5

Conclusion

This chapter outlines the most pertinent conclusions obtained after developing the current thesis. Numerous findings are presented following the completion of four main objectives defined in chapter 1. This thesis leveraged novel DL frameworks to effectively and efficiently diagnose breast cancer using histopathology images. The first case study focuses on the binary classification of breast cancer. It was achieved by developing a novel framework based on an ensemble of deep CNN models to classify breast malignancy into non-carcinoma and carcinoma. Whereas, the second case study focuses on the multiclass classification of breast cancer. It was accomplished by developing a new framework that leveraged multilevel features of deep CNN models to classify breast malignancy into normal, benign, in situ carcinoma, and invasive carcinoma. It is worth noting that both the aforementioned case studies used our collected dataset containing microscopy images retrieved from whole slide images of eighty breast cancer female patients. Thus, using DL with histopathology imaging, the objectives which were set in chapter 1 of the thesis were successfully met.

The hypothesis stated earlier in chapter 1 is given below.

- *Optimized deep learning frameworks can effectively and efficiently diagnose breast cancer as* ***non-carcinoma and carcinoma*** *as well as* ***normal, benign, in situ carcinoma, and invasive carcinoma*** *using histopathology images.*

From the results presented in chapters 3 and 4, it can be stated that DL has the potential to assist clinicians during the diagnostic process, which in turn can improve the quality of diagnosis in various crucial health issues including breast cancer.

5.1 Achievements

This thesis work provided several contributions by designing, optimizing, and validating end-to-end DL frameworks for breast cancer diagnosis. This was achieved by the collection of relevant data, preprocessing of data, designing CNN-based architectures, obtaining reliable findings through model optimization, and validating the proposed systems using crucial validation metrics. The key contributions in the current thesis were accomplished by achieving specific objectives presented in chapter 1. All the main objectives of this thesis were fulfilled during the research process, as provided below.

- **Objective 1: To collect and annotate a dataset containing whole slide images retrieved from eighty female patients suffering from breast cancer:** This objective was successfully completed by collaborating with the pathology department of Colsanitas clinic with a dependence of the Sanitas University, Bogotá, Colombia. This resulted in collecting 544 whole slide images stained with hematoxylin and eosin retrieved from eighty female patients suffering from breast cancer. The digitized slides were then annotated by two experienced pathologists from our consortium to create two datasets: one for binary classification and the other for multiclass classification.

- **Objective 2: To define the state-of-the-art developments in the supervised machine and deep learning for breast cancer diagnosis using widely followed medical imaging modalities:** This objective was successfully completed by compiling an article that analyzed cutting-edge supervised ML and DL models applied to medical imaging modalities for the detection, segmentation, and classification of breast lesions, as explained in chapter 2.

- **Objective 3: To design, optimize, and validate a supervised DL framework aimed at effective *binary classification* of breast cancer using our collected dataset:** This objective was successfully accomplished by proposing a novel end-to-end CNN-based framework to effectively and efficiently classify breast lesions into non-carcinoma and carcinoma, as described in chapter 3.

- **Objective 4: To design, optimize, and validate a supervised DL model intended at effective *multiclass classification* of breast cancer as normal, benign, in situ carcinoma, and invasive carcinoma using our collected dataset:** This objective was successfully accomplished by propounding an innovative

end-to-end CNN-based framework to effectively and efficiently classify breast lesions into normal, benign, in situ carcinoma, and invasive carcinoma, as described in chapter 4.

5.2 Scientific contribution

This thesis provided the most pertinent details about the computer-assisted diagnosis of breast cancer. Although the current document is structured as a monograph, two relevant articles have been published in peer-reviewed first-quartile journals with impact factors whereas the third one is under review. Moreover, while pursuing my Ph.D. at the eVida research group, I also worked on a European project related to natural language processing called Access Multilingual Information opinionS (AMIS). In this regard, I published one article in a peer-reviewed top-quartile journal with an impact factor and presented two papers at international conferences. However, this contribution is not documented because its context is not consistent with the main contributions of this thesis. All the aforementioned journal articles and conference papers are summarized in the upcoming subsections 5.2.1 and 5.2.2.

5.2.1 Journal articles

The first paper presented a systematic review that analyzed 142 state-of-the-art studies using supervised ML and DL approaches in computer-aided detection, segmentation, and classification of breast cancer. It is under review in a peer-reviewed top-quartile journal with an impact factor, as given in Table 5.1.

Table 5.1 Article I in a peer-reviewed journal.

Title	A systematic review of supervised machine and deep learning in breast cancer diagnosis using medical imaging		
Authors	**Z. Hameed**, F. Karem, and B. Garcia-Zapirain		
Journal	IEEE Access	**Quartile**	Q1
Date	Under review		
Impact Factor	3.90		
DOI	Under Review		

Similarly, the second paper proposed a novel end-to-end framework based on an ensemble of CNN models for the definite classification of breast cancer non-carcinoma and carcinoma using histopathology images. It has been published in a peer-reviewed first-quartile journal with an impact factor, as stated in Table 5.2.

Table 5.2 Article II in a peer-reviewed journal.

Title	Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network
Authors	**Z. Hameed**, B. Garcia-Zapirain, J. J. Aguirre, and M. A. Isaza-Ruget
Journal	Scientific Reports **Quartile** Q1
Date	September 16, 2022
Impact Factor	4.60
DOI	https://doi.org/10.1038/s41598-022-19278-2

Likewise, the third paper propounded a new end-to-end system by leveraging multilevel features of deep CNN models to effectively classify breast cancer into normal, benign, in situ carcinoma, and invasive carcinoma. It has been published in a peer-reviewed first-quartile journal with an impact factor, as provided in Table 5.3.

Table 5.3 Article III in a peer-reviewed journal.

Title	Breast cancer histopathology image classification using an ensemble of deep learning models
Authors	**Z. Hameed**, S. Zahia, B. Garcia-Zapirain, J. Javier Aguirre, and A. Maria Vanegas
Journal	Sensors **Quartile** Q1
Date	August 05, 2020
Impact Factor	3.90
DOI	https://doi.org/10.3390/s20164373

Finally, the fourth paper presented a comparably simpler yet effective end-to-end architecture based on the bidirectional long short-term memory network for sentiment classification. It has been published in a peer-reviewed top-quartile journal with an impact factor, as given in Table 5.4.

Table 5.4 Article IV in a peer-reviewed journal.

Title	Sentiment classification using a single-layered BiLSTM model
Authors	**Z. Hameed** and B. Garcia-Zapirain
Journal	IEEE Access **Quartile** Q1
Date	April 17, 2020
Impact Factor	3.90
DOI	https://doi.org/10.1109/ACCESS.2020.2988550

5.2.2 Communications in international conferences

The first conference paper followed an ensemble approach using the bidirectional gated recurrent unit for sentiment classification, as given in Table 5.5 and depicted in Figure 5.1.

Table 5.5 Publication I in an international conference.

Title	Sentiment analysis using an ensemble approach of BiGRU model: A case study of AMIS tweets
Authors	**Z. Hameed**, S. Shapoval, B. Garcia-Zapirain, and A.M. Zorilla
Conference	International Symposium on Signal Processing and Information Technology (ISSPIT)
Date	December 09-11, 2020
Publisher	IEEE **Location** Kentucky, USA
DOI	https://doi.org/10.1109/ISSPIT51521.2020.9408866

Similarly, the second conference presented a computationally efficient model by employing the bidirectional long short-term memory network for sentiment classification, as provided in Table 5.6 and illustrated in Figure 5.2.

Table 5.6 Publication II in an international conference.

Title	A computationally efficient BiLSTM based approach for the binary sentiment classification
Authors	**Z. Hameed**, B. Garcia-Zapirain, and I. O. Ruiz
Conference	IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)
Date	December 10-12, 2019
Publisher	IEEE **Location** Ajman, UAE
DOI	https://doi.org/10.1109/ISSPIT47144.2019.9001781

5.3 International mention

For the ***International Mention*** of my Ph.D. degree, I performed a research internship at the Université Laval in Canada from June 08 to September 8, 2022, as stated in Figure 5.3. I was registered as a full-time research trainee during the summer and fall semesters of 2022 where I worked under the supervision of Professor Simon Duchesne from the Department of Radiology and Nuclear Medicine, Université Laval, Canada.

During this period, I actively worked on a project that was funded by the Quebec Bio-Imaging Network entitled “Deep Learning in Histopathological Cerebrovascular

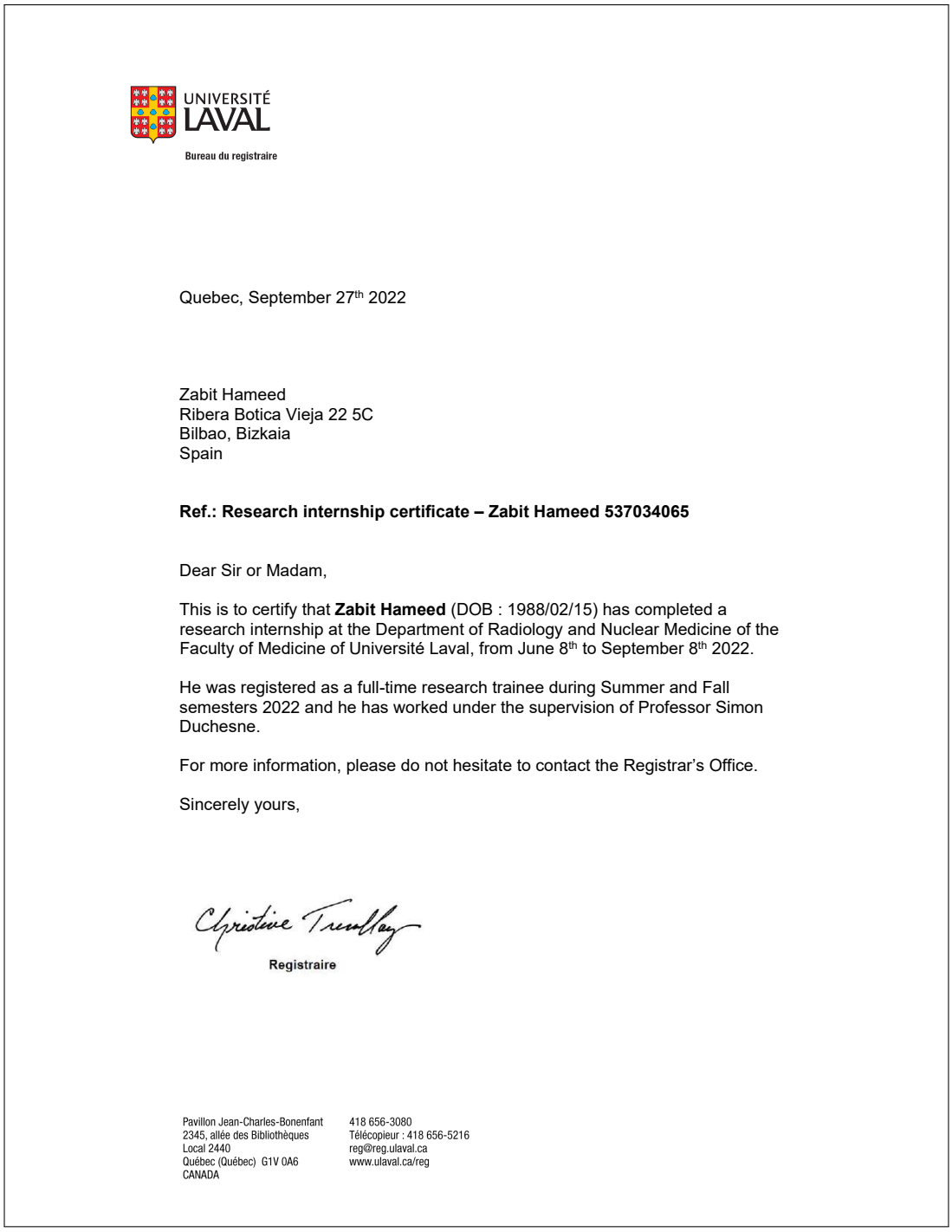
Figure 5.1 The certificate of my presentation at the IEEE ISSPIT 2020 conference.

Lesions Assessment". The aim of this project was to understand the data, analyze the data, and design a deep learning-based algorithm capable of automatically identifying cerebrovascular lesions (e.g. arteriolosclerosis) from whole slide images of human

Figure 5.2 The certificate of my presentation at the IEEE ISSPIT 2019 conference.

brains. It may detect cerebrovascular lesions during a patient's lifetime with great specificity and sensitivity. Consequently, it can allow clinicians to orient treatments accordingly and avoid, as much as possible, the decline to stroke and/or dementia with a vascular component.

To summarize, I analyzed and preprocessed two datasets containing whole slide images obtained during autopsy from healthy brains. During this phase, I participated in discussions with neurologists who collected these datasets. From one dataset, I analyzed, standardized, and preprocessed 504 slides containing arteriolosclerosis or non-arteriolosclerosis lesions from 63 patients. Similarly, from the other dataset, I analyzed, standardized, and preprocessed 442 slides containing arteriolosclerosis or non-arteriolosclerosis lesions from 134 patients. However, due to time constraints and ethical considerations, I could not design a deep learning framework needed to complete this project. My goal in this regard is to investigate the effectiveness of cutting-edge CNN networks in recognizing cerebrovascular lesions, which is analogous to recognizing breast lesions in this thesis.



UNIVERSITÉ LAVAL

Bureau du registraire

Quebec, September 27th 2022

Zabit Hameed
Ribera Botica Vieja 22 5C
Bilbao, Bizkaia
Spain

Ref.: Research internship certificate – Zabit Hameed 537034065

Dear Sir or Madam,

This is to certify that **Zabit Hameed** (DOB : 1988/02/15) has completed a research internship at the Department of Radiology and Nuclear Medicine of the Faculty of Medicine of Université Laval, from June 8th to September 8th 2022.

He was registered as a full-time research trainee during Summer and Fall semesters 2022 and he has worked under the supervision of Professor Simon Duchesne.

For more information, please do not hesitate to contact the Registrar's Office.

Sincerely yours,

Christine Tremblay

Registraire

Pavillon Jean-Charles-Bonenfant
2345, allée des Bibliothèques
Local 2440
Québec (Québec) G1V 0A6
CANADA

418 656-3080
Télécopieur : 418 656-5216
reg@reg.ulaval.ca
www.ulaval.ca/reg

Figure 5.3 The certificate of my international research stay.

5.4 Limitations and recommendations

This thesis proposed novel supervised deep learning frameworks based on deep CNN models to effectively and efficiently diagnose breast cancer using histopathology images. Although the current work offered promising results for the binary and multiclass classification of breast lesions, certain limitations should be considered when evaluating the results outlined within this thesis. These limitations together with future recommendations are presented in their respective chapters; nonetheless, a brief overview is presented here as well.

Firstly, the datasets prepared for both the case studies of this thesis are comparably smaller in contrast to those used in numerous state-of-the-art works. To that end, more images could be included in these datasets as DL models mostly rely on the quality of the input data for the learning process to produce generalized results. Secondly, this thesis leveraged only supervised DL approaches for breast cancer diagnosis using microscopy images. However, it would be interesting to exploit the potential of semi-supervised, unsupervised and self-supervised learning for breast cancer diagnosis using microscopy images as well as whole slide images. Thirdly, the concepts introduced in this thesis could also be applied to diagnose several different cancers, such as lung, colorectal, prostate, stomach, and liver cancers among others.

Apart from cancer, it would be interesting to investigate the potential of DL models for histopathological analysis of other applications like cognitive vascular disease (CVD). In this regard, I accomplished a full-time research internship at the Université Laval, Canada, where I analyzed and preprocessed whole slide images related to CVD of healthy individuals. In the future, our aim would be to impart the concepts introduced within this thesis to diagnose CVD such as arteriolosclerosis which contributes to the impairment in the cognition process of a healthy brain.

Appendix A

Table A.1 Evaluation metrics of the default AlexNet model (baseline) as a feature extractor using the original and normalized datasets.

Dataset	Confusion Matrices					Performance Evaluation					
	Predict → Actual ↓	Ben.	Ins.	Inv.	Nor.	Prec.	Rec.	F1	Test	Accuracy	Kappa
Original	Benign	35	10	3	2	0.73	0.70	0.71	50	82.44%	0.720
	In situ	9	39	2	0	0.70	0.78	0.74	50		
	Invasive	0	5	221	4	0.84	0.96	0.89	230		
	Normal	4	2	38	76	0.93	0.63	0.75	120		
Reinhard	Benign	33	10	3	4	0.52	0.66	0.58	50	76.66%	0.633
	In situ	8	39	2	1	0.64	0.78	0.70	50		
	Invasive	4	9	214	3	0.83	0.93	0.88	230		
	Normal	19	3	39	59	0.88	0.49	0.63	120		
Ruifrok	Benign	31	13	3	3	0.72	0.62	0.67	50	81.55%	0.708
	In situ	9	39	2	0	0.62	0.78	0.69	50		
	Invasive	2	7	215	6	0.85	0.93	0.89	230		
	Normal	1	4	33	82	0.90	0.68	0.78	120		
Macenko	Benign	30	13	4	3	0.75	0.60	0.67	50	81.33%	0.702
	In situ	7	40	3	0	0.63	0.80	0.71	50		
	Invasive	1	6	218	5	0.84	0.95	0.89	230		
	Normal	2	4	36	78	0.91	0.65	0.76	120		
Vahadane	Benign	27	18	2	3	0.77	0.54	0.64	50	78.89%	0.667
	In situ	6	42	2	0	0.53	0.84	0.65	50		
	Invasive	1	8	216	5	0.84	0.94	0.89	230		
	Normal	1	12	37	70	0.90	0.58	0.71	120		

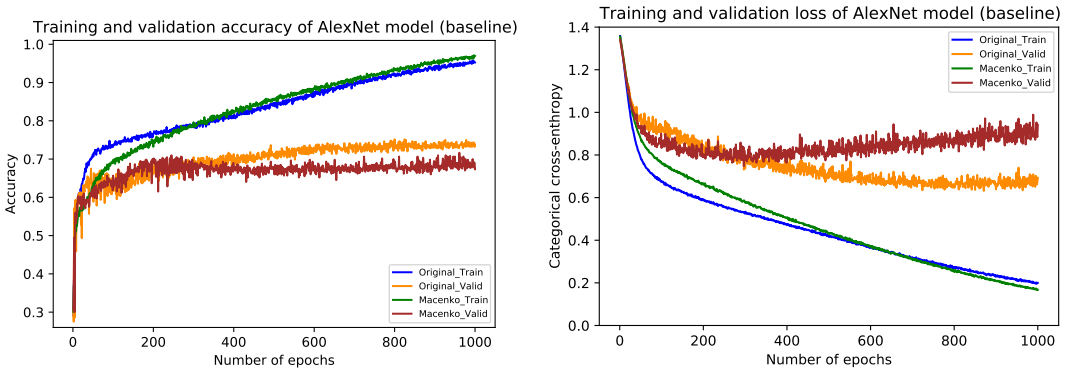


Figure A.1 The left-hand side shows training and validation accuracy curves of the original dataset and Macenko normalization on the default settings of the AlexNet model (baseline) as a feature extractor. Whereas the right-hand side depicts training and validation loss curves of the original dataset and Macenko normalization.

Table A.2 Evaluation metrics of the default VGG16 model as a feature extractor using the original and normalized datasets.

Dataset	Confusion Matrices					Performance Evaluation					
	Predict → Actual ↓	Ben.	Ins.	Inv.	Nor.	Prec.	Rec.	F1	Test	Accuracy	Kappa
Original	Benign	42	5	1	2	0.79	0.84	0.82	50	90.44%	0.852
	In situ	7	40	3	0	0.77	0.80	0.78	50		
	Invasive	1	4	217	8	0.96	0.94	0.95	230		
	Normal	3	3	6	108	0.92	0.90	0.91	120		
Reinhard	Benign	42	5	1	2	0.67	0.84	0.74	50	88.00%	0.814
	In situ	11	35	4	0	0.78	0.70	0.74	50		
	Invasive	2	3	216	9	0.95	0.94	0.94	230		
	Normal	8	2	7	103	0.90	0.86	0.88	120		
Ruifrok	Benign	38	4	3	5	0.67	0.76	0.71	50	87.11%	0.800
	In situ	8	39	2	1	0.81	0.78	0.80	50		
	Invasive	2	3	215	10	0.94	0.93	0.94	230		
	Normal	9	2	9	100	0.86	0.83	0.85	120		
Macenko	Benign	41	5	1	3	0.76	0.82	0.79	50	89.55%	0.839
	In situ	6	41	2	1	0.82	0.82	0.82	50		
	Invasive	2	3	213	12	0.96	0.93	0.94	230		
	Normal	5	1	6	108	0.87	0.90	0.89	120		
Vahadane	Benign	41	6	1	2	0.82	0.82	0.82	50	89.55%	0.838
	In situ	6	41	3	0	0.77	0.82	0.80	50		
	Invasive	1	3	215	11	0.94	0.93	0.94	230		
	Normal	2	3	9	106	0.89	0.88	0.89	120		

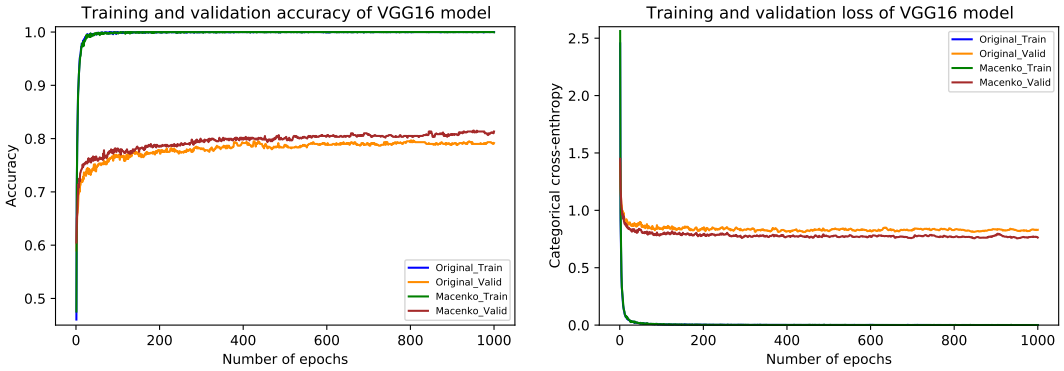


Figure A.2 The left-hand side shows training and validation accuracy curves of the original dataset and Macenko normalization on the default settings of the VGG16 model as a feature extractor. Whereas the right-hand side depicts training and validation loss curves of the original dataset and Macenko normalization.

Table A.3 Evaluation metrics of the default VGG19 model as a feature extractor using the original and normalized datasets.

Dataset	Confusion Matrices					Performance Evaluation					
	Predict → Actual ↓	Ben.	Ins.	Inv.	Nor.	Prec.	Rec.	F1	Test	Accuracy	Kappa
Original	Benign	45	3	0	2	0.66	0.90	0.76	50	87.33%	0.805
	In situ	15	31	3	1	0.82	0.62	0.70	50		
	Invasive	4	1	210	15	0.96	0.91	0.94	230		
	Normal	4	3	6	107	0.86	0.89	0.87	120		
Reinhard	Benign	46	3	1	0	0.64	0.92	0.75	50	88.89%	0.824
	In situ	16	30	2	2	0.86	0.60	0.71	50		
	Invasive	2	1	218	9	0.96	0.95	0.96	230		
	Normal	8	1	5	106	0.91	0.88	0.89	120		
Ruifrok	Benign	40	7	2	1	0.65	0.80	0.71	50	88.00%	0.814
	In situ	13	34	2	1	0.77	0.68	0.72	50		
	Invasive	4	0	217	9	0.95	0.94	0.95	230		
	Normal	5	3	7	105	0.91	0.88	0.89	120		
Macenko	Benign	44	4	1	1	0.70	0.88	0.78	50	89.11%	0.832
	In situ	12	34	3	1	0.83	0.68	0.75	50		
	Invasive	3	1	213	13	0.96	0.93	0.94	230		
	Normal	4	2	4	110	0.88	0.92	0.90	120		
Vahadane	Benign	43	5	1	1	0.67	0.86	0.75	50	89.11%	0.832
	In situ	13	33	2	2	0.82	0.66	0.73	50		
	Invasive	3	1	216	10	0.96	0.94	0.95	230		
	Normal	5	1	5	109	0.89	0.91	0.89	120		

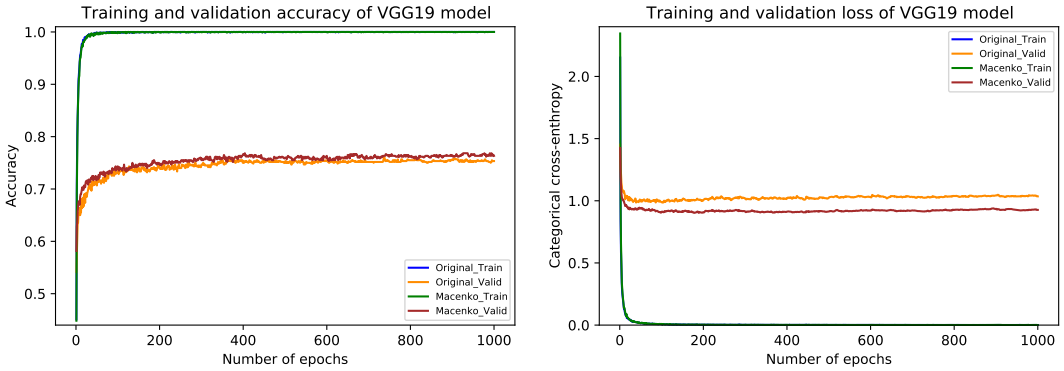


Figure A.3 The left-hand side shows training and validation accuracy curves of the original dataset and Macenko normalization on the default settings of the VGG19 model as a feature extractor. Whereas the right-hand side depicts training and validation loss curves of the original dataset and Macenko normalization.

Table A.4 Evaluation metrics of the default Inception-v3 model as a feature extractor using the original and normalized datasets.

Dataset	Confusion Matrices					Performance Evaluation					
	Predict → Actual ↓	Ben.	Ins.	Inv.	Nor.	Prec.	Rec.	F1	Test	Accuracy	Kappa
Original	Benign	45	2	0	3	0.82	0.90	0.86	50	94.66%	0.917
	In situ	7	41	2	0	0.91	0.82	0.86	50		
	Invasive	1	1	225	3	0.98	0.98	0.98	230		
	Normal	2	1	2	115	0.95	0.96	0.95	120		
Reinhard	Benign	44	3	0	3	0.83	0.88	0.85	50	94.44%	0.914
	In situ	5	43	0	2	0.90	0.86	0.88	50		
	Invasive	2	1	223	4	0.99	0.97	0.98	230		
	Normal	2	1	2	115	0.93	0.96	0.94	120		
Ruifrok	Benign	44	3	0	3	0.83	0.88	0.85	50	94.44%	0.914
	In situ	5	44	1	0	0.88	0.88	0.88	50		
	Invasive	4	2	218	6	1.00	0.95	0.97	230		
	Normal	0	1	0	119	0.93	0.99	0.96	120		
Macenko	Benign	42	4	0	4	0.84	0.84	0.84	50	93.55%	0.900
	In situ	5	43	1	1	0.88	0.86	0.87	50		
	Invasive	1	1	221	7	0.99	0.96	0.97	230		
	Normal	2	1	2	115	0.91	0.96	0.93	120		
Vahadane	Benign	41	5	1	3	0.82	0.82	0.82	50	93.77%	0.904
	In situ	6	43	0	1	0.83	0.86	0.84	50		
	Invasive	2	3	221	4	0.99	0.96	0.98	230		
	Normal	1	1	1	117	0.94	0.97	0.96	120		

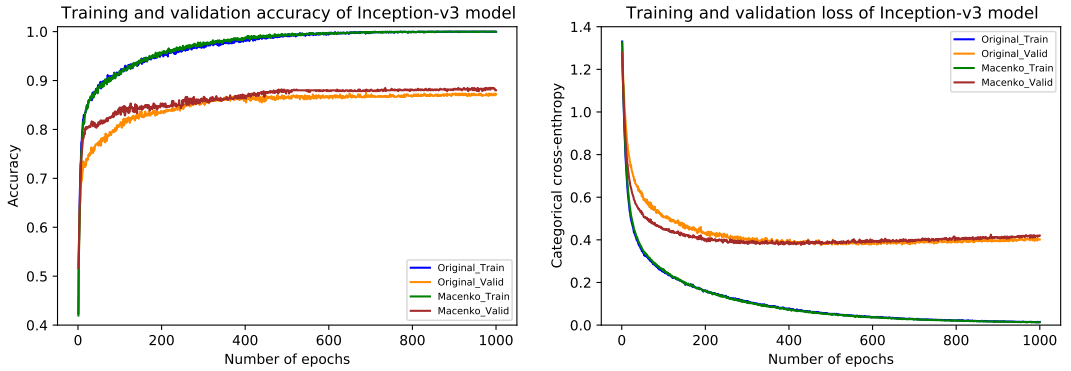


Figure A.4 The left side shows training and validation accuracy curves of the original dataset and Macenko normalization on the default settings of the Incepion-v3 model as a feature extractor. Whereas the right side depicts training and validation loss curves of the original dataset and Macenko normalization.

Table A.5 Evaluation metrics of the default Xception model as a feature extractor using the original and normalized datasets.

Dataset	Confusion Matrices					Performance Evaluation					
	Predict → Actual ↓	Ben.	Ins.	Inv.	Nor.	Prec.	Rec.	F1	Test	Accuracy	Kappa
Original	Benign	47	0	0	3	0.92	0.94	0.93	50	96.44%	0.945
	In situ	3	45	1	1	0.98	0.90	0.94	50		
	Invasive	0	0	225	5	0.99	0.98	0.98	230		
	Normal	1	1	1	117	0.93	0.97	0.95	120		
Reinhard	Benign	47	1	1	1	0.89	0.94	0.91	50	96.66%	0.948
	In situ	5	44	1	0	0.96	0.88	0.92	50		
	Invasive	0	0	227	3	0.99	0.99	0.99	230		
	Normal	1	1	1	117	0.97	0.97	0.97	120		
Ruifrok	Benign	46	2	0	2	0.90	0.92	0.91	50	96.66%	0.948
	In situ	3	47	0	0	0.92	0.94	0.93	50		
	Invasive	0	1	226	3	1.00	0.98	0.99	230		
	Normal	2	1	1	116	0.96	0.97	0.96	120		
Macenko	Benign	47	0	2	1	0.87	0.94	0.90	50	96.00%	0.938
	In situ	4	44	2	0	0.98	0.88	0.93	50		
	Invasive	1	0	225	4	0.98	0.98	0.98	230		
	Normal	2	1	1	116	0.96	0.97	0.96	120		
Vahadane	Benign	46	1	0	3	0.88	0.92	0.90	50	95.56%	0.931
	In situ	4	44	1	1	0.96	0.88	0.92	50		
	Invasive	1	0	225	4	0.98	0.98	0.98	230		
	Normal	1	1	3	115	0.93	0.96	0.95	120		

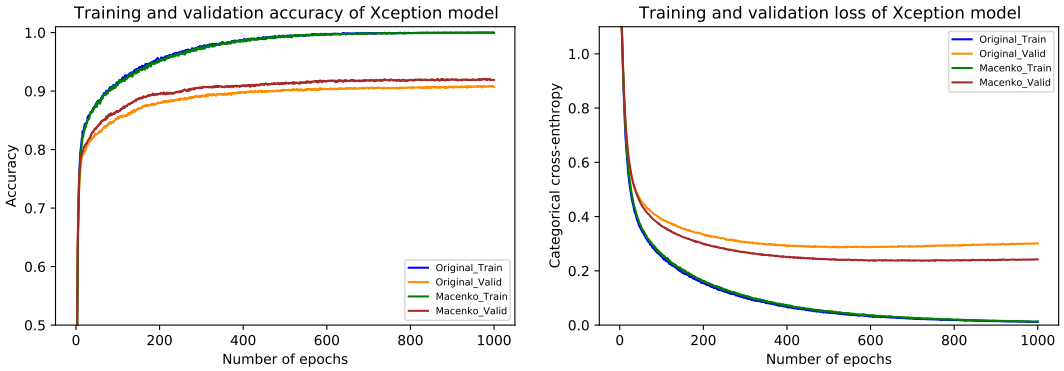


Figure A.5 The left side demonstrates training and validation accuracy curves of the original dataset and Macenko normalization on the default settings of the Xception model as a feature extractor. Whereas the right side presents training and validation loss curves of the original dataset and Macenko normalization.

References

[1] Y. Jiménez-Gaona, M. J. Rodríguez-Álvarez, and V. Lakshminarayanan, "Deep-learning-based computer-aided systems for breast cancer imaging: A critical review," *Applied Sciences*, vol. 10, no. 22, p. 8298, 2020.

[2] M. Tariq, S. Iqbal, H. Ayesha, I. Abbas, K. T. Ahmad, and M. F. K. Niazi, "Medical image based breast cancer diagnosis: State of the art and future directions," *Expert Systems with Applications*, vol. 167, p. 114095, 2021.

[3] J. P. Suckling, "The mammographic image analysis society digital mammogram database," *Digital Mammo*, pp. 375–386, 1994.

[4] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[5] M. Adachi, T. Fujioka, M. Mori, K. Kubota, Y. Kikuchi, W. Xiaotong, J. Oyama, K. Kimura, G. Oda, T. Nakagawa, *et al.*, "Detection and diagnosis of breast cancer using artificial intelligence based assessment of maximum intensity projection dynamic contrast-enhanced magnetic resonance images," *Diagnostics*, vol. 10, no. 5, p. 330, 2020.

[6] Z. Hameed, B. Garcia-Zapirain, J. J. Aguirre, and M. A. Isaza-Ruget, "Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network," *Scientific Reports*, vol. 12, no. 1, pp. 1–21, 2022.

[7] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *International Journal of Surgery*, vol. 88, p. 105906, 2021.

[8] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep learning for whole slide image analysis: an overview," *Frontiers in medicine*, p. 264, 2019.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 7-9, 2015*, 2015.

[10] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[11] A. C. Ruifrok, D. A. Johnston, *et al.*, "Quantification of histochemical staining by color deconvolution," *Analytical and quantitative cytology and histology*, vol. 23, no. 4, pp. 291–299, 2001.

[12] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE international symposium on biomedical imaging: from nano to macro*, pp. 1107–1110, IEEE, 2009.

[13] A. Vahadane, T. Peng, S. Albarqouni, M. Baust, K. Steiger, A. M. Schlitter, A. Sethi, I. Esposito, and N. Navab, "Structure-preserved color normalization for histological images," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 1012–1015, IEEE, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems (NIPS), Lake Tahoe, NV, USA, 3–6 December 2012*, pp. 1097–1105, 2012.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[16] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[17] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[18] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, *et al.*, "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *JAMA*, vol. 313, no. 11, pp. 1122–1132, 2015.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[21] M. Akram, M. Iqbal, M. Daniyal, and A. U. Khan, "Awareness and current knowledge of breast cancer," *Biological research*, vol. 50, no. 1, pp. 1–23, 2017.

[22] D. Vuong, P. T. Simpson, B. Green, M. C. Cummings, and S. R. Lakhani, "Molecular classification of breast cancer," *Virchows Archiv*, vol. 465, no. 1, pp. 1–14, 2014.

[23] C. Dromain, B. Boyer, R. Ferre, S. Canale, S. Delaloge, and C. Balleyguier, "Computed-aided diagnosis (CAD) in the detection of breast cancer," *European journal of radiology*, vol. 82, no. 3, pp. 417–423, 2013.

[24] D. A. Spak, J. Plaxco, L. Santiago, M. Dryden, and B. Dogan, "BI-RADS® fifth edition: A summary of changes," *Diagnostic and interventional imaging*, vol. 98, no. 3, pp. 179–190, 2017.

[25] Y. Gao, K. J. Geras, A. A. Lewin, and L. Moy, "New frontiers: an update on computer-aided diagnosis for breast imaging in the age of artificial intelligence," *American journal of roentgenology*, vol. 212, no. 2, p. 300, 2019.

[26] G. Chugh, S. Kumar, and N. Singh, "Survey on machine learning and deep learning applications in breast cancer diagnosis," *Cognitive Computation*, vol. 13, no. 6, pp. 1451–1470, 2021.

[27] S. Iranmakani, T. Mortezazadeh, F. Sajadian, M. F. Ghaziani, A. Ghafari, D. Khezerloo, and A. E. Musa, "A review of various modalities in breast imaging: technical aspects and clinical outcomes," *Egyptian Journal of Radiology and Nuclear Medicine*, vol. 51, no. 1, pp. 1–22, 2020.

[28] R. Farber, N. Houssami, S. Wortley, G. Jacklyn, M. L. Marinovich, K. McGeechan, A. Barratt, and K. Bell, "Impact of full-field digital mammography versus film-screen mammography in population screening: a meta-analysis," *Journal of the National Cancer Institute*, vol. 113, no. 1, pp. 16–26, 2021.

[29] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review," *Expert Systems with Applications*, vol. 167, p. 114161, 2021.

[30] L. Abdelrahman, M. Al Ghamdi, F. Collado-Mesa, and M. Abdel-Mottaleb, "Convolutional neural networks for breast cancer detection in mammography: A survey," *Computers in Biology and Medicine*, p. 104248, 2021.

[31] S. Beremauro and C. Girio-Fragkoulakis, "Imaging techniques in breast cancer," *Surgery (Oxford)*, 2022.

[32] R. Guo, G. Lu, B. Qin, and B. Fei, "Ultrasound imaging technologies for breast cancer detection and management: a review," *Ultrasound in medicine & biology*, vol. 44, no. 1, pp. 37–70, 2018.

[33] P.-H. Tsui, C.-K. Yeh, C.-C. Chang, and Y.-Y. Liao, "Classification of breast masses by ultrasonic nakagami imaging: a feasibility study," *Physics in Medicine & Biology*, vol. 53, no. 21, p. 6027, 2008.

[34] J. M. Chang, J.-K. Won, K.-B. Lee, I. A. Park, A. Yi, and W. K. Moon, "Comparison of shear-wave and strain ultrasound elastography in the differentiation of benign and malignant breast lesions," *American Journal of Roentgenology*, vol. 201, no. 2, pp. W347–W356, 2013.

[35] R. M. Mann, N. Cho, and L. Moy, "Breast MRI: state of the art," *Radiology*, vol. 292, no. 3, pp. 520–536, 2019.

[36] A. Berger, "How does it work?: Magnetic resonance imaging," *BMJ: British Medical Journal*, vol. 324, no. 7328, p. 35, 2002.

[37] V. P. Grover, J. M. Tognarelli, M. M. Crossey, I. J. Cox, S. D. Taylor-Robinson, and M. J. McPhail, "Magnetic resonance imaging: principles and techniques: lessons for clinicians," *Journal of clinical and experimental hepatology*, vol. 5, no. 3, pp. 246–255, 2015.

[38] R. M. Mann, C. K. Kuhl, and L. Moy, "Contrast-enhanced MRI for breast cancer screening," *Journal of Magnetic Resonance Imaging*, vol. 50, no. 2, pp. 377–390, 2019.

[39] L. Zhang, M. Tang, Z. Min, J. Lu, X. Lei, and X. Zhang, "Accuracy of combined dynamic contrast-enhanced magnetic resonance imaging and diffusion-weighted imaging for breast cancer detection: a meta-analysis," *Acta radiologica*, vol. 57, no. 6, pp. 651–660, 2016.

[40] H. Rahbar and S. C. Partridge, "Multiparametric MR imaging of breast cancer," *Magnetic Resonance Imaging Clinics*, vol. 24, no. 1, pp. 223–238, 2016.

[41] M. Zhang, J. V. Horvat, B. Bernard-Davila, M. A. Marino, D. Leithner, R. E. Ochoa-Albiztegui, T. H. Helbich, E. A. Morris, S. Thakur, and K. Pinker, "Multiparametric mri model with dynamic contrast-enhanced and diffusion-weighted imaging enables breast cancer diagnosis with high accuracy," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 3, pp. 864–874, 2019.

[42] M. Veta, J. P. Pluim, P. J. Van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: A review," *IEEE transactions on biomedical engineering*, vol. 61, no. 5, pp. 1400–1411, 2014.

[43] Z. Hameed, S. Zahia, B. Garcia-Zapirain, J. Javier Aguirre, and A. María Vanegas, "Breast cancer histopathology image classification using an ensemble of deep learning models," *Sensors*, vol. 20, no. 16, p. 4373, 2020.

[44] L. Tabár, P. B. Dean, F. L. Tucker, and A. Vörös, "Can we improve breast cancer management using an image-guided histopathology workup supported by larger histopathology sections?," *European Journal of Radiology*, p. 110750, 2023.

[45] G. Murtaza, L. Shuib, A. W. Abdul Wahab, G. Mujtaba, G. Mujtaba, H. F. Nweke, M. A. Al-garadi, F. Zulfiqar, G. Raza, and N. A. Azmi, "Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges," *Artificial Intelligence Review*, vol. 53, pp. 1655–1720, 2020.

[46] T. G. Debelee, F. Schwenker, A. Ibenthal, and D. Yohannes, "Survey of deep learning in breast cancer image analysis," *Evolving Systems*, vol. 11, pp. 143–163, 2020.

[47] S. M. Shah, R. A. Khan, S. Arif, and U. Sajid, "Artificial intelligence for breast cancer analysis: Trends & directions," *Computers in Biology and Medicine*, p. 105221, 2022.

[48] M. Heath, K. Bowyer, K. Daniel, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," in *Fifth International Workshop on Digital Mammography*, pp. 212–218, Medical Physics Publishing Publishing Corporation, 2001.

[49] M. Elter, R. Schulz-Wendtl and, and T. Wittenberg, "The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process," *Medical physics*, vol. 34, no. 11, pp. 4164–4172, 2007.

[50] M. G. Lopez, N. Posada, D. C. Moura, R. R. Pollán, J. M. F. Valiente, C. S. Ortega, M. Solar, G. Diaz-Herrero, I. Ramos, J. Loureiro, *et al.*, "BCDR: a breast cancer digital repository," in *15th International conference on experimental mechanics*, pp. 01–05, 2012.

[51] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: toward a full-field digital mammographic database," *Academic radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[52] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017.

[53] M. D. Halling-Brown, L. M. Warren, D. Ward, E. Lewis, A. Mackenzie, M. G. Wallis, L. S. Wilkinson, R. M. Given-Wilson, R. McAvinchey, and K. C. Young, "Optimam mammography image database: a large-scale resource of mammography images and clinical data," *Radiology: Artificial Intelligence*, vol. 3, no. 1, p. e200103, 2020.

[54] R. Khaled, M. Helal, O. Alfarghaly, O. Mokhtar, A. Elkorany, H. El Kassas, and A. Fahmy, "Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research," *Scientific Data*, vol. 9, no. 1, pp. 1–10, 2022.

[55] H. Piotrzkowska-Wróblewska, K. Dobruch-Sobczak, M. Byra, and A. Nowicki, "Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions," *Medical physics*, vol. 44, no. 11, pp. 6105–6109, 2017.

[56] M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A. K. Davison, and R. Marti, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1218–1226, 2017.

[57] Y. Zhang, M. Xian, H.-D. Cheng, B. Shareef, J. Ding, F. Xu, K. Huang, B. Zhang, C. Ning, and Y. Wang, "BUSIS: A benchmark for breast ultrasound image segmentation," *Healthcare*, vol. 10, no. 4, p. 729, 2022.

[58] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, *et al.*, "The cancer imaging archive (TCIA): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, pp. 1045–1057, 2013.

[59] E. S. Burnside, K. Drukker, H. Li, E. Bonaccio, M. Zuley, M. Ganott, J. M. Net, E. J. Sutton, K. R. Brandt, G. J. Whitman, *et al.*, "Using computer-extracted image phenotypes from tumors on breast magnetic resonance imaging to predict breast cancer pathologic stage," *Cancer*, vol. 122, no. 5, pp. 748–757, 2016.

[60] A. Saha, M. R. Harowicz, L. J. Grimm, C. E. Kim, S. V. Ghate, R. Walsh, and M. A. Mazurowski, "A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features," *British journal of cancer*, vol. 119, no. 4, pp. 508–516, 2018.

[61] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates," *Cancer letters*, vol. 77, no. 2-3, pp. 163–171, 1994.

[62] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[63] F. Dong, H. Irshad, E.-Y. Oh, M. F. Lerwill, E. F. Brachtel, N. C. Jones, N. W. Knoblauch, L. Montaser-Kouhsari, N. B. Johnson, L. K. Rao, *et al.*, "Computational pathology to discriminate benign from malignant intraductal proliferations of the breast," *PloS one*, vol. 9, no. 12, p. e114885, 2014.

[64] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PloS one*, vol. 12, no. 6, p. e0177544, 2017.

[65] 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, no. 7, pp. 1455–1462, 2015.

[66] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, *et al.*, "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 550–560, 2018.

[67] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, A. Madabhushi, and F. González, "High-throughput adaptive sampling for whole-slide histopathology image analysis (hashi) via convolutional neural networks: Application to invasive breast cancer detection," *PloS one*, vol. 13, no. 5, p. e0196828, 2018.

[68] R. Yan, F. Ren, Z. Wang, L. Wang, T. Zhang, Y. Liu, X. Rao, C. Zheng, and F. Zhang, "Breast cancer histopathological image classification using a hybrid deep neural network," *Methods*, vol. 173, pp. 52–60, 2020.

[69] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, *et al.*, "Bach: Grand challenge

on breast cancer histology images," *Medical image analysis*, vol. 56, pp. 122–139, 2019.

[70] N. Brancati, A. M. Anniciello, P. Pati, D. Riccio, G. Scognamiglio, G. Jaume, G. De Pietro, M. Di Bonito, A. Foncubierta, G. Botti, M. Gabrani, F. Feroce, and M. Frucci, "BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images," *Database*, vol. 2022, pp. 1–10, 2022.

[71] Roy Rosenzweig Center for History & New Media, "Zotero," *Retrieved from https://www.zotero.org/download*, 2016.

[72] A. L. Samuel, "Some studies in machine learning using the game of checkers. II—Recent progress," *IBM Journal of research and development*, vol. 11, no. 6, pp. 601–617, 1967.

[73] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, pp. 137–154, 2004.

[74] E. Michael, H. Ma, H. Li, F. Kulwa, and J. Li, "Breast cancer segmentation methods: current status and future potentials," *BioMed Research International*, vol. 2021, pp. 1–29, 2021.

[75] W. K. Pratt, *Introduction to digital image processing*. CRC press, 2013.

[76] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81–106, 1986.

[77] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[78] P. Hall, B. U. Park, and R. J. Samworth, "Choice of neighbor order in nearest-neighbor classification," *The Annals of Statistics*, vol. 36, pp. 2135–2152, 2008.

[79] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, pp. 131–163, 1997.

[80] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[81] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational Learning Theory: Second European Conference, Barcelona, Spain, March 13–15, 1995 Proceedings 2*, pp. 23–37, Springer, 1995.

[82] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[83] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[84] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[85] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[86] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[87] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[88] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[89] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[90] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[91] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[92] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[93] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, Ieee, 2016.

[94] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[95] 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*, pp. 770–778, 2016.

[96] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[97] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[98] R. Trevethan, "Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice," *Frontiers in public health*, vol. 5, p. 307, 2017.

[99] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, vol. 15, no. 1, pp. 1–8, 2022.

[100] Y. Hu, J. Li, and Z. Jiao, "Mammographic mass detection based on saliency with deep features," in *Proceedings of the International Conference on Internet Multimedia Computing and Service*, pp. 292–297, 2016.

[101] M. Taheri, G. Hamer, S. H. Son, and S. Y. Shin, "Enhanced breast cancer classification with automatic thresholding using svm and harris corner detection," in *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, pp. 56–60, 2016.

[102] T. Kooi, G. Litjens, B. Van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions," *Medical image analysis*, vol. 35, pp. 303–312, 2017.

[103] J.-J. Mordang, A. Gubern-Mérida, A. Bria, F. Tortorella, G. Den Heeten, and N. Karssemeijer, "Improving computer-aided detection assistance in breast cancer screening by removal of obviously false-positive findings," *Medical Physics*, vol. 44, no. 4, pp. 1390–1401, 2017.

[104] M. A. Al-Masni, M. A. Al-Antari, J.-M. Park, G. Gi, T.-Y. Kim, P. Rivera, E. Valarezo, M.-T. Choi, S.-M. Han, and T.-S. Kim, "Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based CAD system," *Computer methods and programs in biomedicine*, vol. 157, pp. 85–94, 2018.

[105] F. F. Ting, Y. J. Tan, and K. S. Sim, "Convolutional neural network improvement for breast cancer classification," *Expert Systems with Applications*, vol. 120, pp. 103–115, 2019.

[106] H. S. Oliveira, J. F. Teixeira, and H. P. Oliveira, "Lightweight deep learning pipeline for detection, segmentation and classification of breast cancer anomalies," in *International Conference on Image Analysis and Processing*, pp. 707–715, Springer, 2019.

[107] R. Agarwal, O. Díaz, M. H. Yap, X. Lladó, and R. Martí, "Deep learning for mass detection in full field digital mammograms," *Computers in biology and medicine*, vol. 121, p. 103774, 2020.

[108] M. A. Al-Antari, S.-M. Han, and T.-S. Kim, "Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital x-ray mammograms," *Computer methods and programs in biomedicine*, vol. 196, p. 105584, 2020.

[109] S. A. Hassan, M. S. Sayed, M. I. Abdalla, and M. A. Rashwan, "Breast cancer masses classification using deep convolutional neural networks and transfer learning," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 30735–30768, 2020.

[110] F. J. Pérez-Benito, F. Signol, J.-C. Perez-Cortes, A. Fuster-Baggetto, M. Pollan, B. Pérez-Gómez, D. Salas-Trejo, M. Casals, I. Martínez, and R. LLobet, "A deep learning system to obtain the optimal parameters for a threshold-based breast and dense tissue segmentation," *Computer Methods and Programs in Biomedicine*, vol. 195, p. 105668, 2020.

[111] L. Viegas, I. Domingues, and M. Mendes, "Study on data partition for delimitation of masses in mammography," *Journal of Imaging*, vol. 7, no. 9, p. 174, 2021.

[112] Y. Yan, P.-H. Conze, M. Lamard, G. Quellec, B. Cochener, and G. Coatrieux, "Towards improved breast mass detection using dual-view mammogram matching," *Medical Image Analysis*, vol. 71, p. 102083, 2021.

[113] G. H. Aly, M. Marey, S. A. El-Sayed, and M. F. Tolba, "Yolo based breast masses detection and classification in full-field digital mammograms," *Computer Methods and Programs in Biomedicine*, vol. 200, p. 105823, 2021.

[114] S. Kulkarni and R. Rabidas, "Squeezeu-net-based detection and diagnosis of microcalcification in mammograms," *Signal, Image and Video Processing*, vol. 17, no. 2, pp. 435–443, 2023.

[115] S. Ramesh, S. Sasikala, S. Gomathi, V. Geetha, and V. Anbumani, "Segmentation and classification of breast cancer using novel deep learning architecture," *Neural Computing and Applications*, vol. 34, no. 19, pp. 16533–16545, 2022.

[116] S. D. Deb and R. K. Jha, "Segmentation of mammogram images using deep learning for breast cancer detection," in *2nd International Conference on Image Processing and Robotics*, pp. 1–6, IEEE, 2022.

[117] R. Almajalid, J. Shan, Y. Du, and M. Zhang, "Development of a deep-learning-based method for breast ultrasound image segmentation," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1103–1108, IEEE, 2018.

[118] Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang, and P. L. Carson, "Medical breast ultrasound image segmentation by machine learning," *Ultrasonics*, vol. 91, pp. 1–9, 2019.

[119] Y. Zhou, H. Chen, Y. Li, S. Wang, L. Cheng, and J. Li, "3D multi-view tumor detection in automated whole breast ultrasound using deep convolutional neural network," *Expert Systems with Applications*, vol. 168, p. 114410, 2021.

[120] M. I. Daoud, A. Al-Ali, R. Alazrai, M. S. Al-Najar, B. A. Alsaify, M. Z. Ali, and S. Alouneh, "An edge-based selection method for improving regions-of-interest localizations obtained using multiple deep learning object-detection models in breast ultrasound images," *Sensors*, vol. 22, no. 18, p. 6721, 2022.

[121] E. Michael, H. Ma, and S. Qi, "Breast tumor segmentation in ultrasound images based on u-net model," in *Proceedings of the ICR'22 International Conference on Innovations in Computing Research*, pp. 22–31, Springer, 2022.

[122] A. S. Podda, R. Balia, S. Barra, S. Carta, G. Fenu, and L. Piano, "Fully-automated deep learning pipeline for segmentation and classification of breast ultrasound images," *Journal of Computational Science*, vol. 63, p. 101816, 2022.

[123] M. U. Dalmış, G. Litjens, K. Holland, A. Setio, R. Mann, N. Karssemeijer, and A. Gubern-Mérida, "Using deep learning to segment breast and fibroglandular tissue in mri volumes," *Medical physics*, vol. 44, no. 2, pp. 533–546, 2017.

[124] M. Benjelloun, M. El Adoui, M. A. Larhmam, and S. A. Mahmoudi, "Automated breast tumor segmentation in dce-mri using deep learning," in *IEEE 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)*, pp. 1–6, IEEE, 2018.

[125] Y. Zhang, J.-H. Chen, K.-T. Chang, V. Y. Park, M. J. Kim, S. Chan, P. Chang, D. Chow, A. Luk, T. Kwong, *et al.*, "Automatic breast and fibroglandular tissue segmentation in breast mri using deep learning by a fully-convolutional residual neural network u-net," *Academic radiology*, vol. 26, no. 11, pp. 1526–1535, 2019.

[126] M. El Adoui, S. A. Mahmoudi, M. A. Larhmam, and M. Benjelloun, "MRI breast tumor segmentation using different encoder and decoder CNN architectures," *Computers*, vol. 8, no. 3, p. 52, 2019.

[127] H. Jiao, X. Jiang, Z. Pang, X. Lin, Y. Huang, and L. Li, "Deep convolutional neural networks-based automatic breast segmentation and mass detection in dce-mri," *Computational and Mathematical Methods in Medicine*, vol. 2020, 2020.

[128] F. Ayatollahi, S. B. Shokouhi, R. M. Mann, and J. Teuwen, "Automatic breast lesion detection in ultrafast dce-mri using deep learning," *Medical Physics*, vol. 48, no. 10, pp. 5897–5907, 2021.

[129] A. Galli, S. Marrone, G. Piantadosi, M. Sansone, and C. Sansone, "A pipelined tracer-aware approach for lesion segmentation in breast dce-mri," *Journal of Imaging*, vol. 7, no. 12, p. 276, 2021.

[130] Y. Zhang, S. Chan, J.-H. Chen, K.-T. Chang, C.-Y. Lin, H.-B. Pan, W.-C. Lin, T. Kwong, R. Parajuli, R. S. Mehta, *et al.*, "Development of u-net breast density segmentation method for fat-sat mr images using transfer learning based on non-fat-sat model," *Journal of digital imaging*, vol. 34, no. 4, pp. 877–887, 2021.

[131] L. Huo, X. Hu, Q. Xiao, Y. Gu, X. Chu, and L. Jiang, "Segmentation of whole breast and fibroglandular tissue using nnu-net in dynamic contrast enhanced mr images," *Magnetic Resonance Imaging*, vol. 82, pp. 31–41, 2021.

[132] Y.-Y. Guo, Y.-H. Huang, Y. Wang, J. Huang, Q.-Q. Lai, and Y.-Z. Li, "Breast mri tumor automatic segmentation and triple-negative breast cancer discrimination algorithm based on deep learning," *Computational and Mathematical Methods in Medicine*, vol. 2022, 2022.

[133] H. Ni, H. Liu, K. Wang, X. Wang, X. Zhou, and Y. Qian, "WSI-Net: Branch-based and hierarchy-aware network for segmentation and classification of breast histopathological whole-slide images," in *International Workshop on Machine Learning in Medical Imaging*, pp. 36–44, Springer, 2019.

[134] S. M. Patil, L. Tong, and M. D. Wang, "Generating region of interests for invasive breast cancer in histopathological whole-slide-image," in *IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 723–728, IEEE, 2020.

[135] C. Li and X. Lu, "Computer-aided detection breast cancer in whole slide image," in *International Conference on Computer, Control and Robotics (ICCCR)*, pp. 193–198, IEEE, 2021.

[136] Y. Lu, J. Zhang, X. Liu, Z. Zhang, W. Li, X. Zhou, and R. Li, "Prediction of breast cancer metastasis by deep learning pathology," *IET Image Processing*, vol. 17, no. 2, pp. 533–543, 2023.

[137] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Computer methods and programs in biomedicine*, vol. 127, pp. 248–257, 2016.

[138] Z. Jiao, X. Gao, Y. Wang, and J. Li, "A deep feature based framework for breast masses classification," *Neurocomputing*, vol. 197, pp. 221–231, 2016.

[139] K. L. Kashyap, M. K. Bajpai, and P. Khanna, "Breast tissue density classification in mammograms based on supervised machine learning technique," in *Proceedings of the 10th Annual ACM India Compute Conference*, pp. 131–135, 2017.

[140] S. Guan and M. Loew, "Breast cancer detection using transfer learning in convolutional neural networks," in *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–8, IEEE, 2017.

[141] P. Sonar, U. Bhosle, and C. Chandrajit, "Mammography classification using modified hybrid svm-knn," in *2017 international conference on signal processing and communication (ICSPC)*, pp. 305–311, IEEE, 2017.

[142] P. U. Hepsağ, S. A. Özel, and A. Yazıcı, "Using deep learning for mammography classification," in *International Conference on Computer Science and Engineering (UBMK)*, pp. 418–423, IEEE, 2017.

[143] N. Antropova, B. Q. Huynh, and M. L. Giger, "A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets," *Medical physics*, vol. 44, no. 10, pp. 5162–5171, 2017.

[144] J.-Y. Yeh and S. Chan, "CNN-based CAD for breast cancer classification in digital breast tomosynthesis," in *Proceedings of the 2nd International Conference on Graphics and Signal Processing*, pp. 26–30, 2018.

[145] H. Chougrad, H. Zouaki, and O. Alheyane, "Deep convolutional neural networks for breast cancer screening," *Computer methods and programs in biomedicine*, vol. 157, pp. 19–30, 2018.

[146] F. Gao, T. Wu, J. Li, B. Zheng, L. Ruan, D. Shang, and B. Patel, "SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis," *Computerized Medical Imaging and Graphics*, vol. 70, pp. 53–62, 2018.

[147] Y. Brhane Hagos, A. Gubern Mérida, and J. Teuwen, "Improving breast cancer detection using symmetry information with deep learning," in *Image analysis for moving organ, breast, and thoracic images*, pp. 90–97, Springer, 2018.

[148] A. A. Mohamed, Y. Luo, H. Peng, R. C. Jankowitz, and S. Wu, "Understanding clinical mammographic breast density assessment: a deep learning perspective," *Journal of digital imaging*, vol. 31, no. 4, pp. 387–392, 2018.

[149] H. Li, S. Zhuang, D.-a. Li, J. Zhao, and Y. Ma, "Benign and malignant classification of mammogram images based on deep learning," *Biomedical Signal Processing and Control*, vol. 51, pp. 347–354, 2019.

[150] S. Laghmati, A. Tmiri, and B. Cherradi, "Machine learning based system for prediction of breast cancer severity," in *International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pp. 1–5, IEEE, 2019.

[151] L. Tsochatzidis, L. Costaridou, and I. Pratikakis, "Deep learning for breast cancer diagnosis from mammograms—a comparative study," *Journal of Imaging*, vol. 5, no. 3, p. 37, 2019.

[152] N. E. Benzebouchi, N. Azizi, and K. Ayadi, "A computer-aided diagnosis system for breast cancer using deep convolutional neural networks," in *Computational intelligence in data mining*, pp. 583–593, Springer, 2019.

[153] X. Li, G. Qin, Q. He, L. Sun, H. Zeng, Z. He, W. Chen, X. Zhen, and L. Zhou, "Digital breast tomosynthesis versus digital mammography: integration of image modalities enhances deep learning-based breast mass classification," *European radiology*, vol. 30, no. 2, pp. 778–788, 2020.

[154] S. Wessels and D. v. d. Haar, "Applying deep learning for the detection of abnormalities in mammograms," in *Information science and applications*, pp. 201–210, Springer, 2020.

[155] D. Saranyaraj, M. Manikandan, and S. Maheswari, "A deep convolutional neural network for the early detection of breast carcinoma with respect to hyperparameter tuning," *Multimedia Tools and Applications*, vol. 79, no. 15-16, pp. 11013–11038, 2020.

[156] D. A. Ragab, O. Attallah, M. Sharkas, J. Ren, and S. Marshall, "A framework for breast cancer classification using multi-dcnns," *Computers in Biology and Medicine*, vol. 131, p. 104245, 2021.

[157] M. Heidari, S. Lakshmivarahan, S. Mirniaharikandehei, G. Danala, S. K. R. Maryada, H. Liu, and B. Zheng, "Applying a random projection algorithm to optimize machine learning model for breast lesion classification," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 9, pp. 2764–2775, 2021.

[158] S. J. Malebary and A. Hashmi, "Automated breast mass classification system using deep learning and ensemble learning in digital mammogram," *IEEE Access*, vol. 9, pp. 55312–55328, 2021.

[159] R. S. Lee, J. A. Dunnmon, A. He, S. Tang, C. Re, and D. L. Rubin, "Comparison of segmentation-free and segmentation-dependent computer-aided diagnosis of breast masses on a public mammography dataset," *Journal of biomedical informatics*, vol. 113, p. 103656, 2021.

[160] L. Tsochatzidis, P. Koutla, L. Costaridou, and I. Pratikakis, "Integrating segmentation information into cnn for breast cancer diagnosis of mammographic masses," *Computer Methods and Programs in Biomedicine*, vol. 200, p. 105913, 2021.

[161] E. M. El Houby and N. I. Yassin, "Malignant and nonmalignant classification of breast lesions in mammograms using convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 70, p. 102954, 2021.

[162] J. Song, Y. Zheng, M. Zakir Ullah, J. Wang, Y. Jiang, C. Xu, Z. Zou, and G. Ding, "Multiview multimodal network for breast cancer diagnosis in contrast-enhanced spectral mammography images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, no. 6, pp. 979–988, 2021.

[163] F. A. Al-Fahaidy, B. Al-Fuhaidi, I. AL-Darouby, F. AL-Abady, M. AL-Qadry, and A. AL-Gamal, "A diagnostic model of breast cancer based on digital mammogram images using machine learning techniques," *Applied Computational Intelligence & Soft Computing*, 2022.

[164] N. A. Samee, A. A. Alhussan, V. F. Ghoneim, G. Atteia, R. Alkanhel, M. A. Al-Antari, and Y. M. Kadah, "A hybrid deep transfer learning of cnn-based lr-pca for breast lesion diagnosis via medical breast mammograms," *Sensors*, vol. 22, no. 13, p. 4938, 2022.

[165] K. Marathe, C. Marasinou, B. Li, N. Nakhaei, B. Li, J. G. Elmore, L. Shapiro, and W. Hsu, "Automated quantitative assessment of amorphous calcifications: Towards improved malignancy risk stratification," *Computers in biology and medicine*, vol. 146, p. 105504, 2022.

[166] H. Singh, V. Sharma, and D. Singh, "Comparative analysis of proficiencies of various textures and geometric features in breast mass classification using k-nearest neighbor," *Visual Computing for Industry, Biomedicine, and Art*, vol. 5, pp. 1–19, 2022.

[167] R. Karthiga, K. Narasimhan, and R. Amirtharajan, "Diagnosis of breast cancer for modern mammography using artificial intelligence," *Mathematics and Computers in Simulation*, vol. 202, pp. 316–330, 2022.

[168] A. A. Hekal, H. E.-D. Moustafa, and A. Elnakib, "Ensemble deep learning system for early breast cancer detection," *Evolutionary Intelligence*, pp. 1–10, 2022.

[169] M. M. Alshammari, A. Almuhanna, and J. Alhiyafi, "Mammography image-based diagnosis of breast cancer using machine learning: a pilot study," *Sensors*, vol. 22, no. 1, p. 203, 2021.

[170] J. Song, Y. Zheng, J. Wang, M. Z. Ullah, X. Li, Z. Zou, and G. Ding, "Multi-feature deep information bottleneck network for breast cancer classification in contrast enhanced spectral mammography," *Pattern Recognition*, vol. 131, p. 108858, 2022.

[171] J. Shan, S. K. Alam, B. Garra, Y. Zhang, and T. Ahmed, "Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods," *Ultrasound in medicine & biology*, vol. 42, no. 4, pp. 980–988, 2016.

[172] B. K. Singh, K. Verma, L. Panigrahi, and A. Thoke, "Integrating radiologist feedback with computer aided diagnostic systems for breast cancer risk prediction in ultrasonic images: An experimental investigation in machine learning paradigm," *Expert Systems with Applications*, vol. 90, pp. 209–223, 2017.

[173] L. R. Sultan, S. M. Schultz, T. W. Cary, and C. M. Sehgal, "Machine learning to improve breast cancer diagnosis by multimodal ultrasound," in *2018 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2018.

[174] M. Byra, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, and M. Andre, "Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion," *Medical physics*, vol. 46, no. 2, pp. 746–755, 2019.

[175] Y. Wang, E. J. Choi, Y. Choi, H. Zhang, G. Y. Jin, and S.-B. Ko, "Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning," *Ultrasound in medicine & biology*, vol. 46, no. 5, pp. 1119–1132, 2020.

[176] M. I. Daoud, S. Abdel-Rahman, T. M. Bdair, M. S. Al-Najar, F. H. Al-Hawari, and R. Alazrai, "Breast tumor classification in ultrasound images using combined deep and handcrafted features," *Sensors*, vol. 20, no. 23, p. 6838, 2020.

[177] A. K. Mishra, P. Roy, S. Bandyopadhyay, and S. K. Das, "Breast ultrasound tumour classification: A machine learning—radiomics based approach," *Expert Systems*, vol. 38, no. 7, p. e12713, 2021.

[178] Y. Eroğlu, M. Yildirim, and A. Çinar, "Convolutional neural networks based classification of breast ultrasonography images by hybrid method with respect to benign, malignant, and normal using mrmr," *Computers in biology and medicine*, vol. 133, p. 104407, 2021.

[179] W.-C. Shia and D.-R. Chen, "Classification of malignant tumors in breast ultrasound using a pretrained deep residual network model and support vector machine," *Computerized Medical Imaging and Graphics*, vol. 87, p. 101829, 2021.

[180] S. Misra, S. Jeon, R. Managuli, S. Lee, G. Kim, C. Yoon, S. Lee, R. G. Barr, and C. Kim, "Bi-modal transfer learning for classifying breast cancers via combined b-mode and ultrasound strain imaging," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 1, pp. 222–232, 2021.

[181] R. Hoffmann, C. Reich, and K. Skerl, "Evaluating different combination methods to analyse ultrasound and shear wave elastography images automatically through discriminative convolutional neural network in breast cancer imaging," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–7, 2022.

[182] A. K. Mishra, P. Roy, S. Bandyopadhyay, and S. K. Das, "Feature fusion based machine learning pipeline to improve breast cancer prediction," *Multimedia Tools and Applications*, vol. 81, no. 26, pp. 37627–37655, 2022.

[183] W. Liu, M. Guo, P. Liu, and Y. Du, "Mfdcmodel: A novel classification model for classification of benign and malignant breast tumors in ultrasound images," *Electronics*, vol. 11, no. 16, p. 2583, 2022.

[184] Q. Wang, H. Chen, G. Luo, B. Li, H. Shang, H. Shao, S. Sun, Z. Wang, K. Wang, and W. Cheng, "Performance of novel deep learning network with the incorporation of the automatic segmentation network for diagnosis of breast cancer in automated breast ultrasound," *European Radiology*, vol. 32, no. 10, pp. 7163–7172, 2022.

[185] M. Razavi, L. Wang, T. Tan, N. Karssemeijer, L. Linsen, U. Frese, H. K. Hahn, and G. Zachmann, "Novel morphological features for non-mass-like breast lesion classification on dce-mri," in *International Workshop on Machine Learning in Medical Imaging*, pp. 305–312, Springer, 2016.

[186] H. Zheng, Y. Gu, Y. Qin, X. Huang, J. Yang, and G.-Z. Yang, "Small lesion classification in dynamic contrast enhancement mri for breast cancer early detection," in *MICCAI: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pp. 876–884, Springer, 2018.

[187] L. Luo, H. Chen, X. Wang, Q. Dou, H. Lin, J. Zhou, G. Li, and P.-A. Heng, "Deep angular embedding and feature correlation attention for breast mri cancer analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 504–512, Springer, 2019.

[188] Y. Ji, H. Li, A. V. Edwards, J. Papaioannou, W. Ma, P. Liu, and M. L. Giger, "Independent validation of machine learning in diagnosing breast cancer on magnetic resonance imaging within a single institution," *Cancer Imaging*, vol. 19, no. 1, pp. 1–11, 2019.

[189] C. Haarburger, M. Baumgartner, D. Truhn, M. Broeckmann, H. Schneider, S. Schrading, C. Kuhl, and D. Merhof, "Multi scale curriculum CNN for context-aware breast mri malignancy classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 495–503, Springer, 2019.

[190] H. Feng, J. Cao, H. Wang, Y. Xie, D. Yang, J. Feng, and B. Chen, "A knowledge-driven feature learning and integration method for breast cancer diagnosis on multi-sequence mri," *Magnetic resonance imaging*, vol. 69, pp. 40–48, 2020.

[191] S. Thakran, R. K. Gupta, and A. Singh, "Characterization of breast tumors using machine learning based upon multiparametric magnetic resonance imaging features," *NMR in Biomedicine*, vol. 35, no. 5, p. e4665, 2022.

[192] T. Fujioka, Y. Yashima, J. Oyama, M. Mori, K. Kubota, L. Katsuta, K. Kimura, E. Yamaga, G. Oda, T. Nakagawa, *et al.*, "Deep-learning approach with convolutional neural network for classification of maximum intensity projections of dynamic contrast-enhanced breast magnetic resonance imaging," *Magnetic Resonance Imaging*, vol. 75, pp. 1–8, 2021.

[193] S. A. Amin, H. Al Shanabari, R. Iqbal, and C. Karyotis, "An intelligent framework for automatic breast cancer classification using novel feature extraction and machine learning techniques," *Journal of Signal Processing Systems*, pp. 1–11, 2022.

[194] H. U. Rashid, T. Ibrikci, S. Paydaş, F. Binokay, and U. Çevik, "Analysis of breast cancer classification robustness with radiomics feature extraction and deep learning techniques," *Expert Systems*, vol. 39, no. 8, p. e13018, 2022.

[195] H. Gui, T. Su, Z. Pang, H. Jiao, L. Xiong, X. Jiang, L. Li, and Z. Wang, "Diagnosis of breast cancer with strongly supervised deep learning neural network," *Electronics*, vol. 11, no. 19, p. 3003, 2022.

[196] M. Tsuchiya, T. Masui, K. Terauchi, T. Yamada, M. Katyayama, S. Ichikawa, Y. Noda, and S. Goshima, "Mri-based radiomics analysis for differentiating

phyllodes tumors of the breast from fibroadenomas," *European Radiology*, vol. 32, no. 6, pp. 4090–4100, 2022.

[197] N. Modi and K. Ghanchi, "A comparative analysis of feature selection methods and associated machine learning algorithms on wisconsin breast cancer dataset (wbcd)," in *Proceedings of International Conference on ICT for Sustainable Development*, pp. 215–224, Springer, 2016.

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

[199] W. Zhi, H. W. F. Yueng, Z. Chen, S. M. Zandavi, Z. Lu, and Y. Y. Chung, "Using transfer learning with convolutional neural networks to diagnose breast cancer from histopathological images," in *International Conference on Neural Information Processing*, pp. 669–676, Springer, 2017.

[200] S. A. Adeshina, A. P. Adedigba, A. A. Adeniyi, and A. M. Aibinu, "Breast cancer histopathology image classification with deep convolutional neural networks," in *IEEE 14th international conference on electronics computer and computation (ICECCO)*, pp. 206–212, IEEE, 2018.

[201] P. Jonnalagedda, D. Schmolze, and B. Bhanu, "MVPNets: Multi-viewing path deep learning neural networks for magnification invariant diagnosis in breast cancer," in *IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 189–194, IEEE, 2018.

[202] L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," in *International Conference on Robots & Intelligent System (ICRIS)*, pp. 157–160, IEEE, 2018.

[203] Y. Huang and A. C.-S. Chung, "Improving high resolution histology image classification with deep spatial fusion network," in *Computational Pathology and Ophthalmic Medical Image Analysis*, pp. 19–26, Springer, 2018.

[204] R. Awan, N. A. Koohbanani, M. Shaban, A. Lisowska, and N. Rajpoot, "Context-aware learning using transferable features for classification of breast cancer histology images," in *International conference image analysis and recognition*, pp. 788–795, Springer, 2018.

[205] S. S. Chennamsetty, M. Safwan, and V. Alex, "Classification of breast cancer histology image using ensemble of pre-trained neural networks," in *International conference image analysis and recognition*, pp. 804–811, Springer, 2018.

[206] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of breast cancer based on histology images using convolutional neural networks," *IEEE Access*, vol. 6, pp. 24680–24693, 2018.

[207] Z. Yang, L. Ran, S. Zhang, Y. Xia, and Y. Zhang, "EMS-Net: Ensemble of multiscale convolutional neural networks for classification of breast cancer histology images," *Neurocomputing*, vol. 366, pp. 46–53, 2019.

[208] Y. Li, J. Wu, and Q. Wu, "Classification of breast cancer histology images using multi-size and discriminative patches based on deep learning," *IEEE Access*, vol. 7, pp. 21400–21408, 2019.

[209] J. De Matos, A. d. S. Britto, L. E. Oliveira, and A. L. Koerich, "Double transfer learning for breast cancer histopathologic image classification," in *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[210] M. Abdar, M. Zomorodi-Moghadam, X. Zhou, R. Gururajan, X. Tao, P. D. Barua, and R. Gururajan, "A new nested ensemble technique for automated diagnosis of breast cancer," *Pattern Recognition Letters*, vol. 132, pp. 123–131, 2020.

[211] H. Hajiabadi, V. Babaiyan, D. Zabihzadeh, and M. Hajiabadi, "Combination of loss functions for robust breast cancer prediction," *Computers & Electrical Engineering*, vol. 84, p. 106624, 2020.

[212] Z. A. El-Shair, L. A. Sánchez-Pérez, and S. A. Rawashdeh, "Comparative study of machine learning algorithms using a breast cancer dataset," in *IEEE International Conference on Electro Information Technology (EIT)*, pp. 500–508, IEEE, 2020.

[213] S. Ray, A. AlGhamdi, K. Alshouiliy, and D. P. Agrawal, "Selecting features for breast cancer analysis and prediction," in *International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pp. 1–6, IEEE, 2020.

[214] H. Dhahri, I. Rahmany, A. Mahmood, E. Al Maghayreh, and W. Elkilani, "Tabu search and machine-learning classification of benign and malignant proliferative breast lesions," *BioMed research international*, vol. 2020, 2020.

[215] H. K. Mewada, A. V. Patel, M. Hassaballah, M. H. Alkinani, and K. Mahant, "Spectral–spatial features integrated convolution neural network for breast cancer classification," *Sensors*, vol. 20, no. 17, p. 4747, 2020.

[216] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast tumor classification using an ensemble machine learning method," *Journal of Imaging*, vol. 6, no. 6, p. 39, 2020.

[217] S. Saxena, S. Shukla, and M. Gyanchandani, "Pre-trained convolutional neural networks as feature extractors for diagnosis of breast cancer using histopathology," *International Journal of Imaging Systems and Technology*, vol. 30, no. 3, pp. 577–591, 2020.

[218] A. M. Ibraheem, K. H. Rahouma, and H. F. Hamed, "3PCNNB-Net: Three parallel cnn branches for breast cancer classification through histopathological images," *Journal of Medical and Biological Engineering*, vol. 41, no. 4, pp. 494–503, 2021.

[219] M. S. K. Inan, R. Hasan, and F. I. Alam, "A hybrid probabilistic ensemble based extreme gradient boosting approach for breast cancer diagnosis," in *IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 1029–1035, IEEE, 2021.

[220] S. Boumaraf, X. Liu, Z. Zheng, X. Ma, and C. Ferkous, "A new transfer learning based approach to magnification dependent and independent classification of breast cancer in histopathological images," *Biomedical Signal Processing and Control*, vol. 63, p. 102192, 2021.

[221] P. Ghosh, S. Azam, K. M. Hasib, A. Karim, M. Jonkman, and A. Anwar, "A performance based study on deep learning algorithms in the effective prediction of breast cancer," in *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2021.

[222] Y. Yari, H. Nguyen, and T. V. Nguyen, "Accuracy improvement in binary and multi-class classification of breast histopathology images," in *IEEE Eighth International Conference on Communications and Electronics (ICCE)*, pp. 376–381, IEEE, 2021.

[223] R. Rashmi, K. Prasad, and C. B. K. Udupa, "Bchisto-net: Breast histopathological image classification by global and local feature aggregation," *Artificial Intelligence in Medicine*, vol. 121, p. 102191, 2021.

[224] C. Hu, X. Sun, Z. Yuan, and Y. Wu, "Classification of breast cancer histopathological image with deep residual learning," *International Journal of Imaging Systems and Technology*, vol. 31, no. 3, pp. 1583–1594, 2021.

[225] N. A. Mashudi, S. A. Rossli, N. Ahmad, and N. M. Noor, "Comparison on some machine learning techniques in breast cancer classification," in *IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pp. 499–504, IEEE, 2021.

[226] S. Alkassar, B. A. Jebur, M. A. Abdullah, J. H. Al-Khalidy, and J. A. Chambers, "Going deeper: magnification-invariant approach for breast cancer classification using histopathological images," *IET Computer Vision*, vol. 15, no. 2, pp. 151–164, 2021.

[227] H. Saleh, H. Alyami, W. Alosaimi, *et al.*, "Predicting breast cancer based on optimized deep learning approach," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[228] A. M. Zaalouk, G. A. Ebrahim, H. K. Mohamed, H. M. Hassan, and M. M. Zaalouk, "A deep learning computer-aided diagnosis approach for breast cancer," *Bioengineering*, vol. 9, no. 8, p. 391, 2022.

[229] P. Madhulika and N. Sampath, "An elaborative approach for the histopathological classification of the breast cancer using residual neural networks," in *Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021*, pp. 447–456, Springer, 2022.

[230] M. Sharma, A. Mandloi, and M. Bhattacharya, "A novel deepml framework for multi-classification of breast cancer based on transfer learning," *International Journal of Imaging Systems and Technology*, vol. 32, no. 6, pp. 1963–1977, 2022.

[231] A. Labrada and B. D. Barkana, "Breast cancer diagnosis from histopathology images using supervised algorithms," in *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 102–107, IEEE, 2022.

[232] M. R. Abbasniya, S. A. Sheikholeslamzadeh, H. Nasiri, and S. Emami, "Classification of breast tumors based on histopathology images using deep features and ensemble of gradient boosting methods," *Computers and Electrical Engineering*, vol. 103, p. 108382, 2022.

[233] H. Aljuaid, N. Alturki, N. Alsubaie, L. Cavallaro, and A. Liotta, "Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning," *Computer Methods and Programs in Biomedicine*, vol. 223, p. 106951, 2022.

[234] O. El Alaoui, H. Zerouaoui, and A. Idri, "Deep stacked ensemble for breast cancer diagnosis," in *Information Systems and Technologies: WorldCIST 2022, Volume 1*, pp. 435–445, Springer, 2022.

[235] F.-Z. Nakach, H. Zerouaoui, and A. Idri, "Hybrid deep boosting ensembles for histopathological breast cancer classification," *Health and Technology*, pp. 1–18, 2022.

[236] S. I. Khan, A. Shahrior, R. Karim, M. Hasan, and A. Rahman, "Multinet: A deep neural network approach for detecting breast cancer through multi-scale feature fusion," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 8, pp. 6217–6228, 2022.

[237] Y. Xu, M. A. dos Santos, L. F. F. Souza, A. G. Marques, L. Zhang, J. J. da Costa Nascimento, V. H. C. de Albuquerque, and P. P. Rebouças Filho, "New fully automatic approach for tissue identification in histopathological examinations using transfer learning," *IET Image Processing*, vol. 16, no. 11, pp. 2875–2889, 2022.

[238] D. Silva and O. Cortes, "On convolutional neural networks and transfer learning for classifying breast cancer on histopathological images using GPU," in *Brazilian Congress on Biomedical Engineering: Proceedings of CBEB 2020, October 26–30, 2020, Vitória, Brazil*, pp. 1993–1998, Springer, 2022.

[239] A. Baccouche, B. Garcia-Zapirain, Y. Zheng, and A. S. Elmaghraby, "Early detection and classification of abnormality in prior mammograms using image-to-image translation and yolo techniques," *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106884, 2022.

[240] C. Fitzmaurice, "Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years

for 29 cancer groups, 1990 to 2017: A systematic analysis for the global burden of disease study," *JAMA Oncology*, vol. 5, no. 12, pp. 1749–1768, 2019.

[241] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[242] B. Weigelt, F. C. Geyer, and J. S. Reis-Filho, "Histological types of breast cancer: How special are they?," *Molecular Oncology*, vol. 4, no. 3, pp. 192–208, 2010.

[243] L. Wang, "Early diagnosis of breast cancer," *Sensors*, vol. 17, no. 7, p. 1572, 2017.

[244] M. Aswathy and M. Jagannath, "Detection of breast cancer on digital histopathology images: Present status and future possibilities," *Informatics in Medicine Unlocked*, vol. 8, pp. 74–79, 2017.

[245] S. Robertson, H. Azizpour, K. Smith, and J. Hartman, "Digital image analysis in breast pathology—from image processing techniques to artificial intelligence," *Translational Research*, vol. 194, pp. 19–35, 2018.

[246] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, pp. 1150–1157, Ieee, 1999.

[247] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture notes in computer science*, vol. 3951, pp. 404–417, 2006.

[248] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[249] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[250] S. Zahia, M. B. G. Zapirain, X. Sevillano, A. González, P. J. Kim, and A. Elmaghraby, "Pressure injury image analysis with machine learning techniques: A systematic review on previous and possible future methods," *Artificial intelligence in medicine*, vol. 102, p. 101742, 2020.

[251] A. Bour, C. Castillo-Olea, B. Garcia-Zapirain, and S. Zahia, "Automatic colon polyp classification using convolutional neural network: a case study at basque country," in *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 1–5, IEEE, 2019.

[252] M. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz, and R. Monczak, "Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images," *Computers in biology and medicine*, vol. 43, no. 10, pp. 1563–1572, 2013.

[253] P. Filipczuk, T. Fevens, A. Krzyżak, and R. Monczak, "Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies," *IEEE transactions on medical imaging*, vol. 32, no. 12, pp. 2169–2178, 2013.

[254] Y. Zhang, B. Zhang, F. Coenen, and W. Lu, "Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles," *Machine vision and applications*, vol. 24, no. 7, pp. 1405–1420, 2013.

[255] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[256] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *2016 23rd International conference on pattern recognition (ICPR)*, pp. 2440–2445, IEEE, 2016.

[257] P. Bankhead, M. B. Loughrey, J. A. Fernández, Y. Dombrowski, D. G. McArt, P. D. Dunne, S. McQuaid, R. T. Gray, L. J. Murray, H. G. Coleman, *et al.*, "Qupath: Open source software for digital pathology image analysis," *Scientific reports*, vol. 7, no. 1, pp. 1–7, 2017.

[258] F. Bianconi, J. N. Kather, and C. C. Reyes-Aldasoro, "Evaluation of colour pre-processing on patch-based classification of h&e-stained images," in *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, pp. 56–64, Springer, 2019.

[259] H. Yao, X. Zhang, X. Zhou, and S. Liu, "Parallel structure deep neural network using cnn and rnn with an attention mechanism for breast cancer histology image classification," *Cancers*, vol. 11, no. 12, p. 1901, 2019.

[260] F. Chollet *et al.*, "Keras: Deep learning library." https://keras.io/api/preprocessing/image/, 2015.

[261] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[262] J. Xie, R. Liu, J. Luttrell IV, and C. Zhang, "Deep learning based analysis of histopathological images of breast cancer," *Frontiers in genetics*, vol. 10, p. 80, 2019.

[263] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[264] 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, no. 1, p. 4172, 2017.

[265] A.-A. Nahid, M. A. Mehrabi, and Y. Kong, "Histopathological breast cancer image classification by deep neural network techniques guided by local clustering," *BioMed research international*, vol. 2018, 2018.

[266] E. Deniz, A. Şengür, Z. Kadiroğlu, Y. Guo, V. Bajaj, and Ü. Budak, "Transfer learning based histopathologic image classification for breast cancer detection," *Health information science and systems*, vol. 6, pp. 1–7, 2018.

[267] Y. Benhammou, B. Achchab, F. Herrera, and S. Tabik, "Breakhis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights," *Neurocomputing*, vol. 375, pp. 9–24, 2020.

[268] Y. Feng, M. Spezia, S. Huang, C. Yuan, Z. Zeng, L. Zhang, X. Ji, W. Liu, B. Huang, W. Luo, *et al.*, "Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis," *Genes & diseases*, vol. 5, no. 2, pp. 77–106, 2018.

[269] C. Elfgen, B. Papassotiropoulos, Z. Varga, L. Moskovszky, M. Nap, U. Güth, A. Baege, E. Amann, F. Chiesa, and C. Tausch, "Comparative analysis of confocal microscopy on fresh breast core needle biopsies and conventional histology," *Diagnostic pathology*, vol. 14, no. 1, pp. 1–8, 2019.

[270] A. Ibrahim, P. Gamble, R. Jaroensri, M. M. Abdelsamea, C. H. Mermel, P.-H. C. Chen, and E. A. Rakha, "Artificial intelligence in digital breast pathology: techniques and applications," *The Breast*, vol. 49, pp. 267–273, 2020.

[271] Y. Jiang, L. Chen, H. Zhang, and X. Xiao, "Breast cancer histopathological image classification using convolutional neural networks with small se-resnet module," *PloS one*, vol. 14, no. 3, p. e0214587, 2019.

[272] H. Elmannai, M. Hamdi, and A. AlGarni, "Deep learning models combining for breast cancer histopathology image classification," *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, pp. 1003–1013, 2021.

[273] S. Sharma and S. Kumar, "The xception model: A potential feature extractor in breast cancer histology images classification," *ICT Express*, vol. 8, no. 1, pp. 101–108, 2022.

[274] F. Bianconi, J. N. Kather, and C. C. Reyes-Aldasoro, "Experimental assessment of color deconvolution and color normalization for automated classification of histology images stained with hematoxylin and eosin," *Cancers*, vol. 12, no. 11, p. 3337, 2020.

[275] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Breast cancer diagnosis with transfer learning and global pooling," in *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 519–524, IEEE, 2019.

[276] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis," *Computers in biology and medicine*, vol. 128, p. 104129, 2021.

[277] H. O. Lyon, A. De Leenheer, R. Horobin, W. Lambert, E. Schulte, B. Van Liedekerke, and D. Wittekind, "Standardization of reagents and methods used in cytological and histological practice with emphasis on dyes, stains and chromogenic reagents," *The Histochemical Journal*, vol. 26, no. 7, pp. 533–544, 1994.

[278] S. Roy, A. kumar Jain, S. Lal, and J. Kini, "A study about color normalization methods for histopathology images," *Micron*, vol. 114, pp. 42–61, 2018.

[279] C. Zhu, F. Song, Y. Wang, H. Dong, Y. Guo, and J. Liu, "Breast cancer histopathology image classification through assembling multiple compact cnns," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–17, 2019.

[280] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015.

[281] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[282] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015.

[283] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[284] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.

[285] Ł. Rączkowski, M. Możejko, J. Zambonelli, and E. Szczurek, "Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[286] Y. Hao, L. Zhang, S. Qiao, Y. Bai, R. Cheng, H. Xue, Y. Hou, W. Zhang, and G. Zhang, "Breast cancer histopathological images classification based on deep semantic features and gray level co-occurrence matrix," *Plos one*, vol. 17, no. 5, p. e0267955, 2022.