

**DEEP LEARNING-BASED BREAST CANCER DETECTION
AND CLASSIFICATION USING HISTOPATHOLOGY
IMAGES**

GHULAM MURTAZA

**FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR**

2021

**DEEP LEARNING-BASED BREAST CANCER DETECTION
AND CLASSIFICATION USING HISTOPATHOLOGY
IMAGES**

GHULAM MURTAZA

**THESIS SUBMITTED IN FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF DOCTOR OF
PHILOSOPHY**

**FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR**

2021

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Ghulam Murtaza

Matric No: WVA170003 (New Matric No: 17043591/1)

Name of Degree: Doctor of Philosophy

Title of Thesis: Deep Learning-based Breast Cancer Detection and Classification using Histopathology Images

Field of Study: Machine Learning

I do solemnly and sincerely declare that:

- (1) I am the sole author/writer of this Work;
- (2) This Work is original;
- (3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyrighted work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
- (4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyrighted work;
- (5) I hereby assign all and every right in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be the owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
- (6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature

Date:09/02/2021

Subscribed and solemnly declared before,

Witness's Signature

Name:

Name:

Designation:

Designation:

Date: 10/2/2021

Date: 10/2/2021

DEEP LEARNING-BASED BREAST CANCER DETECTION AND CLASSIFICATION USING HISTOPATHOLOGY IMAGES

ABSTRACT

Cancer disease is drastically increasing worldwide over the past few years. Among all types of cancers in women, breast cancer (BrC) is the main cause of abnormal deaths. For a confident diagnosis of BrC, histopathology (Hp) images are usually suggested by the doctors. BrC detection is a diagnostic test for benign (non-cancerous) and malignant (cancerous) breast tumors (BrT). Once the BrT is diagnosed, then it needs to be classified for subtypes of benign and malignant to start specific treatment. Several studies developed BrC detection and classification models using Hp images. However, the existing models required high computational resources, long training time, and their performance is compromised due to a higher misclassification rate. Thus, this research is aimed to develop two models. First, the BrC detection model is developed to diagnose BrT basic types like benign and malignant. Second, the BrT classification model is developed to diagnose subtypes of benign and malignant tumors. To perform overall experiments, Hp images of the BreakHis dataset are utilized. BreakHis is a large and complex dataset (i.e., four subtypes of each benign and malignant BrTs) that publicly available. For BrC detection, an efficient and reliable model namely Ensemble BrC Detection Network (EBrC-Net) and three misclassification reduction (McR) algorithms are developed. The proposed EBrC-Net model is based on deep learning (DL) based approach. EBrC-Net architecture is designed to require less training time and computational resources like a normal desktop computer. The trained EBrC-Net is used to extract discriminative features. The extracted features are evaluated through six machine learning (ML) classifiers namely softmax, k-nearest neighbor (kNN), support vector machine, linear discriminant analysis, decision tree, and naive Bayes. Experimentally, it has been

observed that kNN outperformed the rest of the five ML classifiers. Furthermore, three McR algorithms are developed and implemented in a cascaded manner to reduce the false predictions (i.e., misclassification) of the aforementioned six ML classifiers. The proposed BrC detection model for five folds of features achieved mean accuracy, sensitivity, and patient recognition rate by 97.78%, 97.28%, and 97.92% respectively. On the other hand, BrT classification is aimed to develop an efficient and reliable model namely Biopsy Microscopic Image Cancer Network (BMIC-Net) to classify Hp images into eight subtypes of BrT through a DL-based hierarchical classification approach. BMIC-Net model can be trained using less computational resources in less time. The trained BMIC-Net is used to extract discriminative features from Hp images. To reduce the misclassification, a feature selection algorithm (using *information gain* and *principal component analysis* schemes) is developed to elicit the most discriminative feature subset. Finally, the aforementioned six ML classifiers are analyzed to acquire the best performing classifier. The experimental results revealed that BMIC-Net outperformed for five folds of features by obtaining a mean accuracy of 95.33% for first-level hierarchical classifier and 94.70%, 92.53% for second-level hierarchical classifiers. Moreover, the performances of both BrC detection and BrT classification are compared with existing state-of-art baseline studies. Findings discovered that the proposed models are efficient (i.e., consume less computational resources and training time) and reliable (i.e., reduce misclassification to show better and unbiased results even using a complex dataset) in comparison with the existing SoA baseline studies. Thus, the proposed BrC detection and classification models can assist doctors to serve on the basis of the second opinion for early diagnosis of BrC.

Keywords: Breast Cancer Detection, Medical Image Classification, Deep Learning, Histopathology Images.

PENGESANAN DAN PENGELASAN KANSER PAYUDARA BERASASKAN PEMBELAJARAN DALAM MENGGUNAKAN IMEJ HISTOPATOLOGI

ABSTRAK

Penyakit kanser meningkat secara mendadak di seluruh dunia sejak beberapa tahun yang lalu. Di antara semua jenis kanser pada wanita, kanser payudara (BrC) adalah penyebab utama kematian yang tidak normal. Untuk diagnosis BrC yang meyakinkan, imej-imej histopatologi (Hp) biasanya disarankan oleh doktor-doktor. Pengesanan BrC adalah ujian diagnostik untuk “benign” (tidak kanser) dan “malignant” (kanser) ketumbuhan payudara (BrT). Apabila BrT didiagnosis, maka ia perlu diklasifikasikan kepada sub jenis “benign” dan “malignant” untuk memulakan rawatan tertentu. Beberapa kajian telah membangunkan model pengesanan dan pengelasan BrC menggunakan imej Hp. Walau bagaimanapun, model-model sedia ada memerlukan sumber pengiraan yang tinggi, masa latihan yang panjang dan prestasi mereka juga telah terjejas kerana kadar ralat klasifikasi yang lebih tinggi. Oleh itu, kajian ini bertujuan untuk membangunkan dua model. Pertama, model pengesanan BrC dibangunkan untuk mendiagnosis jenis asas BrT seperti “benign” dan “malignant”. Kedua, model pengelasan BrT dibangunkan untuk mendiagnosis sub jenis tumor “benign” dan “malignant”. Untuk melaksanakan eksperimen secara keseluruhan, imej Hp set data BreakHis telah digunakan. BreakHis adalah set data yang besar dan kompleks (iaitu, empat sub jenis daripada setiap BrTs “benign” dan “malignant”) yang adasecara terbuka. Untuk pengesanan BrC, model yang cekap dan boleh dipercayai iaitu Rangkaian Pengesanan Kumpulan BrC (EBrC-Net) dan tiga algoritma pengurangan ralat klasifikasi (McR) telah dibangunkan. Model EBrC-Net yang dicadangkan adalah berdasarkan pendekatan pembelajaran dalam (DL). Senibina EBrC-Net direka untuk memerlukan masa latihan yang kurang dan sumber komputasi seperti komputer biasa. EBrC-Net yang terlatih digunakan untuk mengekstrak ciri berorientasikan hasil. Ciri-ciri yang diekstrak telah dinilai melalui enam pengelas

pembelajaran mesin (ML) yaitu softmax, k-nearest neighbor (kNN), mesin vektor sokongan, analisis diskriminasi linear, pepohon sokongan, dan naive Bayes. Secara eksperimen, kNN mengatasi lima pengelas ML yang lain. Tambahan lagi, tiga algoritma McR telah dibangun dan dilaksanakan dengan cara yang tersusun untuk mengurangkan ramalan palsu (iaitu, ralat klasifikasi) untuk keenam-enam pengelas ML. Model pengesanan BrC yang dicadangkan untuk lima lapisan ciri telah mencapai purata ketepatan, kepekaan dan kadar pengiktirafan pesakit, masing-masing sebanyak 97.78%, 97.28% dan 97.92%. Selain itu, klasifikasi BrT bertujuan untuk membangunkan model yang cekap dan boleh dipercayai iaitu Rangkaian Kanser Mikroskopik Biopsi (BMIC-Net) untuk mengelaskan imej Hp kepada lapan sub jenis BrT melalui pendekatan pengelasan hierarki DL. Model BMIC-Net boleh dilatih menggunakan sumber kurang pengiraan dalam masa yang singkat. BMIC-Net terlatih telah digunakan untuk mengeluarkan ciri-ciri yang berbeza daripada imej Hp. Untuk mengurangkan ralat klasifikasi, algoritma pemilihan ciri telah dibangun (menggunakan *information gain* dan *principal component analysis* skema) untuk memperoleh subset ciri yang paling diskriminatif. Akhirnya, keenam-enam kelas ML yang dinyatakan di atas telah dianalisis untuk memperoleh pengelas terbaik. Keputusan eksperimen menunjukkan bahawa BMIC-Net telah mengatasi untuk lima lapisan ciri dari segi prestasi dengan memperoleh purata ketepatan yang lebih baik daripada 95.33% untuk pengelas hierarki peringkat pertama dan 94.70%, 92.53% untuk pengelas hierarki peringkat kedua. Selain itu, prestasi bagi kedua-dua pengesanan dan pengelasan BrC telah dibandingkan dengan kajian dasar terkini yang sedia ada. Hasil penyelidikan mendapati bahawa kedua-dua model yang dicadangkan adalah cekap (iaitu, menggunakan kurang sumber pengiraan dan masa latihan) dan boleh dipercayai (iaitu, mengurangkan salah klasifikasi untuk menunjukkan hasil yang lebih baik dan tidak berat sebelah walaupun menggunakan set data yang kompleks) berbanding dengan dasar kajian terkini yang sedia ada. Oleh itu, model

pengesanan dan pengelasan BrC yang telah dicadangkan boleh membantu para doktor sebagai asas pendapat kedua untuk diagnosis awal BrC.

Kata kunci: Pengesanan Kanser Payudara, Pengelasan Imej Perubatan, Pembelajaran Dalam, Imej Histopatologi.

Universiti Malaya

ACKNOWLEDGMENTS

In the name of Allah, the Most Gracious and the Most Merciful, Peace and blessings of Allah upon his Prophet Muhammad S.A.W. Alhamdulillah, all praise to Allah for endowing me the strength, wisdom, and endless blessings to do my Ph.D. research work.

Special appreciation goes to my supervisors, Dr. Nor Liyana Mohd Shuib and Associate Professor Dr. Ainuddin Wahid Abdul Wahab for their precious support, supervision, encouragement, and inspiration during my Ph.D. journey of three years. Their consistent support and guidance helped me to produce a valuable piece of research delivered in this thesis. I am also thankful to Dr. Ghulam Raza for helping me as a medical advisor to understand the BrC medical imaging modalities and related concepts. I would also like to extend my heartiest appreciation to my employer Sukkur Institute of Business Admiration University (SIBAU), which has financially supported my Ph.D. studies throughout my tenure.

I would like to extend wholehearted appreciation to my beloved parents who have sacrificed their time, efforts to enable me to become a person of importance and value to society. They always offered me continuous help and support in the ups and downs of real-life. I also pay gratitude to my dear companion wife Rubeena Murtaza, my two little daughters Maidah Murtaza and Manaal Murtaza for their cooperation and prayers during my Ph.D. work. I wish to express my acknowledgments to my friends, colleagues, and other contributors whose names are not mentioned above. Thanks to all for being very supportive.

This Ph.D. work is dedicated to my grandmother Late Janul Memon, my father Late Ghulam Mustafa Memon, my mother Rashida Memon, my wife, and employer SIBAU for their endless support and motivation. Special thanks to my beloved mother, siblings, mother-in-law, and wife for their moral support, unconditional love, and prayers throughout the journey.

TABLE OF CONTENTS

Abstract	iii
Abstrak	v
Acknowledgments.....	viii
Table of Contents	ix
List of Figures	xv
List of Tables	xviii
List of Symbols and Abbreviations.....	xx
List of Appendices	xxii
CHAPTER 1: INTRODUCTION.....	1
1.1 Background.....	1
1.2 Research Motivation.....	9
1.3 Statement of Problem	10
1.4 Research Aims and Objectives	11
1.5 Research Questions.....	11
1.6 Research Methodology	13
1.7 Research Scope.....	16
1.8 Research Contribution	17
1.9 Research Significance.....	19
1.9.1 Significance to Doctors and Pathologists	19
1.9.2 Research Significance for Researchers for BrC Detection and Classification Models Development	20
1.10 Thesis Overview	21
CHAPTER 2: LITERATURE REVIEW.....	24

2.1	Introduction	24
2.2	Breast Cancer.....	24
2.3	Medical Imaging Modalities.....	25
2.3.1	Mammogram	27
2.3.2	Ultrasound	33
2.3.3	Magnetic Resonance Imaging	35
2.3.4	Histopathology Images.....	36
2.3.5	Multimodalities	38
2.4	Breast Cancer Classification Dataset Analysis.....	39
2.5	Medical Image Preprocessing.....	45
2.5.1	Augmentation	49
2.5.2	Image Region of Interest Extraction	50
2.5.3	Scaling	50
2.5.4	Normalization and Enhancement	51
2.5.5	Removing Artifacts	52
2.5.6	Stain Normalization.....	53
2.6	Machine Learning Based Classification Model Types Used for BrC Detection and Classification	54
2.6.1	Traditional Machine Learning Classification Models Used for BrC Detection and Classification.....	54
2.6.2	Artificial Neural Network Used in BrC Detection and Classification	57
2.6.2.1	Shallow Neural Network.....	58
2.6.2.2	Deep Neural Networks	61
2.6.3	Empirical Analysis of Traditional Machine Learning Models Vs. Deep Learning Models for BrC Detection and Classification.....	73

2.6.4	Empirical Evaluation of BrC Deep Neural Network Models Using Different Datasets	74
2.7	Evaluation Metrics Analysis and Review	81
2.7.1	Accuracy	81
2.7.2	Sensitivity	81
2.7.3	Specificity	82
2.7.4	Precision	82
2.7.5	FMeasure	82
2.7.6	Area Under the ROC Curve	82
2.7.7	The Volume Under the ROC Surface	83
2.7.8	Patient Recognition Rate	84
2.7.9	Cross-entropy Loss	84
2.8	Limitations Related to the Existing Literature	87
2.8.1	Limitations of Artificial Neural Networks Based Models	87
2.8.2	Limitations of Performance Evaluation Metrics	89
2.8.3	Low Model Performance (i.e., Higher Misclassification)	90
2.9	Research Gap Analysis for Problem Identification	90
2.10	Summary	92
CHAPTER 3: METHODOLOGY AND EXPERIMENTAL SETUP		95
3.1	Introduction	95
3.2	Methodology	95
3.2.1	Breast Cancer Detection Model Construction Methodology	95
3.2.1.1	Data Collection	97
3.2.1.2	Image Preprocessing	99
3.2.1.3	Development of BrC Detection Model	102
3.2.1.4	Breast Cancer Detection Techniques	104

3.2.1.5	Model Construction and Evaluation.....	105
3.2.2	Breast Tumor Classification Model Construction Methodology	107
3.2.2.1	Data Collection.....	108
3.2.2.2	Image Preprocessing	109
3.2.2.3	Development of BrT Classification Model	110
3.2.2.4	Breast Cancer Classification Technique	112
3.2.2.5	Model Construction and Evaluation.....	113
3.3	Experimental Setup.....	115
3.3.1	Experimental Setup of BrC Detection Model	115
3.3.2	Experimental Setup of BrT Classification Model	117
3.4	Summary.....	119
CHAPTER 4: DEVELOPMENT OF BREAST CANCER DETECTION AND CLASSIFICATION MODELS.....		121
4.1	Introduction	121
4.2	Development of BrC Detection Model.....	122
4.2.1	Pre-trained AlexNet Architecture.....	123
4.2.2	Proposed DL-based BrC Detection Model.....	126
4.2.2.1	EBrC-Net Architecture and Model Structure.....	127
4.3	Performance Enhancement of BrC Detection Model	128
4.3.1	Selection of Image Augmentation Method	129
4.3.2	Misclassification Reduction Algorithms	129
4.4	Development of BrT Classification Model	133
4.4.1	Proposed BMIC-Net.....	134
4.4.1.1	BMIC-Net Architecture and Model Structure.....	135
4.4.1.2	BMIC-Net Model Structure	135
4.5	Performance Enhancement of BrT Classification Model.....	138

4.5.1	Feature Reduction and Selection Algorithm	138
4.6	Summary.....	141
CHAPTER 5: EXPERIMENTAL RESULTS AND DISCUSSION.....		142
5.1	Introduction	142
5.2	Experimental Results.....	142
5.2.1	Experimental Results of BrC Detection Model.....	143
5.2.1.1	Experimental Results of Setting I.....	143
5.2.1.2	Experimental Results of Setting II	145
5.2.1.3	Experimental Results of Setting III.....	147
5.2.1.4	Experimental Results of Setting IV.....	149
5.2.2	Experimental Results of BrT Classification Model	151
5.2.2.1	Setting I Experimental Results.....	151
5.2.2.2	Setting II Experimental Results.....	152
5.2.2.3	Setting III Experimental Results	155
5.2.2.4	Setting IV Experimental Results	158
5.3	Discussion.....	162
5.3.1	State-of-the-art BrC Detection and Classification Models Analysis.....	162
5.3.2	Proposed BrC Detection Model Discussion.....	164
5.3.3	Proposed BrC Detection Model Baseline Comparison	169
5.3.4	Proposed BrT Classification Model Discussion.....	171
5.3.5	Proposed BrT Classification Model Baseline Comparison.....	172
5.4	State-of-the-art versus proposed models	174
5.5	Summary.....	176
CHAPTER 6: CONCLUSION.....		179
6.1	Introduction	179

6.2	Reappraisal of Research Objectives and Research Questions.....	182
6.3	Limitations of Proposed BrC Detection and Classification Models.....	188
6.4	Future Research Directions	189
6.5	Summary.....	189
	List of Publications and Papers Presented	190
	Appendix-A: Breast Cancer detection	191
	Appendix-B: Breast Tumor Classification.....	194
	References.....	195

Universiti Malaya

LIST OF FIGURES

Figure 1.1: Steps involved in research methodology	13
Figure 2.1: (A) Mammogram Screening: Masses with areas of varying density reflecting the presence of elements which are of fat and soft-tissue density (Jonathan J. James, 2016). (B) Left: A mammogram image view, Right: A clustered micro-calcifications in magnified view (Jing, Yang, & Nishikawa, 2012).....	28
Figure 2.2: (A) Well-defined rounded mass mammogram. (B) The absence of internal echoes and the posterior enhancement of the ultrasound beam are diagnostic of a cyst or lump or mass (Jonathan J. James, 2016)	33
Figure 2.3: Left side US image (B-Mode). Shear-wave elastography image on the right side shows an irregular mass in red color, known as heterogeneous elasticity. The statistical parameters (e.g. Mean, Minimum, Maximum, etc.) of ROI (a large circle) are calculated (Youk et al., 2017)	34
Figure 2.4: Left US image (B-mode) of a lesion reconstructed using the RF data on the right side corresponding Nakagami map (Byra et al., 2017)	35
Figure 2.5: Samples of Breast MRI images (Breast Cancer Imaging, 2018).....	35
Figure 2.6: Histopathology WSI is shown on the left at low magnification and a cropped region is shown on the right at high magnification (Liu, Hernandez-Cabronero, Sanchez, Marcellin, & Bilgin, 2017).....	37
Figure 2.7: Histopathology image patches showing eight subtypes of breast cancer (Spanhol et al., 2016b)	37
Figure 2.8: Multimodalities used for BrT classification. The left image is a mammogram showing a solid mass. The Center image is the US image showing stiff tissues as black. The right side image is MRI providing a clear view of breast mass (Breast Cancer Imaging, 2018)	39
Figure 2.9: Different artifacts in a mammogram (Left image) and MRI (right image) (Saidin, Sakim, Ngah, & Shuaib, 2012; Breast Cancer Imaging, 2018).....	53
Figure 2.10: Source image stain normalized by using a reference image through three techniques.....	53
Figure 2.11: Left: An artificial neuron. Right: sample of an artificial neural network...	58
Figure 2.12: Type of ANNs used for BrC detection and BrT classification.....	59
Figure 2.13: A sample illustration of Multi-Layer Neural Network.....	63

Figure 2.14: A restricted Boltzmann machine (RBM) with fully connected visible and hidden units (a), a sample diagram of supervised PGBM shown (b) (Sohn, Zhou, Lee, & Lee, 2013) and (c) shows a sample diagram of supervised DBN	67
Figure 2.15: Left side figure, a sample network diagram of the traditional autoencoder. Right side figure, a network diagram of stacked denoising autoencoder	68
Figure 2.16: A two-staged PCANet block diagram sample (Chan et al., 2015).....	69
Figure 2.17: An illustration of deep CNN-based model for BrT classification using mammograms.....	71
Figure 2.18: (a) A sample ROC diagram, comparing the performance of four classification models of breast cancer. (b) Illustration of sample VUS diagram for three classes.....	83
Figure 3.1: BrC detection model construction methodology.....	96
Figure 3.2: Reinhard method used to normalize source image through reference image	100
Figure 3.3: BrT classification model construction methodology.....	107
Figure 3.4: The experimental setups distribution overview.....	115
Figure 4.1: Contributions for BrC detection and classification	121
Figure 4.2: AlexNet architecture.....	124
Figure 4.3: Network architecture of proposed EBrC-Net	128
Figure 4.4: Network architectures' of BMIC-Net model classifiers' for hierarchical BrT classification.....	136
Figure 4.5: Proposed hierarchical classification flow diagram.....	137
Figure 5.1: The experimental results distribution overview	142
Figure 5.2: Input image size optimization for EBrC-Net.....	144
Figure 5.3: Epoch-wise comparison of the AlexNet model with EBrC-Net model. (a) Validation loss comparison. (b) The validation set accuracy comparison.....	145
Figure 5.4: Parameter optimization of kNN and SVM classifiers	146
Figure 5.5. Performance of six ML classifiers.....	147

Figure 5.6. Analysis of PRR before and after McR and performance of three misclassification algorithms.....	150
Figure 5.7: Epoch-wise comparison of BMIC-Net with non-hierarchical model.....	152
Figure 5.8: Parameter optimization for kNN and SVM.....	153
Figure 5.9. Model wise six traditional ML classifiers accuracies.....	154
Figure 5.10: Feature reduction and selection with overall accuracies	156
Figure 5.11: ROCs after feature reduction.....	158
Figure 6.1: Schematic mapping of research objectives.....	183

Universiti Malaysia

LIST OF TABLES

Table 2.1: Distribution of studies for various medical imaging modalities.	26
Table 2.2: Studies, imaging modality, strengths, weakness, and applications of various medical imaging modalities used in BrT classification	30
Table 2.3: Publically available datasets and corresponding URL	40
Table 2.4: Detailed analysis of public datasets used in breast cancer classification.....	42
Table 2.5: Distribution of studies among preprocessing methods and their advantages	45
Table 2.6: Distribution of studies using various machine learning classifiers for BrC detection and classification	56
Table 2.7: Brief description of popular activation functions	60
Table 2.8: ANN models used in selected studies for BrT classification.....	63
Table 2.9: Study-wise performance of ANNs for breast cancer detection and classification	77
Table 2.10: Frequency count of performance metrics used in each selected primary study	85
Table 3.1: BreakHis dataset images distribution.....	98
Table 3.2: Patient-wise split of BreakHis (40× magnification) dataset	98
Table 3.3: Augmented training set distribution using BreakHis (40× magnification) dataset.....	101
Table 3.4: Utilized BreakHis (40× magnification) dataset distribution.....	102
Table 3.5: Images selected for training and testing.....	109
Table 5.1: Performance comparison of McR algorithms using machine learning classifiers	149
Table 5.2: AUC values for proposed BC ₁ , B ₂ , and M ₂ classifiers	155
Table 5.3: AUC values for proposed BC ₁ , B ₂ , and M ₂ classifiers after feature reduction	157
Table 5.4: Non-hierarchical model performance before and after feature reduction....	159

Table 5.5: Performance comparison of the proposed hierarchical model with the non-hierarchical model, before and after feature reduction	160
Table 5.6: Performance comparison of proposed EBrC-Net model to the state-of-the-art existing models using BreakHis dataset.....	170
Table 5.7: Performance comparison of proposed BMIC-Net model to the state-of-the-art existing models using BreakHis Dataset.....	173

Universiti Malaya

LIST OF SYMBOLS AND ABBREVIATIONS

Abbreviation	Full form
A	: Adenosis
Ac	: Accuracy
ANN	: Artificial Neural Network
AUC	: Area Under the Curve
BCBH	: Bioimaging Challenge 2015 Breast Histology
BCDR	: Breast Cancer Digital Repository
BMIC-Net	: Biopsy Microscopic Image Cancer Network
BrC	: Breast Cancer
BreakHis	: The Breast Cancer Histopathological
BrT	: Breast Tumor
CBIS	: Curated Breast Imaging Subset
DBT	: Digital breast tomography
CNN	: Convolutional Neural Network
CT	: Computerized Tomography
DC	: Ductal Carcinoma
DCE	: Dynamic Contrast-enhanced
DDSM	: Digital Database for Screening Mammography
DeCAFs	: Deep Convolution Activation Features
DL	: Deep Learning
DNN	: Deep Neural Network
DP	: Digital Pathology
EBrC-Net	: Ensembled Breast Cancer Network
F	: Fibroadenoma
FN	: False Negative
FP	: False Positive
GPU	: Graphics Processing Unit
H&E	: Hematoxylin-Eosin
HeFs	: Hand-engineered Features
Hp	: Histopathology
IG	: Information Gain
kNN	: K-Nearest Neighbor
LC	: Lobular Carcinoma
LR	: Linear Regression
MC	: Mucinous Carcinoma
McR	: Misclassification Reduction
McRC	: Misclassification Reduction Confidence-wise
McRI	: Misclassification Reduction Image-wise
McRP	: Misclassification Reduction Patient-wise
MFV	: Master Feature Vector
MG	: Mammogram
MIAS	: Mammographic Image Analysis Society
ML	: Machine Learning
MRI	: Magnetic Resonance Imaging
PC	: Papillary Carcinoma
PCA	: Principal Component Analysis
PEMs	: Performance Evaluation Metrics
Pr	: Precision
PRR	: Patient Recognition Rate

Abbreviation	Full form
PT	: Phyllodes Tumor
Rc	: Recall
RF	: Random Forests
RO	: Research Objective
ROI	: Region Of Interest
RQ	: Research Question
SFM	: Screen-film mammograms
Sn	: Sensitivity
SNN	: Shallow Neural Network
SoA	: State-of-the-art
Sp	: Specification
SVM	: Support Vector Machine
SWE	: Shear-wave elastography
TA	: Tubular Adenoma
TL	: Transfer Learning
TN	: True Negative
TP	: True Positive
US	: Ultrasound
VUS	: Volume Under Surface
WHO	: World Health Organization
WSI	: Whole Slide Image

Universiti Malaysia

LIST OF APPENDICES

Appendix A: Breast cancer detection.....	191
Appendix B: Breast Tumor Classification.....	194

Universiti Malaya

CHAPTER 1: INTRODUCTION

This chapter discusses the overall research background and underlying motivation. It also presents the problem statement, followed by the research questions and research objectives. Moreover, it also briefly describes the research design, scope, contribution, and significance of the overall research. Finally, it states the organization of the overall research work presented in this thesis.

1.1 Background

Cancer is the most prevailing cause of abnormal deaths and a massive problem for public health around the globe. In 2015, 8.8 million deaths caused by cancer and 27 million new cases of cancer are expected till 2030, reported by the International Agency of Research on Cancer, affiliated with World Health Organization (WHO, 2018). Moreover, among all types of cancers, breast cancer (BrC) is the foremost cause of mortality (i.e., 571000 deaths) in women. Whereas, 30% to 50% of cancer burden can be reduced by early diagnosis of cancer, reported by WHO. Initially, BrC detection is a diagnostic test to identify two main types of breast tumors (BrTs) like benign and malignant, which can be classified into further subtypes. Benign is known as a non-invasive/non-cancerous tumor, whereas malignant is an invasive/cancerous type of tumor. Non-invasive BrTs have not spread to nearby tissue or beyond. Conversely, invasive (known as cancerous) BrTs spread to the surrounding breast tissues and other parts of the body, thus can cause abnormal death if does not diagnose at an early stage.

Medical images, such as histopathology (Hp) and radiology images, are used as a diagnostic test for BrC (i.e., benign or malignant). Radiology images, for instance, mammograms, can locate BrC lesions but cannot verify whether a highlighted location is cancerous. However, in a breast biopsy, a small sample of tissue is obtained from a suspicious area of the breast and fixed into slides for manual examination under a

microscope. Microscopic manual examination of the breast tissue gives a more credible cancer diagnosis in comparison with radiology images. Breast Hp slides enable tissue-level analysis, enabling pathologists to distinguish types of cell nuclei and the shapes and architectures of specific patterns. Moreover, Hp slide manual analysis allows visual examination of cell shape abnormality and distribution and helps in determining the breast lesion classification up to eight subtypes of BrT namely adenosis (A), fibroadenoma (F), tubular adenoma (TA), phyllodes tumor (PT), ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). However, the reliability of Hp sample manual analysis solely depends upon the daily workload, laboratory environment, and pathologists' domain knowledge, expertise, and field experience (Vestjens et al., 2012; Ehteshami, Veta, Johannes van Diest, et al., 2017).

In Elmore, Longton, and Carney (2015) study, 6900 cases were diagnosed by 115 pathologists and compared with expert consensus-driven ground truth; 20% of benign cases were misdiagnosed (i.e., misclassified) as malignant, whereas 10% of malignant cases were misclassified as benign. In addition, a pathologist working as a general pathologist and performing manual analysis with a small number of cases in a week makes more diagnostic errors than does an expert pathologist (Allison et al., 2014). Moreover, arriving at a diagnostic consensus is a major issue among expert pathologists due to professional differences of opinion on features meeting diagnostic criteria. Hp manual analysis has other issues, such as the availability of expert pathologists in healthcare institutions, especially in underdeveloped countries (Sophie Softley Pierce, 2017). Furthermore, Hp slide manual analysis is a time-consuming and cumbersome task and hence prone to human errors (Evans, 2011). Therefore, to avoid these issues, digital pathology laboratories convert Hp breast tissue slides into digital images known as digital Hp images (or simply Hp images) by using scanners with various zooming factors.

With the invention of digital Hp images, researchers have developed machine learning (ML) based BrC detection (i.e., benign or malignant) and classification (i.e., up to eight subtypes of BrT) models, which can perform an efficient and reliable automatic BrC diagnosis using Hp image. Thus, ML-based BrC detection and classification can overcome the aforementioned issues of a microscopic manual analysis of breast Hp slides. Moreover, it can assist the pathologists/doctors and serve as the basis of the second opinion in BrC diagnosis at an early stage.

In medical science, two types of Hp image analysis are found to diagnose benign or malignant BrT types, for instance, image-level and patient-level analyses (Spanhol, Oliveira, Petitjean, & Heutte, 2016a). The commonly adopted analysis is the image-level analysis where ML-based model performance is measure in terms of accuracy, sensitivity, FMeasure, etc., without using the number of patients, i.e., how accurately instances are correctly predicted for its intended class label like benign or malignant. However, the patient-level analysis is defined in terms of patient recognition rate (PRR) for benign and malignant cases by the BreakHis dataset host (Spanhol et al., 2016a), which is the sum of all patient scores divide by the total number of patients. Where, the patient score is the number of images correctly predicted (i.e., benign or malignant) for a patient upon the total number of images of a patient. Therefore, most of the studies used image-level analysis, and some of the studies used patient-level analysis for BrC detection.

Recent studies have developed BrC detection and classification models by exploiting Hp images (Araujo et al., 2017; Chang, Yu, Han, Chang, & Park, 2017; Samala et al., 2017; Spanhol, Oliveira, Cavalin, Petitjean, & Heutte, 2017; Zheng et al., 2017; Feng, Zhang, & Yi, 2018; Gandomkar, Brennan, & Mello-Thoms, 2018; Nahid, Mehrabi, & Kong, 2018). Two approaches are generally used to detect and classify BrC are traditional ML-based and deep learning (DL) based approaches. BrC detection and classification

models, which are based on traditional ML, often follow four main steps: image preprocessing, feature extraction, training, and evaluation of classification model.

For instance, Spanhol, Oliveira, Petitjean, and Heutte (2016b) developed and hosted a BrC Hp image dataset named BreakHis for researchers. The authors preprocessed an image by removing unwanted areas and saving it as an 8-bit image portable network graphics format and sized to $700 \times 460 \times 3$ with no compression. Image acquisition is conducted on four different magnifications using a region of interest (ROI) marked by an expert pathologist. In the preprocessing step, features were extracted through six types of commonly used textural descriptor methods, and a master feature vector (MFV) is created. Thereafter, four classifiers were trained and evaluated using MFV: k-nearest neighbor (kNN), quadratic linear analysis, support vector machine (SVM), and random forests (RF). The authors reported overall accuracy ranging from 80% to 85% and PRR is 83.8 ± 4.1 . Nonetheless, the success of traditional ML-based BrC detection models depends upon the discriminative feature (Mujtaba, Shuib, Raj, Majeed, & Al-Garadi, 2017; Nweke, Teh, Al-Garadi, & Alo, 2018; Ishtiaq et al., 2019) extraction key step (Domingos, 2012; Duda, Hart, & Stork, 2012). The limitations of hand-engineered features (HeFs) extraction are as follows.

1. The extraction of discriminative features requires domain knowledge and is a difficult and time-consuming task.
2. Rework is often needed when a similar type of dataset collected from different sites is utilized. Hence, robust feature extraction is not a trivial task.
3. Images, especially neighboring pixels, possess highly correlated information (Shen, Wu, & Suk, 2017). HeFs extraction methods may lose such kind of important information related to normal and abnormal breast tissues.

Therefore, DL-based approaches bypassed (by embedding the features extraction while model training) the HeFs extraction step involved in traditional ML-based approaches. Furthermore, DL-based approaches require minimal data preprocessing tasks (if ever) and identify relevant information in a self-taught manner (Shen et al., 2017). Thus, the fundamental goal of using DL-based approaches is to automate the feature extraction step during the model training process. DL-based BrC detection and classification approaches are of three types. The first type comprises DL-based models created and trained from scratch, which are also known as de-novo (Hadad, Bakalo, Ben-Ari, Hashoul, & Amit, 2017) models. The second type involves models created and retrained after fine-tuning the pre-trained models (e.g., AlexNet), which are called transfer learning (TL)-based models. Whereas third approach models were created by ensembling layers of first and second types of models i.e., de-novo and TL-based layers.

Several existing studies have exploited de-novo models (Araujo et al., 2017; Samala et al., 2017; Feng et al., 2018; Nahid et al., 2018). For instance, in Feng et al. (2018), image patches were fed into a DL-based model of stacked denoising auto-encoders with three hidden layers. The author trained the model using a graphics processing unit (GPU) for three hours and reported better accuracy where benign was 97.98 ± 0.69 and malignant was given as 88.37 ± 1.90 . However, the results were biased because the author used an exclusive small-sized dataset, i.e., 58 breast Hp images. In Araujo et al. (2017), the author used the Bioimaging Challenge 2015 Breast Histology dataset and performed many image preprocessing tasks, such as stain normalization, histogram stretching, division of images into patches with 50% overlap, intensity normalization, and image augmentation. Furthermore, a convolutional neural network (CNN) model was created and trained from scratch, deep convolution activation features (DeCAFs) were extracted, and classification was performed through SVM and softmax; results showed that the former classifier outperformed the latter. The model training time and resources were not mentioned.

Whereas reported accuracy was 77.8% for four classes and 83.3% for BrC detection. However, the classification accuracy needs to be improved for the method to be deployed in real-life scenarios.

The aforementioned studies revealed that de-novo model sizes are often small, and most of the models are created according to the volume of data because a large DL model size needs a large amount of data and is likely to overfit. However, the major advantage of using a de-novo model is that the customized robust layers can be created to improve the performance of a model for the specific type of data like medical images. Nevertheless, de-novo models require a large amount of labeled data to be trained properly and avoid overfitting issue, but the annotated medical images are rarely available in large quantities. Furthermore, model training with large data and model size will require considerable time and computational resources (e.g., high-capacity storage devices, random access memory, and GPU), which are very costly.

Thus, to avoid the limitation of DL-based de-novo models, many researchers adopted the TL-based DL approach (Chang et al., 2017; Spanhol et al., 2017; Zheng et al., 2017; Gandomkar et al., 2018), where, the pre-trained model is exploited followed by fine-tuning step. The fine-tuning step mostly involves the replacement of the last layer of a pre-trained model for the target number of classes. For instance, in Gandomkar et al. (2018), the author used the BreakHis dataset after removing the borderline patient. A TL-based hierarchical model was trained and evaluated followed by preprocessing tasks such as scaling, stain normalization, augmentation, and patch generation. The author achieved 95.70% accuracy for BrT classification using a single fold of data. However, a pre-trained 150-layer Residual neural network (ResNet) was used, and three models were created hierarchically. Hence, the author deployed model was large enough but complex and required plenty of resources and training time. Chang et al. (2017) proposed a TL-based

model by exploiting a pre-trained Google inception v3 model. The BreakHis dataset was used after applying preprocessing tasks, such as rescaling and various image augmentation techniques. Many cutoff values are tested to acquire reliable accuracy rates, and the cutoff value method is optimized for asymmetric misclassification costs. The findings showed accuracy rates of 83% and 89% for benign and malignant cases. Nonetheless, the Google inception v3 model contains a large number of layers (i.e., 48) and needs ample resources and time for training.

The aforementioned studies discovered that the TL-based models may become overfitted if the target data size is very small because pre-trained models were trained on a very large amount of data (Sert, Ertekin, & Halici, 2017; Shen et al., 2017; Gandomkar et al., 2018). Therefore, for retraining, such models cannot properly learn the new features from a few instances of target data. Oftentimes, pre-trained models are large (e.g., GoogLeNet has 152 layers.) and thus require a large amount of data and consume considerable computational resources and training time. Conversely, if the TL-based model size is small, such as AlexNet, and the data size is not too small, then it can be used to create a classification model by using limited resources (Spanhol et al., 2016a; Spanhol et al., 2017). Otherwise, the classification results may not be sufficiently accurate and reliable to be implemented in real-life applications.

Therefore, some researchers created an ensemble DL-based model by combining TL-based layers and few newly created layers trained from scratch (i.e., de-novo model layers) (Kumar, Bhadauria, Virmani, & Thakur, 2017; Rasti, Teshnehlal, & Phung, 2017; Sert et al., 2017; Wan, Cao, Chen, & Qin, 2017). For instance, Wan et al. (2017), exploited an exclusive dataset of 106 Hp images for BrC grading into low, intermediate, and high classes. The author performed nuclei segmentation by an enhanced hybrid active contour model. Thereafter, a CNN-based ensemble model was used to extract an integrated set of

pixel-level texture features and object-level architectural features. Finally, SVM was used in a cascaded fashion to combine two types of features and thus maximize classification performance. The ensemble model is capable of adopting advantages of both TL-based and de-novo models like custom layers can be created to learn better features from medical images compared to TL-based models, model size can be increased or decreased by adding or removing layers to get better features. Moreover, ensemble models require fewer images and can be trained in less time using fewer resources compared to de-novo models. Thus, it can be concluded that TL-based and ensemble models are suitable for medical image BrC detection and classification compared to de-novo models.

Generally, it has been observed in the aforementioned DL-based models of BrC detection and classification that the results are compromised due to a higher number of false negative and false positive predictions, also known as wrong predictions or simply misclassification. Whereas, misclassification using BrC Hp images for eight subtypes of BrT maybe because of three reasons. First, there is a high correlation among the features of many subtypes of BrT Hp images (Han, Wei, et al., 2017). Which may create complexity (i.e., low interclass similarity and low intraclass dissimilarity) for the classifier to differentiate among multiple subtypes of BrT. Therefore, the misclassification rate can be higher (Han, Wei, et al., 2017). Second, a large number of features are extracted through DL-based models. Such a large number of features/dimensions can easily distract the training process of a classifier that can increase the misclassification rate (Fan & Fan, 2008). Third, the DL-based models are normally trained using augmented images along with original images. Whereas the quantity of augmented images is huge than the original images, therefore model maybe got better training for augmented images instead of original images (Simard, Steinkraus, & Platt, 2003). However, testing data contains only original images, thus it can be easily misclassified by the model which was largely trained on augmented images.

1.2 Research Motivation

Cancer-related mortality has drastically increased in recent years. As per WHO report (WHO, 2018), cancer is the leading cause of death, and approximately 8.8 million people have died globally in 2015 due to this disease. In addition, the number of new cancer cases is expected to increase by 70% in the next two decades. Among the various types of cancer, breast cancer is the most common among women and is the third leading cause of cancer-related deaths (1.7 million, 11.3%) (WHO, 2018). In addition, an early and precise diagnosis is important to improve the prognosis and increase the survival rate of patients with BrC by 50%.

Hp imaging is more commonly used for the detection and classification of BrC compared with other medical imaging modalities like mammogram (MG), magnetic resonance imaging (MRI), ultrasound (US), and computerized tomography (CT). Nonetheless, Hp image manual analysis has three major limitations (Gurcan et al., 2009). First, expert pathologists are rare in healthcare organizations in several developing countries. Second, the procedure is cumbersome and time-consuming for pathologists. Therefore, pathologists may experience fatigue and reduced attention during the image manual analysis. Finally, a reliable manual analysis is highly dependent on the professional experience, expertise, and domain knowledge of pathologists.

Thus, the aforementioned limitations may cause misdiagnosis/misclassification of Hp image manual analysis for BrC and may lead to an incorrect treatment plan. Hence, to address the above mentioned limitations of BrC Hp image manual analysis, ML-based diagnostic systems (i.e., detection and classification models) can be used for automatic analysis. Moreover, ML-based BrC detection and classification models can assist doctors to serve as a second opinion to analyze the Hp images efficiently and more accurately compared to manual analysis.

1.3 Statement of Problem

Several researchers classified breast cancer from Hp images (Brook et al., 2008; Zhang, 2011; Spanhol et al., 2016a; Xu, Luo, Wang, Gilmore, & Madabhushi, 2016; Araujo et al., 2017; Ehteshami, Veta, Johannes, & et al., 2017; Han, Wei, et al., 2017; Spanhol et al., 2017; Wan et al., 2017; Zheng et al., 2017; Bardou, Zhang, & Ahmad, 2018; Nahid & Kong, 2018; Nahid et al., 2018). However, there are four major limitations.

1. First, these studies mostly employed image-level BrC detection. However, patient-level BrC detection can pose different results.
2. Second, these studies mostly classified BrT up to four classes. However, there are eight subtypes of BrT which are inherently more complex to classify. Therefore, a higher misclassification rate is observed in the aforementioned studies.
3. Third, these studies mostly employed the accuracy metric to measure classification model performance. However, this metric can be biased (due to misclassification) in measuring the overall classification performance (Powers, 2011).
4. Finally, the majority of the existing DL-based models are trained on high computational resources for longer training time to classify BrC histopathological images. Therefore, computationally efficient classification models are needed.

Hence, to overcome the aforementioned limitations, efficient (i.e., consume less computational resources and training time) and reliable (i.e., reduce misclassification to show better and unbiased results even using complex dataset) BrC detection and classification models are needed, which can be trained on complex, publicly available standard datasets using low computational resources in less time and able to show better results due to reduced misclassification rate. In addition, more robust metrics need to be used to accurately measure the performance of BrC detection and classification models.

1.4 Research Aims and Objectives

The primary goal of this research is to provide efficient (i.e., consume less computational resources and training time) and reliable (i.e., reduce misclassification to show better and unbiased results even using complex dataset) DL-based breast cancer detection and classification models by using Hp images. Thus to accomplish these goals this research has the following research objectives (RO).

RO1: To investigate the existing DL-based models for breast cancer detection and classification, using Hp images for early diagnosis.

RO2: To develop an efficient (i.e., consumes less computational resources and training time) and reliable (i.e., reduces misclassification to show better and unbiased results even using complex dataset) DL-based model for BrC detection at patient-level using Hp images.

RO3: To develop an efficient (i.e., consumes less computational resources and training time) and reliable (i.e., reduces misclassification to show better and unbiased results even using complex dataset) DL-based model for BrT classification (up to eight classes) using Hp images.

RO4: To evaluate the performance of proposed BrC detection and classification models by comparing their performances with existing state-of-the-art (SoA) BrC detection and classification models.

1.5 Research Questions

The research questions (RQ) belong to each research objective are given as follows:

RO1: To investigate the existing DL-based models for breast cancer detection and classification, using Hp images for early diagnosis.

RQ1: What are the existing DL-based models for breast cancer detection and classification, using Hp images for early diagnosis?

RQ2: What are the common medical imaging modalities used for BrC detection and classification?

RO2: To develop an efficient (i.e., consumes less computational resources and training time) and reliable (i.e., reduces misclassification to show better and unbiased results even using complex dataset) DL-based model for BrC detection at patient-level using Hp images.

RQ3: How to develop an efficient (i.e., consumes less computational resources and training time) and reliable (i.e., reduces misclassification to show better and unbiased results even using complex dataset) DL-based model for BrC detection at patient-level using Hp images?

RO3: To develop an efficient (i.e., consumes less computational resources and training time) and reliable (i.e., reduces misclassification to show better and unbiased results even using complex dataset) DL-based model for BrT classification (up to eight classes) using Hp images.

RQ4: How to develop an efficient (i.e., consumes less computational resources and training time) and reliable (i.e., reduces misclassification to show better and unbiased results even using complex dataset) DL-based model for BrT classification (up to eight classes) using Hp images?

RO4: To evaluate the performance of proposed BrC detection and classification models by comparing their performances with existing state-of-the-art BrC detection and classification models.

RQ5: How to evaluate the performance of proposed BrC detection and classification models?

1.6 Research Methodology

The research is conducted to perform BrC detection and classification using a DL-based approach for Hp images. The general research design implemented in this research work is shown in Figure 1.1.

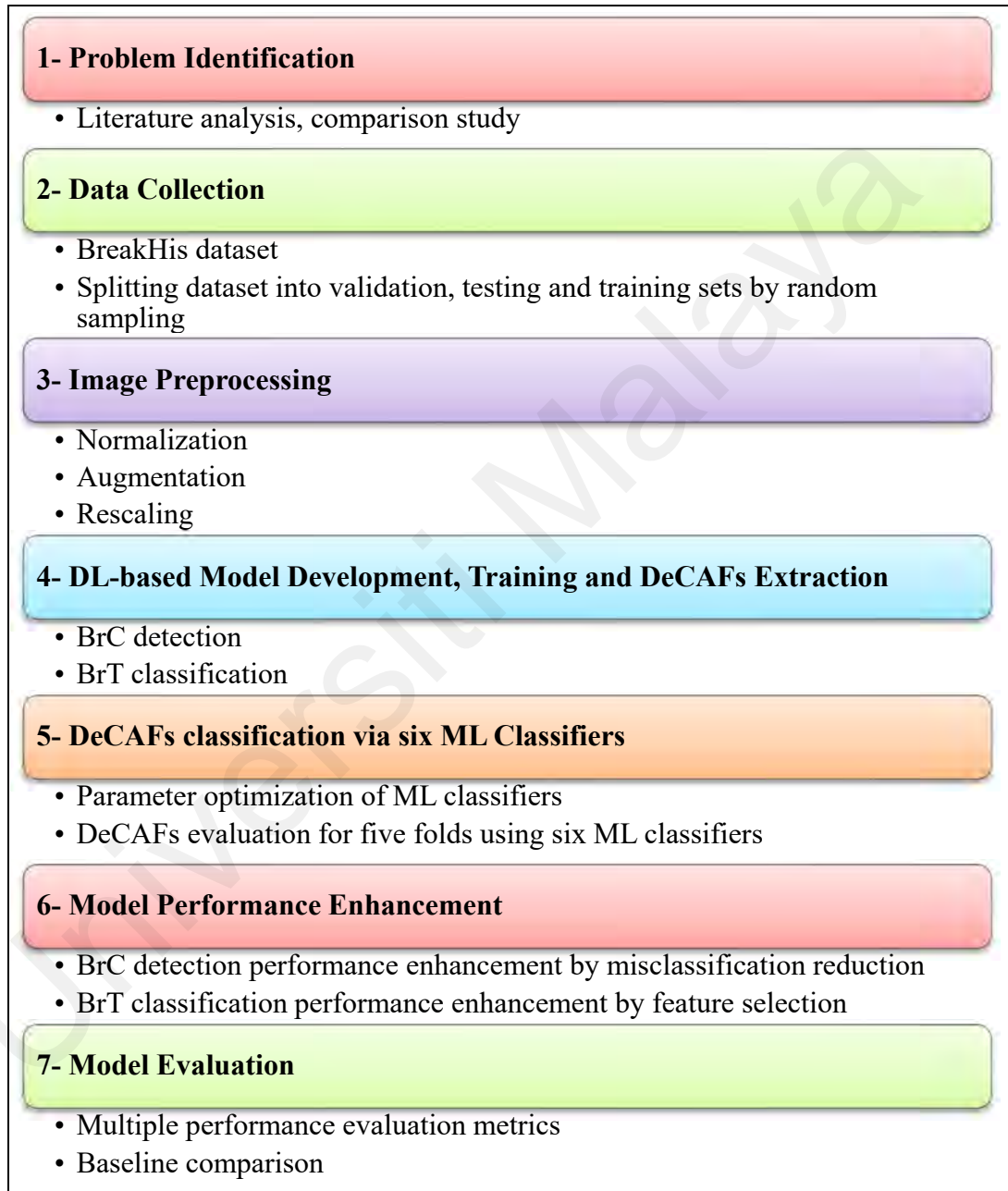


Figure 1.1: Steps involved in research methodology

The presented research design consisting of seven steps, like problem identification, data collection, image preprocessing, DL-based model development, training, and DeCAFs/features extraction, best ML classifier selection, model performance

enhancement, and model evaluation. These research design steps are briefly explained as follows:

1- Problem Identification

This step demonstrated the research problem identified through a literature review on DL-based breast cancer detection and classification through various medical imaging modalities. This review involves the analysis of five aspects related to BrC detection and classification namely medical imaging modalities, medical imaging datasets, medical image preprocessing techniques, BrC deep neural network model types currently implemented and performance evaluation metrics used to assess the performance of the model. The details of this section are discussed in Chapter 2.

2- Data Collection

In this step, the publicly available standard BreakHis corpus (Spanhol et al., 2016b) is collected for DL-based BrC detection and classification. BreakHis consist of a larger number of Hp images compared to other public datasets. Here, Hp images are distributed into eight subtypes of breast tumors in a patient-wise fashion. It is a multifaceted standard dataset, which enabled us to compare directly the results of proposed BrC detection and classification models with existing SoA baseline models. Due to these inherent characteristics of the BreakHis dataset, the results of the proposed DL-based model become more applicable and reliable compared to any small-sized public or exclusive dataset. The details of this section are discussed in Chapter 3 (Sections 3.2.1.1 and 3.2.2.1).

3- Image Preprocessing

In this step, the necessary medical image preprocessing tasks like Hp image stain normalization, image augmentation, and image rescaling are performed before initiating the training process of proposed DL-based models. However, before performing any

preprocessing task, the BreakHis dataset is split into training, validation, and testing sets. The image augmentation is applied to training sets only to increase the number of instances by applying basic image processing techniques. This is due to a large number of training instances that are required to avoid biased training and overfitting issues of the DL-based model. In addition, all images are rescaled to fit into the first layer of the proposed DL-based models for BrC detection and BrT classification. The details of this section are discussed in Chapter 3 (Sections 3.2.1.2 and 3.2.2.2).

4- DL-based Model Development, Training, and DeCAFs Extraction

In this step, proposed DL-based models are created and trained multiple times by adjusting random hyperparameters until minimum validation loss or maximum validation accuracy is not observed. Thus, the models are developed and trained efficiently to consume fewer resources (like a normal desktop computer instead of GPU) and training time. Finally, the trained DL-based models are used to extract features/DeCAFs from Hp images for BrC detection (i.e., benign or malignant) and classification for eight subtypes of BrT. Whereas, further analyses are made by using five folds of extracted DeCAFs. The details of this section are discussed in Chapter 3 (Sections 3.2.1.3 and 3.2.2.3).

5- Best Machine Learning Classifier Selection

In this step, the extracted DeCAFs are evaluated for five folds through six ML classifiers (i.e., softmax, k nearest neighbor (kNN), linear discriminant analysis (LDA), naive Bayes (NB), decision tree (DT), and support vector machine (SVM)) to ensure that the DL-based classification models were trained properly to extract discriminative features from Hp images. Moreover, this step helps to evaluate the performance of six ML classifiers by using six performance evaluation metrics (PEMs) namely accuracy (Ac), sensitivity (Sn), precision (Pr), FMeasure (Fm), patient recognition rate (PRR), and area under the curve (AUC). The details of this section are discussed in Chapter 3 (Sections 3.2.1.4 and 3.2.2.4).

6- Model Performance Enhancement

In this step, the performance of six ML classifiers is improved. For BrC detection, three misclassification reduction algorithms are implemented to enhance six ML classifiers' performance by using multiple Hp images of each patient provided in the BreakHis dataset. Whereas, for BrT classification, a feature selection algorithm is implemented by using two feature reduction schemes (i.e., Information Gain (IG) and Principal Component Analysis (PCA)) to select a minimum number of features to enhance overall classification performance. The details of this section are discussed in Chapter 4 (Sections 4.3 to 4.5).

7- Model Evaluation

This step evaluates the performance of trained BrC detection and classification models by using multiple PEMs like Ac, Sn, Pr, Fm, PRR, and AUC. Where PRR is used for patient-level analysis and rest are used for image-level analysis. This step also compares the performance of the proposed models with existing SoA baseline studies. Finally, it identifies the best model for BrC detection and classification using Hp images. The details of this step are discussed in Chapter 5.

1.7 Research Scope

This research is conducted on the basis of a certain definition to maximize the specialty of work for a certain area. The limitations are discussed as follows:

1. The BreakHis dataset images are actually patches taken from BrC whole slide images (WSI) marked by an expert pathologist. Thus, the proposed model is trained on image patches which possess dependency on expert pathologists. However, the generalizability of the proposed models should be applied to WSI images to minimize the dependency of expert pathologists.

2. The proposed models are capable to classify Hp images of BrC only. However, Hp images can be of many diseases like lung, liver, and bladder cancers.
3. The proposed BrC models are able to detect and classify Hp images only. However, there are many other medical image modalities used for BrC diagnosis.
4. The proposed models are based on CNN model, however, there are many other types of DNN model can be used for BrC detection and classification.

1.8 Research Contribution

The contributions of this research in current literature are as follows.

- **Literature Analysis**

This aims to identify the weaknesses of existing models related to BrC detection and classification. Moreover, deep neural network (DNN) based (i.e., DL-based) BrC detection and classification models that can classify various types of medical imaging modalities namely, MG, MRI, US, CT, and Hp digital images are comprehensively and systematically reviewed. This extensive review is based on the following five aspects namely

1. Medical imaging modalities used for BrC detection and classification.
2. Medical image datasets were used in the development of DL-based detection and classification models.
3. Preprocessing techniques adopted to improve medical image quality.
4. DL-based (i.e., DNN-based) model types currently applied to BrC detection and classification using various medical imaging modalities.
5. Evaluation metrics used to assess the performance of DL-based BrC detection and classification models.

Moreover, the current challenges and future directions related to BrC detection and classification are also discussed.

- **Breast Cancer Detection Model**

1. This research developed a DL-based ensemble model (with larger input image size and unfreezed layers) for BrC detection (i.e., benign or malignant) at the patient-level for Hp images using less computational resources (i.e., normal desktop computer instead of GPU) in less training time. In addition, the proposed model is used to extract discriminative features used for BrC detection.
2. Three misclassification reduction (McR) algorithms are developed to improve the performance of the BrC detection model at the patient-level.

1. **McR image-wise (McRI) Algorithm:** Reduces wrong predictions using many augmented images of the single original image and computes image-wise confidence of each patient.
2. **McR patient-wise (McRP) Algorithm:** Reduces wrong predictions using multiple augmented images of many original images of a patient and computes patient-wise confidence using all images of each patient.
3. **McR confidence-wise (McRC) Algorithm:** Reduces wrong predictions using the average of image-wise confidence and patient wise confidence computed in McRI and McRP algorithms.

- **Breast Tumor Classification Model**

1. This research developed a DL-based hierarchical BrT classification model to solve the multiclassification (i.e., eight classes) problem for breast Hp images using less computational resources (i.e., normal desktop computer instead of GPU) in less training time. Moreover, the proposed model is used to extract discriminative features for BrT classification.
2. To reduce the misclassification of the BrT classification model, a feature selection algorithm is implemented using two feature reduction schemes namely IG and

PCA to select minimum features subset to enhance the overall performance of the BrT classification model.

The proposed BrC detection and classification models have been published in reputable ISI-indexed journals and conferences (refer to page number 190 for the overall list of publications).

ISI-indexed Journal Publications

- **Paper 1 (Literature Review)** Murtaza, G., Shuib, L., Abdul Wahab, A.W. et al. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artif Intell Rev* 53, 1655–1720 (2020). <https://doi.org/10.1007/s10462-019-09716-5>. (**Published**)
- **Paper 2 (BrC Detection):** Murtaza, G., Shuib, L., Wahab, A.W.A. et al. Ensembled deep convolution neural network-based breast cancer classification with misclassification reduction algorithms. *Multimed Tools Appl* 79, 18447–18479 (2020). <https://doi.org/10.1007/s11042-020-08692-1>. (**Published**)
- **Paper 3 (BrT Classification):** Murtaza, G., Shuib, L., Mujtaba, G. et al. Breast Cancer Multi-classification through Deep Neural Network and Hierarchical Classification Approach. *Multimed Tools Appl* 79, 15481–15511 (2020). <https://doi.org/10.1007/s11042-019-7525-4>. (**Published**)

1.9 Research Significance

The significance of this research work is perceived in two domains, namely significance to doctors/pathologists and significance to researchers for BrC detection and classification models' development.

1.9.1 Significance to Doctors and Pathologists

As discussed in Section 1.1, Hp images are used with more confidence for the detection and classification of BrC compared with other medical imaging modalities. However, the prognosis and treatment of BrC solely depend on the diagnostic analyses report made by pathologists. Traditionally, Hp images are manually analyzed by more than one expert pathologist to diagnose the proper subtype of BrC. Whereas, the analysis of two pathologists may vary due to their domain knowledge, expertise, workload, and working environment. Therefore, pathologists' manual analysis can produce 24% incorrect

classification (Vestjens et al., 2012). Moreover, to arrange expert pathologist, and to develop consensus among their analysis is an expensive and time-consuming task (Evans, 2011). Hence, it may cause delay for specific diagnosis followed by treatment plan. Thus, there is a need for DL-based BrC detection and classification models to diagnose BrC automatically. The DL-based BrC detection and classification models are a more accurate, cost-effective, and less time-consuming method for early diagnosis of breast cancer (Schneider & Yaffe, 2000; Sadaf, Crystal, Scaranelo, & Helbich, 2011). Moreover, BrC detection and classification models can assist doctors and serve as the second opinion to make the decision more confidently in less time for the prognosis and treatment of BrC at an early stage.

1.9.2 Research Significance for Researchers for BrC Detection and Classification Models Development

As mentioned in Section 1.1, the existing DL-based BrC detection and classification model requires a large number of labeled images to be trained from scratch (i.e., de-novo model), thus it consumes very high computational resources like GPU and needs longer training time. However, medical images are usually not available in large quantities. Thus, collecting a large number of labeled medical images is not an easy task. Due to these unavoidable limitations of the de-novo model, the TL-based or ensemble model is a better choice, because it requires less number of labeled images, consumes fewer resources (like a normal desktop computer), and needs less time to train the BrC detection and classification models. Thus, the proposed TL-based or ensemble DL-based models are used to extract the features/DeCAFs. The DeCAFs are further analyzed by using ML classifiers to improve the BrC detection and classification results for breast biopsy Hp images. Moreover, robust feature extraction and reduction algorithms are developed to enhance the performance of classifiers for proposed BrC detection and classification

models. Finally, the results are justified by comparing them with existing baseline models.

1.10 Thesis Overview

The rest of the structure of this thesis work is organized as follows:

Chapter 2: This chapter demonstrates the literature analysis conducted for BrC diagnosis. This review focuses on breast cancer detection and classification by using medical imaging multimodalities through SoA DNN approaches. It is anticipated to maximize the procedural decision analysis in five aspects, such as types of medical imaging modalities, datasets, and their categories, preprocessing techniques, types of DNNs, and PEMs used for breast cancer detection and classification. In addition, this study provided quantitative, qualitative, and critical analyses of the five aspects. This review showed that mammograms and Hp images were mostly used to classify breast cancer. Moreover, about 55% of the selected studies used public datasets, and the remaining used exclusive datasets. Several studies employed augmentation, scaling, and image normalization preprocessing techniques to minimize inconsistencies in breast cancer images. Several types of machine learning BrC detection/classification models were implemented and are categorized into two main types like traditional ML-based models and Artificial Neural Network (ANN) based models. In traditional ML BrC detection/classification models the most widely used classifiers are kNN, LDA, NB, DT and SVM. Whereas, in ANN the shallow and DNN were employed to classify breast cancer using medical images. The convolutional neural network is utilized frequently to construct an effective breast cancer classification model. Some of the selected studies employed a pre-trained network or developed new deep neural networks to classify breast cancer. However, the DL-based model required high computational resources and a large number of images to get the desired results. Most of the studies used the accuracy measure

to compare the results. Whereas, a fewer number of studies used AUC metrics followed by Sn, Pr, Fm, and PRR to evaluate the performance of the developed breast cancer detection and classification models. Finally, this review presented research challenges for problem identification.

Chapter 3: This chapter is mainly divided into two sections. Section one presents the overall research methodology employed to develop the BrC detection model. Whereas section two discusses the entire research methodology implemented for the BrT classification model.

Section one converses in detail the dataset used for experiments. It also explains the various Hp image preprocessing tasks like stain normalization, augmentation, rescaling, and splitting the dataset into training, validation, and testing sets. Afterward, it elaborates on the DL-based model development and the DeCAFs extraction process. In addition, it demonstrates the evaluation made for five folds of DeCAFs using six traditional ML classifiers through six PEMs. Finally, it explains the BrC detection model performance enhancement that is achieved by developing and implementing three misclassification reduction (McR) algorithms in a cascade manner.

Whereas, the second section elaborates on the detailed research methodology implemented to construct the BrT classification model to diagnose eight subtypes of BrT. Furthermore, it discusses the data collection, selection of images, and splitting of images into training, validation, and testing sets. Moreover, this section explains the adopted image preprocessing tasks like augmentation, selection, and rescaling. It also discusses the entire methodology used for the development and training of the DL-based hierarchical BrT classification model with DeCAFs extraction. The extracted DeCAFs are evaluated for five folds by using the aforementioned six traditional ML classifiers through three PEMs like Ac, Sn, and AUC. Furthermore, to enhance the performance of

traditional ML classifiers a feature selection algorithm is developed using feature reduction schemes namely IG and PCA. The proposed feature selection algorithm enhanced the BrT classification model performance.

Chapter 4: This chapter explains the development of the proposed model to show the entire contribution made in this research work. It explains the research contribution in terms of problem identification through literature review, development of proposed BrC detection, and BrT classification models. It also discusses the developed algorithms implemented to enhance the performance of proposed models through misclassification reduction.

Chapter 5: This chapter covers two main parts of this research such as experimental results and discussion with baseline comparison. The first part represents the mean results (using five folds of features) for four experimental setups for each BrC detection and BrT classification model. For, BrC detection Ac, Sp, Sn, Pr, Fm, and PRR are reported for four experimental setups. Whereas, for BrT classification Ac, Sn and AUC are shown for four experimental setups. However, the second part provides a detailed discussion about the advantages and limitations of existing models' and compares the proposed models' results with exiting SoA BrC detection and classification baseline models.

Chapter 6: This chapter concludes the thesis by reevaluating the research objective. The main contributions are summarized. It discusses the limitations of the research work and proposed future directions.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter presents a review of existing relevant literature on ML-based BrC detection and classification through medical imaging modalities. It covers the overall analysis of BrC detection and classification by discussing almost all major studies. This review can assist researchers in BrC detection and classification to gain a better, concise perspective of existing problems and future directions. In this regard, many studies were scrutinized from eight journal¹ repositories to achieve five goals: imaging modalities, datasets, preprocessing techniques, machine learning classification models applied, and PEMs used for BrC detection and classification.

This chapter begins with a brief introduction to breast cancer. Whereas description, purpose, and types of medical imaging modalities used to diagnose BrC are discussed in Section 2.2. Section 2.3 elaborates a thorough analysis of BrC medical image datasets that were utilized in various studies. Section 2.4 covers the preprocessing techniques applied to BrC medical imaging modalities. Section 2.5, focuses on medical image preprocessing tasks. Whereas, Section 2.6 presents a comprehensive review of machine learning based classification models used in BrC detection and classification. The analysis of various evaluation metrics is discussed in Section 2.7. Section 2.8 and 2.9 identifies the research limitations of existing work to find the research gap. Finally, Section 2.10 presents a summary of this chapter.

2.2 Breast Cancer

Among all the cancers, BrC is a fatal cause of death in women around the world (WHO, 2018). BrC is usually caused by breast tumors (BrTs). The BrTs are mainly of

¹ Web of Science, scopus, IEE Xplore, PubMed, MedLine, Science Direct, ACM Digital Library, and springerLink

two types benign and malignant. However, both BrTs also have further subtypes and each subtype needs to be diagnosed individually. The benign type of BrTs are known as non-cancerous, thus do not invade other parts of the body. Whereas malignant BrTs are cancerous and aggressive which can spread to other parts of the body and can cause death if not diagnosed properly at its early stage. Initially, the doctor performs a physical examination of a breast to analyze suspicious symptoms. If any abnormality is found then mammograms or other medical imaging modalities (i.e., MRI, US, and CT) are suggested as breast screening tests to detect the BrC. Moreover, if a further detailed analysis of BrT is required then doctors usually suggest a breast biopsy Hp test, which allows BrT definite analysis at the tissue level. Furthermore, each subtype of BrT has a different treatment plan thus needs to diagnose confidently.

2.3 Medical Imaging Modalities

The BrT classification is composed of five unique types of medical imaging modalities and their combinations known as multimodalities. The distribution of chosen studies among various modalities and several studies is shown in Table 2.1. For clarity, imaging modalities can be bifurcated into colored images and grayscale images. Table 2.1 indicates that most of the work had been performed in either breast Hp biopsy colored images or using breast X-ray grayscale images, also known as mammograms (MGs). Table 2.1 shows that most of the studies are based on MG imaging modality. The main reason for a large number of studies using MGs may be the availability of images. MGs imaging technology has been adopted for the last two decades. MG-based studies mostly explored the breast density grading or classification for two classes. Moreover, the second-highest number of articles was published on Hp images. In these studies, researchers usually classified BrC not only into two main BrT types (i.e., benign or malignant) but also into further subtypes of each benign and malignant BrT. However, the third-highest number of studies were published for US images. Fewer studies

compared with US images were found for MRI images. Moreover, very few studies used multimodalities for BrT classification. For instance, one study was found for each combination, such as MG with US and US with CT images. Unfortunately, none of the studies used only CT or positron emission tomography (PET). However, CT and PET have been used for BrT classification for many years and played a significant role (Ahn et al., 2013; Lebron, Greenspan, & Pandit-Taskar, 2015). CT and PET images may be used if evidence shows that BrC has spread or reoccurred outside the breast. The detailed distribution of study references, modality type, a brief description of each modality used, and the number of studies is shown in Table 2.1.

Table 2.1: Distribution of studies for various medical imaging modalities.

Medical Imaging Modalities	Brief Description	Studies
MG (Breast X-rays)	Mammograms are found in three forms, such as screen-film mammograms (SFMs), digital mammograms (DMs), and digital breast tomography (DBT). SFMs and DMs are 2D grayscale in nature, but DBT provides multiple frames of 2D grayscale images that appear like a black-and-white video.	(Arefan, Talebpour, Ahmadinejad, & Asl, 2015; Arevalo, González, Ramos-Pollán, Oliveira, & Lopez, 2015; Fonseca et al., 2015; Rouhi, Jafari, Kasaei, & Keshavarzian, 2015; Kim, Kim, & Ro, 2016; Leod & Verma, 2016; Bakkouri & Afdel, 2017; Carneiro, Nascimento, & Bradley, 2017; Dhungel, Carneiro, & Bradley, 2017; Duraisamy & Emperumal, 2017; Jaffar, 2017; Khan, 2017; Kumar, Bhadauria, et al., 2017; Qiu et al., 2017; Samala et al., 2017; Sert et al., 2017; Sun, Tseng, Zhang, & Qian, 2017; Zhang et al., 2017; Samala et al., 2018a)
US	US images are also known as Sonograms. The US images are used in three combinations: simple 2D grayscale US images, US images along with additional additive features of shear-wave elastography (SWE) color images, and US images along with Nakagami colored images.	(Nascimento et al., 2016; Zhang et al., 2016; Byra, Piotrkowska-Wroblewska, Dobruch-Sobczak, & Nowicki, 2017; Han, Kang, et al., 2017)

Medical Imaging Modalities	Brief Description	Studies
MRI	MRI is used with pre and post-contrast [Dynamic Contrast-enhanced (DCE) MRI] images to diagnose the BrC. Post-contrast images are colored images but are usually converted into grayscale to feed into ANN.	(Bevilacqua et al., 2016; Amit et al., 2017; Han, Wei, et al., 2017)
Hp Images	Hp Images are H&E stained colored images and subdivided into two categories: whole slide images and image patches extracted from WSI by an expert pathologist.	(Cao, Qin, Jing, Chen, & Wan, 2016; Spanhol et al., 2016a; Wu, Shi, Li, Suo, & Zhang, 2016; Xu et al., 2016; Abdullah-Al, Bin Ali, & Kong, 2017; Araujo et al., 2017; Bayramoglu, Kannala, & Heikkila, 2017; Bejnordi et al., 2017; Chang et al., 2017; Han, Wei, et al., 2017; Nahid & Kong, 2017; Nejad, Affendey, Latip, & Ishak, 2017; Spanhol et al., 2017; Wan et al., 2017; Zheng et al., 2017; Abdullah-Al, Mehrabi, & Kong, 2018; Bardou et al., 2018; Feng et al., 2018; Gandomkar et al., 2018; Nahid et al., 2018)
Multimodalities	Some studies used the combination of two modalities of grayscale images named multimodalities for BrT classification. These combinations are MG with MRI and US with CT.	US with CT (Cheng et al., 2016) MG with MRI (Hadad et al., 2017)

2.3.1 Mammogram

Mammograms (MGs), also known as low-dose breast X-ray images, enable radiologists to investigate breast tissues for anomalies. MGs have been studied for the last two decades and are usually suggested in early stages called MG screening, see Figure 2.1 (A). In MG analysis, a radiologist looks for the presence of mass (cyst or lump, Figure 2.1 (A)) and tiny deposits of calcium (specifically with an irregular shape) called micro-calcifications that appear like small white spots or flecks, see Figure 2.1 (B). However, due to imaging technology advancement, MGs fall into three categories, namely, screen-

film mammography (SFM), full-field digital mammograms (FFDM), and digital breast tomosynthesis (DBT).

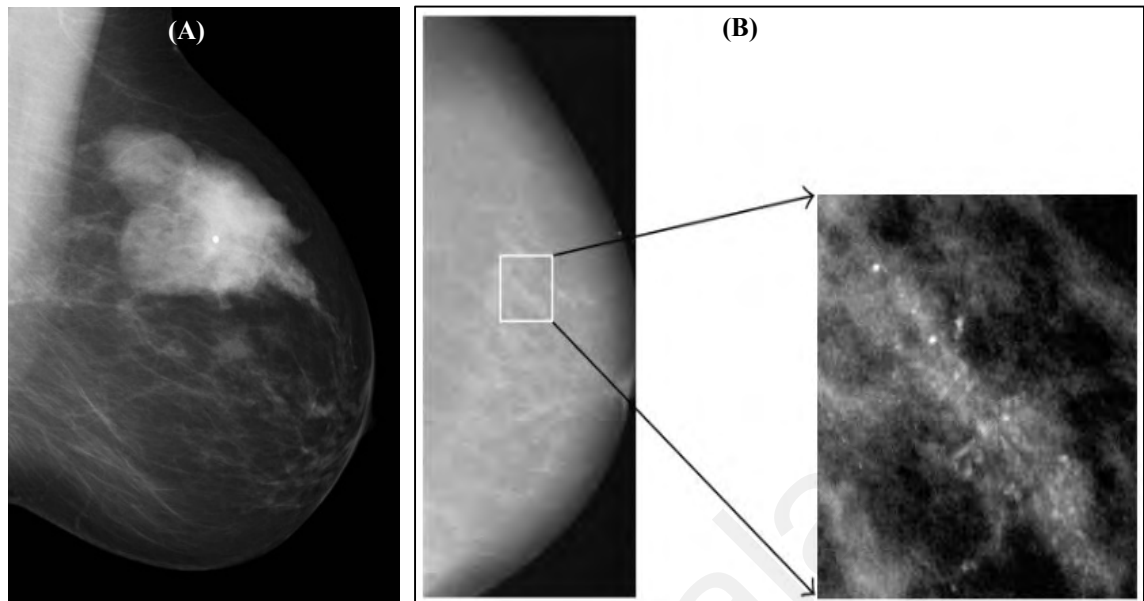


Figure 2.1: (A) Mammogram Screening: Masses with areas of varying density reflecting the presence of elements which are of fat and soft-tissue density (Jonathan J. James, 2016). (B) Left: A mammogram image view, Right: A clustered micro-calcifications in magnified view (Jing, Yang, & Nishikawa, 2012)

The traditional SFM images were used for BrT classification in many studies (Arevalo et al., 2015; Dhungel et al., 2017; Duraisamy & Emperumal, 2017; Jaffar, 2017; Khan, 2017). Dhungel et al. (2017) proposed an integrated model for the detection, segmentation, and classification of BrC into benign or malignant masses using SFM. Similarly, Duraisamy and Emperumal (2017) proposed a novel method by using the Chan-Vese level set method to segment SFM images before classifying BrC into normal, benign, or malignant cases.

The second category of MGs, FFDM (simply called digital MG or DM), is a well-adopted technology used by several researchers for BrT classification (Arefan et al., 2015; Carneiro, Nascimento, & Bradley, 2015; Leod & Verma, 2016; Carneiro et al., 2017; Hadad et al., 2017; Kumar, Bhadauria, et al., 2017; Qiu et al., 2017; Sert et al., 2017; Sun et al., 2017; Zhang et al., 2017). Carneiro et al. (2017) developed a holistic approach to classify unregistered DM and corresponding segmentation maps into normal, benign, or malignant breast lesions. Moreover, Qiu et al. (2017) proposed a model to classify benign

and malignant masses using DM without lesion segmentation, feature extraction, or feature selection.

In the third category, the most advanced MG technology is 3D MG, known as DBT. The DBT machine takes many views by moving over the breast and integrates images to look like a video. Nonetheless, due to the limited availability of datasets, few studies used DBT for BrT classification. Kim et al. (2016) implemented a BrT classification model to discover the latent bilateral feature representations of masses using the volume of interest in DBT. Similarly, Samala et al. (2018b) developed an efficient model by reducing the number of computations to perform BrC binary classification using all types of MGs, such as SFM, FFDM, and DBT.

Apart from DBT image classification, most research used either SFM or DM. The prime advantage of the popularity of SFM is that the images are directly printed on large sheets of film; in addition, it is a more cost-effective and frequently available imaging technology than FFDM and DBT. By contrast, FFDM images are easier to view, store, print, and manipulate using a desktop computer. Therefore, digital MG images can be viewed on a computer screen using many options, such as zooming, contrast enhancement, and highlighting the affected regions. Hence, due to the efficient processing of digital images, most of the recent public datasets utilized by researchers are digital MGs instead of SFM.

However, researchers started to use DBT because of many reasons; for instance, DBT may give a clear view of the breast from multiple angles to diagnose cancer with higher confidence and reduce the chance of follow-up testing as compared with FM or DM (Radiological Society of North America, 2018). Moreover, the availability of a large number of images per subject in video form provides better analysis opportunities to reduce the FNs in MGs.

Table 2.2, lists the detailed advantages and limitations of MGs. Regardless of MG diagnosis popularity, some cases may have dense tissues (bulky patient) or thick breast skin, such as in younger women, rendering the cancerous area almost invisible. Hence, macro-classification can be overlooked or misinterpreted during image analysis and may increase the FN rate. When image analysis is suspicious, the doctor may suggest some complementary tests, such as US, CT, PET, MRI, or biopsy, to acquire a detailed view of suspicious breast regions.

Table 2.2: Studies, imaging modality, strengths, weakness, and applications of various medical imaging modalities used in BrT classification

Imaging modality	Applications	Limitations
Mammogram (MG)	<ul style="list-style-type: none"> • Most studies employed SFM (Arevalo et al., 2015; Dhungel et al., 2017; Duraisamy & Emperumal, 2017; Jaffar, 2017; Khan, 2017; Samala et al., 2018b) or DM (Arefan et al., 2015; Carneiro et al., 2015; Fonseca et al., 2015; Leod & Verma, 2016; Carneiro et al., 2017; Hadad et al., 2017; Kumar, H.S, Virmani, & Thakur, 2017; Qiu et al., 2017; Sert et al., 2017; Sun et al., 2017; Zhang et al., 2017; Samala et al., 2018b) instead of DBT (Kim et al., 2016) for BrC diagnosis. • Relative to Hp, DM technology is an efficient, highly standardized, and cost-effective method to capture, store, and process images. • Needs less expertise and professional knowledge to diagnose and categorize an image compared with Hp. • A large variety of ML-based models are available to serve as a second opinion. • DBT shows a significantly higher rate of screen-detected cancer compared with DM screening (Hofvind et al., 2018). 	<ul style="list-style-type: none"> • Micro-calcifications are very small, isolated, with various sizes, shapes, dispersed, looks similar to their surroundings; thus, they cannot be identified in mammograms from high-frequency noise. • Several preprocessing tasks are needed before performing classification because of the presence of many factors, artifacts, and structure, such as film emulsion error, digitization artifacts, fibrous strands, borders of breast, and hypertrophied lobules, causes misinterpretation. • High breast density complicates the visualization of cancer in mammograms. However, the deeper breast is usually prone to cancer, and a radiologist can overlook or misinterpret the findings (Elmore et al., 2009). Hence, US or MRI can be preferred for a dense breast.

Imaging modality	Applications	Limitations
Magnetic Resonance Imaging (MRI)	<ul style="list-style-type: none"> • An MRI scan does not use potentially harmful ionizing radiation like CT scans and X-rays. • MRI images show more details of tissues (e.g., soft tissues of the breast) than CT scans (Tessa & Keith, 2018). • MRI can identify suspicious areas that can be further used for biopsy, known as MRI-guided biopsy. • DCE MR imaging uses contrast agents to show a clear and detailed view of affected breast regions. 	<ul style="list-style-type: none"> • MRI can still miss some tumors that a mammogram can detect. Thus, MRI is usually suggested in addition to a mammogram test. • An MRI is not generally recommended for women who are pregnant (MFMER, 2018). • May increase body temperature during long MRI (Tessa & Keith, 2018). • Contrast agents usually injected to enhance MRI images may create allergies or any complications, especially for kidney patients (MFMER, 2018).
Ultrasound (US)	<ul style="list-style-type: none"> • Very few articles are found using US images (Silva, Costa, Pereira, W.C, & Filho, 2015; Cheng et al., 2016; Nascimento et al., 2016; Zhang et al., 2016; Byra et al., 2017; Han, Kang, et al., 2017) for breast cancer diagnosis. • Images are taken in a real-time fashion. Hence, a breast lesion can be viewed from multiple angles, reducing the FN rate in diagnosis. • Widely available, extremely safe (noninvasive and no exposure to radiation) technology. Hence, preferred for a routine checkup among pregnant women. • It can detect invasive cancer areas that can be further used for biopsy, known as US-guided biopsy. Additional features, such as color-coded SWE images and Nakagami parametric images, can be captured along with traditional US images to identify breast lesion ROI. 	<ul style="list-style-type: none"> • Poor image quality is usually observed when a great amount of tissues is examined by ultrasound images (Radiological Society of North America, 2018; Ultrasound, 2018). • SWE images can cause misinterpretation if the probe is pressed harder (Barr, 2012; Youk, Gweon, & Son, 2017). • Solely single Nakagami parameters cannot distinguish between benign and malignant tissues (Tsui, Yeh, Chang, & Liao, 2008). The shadowing effect due to high attenuation makes the tumor contour unclear; thus, selecting the proper ROI and estimating tumor Nakagami parameters are difficult (Tsui et al., 2008).

Imaging modality	Applications	Limitations
Histopathology (Hp) Images	<ul style="list-style-type: none"> • Many studies employed Hp images (Cao et al., 2016; Spanhol et al., 2016a; Wu et al., 2016; Xu et al., 2016; Abdullah-Al et al., 2017; Araujo et al., 2017; Bayramoglu et al., 2017; Bejnordi et al., 2017; Chang et al., 2017; Han, Wei, et al., 2017; Nahid & Kong, 2017; Nejad et al., 2017; Spanhol et al., 2017; Wan et al., 2017; Zheng et al., 2017; Bardou et al., 2018; Feng et al., 2018; Gandomkar et al., 2018; Nahid & Kong, 2018; Nahid et al., 2018). • Hp images can be used in two forms, namely, whole slide images or ROI extracted from WSI. • Images are colored, can diagnose multiple types of cancers (Han, Wei, et al., 2017) instead of detecting malignancy only (Qiu et al., 2017) through grayscale imaging modalities. Ultimately, it leads to better prognosis and treatment at an early stage of BrC. • An in-depth study of BrC tissues is possible. Hp images enable us to provide more confident diagnosis results than any other imaging modalities. • Multiple ROI images can be created from WSI, which results in less probability to miss the cancer tissue detection, especially early-stage, and reduces the FN rate. • Images can be shared electronically to obtain an opinion from experts, especially for borderline cases, where two cancer types are hard to characterize. • It can be stored for a long time for future analysis or reference 	<ul style="list-style-type: none"> • A breast biopsy is an invasive method and thus has higher risks than other modalities. • Manual analysis of Hp images is time-intensive and requires high expertise; it depends on the professional experience and knowledge of a pathologist (Farahani, Parwani, & Pantanowitz, 2015). • Manual image inspections are tedious; thus, analysis reports are also affected by factors, such as fatigue and reduced pathologist attention (Spanhol et al., 2016b). • Hp image appearance variability causes misdiagnosis due to variability, different lab protocols, fixation, sample orientation in the block, human expertise in tissue preparation, microscopy maintenance, and color variation due to differences in staining procedures (McCann, Ozolek, Castro, Parvin, & Kovacevic, 2015). • For multiclass classification, traditional machine learning algorithms produce poor results because of high variability among images of the same cancer subtype (McCann et al., 2015). Hence, complex methods and high computational resources are required to improve computer-aided diagnosis.

2.3.2 Ultrasound

Ultrasound (US) images are also known as sonograms. Breast US (Figure 2.2(B)) is an imaging test that sends high-frequency sound waves into the breast and converts them into images without radiation involvement, unlike MGs and MRI. Apart from breast tests, the US image test can help diagnose anomalies, such as pain, swelling, and infections into the human body's internal organs, including a baby in the mothers' womb, brain, lung, heart, and hips. In addition, the US images can help perform breast needle biopsy (Section 2.2.4) for the intrinsic analysis of breast tissues.

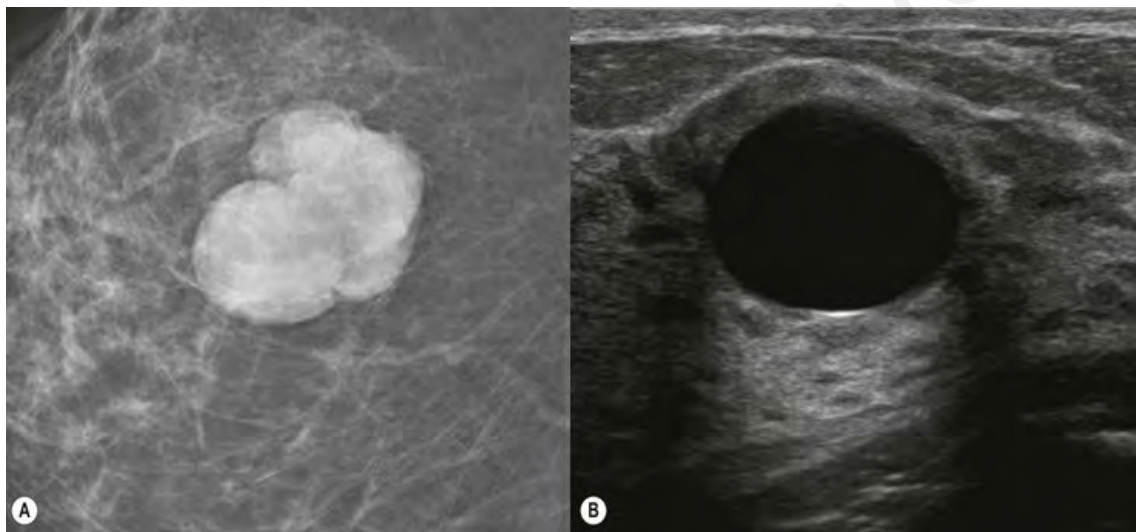


Figure 2.2: (A) Well-defined rounded mass mammogram. (B) The absence of internal echoes and the posterior enhancement of the ultrasound beam are diagnostic of a cyst or lump or mass (Jonathan J. James, 2016)

As per common clinical practices, the US image is not used like MG as its own for only breast screening purposes. Therefore, US may be the best approach to find abnormalities in MG or physical examination (such as benign, i.e., noninvasive cancer) in the form of a solid lump (mass) or fluid-filled regions (cysts) (Silva et al., 2015; Cheng et al., 2016; Nascimento et al., 2016; Han, Kang, et al., 2017). However, the US image cannot distinguish a cancerous mass from calcifications. Some researchers found that breast US is the better choice to diagnose BrC, especially when an MG is unable to highlight BrC lesions clearly, in young subjects with thick, fatty, or bulky breast skin. Detailed advantages and limitations of using US images are discussed in Table 2.2. Cheng

et al. (2016) deployed a model to extract features automatically from breast US images directly to perform accurate breast lesion classification as benign or malignant. Similarly, Nascimento et al. (2016) extracted handcrafted morphological features from breast US images and fed them into ANN for BrC binary classification (benign or malignant). Moreover, due to new developing imaging technologies, the US image has been equipped with more advanced features, such as US image with shear-wave elastography (SWE) (Figure 2.3) and the US image with Nakagami images.

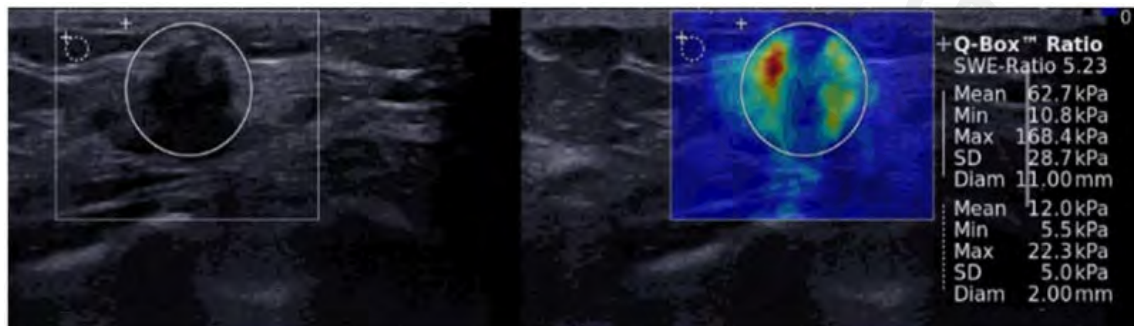


Figure 2.3: Left side US image (B-Mode). Shear-wave elastography image on the right side shows an irregular mass in red color, known as heterogeneous elasticity. The statistical parameters (e.g. Mean, Minimum, Maximum, etc.) of ROI (a large circle) are calculated (Youk et al., 2017)

Elastography is a recently developed US technique used to visualize and measure tissue elasticity. Elastographic images are based on tissue stiffness or hardness (such as in liver or breast) and are used to differentiate between benign and malignant lesions (Youk et al., 2017). It is a supportive parameter to the US image and adopted to quantify tumor grade by using a standardized color scheme. Hence, Zhang et al. (2016) used US SWE images to learn features directly by using a deep belief network to classify images (with higher accuracy) into benign or malignant BrC.

Moreover, US images are used with Nakagami images (Figure 2.4) for BrC analysis. US Nakagami parametric images are used with Nakagami distribution to model echo amplitude distribution to represent tissue characteristics (Tsui et al., 2016). These color-coded images can be captured along with traditional US images. The color-coded US images enable radiologists to quantify the stiffness or hardness of tissues. Hence, SWE

and Nakagami features play an additional role to enhance BrT classification diagnosis. However, very few studies used the new US technology. Byra et al. (2017) developed a model and extracted the scattering properties of breast tissues from parametric maps of Nakagami images to perform BrT classification by using a convolutional neural network (CNN). Data collection, particularly the difficulty in collecting a large number of medical images from any medical institution, maybe one of the reasons for the few publications.

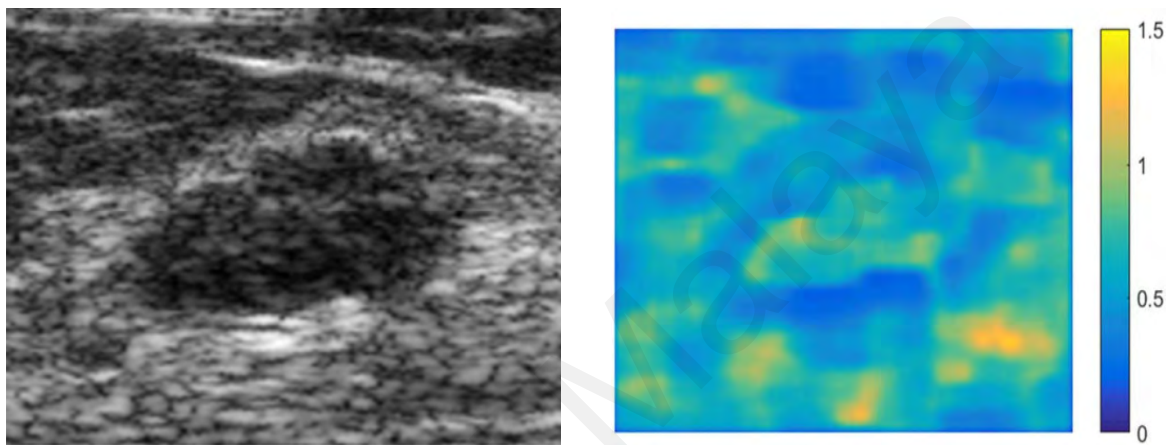


Figure 2.4: Left US image (B-mode) of a lesion reconstructed using the RF data on the right side corresponding Nakagami map (Byra et al., 2017)

2.3.3 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a diagnostic technology that uses magnetic fields and radio waves to capture a detailed image of the body's soft tissue, such as breast (Figure 2.5), liver, or lung, and bones. Therefore, breast MRI images can show more clear views of breast soft tissues than MGs, US, or CT images (Tessa & Keith, 2018).

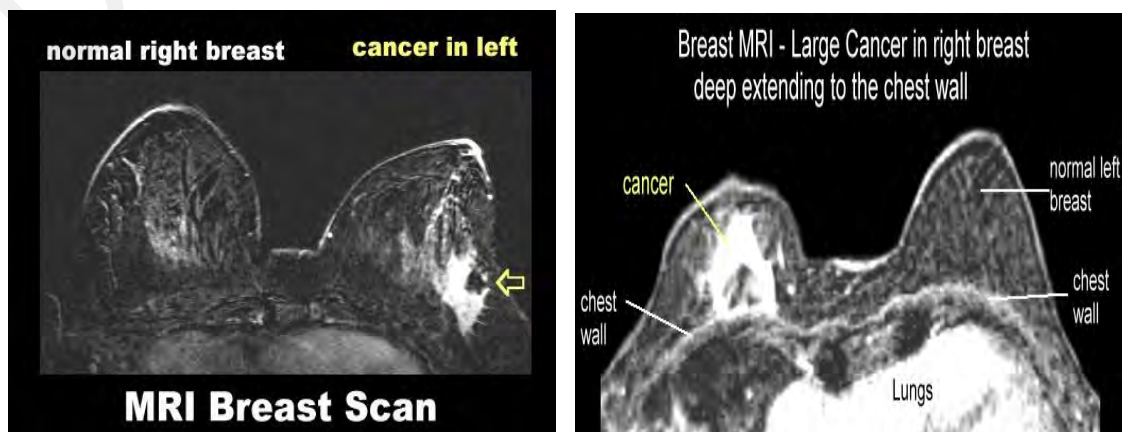


Figure 2.5: Samples of Breast MRI images (Breast Cancer Imaging, 2018)

Table 2.2, lists the advantages and limitations of MRI. Furthermore, MRI can identify suspicious areas that can be used for breast biopsy, known as MRI-guided biopsy. MRI machine captures many breast images of a single subject and combines as a detailed view. MRI is usually requested once cancer has been diagnosed and the doctor wants to obtain detailed information about the extent of the disease (MFMER, 2018).

However, very few studies used MRI to classify BrC (Bevilacqua et al., 2016; Amit et al., 2017; Rasti et al., 2017) possibly because of the unavailability of the public datasets. Bevilacqua et al. (2016) extracted features from MRI-segmented images and inputted them into an ANN for benign and malignant BrC identification. Analogously, Amit et al. (2017) extracted regions of interest (ROIs) from breast MRI images and inputted them into a CNN for multiclass classification.

To enhance image quality, a contrast agent is usually injected into the human body before the dynamic contrast-enhanced MRI (DCE-MRI). This procedure can produce colored parametric views along with contrast-enhanced grayscale images to provide detailed information about cancerous tissues (Moon, Cornfeld, & Weinreb, 2009). However, only one study benefitted from DCE-MRI for BrT classification. Rasti et al. (2017) employed a deep learning ensemble CNN model to classify breast tumors using segmented DCE-MRI images of an exclusive dataset.

2.3.4 Histopathology Images

In histopathology (Hp) biopsy imaging, tissue samples are collected from an abnormal region of the breast and fixed across glass microscope slides. These slides are stained by using hematoxylin-eosin (HE) and examined under a microscope by a pathologist for cancerous tissue diagnosis. Moreover, these stained slides are scanned and converted into digital colored images called WSIs, see Figure 2.6. Expert pathologists usually extract ROI patches from WSI with various zooming factors (Figure 2.6) to diagnose multiple subtypes of noninvasive cancer (benign) or invasive BrC (malignant) (Figure 2.7), which

is impossible by using grayscale images. Due to tissue level image analysis, apart from BrC diagnosis, biopsy imaging is a gold standard for many types of cancers, including liver, lung, and bladder cancer (Rubin, Strayer, & Rubin, 2008).

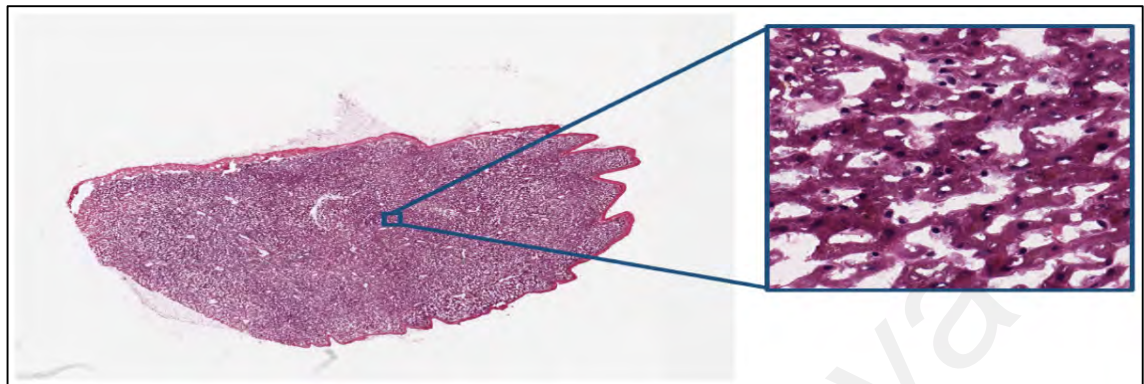


Figure 2.6: Histopathology WSI is shown on the left at low magnification and a cropped region is shown on the right at high magnification (Liu, Hernandez-Cabronero, Sanchez, Marcellin, & Bilgin, 2017)

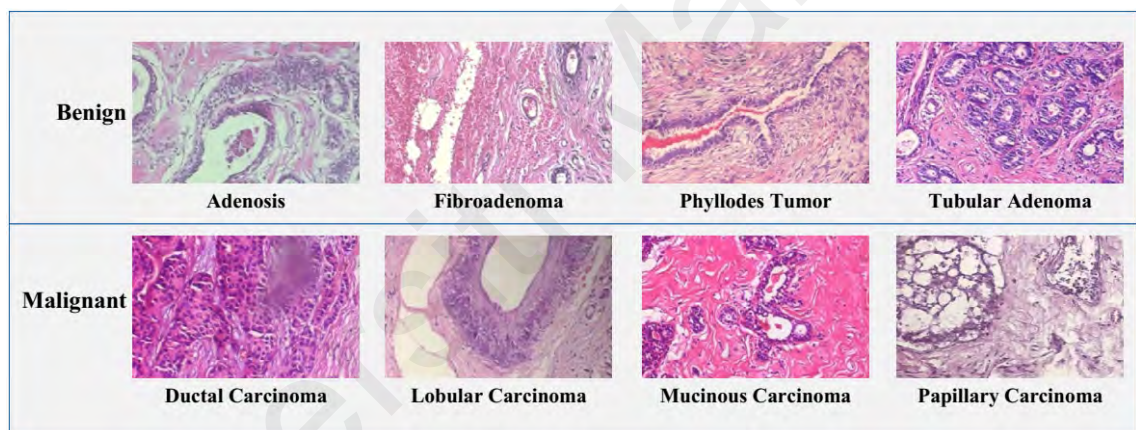


Figure 2.7: Histopathology image patches showing eight subtypes of breast cancer (Spanhol et al., 2016b)

Therefore, many researchers employed Hp images to classify BrC multiclass accurately (Cao et al., 2016; Spanhol et al., 2016a; Wu et al., 2016; Xu et al., 2016; Abdullah-Al et al., 2017; Araujo et al., 2017; Bayramoglu et al., 2017; Bejnordi et al., 2017; Chang et al., 2017; Han, Wei, et al., 2017; Nahid & Kong, 2017; Nejad et al., 2017; Spanhol et al., 2017; Wan et al., 2017; Zheng et al., 2017; Bardou et al., 2018; Feng et al., 2018; Gandomkar et al., 2018; Nahid & Kong, 2018; Nahid et al., 2018). For instance, Han, Wei, et al. (2017) used Hp images to classify BrC into eight types. Araujo et al. (2017) used Hp images to develop a model that classifies BrC into four subtypes. The

above-listed studies reported that the use of Hp images is beneficial for specific subtypes of benign or malignant BrC.

Automatic breast classification through Hp images has several advantages over MGs and other imaging modalities Table 2.2. For instance, Hp images enable the classification of BrC into many subtypes instead of binary classes and the monitoring of treatment effects, whereas WSI images allow the creation of a large number of ROI images, which are required to train DNN models. Images can be shared electronically to obtain the opinion of any far distant expert pathologist and thus form an accurate diagnosis. Although Hp images are authentic for automatic BrT classification, such images have some drawbacks in automatic image classification. For instance, a biopsy is an invasive method. In addition, a long time is needed to create digital images from collected biopsy samples, and high expertise is needed to distinguish between subtypes of BrC. Moreover, color variation is high because of the staining process, lab protocols, and scanner brightness in the development of Hp images, which complicate training a multiclass DNN model efficiently, especially when using borderline cases. Details of the imaging modalities used in previous studies are listed in Table 2.2.

2.3.5 Multimodalities

Apart from classifying BrC by using a single medical imaging modality, few researchers preferred to use at least two different imaging modalities, see Figure 2.8. Hadad et al. (2017) trained various classification models by using two modalities, namely, MGs and MRIs. This study performed a binary classification by identifying a breast image possessing either mass or non-mass regions. Moreover, images were classified as normal, benign, or malignant by (Khan, 2017) through multimodalities, such as MGs with US images. Many imaging modalities for BrT classifications are usually adopted when the size of the collected exclusive dataset is small. Moreover, a model trained on multi-site, multi-datasets using multi-modalities is highly robust to classify real-life images.

Eventually, the performance of the BrC classifier is unaffected by the images captured on various machines, different imaging protocols, and the environment for handling images. Hence, such types of models are more reliable to be implemented in real-life.

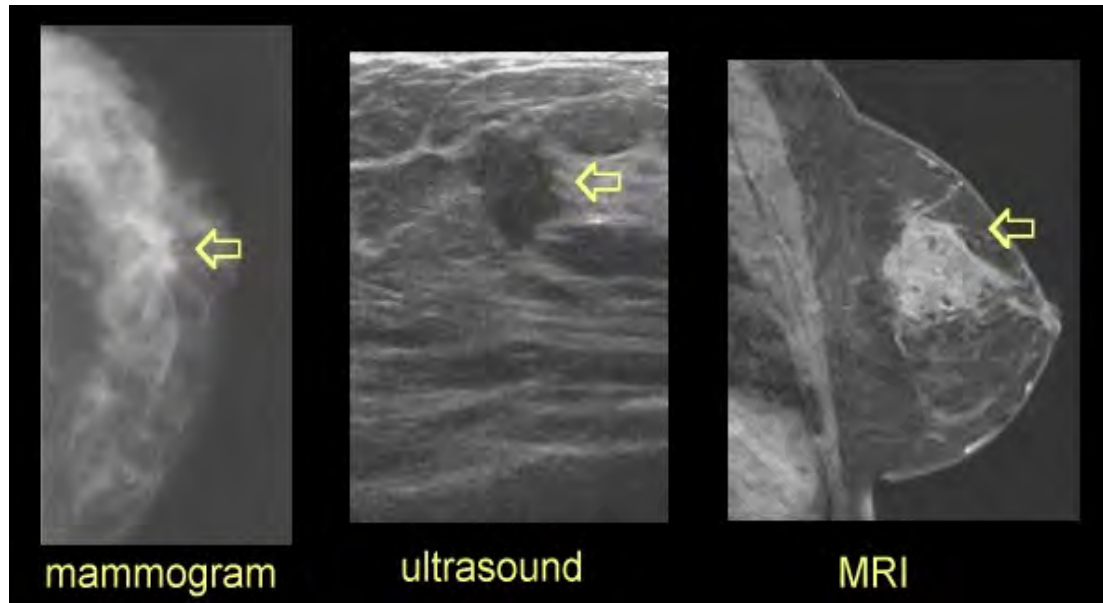


Figure 2.8: Multimodalities used for BrT classification. The left image is a mammogram showing a solid mass. The Center image is the US image showing stuff tissues as black. The right side image is MRI providing a clear view of breast mass (Breast Cancer Imaging, 2018)

2.4 Breast Cancer Classification Dataset Analysis

This section elaborates on a thorough analysis of public datasets that were utilized in various studies for BrT classification. Table 2.3 shows that eight public datasets were employed for BrT classification, namely, Breast Cancer Data Repository (BCDR), Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM), Digital Database for Screening Mammography (DDSM), INBreast, Mammographic Image Analysis Society (MIAS)/mini-MIAS, UCI Machine Learning Repository, Bioimaging Challenge 2015 Breast Histology (BCBH), and Breast Cancer Histopathological Image (BreakHis).

Table 2.3: Publically available datasets and corresponding URL

#	Dataset Name/Authors	Description	Link to Dataset
1	BCDR (Moura & López, 2013)	Breast cancer digital repository (BCDR) is a public dataset contains MGs and US images of 1734 patients. The images are classified and annotated by a specialized radiologist for researchers to develop computer-aided diagnostic systems.	https://bcdr.ceta-ciemat.es/information/about
2	CBIS-DDSM (Clark et al., 2013; Rebecca Sawyer Lee, 2016)	Curated Breast Imaging Subset of DDSM (CBIS-DDSM) dataset is an updated version of the Digital Database for Screening Mammography (DDSM). It possesses 2620 scanned MGs labeled as normal, benign, and malignant.	https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM
3	DDSM (Chris Rose, 2006)	The Digital Database for Screening Mammography (DDSM) dataset provided 2500 studies of MG images to facilitate the research community.	http://www.eng.usf.edu/cvprg/Mammography/Database.html
4	INBreast (Moreira et al., 2012)	The INBreast dataset contains 410 MGs images of 115 cases. Moreover, the images are categorized into various types of lesions like (masses, calcifications, asymmetries, and distortions).	http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database
5	MIAS (Suckling et al., 2015)	The Mammographic Image Analysis Society (MIAS) database of digital mammograms (v1.21) provides the 322 images, associated truth data.	https://www.repository.cam.ac.uk/handle/1810/250394
6	mini-MIAS (Suckling et al., 1994)	Mini-MIAS is originally a subset of the MIAS Database. The images were digitized and clipped/padded so that every image is 1024×1024 pixels.	http://peipa.essex.ac.uk/info/mias.html
7	UCI (Dua, 2017)	Dataset possesses 700 instances for benign and malignant classification.	https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

#	Dataset Name/Authors	Description	Link to Dataset
8	BCBH (Araujo et al., 2017)	Bioimaging Challenge 2015 Breast Histology (BCBH) dataset possesses four classes normal, benign, in situ carcinoma, and carcinoma. Overall 285 Hp images are provided and split into training and testing sets.	https://rdm.inesctec.pt/dataset/nis-2017-003
9	BreakHis (Spanhol et al., 2016b)	The Breast Cancer Histopathological Image Classification (BreakHis) provided 7909 Hp images of 81 patients. The dataset is divided into eight subtypes of breast tumors.	https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

Most of the articles utilized public datasets, usually based on MG, US, or Hp images. By contrast, a fewer number of studies employed exclusive datasets. In exclusive datasets, imaging modalities that are not publicly available similar to CT scan images were also used. Public datasets provided more annotated images than exclusive datasets. Hence, researchers can prepare BrT classification models by comparing the performance of developed classification models. Therefore, the model tested on public datasets is more reliable than the models tested on exclusive datasets. Regardless of database type (exclusive or public) at the abstract level, grayscale (e.g., MG, US, and MRI) or colored images (e.g., Hp images) are used for BrT classification. Moreover, most studies performed binary classification, and very few studies focused on multiclass problems for BrT classification. By contrast, some studies performed breast density grading (Cao et al., 2016; Wan et al., 2017) into three classes, namely, low, high, and medium grade. Detailed analysis of public datasets used for BrT classification is given in Table 2.4.

Table 2.4 shows the dataset name and type of imaging modality along with several images, number of patients, number of classes, and class labels for each dataset. This table also shows the reference of studies in which a particular dataset was used. The

investigation of the dataset reveals that most previous research used MG datasets and usually addressed either binary classification (benign or malignant) or tertiary classification (normal, benign, and malignant) of BrC. In this regard, most of the studies used MG datasets. Where MG-based studies(Carneiro et al., 2015; Rouhi et al., 2015; Leod & Verma, 2016; Bakkouri & Afdel, 2017; Carneiro et al., 2017; Jaffar, 2017; Kumar, H.S, et al., 2017; Samala et al., 2017; Sert et al., 2017) used DDSM datasets and studies(Arefan et al., 2015; Rouhi et al., 2015; Duraisamy & Emperumal, 2017; Jaffar, 2017; Khan, 2017; Kumar, Kumar, & Shao, 2017) used MIAS datasets.

Moreover, the MGs of both INBreast and BCDR datasets utilized by studies (Arevalo et al., 2015; Bakkouri & Afdel, 2017; Carneiro et al., 2017; Dhungel et al., 2017; Duraisamy & Emperumal, 2017; Khan, 2017; Kumar, Kumar, et al., 2017). Whereas, study (Kumar, Kumar, et al., 2017) classified MGs of CBIS-DDSM datasets. However, multimodality (US and MG)-based BCDR-F03 datasets were used by studies (Arevalo et al., 2015; Duraisamy & Emperumal, 2017) for BrT classification.

Table 2.4: Detailed analysis of public datasets used in breast cancer classification.

Imaging Modality	Dataset Name	No. of Images	No. of Patients	No. of Classes	Class Labels	Study References
Mammograms	BCDR	1734	---	3	Normal, Benign, Malignant	(Khan, 2017)
	BCDR-F03	736	344	2 or 10	(Benign, Malignant) or (Normal, Benign-calcification, Malignant-calcification, Benign-circumscribed masses, Malignant-circumscribed masses, Speculated masses, Ill-defined masses, Benign-architectural distortion, Malignant-	(Arevalo et al., 2015; Duraisamy & Emperumal, 2017)

Imaging Modality	Dataset Name	No. of Images	No. of Patients	No. of Classes	Class Labels	Study References
					architectural distortion, Asymmetry)	
	CBIS-DDSM	4067	---	2	Benign, Malignant	(Kumar, Kumar, et al., 2017)
	DDSM	10480	2620	2	Benign, Malignant	(Carneiro et al., 2015; Rouhi et al., 2015; Leod & Verma, 2016; Bakkouri & Afdel, 2017; Carneiro et al., 2017; Jaffar, 2017; Kumar, H.S, et al., 2017; Samala et al., 2017; Sert et al., 2017)
	INBreast	419	115	2 or 3	(Benign, Malignant) or (Normal, Benign, Malignant)	(Carneiro et al., 2017; Dhungel et al., 2017; Kumar, Kumar, et al., 2017)
	MIAS/Mini-MIAS	322	161	2	Benign, Malignant	(Arefan et al., 2015; Rouhi et al., 2015; Duraisamy & Emperumal, 2017; Jaffar, 2017; Khan, 2017; Kumar, Kumar, et al., 2017)
Mammograms and Ultrasound Images	BCDR	3703	1010	2	Benign, Malignant	(Bakkouri & Afdel, 2017)

Imaging Modality	Dataset Name	No. of Images	No. of Patients	No. of Classes	Class Labels	Study References
Histopathology Images	BCBH	269	---	4	Normal, Benign, In situ carcinoma, Carcinoma	(Araujo et al., 2017)
	BreakHis	7909	82	2 or 8	Four Benign Tumors (Adenosis, Fibroadenoma, Phyllodes tumor, Tubular adenoma), Four Malignant Tumours (Ductal carcinoma, Lobular carcinoma, Mucinous carcinoma, and Papillary carcinoma).	(Spanhol et al., 2016a; Abdullah-Al et al., 2017; Bayramoglu et al., 2017; Han, Wei, et al., 2017; Nahid & Kong, 2017; Nejad et al., 2017; Spanhol et al., 2017; Bardou et al., 2018; Feng et al., 2018; Gandomkar et al., 2018; Nahid & Kong, 2018; Nahid et al., 2018)

Apart from MG-based binary or ternary classification, Hp images played a prominent role to solve multiclass (up to eight subtypes) problems for BrT classification. In this respect, many studies performed classification by using the BreakHis dataset, as shown in Table 2.4. Unfortunately, most studies performed binary classification, and very few obtained better results to solve multiclass problems. Moreover, only one study used Bio-Imaging Challenge 2015 Breast Histology dataset and tackled the multiclass BrC issue. As per our review, the most widely used and authentic dataset in MG, US, and Hp imaging modalities are DDSM, BCDR, and BreakHis, respectively, because these datasets contain a large number of images of many patients, which are required to train DNN classification models with confidence. Unlikely, no publicly available datasets have been employed for CT, MRI, PET modalities. Hence, the unavailability of online datasets might be a reason for publically available datasets that may contain an insufficient number of images for training a DNN-based BrT classification model.

2.5 Medical Image Preprocessing

This section covers the preprocessing techniques adopted for medical image multimodalities in BrT classification. In general, BrC image preprocessing tasks involve augmentation, ROI extraction, scaling, image normalization, and enhancement to remove artifacts or cropping, stain normalization, feature reduction, and image registration. However, the use of raw images (without preprocessing) usually distracts the classification model and may lead to a high misclassification rate. The distribution of studies among preprocessing methods and their advantages are summarized in Table 2.5.

Table 2.5: Distribution of studies among preprocessing methods and their advantages

Preprocessing Method	Methodology	Advantages	References
Augmentation	Geometric Transform like rotation, flip	To avoid the DNN model overfitting issue. To overcome the class imbalance training problem. The network can learn lesions from many angles like a pathologist usually does in real-life for better analysis of Hp images.	(Arevalo et al., 2015; Bevilacqua et al., 2016; Cheng et al., 2016; Kim et al., 2016; Spanhol et al., 2016a; Xu et al., 2016; Amit et al., 2017; Araujo et al., 2017; Bakkouri & Afdel, 2017; Bayramoglu et al., 2017; Bejnordi et al., 2017; Byra et al., 2017; Carneiro et al., 2017; Chang et al., 2017; Dhungel et al., 2017; Duraisamy & Emperumal, 2017; Hadad et al., 2017; Han, Kang, et al., 2017; Jaffar, 2017; Kumar, Kumar, et al., 2017; Nejad et al., 2017; Rasti et al., 2017; Samala et al., 2017; Sert et al., 2017; Spanhol et al., 2017; Zhang et al., 2017; Zheng et
	Add noise/ Distortion (Gaussian noise, Barrel or Pin Cushion transforms)	Enables DNN to be trained robustly. Therefore, it can predict with higher accuracy even if images are noisy, as found in real-life. Hence there will be improved class label prediction for noisy images. DNN requires the least preprocessing steps at the time of prediction.	
	Patch creation Methods (Patches with 50% overlapping, no overlapping or randomly selected patches)	Many images can be generated from the original images. Moreover, it can preserve the image aspect ratio, architecture or shape of the lesion, and subjective information. Hence, it increases the performance of	

Preprocessing Method	Methodology	Advantages	References
		<p>the classifier and reduces the chance of false negatives.</p> <p>One can avoid artificial images generated by geometric transform or noise addition methods.</p> <p>No need to rescale images before inputting them to ANN. Hence, it may reduce the chance of information loss due to rescaling.</p>	al., 2017; Bardou et al., 2018; Feng et al., 2018; Gandomkar et al., 2018; Samala et al., 2018b; Kumar et al., 2020)
	Synthetic Minority Over-sampling technique (SMOTE)	To increase the number of samples (vectors) to the minority class, in order to handle the class imbalance problem before DNN training.	
ROI Extraction	Methods used like region growing, Nuclei Segmentation, Otsu Method, Markov Random Model	<p>Enables to increase the number of positive and negative image samples.</p> <p>Help the DNN model to learn better representation related to abnormal and abnormal regions and reduces the chances of overfitting.</p> <p>Saves computation time and resources.</p>	(Arefan et al., 2015; Arevalo et al., 2015; Fonseca et al., 2015; Rouhi et al., 2015; Bevilacqua et al., 2016; Cao et al., 2016; Cheng et al., 2016; Kim et al., 2016; Leod & Verma, 2016; Nascimento et al., 2016; Amit et al., 2017; Duraisamy & Emperumal, 2017; Han, Kang, et al., 2017; Khan, 2017; Kumar, H.S, et al., 2017; Rasti et al., 2017; Samala et al., 2017; Wan et al., 2017; Zheng et al., 2017; Feng et al., 2018; Samala et al., 2018b)
Scaling	Methods like Gaussian Pyramid, Bi-cubic interpolation,	Required to resize the image before served as input to the DNN.	(Arefan et al., 2015; Fonseca et al., 2015; Cheng et al., 2016; Kim et al., 2016; Spanhol et al.,

Preprocessing Method	Methodology	Advantages	References
	Bilinear interpolation	<p>Carefully selected interpolation methods can avoid the loss of information in mapping to the new pixel grid.</p> <p>Gaussian pyramid can help to increase the number of images along with resizing.</p>	<p>2016a; Xu et al., 2016; Zhang et al., 2016; Abdullah-Al et al., 2017; Bakkouri & Afdel, 2017; Bayramoglu et al., 2017; Carneiro et al., 2017; Chang et al., 2017; Dhungel et al., 2017; Duraisamy & Emperumal, 2017; Han, Kang, et al., 2017; Jaffar, 2017; Kumar, Kumar, et al., 2017; Nejad et al., 2017; Wan et al., 2017; Gandomkar et al., 2018; Yao, Zhang, Zhou, & Liu, 2019)</p>
Normalization & Enhancement	Histogram equalization, adaptive Mean, Median filters, Log transforms, CLAHE method, Wiener Filter	<p>Normalize the low-value and high-value intensity/contrast present in an image.</p> <p>Adaptive filters remove noise by mean, variance, and spatial correlations.</p> <p>Reduces US image blurring effects and impulse noise.</p> <p>DNN usually shows better performance on a normalized image, which helps to minimize loss while backpropagation.</p>	<p>(Arefan et al., 2015; Arevalo et al., 2015; Rouhi et al., 2015; Bejnordi et al., 2017; Duraisamy & Emperumal, 2017; Han, Kang, et al., 2017; Jaffar, 2017; Khan, 2017; Nejad et al., 2017; Rasti et al., 2017; Sert et al., 2017; Krishna & Rajabhushnam, 2019)</p>
Remove Artifacts	Using binary images and thresholding the pixel intensity, cropping border, Extracting Bigger regions, using geometric parabola	<p>Help to eliminated non-breast regions (labels, wages, opaque markers, white strips/borders, thorax, lungs, chest wall, and pectoral muscle) in mammograms, US and MRI.</p>	<p>(Cao et al., 2016; Abdullah-Al et al., 2017; Wan et al., 2017; Gandomkar et al., 2018; Mulooly et al., 2019)</p>

Preprocessing Method	Methodology	Advantages	References
	around the rib cage.		
Stain Normalization or Removal	Stain Normalization	<p>To make variable color (due to H&E staining of Hp images) uniform in all images of all patients. So that DNN will not distract due to brightness and color stain inconsistencies and show better classification results for multiclass BrC.</p> <p>Contrast, intensity, and color statistics of source images are almost like the reference image.</p> <p>The Reinhard method preserves the structures of Hp images. Therefore, suitable for BrT classification.</p> <p>Khan's supervised method works at the pixel level and thus achieves, a good result for stain separation.</p>	(Arefan et al., 2015; Bevilacqua et al., 2016; Bayramoglu et al., 2017; Sert et al., 2017; Kumar et al., 2020)
	Color Deconvolution	<p>To extract intensities of hematoxylin-eosin (H&E) staining from Hp images and convert them into optical density space images without being significantly influenced. Hence it reduces the image dimensionality and uses the least resources and enhances the performance of classification.</p> <p>By adopting filtered and independent observations it reduces the impurity of the signal when estimating the stain matrix.</p> <p>It preserves texture information that is associated with stain colors in Hp images.</p>	

2.5.1 Augmentation

Augmentation creates new images by increases the number of instances (BrC images) using basic image preprocessing techniques. In general, a DNN model requires a large number of images to be trained to produce reliable results. Indeed, image augmentation is required when the target dataset does not contain enough images for training a DNN model properly. This review identified four types of augmentation techniques, of which geometric transforms, noise addition, and patch extraction were implemented over breast images directly and synthetic minority over-sampling technique was adopted for feature vector data (manually extracted from breast images) before feeding to any ANN. For instance, some studies (Arevalo et al., 2015; Kim et al., 2016; Amit et al., 2017; Bakkouri & Afdel, 2017; Bayramoglu et al., 2017; Bejnordi et al., 2017; Byra et al., 2017; Carneiro et al., 2017; Chang et al., 2017; Dhungel et al., 2017; Duraisamy & Emperumal, 2017; Hadad et al., 2017; Han, Wei, et al., 2017; Jaffar, 2017; Nejad et al., 2017; Rasti et al., 2017; Samala et al., 2017; Sert et al., 2017; Zhang et al., 2017; Zheng et al., 2017; Bardou et al., 2018; Gandomkar et al., 2018; Samala et al., 2018b; Kumar et al., 2020) utilized geometric transform (e.g., rotation at various angles, flip horizontally and vertically).

However, other studies (Cheng et al., 2016; Spanhol et al., 2016a; Xu et al., 2016; Araujo et al., 2017; Duraisamy & Emperumal, 2017; Kumar, Kumar, et al., 2017; Spanhol et al., 2017; Zheng et al., 2017; Feng et al., 2018; Gandomkar et al., 2018) extracted many patches from the original image. Whereas, patches are extracted by using three strategies, namely, a random number of patches (Spanhol et al., 2016a; Spanhol et al., 2017), patches with 50% overlapping (Spanhol et al., 2016a; Araujo et al., 2017), and patches with no overlapping (fixed size window) (Cheng et al., 2016; Xu et al., 2016; Duraisamy & Emperumal, 2017; Kumar, Kumar, et al., 2017; Zheng et al., 2017; Feng et al., 2018; Gandomkar et al., 2018). Furthermore, augmentation by using noise addition or color

variation was adopted in previous studies (Bejnordi et al., 2017; Chang et al., 2017) to train a model robustly to handle noisy images while performing class label prediction. For instance, Chang et al. (2017) added a random distortion to original images while creating new images.

2.5.2 Image Region of Interest Extraction

An original breast image may contain many regions of normal and abnormal tissues, and the segregation of these regions is known as ROI extraction. ROI extraction has two major advantages. First, it increases the number of training and testing images required for DNNs. Second, it supports DNNs to learn only normal and abnormal regions instead of irrelevant regions. Many studies (Arefan et al., 2015; Arevalo et al., 2015; Fonseca et al., 2015; Rouhi et al., 2015; Bevilacqua et al., 2016; Cao et al., 2016; Cheng et al., 2016; Kim et al., 2016; Leod & Verma, 2016; Nascimento et al., 2016; Amit et al., 2017; Duraisamy & Emperumal, 2017; Han, Kang, et al., 2017; Khan, 2017; Kumar, H.S, et al., 2017; Rasti et al., 2017; Samala et al., 2017; Wan et al., 2017; Zheng et al., 2017; Feng et al., 2018; Samala et al., 2018b) extracted ROIs from the original image before BrT classification. For instance, Samala et al. (2018b) extracted thousands of ROI from 3D MG DBT images. Similarly, Rouhi et al. (2015) cropped the ROI of abnormal tissues and mass regions before BrT classification.

2.5.3 Scaling

Scaling or resizing is an important preprocessing task applied to images before they are fed directly into a DNN. Image scaling or interpolation occurs when an image is resized from the one-pixel grid to another. It increases or decreases the number of pixels by remapping. Most of the selected studies (Arefan et al., 2015; Fonseca et al., 2015; Cheng et al., 2016; Kim et al., 2016; Spanhol et al., 2016a; Xu et al., 2016; Zhang et al., 2016; Abdullah-Al et al., 2017; Bakkouri & Afdel, 2017; Bayramoglu et al., 2017;

Carneiro et al., 2017; Chang et al., 2017; Dhungel et al., 2017; Duraisamy & Emperumal, 2017; Han, Wei, et al., 2017; Jaffar, 2017; Kumar, Kumar, et al., 2017; Nejad et al., 2017; Wan et al., 2017; Gandomkar et al., 2018; Yao et al., 2019) adopted interpolation methods, such as nearest neighborhood, bilinear, or bi-cubic. For instance, Dhungel et al. (2017) adopted the bi-cubic interpolation method to rescale images before feeding into a five-layered CNN for BrC binary classification. Zhang et al. (2016) utilized bilinear interpolation to resize US BrC images for binary classification. However, Bakkouri and Afdel (2017) adopted Gaussian pyramids to reduce and expand image size using MG images before classification.

2.5.4 Normalization and Enhancement

Medical image acquisition and digitization are affected by involving color and light conditions. Hence, different color and light conditions affect all pixel values present in an image. To overcome these issues, researchers adopted many techniques, which can be broadly divided into two categories: global or local image normalization and enhancement techniques.

Global image normalization and enhancement techniques perform the same operation on all pixels of images, such as histogram, mean, and median contrast/intensity normalization. By contrast, local image normalization and enhancement techniques operate on any pixel depending on the contrast or intensity of the neighboring pixels. DNNs usually perform better when the input images are normalized and decorrelated because these properties help gradient-based optimization and learning (Jarrett, Kavukcuoglu, & LeCun, 2009; Krishna & Rajabhushnam, 2019).

Whereas, some studies (Arefan et al., 2015; Arevalo et al., 2015; Bejnordi et al., 2017; Duraisamy & Emperumal, 2017; Han, Kang, et al., 2017; Jaffar, 2017; Khan, 2017; Nejad et al., 2017; Rasti et al., 2017; Sert et al., 2017) utilized the techniques to improve image quality before feeding into any type of DNN for BrT classification. For instance, some

studies (Bejnordi et al., 2017; Duraisamy & Emperumal, 2017; Nejad et al., 2017) employed global contrast normalization by using mean filters to solve the multiclass BrT classification problem. Khan (2017) removed US image speckle noise and blurring effect by adopting Wiener and adaptive filters (e.g., mean, variance, and spatial correlations). The author reduced the impulse noise usually found in US images by using the mean filter and wavelet shrinkage. Moreover, image local contrast enhancement was performed by contrast limited adaptive histogram equalization (CLAHE). However, Jaffar (2017) adopted a hybrid of a bilateral filter with log transformation to preserve edges while performing image normalization.

2.5.5 Removing Artifacts

Artifacts are removed from breast images to eliminate all non-breast regions from the original raw image. Some imaging modalities, such as MG, US, and MRI, possess many artifacts (e.g., labels, wages, opaque markers, white strips/borders, thorax, lungs, chest wall, and pectoral muscle) (Figure 2.9) that should be removed before starting the BrT classification task. Few studies (Arefan et al., 2015; Bevilacqua et al., 2016; Bayramoglu et al., 2017; Sert et al., 2017; Mullooly et al., 2019) adopted preprocessing techniques to remove non-breast regions because they may not use the entire raw image but breast image ROIs for classification. For instance, Arefan et al. (2015) extracted non-breast regions from MGs in two steps, namely, the creation of binary images created by pixel thresholding to detect connected areas and the deletion of small disconnected areas. Hence, the breast region is separated from the rest of the background before performing breast density multiclassification, such as fatty, glandular, or dense breasts. Bevilacqua et al. (2016) classified breast US images after eliminating the thorax part by considering a geometric parabola that follows the rib cage border. Moreover, Sert et al. (2017) removed white strips found at MG borders by thresholding the intensity value to 150.

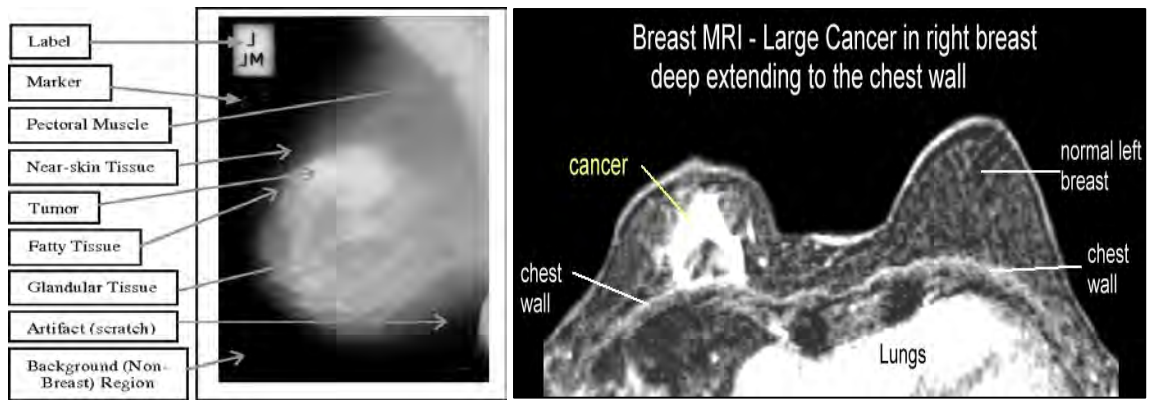


Figure 2.9: Different artifacts in a mammogram (Left image) and MRI (right image) (Saidin, Sakim, Ngah, & Shuaib, 2012; Breast Cancer Imaging, 2018)

2.5.6 Stain Normalization

In digital pathology (DP) labs, the preparation of Hp biopsy images involves different chemicals, stains, lighting effects, and scanners to develop digital images from collected breast tissue samples. The inconsistencies in Hp images may be introduced by using different chemicals for staining, concertation of colors, or different scanners from many vendors. Moreover, these factors may create major inconsistencies in images of two patients even if images are prepared in the same DP lab. To eliminate these inconsistencies, previous studies used RGB histogram specification, Reinhard's (Reinhard, Adhikhmin, Gooch, & Shirley, 2001), Macenko's (Macenko et al., 2009), and Khan's methods (Khan, Rajpoot, Treanor, & Magee, 2014) to normalize the Hp images before classification, see Figure 2.10.

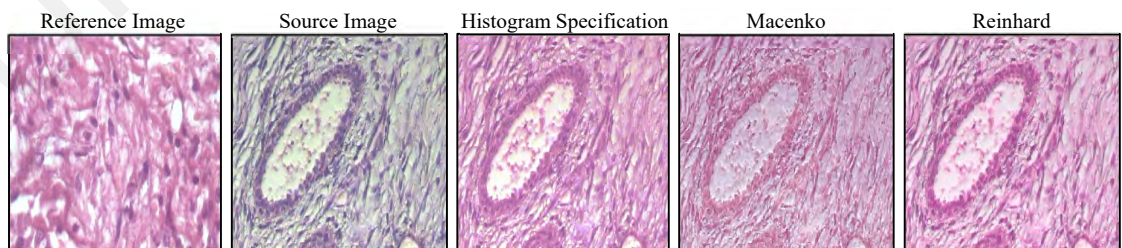


Figure 2.10: Source image stain normalized by using a reference image through three techniques

Many studies (Cao et al., 2016; Abdullah-Al et al., 2017; Araujo et al., 2017; Wan et al., 2017; Zheng et al., 2017; Gandomkar et al., 2018; Kumar et al., 2020) employed stain normalization or removal techniques before proceeding toward BrT classification. For

instance, Abdullah-Al et al. (2017) used Retinex operation to perform a non-linear transformation to normalize illumination. Wan et al. (2017) adopted Khan's method to perform a non-linear mapping-based stain normalization. Gandomkar et al. (2018) employed two stain normalization methods, namely, the histogram specification-based method and Reinhard's method; the latter uses mean and standard deviation to match RGB channels with the reference image. Furthermore, (Zheng et al., 2017) removed the color stain by using the color deconvolution method proposed by Ruifrok and Johnston (2001). This method separates the color information acquired by H&E staining. It determines the contribution of all applied stains according to the stain specific RGB absorption.

2.6 Machine Learning Based Classification Model Types Used for BrC Detection and Classification

This literature review reveals that two types of predictions are made for BrC diagnosis. First is related to predicting either the BrT is benign or malignant named BrC detection. Whereas, the second one is to diagnose the subtypes of BrT for each benign (i.e., A, F, TA, and PT) and malignant (i.e., DC, LC, MC, and PC) commonly referred to as BrT classification. Moreover, this literature review reveals that BrC detection and BrT classification models are mainly of two types. The first type of model is based on ML classifiers like softmax, kNN, LDA, linear regression (LR), NB, DT, and SVM. Whereas, the second type of models are based on artificial neural networks (ANN). However, ANN is mostly using the softmax classifier for the detection/classification of BrC Hp images.

2.6.1 Traditional Machine Learning Classification Models Used for BrC Detection and Classification

It is observed while performing the review, that the most commonly used ML classifiers for BrC detection/classification are kNN, LDA, LR, NB, DT, SVM, and

softmax as shown in Table 2.6. This table shows the distribution of studies where various types of ML classifiers are used for BrC detection/classification. The basic concept behind each of the aforementioned ML classifier is given as follows.

The kNN (Fix, 1951) is one of the simplest classifier also known as non-parametric (i.e., little or no prior knowledge about the distribution of the data is required), lazy learning algorithm. Simply, the kNN separates data points into several classes in order to predict a class label for a new testing point. It is lazy because it does not use training data points for generalization. It determines the feature similarity i.e. how much a new testing point is similar to the training points. Where class label or discrete value for a new testing point is decided by majority voting of k nearest neighbors.

The LDA (Fisher, 1936) is basically a dimensionality reduction method, which is also used to solve classification problems in supervised learning (requires labeled data for training) manner. LDA converges data points into lower dimension space from a higher dimension space. Moreover, LDA can make groups of data points in lower dimension space to separate them into two or more classes.

In statistics, the LR is a linear approach to model the relationship between the dependent and independent variables. As a machine learning classifier, LR fits a line into a data point and maps numeric inputs to numeric outputs. It creates a model by creating a relationship between one or more variables. LR model ensures generalization to predict outputs for unseen inputs.

The NB is a classification technique based on the Bayes theorem (i.e., calculating posterior probability). NB assumes that the attributes of data are independent of each other, therefore called naive or simple. NB classifiers can be built easily and work better on a large dataset.

The DT is an unsupervised machine learning predictive modeling approach. It creates a decision tree as a predictive model from training data to make predictions over testing

data. In DT the leaves represent the class labels, branches are used as features that lead to class labels.

The SVM (Cortes & Vapnik, 1995) is a supervised machine learning classifier used for classification/regression. SVM represents data points in space by dividing with a gap as wide as possible. The unseen testing data points are then mapped into the same space to be predicted a category based on the side of the gap on which they fall.

However, softmax is a function used in the last layer of ANNs (like in CNN for image classification) to predict the probabilities for class labels and to quantify how good or bad a prediction is made. Thus, many researchers (Liao, Xu, Lv, & Zhou, 2015; Qi, Wang, & Liu, 2017; Daghighi, Medini, & Shrivastava, 2019) used softmax function based layer and termed it as a softmax classifier. Therefore, in this research, the term softmax is used as a softmax classifier.

Table 2.6: Distribution of studies using various machine learning classifiers for BrC detection and classification

Study References	Machine Learning Classifiers							BrC Detection/ Classification
	kNN	LDA	LR	NB	DT	SVM	softmax	
Kumar et al. (2020)	No	No	No	No	Yes	Yes	Yes	Detection
Mullooly et al. (2019)	No	No	No	No	Yes	No	Yes	Detection
Krishna and Rajabhushnam (2019)	No	No	No	No	Yes	Yes	No	Detection
Bardou et al. (2018)	Yes	No	No	No	Yes	Yes	Yes	Classification
Nahid and Kong (2018)	No	Yes	No	No	No	Yes	Yes	Detection
Araujo et al. (2017)	No	No	No	No	No	Yes	Yes	Classification
Han, Wei, et al. (2017)	No	No	No	No	No	No	Yes	Classification
Nahid et al. (2018)	No	No	No	No	No	Yes	Yes	Detection
Samala et al. (2017)	No	No	No	No	No	Yes	Yes	Detection
Thirumalai and Manzoor (2017)	No	No	Yes	No	No	No	No	Detection

Study References	Machine Learning Classifiers							BrC Detection/ Classification
	kNN	LDA	LR	NB	DT	SVM	softmax	
Wan et al. (2017)	No	No	No	No	No	Yes	No	Detection
Zheng et al. (2017)	Yes	No	No	No	No	Yes	Yes	Detection
Oleksyuk, Saleheen, Caroline, Pascarella, and Won (2016)	No	No	No	Yes	No	No	No	Detection
Pritom, Munshi, Sabab, and Shihab (2016)	No	No	No	Yes	No	No	No	Detection
Spanhol et al. (2016a)	Yes	No	No	No	Yes	Yes	Yes	Detection
Deng and Perkowski (2015)	No	No	No	Yes	No	No	No	Detection
Rouhi et al. (2015)	No	No	No	Yes	No	No	Yes	Detection
Lo, Shen, Huang, and Chang (2014)	No	No	Yes	No	No	No	No	Detection
Ahn et al. (2013)	No	No	Yes	No	No	No	No	Detection
Zhang (2011)	No	No	No	No	No	Yes	No	Classification

2.6.2 Artificial Neural Network Used in BrC Detection and Classification

The human brain consists of more than 10 billion interconnected neurons. Using chemical reactions, each neuron obtains information, processes it, and responds accordingly. Similarly, artificial neuron (AN) mimics the simple methods of mammal neurons, see AN in Figure 2.11. The first simplified artificial neuron was introduced by (McCulloch & Pitts, 1943). A group of ANs forms a layer, and a group of layers creates an ANN, see Figure 2.11. An ANN is an ML technique that can learn and perform tasks, such as classification, prediction, decision-making, and visualization, by using sample data. Moreover, an ANN can perform multi-disciplinary tasks by using many types of real-life data, including structured (data in vector form), semi-structured (like emails), and unstructured data, such as BrC medical images. Many types of ANNs were developed

to process different types of data. For the classification of BrC medical images, researchers mainly used two types of ANNs, namely, shallow neural networks (SNNs) and DNNs, see Figure 2.12. Most researchers employed DNNs (also known as deep learning-based models) for BrT classification. In subsequent subsections, the types of ANNs used for BrT classification are discussed in the light of selected studies. Moreover, the pros and cons of each model are presented in Table 2.8.

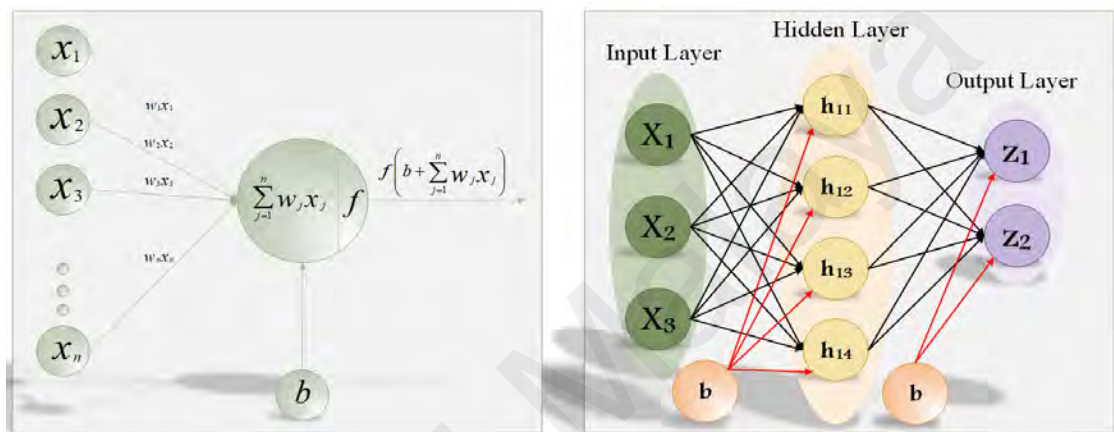


Figure 2.11: Left: An artificial neuron. Right: sample of an artificial neural network

2.6.2.1 Shallow Neural Network

An ANN with a single hidden layer is referred to as an SNN (Bebis & Georgiopoulos, 1994). The basic building block (elementary unit) of an ANN is an artificial neuron, simply referred to as neuron or node or hidden units. A simple ANN is a mathematical function that works similar to a biological neuron. The output of an ANN is represented by connection weights that update the effect of a given input, and the nonlinear characteristics are applied by any transfer function at a particular neuron. Afterward, neuron impulse is calculated by applying a non-linear function (i.e., activation function) on a weighted sum of input data. Simultaneously, a learning algorithm (e.g., backpropagation) is responsible for updating the weight to show the model's learning capability. A simple ANN and the basic structure of SNN are shown in Figure 2.11.

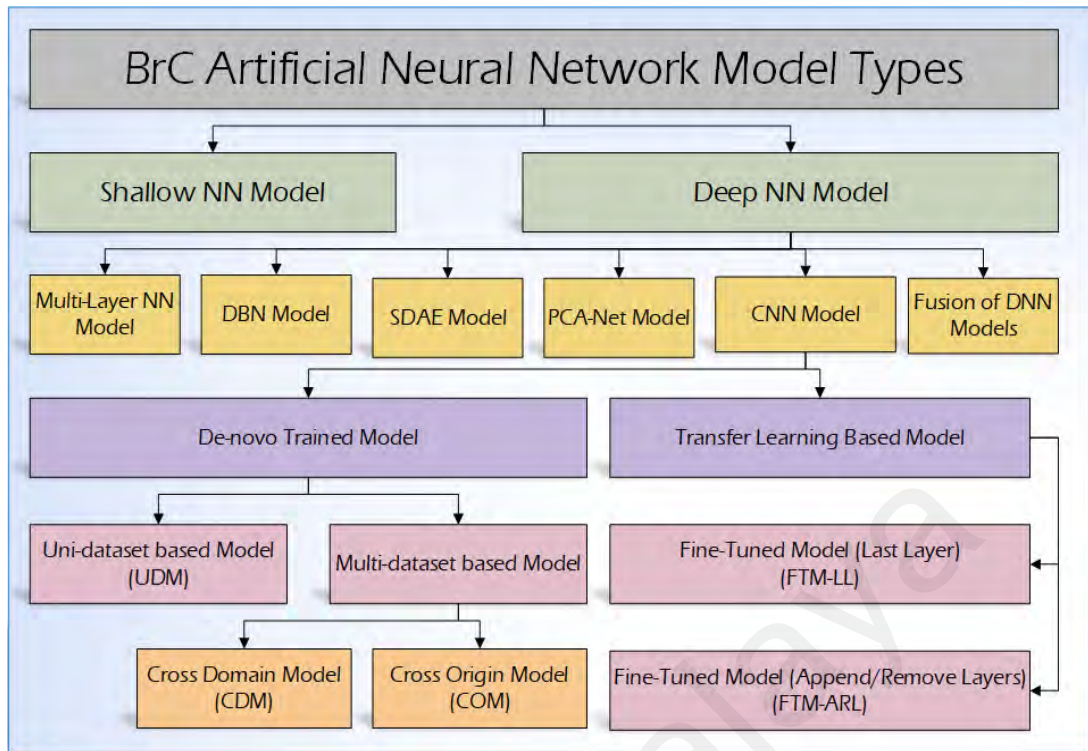


Figure 2.12: Type of ANNs used for BrC detection and BrT classification

In Figure 2.11, a simple ANN obtains unidirectional input, such as $x_1, x_2, x_3, \dots, x_n$, shown by arrows toward the activation function based on the weighted sum of input data. The neuron output is represented by $f(y)$ and has the following relationship:

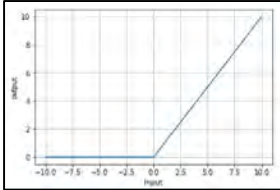
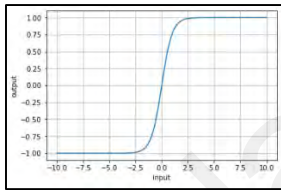
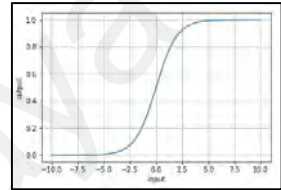
$$f(y) = f\left(b + \sum_{j=1}^n w_j x_j\right), \quad (1)$$

where x_j, w_j represents the input and weight matrix, respectively, b is the bias neuron that allows a classifier to translate its decision boundary, $f(y)$ is an activation function, and y is the sum of the scalar product of the weight matrix and input.

$$y = w_1 x_1 + w_2 x_2 + w_3 x_3 + \dots + w_n x_n. \quad (2)$$

A nonlinearity function is also applied after the sum of the dot product of weighted inputs. The non-linearity function is also known as the activation function (Duch & Jankowski, 1999). The most popular choices for activation functions are rectified linear unit (ReLU), tanh, and sigmoid as shown in Table 2.7.

Table 2.7: Brief description of popular activation functions

Activation Functions	ReLU	Tanh	Sigmoid
Equation	$\phi(x) = \max(x, 0)$	$\phi(x) = \tanh(x)$	$\phi(x) = \frac{1}{1 + e^{-x}}$
Range	$(0, \infty)$	$(-1, 1)$	$(0, 1)$
Type	Discontinuous	Continuous	Continuous
Gradient	<i>if $\phi > 0$ then 1, else 0</i>	$1 - \phi(x)^2$	$\phi(x)(1 - \phi(x))$
Update Suppressed near zero	Yes	Yes	No
Overcome vanishing Gradient	Yes	No	No
Graph			

As shown in Figure 2.11, the nodes are distributed into the input, hidden, and output layers. The input signal flows from the input layer, passes toward the hidden layer, and ends at the output layer. Such type of input flow in a strict feed-forward fashion develops a feed-forward ANN (FF-ANN). However, instead of using an ANN with a single hidden layer, multiple hidden layers can be used, in FF-ANN. Noticeability, none of the nodes possess any connection within the same layer. This independence of neurons within a layer supports parallel computations while training an ANN. The training of an ANN is a learning process where patterns are learned from input data by changing the weights after applying some learning rules. Learning rules, such as backpropagation, delta rule, and perceptron rule, help modulate weights automatically while training the network. The trained ANN can then be used for prediction using real-life data.

Many studies have created SNN models to classify BrC (Rouhi et al., 2015; Leod & Verma, 2016; Kumar, H.S, et al., 2017). Kumar, H.S, et al. (2017), ensembled six binary ANNs for 4-class breast density grading classification using MGs. Rouhi et al. (2015) developed an SNN model to find the threshold for regions growing segmentation and

classification of MGs into benign or malignant cases. These studies highlighted that using SNNs is beneficial for BrT classification. SNNs have some basic advantages owing to their simple structure. They possess a single hidden layer and work in a feed-forward fashion, thereby allowing them to create, implement, and optimize BrT classification. SNNs consume the least computation resources and time among the different types of ANNs. Moreover, SNNs can produce better results than other types of ANNs even if the dataset is small. However, using SNN also has some limitations. For instance, an SNN used for structured data has a limited number of dimensions; otherwise, small networks are unable to show good generalization performance over high-dimensional data, especially when complex patterns need to be learned to solve multiclass problems. Moreover, the performance of the network depends on the designed features and the optimization of the network structure.

2.6.2.2 Deep Neural Networks

DNNs are used for DL as an ML method and AI technique for automatic feature extraction. Usually, the word *deep* is referred to when more than one hidden layer has been deployed between the input and output layers of any NN (Svozil, Kvasnicka, & Pospichal, 1997). DNNs use representation learning to discover complex feature representations automatically (such as diagnosis of BrC using medical images) unlike traditional ML classifiers (e.g., support vector machine, random forest decision tree, and k-nearest neighborhood), which require HcFs to show optimum results. The empirical success of DNN is inherited by its mathematical formulas (Goceri, 2018). Over the years, DNNs focused on applications such as speech recognition (Hannun et al., 2014; Amodei et al., 2016), fraud detection (Paula, Ladeira, Carvalho, & Marzagão, 2016; Wang & Xu, 2018), traffic sign detection (Islam, Raj, & Mujtaba, 2017), face recognition (Sun, Chen, Wang, & Tang, 2014; Parkhi, Vedaldi, & Zisserman, 2015), emotion recognition (Jirayucharoensak, Pan-Ngum, & Israsena, 2014; Kahou et al., 2016), medical image

diagnosis (Wu, Chen, & Ding, 2014; Lakhani & Sundaram, 2017; Siddiqui, Mujtaba, Reza, & Shuib, 2017), and human activity recognition (Nweke, Teh, Alo, & Mujtaba, 2018; Nweke, Teh, Mujtaba, & Al-garadi, 2019).

The upsurge in DL research is fueled by its ability to extract salient features from raw images of BrC without relying on laboriously extracted HcFs. In recent years, an extensive number of DNNs have been proposed. The DNNs can be broadly categorized into multi-layer neural networks (ML-NN), deep belief neural network, stacked denoising auto-encoders (SDAE), principal component analysis network (PCANet), and CNN. Furthermore, CNN models were either trained from scratch called de-novo models or created through TL by using pre-trained models, see Figure 2.12. In subsequent subsections, the types of DNNs used for BrT classification are discussed in the light of selected studies.

(a) Multi-Layer Neural Network

An ML-NN is a type of DNN that is similar to an SNN. Nonetheless, an ML-NN possesses two or more hidden layers between the input and output layers, unlike an SNN (Bengio, 2009; Deng & Yu, 2014), see Figure 2.13. However, ML-NN training must be configured to obtain the desired results. Configuring an ML-NN is actually initializing and modulating the parameters to perform optimum training, such as initializing weights by generating any random number or by using prior domain knowledge before initiating the learning rule. Recently, the most popularly adopted learning rule is backpropagation (Abraham, 2005). In backpropagation, the weights are automatically updated in each pass on the basis of error rate (loss) produced at the output layer by using gradient and chain-rule (Svozil et al., 1997). However, this literature review revealed that very few studies used ML-NN for BrT classification. Like, Kumar, H.S, et al. (2017) proposed an ML-NN model with two hidden layers and optimized by different stopping criteria using 22

morphological features extracted from 100 US images to classify benign or malignant BrC.

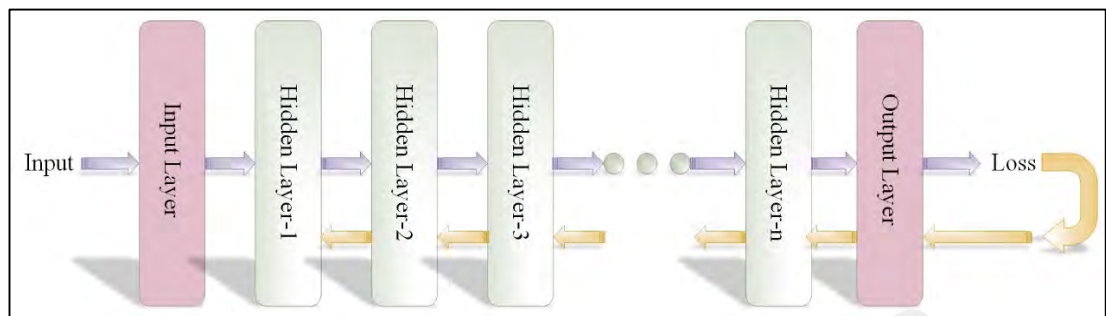


Figure 2.13: A sample illustration of Multi-Layer Neural Network

Furthermore, Arefan et al. (2015) developed an ML-NN model using 2–20 hidden layers. They extracted nine statistical features from 43 MG images to classify breast density as fatty, glandular, or dense. The afore-stated studies showed the urge of using ML-ANN. For instance, increasing the number of hidden layers can improve the generalization performance of the network. However, additional layers require more data instances for better training; otherwise, the network may be overfitted (good performance on validation data but unable to perform on target data). Furthermore, optimizing the number of hidden layers and training hyper-parameters for a larger size of ML-NN become crucial tasks (for further details, see Table 2.8).

Table 2.8: ANN models used in selected studies for BrT classification

ANN Types	Strengths	Weaknesses
SNN	<p>Small size networks.</p> <p>Easy to develop, train and optimize the training parameters.</p> <p>A small amount of data can obtain better generalization performance.</p> <p>Requires less training time, computational power, and memory to store weights.</p>	<p>Do not show good performance on high dimensional data.</p> <p>Performance solely depends upon the designed features and the structure of ANN.</p> <p>Difficult to generalize the predictions.</p>
ML-NN	<p>It includes all advantages of SNN, additionally, the increased hidden</p>	<p>Includes all weaknesses of SNN, additionally, a higher number of</p>

ANN Types	Strengths	Weaknesses
	<p>layers help to get better generalization performance.</p> <p>High Dimensional data can be used for better feature extraction.</p>	<p>hidden layers need more data to get better generalization performance.</p>
DBN	<p>This efficient, greedy learning can be followed by, or combined with, other learning procedures that fine-tune all of the weights to improve the generative or discriminative performance of the whole network.</p> <p>It can be deployed for high dimensional data that possess correlated features.</p>	<p>Unable to track the loss while computing the log-likelihood.</p>
SADE	<p>Automatic denoising from high dimensional data enhances the performance of the BrT classification model, using a real-life medical image.</p> <p>Can track cross entropy which is what is being minimized by the model's learning algorithm like back-propagation.</p>	<p>Denoising works better on high dimensional data compared to low dimensions because of higher dependencies usually found among higher dimensions like BrC medical images.</p>
PCA-Net	<p>Due to the large receptive field, it can extract overall observations of the objects in an image and captures more semantic level information.</p> <p>Due to binary hashing and block histogram, PCANet is flexible for mathematical analysis and justification of its effectiveness.</p>	<p>The use of a simple hashing method cannot provide rich enough information to map the features. Hence affects the representation performance.</p> <p>Preferred when data possess much irrelevant information.</p>
CNN (De-novo)	<p>CNN(UDM): Customized deep CNN models can be created.</p> <p>A model can be created according to the type and number of images.</p> <p>CNN(CDM): Includes the same strengths as in CNN(UM)</p>	<p>CNN(UDM): usually difficult to train a model for a small number of images to solve the multiclass problem.</p> <p>Needs high expertise to design and optimize the deep network for specific data. May consumes lots of time and resources to get optimum results.</p> <p>CNN(CDM): Two times training of the model will take a longer time and may require higher resources.</p>

ANN Types	Strengths	Weaknesses
	<p>Additionally, a model can be effective even if there are fewer target images to solve multiclass classification.</p> <p>CNN(COM): Customized deep CNN models can be created. The training, validation, and testing were performed on a larger number of images of the same modality. Usually, it shows better performance. Preferred when source images (usually exclusive dataset images) are not enough for training. It allows using all target images for testing purposes only.</p>	<p>Hard to optimize model training on two datasets of different domains like ImageNet and BreakHis.</p> <p>Requires a large number of instances (BrC images) with balanced distribution among classes.</p> <p>CNN(COM): Medical images collected from different sites always have different image acquisition protocols. Hence needs extra and carefully adopted preprocessing methodologies to get a reliable generalized model.</p>
CNN (Pre-trained)	<p>A deep CNN model can be trained quickly using the least resources compared to de-novo training.</p> <p>It can show comparable performance even if target data is smaller in size like Hp BrC images.</p>	<p>If the target dataset is very small (like 100 images) then results may be not reliable.</p> <p>Retraining also requires class wise balance data to produce unbiased results, usually not found in real-life medical images.</p>
	<p>CNN(FTM-ARL) possesses the fusion of new layers to be trained from scratch, so flexible to learn more generalized and unbiased weights from a small amount of target data like BrC images compared to FTM-LL.</p>	<p>Limitations are the same as in CNN(FTM-LL) except CNN(FTM-ARL): Training time may increase due to the introduction of new layers to be trained from scratch</p> <p>The optimization of newly appended layers needs to be addressed carefully to get the desired results.</p>

Furthermore, Arefan et al. (2015) developed an ML-NN model using 2–20 hidden layers. They extracted nine statistical features from 43 MG images to classify breast density as fatty, glandular, or dense. The afore-stated studies showed the urge of using ML-ANN. For instance, increasing the number of hidden layers can improve the generalization performance of the network. However, additional layers require more data

instances for better training; otherwise, the network may be overfitted (good performance on validation data but unable to perform on target data). Furthermore, optimizing the number of hidden layers and training hyper-parameters for a larger size of ML-NN become crucial tasks (for further details, see Table 2.9).

(b) Deep Belief Networks

A deep belief network is a type of DNN (Hinton, Osindero, & Teh, 2006) that consists of several layers of restricted Boltzmann machines (RBMs), see Figure 2.14(a) (Fischer & Igel, 2012). An RBM is a generative model that serves as a building block in greedy layer-wise feature learning and training of DNN. RBM maps binary data-vectors using binary latent variables. Hence, the goal is to obtain discriminative representation features. If the RBM network cannot directly be used for medical images (e.g., SWE images), then Point-wise gated Boltzmann machines (PGBM) (Figure 2.14(b)) are adopted to model complex image data (e.g., BrC US-SWE images) while avoiding irrelevant patterns.

Moreover, in unsupervised learning (performed by using unlabeled data), a DBN can learn to probabilistically reconstruct its inputs. Hence, a hidden layer works like a feature extracting entity. All the hidden layers are trained one after the other, i.e., one layer at a time. Finally, a DBN can be trained in a supervised fashion for classification, see Figure 2.14(c). However, only one study utilized the advantages of DBN for BrT classification [28]. Zhang et al. (2016) deployed a two-layered DBN composed of PGBM and RBM for BrC binary classification by using breast US-based SWE colored images. PGBM was equipped to distinguish between relevant and irrelevant features from SWE images. Furthermore, relevant features were supplied to RBM to learn the relationship among the BrC relevant features. Finally, SVM was used to classify benign or malignant BrC cases by using features extracted through RBM.

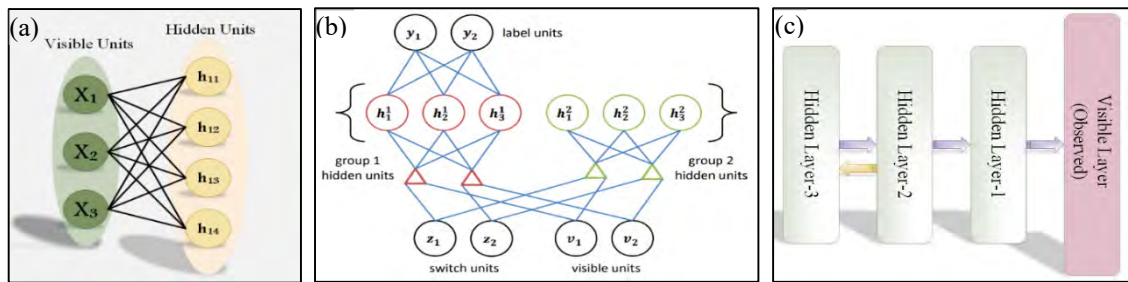


Figure 2.14: A restricted Boltzmann machine (RBM) with fully connected visible and hidden units (a), a sample diagram of supervised PGBM shown (b) (Sohn, Zhou, Lee, & Lee, 2013) and (c) shows a sample diagram of supervised DBN

The main advantage of using a DBN for image classification is that it is mostly trained layer by layer, allowing each layer to be optimized easily for improved feature generalization. In addition, the layers, except the last one, can be trained in an unsupervised fashion. The last hidden layer is usually trained in a supervised manner to fine-tune the network output. Hence, a DBN provides an opportunity to perform better training using a small number of annotated images, also called semi-supervised learning. Semi-supervised learning is useful for medical image classification because finding labeled images for different types of cancers is difficult. However, using RBMs layered deep networks also has some limitations. For instance, a DBN cannot track the loss while computing the log-likelihood for which we care about as the better-trained model.

(c) *Stacked Denoising Autoencoder*

A stacked denoising autoencoder (SDAE) is a type of stacked autoencoder that helps eliminate noisy features, see Figure 2.15. SDAE networks can automatically extract discriminant representative hidden patterns from data using an intrinsic data reconstruction mechanism. The SDAE network can hypothetically address the issues of high variations in either shape or appearance of lumps. As the inherent benefit of automatic feature extraction along with noise tolerance, SDAE-based models can conceivably minimize issues related to image processing inaccuracies, which ultimately lead to non-reliable feature extraction.

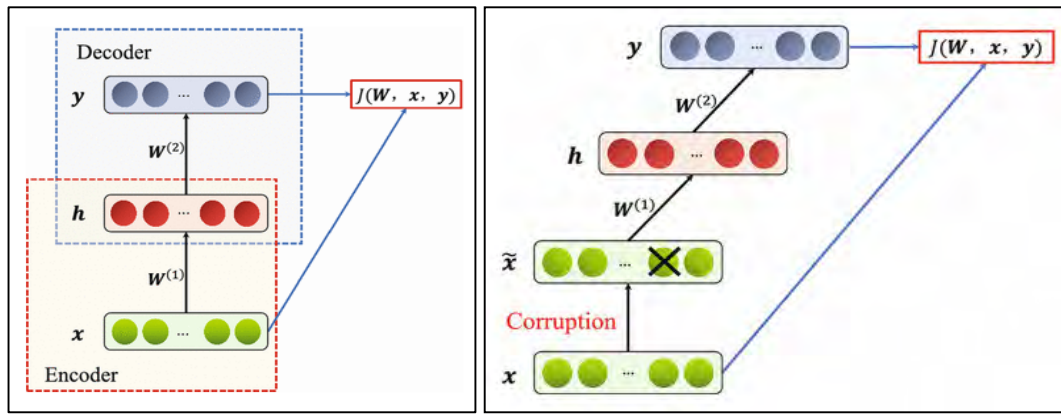


Figure 2.15: Left side figure, a sample network diagram of the traditional autoencoder. Right side figure, a network diagram of stacked denoising autoencoder

Due to noise tolerance nature, few studies (Cheng et al., 2016; Feng et al., 2018) developed an SDAE-based model to classify BrC images. Cheng et al. (2016) developed a model for two-phased training. In the first phase, two-layered SDAE is trained using image ROIs. In the second phase, the pre-trained model is refined by supervised learning with additional neurons to preserve the original image size and aspect ratio. Softmax was used for benign or malignant classification for both breast US and lung CT images, with an Ac and area under the ROC curve (AUC) of $94.4\% \pm 3.2\%$ and $98.4\% \pm 1.5\%$, respectively.

Similarly, Feng et al. (2018) deployed SDAE consisting of three layers along with softmax. An SDAE extracts features layer by layer from breast Hp image ROIs in an unsupervised manner and the model is fine-tuned by using labels to train softmax for benign or malignant BrT classification. The authors obtained $98.28\% \pm 0.12\%$ and $90.54\% \pm 0.45\%$ accuracies for the two classes. These results indicate that the performance of the SDAE-based model is comparable to that of any other type of DNN model because of its integral ability of noise reduction, especially when real-life medical images usually possess noise from different sources.

Hence, auto noise reduction for medical images helps the network to learn more relevant features. Furthermore, layer-by-layer training facilitates easy optimization and regulation of training parameters. Regardless of its major advantages, SDAE also has

some limitations. For instance, SDAE shows poor performance on low-dimensional data or data possessing poor correlation among the dimensions (Vincent, Larochelle, Lajoie, Bengio, & Manzagol, 2010). High-dimensional data, such as medical images, usually inherit a very high correlation.

(d) Principal Component Analysis Network

Principal component analysis network (PCANet) is an easily implementable, two-staged, unsupervised DL technique for image classification (Chan et al., 2015). The two-staged network performs three tasks, namely, cascade PCA, binary hashing, and block-wise histogram. PCA is used to learn multi-stage weights (filter banks), followed by binary hashing and block histograms for indexing and pooling. Binary hashing simply encodes the quantized binary code mapping to the sequence of principal components, see Figure 2.16.

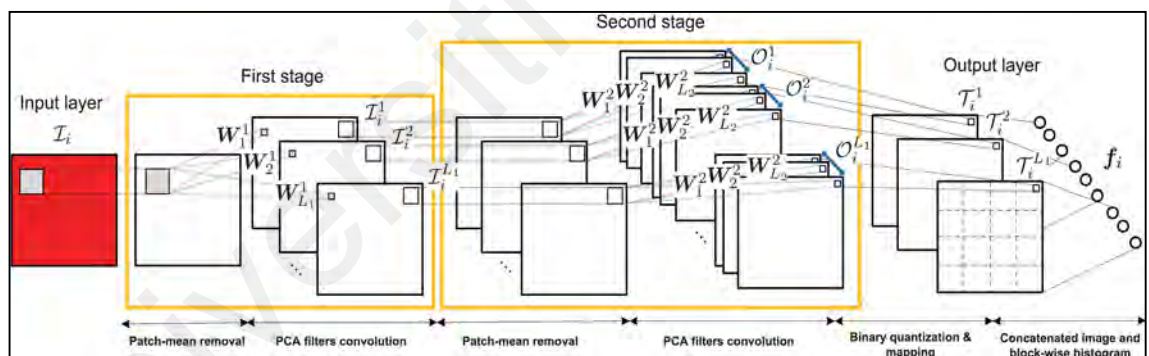


Figure 2.16: A two-staged PCANet block diagram sample (Chan et al., 2015)

According to this review, only one study [34] employed PCANet with some variation of the kernel for breast and liver cancer analysis. Wu et al. (2016) created a PCANet-based model to classify breast/liver cancer Hp images in binary classes. The author used random binary hashing in PCANet instead of simple sequence binary hashing to generate multiple random codes for information extraction. Finally, a low-rank bilinear classifier is used to classify images of two datasets. Compared with other DL-based networks, PCANets are easier to design, implement, and train by using different types of high-

dimensional data. Due to binary hashing and block histogram, PCANet is flexible for mathematical analysis and justification of its effectiveness. Moreover, PCANet has a large receptive field, so that it can extract overall observations of the objects in an image and learn invariance from it. Hence, PCANet can capture pixel-level information.

(e) Convolutional Neural Network

CNN is a type of DL-based ANN technique. This technique has gained attention after work (Hinton & Salakhutdinov, 2006). Moreover, the history of CNN for medical image classification is a long one. Initially, a CNN-based “Neocognitron” model was proposed by (Fukushima & Miyake, 1982). Recently, image classification has been revolutionized after the birth of AlexNet (Krizhevsky, Sutskever, & Hinton, 2012).

A deep CNN model usually consists of some primary layers, such as an input layer, one or more convolution layers, one or more fully connected (FC) layers, and an output layer using softmax to compute label probabilities. Convolution layers are responsible for learning high-level features, such as edges and bobs, whereas FC layers learn pixel-level features. Apart from primary layers, some other layers including a normalization layer (increases network stability) and a pooling layer (progressively reduces the spatial size of the representation to reduce the number of parameters and computation in the network) may be used after convolution layers, and a dropout layer (reduces network overfitting) is usually deployed after the FC layer, see Figure 2.17. However, training is performed in a supervised manner using backpropagation. In addition, hyper-parameters such as input image size and batch size (Goceri & Gooya, 2018) need to be carefully adjusted to obtain optimum results. In brief, the concept of Deep CNN is to make a hierarchical model to represent data at multiple levels of abstraction and enable the model to obtain accurate representations from data in a self-taught manner (Shen et al., 2017).

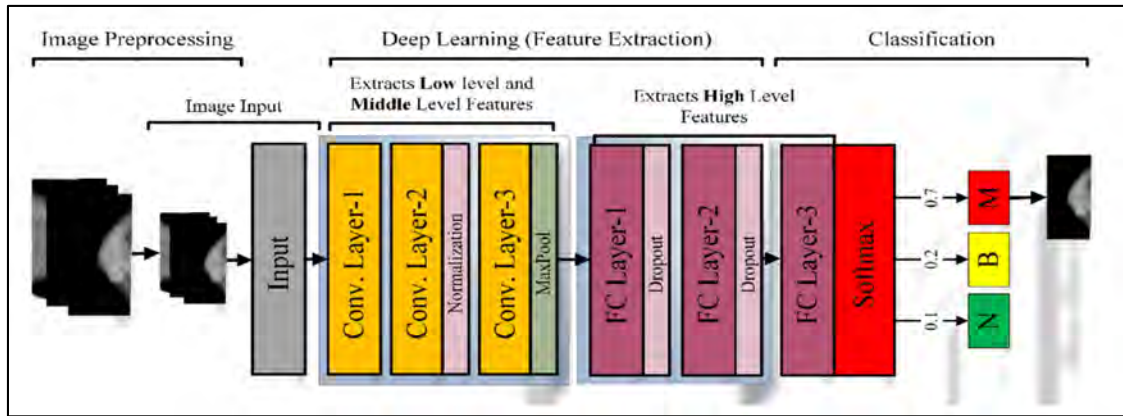


Figure 2.17: An illustration of deep CNN-based model for BrT classification using mammograms

The CCN used for breast classification is divided into two broad categories, namely, the de-novo trained model and the TL-based model, see Figure 2.12. CNN models that were created and trained from scratch are called “de-novo models” (Hadad et al., 2017). Conversely, CNN models that exploited previously trained networks (e.g., AlexNet, VGG-Net, GoogLeNet, and ResNet) are called “TL-based models.”

This survey on BrT classification revealed that many studies (Arevalo et al., 2015; Fonseca et al., 2015; Cao et al., 2016; Kim et al., 2016; Spanhol et al., 2016a; Xu et al., 2016; Abdullah-Al et al., 2017; Amit et al., 2017; Araujo et al., 2017; Bakkouri & Afdel, 2017; Bayramoglu et al., 2017; Bejnordi et al., 2017; Byra et al., 2017; Dhungel et al., 2017; Hadad et al., 2017; Han, Kang, et al., 2017; Kumar, Kumar, et al., 2017; Nahid & Kong, 2017; Nejad et al., 2017; Qiu et al., 2017; Rasti et al., 2017; Sun et al., 2017; Wan et al., 2017; Zheng et al., 2017; Bardou et al., 2018; Nahid & Kong, 2018) used de-novo training.

Conversely, little fewer studies (Bejnordi et al., 2017; Dhungel et al., 2017; Han, Kang, et al., 2017; Kumar, Kumar, et al., 2017; Zheng et al., 2017) employed pre-trained CNN for BrT classification. In this review, the de-novo CNN models are further categorized into two subtypes, namely, uni-dataset and multi-dataset models. Uni-dataset models (UDM) are de-novo models that are trained and tested on a single dataset, whereas cross-domain models (CDM) or cross-origin models (COM) are trained and tested on multiple

datasets, see Figure 2.12. CDM-type models are created from scratch, trained on a dataset of different domains (e.g., nonmedical images), and ultimately retrained (after fine-tuning) for the target dataset, such as BrC images. By contrast, COMs are trained on images of the same domain but collected from different sites, followed by fine-tuning and retraining for the target dataset. However, CDMs are usually smaller in size (possess less number of layers) and created with some special layers to enhance the classification performance compared with pre-trained models such as AlexNet (Han, Kang, et al., 2017).

Apart from models trained from scratch, pre-trained models were also adopted in many studies for BrT classification. The pre-trained models were trained on natural images and mostly possess highly deep structures to learn many class labels; for instance, AlexNet trained for 1000 class labels and contain five convolution layers along with three FC layers, see Figure 4.2. Two strategies were adopted by researchers to perform TL for BrT classification. First, only the last layer was fine-tuned followed by the retraining of the adopted model, named here as the fine-tuned model (last layer) (FTM-LL). Second, one or more layers of the pre-trained network were replaced with newly created layers before retraining the network using target data, named here as the fine-tuned model (append/remove layer) (FTM-ARL), see Figure 2.12.

(f) A fusion of Deep Neural Networks

The review of selected studies showed that most of the CNN-based models use a single type of CNN and are not used in a fused fashion. Some studies (Bejnordi et al., 2017; Nahid & Kong, 2018) deployed models by embedding some residual blocks along with convolutional layers based on pre-trained models, such as ResNet. However, CNN models that were used along with a residual block and were trained from scratch produced good results. For instance, Nahid and Kong (2018) developed a model using a residual block with the convolution layer and obtained an Ac of 92.19%. By contrast, a fusion of

CNNs was prepared by Nahid et al. (2018). The authors deployed three types of model, namely, CNN, long short-term memory (LSTM), and a fusion of CNN and LSTM. The CNN-based model outperformed the other models. Hence, the failure of fused models may be due to the small number of images that are to be fed into a larger fused network. In particular, training from scratch using a small number of images with a large (fused) network may produce unreliable results. Hence, considerable effort is required to assess confidently the effective use of fused CNN type of networks.

2.6.3 Empirical Analysis of Traditional Machine Learning Models Vs. Deep Learning Models for BrC Detection and Classification

It has been discovered in this review, that the aforesaid traditional ML BrC detection/classification models (based on classifiers like kNN, LDA, LR, NB, DT, and SVM) are commonly adopted due to six reasons (Kotsiantis, Zaharakis, & Pintelas, 2007). First, these types of models require fewer computational resources like a normal desktop computer with less training time compared to DL models. Second, the traditional ML classification model usually needs a small number of images for training compared to DL models. Therefore, traditional ML classification models can be trained efficiently using a small number of images to show comparable results. Third, fewer parameter adjustment is required to get almost similar results compared to DL classification models. Fourth, DL-based models get very smaller size input images like AlexNet (Krizhevsky et al., 2012) get 227x227 pixels, whereas Hp images usually are of very high resolution like BCBH Hp image size is 2040 x 1536 pixels. Thus, rescaling is mandatory before feeding into DNNs, which causes loss of information (Komura & Ishikawa, 2018). Sixth, apart from testing and training data DL model required validation data. However, traditional ML-based models do not require validation data. Thus more data instances are required to train a DL-based model compared to traditional ML-based models.

However, a softmax classifier is commonly used for ANN-based DL models for BrC detection and classification using medical images. The traditional ML classification model's performance is highly dependent upon the key step i.e., extraction of handcrafted features(HcFs) (Chen, Jiang, Li, & Li, 2013; Aghdam & Heidari, 2015). HcFs extraction is a highly difficult task for medical images like BrC medical images. Because it requires domain knowledge to get discriminative features that play a vital role in the training of the BrC classification model. However, DL-based models are able to extract medical image correlated features automatically, thus very little or no human expertise (i.e., domain knowledge) is required (Smitha, Shaji, & Mini, 2011).

Thus, for BrC detection and classification DL-based correlated features are more discriminative compared to HcFs. However, traditional ML classifiers can be used after extracting the features from the DL model for BrC detection and classification using medical images like Hp images.

2.6.4 Empirical Evaluation of BrC Deep Neural Network Models Using Different Datasets

This section presents an empirical evaluation of different types of DNN on publicly available datasets. Table 2.9, shows the study-wise DNN models that have been employed on various datasets related to BrT classification. Here, the majority of the studies employed CNN instead of multi-layer NN and SNN to classify BrC. Moreover, most of the studies used MGs followed by Hp images. However, the most common datasets utilized for MG classification are DDSM, INBreast, BCDR-F03, and mini-MIAS.

Carneiro et al. (2017) developed a CNN (FTM-ARL)-based model and achieved the best performance (0.96 ± 0.05 VUS, 0.96 ± 0.05 AUC) by using the DDSM dataset for three classes (normal, benign, or malignant) of BrC. However, using the same DDSM dataset, Rouhi et al. (2015) and Leod and Verma (2016) deployed SNN and reported 0.94 AUC

and 86% Ac for a binary classification problem. These studies show that the CNN(FTM-ARL) model outperforms the SNN model using the same dataset. This finding can be attributed to the fact that CNN pre-trained models along with some new layers are more capable of learning better-generalized activations compared with shallow learning from scratch for BrT classification. Bakkouri and Afdel (2017) and Abdullah-Al et al. (2017) also used the DDSM MG dataset to distinguish between benign or malignant breast lesions. However, a former study adopted a CNN (UDM)-based model and showed a higher Ac of 97.28% compared with that obtained in a later study (i.e., 93.35%) that adopted a CNN (COM)-based model. The reason behind the success of CNN (UDM) maybe because the former study extracted the image ROIs by using Gaussian pyramids, which may enhance the model performance.

Moreover, both Dhungel et al. (2017) and Kumar, Kumar, et al. (2017) used the INBreast MG dataset to distinguish between benign and malignant breast tumors. Although both studies used CNN (COM) models, the former study reported better performance (i.e., Sn=98%, Sp=70%) than the latter study (i.e., Ac=75%, AUC=0.57). Hence, the former study performed better than the latter possibly because of the use of a small network that is more likely to be overfitted instead of a deep-layered network.

Similarly, Duraisamy and Emperumal (2017) and Arevalo et al. (2015) used the BCRDR-F03 MG dataset to classify BrC. Here, the first study used a CNN (FTM-LL)-based model, whereas the second study created a CNN(UDM) model. The first study model outperformed the second one because TL-based models usually perform better on a small number of images (BCRDR-F03 possesses only 736 images) than models trained from scratch. Similarly, Jaffar (2017) and Nahid and Kong (2018) utilized mini-MIAS MGs for two (benign/malignant) and three (normal/benign/malignant) types of BrC predictions. Moreover, the former study created a ML-NN network, whereas the latter study employed a CNN (COM) model type. However, the latter study showed better

performance (i.e., Sn=97%, Sp=100%) than the former (i.e., Sn=93.25%, Sp=90.50%). The better performance of the former study might be due to the smaller size of the network instead of using deep-layered convolutional networks, especially when dealing with a small number of images, such as the mini-MIAS dataset with only 322 images of 161 patients.

Apart from MG datasets, many studies used Hp image datasets, especially for multiclass BrT classification. In addition, the dataset was commonly used for Hp images in BreakHis followed by BCBH (Han, Kang, et al. (2017). Bardou et al. (2018) utilized the BreakHis dataset for multiclass (eight classes) BrT classification. The first study implemented a CNN(CDM) model, whereas the CNN(UDM) network was used by Bardou et al. (2018). Comparative analysis of both studies showed that the first study outperformed (Avg. Ac=93.2%, PRR=97%) the other study because of the pre-training of the newly created model using the ImageNet dataset. However, these studies improved the diagnosis of the eight subtypes of breast lesions.

Similarly, other studies (Spanhol et al., 2016a; Abdullah-Al et al., 2017; Nahid & Kong, 2017; Nejad et al., 2017; Nahid et al., 2018) employed the BreakHis dataset by using the same type of network, such as CNN (UDM), for binary classification. However, the first study showed the highest Ac of 92.19% among all the studies. The author possibly deployed many residual blocks using CNN (for global feature extraction) along with contourlet transform and histogram features (for local feature extraction).

Alongside MG or Hp image datasets for BrT classification, some studies used US (Silva et al., 2015; Cheng et al., 2016; Nascimento et al., 2016; Zhang et al., 2016; Byra et al., 2017; Han, Kang, et al., 2017; Khan, 2017), MRI (Bevilacqua et al., 2016; Amit et al., 2017; Hadad et al., 2017; Rasti et al., 2017), or more than one modality (Hadad et al., 2017). Moreover, most of the datasets used for US and MRI images are exclusive because these modalities are rarely found in publicly available datasets. Zhang et al. (2016)

employed a two-layered DBN for the extraction of features from breast US-SWE images for malignancy detection. The author narrated an Ac of 93.4% (AUC=0.94). Similarly, Nascimento et al. (2016) developed a ML-NN model to classify breast US images into benign or malignant lesions. The author reported a higher Ac of 96.98% (AUC=0.98).

Furthermore, Byra et al. (2017) employed a CNN(UDM) model by using US-based Nakagami images. This study reported 83% Ac (AUC=0.912±0.005) for binary classes of BrC. Few researchers adopted breast MRI modality (Bevilacqua et al., 2016; Amit et al., 2017; Hadad et al., 2017; Rasti et al., 2017) for cancer diagnosis using exclusive datasets. For instance, Bevilacqua et al. (2016) reported an Ac of 89.77%±5.84% for binary classes by deploying ML-NN for breast MRI classification. Similarly, Rasti et al. (2017) implemented a CNN(UDM) model from scratch for benign or malignant breast DCE-MRI classification. They reported the highest Ac of 96.39% for malignancy diagnosis.

Instead of using the single modality, the authors maximized multi-modality to train the NN model. Khan (2017) developed a CNN(COM) model by using two exclusive datasets of different modalities, such as MGs and breast MRI, to perform binary classification. However, model training was performed on MG images, whereas testing results were obtained by using breast MRI. The reported Ac was 94% (AUC=0.98) for benign and malignant classes of breast MRI images. Hence, this review shows that the fusion of multi-modalities can improve the performance of DNN models.

Table 2.9: Study-wise performance of ANNs for breast cancer detection and classification

Reference	ANN Type	Dataset	No. of Classes	Performance	Time, Resource
(Kumar, H.S, et al., 2017)	SNN	DDSM	4	Ac=79.5%	Not given
(Rouhi et al., 2015)	SNN	DDSM, MIAS	2	Avg [(Ac=86.66%, Sn=87.91%, Sp=85.40%, AUC=0.8825) (MIAS),	Not given

Reference	ANN Type	Dataset	No. of Classes	Performance	Time, Resource
				(Ac=95.01%, Sn=96.25%, Sp=93.78%, AUC=0.9499) (DDSM)]	
(Leod & Verma, 2016)	SNN	DDSM, UCI	2	Ac=86% (DDSM), Ac=89.175% (UCI)	Not given
(Feng et al., 2018)	SDAE	ED(Hp Image)	2	Ac=98.28±0.12, 90.54±0.45, Pr=97.88,90.04	6 Hrs 22 Min
(Cheng et al., 2016)	SDAE	ED(US)	2	Ac=94.4±3.2, Sn=90.8±5.3, Sp=98.1±2.2, AUC=98.4±1.5	Not given
(Wu et al., 2016)	PCA-Net	ED(Hp Image)	2	Ac=78.46±3.92, Sn=71.00±4.18, Sp=83.23±5.30	Not given
(Bevilacqua et al., 2016)	Multi-Layer NN	ED(MRI)	2	Avg Ac=89.77±5.84, Min Ac=73.08±0.43, Sn=0.89±0.10, Sp=0.90±0.09	Not given
(Nascimento et al., 2016)	Multi-Layer NN	ED(US)	2	Ac=96.98%, AUC=0.98	Not given
(Arefan et al., 2015)	Multi-Layer NN	mini-MIAS	3	Ac=97.66%	Not given
(Khan, 2017)	Multi-Layer NN	mini-MIAS, BCDR	3	Sn=97%,Sp=100% (mini-MIAS)MG, Sn=98%,Sp=97% (BCDR)MG, Sn=99%,Sp=100% (BCDR)US	Not given
(Zhang et al., 2016)	DBN	ED(US-SWE)	2	Ac=93.4%, Sn=88.6%, Sp=97.1%, AUC=0.947.	1 Hr 11 Min, GPU
(Arevalo et al., 2015)	CNN(UDM)	BCDR-F03	2	AUC=0.86.	Not given, GPU
(Araujo et al., 2017)	CNN(UDM)	BCBH	4,2	Ac=77.8% (4 classes), Ac=83.3% (2 classes), Sn=95.6%.	Not given
(Bardou et al., 2018)	CNN(UDM)	BreakHis	8,2	Ac=83.31% to 88.23% (8 Classes), Ac=96.15%, 98.33% (2 Classes).	1Hr 43 Min, GPU
(Bayramoglu et al., 2017)	CNN(UDM)	BreakHis	3	Avg Ac=80.10%, PRR=83.25%.	Not given
(Spanhol et al., 2016a)	CNN(UDM)	BreakHis	2	Ac=90.0±6.7, PRR=85.6±4.8.	3 Hrs, GPU

Reference	ANN Type	Dataset	No. of Classes	Performance	Time, Resource
(Abdullah-Al et al., 2017)	CNN(UD M)	BreakHis	2	Ac=85.36%, Sp=70.36%, Rc=91.36%, Pr=89%.	2 Hrs, GPU
(Nahid & Kong, 2017)	CNN(UD M)	BreakHis	2	Max Sp=97.18%, Max Sn=99%.	Not given
(Nejad et al., 2017)	CNN(UD M)	BreakHis	2	Ac=77.5%.	Not given
(Nahid & Kong, 2018)	CNN(UD M)	BreakHis	2	Ac=92.19%, Sn=94.94%, Rc=98.20%, Pr=98%.	6 Hrs GPU
(Nahid et al., 2018)	CNN(UD M)	BreakHis	2	Ac=91%, Pr=96%.	Not given
(Bakkouri & Afdel, 2017)	CNN(UD M)	DDSM, BCDR	2	Ac=97.28%, Sn=99.79%, Sp=94.78%	Not Given, GPU
(Kim et al., 2016)	CNN(UD M)	ED(DBT)	2	Avg AUC=0.847±0.012	Not given
(Rasti et al., 2017)	CNN(UD M)	ED(DCE-MRI)	2	Ac=96.39%, Sn=97.73%, Sp=94.87%	Not given, GPU
(Wan et al., 2017)	CNN(UD M)	ED(Hp Image)	3	Avg. Ac=69%	20 Hrs
(Cao et al., 2016)	CNN(UD M)	ED(Hp Image)	2	Ac=90%, 74%, 76%. AUC=0.93	Not given
(Xu et al., 2016)	CNN(UD M)	ED(Hp Image)	2	Ac=84.34, F1=85.21, Max AUC=0.89597	Not given
(Fonseca et al., 2015)	CNN(UD M)	ED(MG)	4	Max Ac=78.35%, Avg Ac=73.05%,	72 Hrs, CPU
(Qiu et al., 2017)	CNN(UD M)	ED(MG)	2	Avg AUC=0.790±0.019, Max AUC=0.836±0.036	Not given, GPU card
(Sun et al., 2017)	CNN(UD M)	ED(MG)	2	Ac=82.43%, AUC=0.8818	Not given
(Hadad et al., 2017)	CNN(UD M)	ED(MG, MRI)	2	Ac=94%, AUC=0.98 (MRI)	7.5 Min, GPU
(Amit et al., 2017)	CNN(UD M)	ED(MRI)	3	Ac=83%, AUC=0.91	2 Min, GPU
(Byra et al., 2017)	CNN(UD M)	ED(US, Nakagami)	2	Ac=83%, Sn=82.4, Sp=83.3, AUC=0.912±0.005	Not given
(Duraisamy & Emperumal, 2017)	CNN(FTM-LL)	BCDR-F03, MIAS	10	Ac=99%, Sn=98.75%, Sp=1.0%, AUC=0.9815	Not given, GPU
(Gandomkar et al., 2018)	CNN(FTM-LL)	BreakHis	8	Max. Ac=95.70% using one fold	Not given

Reference	ANN Type	Dataset	No. of Classes	Performance	Time, Resource
(Chang et al., 2017)	CNN(FTM-LL)	BreakHis	2	Ac=83%,89%, AUC=0.93,	Not given
(Spanhol et al., 2017)	CNN(FTM-LL)	BreakHis	2	Max Ac=84.2%, PRR=86.3%	Not Given
(Sert et al., 2017)	CNN(FTM-LL)	DDSM	2	Ac=94.1%,Pr=95%,Sn=94%	Not given
(Samala et al., 2017)	CNN(FTM-LL)	DDSM, ED(MG)	2	AUC=0.82±0.02,	Not given
(Zhang et al., 2017)	CNN(FTM-LL)	ED(MG)	2	AUC=0.73	1 Hr 10 Min, GPU
(Han, Kang, et al., 2017)	CNN(FTM-LL)	ED(US)	2	Ac=91%, Sn=0.86, Sp=93%, AUC>0.9	Not give, GPU
(Samala et al., 2018b)	CNN(FTM-ARL)	DDSM, ED(DBT)	2	ED AUC=0.90±0.4	Not given, GPU
(Carneiro et al., 2017)	CNN(FTM-ARL)	DDSM, INBreast	3,2	VUS=0.96±0.05(DDSM), 3-class, VUS=0.94±0.05 (INBreast), 3-class, AUC=0.96±0.05(DDSM), 2-class AUC=0.94±0.05 (INBreast), 2-class,	Not given
(Kumar, Kumar, et al., 2017)	CNN(CO M)	CBIS-DDSM, MIAS, INBreast	2	Ac=75%, AUC=0.57 (INBreast)	Not given
(Jaffar, 2017)	CNN(CO M)	DDSM, mini-MIAS	2	Avg Ac=93.35%, Sn=93% (DDSM), Avg Ac=92.85%, Sn=93.25% Sp=90.50%, AUC=0.92 (mini-MIAS).	Not given
(Zheng et al., 2017)	CNN(CO M)	ED(Hp Image)	15,2	Ac=96.4% (15 classes), Ac=95.9%, AUC=0.86306 (2 classes)	Not given
(Bejnordi et al., 2017)	CNN(CO M)	ED(Hp Image-WSI)	3	Ac=81.3%, AUC=0.962	Not given
(Dhungel et al., 2017)	CNN(CO M)	INBreast	2	Sn=98%, Sp=70%	Not given, CPU
(Han, Kang, et al., 2017)	CNN(CD M)	BreakHis, ImageNet	8	Avg Ac=93.2%, PRR=97%	10 Hrs 13 Min, GPU

Exclusive Dataset(ED), Accuracy(Ac), Sensitivity(Sn), Specificity(Sp), Precision(Pr), Average(Avg), Maximum(Max), Patient Recognition Rate (PRR)

2.7 Evaluation Metrics Analysis and Review

After training the DNN model followed by image preprocessing, training, and validation of BrC images, the test images are then served as input to the trained DNN model for classification to evaluate its performance. In general, the evaluation metrics are computed from the confusion matrix. In the confusion matrix, the actual (input) classes are represented with rows, whereas the column represents the predicted (output) class labels. Therefore, the BrC can be classified as true positive (TP) or true negative (TN) when correctly classified and false positive (FP) or false negative (FN) when incorrectly classified. Based on the confusion matrix, the most popularly adopted evaluation measures for BrT classification are A_c , S_n , S_p , P_r , F_m , AUC, the volume under the ROC surface (VUS) (Landgrebe & Duin, 2008), and patient recognition rate. These metrics are briefly defined in subsequent paragraphs.

2.7.1 Accuracy

The accuracy (A_c) measure represents how many of the total instances are correctly classified. It simply shows how much normal patients are correctly predicted and how many abnormal (BrC) patients are correctly diagnosed. It can be expressed by Equation (3):

$$A_c = \frac{(TP + TN)}{(TP + TN + FP + FN)}. \quad (3)$$

2.7.2 Sensitivity

The sensitivity (S_n) or recall (Re) measure indicates how much of the total positive instances are predicted correctly. In simple words, it represents how much BrC patients are correctly predicted from overall abnormal (BrC) patients. Thus, it should be as high as possible. Low S_n means many cancer patients are misdiagnosed and will be treated as normal. Hence, S_n is highly important in medical image diagnosis. It can be computed by using Equation (4):

$$Sn = \frac{TP}{(TP + FN)}. \quad (4)$$

2.7.3 Specificity

Specificity (Sp) measure shows how much of the total negative predictions are correct. It simply represents how much of the normal (BrC) prediction is correct. It should be high as possible but is less important in medical diagnosis than Sn. It can be denoted by Equation (5):

$$Sp = \frac{TN}{(TN + FP)}. \quad (5)$$

2.7.4 Precision

Precision (Pr) denotes how much of the total positive predictions are correct. It simply represents how much of the abnormal (BrC) prediction is correct. Both Sn and Pr should be high for medical image diagnosis to avoid misdiagnosis of cancerous patients. It can be calculated by Equation (6):

$$Pr = \frac{TP}{(TP + FP)}. \quad (6)$$

2.7.5 FMeasure

FMeasure (Fm) reflects the simultaneous impact of both Sn and Pr through harmonic means by applying more penalty over extreme values. It helps to compare two models with high Sn and low Pr and vice versa. It can be measured by Equation (7).

$$Fm = \frac{2 * (Pr * Sn)}{(Pr + Sn)}. \quad (7)$$

2.7.6 Area Under the ROC Curve

A receiver operating characteristic curve (ROC) plots the curve of precision against sensitivity. The area under the ROC curve (AUC) is a common evaluation measure that helps to choose optimal models and ignore suboptimal ones (Figure 2.18 (a)), showing

the performance comparison of four classification models for BrC. The figure shows that model-1 outperforms the three other models. By contrast, model-4 shows the lowest performance. The AUC value can be computed by using Equation (8). An AUC value lies between 1 and 0. However, an AUC value of 1 represents a perfect model and an area of 0.5 or below reflects an ineffective model.

$$AUC = \frac{\sum_i R_i(I_p) - I_p(I_p + 1)/2}{I_p + I_n} \quad (8)$$

where I_p and I_n denote the number of positive and negative BrC images, respectively, and R_i is the rank of the i^{th} positive image in the ranked list.

2.7.7 The Volume Under the ROC Surface

The ROC is a standard tool to evaluate two-class classification problems. It was extended and enabled to evaluate multiclass problems named VUS (for three class VUSs, see Figure 2.18(b)) (He & Frey, 2008). Furthermore, in multi-classes, the independent and dependent (of the same type) classes are grouped, and many ROCs are created. Finally, the decomposed ROCs are interrogated by using cost-sensitive and Neyman–Pearson optimization along with volume under the curve (Ferri, Hernández-Orallo, & Salido, 2003).

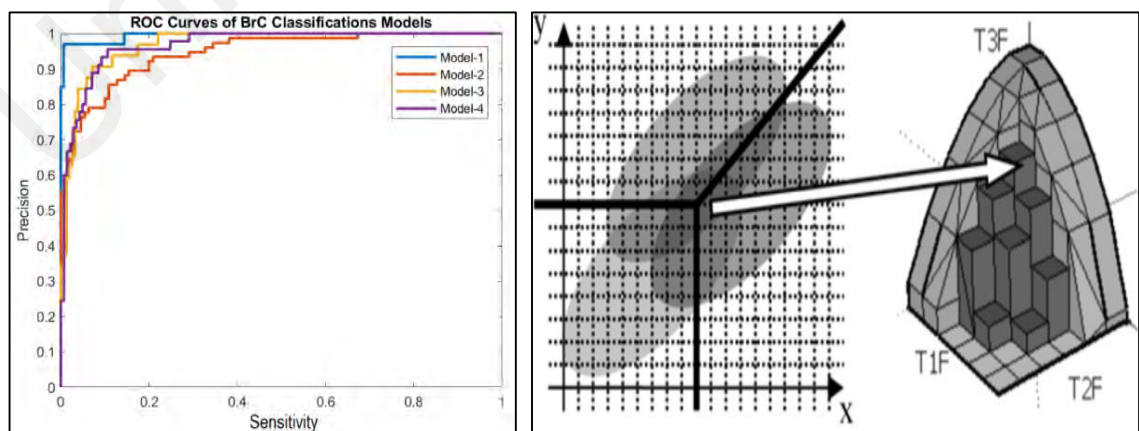


Figure 2.18: (a) A sample ROC diagram, comparing the performance of four classification models of breast cancer. (b) Illustration of sample VUS diagram for three classes

2.7.8 Patient Recognition Rate

It is required to perform a patient-level analysis. The accuracy of a classifier, which is required to decide whether or not the patient is cancerous. Formally, the overall patient recognition rate (PRR) of a classifier is denoted by (Spanhol et al., 2016b) as follows:

$$PRR = \frac{\sum \text{Patient score}}{\text{Total number of patients}}. \quad (9)$$

Where the patient score is calculated by the following:

$$\text{Patient score} = \frac{N_{\text{rec}}}{N_P}. \quad (10)$$

Here, N_P represents the total number of cancer images of patient P, and N_{rec} is the number of images correctly classified for patient P.

2.7.9 Cross-entropy Loss

Cross-entropy loss is used in the classification model to compute the probabilities to measure the performance. If the loss is high means the predicted label is diverging from the actual label. Thus, a perfect prediction is found when the loss is zero. It can be calculated by Equation (11).

$$\text{CrossEntropy} = -\sum_{c=1}^M y_{o,c} \log(p_{o,c}). \quad (11)$$

Where, M , y , o , and p represents the number of classes, binary value indicator(0,1), correct classification for observation c , observation, and predicted probability respectively.

In AlexNet based model, CrossEntropy loss is used to optimize the classification model. It is calculated on training and validation sets. It represents how well the model is doing in these two sets. It is the sum of errors made for each example in training or validation sets termed as training loss or validation loss. Training/validation loss value implies how poorly or well a model behaves after each iteration of optimization. Whereas, validation accuracy metric is used to measure the algorithm's performance in an interpretable way. The accuracy of a model is usually determined after the model parameters and is

calculated in the form of a percentage. It is the measure of how accurate your model's prediction is compared to the true data.

Table 2.10 shows the frequency of studies that used particular performance evaluation measures to compute the performance of BrT classification models. The majority of the studies evaluated the performance by employing the Ac metric. Moreover, studies (Spanhol et al., 2016a; Bayramoglu et al., 2017; Han, Kang, et al., 2017; Nejad et al., 2017; Spanhol et al., 2017; Bardou et al., 2018; Gandomkar et al., 2018) calculated Ac at four magnifications (40×, 100×, 200×, and 400×) based on two levels, such as Ac at the image level and Ac at the patient-level, by using BreakHis Hp images. However, the patient-level Ac (i.e., patient recognition rate) is more important than the image-level Ac in medical science (Spanhol et al., 2017). For instance, previous studies (Spanhol et al., 2016a; Bayramoglu et al., 2017; Han, Kang, et al., 2017; Spanhol et al., 2017; Gandomkar et al., 2018) used the BreakHis dataset and showed Ac at the four magnifications at both levels.

Table 2.10: Frequency count of performance metrics used in each selected primary study

Study Reference	Performance Metrics
(Arefan et al., 2015; Fonseca et al., 2015; Leod & Verma, 2016; Spanhol et al., 2016a; Bayramoglu et al., 2017; Hadad et al., 2017; Han, Kang, et al., 2017; Kumar, H.S, et al., 2017; Nejad et al., 2017; Spanhol et al., 2017; Bardou et al., 2018; Gandomkar et al., 2018)	Ac
(Cao et al., 2016; Nascimento et al., 2016; Amit et al., 2017; Bejnordi et al., 2017; Chang et al., 2017; Kumar, Kumar, et al., 2017; Sun et al., 2017; Wan et al., 2017; Zheng et al., 2017)	Ac, AUC
(Rouhi et al., 2015; Cheng et al., 2016; Zhang et al., 2016; Byra et al., 2017; Han, Kang, et al., 2017; Jaffar, 2017)	Ac, Sn, Sp, AUC
(Arevalo et al., 2015; Kim et al., 2016; Xu et al., 2016; Samala et al., 2017; Zhang et al., 2017; Samala et al., 2018b)	AUC

Study Reference	Performance Metrics
(Bevilacqua et al., 2016; Wu et al., 2016; Bakkouri & Afdel, 2017; Duraisamy & Emperumal, 2017; Rasti et al., 2017)	Ac, Sn, Sp
(Sert et al., 2017; Nahid & Kong, 2018)	Ac, Sn, Pr, Fm
(Carneiro et al., 2017)	AUC, VUS
(Khan, 2017; Nahid & Kong, 2017)	Sn, Sp
(Nahid et al., 2018)	Ac, Pr
(Feng et al., 2018)	Ac, Pr, Fm
(Araujo et al., 2017)	Ac, Sn
(Abdullah-Al et al., 2017)	Ac, Sn, Sp, Pr, Fm
(Dhungel et al., 2017)	Sn
(Qiu et al., 2017)	Sn, Sp, AUC
(Spanhol et al., 2016a; Bayramoglu et al., 2017; Han, Kang, et al., 2017; Spanhol et al., 2017; Gandomkar et al., 2018)	Ac, PRR

The second-highest number of studies used Ac along with AUC. The AUC evaluation measure is usually adopted to analyze the behavior of a model for each class (or for multiple model performance comparison). It reveals the authenticity of the overall predicted Ac and whether a model is biased toward any particular class. However, the studies that created multiple deep CNN de-novo models used exclusive datasets and tried to solve the multiclass BrC problem by reporting the AUC along with Ac to ensure that the newly trained model is unbiased and better than other models. For instance, studies (Amit et al., 2017; Bejnordi et al., 2017; Wan et al., 2017; Zheng et al., 2017) developed de-novo models to classify BrC into more than two classes.

Furthermore, few studies reported either the AUC or AUC along with Ac, Sn, and Sp. Whereas, studies (Cheng et al., 2016; Zhang et al., 2016; Byra et al., 2017; Han, Kang, et al., 2017) used exclusive datasets of breast US images considered the Ac, Sn, Sp, and AUC metrics to test the performance of trained CNN models before deploying their commercial usage. Apart from some basic evaluation measures, few studies used more sophisticated evaluation measures, such as Fm and VUS, for multiclass BrT classification. For instance, Carneiro et al. (2017) used the VUS metric to show the

performance of a TL-based CNN model for three classes of BrC using the INBreast and DDSM datasets. Furthermore, some recent studies (Abdullah-Al et al., 2017; Sert et al., 2017; Feng et al., 2018; Nahid & Kong, 2018) have reported Fm with few other evaluation metrics, such as Ac, Sn, Sp, and Rc.

2.8 Limitations Related to the Existing Literature

This section presents the limitations identified in the review literature. In specific, the current research enhanced the confidence level to make better decisions for BrC image analysis in three aspects, namely, the creation of DL-based models for better feature extraction, performance enhancement by reducing misclassification, and the performance metrics utilized to compare the results.

2.8.1 Limitations of Artificial Neural Networks Based Models

This review identified two major types of artificial neural networks (ANNs), such as SNNs and DNNs (i.e., DL-based), for BrC detection and classification. However, few researchers employed SNNs because their simple network can learn tasks better for both practical and theoretical reasons. In addition, they require less training time, computational power, and memory to store intermediate computational results (e.g., weights). Thus, they can be implemented economically with ease by using a normal desktop machine. Moreover, SNNs can show better generalization performance on a small amount of data than DNNs. SNNs also provide quicker responses than DNNs at the time of testing as required in real-time. However, using SNNs has some limitations. For instance, they may not show better performance on high-dimensional data such as Hp BrC images. Usually, SNNs use structured data; hence, their performance depends on the designed features and the number of neurons used in hidden layers.

Therefore, to avoid the limitations of SNNs, most researchers employed DL-based approaches for BrC detection and classification. This review indicates that DL-based

approaches are based on either a multilayer neural network (ML-NN) or CNN. In ML-NN, the increased number of hidden layers is supported to improve the generalization performance for BrC image detection and classification. However, it requires a larger number of images compared with SNNs. Furthermore, the performance of the network depends on the optimization of parameters, the number of hidden layers used, and the number of neurons per layer employed in the creation of ML-NN. Such a type of network is difficult to optimize, especially in the BrT classification like Hp images.

Alternatively, the majority of researchers used CNN-based approaches to deal with high-dimensional data for BrT classification. CNN approaches used by researchers are often of two types: the establishment of a de-novo model that is trained from scratch or the adoption of a pre-trained model also known as the TL-based model. However, the majority of DL-based models are based on CNN de-novo models because de-novo models are created and optimized according to the size, nature, and type of specific data, such as BrC images. Hence, a small CNN de-novo model can produce better BrT classification results if designed and trained with proper optimization (Goceri & Gooya, 2018). Conversely, employment and training of deeper layers on a small amount of data may face more overfitting issues. Furthermore, de-novo training parameter optimization is difficult and can be achieved by trial-and-error methods. Hence, multiple models may be created and trained, which mostly requires a long time and very high computational resources like GPU. Therefore, to overcome de-novo CNN training issues, many researchers deployed pre-trained models, such as AlexNet. These models are already trained on millions of nonmedical images (natural images) to classify ten hundred natural objects, such as a pen, a tree, and a cap. Moreover, TL-based models are retrained on medical images after fine-tuning. Fine-tuning may involve the removal of the last layers, the use of a small learning rate, and freezing the weights of the first few layers (i.e., ensemble models). The analysis of selected studies reveals that the TL-based models

show comparable performance while using a small number of medical images. Apart from de-novo and TL-based models ensemble models were also used where two or more new layers are added and trained from scratch. However, TL-based and ensemble models can be trained without using high-computational resources, such as GPU in a reasonable time. Whereas, if the dataset is too small (like less than 1000 images), then the pre-trained network may lead to an overfitting issue and will not be able to learn new features properly. Therefore, researchers usually performed image augmentation (rotation, translation, and flipping) to increase the number of images.

In summary, the “no free lunch” theorem of Wolpert and Macready (1997) inferred that no single ML classifiers perform optimally in all domains. Hence, a variety of DL-based techniques should be employed to evaluate which algorithm outperforms on a specific type of data, such as Hp BrC images. The selected primary studies implemented their own customized data set and different experimental setups. Thus, statistically comparing the performance values across the studies is infeasible. Nonetheless, a comparison of the performance of different studies shows that the CNN model outperforms among DL-based models for BrC detection and classification.

2.8.2 Limitations of Performance Evaluation Metrics

This review also reveals that most of the researches used Ac as a primary PEM for BrC detection and classification model comparison. However, the Ac metric can be biased towards a particular class (Powers, 2011). Thus, apart from Ac, there is a need to measure and compare the other PEMs like Sn, Sp, Fm, AUC, and PRR. Because, Sn is a highly important metric in medical science to show the misclassification results for the diagnosis of a cancerous patient i.e., malignant. Whereas, Fm and AUC are also important to show that the BrC detection and classification models are properly trained and able to show unbiased results for multiple classes for BrT. Furthermore, in medical science PRR is more important to detect either the patient is malignant or benign and it may show

different results than image-level classification (Spanhol et al., 2017). Thus appropriate metrics should be used to accurately measure and compare the performance of BrC detection and classification models.

2.8.3 Low Model Performance (i.e., Higher Misclassification)

Generally, it has been observed in the aforementioned DL-based models of BrC detection and classification that the results were compromised due to a higher number of false negative and false positive predictions, also known as false predictions or simply misclassification. Whereas, misclassification using BrC images for multi-subtypes (more than two) of BrT maybe because of three reasons. First, there is a high correlation among the features of many subtypes of BrT images. Which may create complexity (i.e., low interclass similarity and low intraclass dissimilarity) for the classifier to differentiate among multiple subtypes of BrT. Therefore, the misclassification rate can be higher and the model can show compromised accuracy.

Second, a large number of features were extracted through DL-based models. Such a large number of saturated features can easily distract the training process of a classifier that can lead to an increase in false predictions/misclassification rate. Third, the DL-based models were normally trained using augmented images along with original images. Whereas the quantity of augmented images is huge than the number of original images, therefore the model may get better training for augmented images instead of original images. However, testing data contains only original images, thus it can be easily misclassified by the model which was largely trained on augmented images. Thus there is a need to develop a robust algorithm for the reduction of misclassification rate to enhance the model performance.

2.9 Research Gap Analysis for Problem Identification

This review of existing literature revealed that many studies have employed DL-based BrC detection and classification models by using Hp images (Cao et al., 2016; Spanhol

et al., 2016a; Xu et al., 2016; Abdullah-Al et al., 2017; Araujo et al., 2017; Bayramoglu et al., 2017; Bejnordi et al., 2017; Chang et al., 2017; Han, Wei, et al., 2017; Nejad et al., 2017; Spanhol et al., 2017; Wan et al., 2017; Zheng et al., 2017; Bardou et al., 2018; Gandomkar et al., 2018; Nahid & Kong, 2018; Nahid et al., 2018). Because Hp image is a standard medical imaging modality used to diagnose BrC more confidently compared to any other type of modalities like Mg, US, MRI, PET, and CT images. Most of the aforementioned studies utilized high computational resources and longer training time to get better results to perform BrC detection and classification. For instance for BrC detection using Hp images Spanhol et al. (2016a) achieved better average Ac (i.e., 90%) and PRR (i.e., 85.6) using GPU for three hours. Similarly, Nahid and Kong (2018) performed trained for six hours using a GPU to achieve 92.19% Ac. Moreover, Wan et al. (2017) trained a model using GPU for twenty hours for BrC grading like low, medium, or high. Moreover, very few studies have computed PRR for patient-level BrC detection. Which is highly important in medical science to diagnose a patient as benign or malignant instead of just image-level BrC detection.

On the other hand, to solve the BrT classification problem very few studies reported training time and resources for their developed models. For instance, Araujo et al. (2017) developed a CNN-based model and trained from scratch to classify four types of BrT. The author reported low Ac is 77.8% but model training time and resources were not discussed. Moreover, Bardou et al. (2018) created a CNN-based model and trained from scratch using GPU for 1 hour 43 minutes. However, the reported Ac (i.e., 88.31%) was better by using BreakHis Hp images to classify eight types of BrT. Similarly, Wan et al. (2017) developed a CNN-based model and performed 20 hours of training from scratch by using GPU. The reported average accuracy was very low at 69% to perform classification for three classes of BrT. Moreover, Han, Wei, et al. (2017) developed a model from scratch and performed pre-training on ImageNet (a large dataset of natural

images). Afterward model was retrained on GPU for 10 hours 13 minutes using BreakHis images to solve the BrT classification problem. The author reported better Ac (i.e., 93.2%) and PRR (i.e., 96.15%).

Thus it can be concluded from an extensive literature review that DL-based models usually required very high computational resources and longer training time. Moreover, most of the studies reported accuracy and very few studies used other PEMs. Thus, to show results reliability, apart from Ac other PEMs should be computed like PRR, Sn, Sp, Fm, and AUC. PRR is required for patient-level diagnosis for BrC detection. However, Sn, Fm, and AUC metrics are also important to show image-level diagnosis. Thus, PEMs are required to show that the BrC detection and classification models were trained properly and able to produce unbiased results, especially when dealing with multiple classes of BrT using Hp images.

2.10 Summary

This chapter presented a critical analysis of BrC detection and classification by analyzing collectively the major research endeavors presented by current scholars to assist the new researchers in this domain. Many academic studies were carefully selected from eight unique academic repositories. The review was performed based on selected primary studies from five aspects, namely, various medical imaging modalities exploited, datasets used, image preprocessing techniques, types of ANNs (including deep neural networks), and PEMs used to construct and evaluate the BrC detection and classification models. In BrC detection and classification, various types of public and exclusive datasets were used. However, exclusive datasets are usually smaller in size than public datasets. Thus, more researchers preferred to use public datasets over exclusive ones. Whereas, public datasets that contain multimodality images of the same patient along with some other information, such as DNA sequence, are urgently needed. Such a type of dataset can help reduce FPs using automated systems. Furthermore, among all the datasets, MG and Hp imaging

modalities were widely adopted, followed by US images, and very few used MRI and CT breast images. Thus, other modalities (e.g., PET, CT, and thermal images) that may provide different types of lesion characteristics should be explored to improve BrC detection and classification results. Furthermore, in preprocessing tasks, image augmentation, scaling, image intensity/contrast normalization, stain normalization, and stain removal techniques were mostly adopted to remove image inconsistencies before feeding to any DL-based model. However, preprocessing techniques should be adopted carefully so that important information, such as lesion texture-, shape-, and illumination-based information, can be preserved. In this review, several types of DNN architecture were identified to detect and classify BrC. Among these, CNN was the most popular choice of researchers for BrC detection and classification. Of these CNN-based models, de-novo, TL-based, and ensemble models were employed by the researchers, and results showed that de-novo models showed better results but consume high computational resources and training time. By contrast, pre-trained models were also tested and achieved comparable results on smaller datasets after fine-tuning using augmented images for BrC detection and multiclass classification. Moreover, ensemble models also showed better results. However, the TL-based and ensemble models require less computational resources and training time. In addition, such types of models can show better results on a small number of images. Therefore, most of the studies adopted either TL-based or ensemble type of models for BrC detection and classification compared to de-novo models. Nonetheless, DL-based models usually show compromised results due to a higher misclassification rate while using BrC Hp images. To evaluate the DL-based models, various performance metrics are used, such as Ac, Sn, Sp, Fm, AUC, and PRR. Among these, the first three are more common and essential in medical image classification for image-level analysis. However, mostly Ac is reported for baseline comparison. Apart from Ac, Sn, Fm, and AUC are also important metrics, needed to show

that results are reliable and unbiased. While PRR is needed for patient-level analysis. Finally, this review enabled us to find out the research gaps that require extensive efforts to improve DL-based BrC detection and classification models. Thus, it was revealed from the extensive review that there is a need to develop an efficient (i.e., consume fewer computational resources and training time) and reliable (i.e., reduce misclassification to show better and unbiased results even using complex dataset) BrC detection and classification models for early diagnosis of breast cancer to assist doctors to serve as the second opinion in any health care institution.

Universiti Malaya

CHAPTER 3: METHODOLOGY AND EXPERIMENTAL SETUP

3.1 Introduction

This chapter comprises two major divisions namely methodology and experimental setup. In the methodology division, the detailed procedure implemented for the development of both BrC detection and BrT classification models is elaborated. Whereas, in the experimental setup division, the organization of experimental steps followed for BrC detection and BrT classification are reported in detail. Section 3.2 covers the overall research methodology, Section 3.3 demonstrates the entire experimental setup and Section 3.4 gives a summary of this chapter.

3.2 Methodology

This section reports the overall research methodology implemented in this research work. A brief discussion of the research methodology is already presented in Section 1.6. The problem identification is made through the literature review and a detailed discussion is made in Chapter 2. This section presents the overall research methodology in detail to develop the proposed BrC detection (i.e., the subject is benign or malignant) model and BrT classification (i.e., eight subtypes of BrT) model using Hp images.

3.2.1 Breast Cancer Detection Model Construction Methodology

This section elaborates on the methodology (see Figure 3.1) employed to develop a BrC detection model using BreakHis dataset Hp images. The entire methodology for BrC detection is composed of five stages, namely, data collection, image preprocessing, DL-based model development and DeCAFs extraction, BrC detection techniques, model construction, and performance evaluation, see Figure 3.1.

In the data collection stage, the publicly available BreakHis dataset is used. First, all images are divided into benign and malignant categories to perform BrC detection. Overall 82 patients' images of 40× magnification are utilized for further experiments.

Furthermore, the patient-wise BreakHis images are split into training, validation, and testing sets via random sampling. In the image preprocessing stage, few essential tasks are conducted, such as stain normalization, training set augmentation, and selection of equal numbers of augmented training set images for each class and image rescaling. Hence, a comprehensive training set is created by combining the class-wise balanced augmented images and original (i.e., nonaugmented) images.

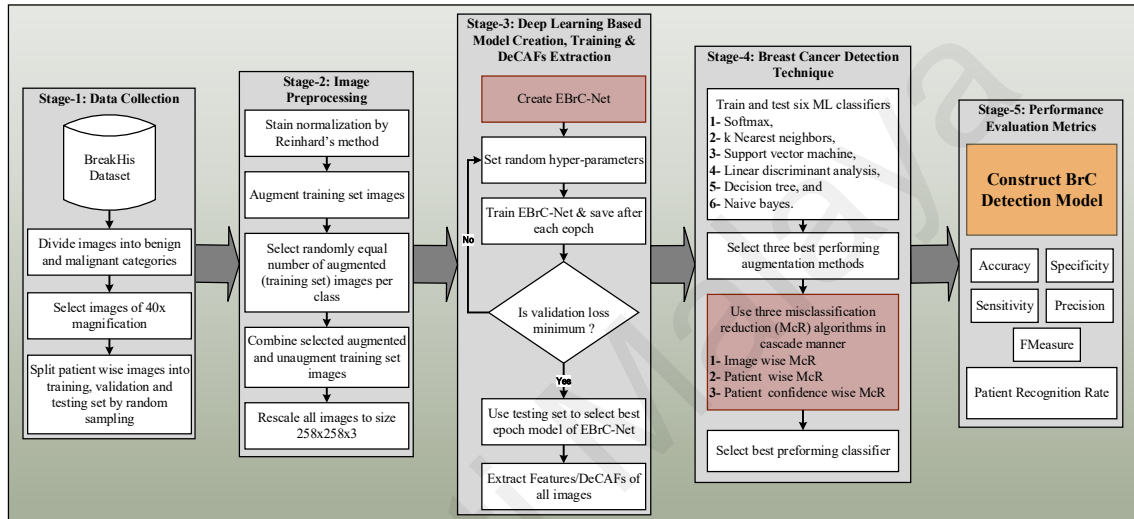


Figure 3.1: BrC detection model construction methodology

In the third stage, ensembled BrC network (EBrC-Net) architecture is created and trained by using random hyper-parameters. The trained EBrC-Net is achieved after using multiple random hyper-parameters via the trial-and-error method. The model is selected when the lowest validation loss has been observed. Ultimately, the DeCAFs of all images are extracted by using the trained EBrC-Net model for further analyses.

In stage four, six ML classifiers (i.e., softmax, kNN, NB, SVM, LDA, and DT) are analyzed using five folds of extracted DeCAFs. Thereafter, the three best augmentation methods are chosen by using Ac, Sn, and Sp metrics to enhance the performance of classifiers. Three McR algorithms (i.e., McRI, McRP, and McRC) are employed to improve the classification performance. Where the McRI algorithm reduces misclassification in image-wise fashion and computes image-wise confidence. The McRP

algorithm further reduces misclassification using multiple images of a patient and computes patient-wise confidence, whereas McRC minimizes misclassification using the average of image-wise and patient-wise confidence. The average of image-wise and patient-wise confidence ensures that if the majority number of one patients' images are cancerous then the patient is determined as cancerous.

In the last stage, performances of the aforementioned six ML classifiers are evaluated on the basis of six PEMs namely Ac, Sn, Sp, Pr, Fm, and PRR. Where PRR shows the patient-level while the rest are representing the image-level performance of the model. Finally, the best classifier is selected to construct the BrC detection model for both image-level and patient-level for BrC detection using the BreakHis dataset, see Figure 3.1. All five stages of the overall research methodology for BrC detection are described in detail in the following sections.

3.2.1.1 Data Collection

This research used the publicly available corpus Breast Cancer Hp Image Classification (BreakHis) (Spanhol et al., 2016b). BreakHis dataset was created by a collaboration of P&D Laboratory and Pathological Anatomy and Cytopathology, Parana Brazil. This dataset was gathered by taking samples through an excisional biopsy of breast tumor tissue from 82 subjects. Each patient possesses many Hp biopsy images. Therefore, the overall dataset consists of 7909 images captured through a microscope with four magnifications: 40×, 100×, 200×, and 400×, see Table 3.1. However, in this research 40x images are used because they have shown the best results in dataset host experiments (Spanhol et al., 2016a). The images are 8-bit RGB of size 700×460 pixels. All patients' images are categorized as either benign or malignant. A benign tumor is a usually noninvasive (non-cancerous) type of tumor; thus, it is localized and the lesion grows gradually. In contrast, malignant is an invasive (cancerous) tumor, spreads farther to other

body parts, and abolishes adjacent structures, leading to an abnormal death. The BreakHis contains 2480 benign and 5429 malignant Hp images, see Table 3.1.

Table 3.1: BreakHis dataset images distribution

BrT types	Magnifications	40×	100×	200×	400×	Total images	Total patients
	BrT subtypes						
Benign	A	114	113	111	106	444	4
	F	253	260	264	237	1014	10
	PT	149	150	140	130	569	3
	TA	109	121	108	115	453	7
	Benign total	625	644	623	588	2480	24
Malignant	DC	864	903	896	788	3451	38
	LC	156	170	163	137	626	5
	MC	205	222	196	169	792	9
	PC	145	142	135	138	560	6
	Malignant total	1370	1437	1390	1232	5429	58
Total images		1995	2081	2013	1820	7909	82

Moreover, BreakHis BrC images are further divided into eight subtypes of BrT. For instance, benign tumor subtypes are adenosis (A), fibroadenoma (F), tubular adenoma (TA), and Phyllodes tumor (PT) whereas malignant tumor is divided into ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC) subtypes. The overall dataset (including a borderline patient images) has been split according to the protocol defined by the dataset host into training (50%), validation (20%), and testing (30%), see Table 3.2. Noticeably in Table 3.2, the patient-wise split is performed to ensure that the images of a patient are not shared among the training, validation, and/or testing sets. Thus patient-level BrC detection required a patient-wise split of dataset images.

Table 3.2: Patient-wise split of BreakHis (40× magnification) dataset

BrT types	BrT subtypes	Training set (50%)	Validation set (20%)	Testing set (30%)
		Images (patients)	Images (patients)	Images (patients)
Benign	A	64 (2)	35 (01)	15 (01)

BrT types	BrT subtypes	Training set (50%)	Validation set (20%)	Testing set (30%)
		Images (patients)	Images (patients)	Images (patients)
	F	117 (5)	49 (02)	87 (03)
	PT	58 (01)	38 (01)	13 (01)
	TA	90 (03)	32 (02)	27 (02)
	Benign total	329 (11)	154 (06)	142 (07)
Malignant	DC	458 (20)	105 (06)	269 (04)
	LC	51 (03)	22 (01)	83 (02)
	MC	84 (04)	75 (02)	46 (03)
	PC	88 (03)	21 (01)	36 (02)
	Malignant total	681 (30)	223 (10)	434 (11)
Total (40x magnification)		1010 (41)	377 (16)	576 (25)

3.2.1.2 Image Preprocessing

Image preprocessing is required to enhance the image quality before performing model training like stain normalization, image augmentation, and rescaling. In general, the Hp image requires stain normalization to normalize the image inconsistencies. Whereas, DL-based models require image augmentation to create data samples in larger quantity to avoid overfitting. However, rescaling of images is needed before feeding the images into a DL-based model.

(a) *Stain Normalization*

Hp images mostly exhibit extremely high color inconsistencies when prepared in a pathology laboratory (lab). These inconsistencies may occur due to the use of different chemicals for staining, the concentration of colors (due to hematoxylin and eosin staining of Hp images), and the use of different scanners from numerous vendors. Given these factors, the Hp images of two patients, although prepared in the same digital pathology lab, may vary in color, intensity, brightness, and contrast. The high variation among images of two patients of the same cancer type may lead to improper training of the proposed DL model. Thus, to eliminate these image inconsistencies, stain normalization is required. In this study, Reinhard's method (Reinhard et al., 2001) is applied for stain

normalization. Reinhard's' stain normalization preserves the structure of cancer lesions better in comparison with other methods, such as Khans' method (Khan et al., 2014) and the Macenko method (Macenko et al., 2009). It uniformes the color, brightness, intensity, and contrast of all images of all patients by using a reference image (Figure 3.2) and used by some studies (Alsubaie, Trahearn, Raza, Snead, & Rajpoot, 2017; Chen et al., 2017; Rasti et al., 2017; Gandomkar et al., 2018). Hence, it supports the DL model in the training process to learn superior generalized features from BrC Hp images.

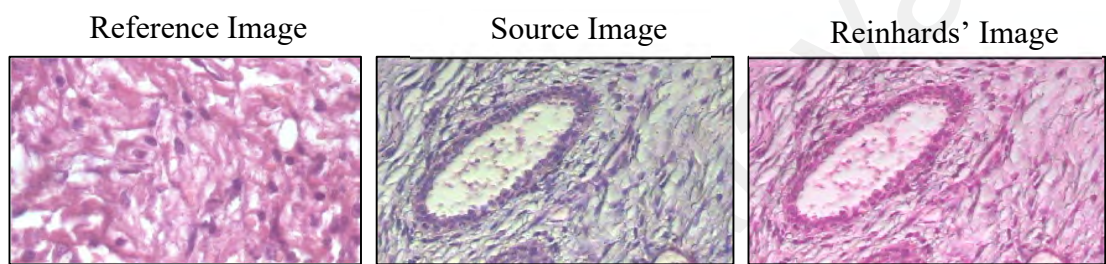


Figure 3.2: Reinhard method used to normalize source image through reference image

(b) Image Augmentation

Data augmentation is required to train the DL-based model properly to avoid overfitting issue, especially for small numbers of images (Shorten & Khoshgoftaar, 2019). Therefore, it is needed to enhance training accuracy for medical images, which are oftentimes not available in large quantities. Images are augmented typically to create more images from original ones, i.e., oversampling. Image augmentation involves basic image processing techniques, such as image rotation, flipping, shifting, rescaling, shearing, and padding. Moreover, by combining two or more of these image augmentation techniques, many new artificial images can be created from the original image. Many studies (Arevalo et al., 2015; Araujo et al., 2017; Bayramoglu et al., 2017; Carneiro et al., 2017; Hadad et al., 2017; Jiang, Liu, Yu, & Xie, 2017; Zheng et al., 2017; Bardou et al., 2018) implemented image augmentation method to train the model properly. In this research, only the training set is augmented through the aforementioned image processing techniques to train EBrC-Net. Each image is augmented 24 times by using rotation 90°

rotation, translation by a fifth of the original image size, image shearing with four affine transforms, vertical and horizontal flipping, and four times of image padding, see Algorithm 3.1. Furthermore, to avoid the class imbalance problem, equal numbers of the augmented images are randomly selected from each class. However, the original training images are used along with the randomly selected augmented training images, see Table 3.3. Therefore, the training set has a total number of 11210 (5429 benign + 5781 malignant) images of 41 patients, see Table 3.4. Thus, the following goals are achieved by adopting the overall image augmentation process:

1. Creating a large training set that is sufficient for proper training of EBrC-Net;
2. Acquisition of a balanced number of images (class-wise) that will help avoid overfitting during model training and thus enable the proposed DL-based EBrC-Net model to show classification results with improved quality and reliability using a small number of images.

Table 3.3: Augmented training set distribution using BreakHis (40× magnification) dataset

BrT types	BrT subtypes	Original images (A)	Total augmented images (B)	Overall training set A + Min(B)
Benign	A	64	1600	1339
	F	117	2959	1392
	PT	58	1450	1333
	TA	90	2250	1365
	Benign total	329	8259	5429
Malignant	DC	458	11450	1733
	LC	51	1275	1326
	MC	84	2100	1359
	PC	88	2200	1363
	Malignant total	681	17025	5781
Total (40x magnification) images		1010	25284	11210

Table 3.4: Utilized BreakHis (40× magnification) dataset distribution

BrT types	Augmented training set Images(patients)	Validation set Images(patients)	Testing set Images(patients)
Benign	5429 (11)	154 (06)	142 (07)
Malignant	5781 (30)	223 (10)	434 (18)
Total	11210 (41)	377 (16)	576 (25)

(c) *Image Rescaling*

The original image size of the BreakHis dataset is too large ($700 \times 460 \times 3$) to fit into the input layer size (i.e., $258 \times 258 \times 3$) of EBrC-Net. Moreover, the image size of the training set has been changed arbitrarily when the aforementioned basic image processing techniques are applied for image augmentation. Hence, all images are rescaled to a size of $258 \times 258 \times 3$ by using the bicubic interpolation method before serving as input to EBrC-Net.

3.2.1.3 Development of BrC Detection Model

Transfer learning is adopting knowledge from other domains to the target domain (Lu et al., 2015). By the definition, the TL-based models possess the same input size (Brownlee, 2020), and no need to be trained from scratch (Tan et al., 2018) except the softmax layer for the target class labels. Thus, fine-tuning is required for the last layer only (i.e., softmax layer) by keeping the rest of the layers freezed and it does not allow to change the input image size. Here, the low-level features (extracted via convolutional layers) and high-level features (extracted via fully connected layers) are extracted from the TL-based model without being trained on target (i.e., medical) images. However, the proposed model EBrC-Net is using TL for convolution layers only, while all fully connected layers are trained from scratch like a de-novo model. In addition, the input layer size is optimized and increased to 258×258 , which is only possible in de-novo model. All the fully connected are trained from scratch to learn the BrC lesion specific feature. Thus, the proposed model is named as ensemble model for two types of

conventions i.e. TL model and de-novo (i.e. layers trained from scratch) model, for further details see section 4.2.2.1.

The ensembling in EBrC-Net will get two advantages, first fully connected layer (which are trained from scratch) will be able to learn domain specific feature from Hp images. Whereas, due to TL-based convolution layers, the proposed ensemble model will be trained using less computational resources in less time with less number of images compared to any de-novo model of similar architecture. Moreover, EBrC-Net possesses the same architecture as AlexNet (Krizhevsky et al., 2012) for a fair comparison of results to show that the proposed model is able to learn better features than pre-trained AlexNet. The details of AlexNet and the proposed DL-based BrT detection model are given in Chapter 4, Section 4.2.

(a) Training and Feature Extraction through EBrC-Net Model

This study used the EBrC-Net model to extract the discriminative features compared to AlexNet from preprocessed Hp BrC images. Several experiments are executed to obtain the optimum results by adjusting a few training options which can be easily used to train EBrC-Net by using a normal desktop machine. The proposed model is trained by using a gradient descent solver with a momentum of 0.9. The other parameter adjustments are as follows: maximum epochs set of 30, mini-batch size is taken as 64, the initial learning rate is 0.001, L2 regularization is set to 0.0001, validation frequency is 50, validation patients are used as 5, learning drop rate is 0.1, and learning drop period is set to 2. The EBrC-Net learning rate is set lower than the AlexNet initial learning rate so that the low-level features of AlexNet CLs are not completely lost while retraining. Furthermore, the model is forced to stop training if validation loss is not reducing in fifteen successive validation iterations. Hence, EBrC-Net is devised in such a way to avoid an overfitting issue using a small number of images than AlexNet. The

aforementioned parameters enabled EBrC-Net to extract more discriminative generalized features than AlexNet. Apart from network training, according to dataset host protocol for BrC detection, a random subsampling approach is adopted for the selection of training, validation, and testing set images, where 50%, 20%, and 30% patient-wise images are used for training, validation, and testing, respectively, see Table 3.4. The validation images are known as seen data and taken separately from testing images because the analysis of the EBrC-Net trained model on unseen data (i.e., testing images) ensures the reliability of performance. Hence, in real-life, the EBrC-Net model can be implemented more confidently, as it has been tested on unseen data. Furthermore, Ac, Sp, Sn, Pr, Fm, and PRR are calculated for testing image analysis. Finally, DeCAFs of training images are extracted from the seventh layer of EBrC-Net. The DeCAFs of all training images are extracted to form a master feature vector (MFV) table. Each testing image MFV consists of 4096 unique features. Furthermore, MVF is divided into five folds to train and test the aforementioned six ML classifiers for further analyses of BrC Hp image detection.

3.2.1.4 Breast Cancer Detection Techniques

This section explains the further experiments performed using five folds of extracted DeCAFs using trained EBrC-Net to enhance the performance of the BrC detection model. Where, six ML classifiers, namely, the softmax, kNN ($k = 1, 3, 5, 7, \text{ and } 9$), SVM (linear, rbf, and polynomial), NB, DT, and LDA, are evaluated by using five folds of DeCAFs of BreakHis dataset. These six ML classifiers are selected on the basis of the literature review as discussed in Chapter 2 (Section 2.6.1). Moreover, according to the no free lunch theorem for optimization (Wolpert & Macready, 1997), not one ML classifier can perform persistently better on all types of data. Therefore, many classifiers are evaluated on the BreakHis dataset to investigate the performance of each classifier individually. Moreover, six performance metric evaluations are computed for six classifiers that enabled the achievement of one best performing classifier for BrC detection. Finally, to

enhance the performance of ML classifier three misclassification reduction algorithms (McR) are developed and implemented for BrC detection. However, before implementing using McR the best augmentation methods are selected to get better results for BrC detection. The details of proposed McR algorithms are discussed in Chapter 4, Section 4.3.

3.2.1.5 Model Construction and Evaluation

Six traditional ML classifiers along with three McR algorithms are evaluated by using six PEMs like Ac, Sp, Sn, Pr, Fm, and PRR. These six PEMs are selected on the basis of the literature review as discussed in Chapter 2 (see Section 2.7) for BrC detection. For medical images, there are two ways to represent the classification results like image-level and patient-level (Spanhol et al., 2016b). Ac, Sp, Sn, Pr, and Fm (see Sections from 2.7.1 to 2.7.5) are required for a fair comparison with the baseline studies at image-level. Here, the Ac metric is the most commonly used to compare the results with existing SoA models. Whereas, in medical science, Sn is more important than any other performance evaluation metric because misclassification of malignancy is not tolerable for diagnosis in follow-up (Van Stralen et al., 2009). While Fm is needed to show that the classifiers are reliable and unbiased to detect BrC. Apart from image-level results, PRR is required to show the performance of the BrC detection model at the patient-level (see Sections 1.1 and 2.7.8). The PRR is defined as the ratio of the sum of patient scores to the total number of patients. Here, the total number of patient score represents the ratio of cancer images correctly classified to the total number of images per patient. In medical science patient-level, BrC detection is more important compared to image-level BrC detection (Spanhol et al., 2017). Finally, on the basis of the aforementioned six PEMs, all aforementioned ML classifiers for the EBrC-Net model are evaluated to select the top-performing model as the BrC detection model.

In this research, the BrC detection model is developed to detect BrC as benign or malignant at the image-level as well as patient-level. However, once the BrC detection is made then there is a need to find the specific subtype of BrT for better prognosis and related treatment. Because each subtype of BrT has a different treatment plan and dosage. Thus, the proposed hierarchical BrT classification model is developed to classify eight subtypes of BrT.

Universiti Malaya

3.2.2 Breast Tumor Classification Model Construction Methodology

This section discusses the detailed research methodology (see, Figure 3.3) to construct the proposed BrT classification model for breast Hp images. The overall research methodology comprises five main phases, namely, data collection, image preprocessing, features extraction through the proposed DL-based model, classification through traditional ML classifiers, feature reduction and selection, BrT classification model construction, and evaluation.

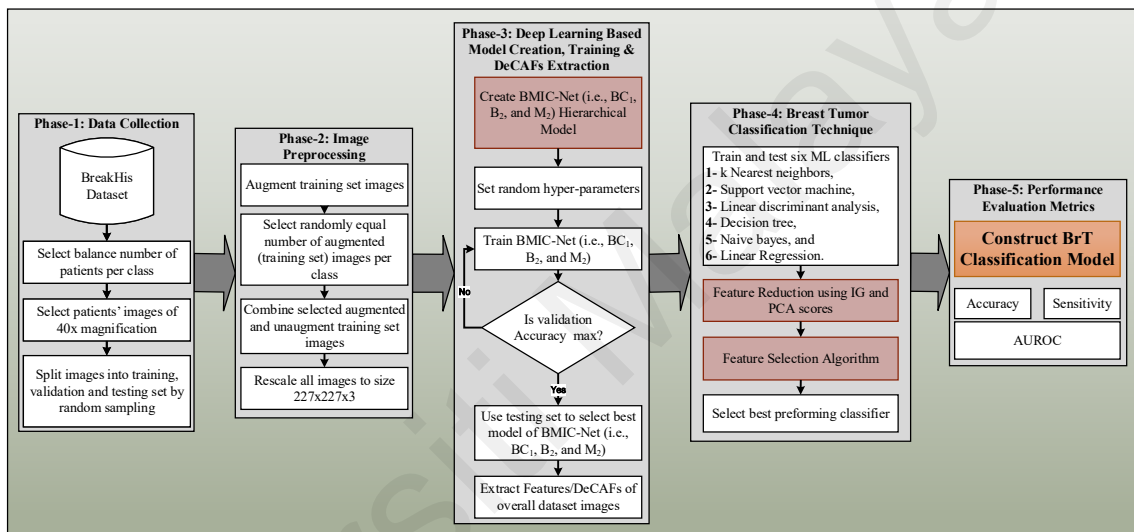


Figure 3.3: BrT classification model construction methodology

In the data collection phase, a publicly available BreakHis dataset is utilized and a balanced number of images per class are selected by the random sampling method. Whereas, in the image preprocessing phase, some essential image preprocessing tasks are performed on the collected image corpus like augmentation, selection of images for training, and rescaling. The image augmentation technique is required to avoid overfitting and class imbalance issues to improve the classification performance. Subsequently, phase three creates the proposed hierarchical BrT classification model. The proposed model comprises three DL-based classifiers namely BC_1 , B_2 , and M_2 . BC_1 can classify benign and malignant images, while B_2 and M_2 can classify further four subtypes each benign and malignant tumor. These three classifiers are created through fine-tuning the

last layer of AlexNet. Moreover, BC_1 , B_2 , and M_2 are trained until maximum validation accuracy is not observed. Finally, in phase three, several discriminative and informative DeCAFs (compared to non-hierarchical AlexNet model) of overall images available in the BreakHis dataset are extracted by using trained BC_1 , B_2 , and M_2 classifiers to perform hierarchical BrT classification. In the fourth phase, the extracted DeCAFs from BC_1 , B_2 , and M_2 DL-based classifiers are evaluated through six traditional ML classifiers namely kNN, SVM, LDA, DT, NB, and LR using three PEMs like Ac, Sn, and AUC. In addition, the overall dataset DeCAFs are evaluated for five folds and mean results are presented in this research. Here, Ac is the primary metric required to compare the results and Sn is important for medical diagnoses to show misclassification of cancerous patients. Whereas, AUC values highlight that the classifier is unbiased and produced reliable results for eight classes of BrT. Moreover, to select the best performing classifier and improve classification results, feature selection and reduction algorithms are developed using two feature reduction schemes namely IG and PCA. The feature reduction schemes and feature selection algorithm reduced the classifier training time and computational cost as well as improve the overall classification performance i.e., misclassification reduction. In the last phase, the BrT classification model is constructed by using best performing ML classifiers evaluated through aforesaid three PEMs with a minimum number of feature sets. All these phases are described in detail in subsequent subsections.

3.2.2.1 Data Collection

The details of the BreakHis dataset are already discussed in Section 3.2.1.1. However, for training BMIC-Net, this research used only images of 58 patients at $40\times$ magnification level because the highest classification performance was achieved at this level by dataset host (Spanhol et al., 2017). Noticeably, the DC class contains 38 patients' images that are 46% of the overall dataset. Thus to avoid class imbalance BMIC-Net training issue, this research randomly selects nine patients. The selected 58 patients' images (including a

borderline patient images) are split into training (i.e., 50%), validation (i.e., 20%), and testing (i.e., 30%) set by using a random sampling method, see Table 3.5.

Table 3.5: Images selected for training and testing

BrT Types	BrT Subtypes	40× (Overall)	40× (Selected)	No. of Patients	Training +Validation	Testi ng
					50% + 20%	30%
Benign	A	114	114	4	80	34
	F	253	253	10	177	76
	PT	109	109	3	76	33
	TA	149	149	7	104	45
Benign total		625	625	24	438	187
Malignant	DC	864	208	14	146	62
	LC	156	156	5	109	47
	MC	205	205	9	144	61
	PC	145	145	6	102	43
Malignant total		1370	714	34	500	214
Total Images		1995	1339	58	937	402

3.2.2.2 Image Preprocessing

The image preprocessing phase involves image augmentation, image selection and splitting into training, validation and testing set, and the rescaling of images. CNN's require a large number of training images to achieve good performance. Therefore, image augmentation can be used to improve training performance by using a small number of original images. Image augmentation creates artificial new images (i.e., over-sampling) by various image processing techniques (e.g., image rotation, shift, shear, flip, and padding) and their random combinations. In this research, the augmentation is only applied to the training set using the aforementioned image processing techniques. Each image is augmented 24 times using rotation 90°, flip in vertical and horizontal directions, translation by the fifth part of image size, shear image using four affine transforms, and image padding, see Algorithm 3.1.

Algorithm 3.1: Image augmentation algorithm

```
Input: path to Source Directory(SD), Target Directory(TD)
Output: Twenty Five Times Augmented images
Procedure ImgsAugment(SD,TD)
1    $n \leftarrow$  count total number of images in SD
2    $i \leftarrow 1$ 
3   while  $i \leq n$  do
4        $img \leftarrow$  read image(i) from SD
5        $imgFH \leftarrow$  flip_horizontally( $Img$ )
6        $imgFV \leftarrow$  flip_vertically( $Img$ )
7        $imgRFH \leftarrow$  rotate90( $ImgFH$ )
8        $imgRFV \leftarrow$  rotate90( $ImgFV$ )
9        $[h, w] \leftarrow$  Image_size( $Img$ )/5
10       $imgFH\_RT \leftarrow$  translate  $ImgFH$  by  $[h, -w]$ 
11       $imgFH\_RB \leftarrow$  translate  $ImgFH$  by  $[h, w]$ 
12       $imgFH\_LT \leftarrow$  translate  $ImgFH$  by  $[-h, -w]$ 
13       $imgFH\_LB \leftarrow$  translate  $ImgFH$  by  $[-h, w]$ 
14       $imgFV\_RT \leftarrow$  translate  $ImgFV$  by  $[h, -w]$ 
15       $imgFV\_RB \leftarrow$  translate  $ImgFV$  by  $[h, w]$ 
16       $imgFV\_LT \leftarrow$  translate  $ImgFV$  by  $[-h, -w]$ 
17       $imgFV\_LB \leftarrow$  translate  $ImgFV$  by  $[-h, w]$ 
18       $[S1, S2, S3, S4] \leftarrow$  shear  $Img$  by affine transform 1,2,3,4
19       $[FxS1, FxS2, FxS3, FxS4] \leftarrow$  flip_Horizontally( $S1, S2, S3, S4$ )
20       $[FyS1, FyS2, FyS3, FyS4] \leftarrow$  flip_Vertically( $S1, S2, S3, S4$ )
21       $imgP =$  padding( $img$ )
22      write all images to disk
23       $i \leftarrow i + 1$ 
24   end
25 end
```

In addition, the original images are used along with the minimum quantity available in any class out of 24 times augmented images. The original number of images taken in each class is shown in Table 3.5. Finally, all the images are rescaled to size $227 \times 227 \times 3$ before initiating the BMIC-Net models training process.

3.2.2.3 Development of BrT Classification Model

The proposed hierarchical BrT classification model named as Biopsy Microscopic Image Cancer Network (BMIC-Net) is created by using a pre-trained model like AlexNet. BMIC-Net is composed of three DL-based classifiers namely BC_1 , B_2 , and M_2 . Moreover, each classifier is created by fine-tuning the last layer of AlexNet for the target number of classes. BC_1 , B_2 , and M_2 classifiers are employed in two leveled hierarchy. At the top level of the hierarchy, BC_1 is placed to classify Hp image as benign or malignant. Whereas, level two of hierarchy possesses B_2 and M_2 classifiers. B_2 classifier is

responsible to classify further four subtypes of benign tumors. On the other hand, M₂ can classify four subtypes of malignant BrC. The hierarchical design and use of pre-trained models enable the BMIC-Net model to be trained with less computational resources (such as a normal desktop machine) in less time using a fewer number of images compared to the de-novo model. For further details of the proposed BrT classification model design, see Chapter 4, Section 4.4.

(a) **Training and Feature Extraction through BMIC-Net Model**

BMIC-Net DL-based classifiers (i.e., BC₁, B₂, and M₂) are trained multiple times with random hyper-parameters using a trial-and-error method. The training process continues until maximum validation accuracy (lies between 0 to 100) is not observed for each of the three DL-based classifiers. In the construction of a fine-tuned BMIC-Net classifiers, several experiments are run recursively to obtain optimum training results by adjusting a few training options, see Algorithm 3.2.

Algorithm 3.2: BMIC-Net model training, classification, and feature extraction algorithm

```

Input: PathTr, PathTs
Output: TrainedBMIC_Net, MFV
Procedure TrainBMIC_Net(TI)
1  [TrainingImage, TrainingLabels] ← Load Training images using ParthTr
2  [TestingImages, TestingLabels] ← Load Testing images using PathTs
3  TrainingImages ← Resize(TrainingImages, [227x227])
4  TestingImages ← Resize(TestingImages, [227x227])
5  Repeat
6      BC1 ← Fine tune and choose random parameters using AlexNet for Binary classes
7      B2 ← Fine tune and choose random parameters using AlexNet for Benign 4 classes
8      M2 ← Fine tune and choose random parameters using AlexNet for Malignant 4 classes
9      TrainedBMIC_Net ← Train BC1, B2, M2 using TrainingImages
10         Stop training if accuracy is not improving in consecutive three epochs
11     PredictedLabels ← Predict(TrainedBMIC_Net, TestingImages, TestingLabels)
12     ConfMat ← confusion_matrix(TestingLabels, PredictedLabels)
13     calculate accuracy, sensitivity
14 Until accuracy is maximum  \\(>=0 AND <=100)
15 TrFeatures ← ExtractFeatures(TrainedBMIC_Net, TrainingImages)
16 TsFeatures ← ExtractFeatures(TrainedBMIC, TestingImages)
17 MFV ← Save (TrFeatures, TrLabels, TsFeatures, TsLabels)
18 Return TrainedBMIC_Net (BC1, B2, M2)
19 Return MFV (BC1, B2, M2)
20 end

```

The BC₁, B₂, and M₂ classifiers are trained using gradient descent. Some of the parameter adjustments are as follows: the momentum of 0.9, maximum epochs of 30, mini-batch size of 50, the initial learning rate of 1e-4, and learning rate drop factor of 0.5. The BC₁, B₂, and M₂ are fine-tuned with stochastic gradient descent with a learning rate adjusted to be lower than the initial learning rate of AlexNet. Hence, the features previously learned from the larger dataset are guaranteed to be not entirely ignored during retraining. Furthermore, the network is compelled to stop training if validation accuracy (lies between 0 to 100) is not improving in the multiple numbers of consecutive validation iterations. Ultimately, BMIC-Net classifiers are devised in such a way to avoid overfitting and underfitting training issues. Moreover, the best feasible validation Ac of BMIC-Net classifiers is achieved to obtain the best possible features.

It should be noted that a random sub-sampling approach is used to select 50% training, 20% validation, and 30% for testing set images. Furthermore, by default, the softmax classifier is used in the validation process for training. Finally, the features that are finalized by BMIC-Net classifiers are extracted before the output layer to form master feature vector (MFV), which is composed of 4096 features for classification and served as an input to traditional ML classifiers (i.e., SVM, kNN, DT, NB, LDA, and LR) to evaluate the classification predictive performance.

3.2.2.4 Breast Cancer Classification Technique

The extracted MFV of each BC₁, B₂, and M₂ classifiers is evaluated using five folds through six ML classifiers namely, the kNN (k=1,3,5,7, and 9), SVM (linear, rbf, and polynomial), NB, DT, LDA, and LR. Apart from these, softmax results are also compared with the proposed model performance for the non-hierarchical model. These six ML classifiers are selected on the basis of the literature review as discussed in Chapter 2 (Section 2.6.1). Moreover, according to the No-Free-Lunch theorem (Wolpert &

Macready, 1997), the best classifier will not be the same for all the data sets. Therefore, these six ML classifiers are applied to select the best performing models for BMIC-Net classifiers for BrT classification on the BreakHis dataset.

(a) Feature Reduction and Selection

A large number of features (i.e., 4096) are extracted through each of the three DL-based classifiers (i.e., BC₁, B₂, and M₂) of the BMIC-Net hierarchical model. Such a large number of features can easily distract the training process of traditional ML classifiers (Fan & Fan, 2008). Therefore, two feature reduction schemes, namely IG and PCA are adopted to reduce the number of features. Afterward, a feature selection algorithm is developed to select a minimum number of feature subsets. The proposed feature selection algorithm selects a minimum number of features (using IG and PCA) without compromising the overall accuracy of the hierarchical BrT classification model. Moreover, it reduces the training time and improves the ML classifier performance by reducing the misclassification. The detailed discussion for feature reduction and selection algorithm is given in Chapter 4, Section 4.5.

3.2.2.5 Model Construction and Evaluation

The best performing traditional ML classifier for BC₁, B₂, and M₂ DL-based classifiers is evaluated by using three (i.e., Ac, Sn, and AUC) PEMs. These three PEMs are selected on the basis of the literature review as discussed in Chapter 2 (see Section 2.7) for BrT classification. Moreover, Ac, Sn, and AUC are also computed for a fair comparison with the baseline study. The Ac metric is mainly required to compare the results with existing SoA models. However, a single metric analysis can be biased like accuracy. On the other hand, in medical science, Sn is more important than any other PEM because misclassification of malignancy is more critical than benign. Whereas, AUC is needed to show that the classifiers are unbiased to classify eight subtypes of BrT. Thus, Ac, Sn, and

AUC metrics will ensure that the performance of the proposed BrT classification model is unbiased (i.e., more reliable) even using a multifaceted dataset. Finally, on the basis of the aforementioned three PEMs, the top-performing traditional ML classifier, best feature reduction scheme, and minimum feature subset are selected to construct a hierarchical BrT classification model for Hp images.

Universiti Malaya

3.3 Experimental Setup

This section presents the experimental setup of proposed DL-based models for BrC detection and BrT classification using Hp images. An extensive set of experiments is carried out using various PEMs. Overall four experimental setups are made for each BrC detection model and BrT classification model, see Figure 3.4. The complete flow of experimental setups is discussed in the following subsections. Sections 3.3.1 and 3.3.1 discusses the overall experimental setup of BrC detection and BrT classification models.

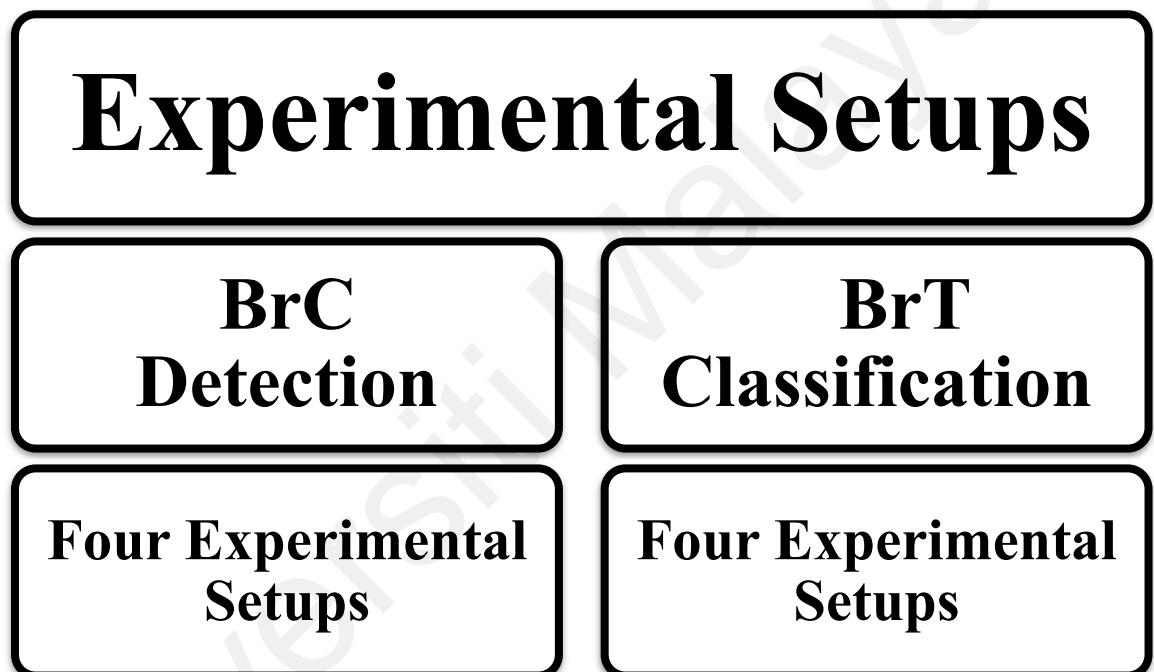


Figure 3.4: The experimental setups distribution overview

3.3.1 Experimental Setup of BrC Detection Model

In this research, four experimental setups are designed to assess the performance of the proposed BrC detection model. The setting I is required to select the best DL-based model for feature extraction. Setting II enables to select the best performing ML classifier and setting III improves the performance of the classification model for BrC detection. Whereas, setting IV shows the improved PRR for patient-level BrC detection. Overall 82 (26 + 14 + 18 + 24) analyses are performed in the following four experimental setups.

1. In Setting I, initially input image size is optimized before initiating the training process of the proposed EBrC-Net model. Afterward, EBrC-Net performance is analyzed and compared with the pre-trained AlexNet model. Thus, during the training of both models, the epoch-wise validation loss and validation accuracy are computed for comparison. In addition, the performance of both models is further examined by using testing images at each epoch. Hence, this setting analysis helps to select the best model for further experiments. Here, 26 analyses are performed (03 for input image optimization + 10 for AlexNet training + 10 for EBrC-Net training), see Figure 5.3.
2. In Setting II, firstly parameter optimization is carried for kNN (1,3,5,7 and 9) and SVM (kernel, rbf, and polynomial) classifiers. Secondly, the performance of extracted DeCAFs from EBrC-Net is evaluated through six ML classifiers, namely, softmax, kNN (k=7), SVM (kernel), NB, DT, and LDA. In this setting, 14 analyses (5 for kNN optimization + 3 for SVM optimization + 6 ML classifiers) are conducted by using five PEMs such as Ac, Sp, Sn, Pr, and Fm. These five PEMs enable us to compare the performance of six ML classifiers for BrC detection. For further experiments, see Figure 5.5.
3. In Setting III, three algorithms are developed and implemented in a cascaded manner for McR to boost up the performance of six ML classifiers for BrT detection. In addition, similar PEMs are adopted for evaluation as used earlier in Setting II. In this setting, overall 18 analyses are conducted (3 McR algorithms x 6 ML classifiers), see Table 5.1.
4. In Setting IV, the PRR is computed to show patient-level improved performance due to McR algorithms. In addition, six ML classifiers' performance is compared before and after applying the proposed McR algorithms using the aforementioned five PEMs. In medical science, the PRR is highly important, because in real-life

the decisions are based on patient-level instead of image-level classification (Spanhol et al., 2017). In this setting, a total of 24 analyses are performed. Where, out of 24, six analyses are performed before, and 18 analyses are performed after applying the McR algorithms by using six ML classifiers, see Figure 5.6.

The image preprocessing steps, development of the EBrC-Net model, all Hp image classification experiments, and McR algorithms are implemented in MATLAB R2017b version. All the experiments are mainly executed on default hyper-parameters except for those hyper-parameters which are specifically mentioned in this research work, see Section 3.2.1.3(a).

3.3.2 Experimental Setup of BrT Classification Model

In this research, four experimental setups are designed to assess the performance of the proposed BMIC-Net hierarchical BrT classification model. Here, an overall 3015 result analyses are carried out. In setting I, II, III, and IV overall 49, 26, 2916, and 24 analyses are made respectively.

1. In the first setting, the training performance of proposed hierarchical (see, **Figure 4.4** and Figure 4.5) model BMIC-Net classifiers (BC₁, B₂, and M₂) [see, Figure 5.7 (a) to (f)] has been analyzed and compared with non-hierarchical (see Figure 4.4) model [see, Figure 5.7 (g) and (h)] Thus, during the training of both models, the epoch-wise validation loss and validation accuracy are computed for comparison. This epoch-wise analysis helps to select the best model for feature extraction. Here, 49 analyses are performed (36 for the proposed BMIC-Net hierarchical model + 13 for the non-hierarchical model), see Figure 5.7. Finally, three MFVs are created from trained BC₁, B₂, and M₂ DL-based classifiers using all images of BreakHis for five-fold analysis.

2. In the second setting, initially, parameters are optimized for kNN (1,3,5,7, and 9) and SVM (linear, rbf, and polynomial) to get the best use of classifiers. Subsequently, the performance of extracted three MFVs is evaluated using six traditional ML classifiers, namely, kNN (k=1), SVM(linear), NB, DT, LDA, and LR. In addition, five folds of each of three MFVs are used to show the mean results in terms of Ac, Sn, and AUC for each of the aforementioned six ML classifiers. Hence, in this setting, 26 analyses are run (5 for kNN optimization + 3 for SVM optimization + 6 traditional ML classifiers \times 3 BMIC-Net classifiers: BC₁, B₂, and M₂), see Figure 5.9.
3. In the third setting, the best feature subset is obtained using IG and PCA. The performance of 50 to all 4096 features is evaluated with an increment of 50 from all three MFVs (BC₁, B₂, and M₂). Thus, 82 sub-MFVs (50 features, 100 features, 150 features, ..., and 4096 features) are prepared overall from each of the three super-MFVs. Moreover, the same six ML classifiers, which are used in setting I, are adopted to evaluate the performance of the prepared 82 \times 3 sub-MFVs. Thus, in this setting, 2916 analyses are run (81 sub-MFVs \times 3 super-MFVs \times 2 feature reduction schemes \times 6 ML classifiers) to perform feature reduction, see Figure 5.10.
4. The fourth setting mainly comprises of two parts. Where, first part shows the mean Ac of five folds to analyze the performance of features extracted from the non-hierarchical model using six ML classifiers for eight subtypes of BrT, see Table 5.4. While, the second part compares the mean Ac of the proposed hierarchical model against the non-hierarchical model using six ML classifiers, see Table 5.5. Therefore, in this experiment, overall 24 analyses are executed (12 for non-hierarchical model + 12 for proposed BMIC-Net hierarchical model).

The image preprocessing, construction of proposed BMIC-Net, all classification experiments, and feature reduction are performed in MATLAB R2017b. All classification experiments are solely executed on default hyper-parameters except a few which are specifically defined for this research work, see Section 3.2.2.3(a).

3.4 Summary

This chapter is divided into two major parts namely methodology and experimental setup. In the first part of this chapter, the overall research methodology used for BrC detection and BrT classification using Hp images is described in detail. Initially, the details of the dataset are discussed. Afterward, the basic medical image preprocessing techniques like stain normalization, augmentation, splitting images into training, validation, and testing, and rescaling are talked about. Furthermore, the development and training of proposed DL-based models namely EBrC-Net and BMIC-Net are discussed in detail. Subsequently, a detailed discussion about DeCAFs extraction and evaluation through six traditional ML classifiers has been made. Moreover, three McR algorithms (i.e., McRI, McRP, and McRC) are briefly explained to enhance the performance of the BrC detection model. A feature reduction and selection algorithm is also discussed to improve the performance of the hierarchical BrT classification model. Finally, performance evaluation metrics are elaborated for evaluation to construct BrC detection and BrT classification models.

The second part of this chapter describes the organization of experimental setups implemented for both BrC detection and BrT classifications models. For BrC detection, four experimental settings are made for each of the proposed models. The experimental setting I ensures that the proposed DL-based EBrC-Net model is able to extract better features compared to AlexNet. Whereas, experimental setting II, enables to evaluate the performance of six traditional ML classifiers using five folds of extracted features from

EBrC-Net. However, in experimental setting III, the performance of the EBrC-Net is enhanced by implementing three McR reduction algorithms. Finally, in setting IV of experimental setups, the improved PRR is also shown for patient-level analysis for BrC detection. Thus, the best performing classifier is obtained for BrC detection. On the other hand, for BrT classification, four experimental settings are implemented. In setting I, proposed BMIC-Net hierarchical classifiers like BC_1 , B_2 and M_2 are trained to extract the best possible MFVs compared to the non-hierarchical model. In setting II, the extracted overall BreakHis MFVs are evaluated using five folds via six ML classifiers. Whereas, in setting III, the performance of the proposed hierarchical classification model is enhanced by implementing feature reduction algorithms. While, in setting IV, the performance of the proposed BMIC-Net hierarchical model is compared with the non-hierarchical model by reporting mean A_c for five folds of extracted MFVs. The details about the main contributions for proposed models are discussed in Chapter 4.

CHAPTER 4: DEVELOPMENT OF BREAST CANCER DETECTION AND CLASSIFICATION MODELS

4.1 Introduction

This chapter represents a detailed discussion about the architectural design and development of proposed BrC detection and BrT classification models to highlight the research contributions. Overall, four research contributions are made in this research, as shown in Figure 4.1.

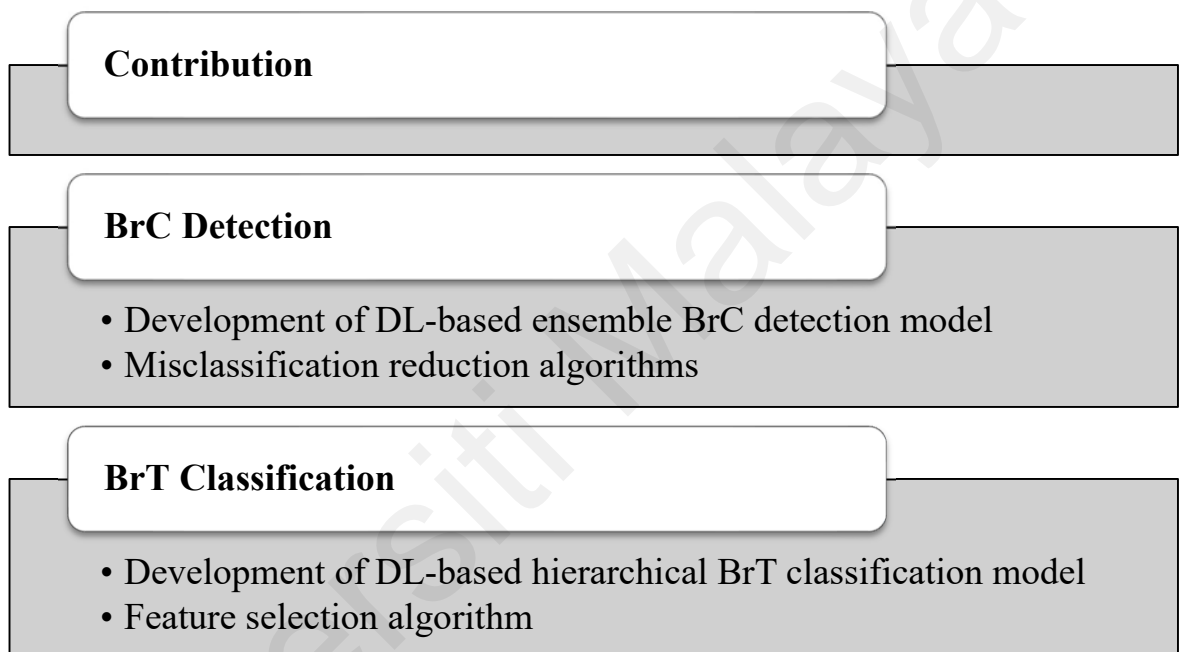


Figure 4.1: Contributions for BrC detection and classification

In BrC detection model the main contributions are:

1. Design and development of an efficient (i.e., consumes less computational resources and training time) and reliable (i.e., produces better and unbiased results even using complex dataset) DL-based BrC detection model (i.e., EBrC-Net) to extract discriminative features (due to larger input image size and unfreezed layers) compared to AlexNet using BreakHis dataset Hp images.
2. Three misclassification reduction algorithms are implemented in a cascade manner to enhance the performance of the proposed BrC detection model.

The main contributions in the BrT classification model,

1. Design and development of an efficient (i.e., consumes less computational resources and training time) and reliable (i.e., produces better and unbiased results even using complex dataset) DL-based BrT classification model (i.e., BMIC-Net hierarchical model) to extract discriminative features compared to non-hierarchical model using BreakHis dataset Hp images.
2. A feature selection algorithm is implemented to elicit a minimum number of feature subsets to reduce misclassification and enhances the overall performance of the proposed BrT classification model.

The development of BrT detection and BrC classification models is described in section 4.2 and 4.4. Whereas, the performance enhancement algorithms for both of the proposed models is described in Sections 4.3 and 4.5.

4.2 Development of BrC Detection Model

The proposed DL-based BrC detection model (i.e., EBrC-Net) is a fusion of TL-based (i.e., pre-trained) and de-novo (i.e., trained from scratch) models, see Section 3.2.1.3. Therefore, it is named as an ensemble breast cancer network for BrC detection using Hp biopsy images. EBrC-Net possesses the same architecture as AlexNet (Krizhevsky et al., 2012) except for two major modifications. First, the input layer image size is larger than that of AlexNet. Second, three fully connected layers are allowed to learn from scratch. Whereas, five convolution layers are adopted from pre-trained AlexNet to perform partial transfer learning. The ensembling in EBrC-Net ensures the extraction of improved generalized and domain specific Hp image features compared to AlexNet using less computational resources and training time. In the following sections, the architecture of AlexNet, the design and development of the proposed EBrC-Net, and the latter proposed model's importance are discussed.

4.2.1 Pre-trained AlexNet Architecture

AlexNet is a pre-trained CNN-based classification model specially trained on 1 million natural images to classify 1000 objects, such as coffee mugs, pencils, pens, keyboards, and animals. At the abstract level, the AlexNet architecture comprises the following layers: an input layer, five convolution layers (CLs), three fully CLs (FCLs), a softmax layer, and an output layer, see Figure 4.2. The input layer directly takes RGB images (e.g., $I = i_1, i_2, i_3, \dots, i_n$) of size $227 \times 227 \times 3$. Additionally, a nonlinear activation layer rectified linear unit (ReLU), and a batch normalization layer is placed after each CL in AlexNet. Similarly, a MaxPool layer is kept after the normalization layer in each CL except third and fourth convolution layers. Furthermore, a ReLU layer and a dropout layer are also placed after the sixth and seventh FCLs, whereas a softmax layer is employed after the last FCL to compute the probabilities of each class label for the output layer. AlexNet is formally denoted by $f: i \rightarrow c$, where i represents the image c , which denotes the classification label. AlexNet contains V convolution layers, and U fully connected layers are defined as follows:

$$f(I; \theta) = f_{out}(FCLs, \theta_{out}) \quad (12)$$

$$FCLs = f_U(\dots, f_1(CLS, \theta_1), \dots, \theta_U) \quad (13)$$

$$CLs = f_V(\dots, f_1(I, \theta), \dots, \theta_V) \quad (14)$$

where $\{f_i(\cdot)\}_{i=1}^V$ represents a convolutional layer, θ_v denotes the parameters of layer v of the AlexNet which consist of a weight matrix $\mathbf{W}_v \in \mathbb{R}^{k_v \times k_v \times n_v \times n_{v-1}}$ and bias vector $\mathbf{b}_v \in \mathbb{R}^{n_v}$, with $k_v \times k_v$ showing the size of the filters in layer v which possess n_{v-1} input channels and n_v output channels. Similarly, f_U is a fully connected layer with weights $\{\mathbf{W}_u\}_{u=1}^U$, where $\mathbf{W}_u \in \mathbb{R}^{n_{u-1} \times n_u}$ representing the connections from fully connected layer $u-1$ to u and biases $\{\mathbf{b}_u\}_{u=1}^U$ where $\mathbf{b} \in \mathbb{R}^{n_u}$ and f_{out} represent a

multinomial logistic regression layer (Krizhevsky et al., 2012) containing weights $\mathbf{W}_{\text{out}} \in \mathbb{R}^{n_u \times C}$ and bias $\mathbf{b}_{\text{out}} \in \mathbb{R}^C$.

The operation performed by each convolutional layer $v \in \{1, \dots, V\}$ of the AlexNet is defined as follows:

$$\mathbf{F}_v = f(I_{v-1}, \theta) = \mathbf{W}_v \otimes \mathbf{H}_{v-1} + \mathbf{b}_v \quad (15)$$

where \otimes used as convolution operator, $\mathbf{H}_v = (\mathbf{h}_{u,1}, \dots, \mathbf{h}_{u,n_u})$ and H_0 represents the input H_p image i . After the convolutional layers, fully connected layers are available that receives vectorized input volume $\mathbf{f}_U \in \mathbb{R}^{|\mathbf{f}_U|}$ from H_v , where $|\mathbf{f}_U|$ represents the length of the vector \mathbf{f}_U and applies V linear transformations defined by (Krizhevsky et al., 2012) as follows:

$$\mathbf{f}_v = f_v(\mathbf{F}_v, \theta_v) = (\mathbf{W}_{v,V}, \dots, (\mathbf{W}_v \mathbf{f}_U + \mathbf{b}_v), \dots + \mathbf{b}_{v,V}) \quad (16)$$

where $\mathbf{f}_v \in \mathbb{R}^{n_{v,v}}$. The final classification layer is defined by a softmax function over a linearity transform input (Krizhevsky et al., 2012) as follows:

$$\mathbf{f}_{\text{out}} = f_{\text{out}}(\mathbf{f}_v, \theta_{\text{out}}) = \text{softmax}(\mathbf{W}_{\text{out}} \mathbf{f}_v + \mathbf{b}_{\text{out}}) \quad (17)$$

where $\text{softmax}(\mathbf{z}) = \frac{e^z}{\sum_j e^{z(j)}}$ and $\mathbf{f}_{\text{out}} \in [0,1]^Y$ denotes the output from the overall extraction of process, with Y representing the number output labels.

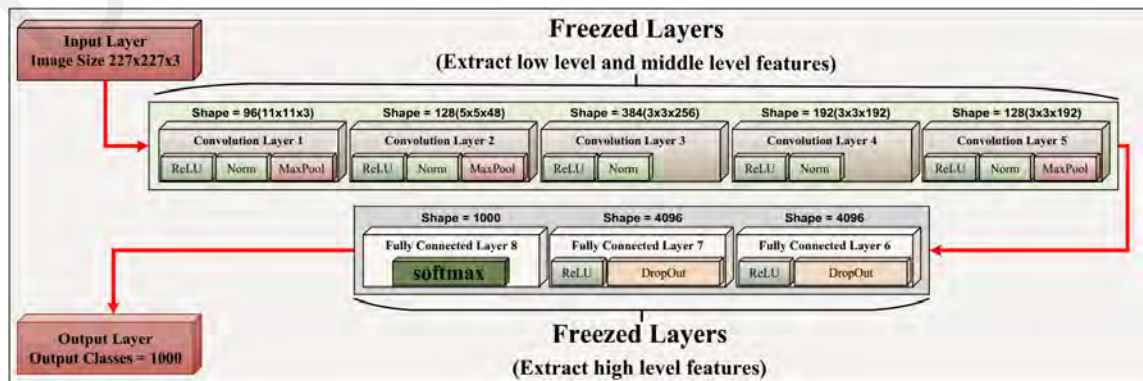


Figure 4.2: AlexNet architecture

The output of the convolutional layer is often forwarded to nonlinearity function, batch normalization layer, and subsampling layers. An activation or nonlinearity function is the function in an artificial neuron that delivers an output that is based on inputs. Many nonlinearity functions are used to model the neuron's output, for instance, tanh [$f(x) = \tanh(x)$] or sigmoid [$f(x) = (1+e^{-x})^{-1}$]. In terms of deep neural training, while using gradient descent to find local minima for calculating validation loss, these nonlinearity functions are much slower than ReLUs [$f(x)=\max(x,0)$] (Nair & Hinton, 2010; Krizhevsky et al., 2012). If x is less than zero, then $f(x)$ is zero; otherwise, $f(x)$ is x . Therefore, AlexNet is equipped with ReLU for expedited processing using gradient descent. ReLU does not require input normalization if some training examples provide positive input. Hence, a batch normalization scheme is introduced after ReLU in each convolutional layer of AlexNet, which can be referred to as “brightness normalization,” and it can reduce the error rate by 2% in a four-layer CNN (Krizhevsky et al., 2012). Mathematically batch normalization is denoted by the following:

$$b_{x,y}^i = \frac{a_{x,y}^i}{\left(k+a \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a_{x,y}^j)^2\right)^\beta} \quad (18)$$

Here, $b_{x,y}^i$, $a_{x,y}^i$ represents batch normalization and activity of the neuron on kernel i at location x, y . Moreover, n denotes adjacent kernel maps at the same location, and N is the total number of kernels in the layer. The constants k, n, a , and β are hyper-parameters, and their values are determined on the basis of the validation set. In addition to the nonlinearity function and batch normalization layers, a subsampling layer (i.e., MaxPool layer) is often placed to summarize the outputs of neighboring groups of neurons in the same kernel map. MaxPool takes the maximum value from an input matrix. The AlexNet avoids the overlap pooling by using stride 2 with 3×3 filter size. Hence, the network is less likely to be overfitted. Furthermore, a dropout layer is used after the first two FCLs in AlexNet, because it reduces the training time and overfitting dispute by assigning zero

to the output of each neuron with a probability of 0.5. Hence, such dropped neurons do not participate in both the forward pass and backpropagation process while training.

4.2.2 Proposed DL-based BrC Detection Model

Transfer learning is the process of using a previously trained (often trained on natural images such as AlexNet) model after fine-tuning and retraining on specific target data such as Hp BrC medical images. In fine-tuning, the last layer of the pre-trained model is often replaced by a newly created classification layer. Moreover, in TL, the weights of all layers remain frozen except the last layer, which is the newly created classification layer. Freezing the layers means not changing the layer weights during the training process in gradient descent optimization. However, due to the freezing of weights, the computations will be minimized, thus it enabled the TL-based model to be trained by using less computational resources in less time using fewer annotated images (like medical images) compared to a de-novo model (Jiang et al., 2017).

Whereas, in a de-novo model, new layers are created and trained from scratch (Hadad et al., 2017). De-novo model layers are able to change their weights while training, therefore referred to as unfrozen layers. Due to the updation of a large number of weights in each iteration with the backpropagation process, a large number of computations are performed, thus requires very high computational resources, longer training time, and a large number of labeled images. However, in a de-novo model, customized layers can be created according to the type and size of data.

On the contrary, TL-based model may not be able to produce good results for specific types of images such as Hp BrC because these models are trained on a large number of natural images like ImageNet (Yosinski, Clune, Bengio, & Lipson, 2014) and unable to learn specific features from a small number of medical images like BreakHis. In contrast,

the de-novo model can show better performance for specific types of images instead of natural images due to customized layers. Moreover, a smaller size de-novo model network can be created and may produce better results by using less computational resources and training time (Hadad et al., 2017). However, the major obstacle of using the de-novo model is that it often requires a large number of annotated images to train from scratch, which is often not possible for medical images. Thus, TL-based or ensemble models can be trained easily instead of the de-novo model for medical image classification.

4.2.2.1 EBrC-Net Architecture and Model Structure

In deep CNN-based models like AlexNet, the CLs are responsible to extract low-level and middle-level features such as edges, curves, bobs, and colors, which are common features in all types of images such as medical and nonmedical (i.e., natural images). On the other hand, the FCLs are capable to extract high-level features, also known as specific features. The specific features of medical images (e.g., BrC lesion structure and geometrical shape features) are entirely different from natural images. Hence, the TL model may lose the specific features of BrC images due to frozen FCLs. Nonetheless, the common features remain unaffected/preserved even though the CLs are frozen. Therefore, by exploiting this property of TL models, the initial five layers of the proposed EBrC-Net are kept frozen (transferred from AlexNet) to adopt image common features. Thus, due to frozen layers, the computations will be reduced at each layer to update the weights. However, the last three layers of EBrC-Net are kept unfrozen as used in a de-novo model, so that they can be trained from scratch to learn the specific features of medical images like BrC lesion shape, size, and structure. Therefore, EBrC-Net is a fusion of the TL and de-novo models, that is, EBrC-Net, see Figure 4.3. The ensembling of two learning techniques enabled EBrC-Net to take advantage of both types of models i.e., TL-based and de-novo models. Thus, EBrC-Net is able to produce better results compared to

the de-novo and TL-based models by using a small number of images without facing an overfitting issue. In addition, EBrC-Net can be trained in less time by consuming fewer resources such as a common desktop CPU. Thus, EBrC-Net is able to produce better generalization of features for BrC Hp images than both de-novo and TL-based models.

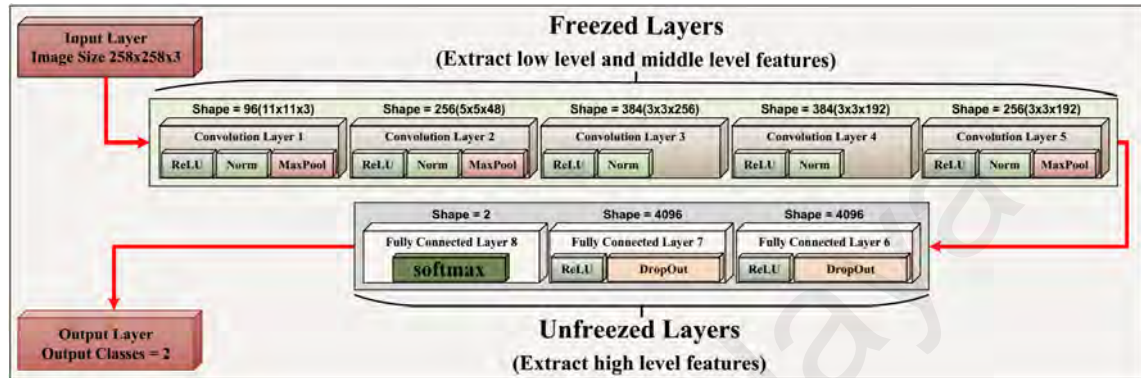


Figure 4.3: Network architecture of proposed EBrC-Net

Finally, the weights of newly created fully connected layers of EBrC-Net are initialized by using the weights of pre-trained AlexNet because AlexNet FCL weights become more supportive to EBrC-Net while training compared to randomly assigned weights. Actually, it minimizes the validation loss faster than the random weight initialization. In addition, the input layer image size of EBrC-Net is taken larger than the pre-trained AlexNet input layer. The input layer image size of EBrC-Net (Figure 4.3) is optimized to $258 \times 258 \times 3$ instead of $227 \times 227 \times 3$ used by AlexNet, see Figure 4.2. Optimized input layer image size enabled the proposed model to extract more discriminative and generalized features for BrC Hp images compared to the AlexNet default input layer image size. In the end, the last fully connected layer is fine-tuned to perform BrC detection (benign/malignant) results for BrT Hp images.

4.3 Performance Enhancement of BrC Detection Model

This section describes the second contribution made for the BrC detection model. In this regard, three misclassification reduction algorithms namely misclassification reduction image-wise (McRI), misclassification reduction patient-wise (McRP), and

misclassification reduction confidence-wise (McRC) are created by using top-performing augmentation methods to enhance the performance of the proposed BrC detection model. Therefore, the next two subsections will describe the augmentation method selection and McR algorithms in a detailed manner.

4.3.1 Selection of Image Augmentation Method

As discussed in Section 3.2.1.2(b), 24 augmentation techniques are used to create augmented images to form the training set. Here, out of 24, three top-performing augmentation techniques are selected to reduce the misclassification for ML classifiers. Because it has been observed by performing numerous experiments that more than three augmentation methods are unable to reduce misclassification. Moreover, the use of more than three augmentation techniques created more saturated and unbiased results. After performing extensive experiments by using ML classifiers, the three best augmentation techniques are selected by using three PEMs, namely, A_c , S_p , and S_n . The evaluation of these metrics enabled the selection of the three best performing unbiased augmentation techniques, which can, therefore, be used for McR.

4.3.2 Misclassification Reduction Algorithms

The DL-based EBrC-Net model is trained using 24 times augmented images along with original images. Whereas the quantity of augmented images is much higher than the original images, thus EBrC-Net got better training for augmented images instead of original images. However, testing data contains only original images, thus it can be easily misclassified by the model which is largely trained on augmented images. Therefore, the misclassification rate needs to be reduced to produce reliable results. In this regard, three McR algorithms namely McRI, McRP, and McRC are developed to minimize the misclassification rate for BrC detection by using BrT Hp images. Noticeably, before applying McR algorithms, the images are sorted in a patient-wise manner.

(a) **McR Image-wise**

In the McRI algorithm (Algorithm 4.1), the input image is augmented three times by using previously selected best augmentation techniques (see Section 4.3.1) and classified through a trained classifier. Because more than three times augmentation is unable to reduce misclassification. Thereafter, the label for the original image is decided on the basis of counting of predicted three labels of augmented images. If the counting of benign prediction is higher than malignant, then the original image is labeled as benign, else malignant. Suppose original testing images are represented by $I_1, I_2, I_3, \dots, I_n$ and each image is augmented three times and denoted as $a_1, a_2,$ and a_3 . Now assume that the augmented images $a_1, a_2,$ and a_3 of original image I_1 are classified as images $a_1,$ and a_3 are benign whereas a_2 is malignant, thus three augmented images are counted as $I_1(2,1)$; likewise assume images $a_1, a_2,$ and a_3 of I_2 are classified as image a_1 is benign whereas images a_2 and a_3 are malignant, therefore counted as $I_2(1,2)$. Let us say seven images of patient P_1 are present, and all are classified and counted as $I_1(2,1), I_2(1,2), I_3(2,1), I_4(0,3), I_5(2,1), I_6(2,1),$ and $I_7(3,0)$. Hence, clearly, five images $I_1, I_3, I_5, I_6,$ and I_7 are classified as benign whereas two images I_2 and I_4 are classified as malignant, see Algorithm 4.1. Thus, using a technique to predict a class label for the original image on the basis of three augmented images reduced the chances of misclassification instead of predicting a single original image directly. However, patient confidence will be $5/7$ (i.e., 71.42%) (see lines 21 and 23 of the McRI Algorithm 4.1).

(b) **McR Patient-wise**

In the McRP algorithm (Algorithm 4.2), initially, all original images of the single patients are read and augmented thrice by using previously selected best augmentation techniques. Thereafter, all the augmented images of one patient are classified by a trained classifier. Subsequently, the predicted labels for benign and malignant are counted and stored in a table named patient-wise counting table (PCT). PCT is populated in such a

way that each row represents the count(s) for malignant or benign predictions for each original image. Thus, PCT contains the overall counts for benign and malignant predictions for all images of a patient.

Algorithm 4.1: McRI Algorithm

```

Input: TrainedClassifier, PatientNoList, TestingImages
Output: AllPredictedLabels_IW, AllPatientsConfidence_IW
Function MisclassificationReduction_ImageWise
1   NoPatients = count(PatientNoList)
2   For PNum = 1:NoPatients
3       SinglePatientImages = GetPatientImages(TestingImages,PNum)
4
5       NoImages = count(SinglePatientImages)
6       For NoI = 1:NoImages
7           Img = ReadImage(SinglePatientImages(NoI))
8           AugImages = AugmentImage3Times(Img)
9           PredictLbIs = Predict(Classifier,AugImages)
10          CntB=CountBenign(PredictLbIs)
11          CntM=CountMalignant(PredictedLbIs)
12          If CntB > CntM
13              OriginalImageLabel = 'Benign'
14              TotalCntB = TotalCntB + 1
15          Else
16              OriginalImageLabel = 'Malignant'
17              TotalCntM = TotalCntM + 1
18          End
19          SinglePatientPredictedLbIs(NoI) = OriginalImageLabel
20      End
21      If TotalCntB >= TotalCntM
22          PatientConfidence = TotalCntB/(TotalCntB+TotalCntM)*100
23      Else
24          PatientConfidence = TotalCntM/(TotalCntB+TotalCntM)*100
25      End
26      Populate AllPredictedLabels_IW by SinglePatientPredictedLbIs
27      Populate AllPatientsConfidence_IW by PatientNo and PatientConfidence
28  End
Return AllPredictedLabels_IW, AllPatientsConfidence_IW
End

```

In the same way, a PCT is created for each patient (e.g., $P_1, P_2, P_3, \dots, P_n$) one after the other. Once the PCT for images of patient P_1 is populated, an overall sum of counts (SoCs) is computed to show the total counts for benign and malignant predictions for patient P_1 . Hence, on the basis of SoCs, all images of a patient are classified as either benign or malignant. Take the same example discussed in McRI algorithm Section 4.3.2(a), where the computed counts for seven images are $I_1(2,1), I_2(1,2), I_3(2,1), I_4(0,3), I_5(2,1), I_6(2,1)$, and $I_7(3,0)$, which are now stored in PCT. Here, in PCT the SoCs for benign and malignant predictions are 12 and 9, respectively. It can be observed that the benign class has higher SoCs than the malignant class. Therefore, all images of patient P_1 are classified as benign. Conversely, in PCT if malignant SoCs become larger than benign SoCs, then

all images of that particular patient are classified as malignant, see McRP Algorithm 4.2. Thus, in the McRP algorithm, all the images are recognized by their relevant group (i.e., augmented patient-wise group of images) instead of classifying each original image individually. Classifying images through its related group can drastically reduce the misclassification rate compared to the McRI algorithm for BrC Hp images using the BreakHis dataset. The same as the McRI algorithm, here patient confidence is also computed for each patient by using SoCs such as 12/21 (i.e., 57.14%) (see lines 17 and 20 of the McRP Algorithm 4.2).

Algorithm 4.2: McRP Algorithm

```

Input: TrainedClassifier, PatientNoList, TestingImages
Output: AllPredictedLabels PW, AllPatientsConfidence PW
Function MisclassificationReduction_PatientWise
1       NoPatients = count(PatientNoList);
2       For PNum = 1: NoPatients
3           SinglePatientImages = GetPatientImages(TestingImages,PNum)
4           NoImages = count(SinglePatientImages)
5           For NoI = 1:NoImages
6               Img = ReadImage(SinglePatientImages(NoI))
7               AugImages = AugmentImage3Times(Img)
8               PredictedLbIs = Predict(TrainedClassifier,TestingDeCAFs)
9               CntB=CountBenignPrediction(PredictedLbIs)
10              CntM=CountMalignantPredictions(PredictedLbIs)
11              Populate PCT by CntB and CntM
12          End
13          SoCB= CountBenignPredictions(PCT)
14          SoCM= CountMalignantPredictions(PCT)
15          If TotalCntB >= TotalCntM
16              Populate SinglePatientPredictedLbIs with 'Benign'
17              PatientConfidence = SoCB /(SoCB + SoCM)*100
18          Else
19              Populate SinglePatientPredictedLbIs with 'Malignant'
20              PatientConfidence = SoCM /(SoCB + SoCM)*100
21          End
22          Populate AllPredictedLabels_PW by SinglePatientPredictedLbIs
23          Populate AllPatientsConfidence PW by PateintNo and PatientConfidence
24      End
25      Return AllPredictedLabels PW, AllPateintsConfidence PW
End

```

(c) *McR Confidence-wise*

In algorithms McRI and McRP, patient confidence is calculated after the labels of all images of each patient are predicted. Patient confidence is formally denoted by the following Equation (19):

$$\text{Patient confidence} = \frac{I_c}{I_t} \quad (19)$$

where I_c represents the number of correctly classified images of a patient and I_t denotes a total number of images of a patient. McRI algorithm (Algorithm 4.1) computes the image-wise confidence (IWC) whereas the McRP algorithm (Algorithm 4.2) calculates the patient-wise confidence (PWC).

However, in the McRC algorithm (Algorithm 4.3), the average patient confidence (APC) is calculated by using IWC and PWC. Suppose, if APC of patient P_1 is above the minimum value found in APC, then patient P_1 images are labeled according to the labels predicted through the McRP algorithm. Otherwise, labels are assigned by using the McRI algorithm. Thus, by using average confidence based on the McRI algorithm and McRP algorithm predictions, the misclassification has been drastically reduced. It is because the misclassification performed at the patient-level (McRP Algorithm) has been corrected by image-level (McRI algorithm) predictions, see Algorithm 4.3.

Algorithm 4.3: McRC Algorithm

<p>Input: AllPatientsConfidence_IW, AllPredictedLabels_IW, AllPatientsConfidence_PW, AllPredictedLabels_PW, PatientNoList</p> <p>Output: AllPredictedLabels_PCW</p> <p>Function MisclassificationReduction PatientConfidenceWise</p> <pre> 1 AvgPatientsConfidence = (AllPatientsConfidence_IW + AllPatientsConfidence_PW)/2 2 NoPatients = count(PatientNoList) 3 For PatientNo = 1 to NoPatients 4 If AvgPatientsConfidence(PatientNo) > minimum(AvgPatientsConfidence) 5 Populate SinglePatientPredictedLbIs by PredictedLabels_PW 6 Else 7 Populate SinglePatientPredictedLbIs by PredictedLabels_IW 8 End 9 Populate AllPredictedLabels_PCW by SinglePatientPredictedLbIs 10 End 11 Return AllPredictedLabels_PCW End </pre>
--

4.4 Development of BrT Classification Model

The proposed DL-based hierarchical BrT classification model (i.e., BMIC-Net) is composed of three classifiers namely BC_1 , B_2 , and M_2 . Each one of the three classifiers is created by using AlexNet after fine-tuning the last layer for their target classes of BrT. For instance, BC_1 is trained to classify the basic types of BrT like benign or malignant. Whereas B_2 and M_2 are enabled to classify further four subtypes of each benign and

malignant BrT. This hierarchical design of BMIC-Net enables the model to classify eight subtypes of BrT in a systematic way (indirectly) the BreakHis dataset is organized. Moreover, to enhance the performance of the BMIC-Net model a feature selection algorithm is implemented to select the minimum number features subset for misclassification reduction to achieve maximum accuracy. The detailed discussion for the design and development of the proposed BMIC-Net and feature selection algorithms are presented in the following section.

4.4.1 Proposed BMIC-Net

As mentioned earlier in Section 3.2.1.1, the collected corpus comprised only 7909 images belonging to eight BrT types. Thus, the available small number of images may not be highly effective for any DL-based model to train from scratch. In such cases, the pre-trained DL-based classification model plays a decisive role in the classification of new types of images, such as medical images. Moreover, it can be easily and quickly retrained on new images to obtain a reasonable classification performance using a smaller number of images, less computational resources (like a personal desktop computer), and training time. Thus, one of the objectives of this research is to train a DL-based classification model on a smaller number of medical images by using the least computation resources with lesser training time. Therefore, this research used the AlexNet pre-trained classification model that can be retrained on a small number of images.

AlexNet architecture possesses a fewer number of layers compared with most of the pre-trained DL-based CNN architectures. To achieve the objectives, this study used TL by fine-tuning AlexNet to construct three classifiers (BC_1 , B_2 , and M_2) for the BMIC-Net hierarchical classification model. In TL, the first part of AlexNet architecture which is based on the convolution layer is responsible to extract low- and middle-level-features.

Whereas, the second part consist of fully connected layers extracts high-level feature. The last layer of AlexNet is trained for 1000 classes of natural images using the ImageNet dataset. Thus need to be fine-tuned for eight subtypes of BrT of the BreakHis dataset. Thus in this research, the last fully connected layer is fine-tuned and retrained for the newly specified number of classes (i.e., eight types of BrT) instead of 1000 classes of natural images used by AlexNet in default.

Hence, the use of a hierarchical model using pre-trained layers reduced the computational resources and training time. Therefore, the proposed model can be trained on a normal desktop computer and requires fewer images to get better results, which enabled the achievement of the main objective of this research.

4.4.1.1 BMIC-Net Architecture and Model Structure

Figure 4.4, shows the architecture of the proposed BMIC-Net hierarchical model three classifiers (i.e., BC₁, B₂, and M₂) and non-hierarchical model to classify eight subtypes of BrT. For the proposed model, each one of the three classifiers is derived from a pre-trained fine-tuned AlexNet model for target BrT classes. Therefore, BC₁ is fine-tuned to classify basic types of BrT namely benign and malignant. However, B₂ is fine-tuned to classify four subtypes of benign tumors namely, A, F, TA, and PT, see B₂ architecture in Figure 4.4. Whereas, M₂ is a fine-tuned form of AlexNet to classify malignant BrC types namely DC, LC, MC, and PC, see the M₂ architecture diagram in Figure 4.4. Moreover, a non-hierarchical model is given to show the classification of eight subtypes of BrT by fine-tuning the last layer of AlexNet for overall eight classes, see Figure 4.4.

4.4.1.2 BMIC-Net Model Structure

The three classifiers (i.e., BC₁, B₂, and M₂) of the BMIC-Net hierarchical classification model are planted into two levels. In addition, a feature reduction and selection algorithm is also used to enhance the classification performance (by reducing misclassification)

after each classifier at each level, see Figure 4.5. At the first level, the BC_1 classifier is placed to classify a Hp image into two basic types of BrT namely, benign and malignant, formally represented by Equation (13). Whereas at the second level rest of the two classifiers B_2 and M_2 are employed. B_2 further classifies a BrC Hp image into four subtypes of benign (i.e., A, F, TA, and PT), formally represented by Equation (14). Whereas the M_2 classifier is responsible to classify BrC Hp images for four subtypes of malignancy (i.e., DC, LC, MC, and PC), formally represented by Equation (15).

$$Y_{BC_1}(I) = \text{Softmax}(I.W + b) \quad (13)$$

$$Y_{B_2}(I|Y_{BC_1}) = \text{Softmax}(I.W + b) \text{ , if } Y_{BC_1} \text{ is Benign} \quad (14)$$

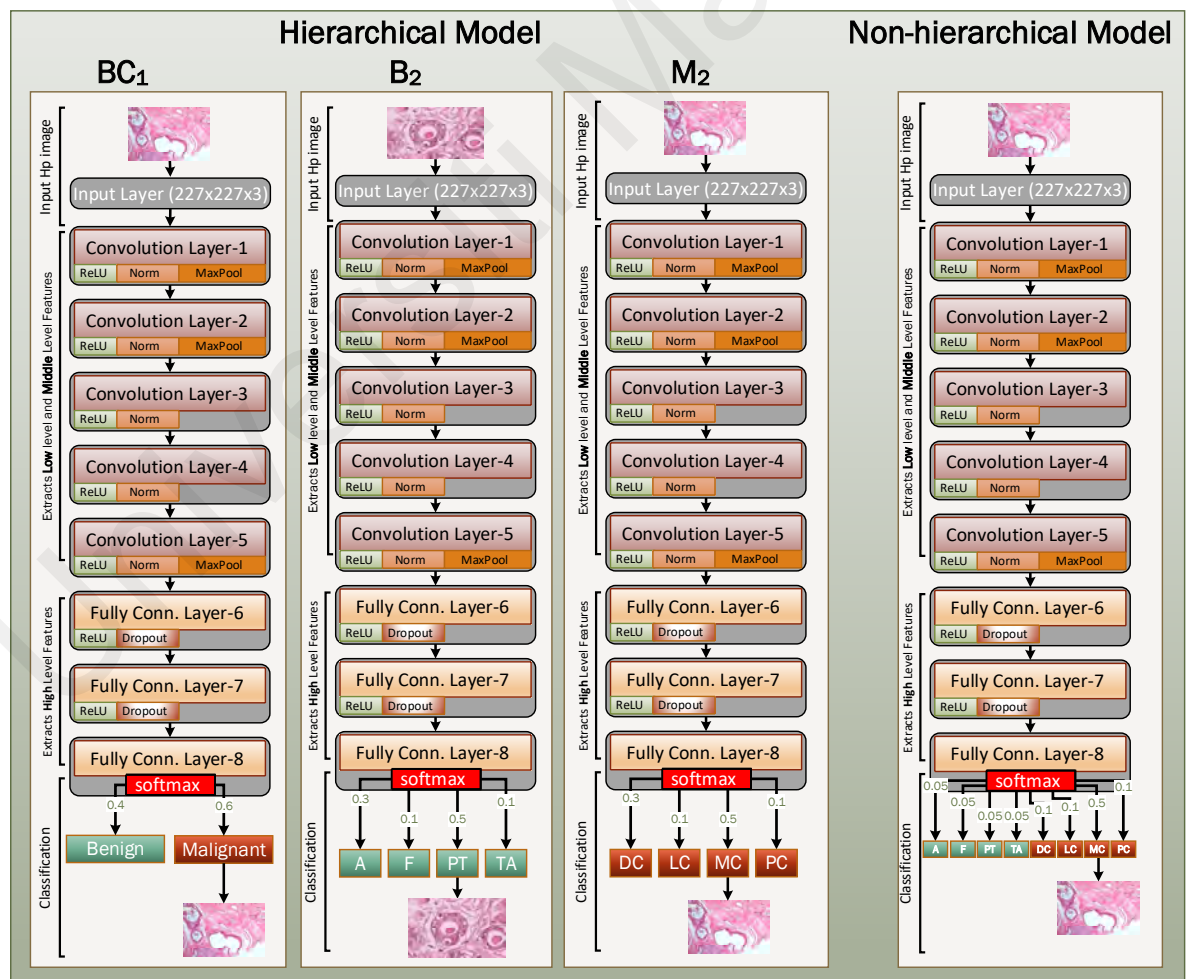


Figure 4.4: Network architectures' of BMIC-Net model classifiers' for hierarchical BrT classification

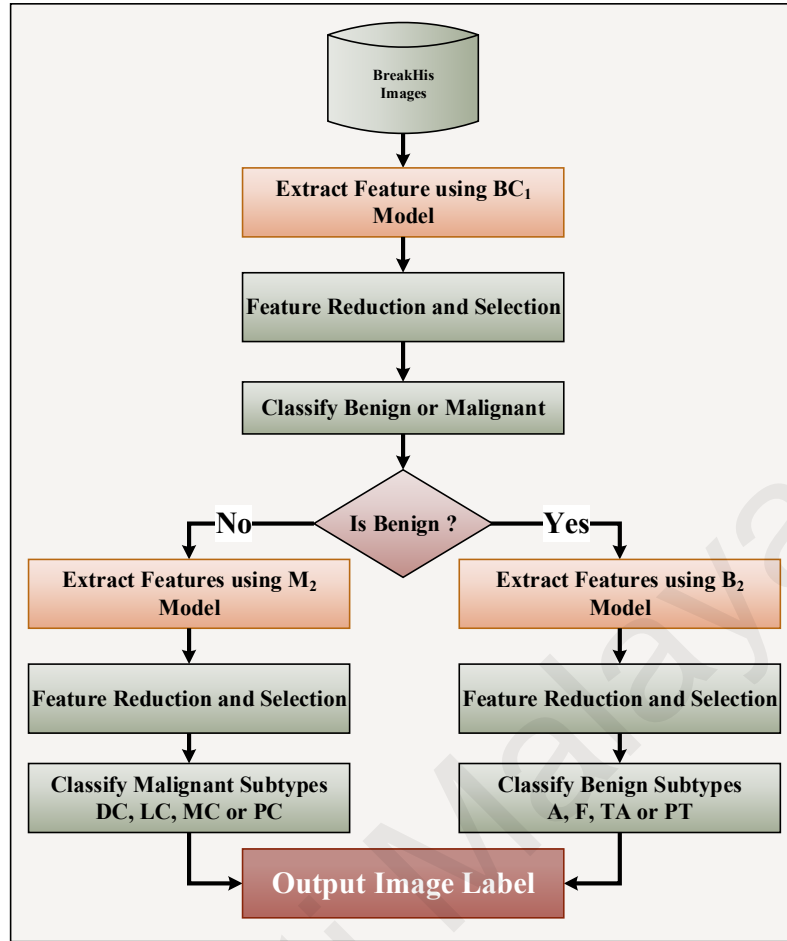


Figure 4.5: Proposed hierarchical classification flow diagram

$$Y_{M_2}(I|Y_{BC_1}) = \text{Softmax}(I.W + b) , \text{ if } Y_{BC_1} \text{ is Malignant} \quad (15)$$

Where, I , W , and b represent input image, weights and biases respectively computed by BC_1 , B_2 , and M_2 classifiers. Y_{BC_1} represents the classification probabilities predicted for benign and malignant classes. Y_{B_2} denotes the classification probabilities predicted for four subclasses of benign BrC. Similarly, Y_{M_2} denotes the classification probabilities predicted for four subclasses of malignant BrC.

In summary, the proposed hierarchical BrT classification model structure is simpler and easier to classify four subtypes of benign separately from four subtypes of malignant BrT instead of overall eight BrT types directly. Thus, an indirect multiclassification approach (i.e., a hierarchical classification approach) produced better results compared to

direct classification (i.e., a non-hierarchical classification approach). This is the main notion of choosing a hierarchical classification model.

4.5 Performance Enhancement of BrT Classification Model

This section describes the second contribution made for the BrT classification model. In this regard, a feature selection algorithm is created using two feature reduction schemes like IG and PCA to get the minimum number of features to reduce the misclassification which enhances the overall performance of the hierarchical BrT classification model. The next subsection 4.5.1 describes the design and working of the feature reduction and selection algorithm.

4.5.1 Feature Reduction and Selection Algorithm

As mentioned in Section 3.2.2.4, after MFVs extraction using three classifiers (i.e., BC_1 , B_2 , and M_2) at each classification level, six ML classifiers are applied, namely, the kNN($k=1$), SVM(linear), NB, DT, LDA classifier, and LR. These six ML classifiers are applied to see the performance of ML classifiers in terms of mean A_c for five folds of extracted MFVs. Furthermore, according to Wolpert and Macready (1997), no single ML algorithm can perform consistently better on all types of data. Thus, the performance of various algorithms on the collected dataset must be evaluated to investigate which one produces better classification results on the collected dataset. Hence, this study selected the six aforementioned ML classifiers to evaluate their performances on the extracted three MFVs using BC_1 , B_2 , and M_2 .

Generally, it has been experimentally observed that the misclassification is because of two reasons. First, there is a high correlation among the features of eight subtypes of BrC Hp images. Which may create complexity for the ML classifier to distinguish eight subtypes of BrT. Therefore, the misclassification rate is higher. Second, a large number of features (i.e., 4096) in MFV is extracted through BMIC-Net for each BC_1 , B_2 , and M_2 .

Such a large number of features can easily distract the training process of a ML classifier that can increase the misclassification rate. Moreover, such a large number of features in MFV may not be feasible for effective ML classifiers to obtain the highest Ac within limited computational time and resources. Thus, the three MFVs are analyzed and optimized using two well-known feature reduction schemes, namely, IG and PCA, to obtain the most informative and discriminative feature subset, see Algorithm 4.4.

The feature reduction process is based on three steps. In the first step, a feature score table (FST) is created using MFV. In the second step, a feature accuracy table is generated using the FST. Finally, the highest accuracy is achieved when the least number of feature subsets is used. There are three major reasons behind the selection of IG and PCA feature reduction methods. First, in several studies, these methods have shown promising results compared to other methods (Bovis, Singh, Fieldsend, & Pinder, 2000; Swiniarski, Lim, Shin, & Skowron, 2006; Naik et al., 2008; Buciu & Gacsadi, 2009; Surendiran & Vadivel, 2010; Buciu & Gacsadi, 2011; Zhang, Tomuro, Furst, & Raicu, 2012; Babu, Sukumar, & Anandan, 2013; Kozegar, Soryani, Minaei, & Domingues, 2013). Second, images mostly have highly correlated features due to similarity among neighboring pixels. However, real-life Hp images usually possess some noise/inconsistencies due to different color, intensity, and lighting effects because of image acquisition protocols and different standards followed in digital pathology labs. Thus entropy-based feature selection (like IG) method helps to find out the purity of contribution for each dimension towards the intended class label (Kent, 1983). Third, for high dimensional data like Hp images, PCA is mostly used in order to handle the curse of dimensionality without losing important information. Moreover, variant information in the data needs to be preserved. Thus, PCA is a well-established mathematical technique for reducing the dimensionality of images and keeps the embedded information variations as its maximum (Abdi & Williams, 2010).

Algorithm 4.4: Feature reduction schemes adopted

```

Input: TrFeatures, TsFeatures, TrLabels, TsLabels, FeatureReductionMethod, FeatureWindowSize
Output: Trained Six Classifiers on optimum features subset
Procedure TrainOnOptimumFeatureSubset(Feature, InputLabels, FeatureReductionMethod)
1  if FeatureReductionMethod is PCA
2    [TrFeatures, TsFeatures] ← PCA(TrFeatures, TsFeatures)
3  elseif FeatureReductionMethod is IG
4    [TrFeatures, TsFeatures] ← InformationGain(TrFeatures, TsFeatures)
5  endif
6  [MinFeaKNN, MaxAccKNN, FeaAccTableKNN] ← OptimumFeatureSubsetExtraction(TrFea, TsFea, TrLbIs, TsLbIs, 'km', FeatureWindowSize)
7  [MinFeaSVM, MaxAccSVM, FeaAccTableSVM] ← OptimumFeatureSubsetExtraction(TrFea, TsFea, TrLbIs, TsLbIs, 'svm', FeatureWindowSize)
8  [MinFeaNB, MaxAccNB, FeaAccTableNB] ← OptimumFeatureSubsetExtraction(TrFea, TsFea, TrLbIs, TsLbIs, 'nb', FeatureWindowSize)
9  [MinFeaDT, MaxAccDT, FeaAccTableDT] ← OptimumFeatureSubsetExtraction(TrFea, TsFea, TrLbIs, TsLbIs, 'tree', FeatureWindowSize)
10 [MinFeaLDA, MaxAccLDA, FeaAccTableLDA] ← OptimumFeatureSubsetExtraction(TrFea, TsFea, TrLbIs, TsLbIs, 'lda', FeatureWindowSize)
11 [MinFeaLR, MaxAccLR, FeaAccTableLR] ← OptimumFeatureSubsetExtraction(TrFea, TsFea, TrLbIs, TsLbIs, 'lr', FeatureWindowSize)
12 Plot 2D line graph by using FeaAccTableKNN, FeaAccTableSVM, FeaAccTableNB, FeaAccTableDT, FeaAccTableLDA, FeaAccTableLR
13 end

```

After applying feature reduction schemes a feature selection algorithm (Algorithm 4.5) is developed and implemented to select the minimum features subset for each of the classifiers used in the hierarchical classification approach. Noticeably, the selected feature subset does not compromise the overall Ac (lies between 0 to 100) of the hierarchical classification model. Overall 82 feature subsets (50, 100, 150, ..., 4096) are created and tested by using the aforementioned six ML classifiers for each feature reduction scheme namely, IG and PCA.

Algorithm 4.5: Feature selection algorithm

```

Input: TrainingFeatures, TestingFeatures, TrainingLabels, TestingLabels, ClassifierName, FeaturesWindowSize
Output: OptimumFeaturesSubset, MaxAccuracy, FeaturesSubsetAccuracyTable
Function OptimumFeatureSubsetExtraction(TrainingFeatures, TestingFeatures, TrainingLabels, TestingLabels,
ClassifierName, FeaturesWindowSize)
1  NoPredictions = NumberOfFeatures(TestingFeatures)/FeaturesWindowSize
2  i ← 1
3  k ← 0
4  while i <= NoPredictions do
5    k ← k+FeaturesWindowSize
6    TrainingFeaturePart ← take k Features Subset from TrainingFeatures
7    TestingFeaturePart ← take k Feature Subset from TestingFeatures
8    TrainedClassifier ← Train(ClassifierName, TrainingFeaturePart, TrainingLabels)
9    PredictedLabels ← Predict(TrainedClassifier, TestingFeaturePart, TestingLabels)
10   ConfMatrix ← confusion_matrix(TestingLabels, PredictedLabels)
11   Accuracy ← Calculate Accuracy by using ConfMatrix
12   FeaturesSubsetAccuracyTable[i] ← table [k, Accuracy]
13   i ← i + 1
14 end
15 MaxAccuracy ← Maximum_Accuracy(FeaturesSubsetAccuracyTable) // (>=0 AND <=100)
16 OptimumFeaturesSubset ← FeatureSubset(Maximum_Accuracy(FeaturesSubsetAccuracyTable))
17 return OptimumFeaturesSubset ,MaxAccuracy, FeaturesSubsetAccuracyTable
18 end

```

The feature selection algorithm enables to select the minimum features subset where highest Ac (lies between 0 to 100) is observed by each of the six ML classifiers for overall 4096 DeCAFs extracted through BC₁, B₂, and M₂ DL-based BMIC-Net model. Finally, a minimum subset of features, top-performing feature reduction scheme and best

performing ML classifier are selected to construct hierarchical Br classification. Thus the main goals of feature reduction and selection are achieved because the proposed model consumes less computational resources and produced better results by reducing misclassification to enhance the overall performance of the proposed hierarchical BrT classification model.

4.6 Summary

This chapter represented the detailed discussion about architectural design (see Section 4.2.2.1 and 4.4.1.1) and algorithms for model performance enhancement (see Sections 4.3.2 and 4.5.1) to highlight the research contributions made for proposed BrC detection and BrT classification models. Overall, four research contributions are discussed, where two contributions are made for each BrC detection model and the BrT classification model. For BrC detection an ensemble DL-based EBrC-Net model is created to extract better features compared to AlexNet using less computational resources, less training time using complex datasets. EBrC-Net is enabled to accept a larger input image size compared to AlexNet. Whereas, the fully connected layers are trained from scratch to learn Br cancer lesion specific features instead of natural image specific features like AlexNet. In addition, to improve the performance of extracted features via EBrC-Net three McR reduction algorithms (i.e., McRI, McRP, and McRC) are developed to reduce the misclassification rate. On the other hand, for BrT classification a DL-based BMIC-Net hierarchical model is developed to extract better features compared to non-hierarchical model for eight subtypes of BrT. Moreover, it consumes less computational resources, less training time, and able to show better results using a complex dataset. Furthermore, a feature reduction and selection algorithm is developed to get a minimum number of features (which reduces misclassification) to enhance the overall performance of the proposed model. The results of the proposed BrC detection and BrT classification models are discussed in Chapter 5.

CHAPTER 5: EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Introduction

This chapter is largely composed of two main parts, the first part reports the experimental results while the second part is discussion. In the first part, the experimental results of both BrC detection and BrT classification models are presented. Whereas, the second part of this chapter presents a vital and hypothetical discussion about existing SoA models for BrC detection and BrT classification. The following sections provide more details of the experimental results and overall discussion.

5.2 Experimental Results

This section reports the experimental results of proposed DL-based BrC detection and BrT classification models using Hp images. An extensive set of experiments are carried out using various PEMs. Overall four experimental setups yielded four experimental results for each BrC detection model and BrT classification model, see Figure 5.1. The complete flow of experimental results is discussed in the next sections.

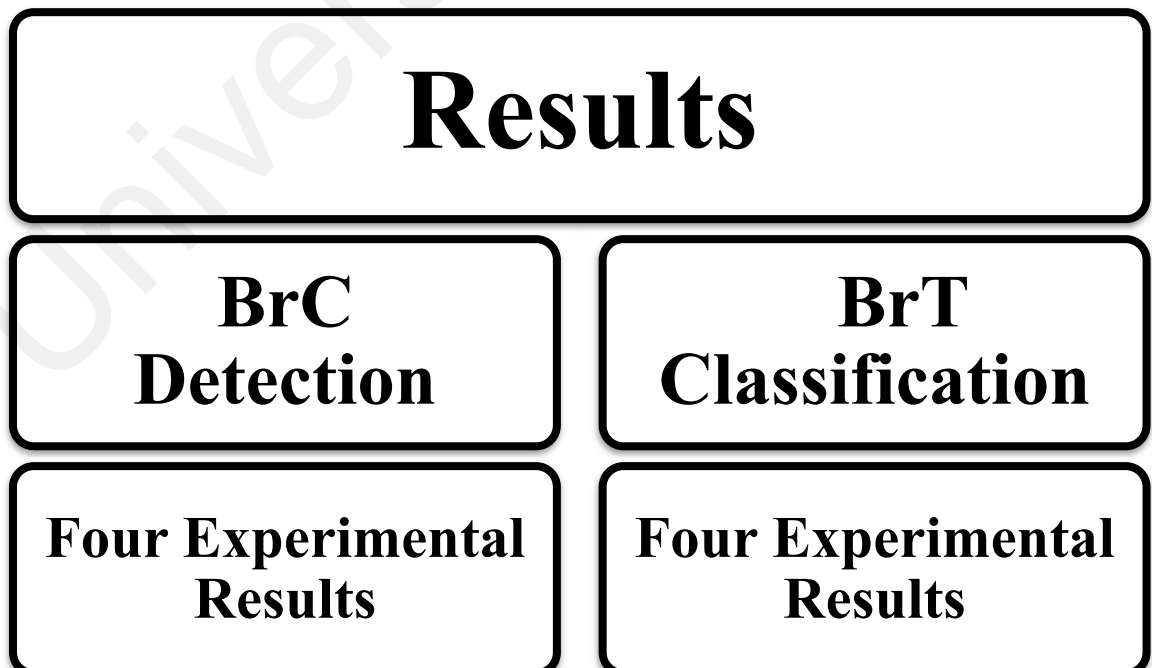


Figure 5.1: The experimental results distribution overview

5.2.1 Experimental Results of BrC Detection Model

In this section, the results of the aforementioned four experimental setups (see Section 3.1.1) are reported and analyzed. Results are discussed in terms of overall mean prediction Ac, Sp, Sn, Pr, Fm, and PRR for five folds of features.

5.2.1.1 Experimental Results of Setting I

This section comprises of two parts, first part shows the results of input image optimization for the proposed EBrC-Net model and the second part compares the training performance to select the best possible trained models for AlexNet and EBrC-Net. Here AlexNet is fine-tuned for the last layer for benign and malignant classes while keeping the rest of the layers freezed. Whereas the proposed EBrC-Net model is similar to AlexNet with few modifications like larger input image size, three unfreezed fully connected layers trained from scratch and the last layer is fine-tuned for benign and malignant classes.

The motivation behind the optimization of input image size is that the rescaling of medical images can cause the loss of BrT lesion related information. Which may reduce the quality of features extracted through CNN models. Thus, the CNN-based model may produce compromised results. Urbaniak and Wolter (2020) experimentally observed that the larger image size usually improves the medical image diagnosis accuracy using CNN. Therefore, the input image size for EBrC-Net is optimized by using three sizes like 138×138 , 227×227 , and 258×258 via minimum validation loss with better accuracy, see Figure 5.2. Here, 227×227 is the standard image size used for AlexNet and 258×258 is the maximum image size supported by the EBrC-Net. However, a smaller size 138×138 is also experimented to ensure that the CNNs using medical images showed better accuracy for large size images, see Figure 5.2. Thus, achieving better accuracy is one of the objectives of this research. From Figure 5.2, it has been experimentally observed that

258×258 image size got 0.6975 minimum validation loss and better accuracy compared to the rest of the image sizes like 138×138 and 227×227. Moreover, the lowest performance is shown by 138x138 image size. For further experiments, 258×258 image size has been selected for input to EBrC-Net on the basis of minimum validation loss as well as maximum validation accuracy, see Figure 5.2.

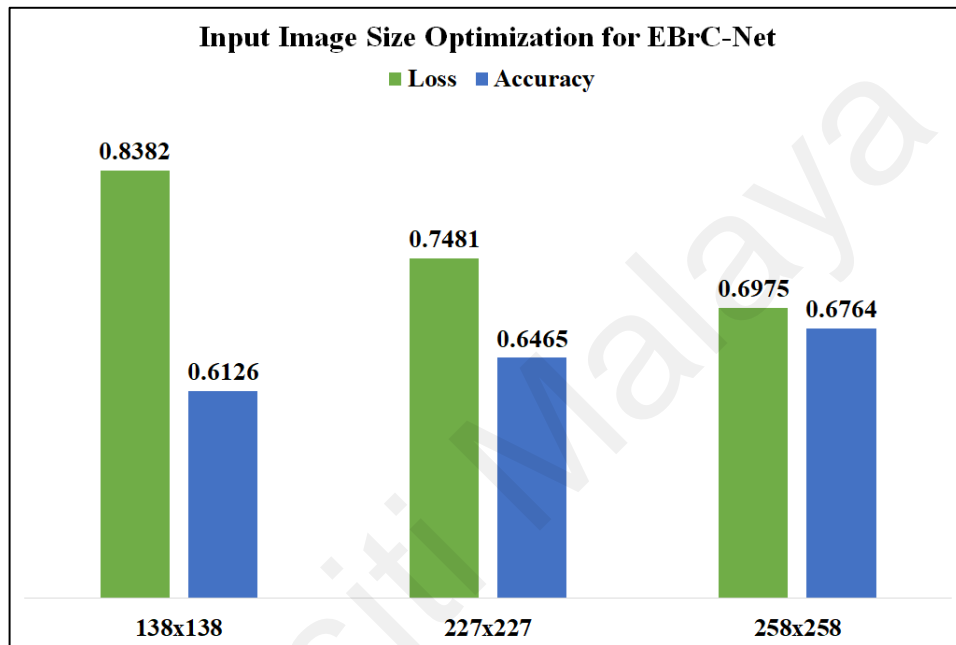


Figure 5.2: Input image size optimization for EBrC-Net

Furthermore, the training of both AlexNet and EBrC-Net is evaluated by validation accuracy followed by validation loss, see Figure 5.3. The model training is stopped if the validation loss is not decreasing until fifteen consecutive iterations.

Based on the results of validation loss (Figure 5.3), the AlexNet model training loss is mostly higher than the proposed model till the third epoch. Thus, Figure 5.3(a) presents that EBrC-Net performed better than the AlexNet, especially at epoch 2 where the minimum validation loss is 0.6975. However, the validation loss is almost the same after 3rd epoch for both models. On other hand, the analysis of validation accuracy shows that the EBrC-Net model initially has achieved lower Ac (i.e., 51.46%, 51.99%, 52.25%, and 49.6%) in comparison with AlexNet (i.e., 42.18%, 54.64%, 55.7%, and 53.58%) in first

epoch. However, after the second epoch, the performance of EBrC-Net is almost retained higher. Hence, the validation accuracy of the proposed model is highest at epoch 2 (i.e., 68.7%), and it is much better than the AlexNet model during the overall training process, see Figure 5.3(b).

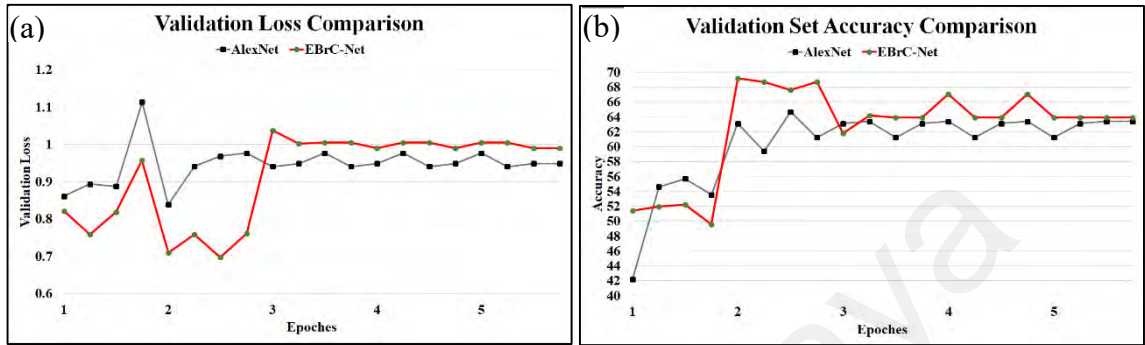


Figure 5.3: Epoch-wise comparison of the AlexNet model with EBrC-Net model. (a) Validation loss comparison. (b) The validation set accuracy comparison

Hence, it can be concluded from the above two (validation loss and validation accuracy) analyses that the EBrC-Net model at epoch 2 has shown better performance than AlexNet. This ensures that the proposed model is better than AlexNet to learn the features from Hp images due to larger input size and unfreezed fully connected layers. Therefore, EBrC-Net epoch 2 trained models are selected for further analyses to solve the BrC detection problem.

5.2.1.2 Experimental Results of Setting II

The performance of extracted DeCAFs from EBrC-Net is evaluated through six ML classifiers, namely, softmax, kNN, SVM, NB, DT, and LDA. Before performing the analysis of the aforementioned classifiers, parameter optimization has been carried out for kNN and SVM classifiers to get the best possible results. The performance of kNN is optimized by using k-values like 1,3,5,7, and 9. Where, it is experimentally noticed that kNN has shown the best results by showing 78.30% Ac, 80.41% Sn, and 71.83% Sp when the k-value is 7, see Figure 5.4(a). Thus, for kNN, the value of k is selected as 7 for further analyses. Similarly, the parameters for SVM are optimized by using different

kernels like linear, rbf and polynomial, in order to achieve the best possible results. Here, the linear kernel outperformed the rest of the two kernels by showing Ac, Sn, and Sp as 75.00%, 74.42%, and 76.76% respectively, see Figure 5.4(b). Therefore, the linear kernel is adopted for SVM for further analyses.

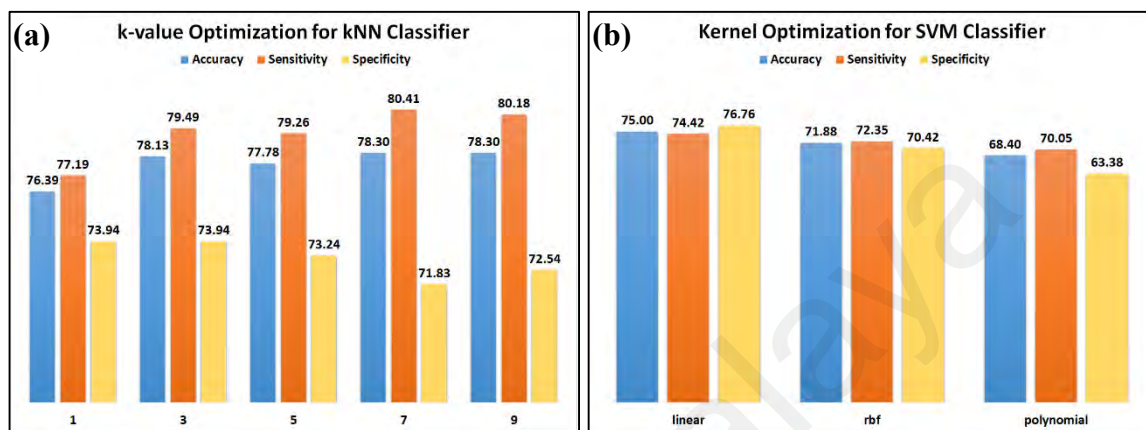


Figure 5.4: Parameter optimization of kNN and SVM classifiers

After parameter optimization, this section reports and examines the experimental results of six analyses by using the aforementioned five PEMs. These PEMs facilitated to show the performance of six ML classifiers, for further analyses of BrC detection. Figure 5.5 shows the mean results in terms of Ac, Sn, Sp, Pr, and Fm for five folds of the dataset using aforesaid six ML classifiers. Where, softmax and kNN ($k = 7$) have achieved better Ac (i.e., 83.96% and 78.26%) than the rest of the four ML classifiers. However, the Sp of softmax is lower (i.e., 56.20%) than kNN (i.e., 71.83%) and in contrast, Sn of softmax is higher (i.e., 93.09%) than kNN (i.e., 80.37%). Thus, it can be concluded that the Ac of softmax is highly biased toward the positive class (malignant class) compared to kNN. Moreover, the Pr of kNN is better than softmax such as 89.71% and 86.65%. Nonetheless, the Fm of softmax is higher than kNN like 89.73% and 84.78%. Furthermore, the rest of the four classifiers like NB, SVM(linear), LDA, and DT have shown almost the same Ac i.e., 74.76%, 74.69%, 73.99%, 73.40%. Whereas, the Sn (i.e., 80.78%) of NB is much better than SVM, LDA, and DT. Conversely, NB has shown the lowest Sp (i.e., 56.34%) than SVM, LDA, and DT. While, Pr of SVM and DT is better among all classifiers like

90.50%, and 90.25%. The Fm of SVM and LDA is almost same like 81.54%, and 81.28%, while Fm of DT is least among all classifiers i.e. 80.43%.

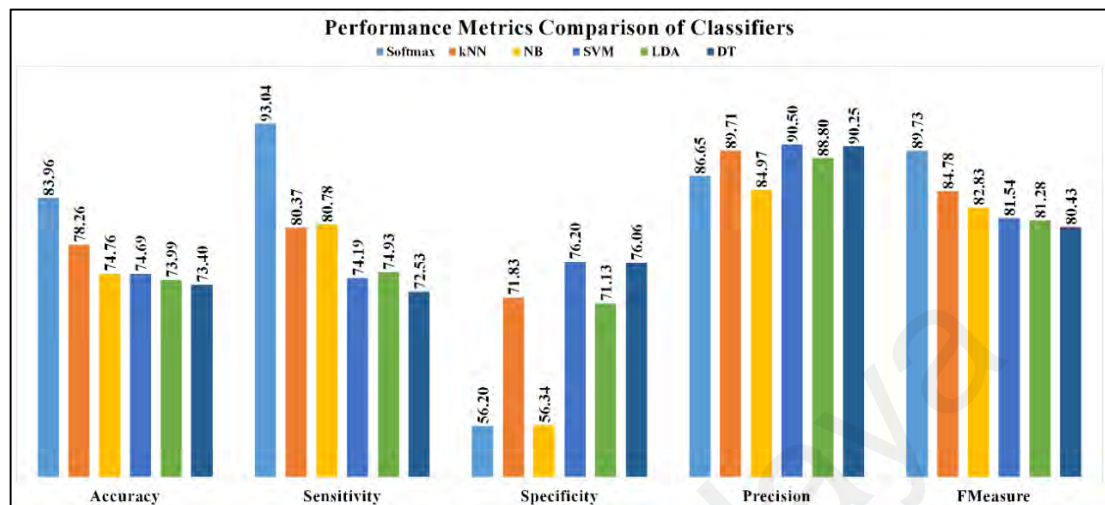


Figure 5.5. Performance of six ML classifiers

Analysis of the six ML classifiers' results has concluded that softmax and kNN have outperformed the rest of the classifiers. However, the Ac, Sn, and Fm of softmax are better than kNN. On the contrary, kNN possesses better Sp and Pr than softmax. Thus further analyses are also made on all aforementioned ML classifiers using five folds of features by implementing McR algorithms. For this experimental setting, the detailed results of five folds with standard deviation using aforesaid six ML classifiers are shown in appendix-A Table-1.

5.2.1.3 Experimental Results of Setting III

In experimental setting III, the performance of three McR algorithms, namely, McRI, McRP, and McRC, is evaluated by using six ML classifiers like kNN, softmax, NB, SVM, LDA, and DT, see Table 5.1. In addition, Table 5.1 represents the mean results in terms of Ac, Sn, Sp, Pr, and Fm for five folds of features with a standard deviation. By comparing the results of all classifiers shown in Table 5.1, it can be observed that the softmax has shown slightly better 83.16(0.31) Ac than kNN i.e., 81.25(0.39). Nonetheless, softmax is more biased toward the positive class than kNN, because softmax

is shown a high Sn (i.e., 87.37(0.31)) with a low Sp (i.e., 70.28(0.53)). In contrast, kNN has shown relatively better unbiased Sn (i.e., 80.6(0.19)) with Sp (i.e., 83.24(1.04)). Therefore, kNN has better Pr (i.e., 93.63(0.38)) than softmax (i.e., 89.99(0.18)). Whereas, Fm reported by softmax (i.e., 88.66(0.22)) is better than kNN (i.e., 86.62(0.27)). On the other hand, when the experimental results of both classifiers for the McRP algorithm are compared, then it is noticed that kNN is achieved much better performance than softmax. For instance, kNN achieved 95.76(0.20) Ac which is higher than softmax with 89.86(0.24). Moreover, kNN is shown 94.52(0.19) Sn with 99.58(0.63) Sp. Whereas, softmax remained highly biased by acquiring 97.14(0.50) Sn with 67.61(0.77) Sp. Therefore, Pr and Fm are more reliable using kNN (i.e., 99.85(0.22) and 97.11(0.13)) rather than softmax (i.e., 90.16(0.18), 93.52(0.17)). Lastly, when the results of McRC algorithm are analyzed using both classifiers, then it can be clearly perceived from Table 5.1, that kNN outperformed the softmax by showing the best Ac with 97.78(0.23); Sn, 97.28(0.19); Sp, 99.30(1.00); Pr, 99.76(0.33); and Fm as 98.51(0.15). Whereas, softmax got Ac as 91.88(0.13); Sn, 98.48(0.50); Sp, 71.69(1.29); Pr, 91.41(0.32); and Fm is noted as 94.81(0.10). Moreover, the softmax remained biased toward the malignant/positive class even after applying McR algorithms. Apart from softmax and kNN performance, it is concluded from Table 5.1 that the performance of the remaining classifiers' is also improved but shown lower results than kNN and softmax when three McR algorithms are applied one after the other. In summary, it can be summarized for this experimental setup that kNN has shown the best performance among all ML classifiers while using McRI, McRP, and McRC algorithms. The detailed results of five folds with standard deviation using three McR algorithms with aforesaid six ML classifiers are shown in appendix-A Table-2.

Table 5.1: Performance comparison of McR algorithms using machine learning classifiers

Classifier	Algorithm	Ac	Sn	Sp	Pr	Fm
kNN	McRI	81.25±0.39	80.6±0.19	83.24±1.04	93.63±0.38	86.62±0.27
	McRP	95.76±0.20	94.52±0.19	99.58±0.63	99.85±0.22	97.11±0.13
	McRC	97.78±0.23	97.28±0.19	99.30±1.00	99.76±0.33	98.51±0.15
Softmax	McRI	83.16±0.31	87.37±0.31	70.28±0.53	89.99±0.18	88.66±0.22
	McRP	89.86±0.24	97.14±0.50	67.61±0.77	90.16±0.18	93.52±0.17
	McRC	91.88±0.13	98.48±0.50	71.69±1.29	91.41±0.32	94.81±0.10
NB	McRI	80.21±0.35	79.77±0.34	81.55±0.59	92.96±0.22	85.86±0.26
	McRP	86.22±0.26	86.77±0.55	84.51±0.86	94.48±0.26	90.46±0.22
	McRC	88.23±0.15	89.26±0.55	85.07±1.44	94.82±0.45	91.95±0.12
SVM	McRI	80.49±0.40	79.91±0.44	82.25±0.59	93.23±0.22	86.05±0.31
	McRP	86.32±0.19	86.77±0.55	84.93±1.18	94.63±0.37	90.53±0.17
	McRC	87.88±0.15	88.80±0.55	85.07±1.44	94.79±0.45	91.69±0.12
LDA	McRI	81.69±3.86	81.20±4.20	83.19±2.86	93.62±1.27	86.94±2.93
	McRP	84.00±3.07	84.10±3.80	83.70±1.31	94.02±0.65	88.75±2.40
	McRC	86.86±0.87	87.49±1.15	84.91±1.21	94.66±0.40	90.93±0.65
DT	McRI	79.97±0.40	79.45±0.44	81.55±0.59	92.94±0.22	85.66±0.31
	McRP	85.45±0.19	86.31±0.55	82.81±1.18	93.89±0.36	89.94±0.17
	McRC	87.01±0.15	87.65±0.55	85.07±1.44	94.73±0.46	91.05±0.13

5.2.1.4 Experimental Results of Setting IV

In experimental setting IV, the performance of three McR algorithms is also examined by using aforesaid six ML classifiers through PRR. The PRR plays a vital role in cancer patient diagnosis because in medical science the patient-level decision is also important than making image-level prediction only.

In this setting, a total of four analyses are presented, and out of 4, 1 and 3 analyses are performed before using McR and after applying McR algorithms. Figure 5.6, shows that initially, PRR of softmax (i.e., 79.25%) is slightly better than kNN (i.e., 76.05%) when none of the three McR algorithms is used. However, when the McRI algorithm is applied, then it has been noticed that PRR of kNN (i.e., 80.82%) is improved, and it is marginally greater than softmax (i.e., 78.49%). Moreover, the application of the McRP algorithm had drastically increased PRR of kNN (i.e., 96.00%) than softmax (i.e., 84.00%). In addition, McRC had shown the best PRR by using kNN (i.e., 97.92%), whereas PRR of softmax (i.e., 91.75%) is improved but highly less than kNN. Apart from PRR analysis of kNN

and softmax classifiers, the other ML classifiers have shown better PRR as the McR algorithms are applied in a cascade manner. The PRR of NB, SVM, LDA and DT is improved from 74.65%,74.65%, 73.45%, 72.41%, and 71.11% to 90.92%, 91.75%, 90.49%, 89.47%, 88.67%, and 87.63% respectively. Furthermore, the trend line of kNN shows that the PRR has been drastically improved as the McR algorithms are applied one after the other. Thus, kNN(k=7) outperformed softmax when three McR algorithms are used in a cascaded manner.

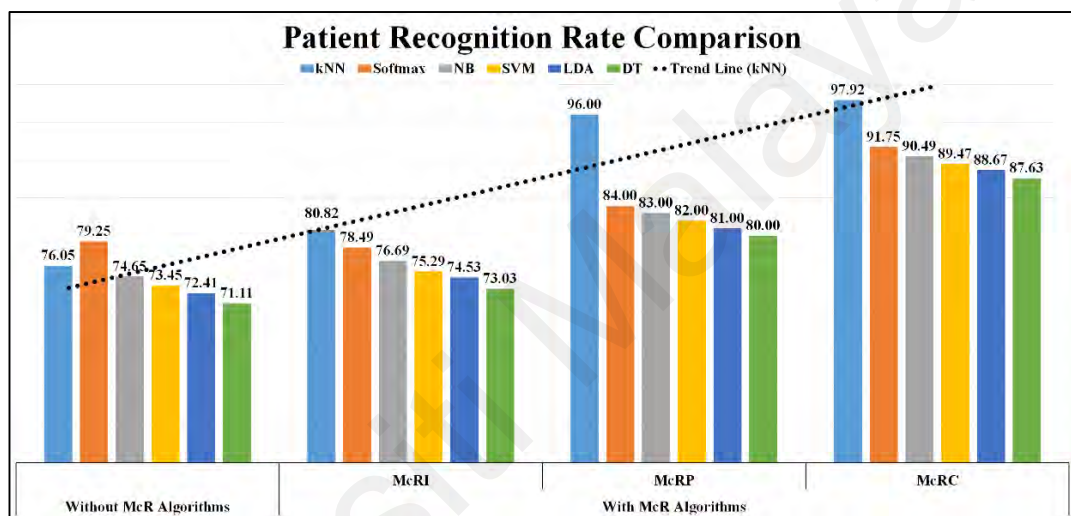


Figure 5.6. Analysis of PRR before and after McR and performance of three misclassification algorithms

In summary of all experimental setups, the DeCAFs of training, validation, and testing sets are extracted through the EBrC-Net model. Thereafter, the aforementioned six ML classifiers are trained and tested using five folds of overall DeCAFs after parameter optimization of kNN and SVM. Where kNN(k=7) has shown reliable performance compared to the rest of the classifiers. Next, the proposed three McR algorithms are used with aforesaid six ML classifiers and it has been observed that kNN has outperformed the remaining five ML classifiers. In addition, the results are compared before and after applying McR algorithms. Here, it is experimentally discovered that the proposed McR algorithms have shown drastically improved results by achieving mean accuracy up to 97.78(0.23). Finally, PRR is computed for each of the six ML classifiers and kNN has

shown the best performance among all. Thus, DeCAFs extracted via EBrC-Net and classified through kNN(k=7) using three McR algorithms have outperformed for BrC detection.

5.2.2 Experimental Results of BrT Classification Model

This section reports and discusses the experimental results of four experimental setups (see Section 3.3.2) for BrT classification in terms of overall predictive mean Ac, Sn, and AUC using five folds of features.

5.2.2.1 Setting I Experimental Results

This section presents results for the selection of the best possible trained DL-based classifiers for the proposed hierarchical BMIC-Net model and non-hierarchical model for BrT classification. The validation loss and validation accuracy are computed and analyzed after each epoch while training both models (i.e., hierarchical and non-hierarchical) for comparison. However, the training process is terminated if validation accuracy is not improving in consecutive three epochs.

It can be observed from Figure 5.7 (a) and (b) that the DL-based classifiers are trained for four epochs. Whereas, lowest validation loss (i.e., 0.3433) and highest validation accuracy (i.e., 83.92%) are observed at epoch 3. Similarly, the B₂ classifier is trained up to 18 epochs as shown in Figure 5.7 (c) and (d). The lowest validation loss (i.e., 0.6308) and highest validation accuracy (i.e., 77.96%) are observed at epoch 15. Whereas, the M₂ classifier is trained for 18 epochs. Where, the lowest validation loss (i.e., 0.66673) and highest validation accuracy (i.e., 74.06%) are observed at epoch 14, see Figure 5.7 (e) and (f). On the other hand, the non-hierarchical classifier is trained till 14 epochs. Here, the lower validation loss (i.e., 0.995) and best validation accuracy (i.e., 62.56%) are observed at epoch 10, see Figure 5.7 (g) and (h). Thus, for AlexNet the model is selected at epoch 10 for further analyses. In summation, by comparing the model training results

it can be concluded that the performance of the proposed hierarchical BrT classification model BMIC-Net is much better than the non-hierarchical classifier. Thus, the DL-based classifiers like AlexNet, BC₁, B₂, and M₂ are selected on the basis of best validation accuracy and are used to extract the features to create MFV for further analyses.

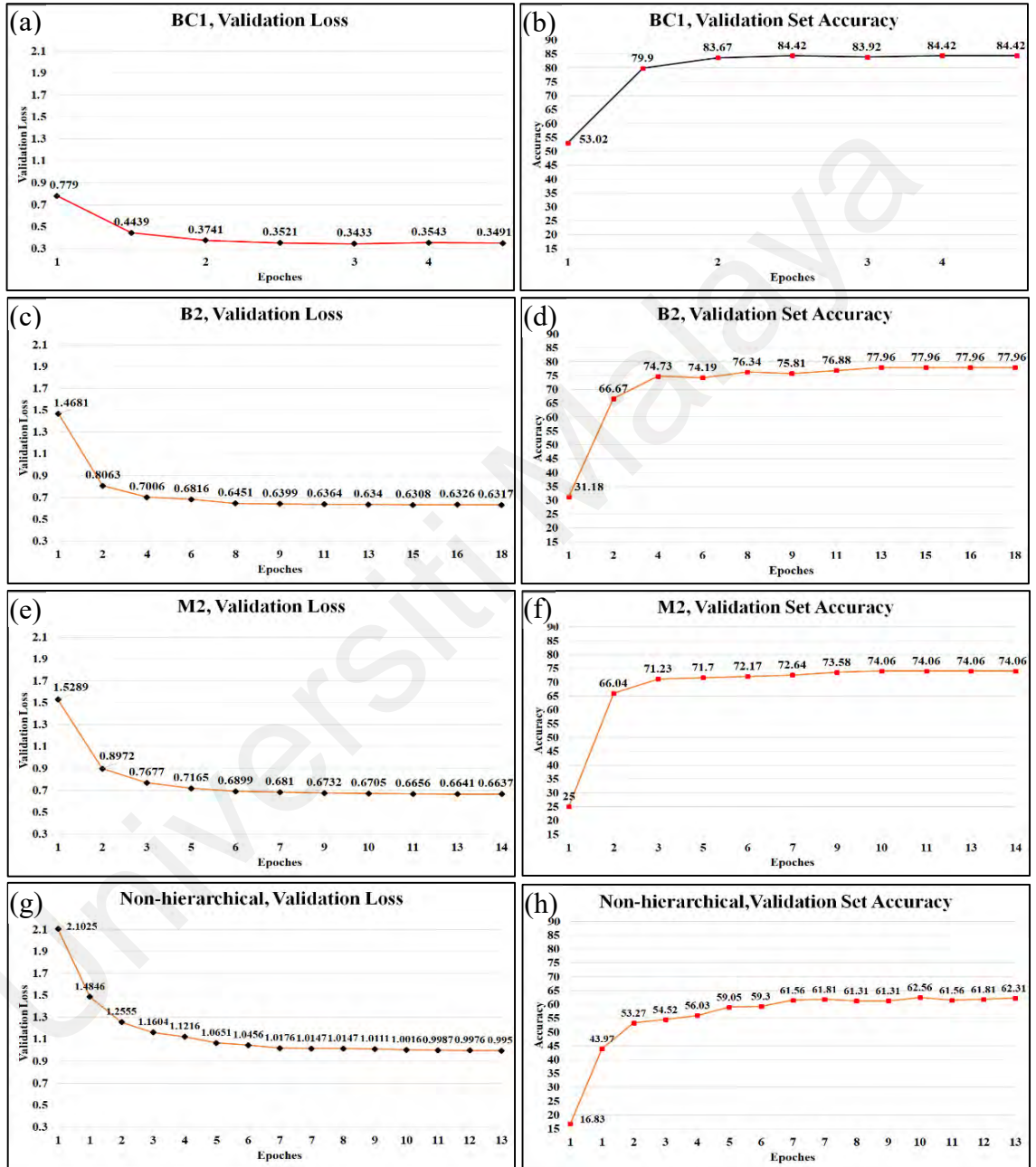


Figure 5.7: Epoch-wise comparison of BMIC-Net with non-hierarchical model

5.2.2.2 Setting II Experimental Results

The performance of extracted features from BMIC-Net (i.e., BC₁, B₂, and M₂) is evaluated through six traditional ML classifiers, namely, kNN, SVM, NB, DT, LDA, and

LR using five folds of MFV. Before performing the analyses of the aforementioned six traditional ML classifiers, the parameter optimization has been carried out for kNN and SVM to get the best possible results. The performance of kNN is optimized by using k-values like 1,3,5,7, and 9. Where, it is experimentally observed that kNN (k=1) has shown the best accuracies for BC₁, B₂, and M₂ like 94.55%, 92.13%, and 91.28% respectively, see Figure 5.8(a). Similarly, the parameters for SVM are optimized by using different kernels like linear, rbf, and polynomial. Here, the linear kernel has been outperformed the rest of the two kernels by showing the best accuracies for BC₁, B₂, and M₂ like 90.37%, 88.40%, and 88.87% respectively, see Figure 5.8(b). Therefore, the linear kernel is adopted for SVM for further analyses.

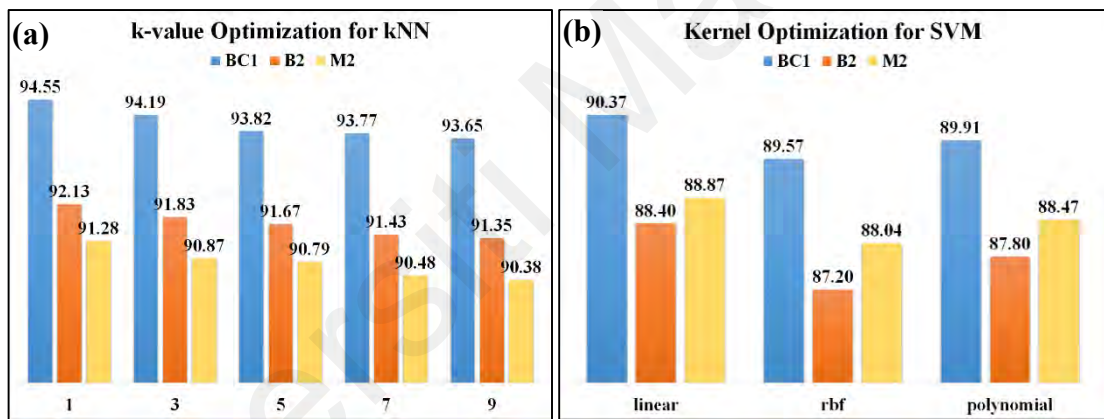


Figure 5.8: Parameter optimization for kNN and SVM

In this section, all results are presented in terms of mean Ac and AUC with standard deviation for five folds of overall 4096 DeCAFs/features. Figure 5.9, shows the mean accuracies of the three proposed BMIC-Net classifiers (i.e., BC₁, B₂, and M₂) using the six traditional ML classifiers. Here, the BC₁ classifiers, kNN (k=1) outperformed the remaining five ML classifiers by obtaining Ac of 94.33% (Sn = 91.74%, 97.17%). Conversely, DT has shown the least Ac for BC₁ classifier i.e., 84.91%. Nonetheless, the Ac and Sn of the SVM and LDA are slightly less than that of kNN. Whereas, NB and LR have shown average accuracy which slightly greater than LDA for the BC₁ classifier. In the B₂ and M₂ of BMIC-Net classifiers, the kNN outperformed the five other traditional

ML classifiers by obtaining mean accuracies of 91.88% (Sn = 95.19%, 95.15%, 85.285, 91.21%) and 91.47% (Sn = 91.80%, 95.74%, 87.79%, 90.69%), respectively, followed by the SVM and LR. In addition, the lowest mean Ac is observed in the NB. To summarize, in setting II, the best performance in all BMIC-Net classifiers is observed through kNN followed by SVM. However, in the BC₁ ML classifiers, the best performance is shown by the kNN followed by the LDA. Thus, the AUC is also shown for best-performing traditional ML classifiers for all three BMIC-Net hierarchical classifiers in Table 5.2. For this experimental setting, the detailed results of five folds with standard deviation using aforesaid six traditional ML classifiers for proposed BMIC-Net model classifiers (i.e., BC₁, B₂, and M₂) are shown in Appendix-B Table-1.

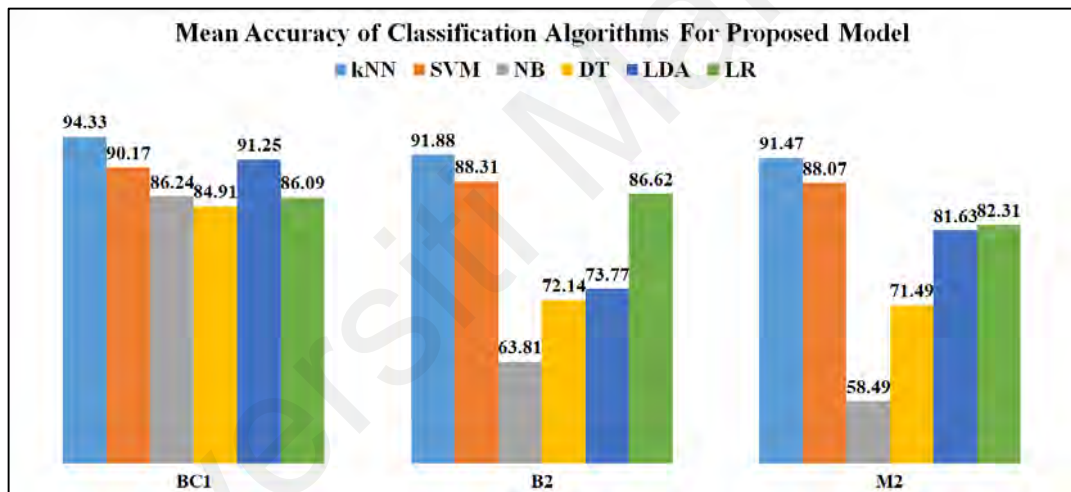


Figure 5.9. Model wise six traditional ML classifiers accuracies

Table 5.2 shows the AUC of all three BMIC-Net classifiers using kNN, SVM, NB, DT, LDA, and LR traditional ML classifiers. The highest AUC values of 0.9750, 0.9484, 0.9121, and 0.9555 are obtained by A, F, PT, and TA classes, respectively, in the B₂ classifier. However, the BC₁ classifier depicts slightly lower AUC values of 0.9455 and 0.9455 for benign and malignant classes, respectively. Conversely, the M₂ classifier shows the lowest AUC values of 0.9358, 0.9544, 0.9245, and 0.9505 for the DC, LC, MC, and PC classes, respectively. As can be witnessed from Table 5.2, the intraclass performances of each class across all three models are reasonable. Moreover, the AUC

figures of each BMIC-Net classifier shows that all three classifiers are good enough to predict BrC across the eight classes. Furthermore, the AUC shows that the proposed BMIC-Net classifiers are not either over-fitted or under-fitted or biased towards any particular class. Thus, the performance of all three proposed BMIC-Net classifiers is satisfactory and reliable. However, the AUC values of LDA are at the second-highest level followed by SVM for BC₁. Whereas, SVM got better AUC for B₂ and M₂ classifiers compared to LDA. In contrast, DT has shown the lowest AUC values for proposed model classifiers i.e., BC₁, B₂, and M₂. In conclusion, kNN(k=1) has shown better results in terms of mean Ac, Sn, and AUC for BC₁, B₂, and M₂ classifiers using five folds of features.

Table 5.2: AUC values for proposed BC₁, B₂, and M₂ classifiers

Proposed Model	Labels	kNN	SVM	NB	DT	LDA	LR
BC ₁	Benign	0.9455	0.9017	0.8644	0.8390	0.9158	0.8623
	Malignant	0.9455	0.9093	0.8644	0.8590	0.9158	0.8623
B ₂	A	0.9750	0.8874	0.6359	0.7263	0.7383	0.8619
	F	0.9484	0.8901	0.6411	0.7222	0.7306	0.8661
	PT	0.9121	0.8826	0.6336	0.7345	0.7331	0.8590
	TA	0.9556	0.8808	0.6448	0.7251	0.7378	0.8680
M ₂	DC	0.9358	0.8545	0.5841	0.7082	0.8222	0.8316
	LC	0.9545	0.9876	0.5901	0.7274	0.8190	0.8218
	MC	0.9245	0.8811	0.5775	0.7208	0.8046	0.8273
	PC	0.9505	0.9854	0.5859	0.7115	0.8119	0.8297

5.2.2.3 Setting III Experimental Results

This section reports and discusses the results using mean Ac and AUC for five folds of each of three MFVs. To recapitalize, the aim of this setting II is to obtain the best feature subset for high classification Ac and to reduce the computational time. Thus, in this setting, various feature subsets are tested with the six aforementioned ML classifiers across BC₁, B₂, and M₂ of the proposed BMIC-Net model to observe their classification performance, see Figure 5.10. Moreover, two feature reduction schemes like IG and PCA are compared to see which one elicits the best subset of features for classification.

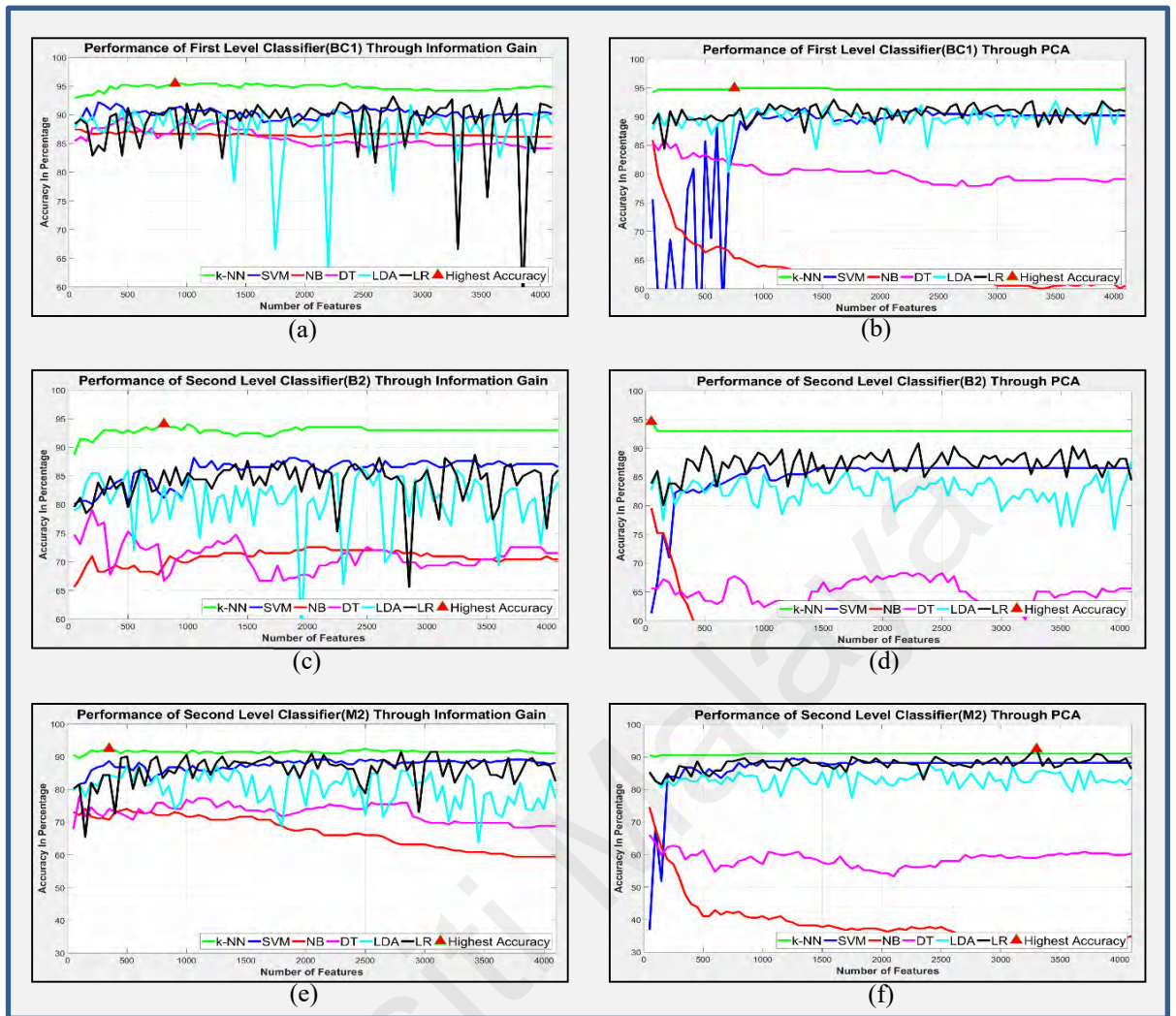


Figure 5.10: Feature reduction and selection with overall accuracies

The experimental results of these analyses are shown in Figure 5.10(a) to Figure 5.10(f). Figure 5.10(a) and Figure 5.10(b) show the overall predictive accuracies of the BC₁ classifiers across the six ML classifiers, 82 feature subsets (50, 100, 150, ... 4096), and two feature reduction schemes. IG outperformed PCA by obtaining the highest Ac of 95.33(0.31) using 900 features through kNN. In addition, Figure 5.10(c) and Figure 5.10(d) show the overall predictive accuracies of the B₂ classifier across the six ML classifiers, 82 feature subsets (50, 100, 150, ... 4096), and two feature reduction schemes. PCA performed slightly better than IG by obtaining a higher Ac of 94.70(0.62) using only 50 features through kNN. Finally, Figure 5.10(e) and Figure 5.10(f) show the overall predictive accuracies of the M₂ classifier across the six ML classifiers, 82 feature subsets (50, 100, 150, ... 4096), and two feature reduction schemes. IG marginally performed

better than PCA by obtaining a higher Ac of 92.53(0.73) using only 350 features through kNN. In sum, compared with all 4096 features, 900 feature subsets show the best performance (95.33% mean Ac, Sn = 93.45%, 97.06%) using IG and kNN in the BC₁ classifiers. In addition, compared with all 4096 features, only 50 feature subsets show the best performance (94.70% mean Ac, Sn = 96.97%, 96.55%, 93.85%, 91.11%) using PCA and kNN in the B₂ classifiers. Finally, in the M₂ classifiers, 350 out of 4096 feature subsets show the best performance (92.53% overall Ac, Sn = 88.72%, 97.87%, 91.85%, 93.02%) using IG with kNN.

Apart from mean Ac, the AUC is also calculated in Table 5.3 to observe the intraclass performance across the BC₁, B₂, and M₂ classifiers. The AUC values of the BC₁ classifiers using 900 feature subsets extracted by IG with kNN are 0.9536 and 0.9536 for the benign and malignant classes. In addition, the AUC values of the B₂ classifiers using 50 feature subsets extracted by PCA and kNN are 0.9718, 0.9621, 0.9623, and 0.9556 for the A, F, PT, and TA classes, respectively. Finally, the AUC values of the M₂ classifiers using 350 feature subsets extracted by IG and kNN are 0.9294, 0.9651, 0.9524, and 0.9529 for the DC, LC, MC, and PC classes, respectively. As shown in Figure 5.11, the intraclass performance of each class across all three BMIC-Net classifiers is satisfactory. The detailed results of five folds with standard deviation using aforesaid six traditional ML classifiers with feature reduction algorithm for proposed BMIC-Net model classifiers (i.e., BC₁, B₂, and M₂) are shown in Appendix-B Table-2.

Table 5.3: AUC values for proposed BC₁, B₂, and M₂ classifiers after feature reduction

Proposed Models	Labels	kNN	SVM	NB	DT	LDA	LR
BC ₁	Benign	0.9536	0.9357	0.8712	0.8570	0.9213	0.8759
	Malignant	0.9536	0.9293	0.8811	0.8590	0.9218	0.8759
B ₂	A	0.9718	0.9174	0.6925	0.7563	0.7783	0.8990
	F	0.9621	0.9211	0.6783	0.7622	0.7906	0.8961
	PT	0.9623	0.9126	0.6785	0.7545	0.7831	0.8890
	TA	0.9556	0.9008	0.6892	0.7451	0.7778	0.8880

Proposed Models	Labels	kNN	SVM	NB	DT	LDA	LR
M ₂	DC	0.9294	0.8945	0.6141	0.7372	0.8292	0.8716
	LC	0.9651	0.8967	0.6211	0.7274	0.8390	0.8598
	MC	0.9524	0.8825	0.6018	0.7208	0.8346	0.8573
	PC	0.9592	0.8934	0.6149	0.7415	0.8319	0.8697

Moreover, the AUC figures of each model show that all three BMIC-Net classifiers can predict BrC across the eight classes using the reduced feature subset. Furthermore, the proposed BMIC-Net model classifiers with reduced features are neither over-fitted nor under-fitted or biased toward any particular class or classes. Thus, the performance of all three proposed BMIC-Net model classifiers with reduced feature subset is better and more accurate compared with overall 4096 features. Furthermore, the top 900 features extracted through IG from the 4096 overall features should be consumed as an input to kNN when constructing the top-level BC₁ classifier. In addition, the top 50 features extracted through PCA from the 4096 overall features should be given as an input to kNN when constructing the second-level B₂ classifier. Moreover, the top 350 features extracted through IG from the 4096 overall features should serve as an input to kNN when constructing the second-level M₂ classifier. Ultimately, the constructed models should be deployed in a cascading manner to predict the eight BrT types.

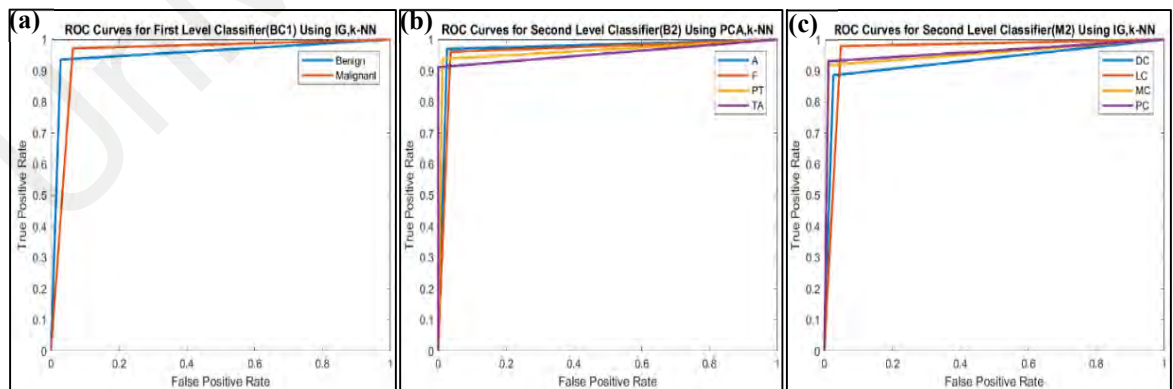


Figure 5.11: ROCs after feature reduction

5.2.2.4 Setting IV Experimental Results

This section comprises two parts, where the first part, in particular, presents 12 analyses for the non-hierarchical classifier to predict eight classes of BrT. In the second

part, further 12 analyses are made to compare the performance of the proposed BMIC-Net hierarchical model with the non-hierarchical model.

Table 5.4 represents the performance of the non-hierarchical model for both before and after feature reduction schemes. In addition, six traditional ML classifiers are examined through mean accuracy and standard deviation for five folds of features. Here, the ML classifiers parameter are optimized and remain the same as applied in the hierarchical model. Initially, before applying the feature reduction algorithm, kNN is shown slightly better Ac (i.e., 86.87(0.58)) compared to SVM like 86.80(0.18). Moreover, LDA is shown the third highest performance among all ML classifiers. Conversely, NB is shown the worst results by showing Ac like 38.43(0.64). Whereas, DT and LR are shown average performance when the feature reduction algorithm is not applied for five folds of the dataset. On the flip side, when a feature reduction algorithm is applied the kNN is shown drastically improved Ac i.e., 90.18(0.20) compared to SVM like 85.79(1.14), see Table 5.4. In contrast, NB reported the least performance by acquiring 44.06(1.10) Ac. Precisely, in 12 analyses the best performance has been observed by kNN whereas SVM got the second-highest results in terms of mean Ac and standard deviation for five folds of features via feature reduction algorithm. However, NB, DT, and LR are unable to give better results at all.

Table 5.4: Non-hierarchical model performance before and after feature reduction

	Classifier	Five Folds					Mean	Std. Dev.
		1	2	3	4	5		
Before Feature Reduction	kNN	86.84	85.94	86.90	87.24	87.44	86.87	0.58
	SVM	86.97	86.80	86.71	86.55	86.96	86.80	0.18
	NB	38.39	37.49	38.29	38.79	39.20	38.43	0.64
	DT	62.43	61.53	60.18	62.83	63.03	62.00	1.17
	LDA	85.53	84.63	84.63	85.93	86.13	85.37	0.71
	LR	60.01	59.11	59.91	61.20	60.61	60.16	0.79
After Feature Reduction	kNN	90.45	90.26	90.10	89.90	90.18	90.18	0.20
	SVM	87.65	85.00	86.10	85.30	84.90	85.79	1.14
	NB	44.93	44.03	44.83	44.33	42.20	44.06	1.10
	DT	69.33	68.43	60.18	69.73	69.93	67.52	4.14

Classifier	Five Folds					Mean	Std. Dev.
	1	2	3	4	5		
LDA	87.49	86.59	86.59	87.89	88.09	87.33	0.71
LR	64.79	63.89	64.69	61.20	65.39	63.99	1.65

Table 5.5 gives a comprehensive performance analysis by summarizing the results of the proposed hierarchical model with non-hierarchical for both before and after feature reduction approaches using mean Ac and standard deviation for five folds of features.

Table 5.5: Performance comparison of the proposed hierarchical model with the non-hierarchical model, before and after feature reduction

		Proposed Hierarchical Model			Non-Hierarchical Model
		BC ₁	B ₂	M ₂	
Before Feature Reduction	kNN	94.33±0.48	91.88±0.67	91.47±0.71	86.87±0.58
	SVM	90.17±0.43	88.31±0.36	88.07±0.67	86.8±0.18
	NB	86.24±0.51	63.81±0.78	58.49±0.82	38.43±0.64
	DT	84.91±0.84	72.14±0.49	71.49±0.63	62.00±1.17
	LDA	91.25±0.87	73.77±0.71	81.63±1.13	85.37±0.71
	LR	86.09±0.73	86.62±0.48	82.31±0.95	60.16±0.79
After Feature Reduction	kNN	95.33±0.31	94.7±0.62	92.53±0.73	90.18±0.20
	SVM	93.02±0.23	90.46±0.25	88.98±1.40	85.79±1.14
	NB	87.73±0.51	68.31±0.73	61.68±0.87	44.06±1.10
	DT	85.8±0.36	75.29±0.30	73.16±0.52	67.52±4.14
	LDA	92.11±0.74	78.10±0.99	83.79±0.65	87.33±0.71
	LR	87.55±0.24	89.5±0.31	86.31±0.80	63.99±1.65

Here, it can be concluded that the proposed model BMIC-Net classifiers (i.e., BC₁, B₂, and M₂) outperformed using kNN(k=1) when the feature reduction algorithm is applied. However, SVM(linear) has shown the second-highest performance. While remaining ML classifiers are shown compromised mean Ac. Lastly, in comparison, the non-hierarchical model has shown lower performance than the proposed BMIC-Net hierarchical model classifiers.

In summation of above mentioned four experimental setups, the proposed hierarchical model is trained and selected on the basis of maximum validation accuracy. The best possible trained models of BMIC-Net are used for feature extraction for further analyses via six ML classifiers. Next, it is experimentally noticed that kNN(k=1) has outperformed

for five folds of dataset among all six ML classifiers. Furthermore, a feature reduction algorithm is used to reduce the feature set to enhance the performance of ML classifiers using PCA and IG schemes. Where a large number of features are reduced and the performance is significantly improved by proposed model classifiers. Afterward, six ML classifiers are analyzed using reduced features. Here, it is experimentally observed that the proposed feature reduction algorithm has improved the BrT classification drastically. Finally, the results of the proposed BMIC-Net hierarchical model are compared with the non-hierarchical model to highlight the contribution of the proposed model and feature reduction algorithm. Where it is clearly examined that the proposed hierarchical model has shown better results compared to the non-hierarchical model using five folds of features, see Table 5.5.

The proposed hierarchical BrT classification model outperformed compare to the non-hierarchical classification model using Hp images due to many reasons. First, there is a great challenge in classification due to broad inconsistency in high-resolution image appearance. Second, there is a greater similarity of cancerous tissues between two borderline types of cancers like in the BreakHis dataset one of the subjects (ID: 13412) is a borderline case. The subject possesses characteristics of two cancer types like ductal and lobular carcinoma. Third, due to inhomogeneous staining, the color distribution in image slides varies among patients. Due to these inherent Hp image issues, a classifier can be jumbled and leads to a higher misclassification rate. Therefore, it is simpler and easier to classify among four subtypes of cancer instead of eight subtypes of BrT directly. Thus indirect classification using the proposed hierarchical approach enhanced the BrT classification performance using less computational resources and training time. This is the main notion of choosing a hierarchical classification model.

5.3 Discussion

This section presents a vital and hypothetical discussion about existing SoA BrC detection and BrT classification models. Moreover, it highlights the pros and cons of various types of existing models developed by the research community. This section also compares the results of proposed BrC detection and BrT classification models with existing SoA models. The following sections provide further details of the overall discussion.

5.3.1 State-of-the-art BrC Detection and Classification Models Analysis

This section reveals the theoretical analysis and imperative results of the proposed BrC detection and classification model by using the Hp images of the BreakHis dataset. The proposed model had shown enhanced performance and obtained reliable BrC diagnostic results. BreakHis dataset is a multifaceted, challenging, and publically available standard dataset. Thus, the evaluation of the proposed classification model on such type of complex dataset proved that the model is simple, computationally cost-effective, reliable, and relatively more accurate than the existing baseline SoA BrC detection and classification models. Several studies have employed BrC detection and classification models and reported high accuracy, which assists doctors in early diagnosis.

The current studies have deployed several types of BrC detection and classification models by using DL-based approaches. DL-based classification models specifically embed the feature extraction task into its training process. The auto-feature extraction methodology enables the DL-based detection/classification model to learn discriminative features in a self-taught manner. However, recent studies exploited the following three types of DL-based approaches for BrC detection and classification.

1. Models are created and trained from scratch (i.e., de-novo models).

2. Models are created using transferred learning by adopting pre-trained models, followed by a fine-tuning step (i.e., TL-based models).
3. Ensemble models are created by combining both de-novo and TL-based layers.

However, all the above mentioned types of DL-based models have some limitations. Likewise, DL de-novo models inherently need a large number of annotated (class-wise balanced) instances to avoid overfitting issues in the training process. Moreover, such type of models requires a large volume of storage capacity, very high computational power, and a longer time to train the model properly. Nonetheless, it will not be a realistic solution for medical images, because collecting medical images in large volumes with class-wise balance instances is a highly difficult task.

By contrast, TL-based and ensemble models provide computationally feasible, reliable, and faster solution for the detection/classification of smaller datasets. The main advantage of using TL-based or ensemble models are, for instance, most of the pre-trained models exploited TL is larger in size, but they can be retrained after fine-tuning for target data in very less time than the similar size of de-novo models. Moreover, they require less computational power as well as storage capacity because in TL usually the last layer is fine-tuned and retrained while keeping most of the layers frozen and their weights are not changed while retraining process. Thus, computation is required for only a last layer that is kept unfrozen, and their weights are computed in the backpropagation process. Furthermore, the frozen layers of pre-trained models adopted in the creation of TL-based ensemble models are already trained on natural images for many classes. Therefore, they can be retrained using a small number of images more efficiently and show better results than de-novo models. Thus, to overcome the limitations of the de-novo model the BrC detection model is created by the fusion of the de-novo model with the TL-based model.

Whereas, the BrT classification model is created by using a TL-based hierarchical classification approach.

5.3.2 Proposed BrC Detection Model Discussion

In this research, to overcome the issues of de-novo, a fused model EBrC-Net has been proposed. EBrC-Net is a combination of the smallest size pre-trained model (i.e., AlexNet) and the de-novo model. In AlexNet the convolution layers are responsible for the extraction of low- and medium-level features such as edges, corners, and bobs whereas fully connected layers are involved to extract high-level features or semantic features. The low- and medium-level features are most common in all types of images such as natural and medical images. However, the semantic features of specific images, namely, medical images, are entirely different from the natural images. The specific feature of BrC Hp images is like lesion size, geometrical structure, shape, etc., which are discriminative features related to medical images only. Therefore, the features learned by convolution layers of AlexNet are generic and can be adopted for BrC Hp medical images. Nonetheless, the features learned by fully connected layers of AlexNet are specific to natural images, hence cannot be used for Hp images of BrC.

Therefore, in the EBrC-Net model, the convolutional layers are adopted through TL, whereas fully connected layers are created and trained from scratch like the de-novo model. Due to TL, the weights of convolutional layers of EBrC-Net are frozen whereas fully connected layer weights remained unfrozen. The unfrozen weight of fully connected layers in EBrC-Net will allow the model to learn specific features of Hp images from scratch whereas frozen layers of convolution layers will remain the same. Hence, this ensembling of the TL-based and de-novo model training strategy reduces the computational time and consumes limited resources. Moreover, the proposed model will

be able to learn discriminative features using a small number of images like the BreakHis dataset.

Furthermore, to avoid the overfitting issue, the training set images are augmented, and an equal number of images per class has been utilized to avoid class imbalance disputes while training the EBrC-Net. Apart from image augmentation as a preprocessing task, some other necessary preprocessing tasks such as stain normalization and scaling are also carried out before initiating the training process. Reinhard's' stain normalization method is applied to remove the inconsistencies of BreakHis Hp images. The Hp images inherently carry inconsistencies due to the use of the variable quantity of coloring chemicals in staining, the concentration of colors, and preservatives in the preparation of microscopic slides. Moreover, these slides are converted into digital images by using scanners of different vendors in digital Hp labs. Therefore, the Reinhard method is used to harmonize all images without losing the structural features of the BrC lesion. However, other stain normalization methods are prone to lose such kind of important information such as Khans' (Khan et al., 2014) and Macenko method (Macenko et al., 2009).

The goal of stain normalization is to train the EBrC-Net without being distracted by Hp image inconsistencies and able to learn discriminative and more generalized features. In addition, the images are rescaled to size $258 \times 258 \times 3$ and served as input to EBrC-Net first layer, because the input image size of the first layer is modified in EBrC-Net to extract better features than AlexNet. Thereafter, EBrC-Net is trained multiple times after changing the hyper-parameters based on the trial-and-error technique. The training process continues until the minimum loss has been achieved. While training, a trained model is stored on the completion of each epoch. Finally, the best performing trained epoch model of EBrC-Net has been selected by evaluating it on validation and testing set. Lastly, the DeCAFs of training and testing sets are extracted from the trained EBrC-Net

model and classified through six ML classifiers like softmax, kNN(k=7), NB, SVM (linear), LDA, and DT after parameter optimization.

The performance of six ML classifiers has been evaluated by using five performance metrics (i.e., mean Ac, Sp, Sn, Pr, and Fm) for BrC detection (i.e., benign or malignant cases) using five folds of features. It has been experimentally observed that softmax and kNN are outperformed the rest of the classifiers. Noticeably, the mean accuracy of softmax (i.e., 83.96%) is higher than kNN (i.e., 78.26%), but the accuracy of softmax is biased and inclined toward the malignant class as has been observed by analyzing Sn (i.e., 80.37%) and Sp (i.e., 56.20%) measures. It might be due to the majority of patients (i.e., 57 out of 81) who belong to the malignant class in the BreakHis dataset. Thus, the malignant class possesses a better representation of overall cancer cases than the benign class. However, kNN possess more reliable and unbiased results (Sn = 80.78% and Sp = 71.83%) than softmax and all other classifiers.

The following reasons show prominent and reliable results by kNN classifier:

1. It often performs better if the number of instances is large enough.
2. It can be properly applied to the data which possesses a higher dimension, even if it is dispersed and have an inseparable linear boundary.
3. It works well even if data is noisy, hence show a lower misclassification rate.
4. Inherently, the kNN classifier is flexible for distance choices and well suited for multiclass data.
5. Classification model development is simpler, faster, and computationally cost-effective, producing the best results on image data (Kuramochi & Karypis, 2005).

Experiments show that the NB, SVM, LDA, and DT have shown a lower performance related to softmax and kNN. NB produced lower results due to some facts. First, NB often

works well on independent features, whereas in images neighbor pixel is highly correlated. Second, it is highly sensitive to class imbalance issues. Third, NB estimates possible likelihood values between 0 and 1, hence causes unstable results. Due to these reasons, NB might be unable to extract informative features from images (Rennie, Shih, Teevan, & Karger, 2003).

The reason behind the poor performance of the DT classifier might be because it often trains a weak and noisy classifier. It cannot generalize well and optimized DT is highly affected even if the minor change is conducted in the training set. In addition, it shows unstable performance on numeric data such as images and makes complex, larger tree splits, that is, needed to be pruned, which causes loss of useful information (Kotsiantis et al., 2007).

The causes of weak results shown by LDA maybe because it shows poor results if the data is a slightly skewed or class wise imbalance. It would be more sensitive in binary classification if a dataset is imbalanced. Furthermore, it is unsuitable for nonlinear problems, such as images that have nonlinearly spread information. It also shows better results if the interclass distance is higher, which is often not found in medical images. LDA is highly sensitive to overfitting, thus requiring careful validation and testing (Kotsiantis et al., 2007), whereas SVM is unable to perform for images because it is not appropriate for nonlinear problems and not a suitable choice for data to possess a large number of features. Furthermore, if the inter-class difference is low and data sparsity is very high, then SVM often show weak results (Byun & Lee, 2002).

The results of kNN show a high misclassification rate and room for BrC detection improvement. In this regard, three McR algorithms are developed and implemented in a cascaded manner. The McRI algorithm minimized the misclassification rate at the image level, where each image is augmented three times and classified through kNN. The

original image is classified on the basis of the majority count of three classified augmented images. Whereas the McRP algorithm reduced false predictions at the patient-level. Here, all images of a patient are augmented thrice and classified through kNN. The overall original images of a patient are classified on the basis of the maximum count of classified augmented images of a patient. In addition, both algorithms also computed the patient-level confidence. Patient-level confidence represents the ratio of correctly classified augmented images to overall augmented images. Finally, the McRC algorithm further reduces the patient-level misclassification (i.e., McRP algorithm) by using image-level McR (i.e., McRI algorithm). Thus, if any misclassification is made by the McRP algorithm will be recovered by the McRI algorithm, on the basis of patient-level confidence.

The results show that three McR algorithms successively improved the BrC classification performance. For instance, the mean accuracy before applying McR algorithms is 78.26%. However, the accuracy improved gradually using three McR algorithms (applied one after the other) such as 81.25%, 95.76%, and 97.78%. Similarly, PRR is also improved from 76.05% to 97.92% by using three McR algorithms.

The success of the proposed BrC detection model also lies in the inherent distribution of images in a patient-wise fashion. The BreakHis dataset possesses multiple images of each of the 82 patients. Thus, if a majority of the images of a patient are classified as cancerous, then it will be easier to diagnose the malignancy. Moreover, this is the common practice of doctors which is adopted for manual analysis of images to diagnose patients' malignancy. To take advantage of this idea which is based on dataset characteristics, this research came up with McR algorithms. However, it has been observed during experiments that the second and third McR algorithms have shown the best results when the maximum number of images per patient are utilized. Because a

larger number of images makes it possible to detect BrC for a patient more accurately compared to the patient who has fewer images.

5.3.3 Proposed BrC Detection Model Baseline Comparison

Table 5.6 summarizes some existing SoA studies that had developed BrC detection models by using Hp images for two classes. For instance, (Spanhol et al., 2016a; Spanhol et al., 2017; Nahid & Kong, 2018; Nahid et al., 2018) utilized the BreakHis dataset for BrC detection, i.e., benign or malignant. Spanhol et al. (2016a) used the TL-based model to achieved better average Ac (i.e., 90%) and PRR (i.e., 85.6) using GPU for three hours. Furthermore, Spanhol et al. (2017) used pre-trained AlexNet to extract DeCAFs from the last three layers. Thereafter, a feature fusion of a three-layer is performed to obtain better results. The reported Ac is 84.2%. Similarly, Nahid and Kong (2018) extracted both local and global features from BreakHis images and classified them through the CNN model containing a residual block. The Ac achieved is 91.19% by using GPU for six hours of model training. Nahid et al. (2018) developed the CNN model, guided by the unsupervised clustering method like k-means and mean-shift to reduced misclassification. After feature extraction, softmax and SVM are used for classification. The author presented 91% Ac. It can be noticed from Table 5.6, that some of the studies did not provide Sn, Fm, and PRR, which are highly important in medical diagnosis for BrT detection. However, most of the studies used high computational resources for longer training time. Whereas few of the baseline studies did not the mentioned the training time.

Conversely, the proposed EBrC-Net model is enabled to extract better features of BrT lesion due to larger input image size and unfreezed fully connected to learn specific features from scratch using Hp images. However, Hp images' general features are learned via TL-based freezed convolutional layers. Moreover, to enhance the performance of the proposed BrT detection model, three McR algorithms are implemented using kNN(k=7).

Which, reduced the misclassification rate drastically when applied in a cascade manner. Thus, the proposed EBrC-Net model is able to show better performance by extracting better features in less time duration with fewer resources compared to the aforementioned baseline models.

Table 5.6: Performance comparison of proposed EBrC-Net model to the state-of-the-art existing models using BreakHis dataset

Study Reference	Model Type	Training Duration, Resources	Ac (%)	Sn (%)	Fm (%)	PRR (%)	Limitations
(Spanhol et al., 2016a)	De-novo (CNN)	3 Hrs, GPU	85.6	Not provided	Not provided	88.6	<ul style="list-style-type: none"> • Sn and Fm are not provided. • Needs to improve model performance.
(Spanhol et al., 2017)	TL (CNN)	Not provided	84.2	Not provided	88.7	86.3	<ul style="list-style-type: none"> • Model training time duration, resources, Sn and are not provided. • Needs to improve model performance
(Nahid & Kong, 2018)	De-novo (CNN)	6.25 Hrs, GPU	92.19	94.94	98.00	Not provided	<ul style="list-style-type: none"> • PRR is not provided. • Required very high resources
(Nahid et al., 2018)	De-novo (CNN, SVM)	Not provided, CPU	91.00	Not provided	93.00	Not provided	<ul style="list-style-type: none"> • Model training time, Sn, and PRR are not provided. • Needs to improve model performance
The proposed model (EBrC-Net)	Ensemble d (CNN, kNN)	04 Hrs, CPU	97.74	97.01	98.48	97.98	<ul style="list-style-type: none"> • Trained on image patches created from WSI. Therefore, results may be different for WSI images.

It can be seen in Table 5.6, the results of the proposed BrC detection model are comparatively better (Ac=97.74%, Sn=97.01%, Fm=98.48%, PRR=97.98%) than all of the baseline studies by using less computational resources (i.e., CPU instead of GPU) in less time (i.e., four hours) (RO2 and RO4 are achieved). Hence, the proposed EBrC-Net

model is efficient and reliable and can be deployed using a desktop machine to assist doctors as a second opinion for BrC detection using Hp images.

5.3.4 Proposed BrT Classification Model Discussion

This section presents the hypothetical analysis and significant results of the proposed BMIC-Net hierarchical BrT classification model using Hp images. The proposed model obtained reliable and improved classification performance for eight subtypes of BrT. The rigorous experimental evaluation on complex, challenging, and standard publicly available datasets proved that the proposed BMIC-Net model is less complex, computationally effective, reliable, and more accurate compared with existing baseline classification models for BrT classification.

Numerous studies have proposed classification models that claim high accuracies for the early diagnosis of BrC. However, such models suffer from three major limitations. First, these models are mostly capable of predicting only two classes of BrC, namely, benign and malignant. Second, several studies have evaluated the performances of those classification models on exclusive datasets containing a low amount of training images. Thus, the reported results in existing studies may not be directly comparable and applicable on a wider scale. Finally, most of the existing classification models have been developed using traditional ML approaches, whereby the handcrafted feature extraction and selection process are performed with the help of domain experts. Thus, extracting and selecting the features manually are tiring and time-consuming tasks. To overcome the issues of classification models developed using traditional ML approaches, recent studies have developed several classification models through DL-based approaches to produce accurate predictions by involving an auto-feature extraction step. Therefore, the proposed model BMIC-Net hierarchical model is developed by fine-tuning the last layer of AlexNet for each of the classifiers like BC₁, B₂, and M₂. The model performance is enhanced by

implementing a feature selection algorithm via IG and PCA schemes. The reduced features with kNN(k=1) have drastically improved the performance of the proposed model for BrT classification using Hp images.

5.3.5 Proposed BrT Classification Model Baseline Comparison

Some studies (Han, Wei, et al., 2017; Bardou et al., 2018; Nahid et al., 2018) had employed a DL-based BrT classification model to classify eight subtypes of BrT using BreakHis dataset, see Table 5.7. For instance, Bardou et al. (2018) created a CNN-based model and trained from scratch using GPU. However, the model training time is not mentioned by the author. The classification is made through SVM and RF classifiers for eight subtypes of BrT using the BreakHis dataset. Here, the author reported Ac ranging from 83.31% to 88.23% with 84.48% Sn. Nahid et al. (2018) deployed three types of model, namely, CNN, long short-term memory (LSTM), and a fusion of CNN and LSTM using the BreakHis dataset. Softmax and SVM are used for classification and acquired 91.00% Ac and 96.00% Pr. However, in the aforementioned studies, there is a need to show model training time and other performance metrics like Sn and AUC for image-level analysis and PRR for patient-level analysis. Moreover, Han, Wei, et al. (2017) classified BrC using the BreakHis dataset through a class structure-based deep convolutional neural network (CSDCNN) model and obtained a classification Ac of 93.2% for eight subtypes of BrT. Nonetheless, the Ac obtained through CSDCNN is higher than the aforementioned baseline studies. Where distance constraint of feature space is proposed to formulate the feature space similarities of Hp images by leveraging intra-class and inter-class labels of BrC as prior knowledge. It optimized the distance of different classes features space to select the desired features. However, CSDCNN has shown poor results when directly trained on the BreakHis dataset. Thus, it was trained on the ImageNet dataset (possess 14 million natural images of 1000 categories) to construct a pre-trained CSDCNN. Afterwards, TL is performed to retrained the pre-trained

CSDCNN on the BreakHis dataset for 10 hours and 13 minutes on GPU. Whereas, the pre-training time is not mentioned by the author and it can be very long due to a large number of computations for huge ImageNet dataset based pre-training. Thus, this model is computationally very expensive and needs extensive resources in training and requires huge number of images for extracting useful features from the BreakHis dataset.

Table 5.7: Performance comparison of proposed BMIC-Net model to the state-of-the-art existing models using BreakHis Dataset

Study Reference	Model	Training Duration, Resources	Ac (%)	Sn (%)	AUC	Limitations
(Bardou et al., 2018)	De-novo (CNN, SVM, RF)	Not given, GPU	83.31 to 88.23	84.48	Not provided	<ul style="list-style-type: none"> • Needs to improve model performance. • AUC is not provided. • The model is computationally expensive.
(Nahid et al., 2018)	De-novo (CNN, SVM)	Not given, CPU	91.00	Not provided	Not provided	<ul style="list-style-type: none"> • Needs to improve model performance. • Sn and AUC are not provided. • Only image-wise classification, no PRR.
(Han, Wei, et al., 2017)	TL (CNN)	10 Hrs 13 Min + Pre-training Time, GPU	93.80	Not provided	Not provided	<ul style="list-style-type: none"> • Sn and AUC are not provided. • Computational expensive model. • Model is complex and pre-trained on ImageNet so longer pre-training time, requires a large number of images.
Proposed hierarchical model (BMIC-Net)	Hierarchical Model (CNN, kNN)	5 Hours /classifier, CPU	BC ₁ (95.33), B ₂ (94.70), M ₂ (92.53)	BC ₁ (93.45, 97.06), B ₂ (96.97, 96.55, 93.85, 91.11), M ₂ (88.7, 97.87, 91.85, 93.02)	BC ₁ (0.9536, 0.9536), B ₂ (0.9718, 0.9621, 0.9623, 0.9556), M ₂ (0.9294, 0.9651, 0.9529)	<ul style="list-style-type: none"> • Trained on BreakHis images, which are patches created from WSI. Therefore, results may be different for WSI images. • If a borderline patient is excluded in model training then it can improve the performance.

Conversely, the proposed model BMIC-Net model uses a hierarchical approach to extract discriminative features. Here, the B_2 classification model is responsible to extract features of four subtypes of BrT if the Hp image is classified as benign by the BC_1 model otherwise M_2 classification model is implemented to extract features of four subtypes of malignant BrT. Thus, due to this hierarchical approach implementation, the BMIC-Net model is efficient (i.e., used fewer computational resources and time) to extract features more accurately as compared to the non-hierarchical model. Here, a large number of features are extracted due to three models like BC_1 , B_2 , and M_2 . Thus, to extract discriminative features, which are mainly contributing to the model classification, a feature selection algorithm is implemented using IG and PCA schemes to enhance the performance of ML classifiers like kNN.

Therefore, the results of the proposed hierarchical BrT classification model are comparatively better i.e., BC_1 (mean $Ac=95.33\%$, $Sn=95.25\%$, $AUC=95.36\%$), B_2 (mean $Ac=94.70\%$, $Sn=94.62\%$, $AUC=96.3\%$), and M_2 ($Ac=92.53$, $Sn=92.87$, $AUC=95.15$), than the baseline study (Han, Wei, et al., 2017) by using less computational resources (i.e., CPU instead of GPU) in less time (i.e., used CPU for five hours per classifier) (RO3 and RO4 are achieved). The baseline study (Han, Wei, et al., 2017) used a larger ImageNet dataset (14 million natural images of 1000 classes, needs large computations to get desired results) for pre-training while the proposed model acquired better results using only BreakHis dataset, thus proposed model is trained on fewer images in less time using fewer resources. Hence, the proposed model is efficient and reliable and can be deployed using a desktop machine to assist doctors as a second opinion for BrT classification using Hp images.

5.4 State-of-the-art versus proposed models

The SoA baseline studies and the proposed BrC detection and classification models are developed to extract both local and global features using a CNN-based architecture.

Where the BrC detection SoA baseline studies used some techniques to improve their model performance. For instance, Nahid and Kong (2018) extracted features from BreakHis images and classified them through the CNN model containing a residual block. The model trained on GPU for 6.25 Hrs with 92.19% accuracy. Similarly, Nahid et al. (2018) developed a CNN model, guided by the unsupervised clustering method. After feature extraction, softmax and SVM are used for classification. The author reported 91.00% accuracy using the CPU. However, the proposed EBrC-Net is an AlexNet based CNN model used for local and global features extraction. The EBrC-Net is based on AlexNet and trained on BreakHis after two modifications. First, the input layer size is optimized and kept larger than the AlexNet to extract distinct features. Second, all fully connected layers are trained from scratch to extract better local features, nonetheless convolution layers are used from AlexNet as transfer learning to extract better global features of BrC lesion. Moreover, to improve BrC detection model performance, McR algorithms are implemented and classified through kNN. The proposed model achieved better 97.74% accuracy in less time (4 hrs) using CPU compared to aforementioned SoA baseline studies (see Table 5.6).

For BrT classification, the SoA baseline model (Han, Wei, et al., 2017) implemented CNN-based CSDCNN model. Where a distance constraint of feature space is proposed to formulate the feature space similarities of Hp images by leveraging intra-class and inter-class labels of BrC as prior knowledge. However, the model had shown poor results when directly trained on the BreakHis dataset. Therefore, it was initially trained on the ImageNet dataset to construct a pre-trained CSDCNN. Apart from pre-training time duration, CSDCNN is trained for 10 Hrs and 13 min on GPU and had shown comparable accuracy. Whereas the proposed BMICT-Net hierarchical classification model is based on three classifiers BC_1 , B_2 , and M_2 , which are created from AlexNet transfer learning. Here, BC_1 extracts benign and malignant features, while B_2 and M_2 are trained to extract

BrT features related to benign and malignant subtypes. The features extracted through the proposed hierarchical model are discriminative than non-hierarchical models for eight subtypes of BrT classification. Furthermore, to enhance the proposed classification model performance a feature selection algorithm is implemented and classified via kNN. The proposed model is shown better results compared to CSDCNN using fewer images and trained on CPU for 5hr per classifier (see Table 5.7).

5.5 Summary

This chapter focus on two parts of this research, the first experimental results and the second part is discussion. The first part of this chapter covers the entire experimental results. Where the proposed BrC detection and BrT classification model experimental setup and results are analyzed. For BrC detection the EBrC-Net is trained on the normal desktop computer till minimum validation loss is not observed. Whereas, the best performing model is selected at epoch 2 of the overall training process. The DeCAFs of the training and testing sets are extracted from EBrC-Net, and six ML classifiers (i.e., softmax, kNN, NB, SVM, LDA, and DT) are evaluated using six PEMs (i.e., Ac, Sn, Sp, Pr, Fm, and PRR) for five folds of DeCAFs. Where kNN(k=7) and softmax has shown better results compare to the rest of the four classifiers. Furthermore, three McR algorithms are developed to improve the classification results of the six ML classifiers. Next, the three McR algorithms (i.e., McRI, McRP, and McRC) are implemented one after the other to reduce the misclassification. McRI algorithm reduces wrong predictions in an image-wise fashion. Successively, the McRP algorithm further minimizes misclassification in a patient-wise manner. Whereas, the McRC algorithm utilizes the average confidence of predictions made by McRI and McRP algorithms to reduce further misclassification. The kNN results are much better and reliable than the softmax when the McR algorithms are applied in a cascaded manner. The best mean Ac shown by kNN

is 97.78%. Meanwhile, 97.28%, 99.30%, 99.76%, 98.51%, and 97.92% are mean results achieved for the other PEMs, i.e., Sn, Sp, Pr, Fm, and PRR, respectively.

The proposed BrT classification model produced promising results in comparison with the baseline models. However, for BrT classification a DL-based hierarchical classification model BMIC-Net is developed which contains three classifiers like BC₁, B₂, and M₂. The three classifiers of BMIC-Net are trained on a normal desktop computer until maximum validation accuracy is not achieved. The features are extracted to obtain the MFVs. These MFVs contained 4096 enormous features. Thus, the most discriminative features are elicited through a feature selection algorithm using IG and PCA to reduce the misclassification. Finally, the six ML classifiers are applied on extracted subsets of features to evaluate the classification performance for five folds of MFVs.

The results of several analyses showed that IG outperformed PCA in obtaining the most discriminative subset of features. Furthermore, kNN outperformed then all other ML classifiers and obtained the mean accuracies of 95.33% (Sn = 93.45%, 97.06%), 94.70% (Sn = 96.97%, 96.55%, 93.65%, 91.11%) and 92.53% (Sn = 88.72%, 97.87%, 91.85%, 93.02%) in the BC₁, B₂, and M₂ models, respectively. Finally, the results of both BrC detection and BrT classification models are compared with baseline studies and it has been concluded that proposed models are efficient (i.e., consume less computational resources and training time) and produced reliable (i.e., reduce misclassification to show better and unbiased results even using complex dataset) results and used less number of image i.e., only BreakHis dataset is used instead of large dataset like ImageNet. Thus proposed BrT classification model can be deployed on any normal desktop computer to serve as a second opinion for a doctor.

Whereas, the second part of this chapter represents an important and theoretical discussion about existing SoA BrC detection and BrT classification models. Three types

of models namely de-novo, TL-based, and ensemble models are discussed and pros, cons are also categorically emphasized. The development of the proposed BrC detection model by ensembling both the de-novo model and the TL-based model are defended. Moreover, TL-based model development for BrT classification is also justified with reasoning. Prominently, this chapter also compared the results of proposed BrC detection and BrT classification models with existing SoA baseline models. Thus it has been elucidated that the proposed BrC detection and classification models are efficient (i.e., consume less computational resources and training time), reliable (i.e., reduce misclassification to show better and unbiased results even using complex dataset) fewer images compared to the existing SoA baseline models. Chapter 6 gives a conclusion of this research work carried out for DL-based BrC detection and classification using Hp images.

CHAPTER 6: CONCLUSION

6.1 Introduction

In this thesis, initially, a literature analysis is conducted to understand the existing SoA research methodologies and to dig out the present research challenges related to BrC detection and classification using medical images. The literature analysis aims to perform hypothetical and statistical analysis of existing medical imaging modalities used, publically available standard medical imaging datasets utilized, medical image preprocessing techniques adopted, DL-based BrC classification models developed, and performance evaluation metrics used for BrC detection and classification. While conducting this extensive analysis, future directions are also identified for problem identification for this thesis work. The details of the literature review are discussed in Chapter 2.

Besides the identification of research problems of BrC detection and classification, the publically available standard dataset BreakHis is collected. BreakHis contains breast biopsy Hp images split into two main types of BrT namely benign and malignant. However, each benign and malignant BrT is further divided into four subtypes. Thus overall eight subtypes of BrT Hp images are collected in the BreakHis dataset. Moreover, BreakHis possess 7909 BrT Hp images of 82 patients with four different magnifications. Whereas, in this research benign and malignant BrT types are used for BrC detection and eight subtypes of BrT are utilized for BrT classification. Moreover, the dataset is split into training, validation, and testing sets using a random sampling method before performing any preprocessing tasks. The details of the collected dataset are discussed in Chapter 3, Sections 3.2.1.1 and 3.2.2.1.

Apart from data collection and splitting into training, validation, and testing sets, some necessary preprocessing tasks like Hp image stain normalization, image augmentation,

selection of an equal number of augmented images for each class, and the rescaling of overall images are performed before initiating the training process of DL-based BrC detection and BrT classification models. The stain normalization is required to minimize the color abnormalities inherently found in original raw Hp images so that the BrC model will not be distracted by unwanted features. Image augmentation is required to increase the number of training instances by applying basic image processing techniques. Because DL-based BrC detection and classification models will be trained properly using augmented images without facing an overfitting issue. The selection of an equal number of augmented training images per class (by random sampling) is made to avoid the biased (toward majority class) training of proposed models. However, due to the image large size of the BreakHis dataset, there is a need for image rescaling before feeding into the input layer of proposed DL-based BrC detection and BrT classification models. The details of image preprocessing tasks are discussed in Chapter 3, Sections 3.2.1.2 and 3.2.2.2.

The BrC detection (i.e., EBrC-Net) and BrT classification (i.e., BMIC-Net) models are created using AlexNet architecture and trained multiple times (using the trial-and-error method) over randomly selected hyper-parameters until the minimum validation loss or maximum validation accuracy is not observed. The EBrC-Net model is based on the ensembling of de-novo and TL-based layers. EBrC-Net is enabled to accept larger input image size compared to AlexNet to extract better features. Whereas, the fully connected layers are trained from scratch to learn Br cancer lesion related specific features instead of natural image specific features already learned by AlexNet. Whereas the BrT classification model is developed using TL with the hierarchical classification approach. Thus both models are designed to be trained efficiently using less computational resources like a normal desktop computer instead of GPU in less time. After performing extensive attempts of training (due to the trial-and-error method) using randomly selected

multiple hyper-parameters finally trained models for BrC detection and BrT classification are achieved and used for DeCAFs extraction. The details of BrC detection and BrT classification model training are discussed in Chapter 3 (Sections 3.2.1.3(b), and 3.2.2.3(b)) and Chapter 4 (Section 4.2.2 and 4.4.1).

The extracted DeCAFs are evaluated in five folds by using multiple traditional ML classifiers to ensure that the extracted features are well generalized to represent all subtypes of BrT. It has been observed that there is a higher number of misclassification is made by all ML classifiers. Thus to reduce the misclassification rate three McR algorithms (see Section 4.3.2) are developed and implemented to enhance the performance of the BrC detection model. On the other hand, due to the large number (i.e., 4096) of DeCAFs extracted for BMIC-Net hierarchical BrT classification the misclassification rate was higher. Thus, to enhance the performance of the BMIC-Net BrT classification model, the selection algorithm is developed using feature reduction schemes like IG and PCA, see Section 4.5.1. The feature selection algorithm reduces the misclassification by eliminating the unwanted features without being compromising the overall performance of the BMIC-Net hierarchical BrT classification model. The details of BrC detection and BrT classification models performance enhancement are discussed in Chapter 3 (Sections 3.2.1.4 and 3.2.2.4) and Chapter 4 (Sections 4.3 and 4.5).

Finally, multiple PEMs like Ac, Sp, Sn, Fm, PRR, and AUC are used to evaluate the performance of proposed BrC detection and BrT classification models. The use of many PEMs allows comparing the results of proposed models with existing SoA baseline models. Because, the multiple evaluation metrics have ensured that the models trained are reliable and unbiased even if the dataset is complex i.e., eight subtypes of BrT. The details of PEMs results are discussed in Chapter 3 (Sections 3.2.1.5 and 3.2.2.5). In addition, the computational resources and training time of proposed models are also

compared with baseline studies. In summary, the comparison in terms of multiple PEMs, computational resources, and training time of proposed models with baseline studies revealed that the proposed models are efficient and reliable. Hence the proposed models can be implemented where limited computer resources are available. The proposed models can assist doctors as the second opinion to detect BrC at the patient-level and classify Hp images for eight subtypes of BrT.

Due to BMICT-Net hierarchical model, the number of features is three times higher than EBrC-Net thus feature reduction/selection was essential. Moreover, the feature selection algorithm is designed to solve multiclassification problems in a hierarchical manner thus does not fit for EBrC-Net. Whereas, in EBrC-Net, the McR algorithms are designed to work for two classes only, thus does fit for the BMIC-Net hierarchical multiclassification model.

With the context of this study, each research question is answered and discussed in Chapters 2, 3, 4, and 5. This thesis concludes by revisiting the research objectives and RQs presented in Chapter 1 describing how they are achieved. The core contributions of this thesis and the limitations and future research directions are also discussed.

6.2 Reappraisal of Research Objectives and Research Questions

This section revisits the research objectives and research questions for this thesis. Moreover, it discusses the findings of each RQ of each objective briefly. Figure 6.1, shows the relationship between the research objectives and the chapters of the thesis in which these objectives are achieved and presented. It also shows the list of publications where these objectives are achieved and published.

RO1: To investigate the existing DL-based models for breast cancer detection and classification, using Hp images for early diagnosis.

To achieve this research objective, the academic literature in the field of DL-based BrC detection and classification using medical imaging modalities is reviewed by exploiting the analysis of the procedural decision in five aspects namely types of medical imaging modalities, medical imaging datasets, preprocessing techniques, types of DL-based classification models, and performance evaluation metrics. To achieve the first research objective, many studies are selected and thoroughly reviewed in the scope of the aforementioned five aspects. The findings of each RQ of objective 1 are given below:

<p>Research Objective 1 (Achieved in Chapter 2)</p> <ul style="list-style-type: none"> • Journal Paper-1: Murtaza, G., Shuib, L., Abdul Wahab, A.W. et al. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. <i>Artif Intell Rev</i> 53, 1655–1720 (2020). https://doi.org/10.1007/s10462-019-09716-5 (Published)
<p>Research Objective 2 (Achieved in Chapter 3 [Section 3.2.1] and Chapter 4 [Sections 4.2, 4.3])</p> <ul style="list-style-type: none"> • Journal Paper-2: Murtaza, G., Shuib, L., Wahab, A.W.A. et al. Ensembled deep convolution neural network-based breast cancer classification with misclassification reduction algorithms. <i>Multimed Tools Appl</i> 79, 18447–18479 (2020). https://doi.org/10.1007/s11042-020-08692-1. (Published) • Conference Paper-1: Murtaza, G., Shuib, L., Wah, T. Y., Mujtaba, G., & Mujtaba, G. (2018). Breast Cancer Classification from Histopathology Images using Deep Neural Network. Paper presented at the Data Science Research Symposium 2018. (Published) • Conference Paper-2: Murtaza, G., Shuib, L., Wahab, A. W. A., Mujtaba, G., & Raza, G. (2019). Breast cancer classification using digital biopsy histopathology images through transfer learning. Paper presented at the First International Conference on Computer Science and Engineering 2019, Indonesia. (Published).
<p>Research Objective 3 (Achieved in Chapter 3 [Section 3.2.2] and Chapter 4 [Sections 4.4, 4.5])</p> <ul style="list-style-type: none"> • Journal Paper-3: Murtaza, G., Shuib, L., Mujtaba, G. et al. Breast Cancer Multi-classification through Deep Neural Network and Hierarchical Classification Approach. <i>Multimed Tools Appl</i> 79, 15481–15511 (2020). https://doi.org/10.1007/s11042-019-7525-4 (Published)
<p>Research Objective 4 (Achieved in Chapter 5 and 6)</p> <ul style="list-style-type: none"> • Journal Paper-2: Murtaza, G., Shuib, L., Wahab, A.W.A. et al. Ensembled deep convolution neural network-based breast cancer classification with misclassification reduction algorithms. <i>Multimed Tools Appl</i> 79, 18447–18479 (2020). https://doi.org/10.1007/s11042-020-08692-1 (Published) • Journal Paper-3: Murtaza, G., Shuib, L., Mujtaba, G. et al. Breast Cancer Multi-classification through Deep Neural Network and Hierarchical Classification Approach. <i>Multimed Tools Appl</i> 79, 15481–15511 (2020). https://doi.org/10.1007/s11042-019-7525-4 (Published)

Figure 6.1: Schematic mapping of research objectives

RQ1: What are the existing DL-based models for breast cancer detection and classification, using Hp images for early diagnosis?

In the literature review, mainly two types of DL-based models are created the first de-novo and the other one is based on TL models. However, few models used the fusion of both de-novo and transferred learning model for Hp images. Each type of model has its pros and cons to be developed and used for a specific type of data. The detailed answer is given in Section 2.6.2.2.

RQ2: What are the common medical imaging modalities used for BrC detection and classification?

The literature review discovered that there are many BrC medical imaging modalities [i.e., breast X-rays (mammograms), Hp, MRI, US, and CT] used by the researchers to detect and classify breast cancer. However, mammograms and Hp images are the most commonly used medical imaging modalities for BrC detection and classification. Moreover, Hp images will give more detailed breast tissue level analysis to diagnose BrC more confidently compared to mammograms. For further details, please refer to Section 2.3.

RO2: To develop an efficient (i.e., consumes less computational resources and training time) and reliable (i.e., reduces misclassification to show better and unbiased results even using complex dataset) DL-based model for BrC detection at patient-level using Hp images.

RQ3: How to develop an efficient (i.e., consumes less computational resources and training time) and reliable (i.e., reduces misclassification to show better and unbiased results even using complex dataset) DL-based model for BrC detection at patient-level using Hp images?

To achieve this research objective, a DL-based ensembled breast cancer detection model EBrC-Net is developed using preprocessed Hp images in a patient-wise fashion. EBrC-Net is designed in such a way so that it can be trained in less time using less computational resources. EBrC-Net possesses frozen and unfrozen layers. No computation is required for the frozen weights of convolution layers in EBrC-Net. Thus, it reduced the computational time, resources and does not require a large number of images like de-novo models. The training of fully connected unfrozen layers of EBrC-

Net is performed rigorously to get generalized, discriminative DeCAFs using a normal desktop computer. The extracted DeCAFs are evaluated using five folds with six ML classifiers. The top-performing ML classifier is selected and BrC detection results are improved by implementing three misclassification reduction algorithms to detect BrC. For details of model development and training, see Chapter 3 (Sections 3.2.1.3 and 3.2.1.4) and Chapter 4 (Sections 4.2 and 4.3). The mean results of five folds are evaluated using five performance evaluation metrics and compared with baseline models. The five performance evaluation metrics shown better and unbiased results compared to existing SoA baseline models see Chapter 5, (Sections 5.3.2 and 5.3.3). Moreover, computational resources and training time of BrC detection model is also compared with baseline studies. It is observed that the proposed model showed better results using less computational time and resources. Thus it can be concluded that the proposed BrC detection model is efficient (i.e., consumes less computational resources and training time) and reliable (i.e., reduces misclassification to show better and unbiased results even using complex dataset) to be implemented for early diagnosis of BrC using Hp images. For detailed results analysis, see Chapter 5 (Sections 5.3.3).

RO3: To develop an efficient (i.e., consumes less computational resources and training time) and reliable (i.e., reduces misclassification to show better and unbiased results even using complex dataset) DL-based model for BrT classification (up to eight classes) using Hp images.

RQ4: How to develop an efficient (i.e., consumes less computational resources and training time) and reliable (i.e., reduces misclassification to show better and unbiased results even using complex dataset) DL-based model for BrT classification (up to eight classes) using Hp images?

To achieve this research objective, a TL-based hierarchical breast cancer classification model BMIC-Net is developed using preprocessed Hp images to classify eight subtypes of BrT. BMIC-Net model is composed of three classifiers (i.e., BC_1 , B_2 , and M_2). Each classifier is created by fine-tuning the last layer of AlexNet for the target number of classes. Due to hierarchical design, the BMIC-Net is simple and able to classify four subtypes of each benign and malignant tumor separately instead of classifying eight subtypes collectively. Moreover, due to separate classifiers (i.e., B_2 and M_2) in hierarchical model design, it is faster and easier to train the model for a maximum of four classes [i.e., four for benign (i.e., B_2) and four for malignant (i.e., M_2)] compared to entire eight classes of BrT collectively. Thus, due to the TL and hierarchical design of BMIC-Net, the model is efficient and required less computational resources, training time, and required fewer images to show better results. Furthermore, the training of BMIC-Net is performed thoroughly to get discriminant DeCAFs using normal desktop computers. The extracted DeCAFs are evaluated by using six ML classifiers. The top-performing ML classifier is selected and BrT classification mean results for five folds. Moreover, a feature selection algorithm is developed to reduce the misclassification of ML classifiers using feature reduction schemes like IG and PCA. For details of BrT classification model development and training see Chapter 3 (Sections 3.2.2.3 and 3.2.2.4) and Chapter 4 (Sections 4.4 and 4.5). The mean results for five folds are evaluated using three PEMs (like A_c , S_n , and AUC) and compared with baseline models. The computed performance evaluation metrics shown better and unbiased results for eight classes compared to existing SoA baseline models see Chapter 5, (Sections 5.3.4 and 5.3.5). Thus it can be concluded that the proposed hierarchical BrT classification model is efficient (i.e., consumes less computational resources and training time) and reliable (i.e., shows better and unbiased results even using complex dataset) to be implemented in any healthcare

center for early diagnosis of BrC using Hp images. For detailed results analysis, see Chapter 5, (Sections 5.3.5)

RO4: To evaluate the performance of proposed BrC detection and classification models by comparing their performances with existing state-of-the-art BrC detection and classification models.

RQ5: How to evaluate the performance of the proposed models?

To achieve this research objective, the proposed BrC detection and classification models are evaluated by using multiple evaluation metrics (like Ac, Sn, Pr, Fm, AUC, and PRR) and results are compared with the existing SoA baseline models. The Ac is the most common single value metric required to compare the results directly with baseline models. However, accuracy can be biased and may be inclined toward the majority class. Thus other performance evaluation metrics like Sn, Pr, Fm, AUC, and PRR need to be measured with accuracy. However, Sn is very important to measure in medical science to avoid misdiagnosis of a cancerous patient compared to noncancerous. Whereas, Fm and AUC metrics used in this study show that the results of the proposed model are not biased even for eight subtypes of breast cancer. Moreover, for BrC patient detection PRR is more important than image-level classification accuracy to detect cancerous patients. Thus, in this research, the use of multiple evaluation metrics shows that the proposed models have shown better and reliable performance to be implemented in real-life scenarios. The details of experimental results and baseline comparison are discussed in Chapter 5, Section 5.3.

6.3 Limitations of Proposed BrC Detection and Classification Models

Certain limitations are identified in this work

1. **Limitation due to BreakHis dataset:** The proposed DL-based BrC detection and classification models are trained on Hp images as provided by the collected BreakHis dataset. However, the images of the BreakHis dataset are image patches marked by the groups of expert pathologists from WSI Hp images. Obviously, the proposed models are trained on selected BrC Hp image patches. Therefore, the proposed models are not trained over WSI images and may show different results.
2. **Borderline patients:** In the BreakHis dataset the images of a borderline patient (ID:13412) are placed in two subtypes of malignant class (i.e., DC and LC) of BrT in the BreakHis dataset. Therefore, the proposed models are trained using a borderline case accordingly. Here, the BrT detection model will have duplicate images for malignant class, thus no effect on results. However, the proposed BrT classification model can show improved results if the borderline patient is removed.
3. **Requires many images per patient:** The BreakHis dataset provides many images of a BrC patient in order to develop patient-level BrC detection models. It is already discussed that in medical science the patient-level BrC detection is more important than the image-level diagnosis. Therefore, in this research, the proposed BrC detection model is trained by using multiple images per patient, because, it may show different results for image-level BrC detection.
4. **Multimodality medical imaging BrC diagnosis:** The proposed BrC detection and classification models are trained on single medical imaging modalities like Hp images of BrT. Therefore, the proposed model may not be able to diagnose any other type of medical imaging modalities like Breast MRI, US, and CT images.
5. **Multi-cancer Hp image diagnosis:** The proposed BrC detection and classification models are trained to diagnose BrC only for Hp images. However, Hp images can

be of other cancer types like liver, lung, or bladder. Therefore, the proposed models may not show better results for other cancer types for Hp images.

6.4 Future Research Directions

The future direction to enhance the capability for BrC detection and BrT classification models are given as follows

1. Mostly BrC detection and classification models used either Hp image patches taken from WSI or directly WSI images. Thus there is a need to develop robust BrC detection and classification models to classify both types of Hp images simultaneously. Because it will minimize the dependency of an expert pathologist to mark WSI for the extraction of Hp image patches.
2. Usually, single modality based BrC detection and classification models are created. However, multiple modalities are used by doctors to diagnose BrC. Therefore, there is a need to develop robust BrC detection models that can classify multiple imaging modalities concurrently like MG, MRI, US, CT, and PET images.
3. Hp images can be used to diagnose cancer of various parts of the body like the liver, lung, or bladder. However, the current studies mostly diagnosed only a single cancer type. Thus there is a need to develop a generic model for detection and classification to diagnose Hp images of multiple types of cancer (i.e., liver, lung, or bladder).

6.5 Summary

This chapter concludes the overall research work presented in this thesis by revisiting the research objectives and research questions. This chapter also discussed the various limitations of the proposed research and presented the future research directions in the field of BrC detection and classification using medical images.

REFERENCES

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4), 433-459.
- Abdullah-Al, N., Bin Ali, F., & Kong, Y. N. (2017). *Histopathological Breast-Image Classification With Image Enhancement by Convolutional Neural Network*. Paper presented at the 2017 20th International Conference of Computer and Information Technology, New York.
- Abdullah-Al, N., Mehrabi, M. A., & Kong, Y. N. (2018). Histopathological Breast Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering. *Biomed Research International*, 20. doi: 10.1155/2018/2362108
- Abraham, A. (2005). Artificial neural networks. *handbook of measuring system design*. doi: 10.1002/0471497398.mm421
- Aghdam, M. H., & Heidari, S. (2015). Feature selection using particle swarm optimization in text categorization. *Journal of Artificial Intelligence and Soft Computing Research*, 5(4), 231-238.
- Ahn, S. J., Kim, Y. S., Kim, E. Y., Park, H. K., Cho, E. K., Kim, Y. K., . . . Choi, H. Y. (2013). The value of chest CT for prediction of breast tumor size: comparison with pathology measurement. *World journal of surgical oncology*, 11, 130-130. doi: 10.1186/1477-7819-11-130
- Allison, K. H., Reisch, L. M., Carney, P. A., Weaver, D. L., Schnitt, S. J., O'Malley, F. P., . . . Elmore, J. G. (2014). Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. *Histopathology*, 65(2), 240-251. doi: doi:10.1111/his.12387
- Alsubaie, N., Trahearn, N., Raza, S. E. A., Snead, D., & Rajpoot, N. M. (2017). Stain deconvolution using statistical analysis of multi-resolution stain colour representation. *Plos One*, 12(1), e0169875.
- Amit, G., Ben-Ari, R., Hadad, O., Monovich, E., Granot, N., & Hashoul, S. (2017). *Classification of breast MRI lesions using small-size training sets: Comparison of deep learning approaches*. Paper presented at the Progress in Biomedical Optics and Imaging - Proceedings of SPIE.
- Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., . . . Chen, G. (2016). *Deep speech 2: End-to-end speech recognition in english and mandarin*. Paper presented at the International Conference on Machine Learning.
- Araujo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., . . . Campilho, A. (2017). Classification of breast cancer histology images using Convolutional Neural Networks. *Plos One*, 12(6), 14. doi: 10.1371/journal.pone.0177544
- Arefan, D., Talebpour, A., Ahmadinejad, N., & Asl, A. K. (2015). Automatic breast density classification using neural network. *Journal of Instrumentation*, 10(12). doi: 10.1088/1748-0221/10/12/T12002
- Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Lopez, M. A. G. (2015, 25-29 Aug. 2015). *Convolutional neural networks for mammography mass lesion classification*. Paper presented at the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).
- Babu, J. S., Sukumar, L. B., & Anandan, K. (2013). Quantitative analysis of digitized mammograms using nonsubsampling contourlets and evolutionary extreme learning machine. *Journal of Medical Imaging and Health Informatics*, 3(2), 206-213.
- Bakkouri, I., & Afdel, K. (2017). *Breast tumor classification based on deep convolutional neural networks*. Paper presented at the Proceedings - 3rd International

- Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2017.
- Bardou, D., Zhang, K., & Ahmad, S. M. (2018). Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks. *IEEE Access*, 1-1. doi: 10.1109/ACCESS.2018.2831280
- Barr, R. G. (2012). Sonographic breast elastography: a primer. *Journal of Ultrasound in Medicine*, 31(5), 773-783.
- Bayramoglu, N., Kannala, J., & Heikkila, J. (2017). *Deep learning for magnification independent breast cancer histopathology image classification*. Paper presented at the Proceedings - International Conference on Pattern Recognition.
- Bebis, G., & Georgiopoulos, M. (1994). Feed-forward neural networks. *IEEE Potentials*, 13(4), 27-31.
- Bejnordi, B. E., Zuidhof, G., Balkenhol, M., Hermsen, M., Bult, P., van Ginneken, B., . . . van der Laak, J. (2017). Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *Journal of Medical Imaging*, 4(4), 8. doi: 10.1117/1.jmi.4.4.044504
- Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1), 1-127. doi: 10.1561/22000000006
- Bevilacqua, V., Brunetti, A., Triggiani, M., Magaletti, D., Telegrafo, M., & Moschetta, M. (2016). *An optimized feed-forward artificial neural network topology to support radiologists in breast lesions classification*. Paper presented at the GECCO 2016 Companion - Proceedings of the 2016 Genetic and Evolutionary Computation Conference.
- Bovis, K., Singh, S., Fieldsend, J., & Pinder, C. (2000). *Identification of masses in digital mammograms with MLP and RBF nets*. Paper presented at the Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on.
- Breast Cancer Imaging. (2018). Breast Cancer Imaging. Retrieved 09 Sept 2018, from http://www.aboutcancer.com/breast_cancer_imaging.htm
- Brook, A., El-Yaniv, R., Isler, E., Kimmel, R., Meir, R., & Peleg, D. (2008). Breast cancer diagnosis from biopsy images using generic features and SVMs: Computer Science Department, Technion.
- Brownlee, J. (2020, Jan 10, 2020). How to Improve Performance With Transfer Learning for Deep Learning Neural Networks. from <https://machinelearningmastery.com/how-to-improve-performance-with-transfer-learning-for-deep-learning-neural-networks/#:~:text=In%20deep%20learning%2C%20transfer%20learning,on%20the%20problem%20of%20interest.>
- Buciu, I., & Gacsadi, A. (2009). *Gabor wavelet based features for medical image analysis and classification*. Paper presented at the Applied Sciences in Biomedical and Communication Technologies, 2009. ISABEL 2009. 2nd International Symposium on.
- Buciu, I., & Gacsadi, A. (2011). Directional features for automatic tumor classification of mammogram images. *Biomedical Signal Processing and Control*, 6(4), 370-378.
- Byra, M., Piotrkowska-Wroblewska, H., Dobruch-Sobczak, K., & Nowicki, A. (2017). *Combining Nakagami imaging and convolutional neural network for breast lesion classification*. Paper presented at the IEEE International Ultrasonics Symposium, IUS.
- Byun, H., & Lee, S.-W. (2002). Applications of support vector machines for pattern recognition: A survey *Pattern recognition with support vector machines* (pp. 213-236): Springer.

- Cao, J., Qin, Z., Jing, J., Chen, J., & Wan, T. (2016, 13-16 April 2016). *An automatic breast cancer grading method in histopathological images based on pixel-, object-, and semantic-level features*. Paper presented at the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI).
- Carneiro, G., Nascimento, J., & Bradley, A. P. (2015). *Unregistered Multiview Mammogram Analysis with Pre-trained Deep Learning Models*, Cham.
- Carneiro, G., Nascimento, J., & Bradley, A. P. (2017). Automated Analysis of Unregistered Multi-View Mammograms With Deep Learning. *Ieee Transactions on Medical Imaging*, 36(11), 2355-2365. doi: 10.1109/TMI.2017.2751523
- Chan, T., Jia, K., Gao, S., Lu, J., Zeng, Z., & Ma, Y. (2015). PCANet: A Simple Deep Learning Baseline for Image Classification? *IEEE Transactions on Image Processing*, 24(12), 5017-5032. doi: 10.1109/TIP.2015.2475625
- Chang, J., Yu, J., Han, T., Chang, H. j., & Park, E. (2017, 12-15 Oct. 2017). *A method for classifying medical images using transfer learning: A pilot study on histopathology of breast cancer*. Paper presented at the 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom).
- Chen, H., Jiang, W., Li, C., & Li, R. (2013). A heuristic feature selection approach for text categorization by using chaos optimization and genetic algorithm. *Mathematical problems in Engineering*, 2013.
- Chen, J. M., Li, Y., Xu, J., Gong, L., Wang, L. W., Liu, W. L., & Liu, J. (2017). Computer-aided prognosis on breast cancer with hematoxylin and eosin histopathology images: A review. *Tumor Biology*, 39(3), 12. doi: 10.1177/1010428317694550
- Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., . . . Chen, C.-M. (2016). Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Scientific Reports*, 6, 24454-24454. doi: 10.1038/srep24454
- Chris Rose, D. T., Alan Williams, Katy Wolstencroft and Chris Taylor. (2006). DDSM: Digital Database for Screening Mammography. Retrieved 19 Oct 2018, from <http://marathon.csee.usf.edu/Mammography/Database.html>
- Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., . . . Prior, F. (2013). The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6), 1045-1057. doi: 10.1007/s10278-013-9622-7
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
- Daghghi, S., Medini, T., & Shrivastava, A. (2019). Semantic Similarity Based Softmax Classifier for Zero-Shot Learning. *arXiv preprint arXiv:1909.04790*.
- Deng, C., & Perkowski, M. (2015, 18-20 May 2015). *A Novel Weighted Hierarchical Adaptive Voting Ensemble Machine Learning Method for Breast Cancer Detection*. Paper presented at the 2015 IEEE International Symposium on Multiple-Valued Logic.
- Deng, L., & Yu, D. (2014). Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, 7(3-4), 197-387. doi: 10.1561/20000000039
- Dhungel, N., Carneiro, G., & Bradley, A. P. (2017). A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical Image Analysis*, 37, 114-128. doi: <https://doi.org/10.1016/j.media.2017.01.009>
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78-87.

- Dua, D. a. K. T. (2017). UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science. Retrieved 10 Oct 2018, from <http://archive.ics.uci.edu/ml>
- Duch, W., & Jankowski, N. (1999). Survey of neural transfer functions. *Neural Computing Surveys*, 2(1), 163-212.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*: John Wiley & Sons.
- Duraisamy, S., & Emperumal, S. (2017). Computer-aided mammogram diagnosis system using deep learning convolutional fully complex-valued relaxation neural network classifier. *IET Computer Vision*, 11(8), 656-662. doi: 10.1049/iet-cvi.2016.0425
- Ehteshami, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., . . . Venancio, R. (2017). Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *Jama*, 318(22), 2199-2210. doi: 10.1001/jama.2017.14585
- Ehteshami, B., Veta, M., Johannes, v. D. P., & et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22), 2199-2210. doi: 10.1001/jama.2017.14585
- Elmore, J. G., Jackson, S. L., Abraham, L., Miglioretti, D. L., Carney, P. A., Geller, B. M., . . . Rosenberg, R. D. (2009). Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology*, 253(3), 641-651.
- Elmore, J. G., Longton, G. M., & Carney, P. A. (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA*, 313(11), 1122-1132. doi: 10.1001/jama.2015.1405
- Evans, A. J. (2011). Re: Barriers and facilitators to adoption of soft copy interpretation from the user perspective: Lessons learned from filmless radiology for slideless pathology. *J Pathol Inform*, 2011; 2: 1, Patterson et al. *Journal of pathology informatics*, 2.
- Fan, J., & Fan, Y. (2008). High Dimensional Classification Using Features Annealed Independence Rules. *Annals of statistics*, 36(6), 2605-2637. doi: 10.1214/07-AOS504
- Farahani, N., Parwani, A. V., & Pantanowitz, L. (2015). Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathol Lab Med Int*, 7, 23-33.
- Feng, Y., Zhang, L., & Yi, Z. (2018). Breast cancer cell nuclei classification in histopathology images using deep neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 13(2), 179-191. doi: 10.1007/s11548-017-1663-9
- Ferri, C., Hernández-Orallo, J., & Salido, M. A. (2003). *Volume under the ROC surface for multi-class problems*. Paper presented at the European Conference on Machine Learning.
- Fischer, A., & Igel, C. (2012). *An introduction to restricted Boltzmann machines*. Paper presented at the Iberoamerican Congress on Pattern Recognition.
- Fisher, R. (1936). Linear discriminant analysis. *Ann. Eugenics*, 7, 179.
- Fix, E. (1951). *Discriminatory analysis: nonparametric discrimination, consistency properties*: USAF school of Aviation Medicine.
- Fonseca, P., Mendoza, J., Wainer, J., Ferrer, J., Pinto, J., Guerrero, J., & Castaneda, B. (2015). *Automatic breast density classification using a convolutional neural network architecture search procedure*. Paper presented at the Progress in Biomedical Optics and Imaging - Proceedings of SPIE.

- Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition *Competition and cooperation in neural nets* (pp. 267-285): Springer.
- Gandomkar, Z., Brennan, P. C., & Mello-Thoms, C. (2018). MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. *Artificial Intelligence in Medicine*. doi: <https://doi.org/10.1016/j.artmed.2018.04.005>
- Goceri, E. (2018). *Formulas Behind Deep Learning Success*. Paper presented at the International Conference on Applied Analysis and Mathematical Modeling Istanbul, Turkey.
- Goceri, E., & Gooya, A. (2018). *On The Importance of Batch Size for Deep Learning*. Paper presented at the International conference on mathematics, Istanbul, Turkey.
- Gurcan, M. N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., & Yener, B. (2009). Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2, 147.
- Hadad, O., Bakalo, R., Ben-Ari, R., Hashoul, S., & Amit, G. (2017). *Classification of breast lesions using cross-modal deep learning*. Paper presented at the Proceedings - International Symposium on Biomedical Imaging.
- Han, S., Kang, H. K., Jeong, J. Y., Park, M. H., Kim, W., Bang, W. C., & Seong, Y. K. (2017). A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Physics in Medicine and Biology*, 62(19), 7714-7728. doi: 10.1088/1361-6560/aa82ec
- Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., & Li, S. (2017). Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model. *Scientific Reports*, 7(1). doi: 10.1038/s41598-017-04075-z
- Hannun, A., Case, C., Casper, J., Catanzaro, B., Damos, G., Elsen, E., . . . Coates, A. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.
- He, X., & Frey, E. C. (2008). The meaning and use of the volume under a three-class ROC surface (VUS). *Ieee Transactions on Medical Imaging*, 27(5), 577-588.
- Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7), 1527-1554.
- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786), 504-507.
- Hofvind, S., Hovda, T., Holen, Å. S., Lee, C. I., Albertsen, J., Bjørndal, H., . . . Park, D. (2018). Digital Breast Tomosynthesis and Synthetic 2D Mammography versus Digital Mammography: Evaluation in a Population-based Screening Program. *Radiology*, 287(3), 787-794.
- Ishtiaq, U., Abdul Kareem, S., Abdullah, E. R. M. F., Mujtaba, G., Jahangir, R., & Ghafoor, H. Y. (2019). Diabetic retinopathy detection through artificial intelligent techniques: a review and open issues. *Multimedia Tools and Applications*. doi: 10.1007/s11042-018-7044-8
- Islam, K. T., Raj, R. G., & Mujtaba, G. (2017). Recognition of traffic sign based on bag-of-words and artificial neural network. *Symmetry*, 9(8), 138.
- Jaffar, M. A. (2017). Deep Learning based Computer Aided Diagnosis System for Breast Mammograms. *International Journal of Advanced Computer Science and Applications*, 8(7), 286-290.
- Jarrett, K., Kavukcuoglu, K., & LeCun, Y. (2009). *What is the best multi-stage architecture for object recognition?* Paper presented at the Computer Vision, 2009 IEEE 12th International Conference on.

- Jiang, F., Liu, H., Yu, S., & Xie, Y. (2017). *Breast mass lesion classification in mammograms by transfer learning*. Paper presented at the ACM International Conference Proceeding Series.
- Jing, H., Yang, Y., & Nishikawa, R. M. (2012). Regularization in retrieval-driven classification of clustered microcalcifications for breast cancer. *Journal of Biomedical Imaging*, 2012, 3.
- Jirayucharoensak, S., Pan-Ngum, S., & Israsena, P. (2014). EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal*, 2014, 1-10.
- Jonathan J. James, A. R. M. W., Andrew J. Evans. (2016, Mar 2, 2016). The Breast. Retrieved 07 Sept 2018, 2016, from <https://radiologykey.com/the-breast-2/>
- Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., . . . Boulanger-Lewandowski, N. (2016). Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2), 99-111.
- Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika*, 70(1), 163-173. doi: 10.1093/biomet/70.1.163
- Khan, A. M., Rajpoot, N., Treanor, D., & Magee, D. (2014). A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering*, 61(6), 1729-1738.
- Khan, M. H. M. (2017). *Automated Breast Cancer Diagnosis Using Artificial Neural Network (ANN)*. Paper presented at the 2017 3rd Iranian Conference on Signal Processing and Intelligent Systems, New York.
- Kim, D. H., Kim, S. T., & Ro, Y. M. (2016, 20-25 March 2016). *Latent feature representation with 3-D multi-view deep convolutional neural network for bilateral analysis in digital breast tomosynthesis*. Paper presented at the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
- Komura, D., & Ishikawa, S. (2018). Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal*, 16, 34-42. doi: <https://doi.org/10.1016/j.csbj.2018.01.001>
- Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160, 3-24.
- Kozegar, E., Soryani, M., Minaei, B., & Domingues, I. (2013). Assessment of a novel mass detection algorithm in mammograms. *Journal of cancer research and therapeutics*, 9(4), 592.
- Krishna, T. H., & Rajabhushnam, C. (2019). DETECTION OF MAMMOGRAPHIC CANCER USING SUPPORT VECTOR MACHINE AND DEEP NEURAL NETWORK. *Journal of Mechanics of Continua and Mathematical Sciences*, 14(6), 156-167. doi: 10.26782/jmcms.2019.12.00013
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks*. Paper presented at the Advances in neural information processing systems.
- Kumar, A., Singh, S. K., Saxena, S., Lakshmanan, K., Sangaiah, A. K., Chauhan, H., . . . Singh, R. K. (2020). Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer. *Information Sciences*, 508, 405-421. doi: 10.1016/j.ins.2019.08.072
- Kumar, D., Kumar, C., & Shao, M. (2017, 11-14 Dec. 2017). *Cross-database mammographic image analysis through unsupervised domain adaptation*. Paper presented at the 2017 IEEE International Conference on Big Data (Big Data).

- Kumar, I., Bhadauria, H. S., Virmani, J., & Thakur, S. (2017). A classification framework for prediction of breast density using an ensemble of neural network classifiers. *Biocybernetics and Biomedical Engineering*, 37(1), 217-228. doi: 10.1016/j.bbe.2017.01.001
- Kumar, I., H.S. B., Virmani, J., & Thakur, S. (2017). A classification framework for prediction of breast density using an ensemble of neural network classifiers. *Biocybernetics and Biomedical Engineering*, 37(1), 217-228. doi: 10.1016/j.bbe.2017.01.001
- Kuramochi, M., & Karypis, G. (2005). Gene classification using expression profiles: A feasibility study. *International Journal on Artificial Intelligence Tools*, 14(04), 641-660.
- Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2), 574-582.
- Landgrebe, T. C., & Duin, R. P. (2008). Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5), 810-822.
- Lebron, L., Greenspan, D., & Pandit-Taskar, N. (2015). PET Imaging of Breast Cancer: Role in Patient Management. *PET Clinics*, 10(2), 159-195. doi: <https://doi.org/10.1016/j.cpet.2014.12.004>
- Leod, P. M., & Verma, B. (2016). *Polynomial prediction of neurons in neural network classifier for breast cancer diagnosis*. Paper presented at the Proceedings - International Conference on Natural Computation.
- Liao, B., Xu, J., Lv, J., & Zhou, S. (2015). An Image Retrieval Method for Binary Images Based on DBN and Softmax Classifier. *IETE Technical Review*, 32(4), 294-303. doi: 10.1080/02564602.2015.1015631
- Liu, F., Hernandez-Cabronero, M., Sanchez, V., Marcellin, M. W., & Bilgin, A. (2017). The Current Role of Image Compression Standards in Medical Imaging. *Information*, 8(4), 131.
- Lo, C., Shen, Y.-W., Huang, C.-S., & Chang, R.-F. (2014). Computer-aided multiview tumor detection for automated whole breast ultrasound. *Ultrasonic Imaging*, 36(1), 3-17. doi: 10.1177/0161734613507240
- Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., & Zhang, G. (2015). Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 80, 14-23. doi: <https://doi.org/10.1016/j.knosys.2015.01.010>
- Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., . . . Thomas, N. E. (2009). *A method for normalizing histology slides for quantitative analysis*. Paper presented at the Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on.
- McCann, M. T., Ozolek, J. A., Castro, C. A., Parvin, B., & Kovacevic, J. (2015). Automated histology analysis: Opportunities for signal processing. *IEEE Signal Processing Magazine*, 32(1), 78-87.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4), 115-133. doi: 10.1007/BF02478259
- MFMER. (2018, 2018, March 22). Breast MRI. from <https://www.mayoclinic.org/tests-procedures/breast-mri/about/pac-20384809>
- Moon, M., Cornfeld, D., & Weinreb, J. (2009). Dynamic contrast-enhanced breast MR imaging. *Magnetic resonance imaging clinics of North America*, 17(2), 351-362.
- Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: toward a full-field digital mammographic database. *Acad Radiol*, 19(2), 236-248.

- Moura, D. C., & López, M. A. G. (2013). An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. *International Journal of Computer Assisted Radiology and Surgery*, 8(4), 561-574.
- Mujtaba, G., Shuib, L., Raj, R. G., Majeed, N., & Al-Garadi, M. A. (2017). Email classification research trends: Review and open issues. *IEEE Access*, 5, 9044-9064.
- Mullooly, M., Bejnordi, B. E., Pfeiffer, R. M., Fan, S. Q., Palakal, M., Hada, M., . . . Gierach, G. L. (2019). Application of convolutional neural networks to breast biopsies to delineate tissue correlates of mammographic breast density. *Npj Breast Cancer*, 5. doi: 10.1038/s41523-019-0134-6
- Nahid, A. A., & Kong, Y. (2018). Histopathological breast-image classification using local and frequency domains by convolutional neural network. *Information (Switzerland)*, 9(1). doi: 10.3390/info9010019
- Nahid, A. A., & Kong, Y. A. (2017). *Local and Global Feature Utilization for Breast Image Classification by Convolutional Neural Network*. Paper presented at the 2017 International Conference on Digital Image Computing - Techniques and Applications, New York.
- Nahid, A. A., Mehrabi, M. A., & Kong, Y. (2018). Histopathological breast cancer image classification by deep neural network techniques guided by local clustering. *Biomed Research International*, 2018. doi: 10.1155/2018/2362108
- Naik, S., Doyle, S., Agner, S., Madabhushi, A., Feldman, M., & Tomaszewski, J. (2008). *Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology*. Paper presented at the Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on.
- Nair, V., & Hinton, G. E. (2010). *Rectified linear units improve restricted boltzmann machines*. Paper presented at the Proceedings of the 27th international conference on machine learning (ICML-10).
- Nascimento, C. D. L., Silva, S. D. S., da Silva, T. A., Pereira, W. C. A., Costa, M. G. F., & Costa Filho, C. F. F. (2016). Breast tumor classification in ultrasound images using support vector machines and neural networks. *Revista Brasileira de Engenharia Biomedica*, 32(3), 283-292. doi: 10.1590/2446-4740.04915
- Nejad, E. M., Affendey, L. S., Latip, R. B., & Ishak, I. B. (2017). *Classification of histopathology images of breast into benign and malignant using a single-layer convolutional neural network*. Paper presented at the ACM International Conference Proceeding Series.
- Nweke, H. F., Teh, Y. W., Al-Garadi, M. A., & Alo, U. R. (2018). Deep Learning Algorithms for Human Activity Recognition using Mobile and Wearable Sensor Networks: State of the Art and Research Challenges. *Expert Systems with Applications*.
- Nweke, H. F., Teh, Y. W., Alo, U. R., & Mujtaba, G. (2018). *Analysis of Multi-Sensor Fusion for Mobile and Wearable Sensor Based Human Activity Recognition*. Paper presented at the Proceedings of the International Conference on Data Processing and Applications.
- Nweke, H. F., Teh, Y. W., Mujtaba, G., & Al-garadi, M. A. (2019). Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Information Fusion*, 46, 147-170.
- Oleksyuk, V., Saleheen, F., Caroline, D. F., Pascarella, S. A., & Won, C. H. (2016, 3-3 Dec. 2016). *Classification of breast masses using Tactile Imaging System and machine learning algorithms*. Paper presented at the 2016 IEEE Signal Processing in Medicine and Biology Symposium (SPMB).
- Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). *Deep face recognition*. Paper presented at the BMVC.

- Paula, E. L., Ladeira, M., Carvalho, R. N., & Marzagão, T. (2016). *Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering*. Paper presented at the Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on.
- Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.
- Pritom, A. I., Munshi, M. A. R., Sabab, S. A., & Shihab, S. (2016, 18-20 Dec. 2016). *Predicting breast cancer recurrence using effective classification and feature selection technique*. Paper presented at the 2016 19th International Conference on Computer and Information Technology (ICCIT).
- Qi, X., Wang, T., & Liu, J. (2017). *Comparison of support vector machine and softmax classifiers in computer vision*. Paper presented at the 2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE).
- Qiu, Y., Yan, S., Gundreddy, R. R., Wang, Y., Cheng, S., Liu, H., & Zheng, B. (2017). A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. *Journal of X-Ray Science and Technology*, 25(5), 751-763. doi: 10.3233/XST-16226
- Radiological Society of North America, I. R. (2018, March 09, 2018). RadiologyInfo for Patients. from <https://www.radiologyinfo.org/en/info.cfm?pg=genus>
- Rasti, R., Teshnehlal, M., & Phung, S. L. (2017). Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks. *Pattern Recognition*, 72, 381-390. doi: 10.1016/j.patcog.2017.08.004
- Rebecca Sawyer Lee, F. G., Assaf Hoogi, Daniel Rubin. (2016). Curated Breast Imaging Subset of DDSM Dataset. The breast cancer Imaging Archive. Retrieved 19 Oct 2018, from <https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM#4413fe70f2bb4159b326a3f07fa6e6a9>
- Reinhard, E., Adhikhmin, M., Gooch, B., & Shirley, P. (2001). Color transfer between images. *IEEE Computer graphics and applications*, 21(5), 34-41.
- Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). *Tackling the poor assumptions of naive bayes text classifiers*. Paper presented at the Proceedings of the 20th international conference on machine learning (ICML-03).
- Rouhi, R., Jafari, M., Kasaei, S., & Keshavarzian, P. (2015). Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Systems with Applications*, 42(3), 990-1002. doi: 10.1016/j.eswa.2014.09.020
- Rubin, R., Strayer, D. S., & Rubin, E. (2008). *Rubin's pathology: clinicopathologic foundations of medicine*: Lippincott Williams & Wilkins.
- Ruifrok, A. C., & Johnston, D. A. (2001). Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4), 291-299.
- Sadaf, A., Crystal, P., Scaranelo, A., & Helbich, T. (2011). Performance of computer-aided detection applied to full-field digital mammography in detection of breast cancers. *European Journal of Radiology*, 77(3), 457-461.
- Saidin, N., Sakim, H. A. M., Ngah, U. K., & Shuaib, I. L. (2012). Segmentation of breast regions in mammogram based on density: a review. *arXiv preprint arXiv:1209.5494*.
- Samala, R. K., Chan, H.-P., Hadjiiski, L. M., Helvie, M. A., Richter, C., & Cha, K. (2018a). Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Physics in Medicine and Biology*, 63(9), 095005-095005. doi: 10.1088/1361-6560/aabb5b
- Samala, R. K., Chan, H. P., Hadjiiski, L. M., Helvie, M. A., Cha, K. H., & Richter, C. D. (2017). Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms.

- Physics in Medicine and Biology*, 62(23), 8894-8908. doi: 10.1088/1361-6560/aa93d4
- Samala, R. K., Chan, H. P., Hadjiiski, L. M., Helvie, M. A., Richter, C., & Cha, K. (2018b). Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Physics in Medicine and Biology*, 63(9), 8. doi: 10.1088/1361-6560/aabb5b
- Schneider, M., & Yaffe, M. (2000). *Better detection: improving our chances*. Paper presented at the Digital Mammography: 5th International Workshop on Digital Mammography IWDM.
- Sert, E., Ertekin, S., & Halici, U. (2017). *Ensemble of convolutional neural networks for classification of breast microcalcification from mammograms*. Paper presented at the Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS.
- Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. *Annual review of biomedical engineering*, 19, 221-248. doi: 10.1146/annurev-bioeng-071516-044442
- Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60.
- Siddiqui, M. F., Mujtaba, G., Reza, A. W., & Shuib, L. (2017). Multi-class disease classification in brain MRIs using a computer-aided diagnostic system. *Symmetry*, 9(3), 37.
- Silva, S., Costa, M., Pereira, D. A., W.C, d., & Filho, C. (2015). *Breast tumor classification in ultrasound images using neural networks with improved generalization methods*. Paper presented at the Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS.
- Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). *Best practices for convolutional neural networks applied to visual document analysis*. Paper presented at the null.
- Smitha, P., Shaji, L., & Mini, M. (2011). *A review of medical image classification techniques*. Paper presented at the International conference on VLSI, Communication & Intrumrnataiom.
- Sohn, K., Zhou, G., Lee, C., & Lee, H. (2013). *Learning and selecting features jointly with point-wise gated Boltzmann machines*. Paper presented at the Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, Atlanta, GA, USA.
- Sophie Softley Pierce, P. M., Breast Cancer Care. (2017). Three quarters of NHS Trusts and Health Boards say 'not enough' care for incurable breast cancer patients.
- Spanhol, F. A., Oliveira, L. S., Cavalin, P. R., Petitjean, C., & Heutte, L. (2017). *Deep features for breast cancer histopathological image classification*. Paper presented at the Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on.
- Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016a). *Breast cancer histopathological image classification using Convolutional Neural Networks*. Paper presented at the Proceedings of the International Joint Conference on Neural Networks.
- Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016b). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7), 1455-1462.
- Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., . . . Kok, S. (1994). *The mammographic image analysis society digital mammogram database*. Paper presented at the Excerpta Medica. International Congress Series.

- Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., . . . Kok, S. (2015). Mammographic Image Analysis Society (MIAS) database v1. 21.
- Sun, W., Tseng, T. B., Zhang, J., & Qian, W. (2017). Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Comput Med Imaging Graph*, 57, 4-9. doi: 10.1016/j.compmedimag.2016.07.004
- Sun, Y., Chen, Y., Wang, X., & Tang, X. (2014). *Deep learning face representation by joint identification-verification*. Paper presented at the Advances in neural information processing systems.
- Surendiran, B., & Vadivel, A. (2010). Feature selection using stepwise ANOVA discriminant analysis for mammogram mass classification. *International J. of Recent Trends in Engineering and Technology*, 3(2), 55-57.
- Svozil, D., Kvasnicka, V., & Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, 39(1), 43-62.
- Swiniarski, R. W., Lim, H. K., Shin, J. H., & Skowron, A. (2006). *Independent Component Analysis, Principal Component Analysis and Rough Sets in Hybrid Mammogram Classification*. Paper presented at the IPCV.
- Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). *A Survey on Deep Transfer Learning*, Cham.
- Tessa, S., & Keith, F. (2018, August 23, 2018). The difference between an MRI and CT scan. Retrieved 26 Aug 2018, from <https://www.healthline.com/health/ct-scan-vs-mri>
- Thirumalai, C., & Manzoor, R. (2017, 20-22 April 2017). *Cost optimization using normal linear regression method for breast cancer Type I skin*. Paper presented at the 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA).
- Tsui, P.-H., Ho, M.-C., Tai, D.-I., Lin, Y.-H., Wang, C.-Y., & Ma, H.-Y. (2016). Acoustic structure quantification by using ultrasound Nakagami imaging for assessing liver fibrosis. *Scientific Reports*, 6, 33075.
- Tsui, P.-H., Yeh, C.-K., Chang, C.-C., & Liao, Y.-Y. (2008). Classification of breast masses by ultrasonic Nakagami imaging: a feasibility study. *Physics in Medicine & Biology*, 53(21), 6027.
- Ultrasound. (2018, March 09, 2018). General Ultrasound. from <https://www.radiologyinfo.org/en/info.cfm?pg=genus>
- Urbaniak, I., & Wolter, M. (2020). Quality assessment of compressed and resized medical images based on pattern recognition using a convolutional neural network. *Communications in Nonlinear Science and Numerical Simulation*, 105582. doi: <https://doi.org/10.1016/j.cnsns.2020.105582>
- Van Stralen, K. J., Stel, V. S., Reitsma, J. B., Dekker, F. W., Zoccali, C., & Jager, K. J. (2009). Diagnostic methods I: sensitivity, specificity, and other measures of accuracy. *Kidney international*, 75(12), 1257-1263.
- Vestjens, J. H. M. J., Pepels, M. J., de Boer, M., Borm, G. F., van Deurzen, C. H. M., van Diest, P. J., . . . Tjan-Heijnen, V. C. G. (2012). Relevant impact of central pathology review on nodal classification in individual breast cancer patients. *Annals of Oncology*, 23(10), 2561-2566. doi: 10.1093/annonc/mds072
- Vincent, P., Laroche, H., Lajoie, I., Bengio, Y., & Manzagol, P.-A. (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.*, 11, 3371-3408.
- Wan, T., Cao, J., Chen, J., & Qin, Z. (2017). Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features. *Neurocomputing*, 229, 34-44. doi: <https://doi.org/10.1016/j.neucom.2016.05.084>

- Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems, 105*, 87-95.
- WHO, W. H. O. (2018). World Cancer Report.
- Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67-82.
- Wu, J., Shi, J., Li, Y., Suo, J., & Zhang, Q. (2016, Aug. 29 2016-Sept. 2 2016). *Histopathological image classification using random binary hashing based PCANet and bilinear classifier*. Paper presented at the 2016 24th European Signal Processing Conference (EUSIPCO).
- Wu, K., Chen, X., & Ding, M. (2014). Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound. *Optik-International Journal for Light and Electron Optics, 125*(15), 4057-4063.
- Xu, J., Luo, X., Wang, G., Gilmore, H., & Madabhushi, A. (2016). A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing, 191*, 214-223. doi: <https://doi.org/10.1016/j.neucom.2016.01.034>
- Yao, H. D., Zhang, X. J., Zhou, X. B., & Liu, S. Y. (2019). Parallel Structure Deep Neural Network Using CNN and RNN with an Attention Mechanism for Breast Cancer Histology Image Classification. *Cancers, 11*(12). doi: 10.3390/cancers11121901
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). *How transferable are features in deep neural networks?* Paper presented at the Advances in neural information processing systems.
- Youk, J. H., Gweon, H. M., & Son, E. J. (2017). Shear-wave elastography in breast ultrasonography: the state of the art. *Ultrasonography, 36*(4), 300-309. doi: 10.14366/usg.17024
- Zhang, B. (2011). *Breast cancer diagnosis from biopsy images by serial fusion of Random Subspace ensembles*. Paper presented at the Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on.
- Zhang, Q., Xiao, Y., Dai, W., Suo, J. F., Wang, C. Z., Shi, J., & Zheng, H. R. (2016). Deep learning based classification of breast tumors with shear-wave elastography. *Ultrasonics, 72*, 150-157. doi: 10.1016/j.ultras.2016.08.004
- Zhang, X., Zhang, Y., Han, E. Y., Jacobs, N., Han, Q., Wang, X., & Liu, J. (2017, 13-16 Nov. 2017). *Whole mammogram image classification with convolutional neural networks*. Paper presented at the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).
- Zhang, Y., Tomuro, N., Furst, J., & Raicu, D. S. (2012). Building an ensemble system for diagnosing masses in mammograms. *International Journal of Computer Assisted Radiology and Surgery, 7*(2), 323-329.
- Zheng, Y., Jiang, Z., Xie, F., Zhang, H., Ma, Y., Shi, H., & Zhao, Y. (2017). Feature extraction from histopathological images based on nucleus-guided convolutional neural network for breast lesion classification. *Pattern Recognition, 71*, 14-25. doi: 10.1016/j.patcog.2017.05.010