



THE UNIVERSITY OF  
**BUCKINGHAM**

**Image Data Preparation For CNN-based Breast  
Ultrasound Lesion Diagnostic with Reduced  
Overfitting**

By

Tahir Hassan

A thesis Submitted for the Degree of Doctor of Philosophy in Computing  
to the School of Computing at the University of Buckingham.

September 2023

# Abstract

This thesis aims to contribute to efforts of leveraging deep learning (DL) techniques, specifically convolutional neural networks (CNNs), for improved diagnostics of breast lesions in ultrasound (US) images with reduced overfitting manifested by the inability to generalise models to unseen data. Our investigations focus on data preparation factors that influence the performance of CNN models for analysing Breast US (BUS) tumour images. These factors include CNN stipulated fixed input image size, adequate US image quality, and availability of a sufficiently large dataset of adequately labelled samples with good class-diversity for training. Current approaches to deal with these factors are focused on image resizing, relying on unstandardized manual quality assessment by radiology experts, and image augmentation. Most of these solutions rely heavily on knowledge of the natural image domain, which differs from US images. The sizes of US tumour region of interest (RoI) are influenced by the adopted cropping/segmentation procedure and vary significantly with a huge range on both sides of the strictly required input image size for most CNN models. Resizing the many tiny RoIs by several factors seriously impacts their quality. Existing augmentation schemes are designed to enlarge training sets and increase diversity, but the learnt feature patterns by pre-trained CNN models are more relevant to natural images.

We implemented the bicubic image resizing (BiCubic) method and a Compressed Sensing Super Resolution (CSSR) based image resizing known for superior quality resizing methods in terms of human perception. Our expert radiologist testified that CSSR-resized images are of better quality from the clinical point of view. We tested the performance of several pre-trained CNN models trained in fine-tuning mode on a database of BUS recorded and labelled in one clinical centre, whose RoI images were resized by both methods. All models achieved High-to-Excellent diagnostic accuracy, but little or no improvements were noted with the CSSR resizing scheme.

No RoI segmentation was adopted, but optimal cropping of RoI was developed from a set of radiologists' marked lesion border points. We introduced the Convex Hull (CH) lesion border RoI that efficiently minimizes the exclusion of lesion pixels and is easy to expand. We tested the performance of a few pre-trained CNN models and 2 Handcrafted (HC) schemes with various RoI cropping scenarios, including the tumour polygonal shape. We expanded CH at different rates, each with 2 padding schemes for the area between the surrounding rectangular box and the tumour polygon area: zero padding and tissue padding.

While tissue padding of several expanded CH rates had improved performance, zero padding of these schemes was marginally lower. Hence, the inclusion of some external tissue surrounding the lesion border shows promise for enhancing model performance. However, for both padding scenarios, the trained models have very low generalisation when tested on two unseen external datasets, confirming the problem of overfitting when the training dataset is not large and diverse enough.

Training the same CNN models with the larger Modelling dataset, compiled by including BUS images from 4 other clinical centres, didn't improve their validation performance but significantly improved their generalisation to the two unseen datasets. This improvement reflects that the expansion created a more diverse sample of the population resulting in reduced overfitting.

For the challenge of US image quality assessment (IQA), we uncovered the inadequacy of existing IQA metrics defined for natural images. We developed a simple Multi Characteristic Quality Feature Vector (MCIQ) that captures the spatial distribution of individual IQA metrics. MCIQ have shown good tumour class dependency and a high ability to distinguish different image modalities and datasets. An innovative version of MCIQ, extracted from image convolution with only 6 well-conditioned  $5 \times 5$  Hadamard filters, successfully aligned with our expert radiologist quality labelling of an extremely small set of US images.

Finally, to address the scarcity of BUS images beyond recording a larger training dataset, we investigated several existing conventional image augmentation schemes, including Singular Value Decomposition (SVD), besides our innovative Hadamard filters convolution. All these schemes improved the model's ability to generalize to the two unseen datasets but with varied levels of improvement. However, these schemes are not specific to US images, so it is difficult to determine which causes of overfitting these schemes help mitigate. For that, we developed the Tumour Margin Appending (TMA) strategy that combines several locally optimal cropping ratios to enlarge the training dataset aiming to alleviate the lack of generalization due to variation in RoI cropping practice. It successfully mitigated the lack of generalization to unseen datasets for this cause and removed the need to test with many unseen datasets.

# Acknowledgement

As I reach the culmination of this transformative journey, I am overwhelmed with gratitude for the exceptional individuals who have played pivotal roles in shaping my academic and personal growth. This thesis would not have been possible without the unwavering support, guidance, and encouragement of the following remarkable individuals and organizations:

**Prof. Sabah Jassim:** My first and primary supervisor – you are the foundation of my academic success. Your dedication, expertise, and constant belief in my abilities have been the bedrock of my achievements. You've been more than a supervisor; you've been my unwavering support, my mentor, and my guide. Your kindness and immense knowledge have shaped not only my research but also my character. You stood by me through every challenge, offering unwavering support. Your ability to uplift and inspire is truly remarkable. Thank you for being the guiding light on this journey.

**Dr Alaa:** My second supervisor – your insightful feedback and occasional guidance broadened my perspective on the research. Your contributions were invaluable in shaping the direction of this work.

**Prof. Hongbo Du:** Your guidance and support as the project coordinator were instrumental in steering this research in the right direction. Your wisdom enriched the quality of this work.

**TenD-Innovation for Medical Technologies:** Your generous sponsorship bridged the gap between academia and industry, emphasizing the practical significance of my research.

**My Family:** To my mother, siblings, and extended family – your boundless love, understanding, and unwavering support have been my constant source of strength.

**Friends and Peers:** Your camaraderie made the challenges more manageable and the successes more joyous.

**Support Staff:** Your dedicated efforts behind the scenes contributed to the smooth functioning of academic life.

In closing, this thesis stands as a collective effort, a tapestry woven together by the invaluable contributions of these exceptional individuals and organizations. A special and profound thank you to Prof. Sabah Jassim – your impact on my academic trajectory is immeasurable, and I am deeply grateful for the privilege of learning under your guidance.

With profound gratitude,

Tahir Hassan

# Abbreviations

AUC	Area Under the ROC Curve
BiCubic	Bicubic Interpolation
BI-RADS	Breast Imaging Reporting and Data System
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator
BUS	Breast Ultrasound
CAD	Computer-aided Diagnostic Systems
CH	Convex Hull
CNN	Convolution Neural Network
CS	compressed sensing
CSSR	Compressed Sensing Super Resolution
DL	Deep Learning
FCL	Fully Connected Layer
Flip&Rot	Flip and Rotation Augmentation
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Networks
GLCM	Gray-Level Co-occurrence Matrix
Grad-CAM	Gradient-weighted Class Activation Mapping
HC	Handcrafted
HOG	Histogram of Orientation Gradient
HR	High Resolution
IQA	Image Quality Assessment
kNN	k-Nearest Neighbours
LBP	Local Binary Pattern
LR	Low Resolution
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NNR	Nearest Neighbour Replacement
NRIQA	No-Reference Image Quality Assessments
NSS	Natural Scene Statistics

PCA	Principal Component Analysis
PSNR	Peak-Signal-to-Noise-Ratio
ReLU	Rectified Linear Unit
RIP	Restricted Isometry Property
RoI	Region of Interest
SISR	Single Image Super Resolution
SR	Super Resolution
SSIM	Structural Similarity Index Measure
SVD	Singular Value Decomposition
SVM	Support Vector Machine
T	Tissue padding
TDA	Topological Data Analysis
TMA	Tumour Margin Appending
TN	True Negative
TP	True Positive
TripleIQA+Entropy	BRISQUE, NIQE, PIQE and image Entropy
TumourT	RoI: smallest rectangular bounding box around the tumour polygon area, tissue padded.
TumourZ	RoI: smallest rectangular bounding box around the tumour polygon area, zero padded.
UIQI	Universal Quality Image Index
ULBP	Uniform Local Binary Pattern
US	Ultrasound
VAE	Variational Autoencoder
vs.	versus
Z	Zero Padding

# Table of Contents

Abstract.....	i
Acknowledgement.....	iii
Abbreviations .....	iv
Table of Contents .....	vi
List of Figures.....	xi
List of Tables.....	xiii
Declaration of Originality.....	xv
Chapter 1: Introduction.....	1
1.1 Background to Thesis Research Project.....	1
1.2 Thesis Aim and Objectives. ....	5
1.3 Contributions of the Thesis .....	7
1.4 Structure of the Thesis.....	10
Chapter 2: Review of Background Knowledge and Materials .....	12
2.1 Introduction to Breast Cancer and Ultrasound Imaging.....	12
2.1.1 Prevalence, Risk Factors, and Impact on Public Health.....	12
2.1.2 Medical Imaging in Breast Cancer Diagnosis .....	14
2.1.3 Advantages and Limitations of US Imaging for Breast Cancer Diagnosis .....	16
2.2 BUS Lesion Diagnosis: From Manual to Automated Systems .....	17
2.2.1 Deep Learning and Convolutional Neural Networks .....	20
2.2.2 Handcrafted Feature-based Image Analysis Schemes.....	23
2.3 Building Blocks of CNN Models .....	27
2.3.1 Convolutional Layers .....	27
2.3.2 Activation Functions.....	28
2.3.3 Pooling Layers.....	29
2.3.4 Fully Connected layers .....	30

2.3.5 Samples of state-of-the-art CNN Architectures.....	31
2.3.6 Pre-trained CNN model-Transfer Learning and Fine-Tuning Method .....	32
2.4 Research Materials .....	33
2.4.1 Ultrasound Image Datasets .....	33
2.4.1.1 Renmin Dataset .....	34
2.4.1.2 Modelling Dataset .....	35
2.4.1.3 Test1 Dataset.....	36
2.4.1.4 BUSI Dataset .....	36
2.4.2 Data Cleaning, Split and Training/Testing Protocol .....	37
2.4.3 Machine Learning Classification Performance Metrics .....	39
2.5 A Review of DL Approaches for BUS Lesion Classification.....	42
Chapter 3: Deep Learning for Ultrasound Images - Performance Influencing Factors.....	46
3.1 Deep Learning Requirements on RoI Size - Challenges and Trends.....	46
3.2 RoI Size variation and size normalisation.....	48
3.2.1 RoI size variation.....	49
3.2.2 RoI size normalization.....	54
3.2.3 Image Super Resolution.....	56
3.2.3.1 The Mathematical Model of Super Resolution.....	57
3.2.3.2 Compressed Sensing-based Single Image Super Resolution.....	58
3.2.3.3 Experimental Classification Results – BiCubic vs. SR .....	62
3.3 The Challenge of Ultrasound Image Quality Assessment .....	65
3.4 Lack of Training Samples - Solutions.....	72
3.5 Conclusion.....	74
Chapter 4: Lesion Shape Cropping from US Images .....	76
4.1 Introduction and Related Work .....	77
4.2 Cropping Models.....	79
4.2.1 Lesion Cropping by Interpolations .....	79



4.2.2 Lesion Cropping by Curve Fitting.....	81
4.2.3 Lesion Cropping by Convex Hull.....	82
4.2.4 Lesion Margin-appending Scenarios within the Convex Hull Shape.....	86
4.3 Performance of DL/HC Models for the Convex Hull Cropping Strategies.....	89
4.3.1 Renmin Dataset - Performance Testing.....	89
4.3.1.1 Performance of CNN models.....	89
4.3.1.2 Performance of Handcrafted Feature Schemes.....	92
4.3.2 Modelling Dataset - Performance Testing.....	94
4.4 CNN Cropped Lesion Models – Generalization Performance.....	95
4.4.1 Generalisation of the Renmin-trained CNN Models.....	95
4.4.2 Generalisation of the Modelling Dataset-trained CNN Models.....	96
4.5 Grad-CAM Visualization.....	99
4.6 Conclusion.....	104
Chapter 5: A No-reference Multi-Characteristics US Image Quality Descriptor.....	106
5.1 Computer-based Measure of Ultrasound Image Quality.....	107
5.1.1 Existing Work on Ultrasound Image Quality Assessment.....	107
5.1.2 Towards No-Reference Objective US IQA Descriptor.....	110
5.2 No-reference Multi-Characteristic Image Quality Vector (MCIQ).....	116
5.3 Explaining CNN Generalisation Results by MCIQ Feature Vectors.....	119
5.3.1 MCIQ - Generalisation Association: (Renmin, Modelling) vs. Test1.....	120
5.3.2 MCIQ - Generalization Association: (Renmin, Modelling) vs. BUSI.....	121
5.3.3 MCIQ Separability Between Images in Renmin and Modelling Datasets.....	122
5.4 Discriminating Power of MCIQ for Other Purposes.....	122
5.4.1 MCIQ - Tumour Class Association: Renmin (Benign vs. Malignant).....	123
5.4.2 MCIQ - Tissue Type Association: Breast vs. Liver/Bladder.....	124
5.4.3 MCIQ - Image Modality Association (Breast Lesion): US vs. Mammogram.....	125
5.4.4 MCIQ Domain Association: US Breast Tissue vs. Face Images.....	126

5.5 Limitations of MCIQ and Potential Remedies .....	126
5.5.1 MCIQ for the Pilot US dataset (Good vs. Bad).....	127
5.5.2 Modified MCIQs for the Pilot (Good and Bad) US Dataset .....	128
5.5.3 Performance of Modified MCIQ for Tumour Classification.....	131
5.6 Conclusion.....	132
Chapter 6: Image Augmentation for Deep Learning-based Breast Ultrasound Diagnosis .	134
6.1 Data Scarcity in Ultrasound Lesion Images - Introduction.....	135
6.2 A Review of Existing Image Augmentation Techniques.....	138
6.2.1 Conventional Augmentation Techniques .....	138
6.2.2 Synthetic Augmentation Techniques.....	139
6.2.3 Spectral-based Augmentation Techniques .....	141
6.3 Mathematically Inspired Augmentation Techniques for BUS.....	142
6.3.1 SVD-based Image Augmentation.....	142
6.3.2 Hadamard-based Kernel Image Augmentation .....	144
6.3.3 Augmentation Experimental Work.....	146
6.3.3.1 Performance of the Flip&Rot Augmentation Scheme .....	147
6.3.3.2 Performance of the SVD-based Image Augmentation Scheme.....	148
6.3.3.3 Performance of the Hadamard-based Image Augmentation Scheme .....	150
6.4 Mitigating BUS Data Scarcity by Margin Appending Schemes.....	152
6.4.1 Optimal Tumour Cropping in Uncontrolled Scenario – Revisited.....	153
6.4.2 BUS Augmentation Using Tumour Margin Appending Schemes .....	156
6.5 Conclusion.....	158
Chapter 7: Conclusion and Future Research Challenges .....	159
7.1 Main Conclusions.....	160
7.2 Future Research Challenges .....	163
7.2.1 Hybridisation of Multiple HC and Pre-trained CNN Models.....	163
7.2.1.1 Empowering Handcrafted texture features by CNN convolution filters .....	164

7.2.1.2 Incremental Fusion of CNN Models.....	166
7.2.2 Optimal Cropping – Revisited.....	167
7.2.3 Extending MCIQ for US-IQA Aligned with Expert Quality Labelling .....	169
Appendix A .....	170
Appendix B.....	172
References .....	175

# List of Figures

Figure 2.1 Breast Cancer Tumour [14].....	14
Figure 2.2 Medical Ultrasound Machine [18].....	15
Figure 2.3 An Ultrasound Scan Image of a Breast Tumour.....	18
Figure 2.4 The general pipeline of Machine Learning algorithms for image classification. ....	19
Figure 2.5 Deep Learning Family.....	21
Figure 2.6 Image kernel filtering [52]. ....	28
Figure 2.7 Various Activation Functions. ....	29
Figure 2.8 A fully Connected Layer [53].....	30
Figure 2.9 AlexNet architecture[60] .....	31
Figure 2.10 CNN model fine-tuning process for BUS classification. ....	32
Figure 2.11 Samples of breast ultrasound images from the Renmin dataset.....	34
Figure 2.12 Image data cleaning steps. ....	37
Figure 2.13 The data cleaning bounding box. ....	38
Figure 2.14 Area Under the Curve (AUC) [62].....	41
Figure 3.1 RoI (Yellow bounding boxes) size variation among 6 different BUS tumour images.....	47
Figure 3.2 Medical US Data Preparation Process.....	48
Figure 3.3 Steps of preparing and determining the digital representation of tumour RoI. ....	49
Figure 3.4 The distribution of tumour-area in pixel for the four BUS datasets.....	50
Figure 3.5 The distribution of tumour area in pixel for the Renmin dataset. ....	51
Figure 3.6 The distribution of tumour area in pixel for the Modelling dataset.....	51
Figure 3.7 The distribution of tumour area in pixel for the Test1 dataset.....	52
Figure 3.8 The distribution of tumour area in pixel for the BUSI dataset.....	52
Figure 3.9 Resizing two BUS RoIs from 30x45 (A) and 289x280 (B) into 227x227. ....	53
Figure 3.10 Resizing 3 different size RoI tumour images A, B, and C using 3 interpolation methods.....	55
Figure 3.11 The iterative procedure of Super Resolution problem. ....	57
Figure 3.12 Binary display of Sylvester type, Walsh-Paley, and Walsh Matrices of size 16x16. ....	59
Figure 3.13 The steps of our CS-based SISR procedure. ....	60
Figure 3.14 CS-based SISR algorithm [75].....	61
Figure 3.15 BiCubic vs. SR for resizing a LR BUS image.....	61
Figure 3.16 Distribution of tumour size for misclassified cases with BiCubic resizing. ....	64
Figure 3.17 Distribution of tumour sizes for the misclassified cases with SISR resizing.....	65
Figure 3.18 Display of spatial distribution of (Correlation, Luminance, and Contrast) for tumour RoIs and a face image.....	71
Figure 3.19 US image augmentation using rotation, mirror, and noise insertion. ....	74
Figure 4.1 Linear interpolation-based Polygonal lesion shape for (Regular vs. Irregular) lesion border. ....	80
Figure 4.2 Cubic-spline interpolation-based Polygonal lesion shape, in red colour, for (Regular vs. Irregular) lesion border.....	81

Figure 4.3 Fitted Gaussian curve and fitted ellipse, in red colour, for Regular vs. Irregular lesion border. ....	82
Figure 4.4 Steps of forming the CH polygon of a set of lesion boundary points. ....	84
Figure 4.5 Tumour shape approximation of 2 lesions: (A) tumour area polygon, (B) cubic spline border interpolation, (C) Gaussian curve fitting, (D) fitted ellipse, and (E) CH lesion border. ....	85
Figure 4.6 Illustrating tissue-padding tumour cropping scenarios for a breast tumour US image. ....	87
Figure 4.7 Illustrating 0-padding tumour cropping scenarios for a breast tumour US image. ....	88
Figure 4.8 Performance of 4 CNN models for different cropping scenarios – Tissue padding. ....	90
Figure 4.9 Performance of 4 CNN models for different cropping scenarios – Zero padding. ....	91
Figure 4.10 Performance of handcrafted models for cropping scenarios with Tissue padding. ....	92
Figure 4.11 Performance of handcrafted models for cropping scenarios with 0-padding. ....	93
Figure 4.12 Average heatmap scores of true classified and misclassified cases for both Benign and Malignant classes by Xception model of Test1 dataset: (TumourT vs. TumourZ). ....	102
Figure 4.13 Average heatmap scores of true classified and misclassified cases for both Benign and Malignant classes by Xception model of BUSI dataset: (TumourT vs. TumourZ). ....	103
Figure 5.1 Process of building and extracting MCIQ feature vector. ....	119
Figure 5.2 The 6 5x5 Hadamard-based filters. ....	129
Figure 6.1 SVD-based US image augmentation via reduced number of singular values. ....	144
Figure 6.2 Hadamard filter US image augmentation. ....	145
Figure 6.3 Performance of VGG19 trained and tested with cropping at the same ratio (Renmin). ....	153
Figure 6.4 Performance of VGG19 trained with one cropping ratio and tested on all the ratios. ....	155
Figure 6.5 Performance of VGG19 trained with one cropping ratio and tested on all the ratios. ....	155
Figure 6.6 Tumour Margin Appending steps to train/test CNN models. ....	156
Figure 6.7 Performance of VGG19 trained with TMA augmented datasets. ....	157
Figure 7.1 The proposed feature hybridisation methods. ....	163
Figure 7.2 Texture Feature Augmentation with Random Filters. ....	164
Figure 7.3 The four types of tumour posterior features ....	168
Figure 7.4 A breast tumour with its posterior region ....	168
Figure 7.5 Performance of VGG19 trained with one cropping ratio and tested on all the ratios. ....	172
Figure 7.6 Performance of VGG19 trained with one cropping ratio and tested on all the ratios. ....	173
Figure 7.7 Performance of VGG19 trained with one cropping ratio and tested on all the ratios. ....	173
Figure 7.8 Performance of VGG19 trained with one cropping ratio and tested on all the ratios. ....	174
Figure 7.9 Performance of VGG19 trained with TMA augmented datasets. ....	174

# List of Tables

Table 3.1 Performance of CNNs and HC features on Renmin dataset with BiCubic resizing.....	62
Table 3.2 Performance of CNNs and HC features on Renmin dataset with CS-SISR resizing. ....	63
Table 3.3 The computed PSNR, SSIM, UIQI and its three components for a selected 4 US images.....	69
Table 4.1 Average validation performance of CNN models for TumourT and TumourZ scenarios. ....	94
Table 4.2 Classification performance of Renmin-trained CNN models with (TumourT vs. TumourZ) for Test1 dataset.....	96
Table 4.3 Classification performance of Renmin-trained CNN models with (TumourT vs. TumourZ) for BUSI dataset.....	96
Table 4.4 Classification performance of the Modelling dataset-trained CNN models with (TumourT vs. TumourZ) for Test1 dataset.....	97
Table 4.5 Classification performance of the Modelling dataset-trained CNN models with (TumourT vs. TumourZ) for BUSI dataset.....	98
Table 5.1 The loss of correlation measure between pairs of Good/Bad images.....	111
Table 5.2 The luminance distortion measure between pairs of Good/Bad images.....	112
Table 5.3 The contrast distortion measure between pairs of Good/Bad images.....	113
Table 5.4 The computed UIQI measure between pairs of Good/Bad images. ....	114
Table 5.5 The three no-reference quality scores of the images (Good and Bad). ....	115
Table 5.6 Quality inspection of Renmin/Modelling vs. Test1 using MCIQ. ....	120
Table 5.7 Quality inspection of Renmin/Modelling vs. BUSI using MCIQ. ....	121
Table 5.8 Quality inspection of Renmin vs. Modelling using MCIQ.....	122
Table 5.9 MCIQ classification performance of Benign vs. Malignant (Renmin dataset). ....	123
Table 5.10 TripleIQA+Entropy classification of Benign vs. Malignant (Renmin dataset).....	123
Table 5.11 MCIQ classification performance of US Breast Tissue (Renmin) vs. US Liver/Bladder Tissue. ....	124
Table 5.12 MCIQ modality association for breast lesion: US (Renmin) vs. Mammogram (DDSM).....	125
Table 5.13 MCIQ classification performance of US (Renmin) vs. Face dataset.....	126
Table 5.14 The pilot dataset, Good vs. Bad using MCIQ. ....	127
Table 5.15 The pilot dataset, Good vs. Bad using TripleIQA+Entropy features. ....	127
Table 5.16 Good vs. Bad using MCIQ post 6(10) Gaussian filters convolution.....	128
Table 5.17 Bad vs. Good using TripleIQA+Entropy post 6 Gaussian filters convolution. ....	129
Table 5.18 Good vs. Bad using MCIQ post 6 Hadamard filter convolution. ....	130
Table 5.19 Comparisons of condition numbers of the Hadamard filters vs. Gaussian ones. ....	130
Table 5.20 Good vs. Bad using TripleIQA+Entropy post 6 Hadamard filter convolution.....	131
Table 5.21 Benign vs. Malignant (Renmin) using MCIQ post 6 Hadamard filters augmentation. ....	132
Table 5.22 Benign vs. Malignant (Renmin) using MCIQ post 6 Gaussian filters augmentation. ....	132
Table 6.1 Performance of the CNN models retrained on un-augmented Renmin dataset. ....	147
Table 6.2 Generalisation of pre-trained CNNs retrained with Flip&Rot -augmented Renmin dataset.	147
Table 6.3 Generalisation of pre-trained CNNs retrained with SVD-Augmented Renmin dataset. ....	149

<b>Table 6.4 Generalisation of pre-trained CNNs retrained with Hadamard-Augmented Renmin dataset.</b> .....	<b>150</b>
<b>Table 7.1 Classification of Several HC texture features post-convolution with fine-tuned AlexNet. ....</b>	<b>165</b>
<b>Table 7.2 Classification performance of the deep fused features with Cubic SVM. ....</b>	<b>167</b>
<b>Table 7.3 Generalisation of pre-trained CNNs retrained with Flip&amp;Rot-Augmented Modelling dataset.</b> .....	<b>170</b>
<b>Table 7.4 Generalisation of pre-trained CNNs retrained with SVD-Augmented Modelling dataset. ....</b>	<b>171</b>
<b>Table 7.5 Generalisation of pre-trained CNNs retrained with Hadamard-Augmented Modelling dataset.</b> .....	<b>171</b>

# Declaration of Originality

I, Tahir Hassan, hereby declare that this thesis titled "Image Data Preparation For CNN-based Breast Ultrasound Lesion Diagnostic with Reduced Overfitting" is entirely the product of my original research and writing unless otherwise acknowledged. I attest that all sources of information and ideas used in this thesis have been duly cited and referenced. No part of this thesis has been submitted for any other degree or qualification at any university. I take full responsibility for the accuracy of this work's data, figures, and tables. Any contributions made by individuals or organizations have been appropriately recognized in the Acknowledgements section. I affirm that ethical considerations were adhered to during the research, including obtaining necessary approvals and consent. By signing this declaration, I acknowledge the consequences of any breach of academic integrity and assure the authenticity and integrity of this thesis.

Date: 20/08/2023

Signature: *Tahir Hassan*



# Chapter 1: Introduction

Breast cancer continues to be a significant global health concern, impacting the lives of millions of women worldwide. Early detection and accurate diagnosis are crucial in improving patient outcomes and reducing mortality rates associated with this disease, but such tasks are increasingly dependent on advances in medicine in general and medical image analysis algorithms. While such crucial and critical clinical tasks are becoming more intensive and present enormous demands on the healthcare system's stretched resources and staffing, the recent giant advances in the fields of Machine Learning (ML) and Deep Learning (DL) for computer vision hold great promises of exciting opportunities in support of medical diagnostics. Leveraging the benefits of DL models, particularly Convolution Neural Networks (CNN), for medical image analysis requires addressing the challenges that emanate from the fact that high-performing CNN models are primarily designed for natural images that differ significantly from medical images such as ultrasound (US) images. This thesis is devoted to dealing with some of these challenges and exploring/developing novel image data preparation techniques and quality assessment tools for breast ultrasound (BUS) images to improve the efficiency and generalizability of CNN-based breast lesion classification. The main premise of this thesis is that addressing the limitations/shortcomings of current pre-processing methods for BUS helps improve the performance of the ML models for image analysis, particularly CNN-based classification schemes, and reduces variability in diagnosis.

## 1.1 Background to Thesis Research Project

BUS scan imaging has emerged as a valuable diagnostic and clinical examination tool that complements other radiological breast scan images, such as Mammograms and Magnetic Resonance Imaging (MRI) [1]. It offers several advantages, including its non-invasive nature, without using ionizing radiation, and the ability to differentiate solid from cystic lesions. It plays a vital role in various clinical scenarios, such as distinguishing benign and malignant lesions, guiding biopsy procedures, and monitoring treatment response. The real-time visualization nature of BUS scans enables clinicians/radiologists to assess lesions' morphological and vascular features, aiding in speedy and more reliable diagnosis [2].

Despite the benefits of BUS imaging, reliable and accurate classification of breast lesions remains challenging for clinicians/radiologists in their early stages of training. This is due to many factors, including US-specific properties such as often being of low contrast, subjected

to the presence of speckle noise, and other artefacts that make distinguishing benign from malignant lesions difficult for many early career radiologists. Traditional clinical diagnostic process heavily relies on the subjective interpretation of radiologists, whose training experience and expertise could be very wide, which can introduce inter- and intra-observer variability in outcome decisions and could have variation in subsequent treatment decisions [3], [4]. Inter-observer variability refers to differences in the interpretation of US images between different US radiologists/examiners, whereby two or more examiners may interpret the same US image differently, leading to varying diagnoses. It can occur due to differences in experience/expertise and different clinical practices/training that may lead to personal biases. Intra-observer variability, on the other hand, refers to differences in the interpretation of US images by the same examiner due to fatigue or lack of experience.

To address these concerns, overcome shortage of clinical expertise, and enhance the accuracy of breast (or other organs) lesion classification, there has always been an interest in using advances in computer technology, particularly automatic image processing/analysis algorithms. The rapid advancement in image technology and increased computational powers became a driving force for integrating digital technologies into clinical practices. These efforts involved close collaboration between research scientists and medical professionals, resulting in some reasonably successful computer-aided diagnostic (CAD) systems. Designing most of those early systems relied on extracting carefully chosen handcrafted (HC) texture feature vectors and using known mathematical classifiers (e.g., [5], [6]). These CAD systems benefited from the growing knowledge of distinguishing natural image features established over decades of research investigations in pattern recognition, including biometric systems.

The emergence of DL models of image analysis at the turn of the 21st century, and its remarkable success in dealing with very tough computer vision challenges, has led to a growing interest in applying DL techniques, particularly CNNs. This interest only started to be taken seriously as the success stories of CNN in computer vision accumulated in a legendary manner towards the end of the last decade. Coincidentally, the Covid Pandemic of the last few years exposed the vulnerability of most healthcare systems worldwide, showing stressful difficulties in coping with the unprecedented pressure on already strained resources. A huge volume of CNN and HC-ML schemes have been proposed and tested on ML-based Covid-19 detection from Chest CT scans and X-Rays, helping open the way, more than ever, to leverage the power of CNN and HC models for medical diagnostics.

The astounding success of CNN models in computer vision benefited from many factors besides the obvious boost from the tremendous advances in neural networks. These factors include the construction of huge datasets of natural images, such as ImageNet, with 100s of millions of high-quality images of different classes/objects that could be used to train all kinds of sophisticated CNN models [7]. A plethora of increasingly sophisticated CNN architectures have been proposed and demonstrated the ability to automatically learn large hidden discriminative image features, far beyond the capabilities of humans, with remarkable success in various image classification tasks.

These unrivalled capabilities of CNN models to learn hidden image data features through efficient processing of large image datasets raise high hopes for achieving similarly optimal success in medical image analysis and diagnostic tasks. Though achievable eventually, these benefits must be balanced against the fact that in the health service, digitisation of the radiological image data is lagging in constructing sufficiently large databases of radiological tumour scan images of standardised quality characteristics with standardised class labelling and annotation. This is more so for BUS images and even though such scans are routinely conducted in large numbers of clinical centres worldwide every day. Only recently, we started to see active research into the segmentation of BUS lesions. Our emphasis on the need for a large dataset of standardised quality BUS images with standardised class labelling, stems from the notable variation in US prob devices as well as their embedded electronic systems that could contribute to worsening Inter- and Intra- observer variations. Furthermore, there are no standardised global clinical/radiological practices. Consequently, even though a great deal of knowledge has been established on CNN models for image analysis and many state-of-the-art CNN architectures are available, leveraging these technologies for medical diagnosis from radiological tissue scanning is not straightforward. The absence of a sufficiently large dataset of BUS images is the major recognised obstacle in training CNN architectures from scratch for diagnostic purposes. CNN models trained with a few hundred BUS images recorded in a single clinical centre (or even multiple centres that follow similar clinical practices) are not expected to perform well in distinguishing benign from malignant tumours, even if scanned from the same centre(s). Moreover, such models are expected to suffer from the effect of overfitting (i.e., failure to generalise performance to unseen BUS samples recorded in other centres). They are doubtful to be robust against image data noise and adversarial attacks. The traditional approach to using CNN models for BUS (and medical image) analysis without training from scratch is to select a CNN model pre-trained on a large natural image dataset (e.g., subsets of ImageNet [7])

and complement it with an additional retraining process with the BUS in what is known as transfer learning mode (preferably the fine-tuning version). However, doing so does not avoid the overfitting problem. Failure to generalise the performance of CNN models (in transfer learning mode) to unseen data is a known challenge for any CNN models trained with a small image dataset other than BUS, but its impact on BUS (and diagnostics of medical images in general) is more serious than in other less critical applications.

At the time of setting up the TenD-Innovation research project, and this thesis was planned, significant efforts were dedicated to recording a reasonable-size BUS dataset from medical centres that were expected to follow similar clinical practices in relation to this task. To initiate TenD research projects, including the one for this thesis, a temporary alternative to automatic lesion segmentation was devised. Expert radiologists were asked to mark a sufficiently small set of lesion border points to enable a reasonable cropping of the region of interest (RoI) to be automated.

It is not enough to get reasonably large datasets of BUS lesion images, whose boundary is marked with a set of points, to retrain a pre-trained CNN architecture with guaranteed good performance and avoid overfitting problems, the list of requirements on the CNN training image datasets includes (1) the input images being of reasonable quality, (2) providing a reasonably good representation of the application population (sample diversity), and (3) being of stipulated fixed size. In this thesis, we will demonstrate that each of these requirements is a challenge for BUS that need to be addressed prior to retraining any CNN model in transfer learning mode. This list raises a number of different challenges, and non-compliance with is expected to have undesired performance outcomes. The image quality issue is challenging to adhere to for US images in the absence of a standard definition of image quality compatible with clinical expectations. The 2nd requirement is another challenge that reflects the lack of availability of BUS images, which restricts the knowledge about the actual BUS population. The fixed-size image requirement is a challenge as a result of the fact that tumour size varies depending on how early or late the mass was detected and scanned, and it is not a clear class-dependent factor. Resizing the cropped lesion RoI become necessary, but the effect of the adopted image resizing procedure on the image quality may adversely influence the performance of trained CNN models. These stipulated requirements necessitate the adoption of adequate image preparation techniques as a crucial step in developing BUS CNN-based classification models that have the desired high level of accuracy without suffering from overfitting. The decision to avoid automatic tumour segmentation in our TenD projects implied that the RoI size determination needs to use an

adequate lesion-cropping procedure. In relation to the RoI size requirement, image data preparation tasks consist of image pre-processing that encompasses a range of operations, including procedures for lesion cropping, noise reduction, RoI resizing and resolution enhancement. While ensuring adherence to the other two requirements, typically involves selecting appropriate image augmentation schemes and employing adequate image quality assessment (IQA) metrics. Image augmentation schemes are commonly applied through image processing procedures. By carefully selecting and optimizing these pre-processing steps, the quality and discriminative power of the input data for the CNN can be enhanced, leading to improved classification performance [8], [9].

Existing approaches to deal with the above BUS preparation tasks are based on using existing image preprocessing techniques developed to be suitable for natural images. Moreover, relying on natural IQA techniques may not be aligned with radiologist assessment of US image quality. In designing reasonably reliable CNN models for BUS lesion classification, these observations influence the related investigations conducted in this thesis. Accordingly, our research investigations focus on selecting pre-processing and data preparation procedures that are specific, as much as possible, to BUS images by examining their impact on the classification performance of the adopted CNN models. We seek to enhance breast lesion diagnosis's reliability, generalizability, and effectiveness. Although the findings of this study contribute to the advancement of CAD systems for breast cancer, it is expected to be of use for the analysis of US images for other types of tissues/organs.

## 1.2 Thesis Aim and Objectives.

This thesis addresses a set of research problems and questions that revolve around leveraging CNN models for improved diagnostics of breast lesions in US images while mitigating the challenges of overfitting as manifested by limited generalization. Here, we state the overall aim of the thesis and describe the list of research objectives that together ensure the fulfilment of the aimed purpose(s).

**The overall Aim:** This thesis aims to develop CNN-based models of BUS image analysis that achieve improved tumour diagnostics when trained on a dataset recorded in a single/multiple clinical setting while reducing the possibility of overfitting manifested by the lack of generalisation to unseen data.

The main objectives of the conducted research investigations to achieve the stated aim with reasonable success were not all set in a coherent manner at the start of my research, but evolved with time and were refined with the more knowledge I acquired about existing CNN

architectures, their requirements, and about the differences between natural image characteristics and US images. At the beginning of my project, the theory of CNN technology was maturing, and many CNN architectures designed for natural image analysis became available. However, there was limited knowledge and interest in adapting these architectures for computer vision tasks involving non-natural images. Accordingly, the thesis objectives were developed by first identifying the common CNN requirement factors that influence the performance of such models and may contribute to the overfitting problem. Our exploratory investigations identified the following (1) the fixed size input RoI images, (2) IQA, and (3) training dataset size and diversity. We found that complying with these factors is particularly challenging for US images. Accordingly, the research objectives and our investigations targeted the following list and their implications:

- 1.** Determine the extent of variation in lesion RoI sizes of BUS images and how this variation influences the performance of different CNN models trained with BUS datasets. The statistical distribution of RoI sizes needs to be studied regarding tumour class dependency. Furthermore, in the absence of exact RoI segmentation, two image processing issues contribute to the performance influence of RoI size: (1) RoI cropping practice and (2) the resizing procedure used to comply with the adopted CNN model.
- 2.** Investigate various RoI cropping schemes (followed by the most commonly used image resizing procedure) and determine if there is a near-optimal cropping scheme in terms of the performance of the adopted CNN models. Furthermore, we need to determine if the adopted resizing scheme impacts the outcome by repeating this investigation using a Compressed Sensing Super Resolution (CSSR) scheme known to improve image quality compared to traditional interpolation schemes.
- 3.** Develop automatic IQA metrics that align with radiologists' subjective assessment of BUS image quality. This objective need to benefit from a wealth of knowledge established for natural IQA metrics and determine their alignment with the radiologists' subjective assessment of BUS images.
- 4.** Investigate and develop schemes to enlarge and diversify BUS training datasets. This objective may require using a large BUS dataset of recorded samples in multi-clinical centres. It should go beyond using image augmentation procedures proposed for natural images by considering how radiologists make diagnostic decisions.

By achieving these objectives and research questions, this thesis is meant to promote the stated overall aim of this thesis for improving the efficiency, accuracy, and generalization

capabilities of CNN-based BUS lesion classification. The findings and insights gained from this research have potential implications for breast cancer diagnosis, patient outcomes, and clinical decision support systems development. The planned investigations will undoubtedly generate new challenges and opportunities for future research directions towards the advancements of DL-based BUS analysis.

### 1.3 Contributions of the Thesis

The thesis presents several contributions aimed at addressing research problems related to leveraging CNN models for improved diagnostics of breast lesions in US images while tackling challenges of overfitting and limited generalization. These contributions are outlined as follows:

- 1. Identification of Factors Influencing CNN Performance:** Our extensive initial investigations gained valuable insights into factors that impact CNN-based BUS lesion analysis. These insights have been established by conducting a statistical study of the extent of variations in RoI sizes, examining RoI cropping procedures, understanding the challenge of scarcity of well-annotated BUS images, and attempting to understand the challenge of assessing the quality of BUS images in terms of natural image distortion metrics. These findings illustrate that any attempt to leverage DL models into any image analysis tasks involving datasets of non-natural images must begin with a thorough analysis of the specific characteristics of these datasets that distinguish them from natural images in relation to the expected performance of adopted DL models.
- 2. Compressed Sensing Super Resolution (CSSR) Resizing:** We utilised our developed CSSR resizing algorithm that uses well-structured Hadamard-based dictionaries to resize BUS-RoI images and enhance the resolutions. Despite improving CNN models' performance marginally, compared to the BiCubic resizing procedure, CSSR improves the perceptual quality of resized RoIs, especially for low-resolution and degraded RoIs, confirmed by an experienced radiologist.
- 3. Convex Hull (CH) Lesion Border Approximation:** The CH lesion border RoI approximation is an efficient and effective alternative to automatic RoI segmentation, that minimizes the exclusion of lesion pixels and facilitates various cropping scenarios, contributing to improved model performance. The ease with which CH lesions can be expanded, even for highly irregular lesions, is instrumental in designing an effective US-specific augmentation technique (see below).

- 4. Optimal Tumour Cropping:** This research investigates the determination of optimal tumour cropping scenarios for improved CNN model performance. It identifies the TumourZ (tightly cropping tumour polygonal shape area with zero padding) cropping scenario as optimal, which yields superior performance compared to other cropping methods. The findings demonstrate the potential of TumourZ in enhancing the accuracy and generalizability of CNN-based breast lesion analysis, with improved decision quality supported by Grad-Cam visualization.
- 5. Multi-centre versus Single-centre Training Dataset:** TenD managed to source BUS datasets of nearly 4000 images from five medical centres located in Shanghai. Though helpful for achieving the last objective of Section 1.2, not all the images were suitable. We compiled a subset of these images by carefully selecting just under 1600 samples through a comprehensive cleaning process and visual examination. We call this BUS dataset the *Modelling* dataset. We experimentally show that training CNN models with this diverse modelling dataset, containing data from multiple clinical centres rather than a single clinical centre, significantly improves the generalization capabilities of the CNN models. This highlights the importance of multi-centre data in reducing the effect of scarcity and improving Data-Diversity for robust CNN-based breast lesion analysis with reduced overfitting. CNN models trained on a single clinical centre dataset have been shown to suffer from severe overfitting and do not generalize to unseen data from other sources.
- 6. Developing IQA for Medical US images:** We demonstrated that most existing IQA techniques for natural images are unreliable in assessing the quality of BUS images and do not align with radiology experts' assessments. Instead, we introduced a Multi Characteristic Image Quality (MCIQ) feature vector as a tool for US-IQA. The MCIQ captures the spatial distribution of different natural image quality metrics, reflects a good tumour class dependency, and can distinguish various image sources and datasets. MCIQ helps understand the lack of generalization of CNN models to unseen datasets. The limitations of MCIQ in alignment with expert radiologist BUS-IQA were linked to the scarcity of BUS images that were quality labelled by radiologists and the absence of established knowledge on distortions other than speckle noise. An advanced MCIQ version, developed by utilizing a small set of 6 well-conditioned 5x5 Hadamard filters for convolution-based US image augmentation, has been shown to improve MCIQ performance and its alignment with radiologists' IQA even for a small quality labelled BUS dataset.



- 7. Investigating Image Augmentation Strategies:** We explored several image augmentation strategies for US images to mitigate the scarcity of labelled data, evaluating their impact on model performance and generalizability. These strategies encompass the utilization of Hadamard filters convolution, Singular Value Decomposition (SVD) techniques, and other conventional augmentation schemes, such as flip and rotations. The experimental results show that such augmentation schemes effectively enhance the CNN models' generalizability to external datasets. Unfortunately, the success of these augmentation schemes does not provide information on the source of generalization failure that they treat. Ensuring success against the numerous sources of generalization failure requires testing the trained models on many external BUS datasets.
- 8. Tumour Margin Appending (TMA) Scheme:** We introduce the TMA augmentation-like approach to expand small training datasets by combining locally optimal cropping ratios (scenarios). This approach is shown to effectively mitigate the lack of generalization caused by variations in RoI cropping practices, improving the robustness of CNN models for breast lesion analysis in uncontrolled scenarios.

The above contributions collectively advance the efficiency, accuracy, and generalization capabilities of CNN-based breast lesion analysis in US images. The findings of this study have implications for improving breast cancer diagnosis, enhancing patient outcomes, and guiding the development of clinical decision support systems. Furthermore, the research identifies future research directions, emphasizing the significance of data preparation, IQA, augmentation techniques, and standardization of image acquisition protocols and datasets in the field of DL-based BUS analysis.

Finally, the various investigations to achieve the above contributions generated the following list of publications, and a few more manuscripts are planned for publication later. The list of published manuscripts to date are:

- I.** T. Hassan, H. Du, and S. Jassim, 'Enhancing Generalization of CNN Models for Breast Lesion Classification from Ultrasound Images', 2023, Presented at MIUA 2023 to appear in *Frontiers in Medical Technology*.
- II.** Tahir Hassan, Alaa Al Zoubi, Hongbo Du, and Sabah Jassim "Ultrasound image augmentation by tumor margin appending for robust deep learning based breast lesion classification", *Proc. SPIE 12100, Multimodal Image Exploitation and Learning 2022*, 1210008 (27 May 2022); <https://doi.org/10.1117/12.2618656>

- III. Tahir Hassan, Alaa AlZoubi, Hongbo Du, and Sabah Jassim "Towards optimal cropping: breast and liver tumor classification using ultrasound images", Proc. SPIE 11734, Multimodal Image Exploitation and Learning 2021, 117340G (12 April 2021); <https://doi.org/10.1117/12.2589038>

## 1.4 Structure of the Thesis

This thesis consists of 7 chapters, each focusing on specific aspects of CNN-based BUS lesion analysis and the associated challenges and objectives. The rest of the thesis is organised as follows:

**Chapter 2** provides essential background knowledge and context for understanding the research conducted in subsequent chapters. It explores the field of medical image analysis, emphasizing the role of CAD in supporting healthcare systems. The chapter also delves into the rapidly evolving field of DL and its application to medical image analysis, with a focus on CNN models. Additionally, reviewing existing works on CNN-based BUS lesion classification.

**Chapter 3** delves into the factors that influence the performance of DL models when applied to US tumour scan image analysis, emanating from CNN model requirements of training dataset samples. It identifies and describes the nature of the challenges posed by these factors that are specific to the BUS training dataset. For each of these factors, existing solutions from work on training CNN models on datasets of natural images are explored, and the chapter outlines the research strategy to tackle each of these factors.

**Chapter 4** focuses on the critical aspect of lesion cropping in BUS images. We review related work, explore tumour border approximation using interpolation methods from a set of border points marked by an expert radiologist and identify the CH of these points as a simple alternative. The chapter presents the experimental results and performance analysis of different cropping scenarios of both CNN models and HC feature schemes. We also explore the generalization performance of the developed models when tested on external datasets and utilize heatmaps visualization to understand the impact of the cropping scenario on decision quality.

**Chapter 5** addresses the challenge of IQA in BUS images. It introduces a novel approach called the Multi Characteristic Image Quality (MCIQ) feature vector, which captures the spatial distribution of quality metrics and serves as a quality descriptor. The chapter reviews existing IQA schemes and presents experimental findings using MCIQ to explain disparities

observed in DL generalization results. It also explores the use of MCIQ for other quality-related applications in BUS image analysis.

**Chapter 6** addresses the scarcity of labelled US images through image augmentation techniques. It reviews existing augmentation approaches and introduces novel schemes tailored explicitly for BUS images. The chapter investigates their impact on pre-trained CNN models' performance and generalization capabilities. Additionally, it introduces the TMA strategy, a cropping-based augmentation approach, to expand scarce training datasets and enhance generalization capabilities.

**Chapter 7** summarizes the main findings and contributions of the research. It highlights the significance of the research problems addressed, outlines the novel techniques and methodologies proposed, and emphasizes the improvements achieved in CNN-based BUS lesion analysis. The chapter also discusses the practical implications of the research and provides insights into future research directions.

# Chapter 2: Review of Background Knowledge and Materials

In recent years, research activities have intensified toward integrating ML algorithms, ranging from HC texture analysis schemes to emerging DL technologies, into the clinical practices of breast lesion classification using US tumour scan images. Leveraging advances in these technologies is ultimately expected to aid in accurate and efficient early diagnosis for improved patient care.

This chapter provides essential background knowledge and context for understanding the key concepts and techniques necessary, relevant to such integration, to comprehend the research conducted in the subsequent chapters of this thesis. It begins by exploring the field of medical image analysis, highlighting the significance of CAD in supporting healthcare systems. The chapter then delves into the rapidly evolving DL as a powerful approach that has revolutionized various computer vision domains, including medical image analysis. The utilization of pre-trained models and fine-tuning strategies is discussed, emphasizing their effectiveness in adapting CNNs for specific tasks for which training CNN models from scratch is infeasible, such as BUS lesion classification. The chapter ends with a review of existing works on DL for BUS lesion classification.

## **2.1 Introduction to Breast Cancer and Ultrasound Imaging**

In this section, we describe the various aspects of the BUS diagnostic efforts in terms of general health issues and the involvement of medical image technologies.

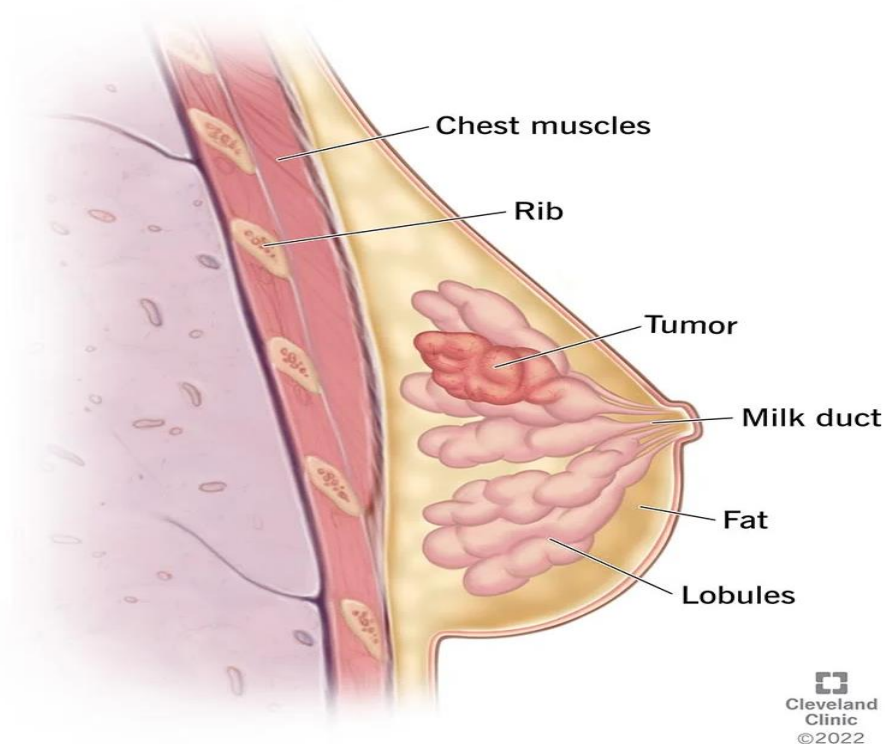
### **2.1.1 Prevalence, Risk Factors, and Impact on Public Health**

Breast cancer is a prevalent and critical global health issue, affecting millions of individuals worldwide, particularly women. It is the most commonly diagnosed cancer among women and a leading cause of cancer-related deaths. To effectively fight or manage this killer disease, we need to comprehend its prevalence, risk factors, and impact on public healthcare systems [10]. The incidence rates of breast cancer vary across different regions globally. Still, this variation has to take into account the significant variation in the level of available medical care in different parts of the world. In 2020 alone, the World Health Organization reported approximately 2.3 million new cases of breast cancer and 685,000 deaths attributed to the disease. These staggering numbers underscore the urgent need for effective strategies to combat breast cancer and alleviate its burden on public health.

Various risk factors contribute to breast cancer development, encompassing genetic and environmental factors. Age is a significant risk factor, as the incidence of breast cancer increases with advancing age. Family history of breast cancer, particularly among first-degree relatives, is another factor indicating the relevance of genetic predisposition. Certain gene mutations, such as BRCA1 and BRCA2, are associated with an elevated risk of developing breast cancer. Hormonal factors, including early onset of menstruation, late menopause, and hormone replacement therapy, also contribute to the risk. Lifestyle choices that may increase the likelihood of developing breast cancer include an inactive lifestyle, unhealthy diet, obesity, excessive alcohol consumption, and long-term use of oral contraceptives. Other potential environmental risk factors include exposure to ionizing radiation and patients having had previous benign breast conditions [10].

Breast cancer profoundly affects the overall quality of life for those diagnosed with the disease. Its impact on public health goes far beyond individual patients to their families and the community healthcare systems. It places significant financial and staffing burdens on healthcare utilization relating to diagnostic tests, treatments, follow-up care, and demands on social care. Physical symptoms, treatment side effects, and psychosocial challenges can significantly impact patients' well-being and daily functioning. This emotional toll extends to their families and caregivers, who provide essential support throughout the journey [11], [12]. Figure 2.1 presents a breast cancer tumour.

Increased public awareness about breast self-examinations, regular clinical breast examinations, mammography and US screenings help identify breast cancer at earlier stages when treatment options are more effective, yielding improved survival rates. By addressing all issues relating to the risk factors and the implications of breast cancer, researchers, healthcare professionals, policymakers, and individuals can collaborate to reduce their consequential burdens and enhance the overall well-being of society as well as those affected by this disease [13]. Next, we shall explain the role of medical imaging in breast cancer early detection and diagnosis.



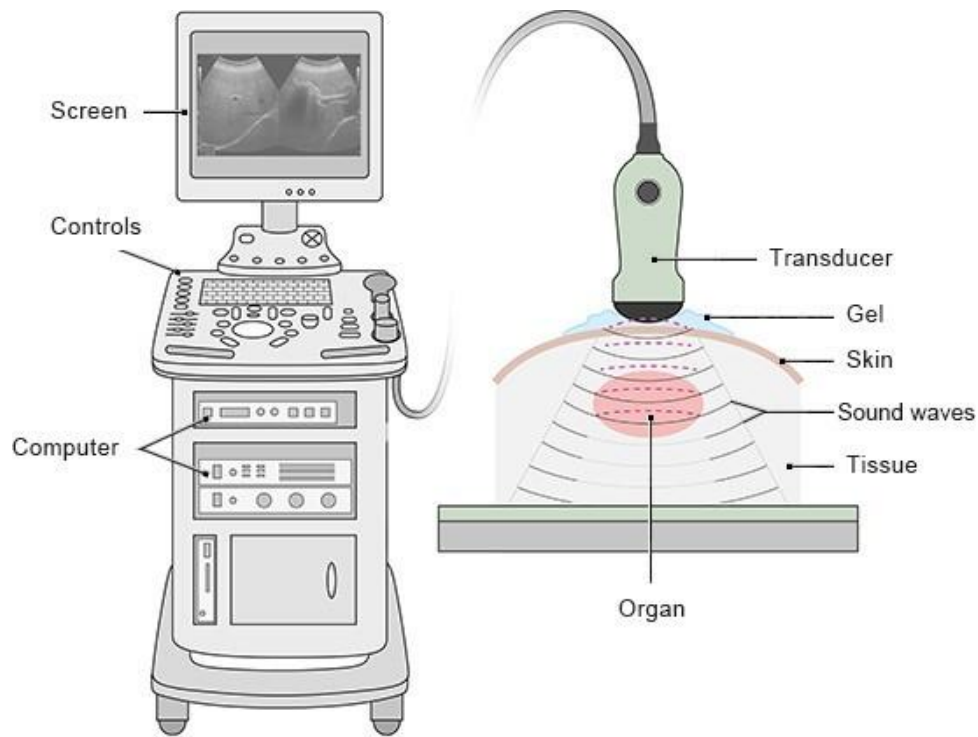
**Figure 2.1 Breast Cancer Tumour [14].**

### **2.1.2 Medical Imaging in Breast Cancer Diagnosis**

Medical imaging plays a crucial role in detecting, diagnosing, and monitoring breast cancer and other diseases. It aids healthcare clinicians as a vital visual tool to examine internal tissue/organ structure, monitor functioning, and detect abnormalities. It provides valuable information needed for making accurate assessments and informed decisions regarding patient disease diagnosis/management. Among the various imaging modalities available, the US has emerged as a vital tool in breast cancer care due to its unique capabilities and advantages. US imaging, also known as sonography, utilizes high-frequency sound waves to create detailed images of breast tissues. Figure 2.2 displays a typical medical US machine. US scanning is a non-invasive, safe, and widely accessible imaging technique that does not involve exposure to ionizing radiation. It is particularly suitable for scanning sensitive populations, such as pregnant women and young individuals [15].

US scanning of the breast for cancer detection is often used to complement other imaging modalities, such as mammography. Mammography is a standard screening tool, but it may have limitations, especially in dense breast tissue or for individuals with a high risk of developing breast cancer. The US can provide additional information, particularly in

distinguishing between solid masses and fluid-filled cysts. It can help evaluate breast abnormalities detected during physical examinations or mammographic screening, aiding in *early detection* [16], [17].



**Figure 2.2 Medical Ultrasound Machine [18].**

In addition to detection and diagnosis, US is also valuable in the monitoring and surveillance of breast cancer patients. It allows for assessing treatment response, including the evaluation of tumour size, vascularity, and changes in the surrounding tissues. Tracking treatment progress can benefit from safe serial US examinations to determine the need for additional interventions or adjustments in disease management [19]. Furthermore, serial US scans aid in evaluating cancer recurrence/metastasis and help identify new suspicious lesions or changes in the previously affected areas, i.e., guide further investigations and appropriate treatment strategies [20].

Enhanced technological advancements and evolving techniques can improve diagnostics efforts. For example, Doppler US scans enable the assessment of blood flow within breast lesions, providing valuable information about their vascularity and potential disease aggressiveness. Other advanced US technologies, such as elastography, can evaluate tissue stiffness and aid in characterizing breast lesions [21].

In conclusion, US imaging plays a pivotal role in breast cancer detection, diagnosis, and monitoring. Its continued advancements hold promise for further enhancing the role of US in breast cancer management.

### **2.1.3 Advantages and Limitations of US Imaging for Breast Cancer Diagnosis**

US imaging is a valuable tool in breast cancer assessment, offering several advantages that make it a relevant and widely used modality in clinical practice. However, it also has certain limitations that researchers and healthcare professionals need to consider. Understanding these advantages and limitations is essential for optimizing the use of US in breast cancer research and patient care.

One of the key advantages of US imaging is its non-invasive and painless nature. It is a safe imaging modality as it does not involve exposure to ionizing radiation. Additionally, US allows for dynamic assessment and visualization of the examined tissue or organ in real-time, providing immediate feedback [22].

Another advantage of the US is its excellent soft tissue contrast, which enables the differentiation between normal and abnormal breast tissues. It helps identify and characterise various breast lesions, such as cysts, solid masses, or benign conditions, aiding in the diagnosis and treatment planning [23].

US is widely accessible and more cost-effective compared to other imaging modalities, such as MRI and CT scans. This accessibility makes it a valuable tool, particularly in resource-limited settings where availability and cost considerations are important factors [22].

However, US imaging also has certain limitations that need to be taken into account. One major limitation is its operator dependency, as the quality of the images and interpretation can vary based on the skills and experience of the operator. Standardization and ongoing training are essential to ensure consistent results to minimize inter-observer variability [24].

Compared to other medical imaging modalities such as MRI or CT scans, US images are often perceived to have lower image quality. The concept of US image quality is somewhat vague and confusing to non-expert observers. This perception arises from differences in textural and structural content, which are essential in clinical settings. US images can be affected by various factors that contribute to reduced image clarity and diagnostic confidence. Speckle noise, caused by the interaction of US waves with tissue structures, can obscure fine details and reduce the visibility of subtle features within the image. Additionally, US images may exhibit different levels of contrast in different regions, making distinguishing between different tissue types or detecting subtle abnormalities difficult.



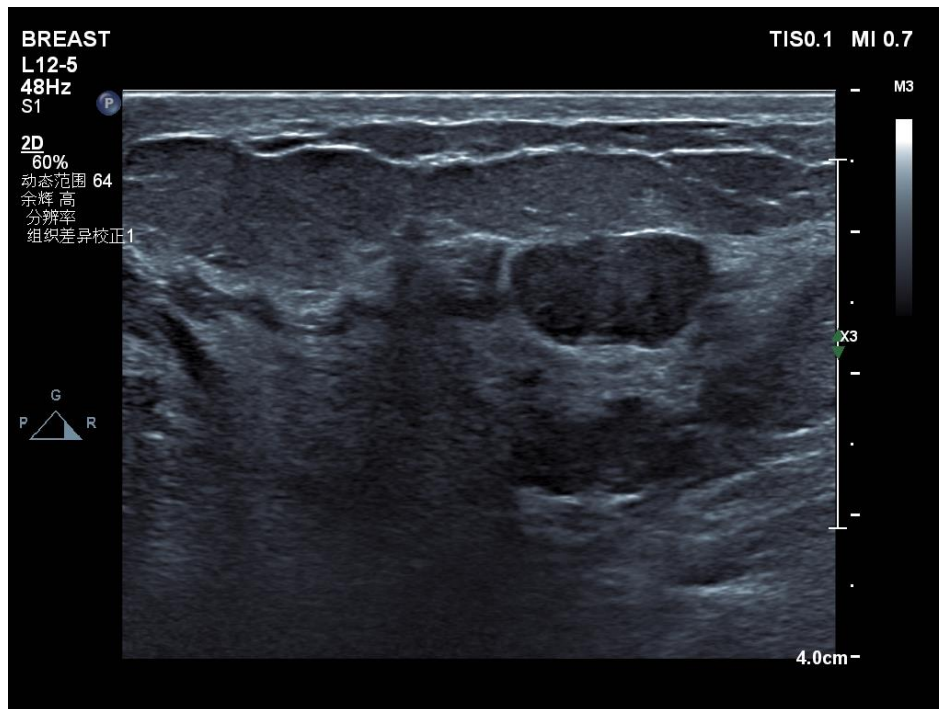
These limitations can hinder the accurate interpretation and analysis of US images, potentially impacting the performance of DL models that rely on high-quality input data [25].

Another limitation of US imaging is its limited tissue penetration and field of view, which could result in a partial assessment of the breast tissue and potentially miss lesions in certain cases [19]. While the US helps differentiate between cystic and solid lesions, it may have limitations in characterizing the nature of solid masses. Additional imaging modalities, such as mammography or MRI, may be required for a more comprehensive evaluation and accurate diagnosis [19]. Moreover, US imaging cannot efficiently detect or evaluate microcalcifications formed by small calcium deposits, often considered a sign of early breast cancer. Mammography is the gold standard for detecting microcalcifications [26].

It is worth noting that advancements in US technology and image processing techniques have been exploited to mitigate some of these limitations, such as speckle noise reduction algorithms and contrast enhancement methods. However, the inherent challenges associated with US image quality should be considered when developing and evaluating automated BUS lesion classification methods. In this thesis, our research focuses on performance-influencing factors of CNN-based BUS lesion classification, including investigating US-related concepts of image quality and data scarcity.

## **2.2 BUS Lesion Diagnosis: From Manual to Automated Systems**

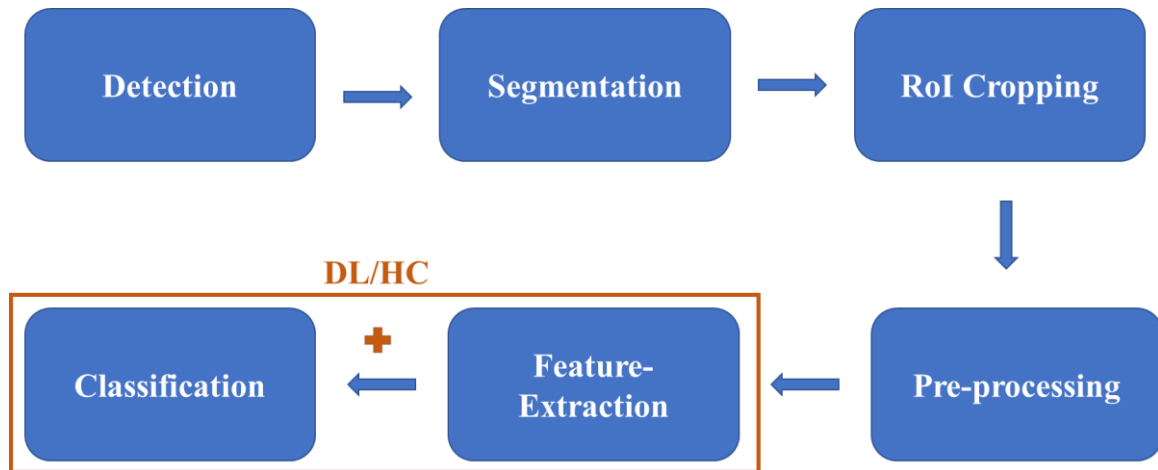
Traditional breast cancer diagnosis heavily relies on the expertise of radiologists for identifying, labelling and classifying breast lesions. Radiologists manually examine and analyse US scan images of breast tissues; they look for specific features relevant to signs of malignancy or otherwise (e.g., shape, size, echogenicity (the ability to reflect sound waves), margin characteristics, and the presence of microcalcifications). Their assessment is recorded in BI-RADS (Breast Imaging Reporting and Data System) reports that categorise each identified lesion sign indicating the level of suspicion for malignancy [27]. This process can be time-consuming, leading to diagnosis and treatment initiation delays. Figure 2.3 shows a breast tumour scan image, from which it is not feasible for non-specialised observers to assess the patient's status credibly accurately.



**Figure 2.3 An Ultrasound Scan Image of a Breast Tumour.**

However, this manual process is subjective, time-consuming, and prone to inter- and intra-observer variability [24]. In recent years, there has been a growing interest in leveraging ML technologies, including HC feature analysis and CNN, to automate and improve BUS lesion diagnosis. The development and deployment of HC feature medical image analysis schemes predate even the emergence of the first CNN model just before the turn of the century for the analysis of natural images. For both types of schemes, the model consists of two components: a *feature extraction* component and a *classification* component. Figure 2.4 presents the general pipeline of both paradigms of ML algorithms for image classification. Both types of diagnostic schemes are based on supervised learning, i.e., their training dataset samples must be class labelled, and their algorithms learn how to map input samples to their labels. HC texture feature analysis relies on extracting certain texture features that are engineered by researchers with good working knowledge of image content and processing, then training a classifier to learn a mapping between the extracted texture features' representation and the labels of the training samples. CNN models are designed to learn large numbers of hidden complex patterns at different scales refined through an elaborate training procedure using several convolution layers designed to meet the investigated objectives and possibly more accurate diagnostic decisions [28]. The success of both approaches relies on having enough training samples reflecting the diversity of the task population distribution.

Still, CNN models require significantly larger training sets, especially for training from scratch. The problem of scarcity of US tumour scan images that are adequately labelled according to an adapted standard is a serious challenge to the work of this thesis in relation to deploying CNN models for BUS diagnostic tasks.



**Figure 2.4** The general pipeline of Machine Learning algorithms for image classification.

CNN models have demonstrated remarkable capabilities in various image analysis tasks, and their potential to enhance the detection/diagnosis of various tumour types by analysing medical image scans of their corresponding tissues/organs is gaining remarkable attention [22]. Naturally, the widely reported astonishing success levels of CNN models in computer vision tasks raise the question about the wisdom/necessity of using HC features schemes anymore. Lin et al. in [29] consider this question with respect to the problem of identifying the adequacy of contrast-enhanced liver MRIs. They demonstrate that some HC schemes perform consistently, in terms of AUC curves, for this application through a range of training sample sizes, while CNN was unable to converge with sizes < 100 samples. The adequacy of these images relates to the fact that images acquired soon after intravenous contrast injection may have insufficient contrast and impaired differentiation between normal liver tissue and focal lesions. Lack of contrast and other clinically relevant image quality characteristics is another challenge to CNN models for US images we investigate in this thesis. In the case of HC schemes, these issues are often dealt with by some image quality enhancement procedures (e.g., sharpening, denoising, histogram equalisation, etc.) or by quality adaptive schemes, see [30]. Another issue is that the two types of ML schemes handle the issue of input image size in different ways. While CNN schemes require all (training and testing) images to be of a fixed size, HC schemes deal with input image size variations by

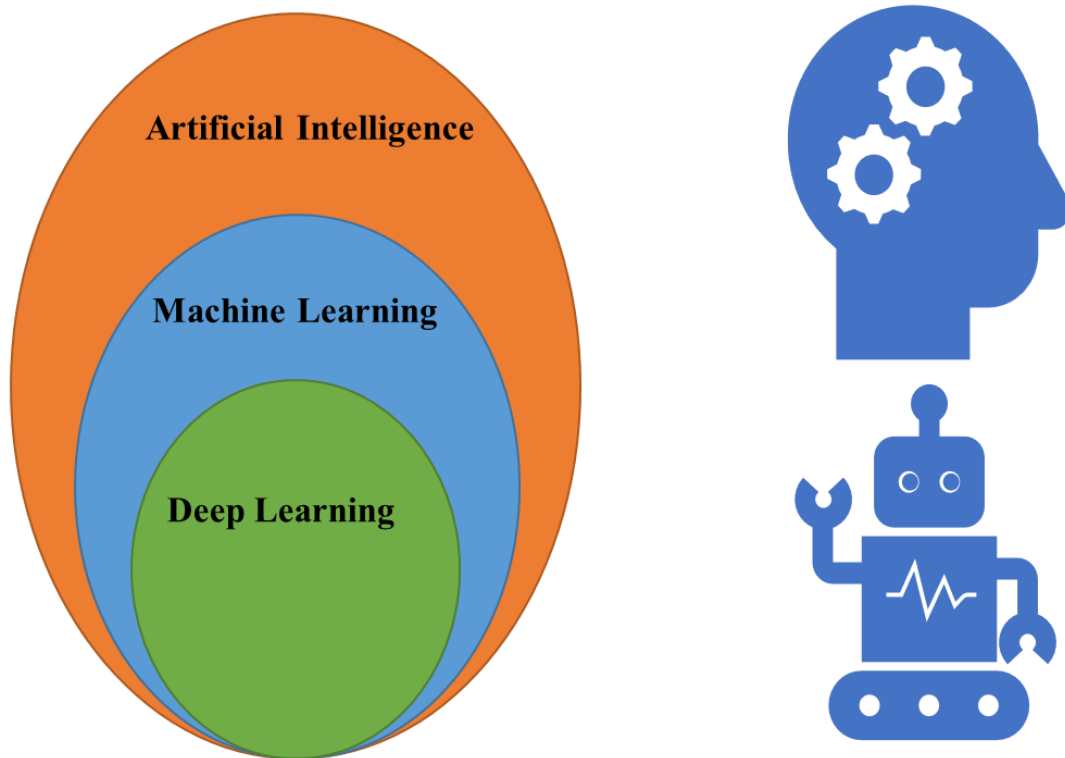
normalising the extracted features' representation. For US tumour images, RoI image size varies within a large range, and resizing seriously impacts the quality of the input images to CNN models.

Instead of looking at these two paradigms as competitors, many researchers have been inspired by the variety of benefits from the two paradigms of ML image analysis, proposed and investigated generic systems that combine HC features with the features learnt by the layers of CNN models, (see, e.g., [31], [32]). In cases where the available training data is limited, or the target classification task is specialized, HC features can provide valuable insights and improve classification performance, e.g., cancer sign research in BUS lesion classification [33], [34]. Moreover, combining HC features with CNNs in a hybrid approach can leverage the strengths of both methodologies, leading to enhanced classification accuracy and interpretability [35]. During the work of this thesis, we have also conducted some pilot work along these lines.

In summary, we recognise that integrating CNNs into BUS lesion clinical diagnostics has great potential for improvement and reliability. However, the transition from radiologist diagnosis to CNN automated diagnosis is some way away due to many challenges and considerations. The CNN models must be extensively validated on independent datasets to ensure their robustness and generalizability [36], [37]. Moreover, most CNN models work as black boxes with little or no interpretation of their decisions, thereby limiting trust and acceptance by clinicians and the public [38]. Next, we shall describe these 2 image-analysis paradigms, highlighting the main structures facilitating their learning aspects.

### **2.2.1 Deep Learning and Convolutional Neural Networks**

DL is a subfield of ML in artificial intelligence that focuses on multi-layered artificial neural networks. These networks have gained substantial attention and popularity since the turn of this century, owing to their remarkable capacity to automatically learn and extract high-level features/representations from complex data. The invention of DL algorithms was, and continues to be, inspired by established knowledge of the neural organization in the visual cortex of living organisms. DL-ML paradigm is often described as an attempt to mimic the functioning of the human brain by enabling computers to process and analyse vast amounts of data with exceptional accuracy [39]. Figure 2.5 below depicts this popular perception of the DL family of algorithms.



**Figure 2.5 Deep Learning Family.**

The DL paradigm has significantly impacted various computational fields through its applications since the first scheme LeNet was published in 1998 [40]. Since then, a race has ensued to create new CNN models for various challenging image processing/analysis tasks like facial recognition, autonomous driving, and video analysis. In this exponentially growing field of image analysis, DL has achieved human-level performance in object detection, image segmentation, and classification. The transformative potentials of the DL paradigm across diverse fields of computer vision application are widely recognised for improved results and for driving future advancements in technology and understanding [39]. Most of these CNN models have been developed for applications related to natural image modalities. Still, in recent years, these successes have been the driving force for leveraging their use in the domain of medical image analysis for automating clinical tasks of interpreting the content of non-natural scan images obtained by X-ray, CT, MRI, and US machines. By extracting/learning intricate features, CNN models are expected to facilitate disease detection, diagnosis, and treatment planning, promising results in anatomical structure segmentation, abnormality identification, and clinical outcome prediction [28]. Most CNN models have the same overall architecture but differ in certain choices of parameters, and

below is a very concise description that applies to the commonly used state-of-the-art CNN models.

The feature extraction component of the various CNN architectures consists of multiple interconnected convolutional layers, each serving a distinct purpose in extracting and learning application-relevant features. These layers consist of grid-like structures of receptive fields, enabling the gradual formation of spatial hierarchies of features, from simple local patterns to more global structural patterns. Each convolution layer employs a large number of convolution filters/kernels, organised in a number of channels, to be used for convolving images and extracting spatial patterns of local data. Post-convolving input images, these layers apply other operations, including activation functions, such as the Rectified Linear Unit (ReLU), designed to introduce non-linearity characteristics that the linear convolutions overlook. The activation functions help create sparse representations of the convolved feature maps and some convolution layers, then apply local pooling as a down-sampling dimension reduction operation before passing it onto the next layer or to the fully connected layer (FCL), also known as fully/dense connected neural network, classification component. The operations of these convolution layers involve other parameters besides the sets of convolution filters, including adding a bias parameter post-convolution besides the stride and padding parameters that determine how the convolution kernels scan the input image data. Furthermore, filter sizes at different layers can vary from one layer to another, and the CNN models differ in their choices of filter sizes as well as the other parameters.

To facilitate high-level learning and decision-making, the FCLs are employed to integrate the extracted/learnt feature maps and capture their semantic relationships in order to map the input image onto the label. These layers establish dense connections between all neurons in consecutive layers, enabling the network to model complex feature combinations and generate final class label predictions. By leveraging weight matrices and activation functions, the FCL transforms the extracted/learnt feature maps into outputs corresponding to specific classes or regression values. Training CNNs on a sufficiently large training dataset of image samples is an elaborate procedure whereby the set of feature maps obtained from convolution layers components are passed through the FCLs, with their activation functions, to be subjected iteratively to the backpropagation procedure, that adjusts the CNN parameters (including the convolution filter weights and FCL weights) according to the differences between the predicted and the target decisions. The parameter adjustment uses optimization algorithms, such as stochastic gradient descent variants, to minimize a defined loss function [28], [41].

### 2.2.2 Handcrafted Feature-based Image Analysis Schemes

The HC paradigm of automated ML algorithms grows from decades of computer vision research into image processing, pattern recognition, biometrics recognition and digital forensics. Much of these efforts rely on identifying local image data patterns (i.e., features) associated with notable visual/frequency changes, referred to as changes in texture, that are amenable to mathematical formulation. In all these fields, researchers recognised the importance of texture features and their statistics. The computer vision literature is awash with many different image features that have been manually engineered (i.e., handcrafted). The design of HC features often involves finding the right trade-off between accuracy and computational efficiency.

Furthermore, when engineering image HC features, for many image analyses, appropriate considerations were given to features that are invariant in certain image operations, such as scaling and rotation. HC features for face recognition should be less affected by occlusions and variations in pose and illumination. The well-known Scale Invariant Feature Transform is used to extract local HC features in digital images by first locating some key landmarks and endowing them with quantitative descriptors that are invariant against object rotation and scale variations [31]. However, this comes with a high computational cost.

HC texture features in images are not confined to the spatial domain, but many HC features have been extracted from image frequency domains (e.g., Fourier, Gabor and Wavelet) and or in many transform domains, e.g., [6], [42]. For detection and classifying Ovarian tumours from US ovary scan images, several HC texture-based schemes were developed and tested individually and fused groups, with various high success rates. These HC features included the Fast Fourier-based Geometric Features, Local Binary Pattern (LBP), Histogram of Orientation Gradient (HOG), Gabor filter, Fractal dimension, seven moments features, and Gray-Level Co-occurrence Matrix (GLCM) (e.g., see [5], [6], [9], [22], [42]). The emergence of the topological data analysis (TDA) paradigm of image analysis [43] provides a new texture HC representation. TDA's persistent homology tool that encapsulates the spatial distribution of certain HC feature landmarks (e.g., LBP) has shown a high level of success in breast tumour diagnoses from mammogram scan images and US liver tumour scans [44], [45]. Recently, the TDA-based HC feature of the Euler Characteristic Curve was used in [29], with considerable success in identifying the adequacy of contrast-enhanced liver MR images.

While DL methods often outperform traditional HC feature-based approaches in medical image classification tasks, during the work of this thesis, we used several of the above-mentioned HC schemes. However, to be brief, we will only describe the 2 most common ones: LBP and HOG. We were investigating their effectiveness when dealing with limited-size dataset US images as a more straightforward and more interpretable ML model. It is worth noting that the performance of HOG and LBP-based approaches may not match the state-of-the-art achieved by DL models in BUS lesion classification. Next, we shall briefly explain them.

### Histogram of Oriented Gradient (HOG)

HOG-HC image descriptor is computed from the gradient map  $G$  of a grayscale image  $f$ , by representing the partial derivative pair at each pixel position  $(i, j)$ , in the polar coordinates:

$$\left(\frac{\partial f(i, j)}{\partial x}, \frac{\partial f(i, j)}{\partial y}\right) = (g_x(i, j), g_y(i, j)) \quad 2.1$$

$$g_x(i, j) = f(i + 1, j) - f(i - 1, j) \quad 2.2$$

$$g_y(i, j) = f(i, j + 1) - f(i, j - 1) \quad 2.3$$

$$g(i, j) = \sqrt{g_x^2(i, j) + g_y^2(i, j)} \quad 2.4$$

$$\theta(i, j) = \arctan\left(\frac{g_y(i, j)}{g_x(i, j)}\right) \quad 2.5$$

Here,  $g_x(i, j)$  and  $g_y(i, j)$  represent the gradient in horizontal and vertical directions, respectively. The polar representation of the Gradient map  $G$  is given at each point by the pair  $(g, \theta)$  of magnitude and orientation (see, e.g., [46], [47]).

To quantify the orientation, the  $[0, 180[$  interval is typically divided into 9 equal bins. The image is then divided into non-overlapping rectangular blocks of equal size. For each block  $c$ , a weighted histogram  $hog_c$  is computed by summing the magnitudes of the gradient pixels in block  $c$  that fall within each orientation bin. The computation of the weighted histogram  $hog_c$  can be defined as:

$$hog_c(k) = \sum_{(i, j)} \{g(i, j) : 20(k - 1) < \theta(i, j) \leq 20k\} \quad 2.6$$

After computing the histogram for each block, it is common to normalize the histogram by dividing each bin by the sum of the contents of the 9 bins. This normalization ensures that



the descriptor is robust to changes in illumination. Finally, the HOG texture feature vector of an image  $f$  is formed by concatenating the histograms  $hogc$  for all blocks. In many applications, the image is subdivided into  $3 \times 3$  equal rectangular blocks, resulting in a HOG feature vector of size 81. However, a larger number of blocks, such as  $4 \times 4$  or  $5 \times 5$ , can also be used.

### Local Binary Patterns (LBP)

LBP is another widely used texture descriptor with applications in various image analysis tasks. Initially proposed by Ojala et al. as a method to characterize local texture [48]. LBP defines an image transform that operates by associating an 8-bit binary code with each pixel in an image based on its order relation to its neighbours. Traditionally, the central pixel in a  $3 \times 3$  image patch is compared to its neighbours in a clockwise manner, starting from the top-left corner. For each neighbour pixel, if it has a value greater than or equal to the central pixel, it is assigned the binary 1 and 0 otherwise, resulting in circular 8-bit binary code. The LBP-transformed image changes each image pixel value to a new decimal value obtained from its 8-bit LBP code using the formula:

$$LBP(x_c, y_c) = \sum_{n=0}^{n=7} s(i_n - i_c) 2^n \quad 2.7$$

where  $i_c$  and  $i_n$  are grayscale values of the central and its 8-neighbour pixel, scanned in clockwise order from the top left corner, and the function  $s(x)$  is defined as:

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad 2.8$$

To differentiate between groups of LBP codes in terms of their geometric interpretation, the 8-bit byte is considered as a circular string in a clockwise order starting from the top-left corner. The number of transitions between 0 and 1 in the circular string is counted. **Uniform LBP codes (ULBP)**, the adopted LBP texture feature in this work, have at most two transitions and indicate the presence of corners, end of lines, and other important features. Interestingly, in face images, ULBP codes form nearly 90% of all LBP codes [49], [50]. There are a total of 58 Uniform LBP codes, with 56 of them having two transitions and two codes having no transitions: "00000000" and "11111111".

Different groups of LBP pixels in an image are used to extract feature vectors. Three common feature vectors are LBP56, representing the 56-bin histogram of all the 56 ULBP codes with two transitions. ULBP represents the 58-bin histogram of all the Uniform LBP codes. LBP59 includes all 58 Uniform LBP codes, with the last bin holding the count of all

other LBP codes. We note that other applications use a 256-bin histogram, one bin for each of 256 different LBP codes, and this is often referred to as the standard LBP feature vector. However, in this work, ULBP is the only adopted LBP feature vector for BUS lesion classification.

For more detail on other relevant texture feature descriptors in this area, including GLCM, Gabor and Fractal, we guide the readers to [5], [6], [42].

Next, we shall briefly describe the two most commonly used classifiers for HC features that we also used in this work.

### **Support Vector Machine (SVM)**

SVM is a supervised ML algorithm for classification and regression analysis [51]. It aims to find an optimal hyperplane in a high-dimensional feature space that separates different classes or groups of data points. SVM operates by maximizing the margin, the distance between the hyperplane and the nearest data points from each category. This margin maximization allows SVM to handle complex decision boundaries and classify new, unseen data points effectively. SVM uses a kernel function to transform the input data into a higher-dimensional space where linear separation is possible. This transformation enables SVM to handle non-linear relationships between features. SVM is known for its ability to handle high-dimensional data and its robustness against overfitting.

### **k-Nearest Neighbours (kNN)**

kNN is a non-parametric ML algorithm used for classification and regression tasks [51]. It operates based on the principle that data points with similar features tend to belong to the same class or have similar output values. In kNN, the "k" represents the number of nearest neighbours to consider. To classify a new data point, kNN searches for the "k" closest training data points in the feature space. The class or value of the new data point is determined by the majority vote or averaging of the classes or values of its k nearest neighbours. kNN is simple yet effective, as it relies on local information and does not make strong assumptions about the underlying data distribution. However, its performance can be sensitive to the choice of the number of neighbours (k) and the distance metric used to measure similarity between data points.

## 2.3 Building Blocks of CNN Models

This section describes the standard building blocks of existing CNN models used for image analysis and explains their structures and functions within the overall CNN architectures. The aim is to highlight the end-to-end process of their learning algorithms. We shall also describe and list samples of the CNN models that we used in this thesis.

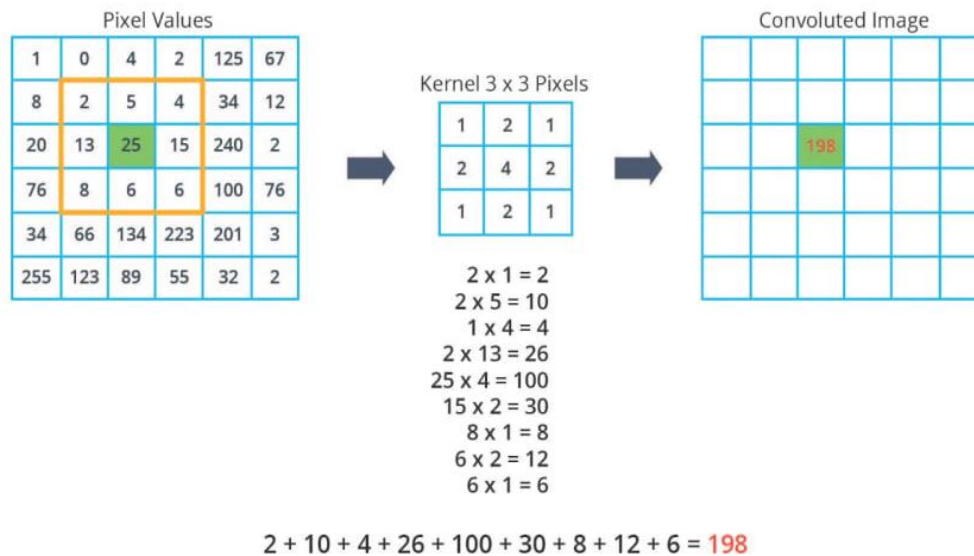
### 2.3.1 Convolutional Layers

Convolutional layers are fundamental components of Convolutional Neural Networks (CNNs) used in computer vision tasks. These layers consist of filters that perform convolutions on input images, extracting local features. Convolutions involve sliding the filters over the input image and computing dot products between filter weights and image pixels. By learning these filter weights through training, convolutional layers can detect important visual patterns, such as edges, textures, or shapes, at different spatial locations. The output of a convolutional layer is a feature map highlighting the presence of these learned features. The use of convolutional layers enables CNNs to effectively capture spatial hierarchies and perform tasks such as object detection, segmentation and image classification, contributing to advancements in computer vision research and applications [41].

Mathematically, convolving an input image patch and a same-size filter is defined as follows:

$$F(i, j) = \sum_m \sum_n I(i + m, j + n) * K(m, n) \quad 2.9$$

Where  $F(i, j)$  represents the value of the convolutional output at position  $(i, j)$ ,  $I(i + m, j + n)$ , denotes the pixel value of the input image at position  $(i + m, j + n)$ , and  $K(m, n)$  corresponds to the filter coefficient at position  $(m, n)$ . Figure 2.6 visually presents the image convolution operation.



**Figure 2.6 Image kernel filtering [52].**

In addition to the convolution operation, convolutional layers incorporate other parameters, such as stride and padding, to control the spatial dimensions of the output feature maps. The stride parameter determines the step size at which the filter is moved across the input image. A larger stride reduces the spatial resolution of the output feature maps, leading to spatial downsampling. Padding is used to preserve spatial dimensions by adding additional border pixels to the input image, preventing information loss at the edges.

### 2.3.2 Activation Functions

Activation functions play a critical role in artificial neural networks by introducing non-linearities to the network's output. These functions are applied to the output of each neuron, determining whether the neuron should be activated or not. The activation function introduces non-linear transformations to the input data, enabling the network to learn complex patterns and make non-linear decisions. Common activation functions include the sigmoid function, which maps input values to a bounded range between 0 and 1, and the ReLU function, which outputs the input value if it is positive and zero otherwise. Other activation functions, such as the hyperbolic tangent (tanh) and SoftMax functions, are also used in specific contexts. The choice of activation function impacts the network's ability to model complex relationships and affects its learning dynamics. By applying activation functions, neural networks can effectively model complex functions and achieve better performance in tasks such as detection and classification [41]. Figure 2.7 presents the graph of the above-mentioned activation functions.

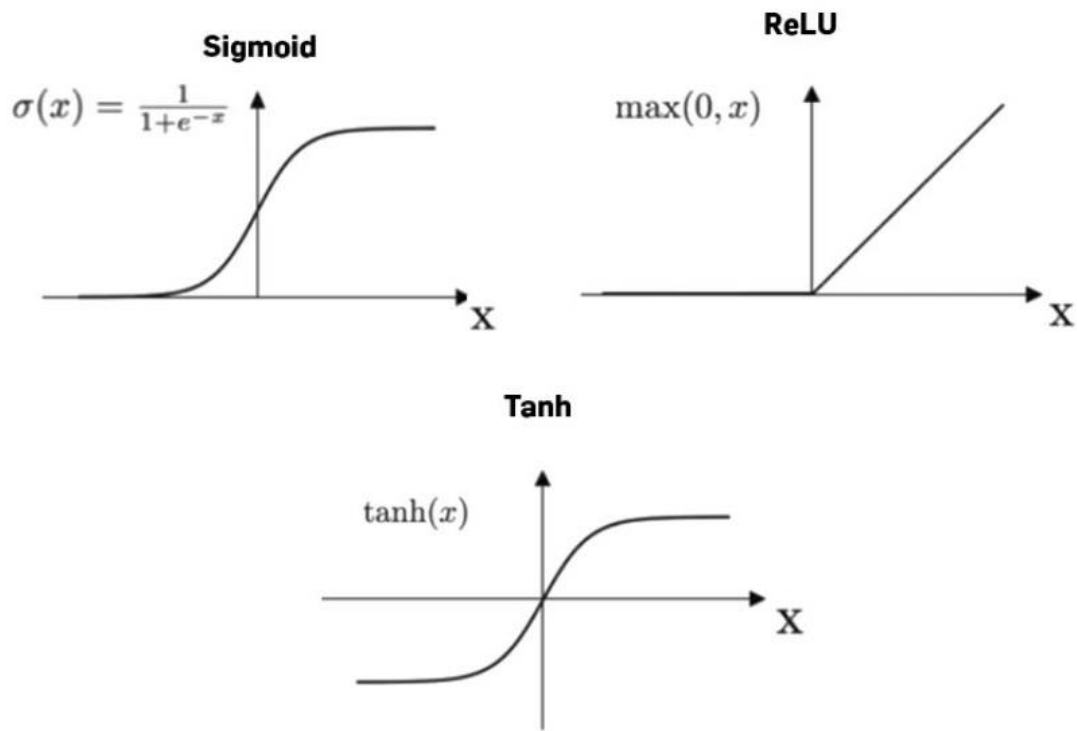


Figure 2.7 Various Activation Functions.

### 2.3.3 Pooling Layers

Pooling layers in convolutional neural networks (CNNs) are utilized to reduce the spatial dimensions of the input feature maps, thereby extracting essential information while reducing computational complexity. These layers divide the input into non-overlapping or overlapping regions and perform an aggregation operation within each region, typically maximum or average pooling. By downsampling the feature maps, pooling layers help in achieving translation invariance, robustness to small spatial variations, and increased computational efficiency. Max pooling selects the maximum value within each pooling region, effectively capturing the most salient features. Average pooling computes the average value, providing a more smoothed input representation. Pooling layers contribute to spatial hierarchies in feature maps, enabling the network to learn increasingly abstract and invariant representations of the input data. While pooling layers discard some spatial information, they enhance the network's ability to detect and recognize important features, improving the network's performance in tasks such as object detection and image classification [41].

### 2.3.4 Fully Connected layers

FCLs, also known as dense/fully connected neural networks, are a fundamental component of DL architectures, including CNN models. Every neuron in a layer is connected to all neurons in the preceding layer, allowing information propagation throughout the network. These layers are crucial in capturing complex relationships and high-level abstractions in the input data. Each neuron in a FCL performs a weighted sum of the inputs, followed by an activation function, generating an output contributing to the subsequent layer's computation. Their dense connectivity pattern enables the CNN model to learn intricate patterns and nonlinear relationships in the data effectively. They output a map of the extracted features to the final decision space, making them well-suited for tasks like classification and regression. However, FCLs introduce a large number of parameters, which can lead to overfitting and increased computational complexity. Regularization techniques, such as dropout, are often applied to mitigate these challenges and improve the generalization capability of the network [41]. Mathematically, the output of a FCL can be computed as follows:

$$y = W \cdot x + b \quad 2.10$$

Where  $y$  represents the output vector,  $W$  is the weight matrix,  $x$  is the input vector, and  $b$  is the bias vector. The weight matrix  $W$  contains the learnable parameters of the layer, and the bias vector  $b$  allows for shifting the activation function. The number of neurons in a FCL is determined by the network architecture design. Each neuron in the layer takes as input the outputs of all neurons in the previous layer and applies a non-linear activation function to enable learning complex representations and decision boundaries. Figure 2.8 depicts a typical structure of a CNN-FCL.

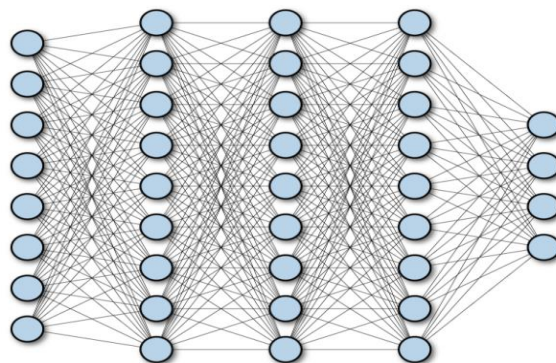


Figure 2.8 A fully Connected Layer [53].

### 2.3.5 Samples of state-of-the-art CNN Architectures

In this section, we provide a brief overview of several prominent CNN architectures employed in this work, highlighting their contributions to the field of DL-based image analysis. AlexNet (2012), [54], stands as an architecture that showcases deep CNNs' potential in image classification with its innovative ReLU activation function and dropout regularization, see Figure 2.9. VGG16 and VGG19 (2014), [55], gained recognition for their simplicity and effectiveness, employing 3x3 convolutional filters extensively. ResNet18 and ResNet50 (2016), [56], introduced skip connections to tackle the challenge of training deeper networks to enable the training of highly complex models. Xception model (2016), [57], extended the Inception architecture by utilizing depth-wise separable convolutions and optimizing parameter usage to facilitate impressive performance with limited training data. InceptionV3 (2016), [58], further enhanced performance by incorporating inception modules with parallel operations to capture diverse information scales efficiently. DenseNet201 (2017), [59], adopted innovative connectivity patterns with dense connections, promoting feature reuse and gradient flow for improved model performance. Utilizing these diverse architectures developed for the analysis of natural images in the research of this thesis allows for a comprehensive exploration and evaluation of their effectiveness in the context of BUS lesion diagnosis.

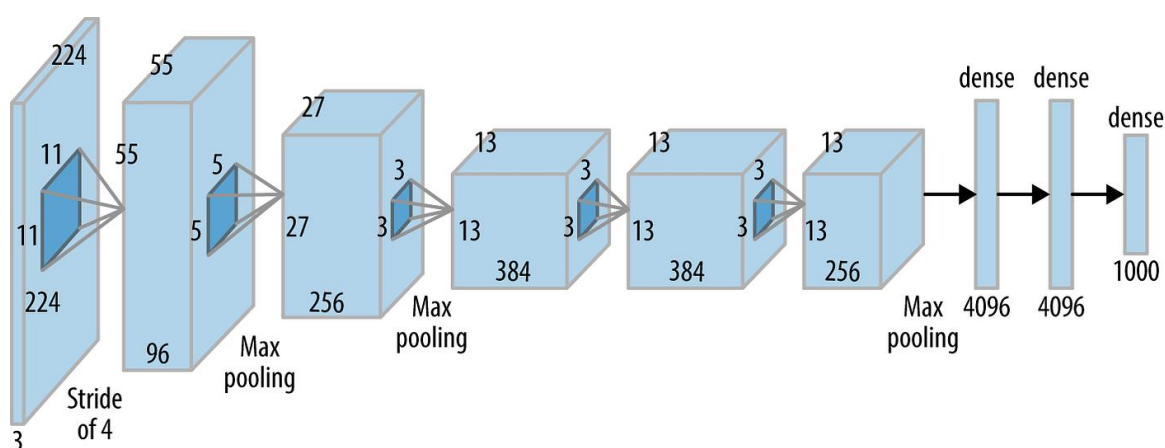


Figure 2.9 AlexNet architecture[60] .

### 2.3.6 Pre-trained CNN model-Transfer Learning and Fine-Tuning Method

CNN models have a well-established property of being useful for image analysis tasks, even with limited data. They can be transferred to different image tasks, even ones different from their original purpose. These models were initially designed for analysing millions of natural images with numerous classes (e.g., ImageNet [7]). Transfer learning mode and its fine-tuning version of DL models are designed to retrain a chosen pre-trained CNN model on a new image dataset of a different modality (and a number of classes) than the original model training dataset. The aim is to adapt an existing pre-trained model to enable analysis of the new dataset. Fine-tuning is a training procedure that is more suitable for CNNs to address the challenge of having a small and non-representative dataset of the domain population for a specific task, such as BUS lesion classification [28]. A simple version of transfer learning often freezes the learnable parameters of the convolution layers during the retraining, and only the FCL weights are updated during the retraining. On the other hand, the fine-tuning version retrains a pre-trained CNN model without freezing any layer/parameter [28]. The fine-tuning training procedure can be summarized as follows; see Figure 2.10.

1. Pre-trained Model Initialization: Start with a pre-trained CNN model.
2. Model Modification: Modify the pre-trained model by replacing the final fully connected layer with a new layer that matches the number of classes in the target dataset (Binary Classification: Benign Vs. Malignant).
3. Training Dataset Preparation: Prepare the training dataset specific to the task at hand, in our case, labelled Benign and Malignant BUS images.
4. Fine-tuning Process: the parameters of all layers in the model are updated during the additional training performed on the BUS images. In this work, the additional training is conducted using the training datasets with 10 epochs of training. This means that both the convolutional layers and the newly replaced layers for the classification task are fine-tuned.
5. Performance Evaluation: Evaluate the performance of the fine-tuned models on external testing datasets to assess their generalisation capability. Compute performance metrics such as accuracy, sensitivity, specificity, F1-score, and AUC to measure the effectiveness of the models.

**Figure 2.10 CNN model fine-tuning process for BUS classification.**



By fine-tuning the pre-trained CNN models on the BUS dataset, we can benefit from the initially learned features while allowing the model to adapt and specialize for the BUS lesion classification task. This approach takes advantage of the pre-trained model's knowledge while refining the weights to better fit the target dataset. Therefore, fine-tuning is the adopted CNN training protocol throughout our work to address the limitations of the small and non-representative BUS dataset and leverage the knowledge encoded in pre-trained models for effective BUS lesion classification.

## **2.4 Research Materials**

In this section, we shall present the relevant research materials that I used in my research work for this thesis. It includes preparing to form BUS datasets for training/testing CNN models and dataset description. We shall also cover the commonly used classification performance measures and evaluation protocols for the CNN models.

### **2.4.1 Ultrasound Image Datasets**

In the field of DL research, the choice of datasets plays a crucial role in evaluating the performance and generalizability of the proposed models. For the experimental work conducted in this thesis, defining the datasets used for training, validation, and testing is essential. This section provides a sufficiently comprehensive overview of the datasets employed, including their characteristics, data collection methods, and relevant pre-processing steps.

To facilitate standardised advanced analysis, we used a special approach to determine the RoI tumour area in all images without relying on manual/automatic segmentation. Experienced radiologists manually marked a sufficient number of lesion boundary points for each instance in all the following US breast tumour datasets. These annotated boundary points serve a crucial purpose, allowing for precise detection of the lesion's shape and location. This information is particularly valuable when considering alternative methods for automatic lesion detection and segmentation [22]. Figure 2.11 presents two BUS images with corresponding lesion boundary points and class labels from the Renmin dataset, which is defined in the next section. Next, we define the BUS datasets used in our investigations.

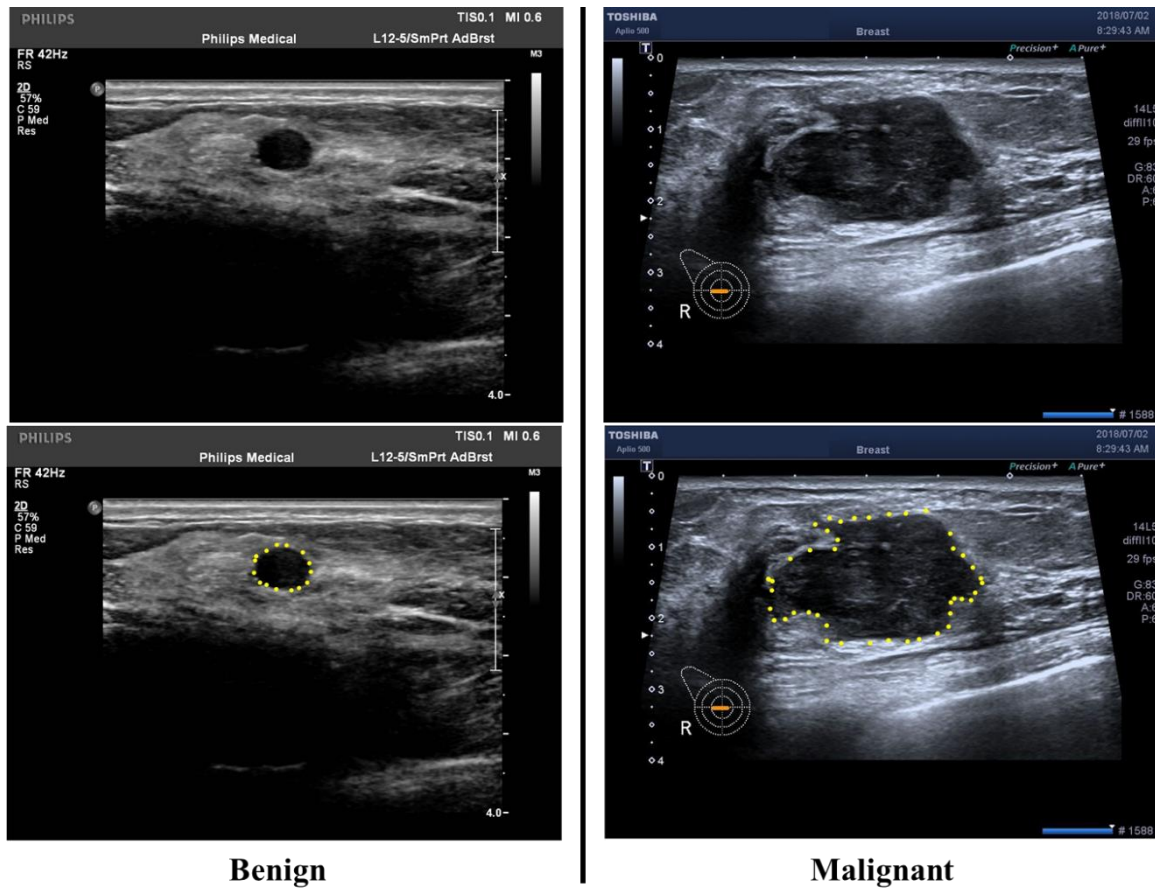


Figure 2.11 Samples of breast ultrasound images from the Renmin dataset.

### 2.4.1.1 Renmin Dataset

The Renmin dataset is of significant value as the foundational dataset utilized in the development of most of the work conducted in this research. It represents a relatively small, yet crucial collection of BUS images meticulously gathered at Pudong New District Renmin Hospital in Shanghai, China. This dataset serves as one of the training datasets.

Comprising a balanced distribution of 524 images, the Renmin dataset consists of 262 benign and 262 malignant cases. Multiple US machines were employed during the data collection process, reflecting the real-world clinical setting. However, it is noteworthy that all images were reviewed and labelled by a single experienced radiologist, ensuring consistency and reducing potential inter-observer variability. The radiologist carefully examined each image, accurately differentiating between benign and malignant lesions based on their expert knowledge and expertise, and backed by biopsy tests were performed to confirm the nature of each lesion.

The main disadvantage of using this database for CNN retraining is its relatively small size and limited sample diversity, which may result in potentially biased predictions. However, it is still valuable for gaining a solid foundation for the development and refinement of algorithms, allowing us to explore the capabilities of DL in accurately distinguishing between benign and malignant breast lesions after training on a relatively small dataset from a single medical centre.

#### **2.4.1.2 Modelling Dataset**

TenD sourced BUS datasets of nearly 4000 images from five Shanghai medical centres. I compiled a subset of these images through a comprehensive cleaning process and visual examination. The resulting TenD *Modelling* dataset comprises 1598 BUS images, 999 benign and 599 malignant cases. It was meticulously curated by collecting data from the five different medical centres in Shanghai, including the renowned Renmin Hospital. Including BUS images from multiple medical centres ensures greater diversity in clinical practices, imaging protocols, and patient populations. It provides a more representative sample of the variations encountered in breast tumour US imaging. This TenD Modelling dataset represents a significant advancement compared to the Renmin dataset, offering a larger and more diverse collection of BUS images. As the second training dataset in this research, it is a valuable resource for training CNN models designed explicitly for BUS lesion classification.

Various US machines were employed throughout the data acquisition process, reflecting the real-world clinical setting where different equipment is used across medical centres. Additionally, to capture a comprehensive range of expertise and perspectives, images were examined and labelled by five experienced radiologists from each medical centre. This multi-radiologist approach helps mitigate potential bias and inter-observer variability, enhancing the dataset's reliability and the model's generalizability. Importantly, all labels provided by the radiologists are supported by pathology reports, confirming the accuracy of the assigned labels.

The diversity encompassed in the TenD Modelling dataset makes it highly suitable for CNN training purposes. It captures the inherent variations in breast tumour US imaging, reflecting real-world scenarios and enhancing the CNN models' ability to generalize to unseen data. Including a larger number of images, along with the diverse nature of the dataset, contributes to developing more robust and reliable CNN models for BUS lesion classification.

### **2.4.1.3 Test1 Dataset**

The Test1 dataset is a valuable external testing dataset, distinct from the five medical centres where the Modelling dataset was collected. Testing CNN, or HC features-based ML algorithms using the Test1 dataset, provides an independent evaluation of the performance of those models trained on either Renmin or Modelling datasets. As for the other datasets, this dataset was carefully curated in a hospital located in Shanghai, China, ensuring diversity and independence from the training dataset sources.

Comprising a total of 306 images, the Test1 dataset consists of 189 benign and 117 malignant cases. Similar to the Renmin and Modelling datasets, the data collection and preparation process for the Test1 dataset follows rigorous protocols. Multiple US machines were employed during the data acquisition phase to account for the variability in imaging equipment encountered in real-world clinical settings.

Evaluating the model's performance on independent and previously unseen data provides a reliable measure of its effectiveness in real-world scenarios. Including a distinct medical centre and radiologist in the data collection process helps eliminate any potential bias or overfitting that may arise from using the same sources for training and testing. This evaluation contributes to the overall credibility and reliability of the developed models, ensuring their applicability and effectiveness in clinical settings. The Test1 dataset represents a critical benchmark for evaluating the performance of CNN models and serves as a bridge between the training phase and real-world deployment.

### **2.4.1.4 BUSI Dataset**

The BUSI dataset is an important external testing dataset in our research, providing an opportunity to evaluate the performance of trained CNN models on the TenD Renmin/Modelling dataset. The BUSI dataset is a publicly available dataset collected from Baheya Hospital for Early Detection and Treatment of Women's Cancer in Cairo, Egypt, in 2018 [61].

The BUSI dataset consists of 780 images, with 487 labelled as benign, 210 as malignant, and 133 as normal/clear. These images were acquired using the LOGIQ E9 and LOGIQ E9 Agile US systems. Each image in the dataset is accompanied by a corresponding label or class, indicating the nature of the breast tissue and a mask that facilitates tumour detection and segmentation.

Being focused on the classification of benign and malignant cases, we disregarded the class of normal/clear breast images, a total of 133 instances, as it was deemed irrelevant to our

research objectives. After this exclusion, a total of 524 images were selected for further analysis. Among these, 343 images were labelled benign, while 181 were labelled malignant. Any images with severe artefacts, such as lines, annotations, and calibre points, were removed from the dataset to ensure data quality and consistency.

Utilizing the BUSI dataset in our research enables the evaluation of trained CNN models on a distinct dataset from a different medical centre in Cairo. By employing this dataset as an external testing benchmark, we can assess the generalization and performance of the models developed using the TenD Renmin/Modelling dataset. The dataset's origins from a different medical centre, the utilization of a distinct US system, and the involvement of other radiologists contribute to the overall generalizability and applicability of the trained models in real-world scenarios.

#### **2.4.2 Data Cleaning, Split and Training/Testing Protocol**

In order to ensure data quality and consistency across all the BUS lesion datasets, a comprehensive data-cleaning process was conducted. This process involved several steps to identify and remove images with severe artefacts, thereby maintaining the integrity of the datasets. The following procedure, as explained in Figure 2.12 and depicted in Figure 2.13, was employed to clean the datasets:

1. **Lesion Boundary Points Drawing:** Each image in the datasets was carefully examined, and the corresponding lesion boundary was marked using yellow points.
2. **Bounding Box Determination:** Based on the marked boundary points, a fitted bounding box was generated, indicated by a yellow-coloured rectangle.
3. **Bounding Box Scaling:** The fitted bounding box was then scaled using a ratio of 1.5, resulting in a red-coloured bounding box.
4. **Artefact Removal:** Any image that contained severe artefacts, such as lines, annotations, or calibre points within the scaled bounding box, was removed from the datasets.

**Figure 2.12 Image data cleaning steps.**

This data-cleaning procedure was applied to all four datasets: Renmin, Modelling, Test1, and BUSI. As a result, the datasets were cleansed, ensuring that only good-quality images

without significant artefacts were included. After the cleaning process, the Renmin and Modelling datasets were selected as the training datasets. These datasets were utilized to train the CNN models for BUS lesion classification. On the other hand, the Test1 and BUSI datasets were designated as external testing datasets. These datasets were used to evaluate the performance of the trained models developed using the training datasets.

A 5-fold cross-validation approach was employed throughout the thesis as the training/testing protocol to conduct the experiments consistently and fairly. The training dataset was divided into five folds, with one fold reserved as an internal testing dataset and the remaining folds used for training and validation in an 80:20 ratio. For each experiment, this results in five trained models. During the testing phase, the performance of each model was evaluated on the testing datasets. The classification performance metrics, including accuracy, sensitivity, specificity, F1-score, and AUC, were computed for each model. The average and standard deviation of these metrics were calculated across the five trained models, providing a robust assessment of the model's performance.

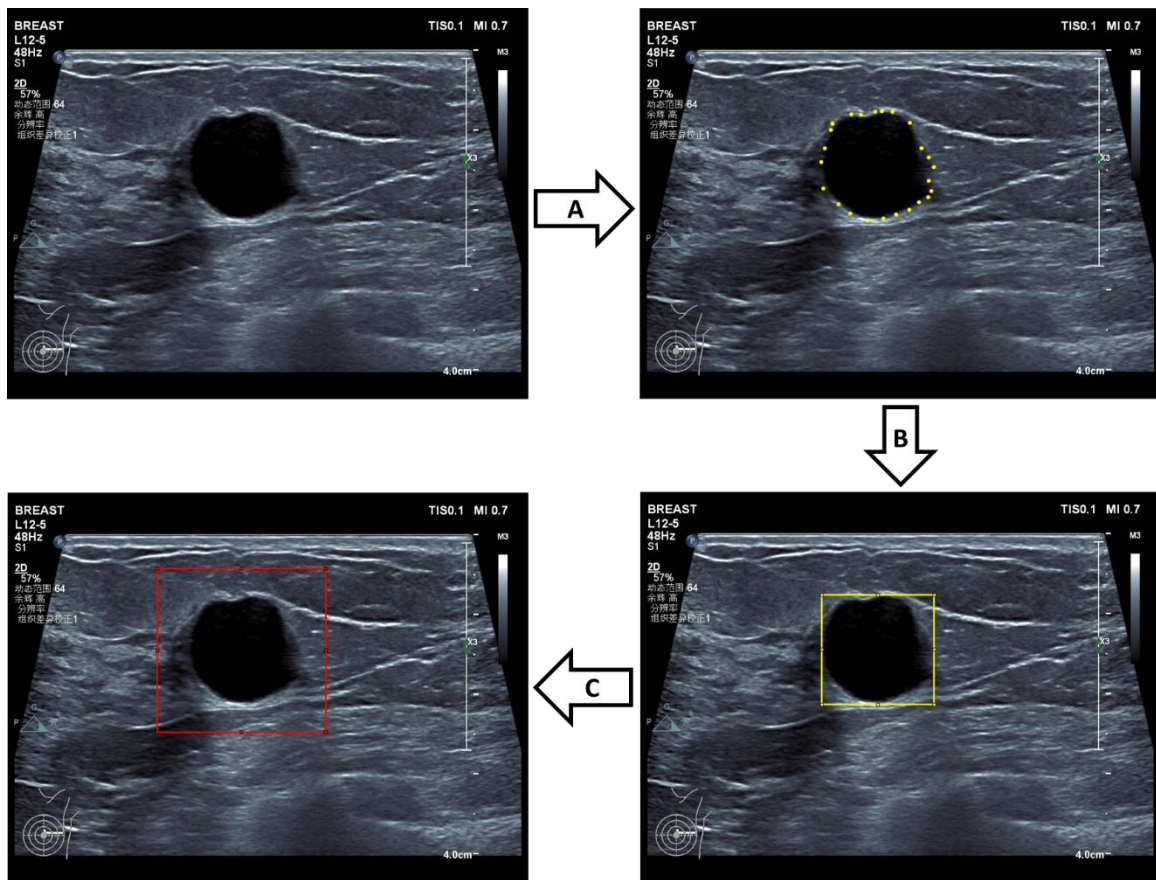


Figure 2.13 The data cleaning bounding box.

In the remaining sections, we will present in detail the classification performance metrics used in this thesis and discuss their relevance in evaluating the effectiveness of the developed CNN models for BUS lesion classification.

### 2.4.3 Machine Learning Classification Performance Metrics

Here, we explain the classification performance metrics used in this research, along with their mathematical formulas and advantages to the binary classification of BUS lesions (Benign vs. Malignant). First, we define the four key components of the metrics as follows:

**True Positive (TP):** In binary classification, TP refers to the number of positive instances that are correctly identified as positive by the model. These are the cases where the model accurately predicts the presence of the condition or event (Malignant).

**True Negative (TN):** TN represents the number of negative instances that are correctly identified as negative by the model. These are the cases where the model accurately predicts the absence of the condition or event (Benign).

**False Positive (FP):** FP indicates the number of negative instances that are incorrectly identified as positive by the model. These are the cases where the model incorrectly predicts the presence of the condition or event when it is actually absent (Benign predicted as Malignant).

**False Negative (FN):** FN represents the number of positive instances that are incorrectly identified as negative by the model. These are the cases where the model incorrectly predicts the absence of the condition or event when it is actually present (Malignant predicted as Benign).

In summary, TP and TN reflect the correct predictions made by the model for positive and negative instances, respectively. At the same time, FP and FN represent the incorrect predictions made by the model for negative and positive instances, respectively. Next, we shall define the five-performance metrics used in our work.

#### **Accuracy:**

Accuracy measures the overall correctness of the model's predictions by calculating the proportion of correctly classified samples out of the total number of samples in the dataset. It is a widely used metric in classification tasks. Mathematically, accuracy is computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad 2.11$$

Accuracy provides a general assessment of the model's correctness in predicting both benign and malignant BUS lesions. It indicates the model's performance overall, making it a valuable metric for evaluating the classifier's effectiveness.

**Sensitivity (Recall):**

Sensitivity, also known as recall or true positive rate, quantifies the model's ability to correctly identify positive instances, specifically malignant BUS lesions. It calculates the proportion of true positives correctly classified out of the total number of actual positive instances. Mathematically, sensitivity is expressed as:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad 2.12$$

Sensitivity is particularly crucial in breast cancer diagnosis as it measures the model's ability to detect malignant lesions accurately. Higher sensitivity implies a better ability to identify true positive cases, minimizing the risk of missing potentially cancerous lesions.

**Specificity:**

Specificity measures the model's ability to correctly identify negative instances, specifically benign BUS lesions. It determines the proportion of true negatives correctly classified out of the total number of actual negative instances. Mathematically, specificity is defined as:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad 2.13$$

Specificity is essential to ensure the accurate identification of benign lesions. A higher specificity indicates a better ability to avoid false positive classifications, reducing the risk of unnecessary invasive procedures.

**F1-Score:**

The F1-score is a composite metric that combines precision and recall (sensitivity) to provide an overall measure of the model's performance. It considers both false positives and false negatives and is particularly useful when dealing with imbalanced datasets. Mathematically, the F1-score is calculated as:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad 2.14$$

Where

$$Precision = \frac{TP}{TP + FP} \quad 2.15$$



The F1-score offers a balanced measure of the model's performance, considering both precision and recall (sensitivity). It is valuable when the dataset exhibits class imbalance, such as a higher number of benign cases than malignant ones.

**AUC (Area Under the ROC Curve):**

The AUC is a widely used performance metric that evaluates the model's ability to distinguish between positive and negative instances. The Receiver Operating Characteristic Curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various classification thresholds. The AUC represents the area under this curve and provides an aggregated measure of the model's performance. A higher AUC value indicates a better-performing model, while 0.5 represents random classification. See Figure 2.14. The AUC metric is advantageous as it assesses the overall discriminatory power of the model. It is insensitive to a specific classification threshold and provides a robust evaluation of the model's ability to differentiate between benign and malignant BUS lesions. A higher AUC implies a better ability to distinguish between the two classes.

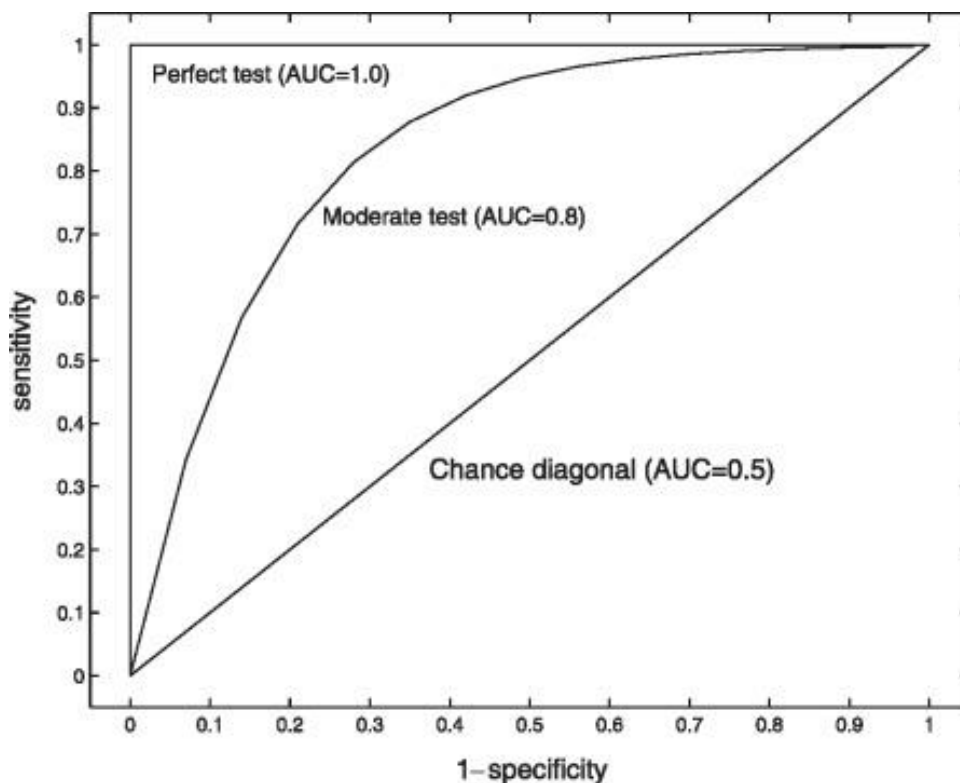


Figure 2.14 Area Under the Curve (AUC) [62].

By utilizing these performance metrics, we can comprehensively evaluate the accuracy, sensitivity, specificity, balance between precision and recall, and discriminatory power of the CNN models for the binary classification of BUS lesions. Each metric offers unique insights into the model's performance and aids in assessing its effectiveness and reliability for clinical decision-making.

## **2.5 A Review of DL Approaches for BUS Lesion Classification**

In this section, we review several articles relevant to applying DL models for BUS lesion classification. The literature offers various approaches for utilizing these models, including transfer learning (including fine-tuning version), training from scratch, and combining deep features with HC features. We aim to gain insights into their effectiveness in classifying lesions in BUS images.

Byra et al. [63] employed two approaches to transfer learning. In the first approach, the pre-trained VGG19 was used as a fixed feature extractor. The CNN's architecture was not modified, and the network was directly applied to the BUS images. Features were extracted from five max pooling layers of the VGG19 model, and these features were then averaged and normalized to form the final feature vector. An SVM classifier was then used for classification based on these extracted features. In the second approach, the CNN was fine-tuned using BUS images, and the architecture of the VGG19 model was modified, replacing the last layers with custom FC layers suitable for binary classification. The last convolutional block and the FCs were fine-tuned while keeping the first four blocks frozen. The fine-tuning was performed using the mini-batch stochastic gradient descent with Nesterov update. To enhance CNN's ability to recognize colour information, the authors introduced a matching layer. The matching layer performed a linear transformation on the grey scale US images, converting them into RGB images before feeding them into the pre-trained CNN. The parameters of the matching layer were learnt during fine-tuning to optimize the classification performance. The results showed that the fine-tuned CNN with the matching layer achieved the highest AUC value, outperforming the fixed feature extractor and SVM classifier. Additionally, the classification performance of the CNN-based approach was compared to the assessments made by four expert radiologists using the BI-RADS categories. The CNN exhibited higher AUC values than the radiologists, indicating its potential clinical usefulness in breast mass classification.

Tanaka et al. in [64] employed three CNN models in their study: VGG19, ResNet152, and an ensemble network. They fine-tuned these pre-trained models using their dataset of BUS

images by modifying the last FCL of the pre-trained models to match the number of classes in the dataset and then retraining the models on the BUS images. For classification, the CNN models took US images in patches as input. The patches were cropped from different views of the breast masses. The CNN models output class probability values for each patch, indicating the likelihood of the patch belonging to the benign or malignant class. The view-level classification was performed by averaging the class probability values from three patches cropped from each view, and the class with the highest probability was selected as the classified class for that view. The mass-level classification took all the class probability values of the patches cropped from all views of a mass and combined them to classify the entire mass. The ensemble network further improved the classification performance by combining the predictions of both VGG19 and ResNet152.

Cao et al. [65] proposed a breast lesion classification method using several CNN models. They collected a dataset of BUS images with annotations indicating benign or malignant lesions. The CNN architectures evaluated include AlexNet, ZFNet, VGG16, GoogLeNet, ResNet, and DenseNet. The experiments compared four scenarios: RoI with random initialization, RoI with transfer learning, full-size images with random initialization, and full-size images with transfer learning. The results showed that DenseNet achieves the best classification performance on their dataset. The authors concluded that transfer learning from the large-scale ImageNet dataset significantly improves classification accuracy for all CNN architectures. The study highlighted the potential benefits of using deep CNNs for breast lesion classification and demonstrated the importance of selecting appropriate architectures and utilizing transfer learning to enhance performance.

Zeimarani et al. [66] proposed a classification method for breast lesions using specially designed CNN architecture that consists of four convolutional layers, followed by two FCLs with ReLU activation functions and a SoftMax activation function for binary classification. The method involves a few preprocessing steps; the US images were resized to 224x224 pixels and balanced to ensure equal representation of benign and malignant cases. Zero-centring and normalization were applied to improve network performance. Image augmentation techniques, such as rotation, crops, and flips, were used to increase the training dataset size, effectively reducing overfitting. Regularization techniques like L2 regularization and dropout were applied to the network to further prevent overfitting. Different optimizers were evaluated, and Stochastic Gradient Descent with Momentum was selected as the candidate optimizer for training the CNN. The results were compared with other pre-trained CNN architectures and traditional ML methods. The proposed CNN-based

approach performed better than conventional ML methods and pre-trained CNN architectures for breast lesion classification in US images. Further improvements were planned, including gathering more data and exploring different CNN architectures with more hidden layers.

Wu et al. [67] proposed a classification method for breast lesions using pre-trained CNN models in transfer learning, mainly ResNet18 and ResNet50 models pre-trained on the ImageNet dataset. Their BUS image dataset consists of 131 images with 109 benign and 22 malignant lesions. For the un-pre-trained models (ResNet18 and ResNet50), the authors directly trained them from scratch on the BUS dataset. In contrast, the first two layers are frozen for the pre-trained models to utilize general features extracted from ImageNet. The convolutional layers' weights were initialized with pre-training, and the FC layer was modified to classify the images into benign or malignant tumours. The results demonstrated that the pre-trained transfer learning models, especially ResNet18, outperformed the un-pre-trained models significantly, and ResNet18 achieved the best performance.

Daoud et al. [35] proposed a method for BUS lesion classification using deep features extracted from a pre-trained VGG19 model and HC texture and morphological features. By utilizing a pre-trained VGG19 model, deep features were extracted from the BUS images at six different deep feature extraction levels. HC texture and morphological features were computed from the BUS images. The deep features extracted from the VGG19 model were combined with the HC texture and morphological features. A feature selection algorithm was applied to choose the most relevant features from the combined feature set. The selected features were used to train an SVM classifier. The study demonstrated that the best classification performance is achieved by combining the deep features from all convolution blocks of the VGG19 model with HC morphological features. The proposed approach outperformed other methods, including fine-tuned VGG19 and HC texture features, and showed promising generalization capabilities to other BUS image datasets.

While most existing approaches select a specific state-of-the-art pre-trained CNN model to retrain on a new dataset, a recent approach emerged that uses optimisation-based search to design the best performing customised CNN architecture for the given training dataset according to user input wish list. This approach led to the development of efficient and lightweight CNN models designed explicitly for BUS lesion classification. Mohammed et al. [68] proposed a classification method for breast lesion classification from US images using the ENAS (Efficient Neural Architecture Search) approach. The ENAS method is designed to automatically discover optimal CNN architectures tailored for the task. The

method involves two main stages: First, in the architecture Search Stage, the ENAS Micro approach was used to search for the best CNN architectures. A subset of their modelling dataset, containing both benign and malignant lesion images, was used for this purpose. The ENAS controller generates a set of cells, and the optimal cells with the highest validation test accuracies were selected to design the final CNN architecture. Second, Training and Generalization Evaluation Stage, the selected ENAS CNN models were trained from scratch on a balanced modelling dataset. Data augmentation techniques were used to enlarge the training set and reduce overfitting. The models' performance was evaluated using 5-fold cross-validation on the internal test set and then tested on two external test datasets. To reduce generalization errors, three methods were explored, including Reducing Model Complexity, Data Augmentation, and Using Unbalanced Data. The authors found that using an unbalanced dataset significantly reduces generalization errors in ENAS models. When comparing the ENAS-generated models with other existing CNN architectures, ENAS models outperformed other CNN models in terms of overall accuracy on both internal and external test sets.

In This thesis, we adopt the CNN fine-tuning approach for BUS lesion classification instead of training a CNN from scratch on our limited dataset or transfer learning. Fine-tuning allows the pre-trained model to adapt and learn relevant features from our BUS datasets, leveraging knowledge from the pre-trained model and improving performance in BUS lesion classification tasks. In the next chapter, we shall focus on the deployment of DL models for BUS lesion classification and the performance influencing factors in relation to data preparation and pre-processing.

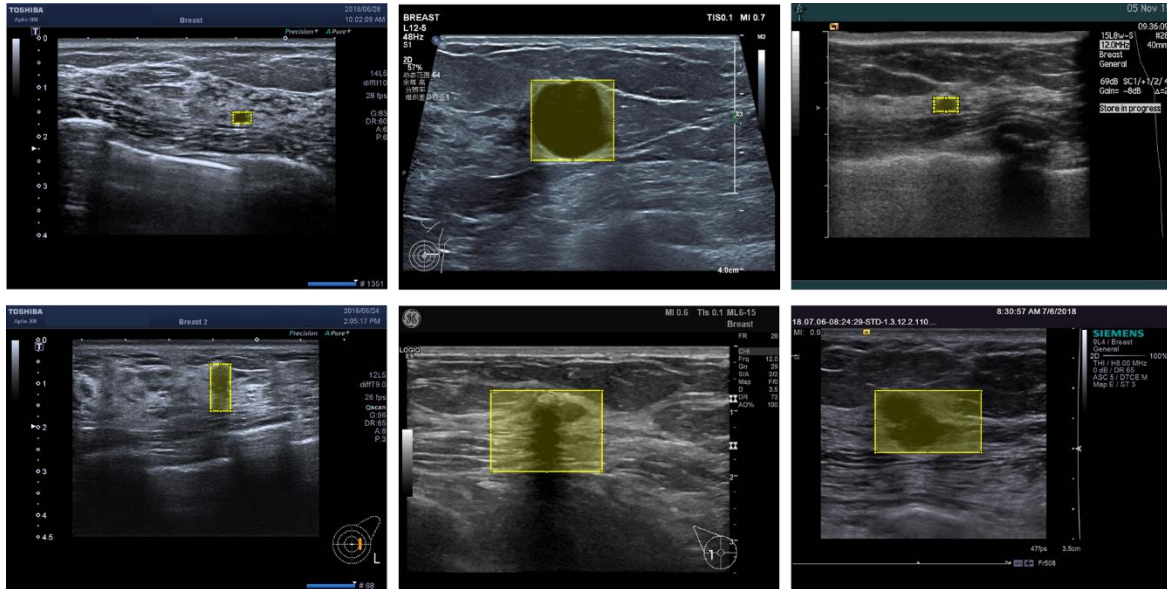
# Chapter 3: Deep Learning for Ultrasound Images - Performance Influencing Factors

Designing ML models for US image analysis is very challenging due to the lack of well-annotated and diverse training samples and the fact that US images are generally perceived as low quality compared to natural images [35]. In this chapter, we shall study the main challenges in developing DL models for the analysis of US B-mode scan images that radiologists and clinicians use for tumour diagnostic tasks. The performance of such DL models is influenced by a variety of factors, including size variation of tumour RoI, other tumour-related variations (shape, border clarity, etc.), variation of Radiologist level of expertise, variation in deployed US devices, variation of clinical practices by different centres, inter- and intra-observer variability, variation of RoI cropping procedures, variation of RoI image quality, and the lack of availability of sufficiently large samples of related images with standardised labelling. These factors are not independent of each other. For example, image quality variation can be influenced by variations in radiologist experience and variations in deployed US devices. This chapter focuses on three challenging factors that directly stem from image pre-processing requirements stipulated by the architecture of DL models for image analysis/classification: ***RoI size***, ***RoI image quality***, and ***availability of samples***. Besides describing the nature of these challenges, we shall discuss existing solutions and outline our strategy to deal with each. The rest of the thesis is devoted to the implementation and analysis of the outcome of this strategy.

## 3.1 Deep Learning Requirements on RoI Size - Challenges and Trends

The architecture of all state-of-the-art DL models for image analysis requires that all input images in the training and testing datasets must be of the same size, and different CNN models only differ marginally in the image size requirements. To deploy any DL model for the analysis of US scan images, all the tumour RoI images need to be resized to the fixed architecture stipulated input size. The input images to the CNN architectures investigated in this work are expected to be square images of size: (AlexNet\_227x227), (VGG16, VGG19, ResNet18, ResNet50, DensNet201\_224x224), and (Xception, InceptionV3\_299x299). Unlike the case of developing DL models for natural image analysis, this is a challenging requirement when these DL models are used to analyse datasets of US tumour scan images. Different patients undergoing US scanning are at different examination stages. Accordingly,

the recorded scan images may differ significantly in lesion size, including none for disease-free patients, and in tumour shape. It is highly unlikely that tumour shapes fit well into square boxes. Figure 3.1 below, displays a small sample of BUS scan images, collected in one hospital, where each tumour ROI region is marked by a yellow bounding box around it.



**Figure 3.1 ROI (Yellow bounding boxes) size variation among 6 different BUS tumour images.**

The cropped RoIs lesions are of different sizes, and the size variation in our experimental datasets is significant. Within the BUS datasets used for our investigations, we found that a cropped tumour ROI can be as small as 20x20 pixels and as large as 500x500 pixels. Figure 3.1 is only a modest illustration of the severity of this challenge facing the use of DL and other ML models to classify/analyse US scan images.

Image resizing is an obvious solution, but it is not harmless and may produce low-quality images, especially when the actual ROI is of low resolution (LR) and degraded. Figure 3.1 also illustrates the variation in ROI image quality/clarity. The variation of the ROI sizes influences the quality of resized images obtained by any resizing procedure. In the remaining part of this section, we shall first determine the severity of the ROI size variation challenge by examining different datasets of US images. Then, we shall present the common practice to deal with this problem, describe a different resizing procedure that was developed previously at the University of Buckingham, and analyse experimental work we conducted to compare the performance of various DL models when using these two different resizing solutions.

### 3.2 RoI Size variation and size normalisation

Prior to conducting statistical analysis on RoI size distribution in our BUS image datasets, we shall first describe the method used to determine the size of an RoI in a BUS image. The pipeline for developing/using any ML, including DL and HC features, the model starts with pre-processing steps that include manual/automatic procedures for detecting and segmenting the RoI (i.e., extracting the sought-after foreground (lesion) from the background). This step is designed to enable the representation of the RoI image in a digital form, and it is an important part of data preparation. Since the output RoIs have no standardized shape, it is customary to draw a rectangular bounding box surrounding the RoI. In our DL-based pipeline for tumour classification, our collaborative partner (TenD-Innovation, China) recorded various raw B-mode US scan images for different cancer diseases from different hospitals that were recruited over several years. Due to time constraints and the desire to produce and test the performance of ML analysis software in a short time, we adopted a non-automatic detection and segmentation procedure. Figure 3.2 below describes this process.

1. A radiologist manually examines the various frames of an US scanning video, recorded earlier and
  - a. identify the best frame (according to his/her view and expertise) that contains the potential tumour tissue, if any, and
  - b. select the tumour RoI by marking a sufficient number of tumour boundary points to be used for cropping the tumour RoI.
2. Input the marked lesion boundary points into a standard computer procedure that
  - a. Compute a polygonal shape (tumour area) defined by the order of the selected points, and
  - b. Draw the smallest rectangular box circumscribing the polygonal shape (RoI).

**Figure 3.2 Medical US Data Preparation Process.**

Figure 3.3 below illustrates this procedure stepwise with an actual breast tumour scan. This process is straightforward and produces a good visual representation of the tumour RoI when it is of reasonable size with sufficient marked lesion boundary points and/or its tumour polygonal boundary is not far from being convex (see Figure 3.1). In the literature, this pre-processing procedure is referred to as *cropping*. Chapter 4 is dedicated to investigating RoI cropping scenarios and determining optimal cropping for better model classification performance in more detail.



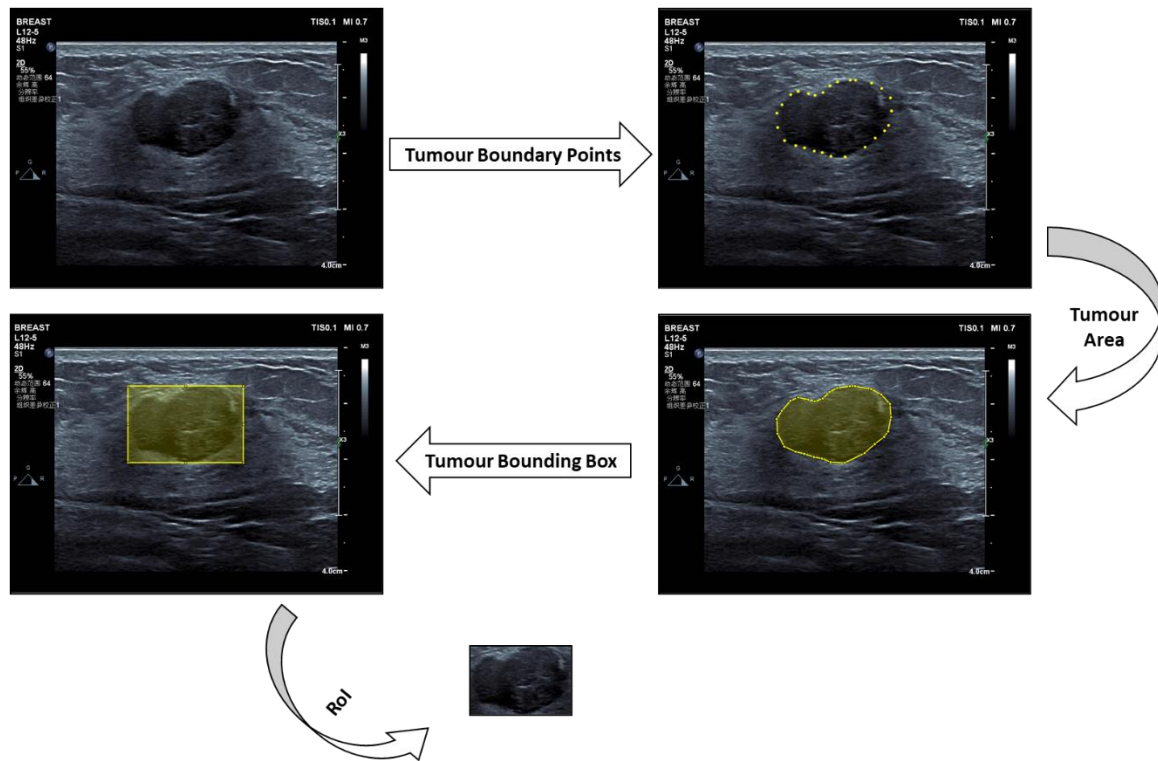


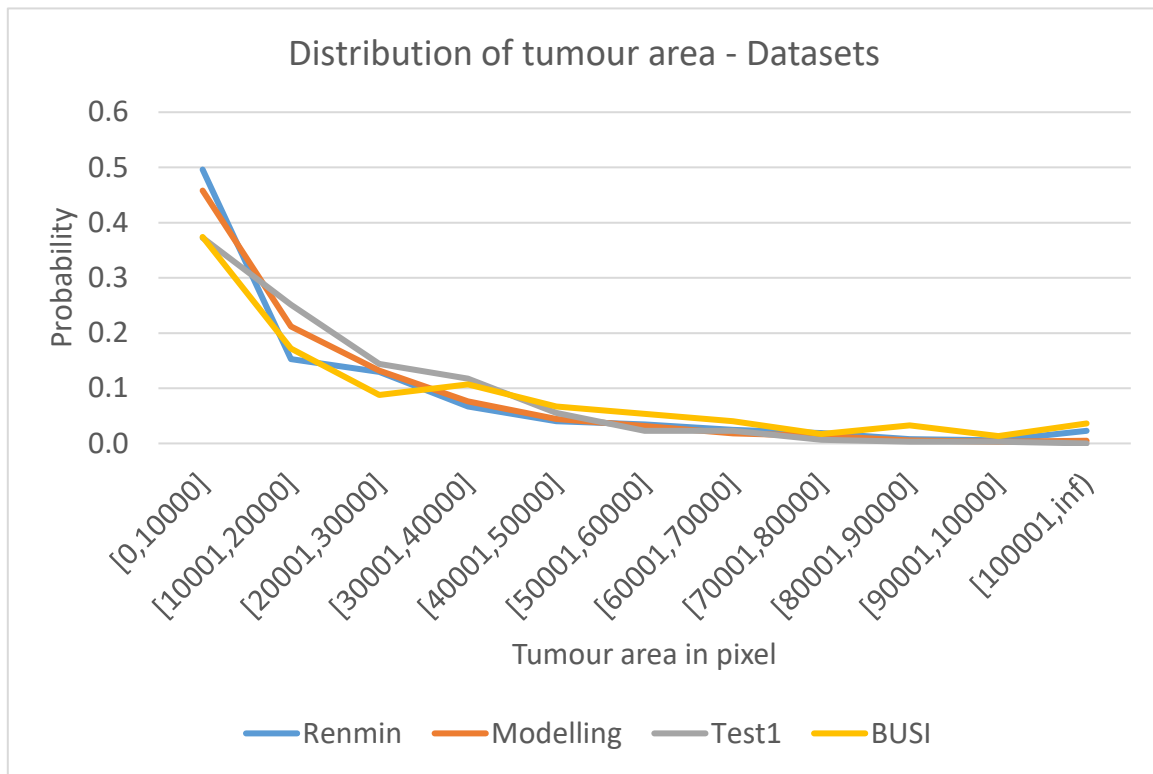
Figure 3.3 Steps of preparing and determining the digital representation of tumour RoI.

### 3.2.1 RoI size variation

This research project is part of a collaborative arrangement between the University of Buckingham and the rather new Chinese company (TenD-Innovation). Over the last few years, TenD-Innovation gradually established links with several hospitals, mainly in the municipality of Shanghai, to record datasets of US scan images of different types of tumour tissues/organs. Accordingly, my research investigations initially had access to the dataset from the first hospital (Renmin), and only much later, datasets from other hospitals became available to us. This section presents the results of several statistical studies conducted for our experimental image datasets to illustrate the severity of the RoI size variation challenges. We mainly use tumour-area measured by the actual number of pixels inside the tumour polygonal shape. We conducted our statistical studies on the four BUS image datasets.

Figure 3.4 presents the breast tumour-area size distribution for each of the four datasets separately. We observe a similar RoI size distribution pattern (negative exponential pattern) across different datasets. In all cases, the majority of the tumours are in the range of  $[0, 10000]$  pixels, which we refer to as the range of small-size tumours, but the proportion of small-size RoI to the total varies for different datasets ranging from around 37% for Test1 and BUSI datasets to 50% for the Renmin dataset, then the number of the tumours in larger

size ranges decreases gradually. This illustrates the fact that the majority of the tumours in the breast datasets are of small sizes and need to be upscaled prior to input into DL models.



**Figure 3.4** The distribution of tumour-area in pixel for the four BUSI datasets.

Next, we computed the corresponding distribution for each tumour class (Benign, Malignant) to determine if these distributions are tumour class dependent. Figures 3.5-8 present the tumour-size distribution for each dataset and their two classes, Benign and Malignant, separately. In all four datasets, the Majority of Benign cases are in the range of the smallest tumour-size interval [0, 10000] pixels, and this number decreases gradually for the larger size ranges similar to the size distribution of the whole dataset. On the other hand, the size distribution of Malignant tumours is different from Benign. The majority of the tumours are in the range of [10000, 40000] pixels, which shows that the overall size of RoI malignant cases is larger than benign cases. Thus, tumour size is relatively class dependent.

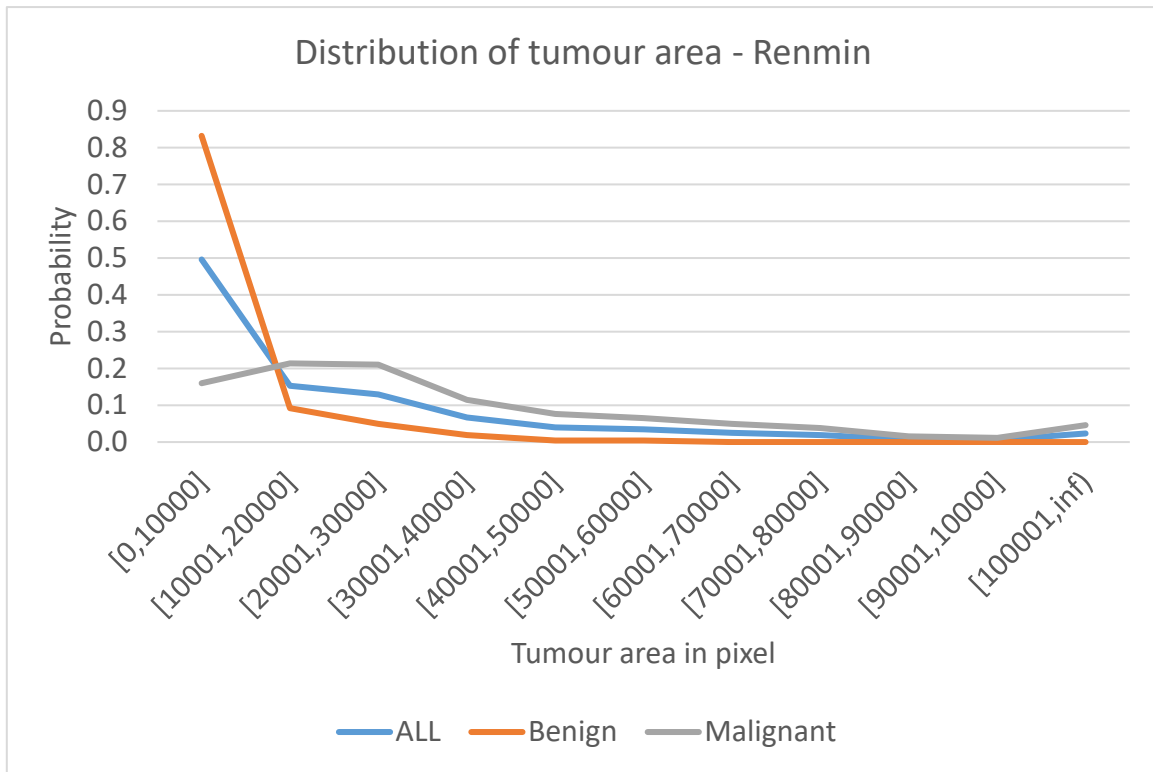


Figure 3.5 The distribution of tumour area in pixel for the Renmin dataset.

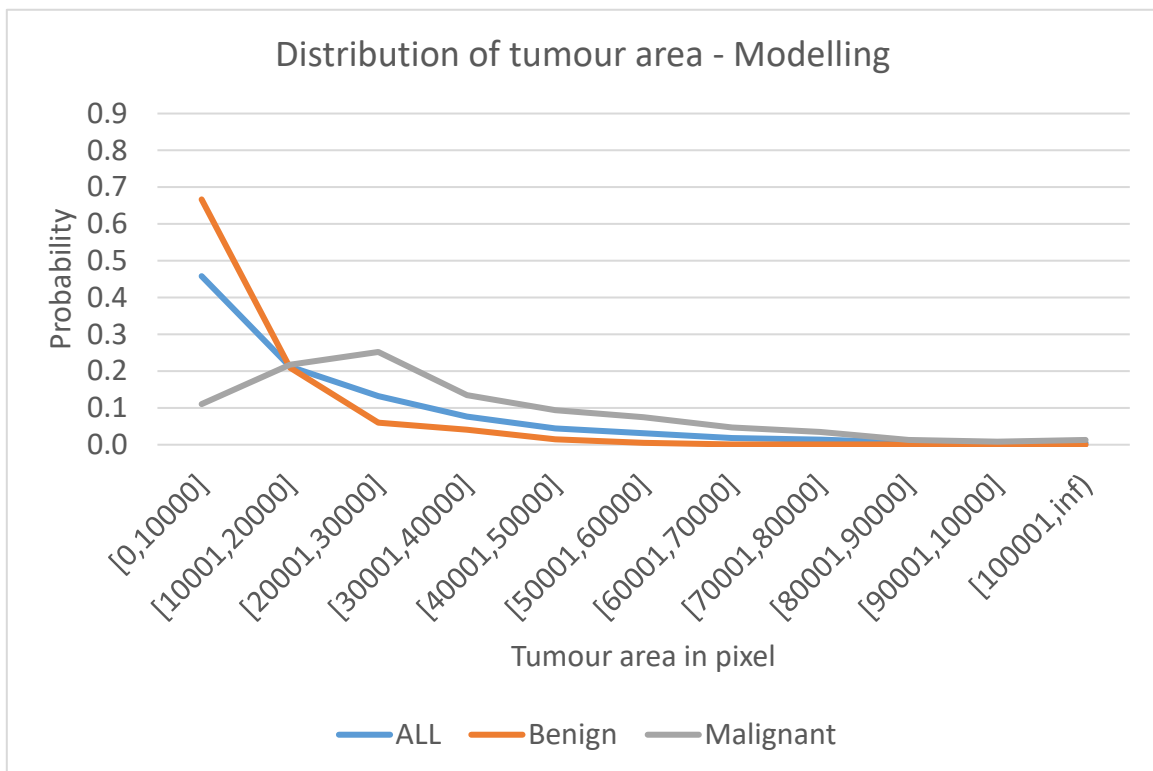


Figure 3.6 The distribution of tumour area in pixel for the Modelling dataset.

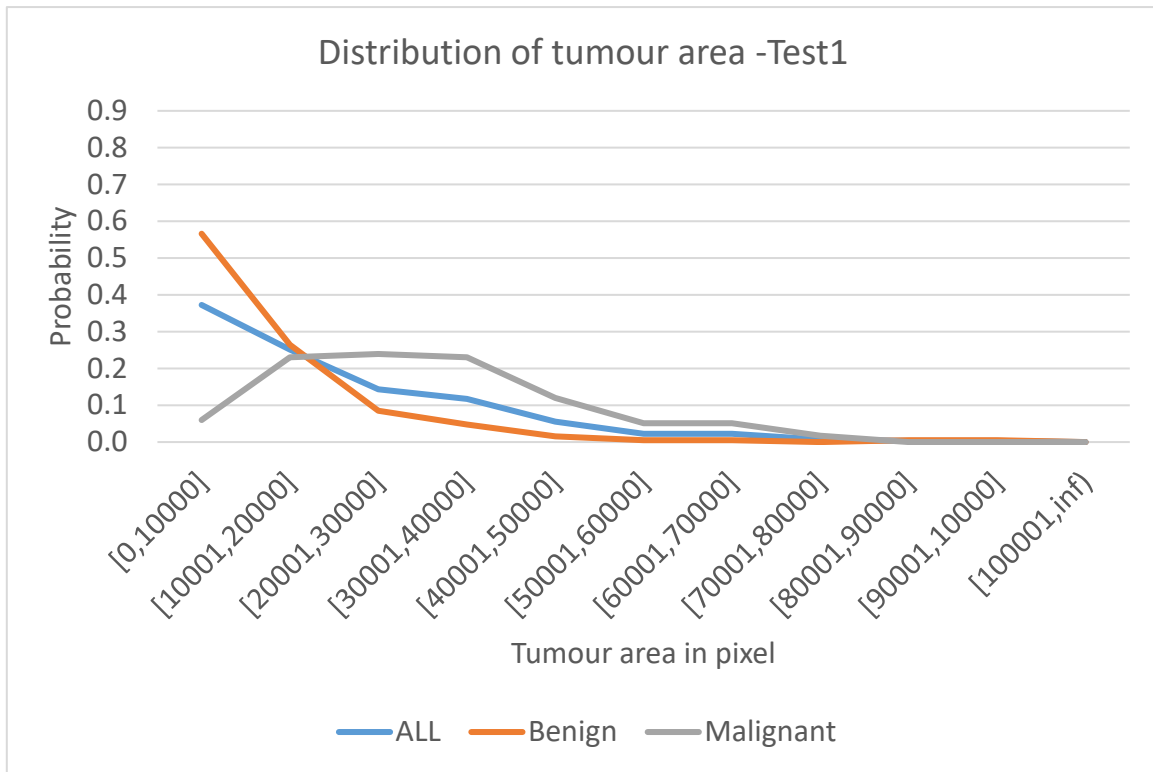


Figure 3.7 The distribution of tumour area in pixel for the Test1 dataset.

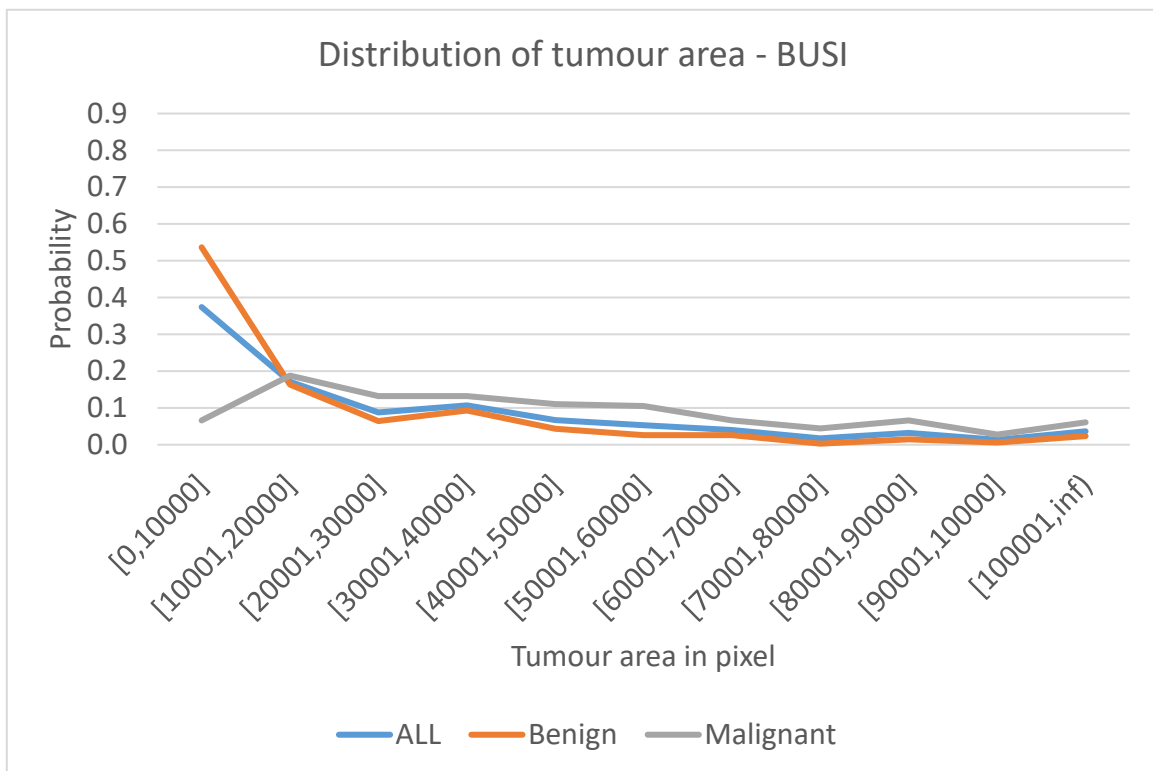
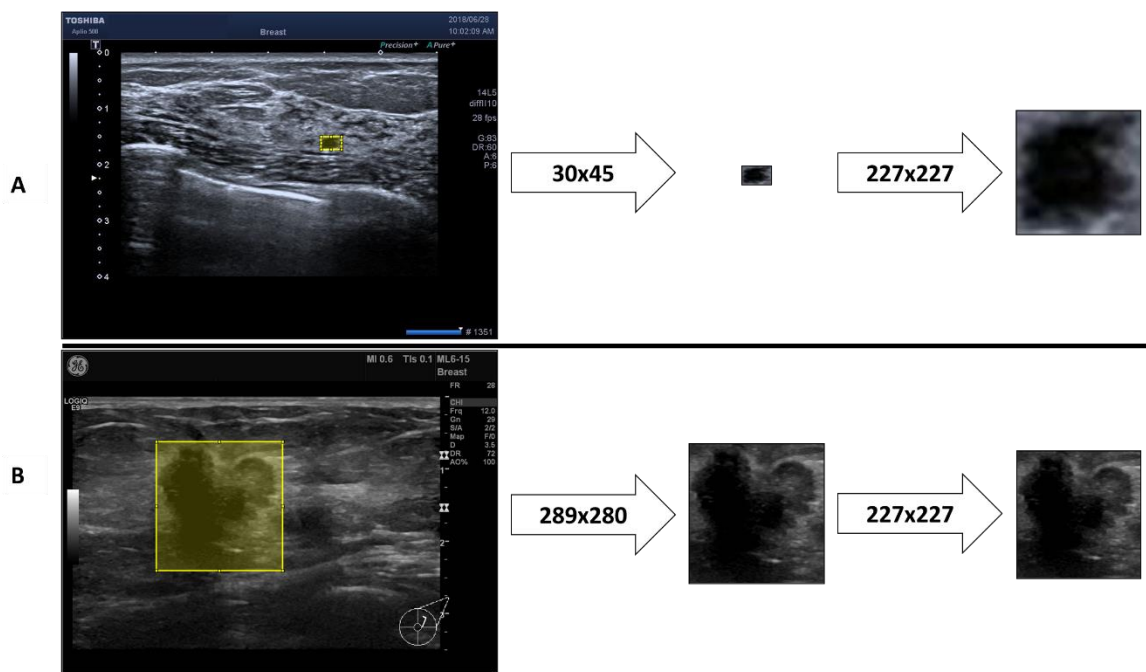


Figure 3.8 The distribution of tumour area in pixel for the BUSI dataset.

Now, a fixed RoI size is required for all DL pipelines, and all the RoIs have to be resized according to the adopted DL architecture input size. In this case, the tumour area of the resized RoIs is required to be approximately in the range [40000, 60000] pixels. All the RoIs are to be resized to the required CNN input size, and the above tables show that most of the RoIs need to be upscaled as they are smaller than the input size. For the majority of actual RoIs (benign and malignant), the tumour area is in the range [0, 10000] pixels which means that they need to be significantly upscaled. Image resizing, a commonly required pre-processing technique, is not harmless, especially when the original resolution is too low. In contrast, image downscaling often results in better contrast. Figure 3.9 illustrates this assertion: when upscaled a 30x45 RoI tumour into 227x227 (the input size of AlexNet), the resulting image is of low quality in that it is severely degraded, blurry, very fuzzy and pixelated in places. On the other hand, the downsampled image of a 289x280 RoI tumour results in seemingly improved quality.



**Figure 3.9 Resizing two BUS RoIs from 30x45 (A) and 289x280 (B) into 227x227.**

After resizing, the quality of middle-size RoIs is not expected to differ significantly from that of the original RoI images. Now, the fact that the majority of benign cases are in the lowest resolution while only 7%-17% of malignant cases fall in this category, the resizing will adversely affect the quality of benign cases more than malignant cases. Therefore, after cropping all the RoIs and resizing small-size RoIs, there may be a need to improve the quality of the resized RoIs by adopting further image pre-processing techniques, including

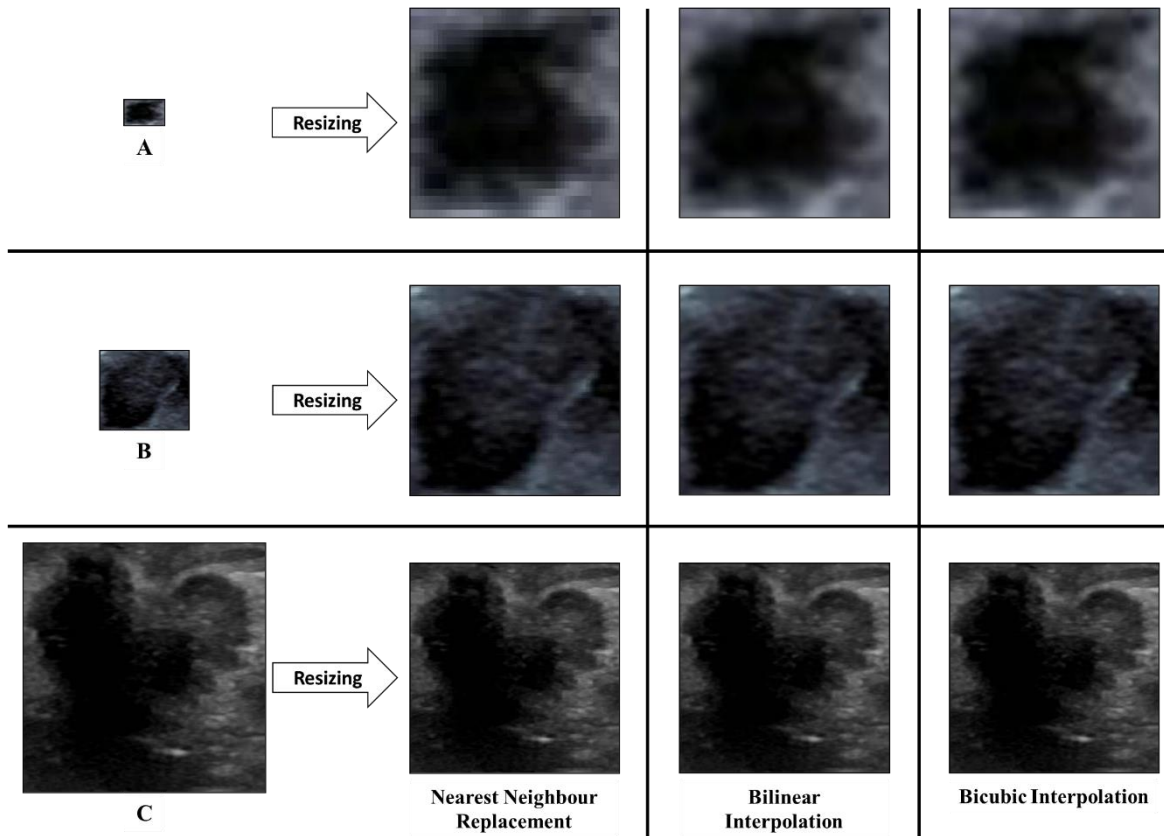
resolution enhancement, denoising, deblurring and other image enhancement techniques; this may help to reduce the distortion caused by size variation and upscaling small-size RoIs input to a CNN model. It is worth noting that US image quality is difficult to be measured using the quality measures for natural images. Section 3.3 will briefly discuss the image quality factor and its influence on the performance of ML in US image analysis. Next, we shall explain the RoI resizing and resolution enhancement techniques.

### **3.2.2 RoI size normalization**

Image resizing is the process of upscaling/downscaling an image in which the pixel values of the resized image are usually determined by an interpolation method after mapping them back to a location in the original image. The resizing techniques are known to produce various artefacts, including blurring and aliasing [69]. An alternative method for image resizing is the Super Resolution (SR) procedure which exploits the benefits of the compressed sensing (CS) paradigm. However, SR is designed to enlarge small natural images while maintaining/improving image quality.

The conventional way for single-image resolution enhancement is interpolation-based resizing techniques, including Nearest Neighbour Replacement (NNR), Bilinear Interpolation, and Bicubic Interpolation (BiCubic). All these methods start by creating a rectangular uniform grid for the resized image and map it onto the original image grid, but they differ in how they define the intensity of the resized image pixels. The NNR resizing technique is straightforward and simply sets the intensity of each new image pixel to the pixel value of the closest neighbour defined by the resizing map. It has a minimum computational cost, but when applied to natural images, the resized image is usually blurry with aliasing/blocking artefacts [70], [71]. The Bilinear interpolation determines the value of the resized image pixels by interpolating in both horizontal and vertical directions as a weighted average of the immediate 4 neighbouring pixels, determined by the resizing map. The computational cost of the Bilinear method is slightly more than that of the NNR method, and for natural images, the resized image is relatively smoother with less aliasing/blocking artefact [72]. The BiCubic method produces sharper and better-quality natural images than the two previous methods. It again determines the intensity of new pixels by interpolating in horizontal and vertical directions. It uses cubic interpolation as a weighted average of the nearest weighted 4x4 neighbouring pixels in the original image for a total of 16 pixels. In each direction, the 4 neighbouring are at various distances from the resized pixel, and closer pixels are given more weights. BiCubic is a balanced combination of image quality and

complexity; therefore, it is the adopted image resizing in many imaging software such as Photoshop [72]. Due to the significant variation in RoI tumour sizes in US images, it is prudent to study the effect on image quality of each of these methods for different classes of image size. Figure 3.10 displays the outputs of size 227x227 images when resizing 3 different sizes of RoI tumour images: small (A), mid-size (B) and large-size (C) tumours.



**Figure 3.10 Resizing 3 different size RoI tumour images A, B, and C using 3 interpolation methods.**

Figure 3.10 illustrates that in all cases, the BiCubic outperforms the other 2 methods but to a lesser extent for the large-size RoIs. As can be seen, the resizing procedure affects the quality of the small lesion (A) significantly, while the quality of the mid-size tumour (B) is maintained, and for the large-size tumour (C) is even slightly improved.

Other well-known interpolation-based image resizing techniques include Cubic B-Spline, Mitchell-Netravali approximation, Catmull-Rom Interpolation, and Lanczos interpolation. These techniques produce good-quality images after resizing comparable to the ones produced by BiCubic; however, they are computationally more expensive and less popular. For more details, see [73], [74].

The fact that small-size RoIs are more prevalent than mid and large-size is an incentive to investigate the use of the Compressed Sensing Super Resolution (CSSR) technique known to maintain/improve the quality of small and degraded images after upscaling [75], [76].

### 3.2.3 Image Super Resolution

Even for natural images, the quality of resized images by any interpolation-based resizing technique is influenced not only by the original image size but also by the quality and the amount of information in the image. This is the nature of the interpolation methods as they combine the existing frequencies to generate the resized image rather than recovering/maintaining high-frequency image parts. However, the high-frequency image parts of the LR images can be recovered by using certain transformations like Wavelet and Fourier in the frequency domain [77]. Recovering such image information is known as Image Restoration, an inverse problem, to recover an assumed high-quality image of the captured object/scene from a blurry/degraded noisy one. The restoration process involves modelling the distortion/degradation and conducting an inverse technique to recover the undistorted image. Inverse and Wiener filters are the conventional ways of image restoration; see [77]. SR is an effective image recovery technique to obtain a high resolution (HR) image of good quality from a single/multiple degraded LR image(s) of the same scene. It combines image registration, interpolation, and restoration in one algorithm [78], [79]. There are several methods of implementing SR. For example, the multi-image SR utilizes multiple LR images of the same object. The process requires registering and misaligning the LR images and combining them into one HR image through fusion/interpolation. Another well-known SR method is reference-based SR. It attains a HR image from a LR image while another HR image with similar content to the LR image is provided, referred to as the reference image. Such a strategy has been shown to work effectively by the recently proposed algorithm of “Image Super-Resolution by Neural Texture Transfer” [80]. The fundamental problem for our work is the unrealistic requirement of a HR reference image. Recovering a HR image from a LR version can also be accomplished via Single Image Super Resolution (SISR) techniques. Interpolation-based approaches for image resizing as a SISR do not offer the best quality possible because they somehow lose the crisp edges, causing the resized image to be blurry. Next, we briefly describe the Mathematical model of SR.



### 3.2.3.1 The Mathematical Model of Super Resolution

Given a LR (i.e., small size) image(s), SR is meant to recover a HR version of the image by adding additional details, modelled as an inverse problem solved by optimization. The main assumption of SR is that the small size degraded image(s) is(are) obtained from a hypothetically assumed good-quality HR image  $X$  by a combined process of Blurring and Subsampling, i.e., the observed LR images are simply obtained by the following formula:

$$Y_k = S * B * X \quad 3.1$$

Where  $k = 1, 2, 3, \dots, P$ , is the number of the LR images.

Here, we only deal with SISR techniques, i.e., when  $k = 1$ . In any case, the functions  $B$  and  $S$  represent the point spread function that results in blurring and down-sampling, respectively. Traditional solutions of the above-described optimization involve an iterative procedure, as explained in Figure 3.11:

1. Combine the input observed  $Y_k$  image(s) (e.g., via interleaving) into a first approximation  $X_0$  of the SR image.
2. Apply blurring/degrading and down sampling procedures to obtain new  $k$  versions of LR images and compute the objective/cost function representing the error between current LR versions and the previous versions.
3. Use the error function to update the observed LR images.
4. Repeat until the objective function reaches the optimal error.

**Figure 3.11 The iterative procedure of Super Resolution problem.**

In order to avoid divergence and infinite loops when using this iterative procedure, it is customary to use a regularization term that enforces certain constraints on the solution, such as smoothness or sparsity. This helps to avoid overfitting and produce a more realistic and coherent HR image.

A major challenge is determining an appropriate image degradation/blurring function method independent of the capturing conditions/devices. The emergence of CNN image analysis models points to a potential method using a set of Gaussian convolution filters to generate different blurred versions of any image. Dong et al. [81] proposed the SR-CNN model to generate HR images from their LR and degraded counterparts. This approach is

outside the current realm of our thesis, but readers interested in related work and follow-ups are referred to [82]–[84]. Instead, we shall adopt the CSSR approach.

### 3.2.3.2 Compressed Sensing-based Single Image Super Resolution

Compressed Sensing (CS), also known as sparse recovery, is a signal sampling method that maintains the important aspects of data without including significant redundancies. The main idea of CS is that high dimensional signals that are sparse (or can be sparsified) can be recovered from much lower measurements/samples than the Nyquist-Shannon Sampling theory stipulated. In terms of SR, a LR image is assumed to be the output of linear dimension reduction applied on a high dimensional super-resolved signal. If the over-complete dimension reduction matrix (referred to as the dictionary) satisfies the Restricted Isometry Property (RIP), then the CS theory guarantees the recovery of the Super-resolved image. The solution by the least square method minimizes the Euclidean norm, but the solution is not unique. Enforcing the sparsity constraint enables solving the system, which minimises the  $L^1$  norm (i.e., Manhattan norm) by linear programming. The RIP is a necessary condition for the unique recovery of the sparse solution [75].

CSSR techniques use dictionary learning approaches that have been studied in recent years to super-resolve single LR images (see, e.g., [85], [86]). These approaches work in multiple steps: (1) overlapping patches of the LR image are densely processed, (2) encode the patches using a Low-Resolution Dictionary (LD) to determine the sparse representation of the flattened patches with few coefficients, (3) the sparse representations are fed into a High-Resolution dictionary (HD), used to recover HR patches. The construction and the performance of the LD and HD highly depend on the training samples, as the columns of such dictionaries are built from a number of LR and HR random patches sampled from the training images.

A useful implementation of the above CS-based SISR approach is the deployment of data-independent CS-compliant dictionaries. Gaussian matrices have been used as a rich source of such dictionaries due to their known blurring effects. For my MSc research project [76], I extensively studied this approach and developed CS dictionaries without training. This approach is ideally suitable for our application as our training dataset is relatively small, and US images are generally low-quality. It is unrealistic to create dictionaries using US images. Our implemented CSSR algorithm process 5x5 image patches, while the overcomplete dictionaries are constructed as submatrices of Hadamard matrices of appropriate size due to

their orthogonality properties that ensure stable generation of super-resolved images. Next, we shall briefly define Hadamard matrices.

An  $n \times n$  matrix  $H$  is a Hadamard matrix if it satisfies the condition ( $H * H^T = H^T * H = n * I$ ), when  $H^T$  is the transpose of  $H$ , and  $I$  is the identity matrix of size  $n \times n$ . This means that the dot product of any row or column with itself is equal to the order/size of the matrix. Square Hadamard matrices have either 1 or -1 entries, and their rows are mutually orthogonal. Figure 3.12 presents the three well-known types of Hadamard matrices: Sylvester-type, Walsh-Paley, and Walsh of size 16x16. For more detail on Hadamard matrices and their generation methods, please see [76], [87].



**Figure 3.12 Binary display of Sylvester type, Walsh-Paley, and Walsh Matrices of size 16x16.**

There are different approaches to constructing the pair of CS-LD and HD from Hadamard matrices, and it is proven that such over-complete dictionaries satisfy the RIP [76]. In the current work, the pair of dictionaries are set as a 25x512-HD and a 100x512-LD. We construct the pair of LD and HD from Walsh matrices of size 512x512 by selecting the top  $k$  rows, where  $k=25$  for the HD and  $k=100$  for the LD. The choice of matrices of 512 columns is independent of the sizes of the images but is based on the fact that full Hadamard matrices are of the order  $2^m$ , where  $m$  is a positive integer. When used for natural images, this method produces a super-resolved image of a desired size with improved/maintained quality from an input-degraded LR image, [75], [76]. Our SR procedure works in 6 steps as follows; see Figure 3.13.

1. Bicubic interpolation is used to upscale the LR image to the desired size.
2. Produce four feature images by applying four 1D filters which are first and second order gradient filters to the interpolated image to emphasise the edges/details in various directions (Horizontal and Vertical).
3. Split the four gradient feature images into overlap patches of 5\*5 pixels and convert each patch into a 25-pixel-long column vector. Then, concatenate the matching column vectors for each patch position to produce a Y-column vector of size of  $100 = 4*25$ .
4. For each Y, solve the L1-minimization problem to find a sparse vector  $\alpha$  of the underdetermined equation  $y = LD*\alpha$ .
5. Using the coefficients of  $\alpha$  and HD dictionary, the output HR X-patch is created as a linear combination of them  $X = HD*\alpha$ .
6. In order to remove potential artefacts from the local sparse and correct any reconstruction error in the HR image, Iterative Back-Project is applied to the single image.

**Figure 3.13 The steps of our CS-based SISR procedure.**

Figure 3.14 illustrates the steps of our CS-SISR algorithm, and Figure 3.15 displays its effect on a resized RoI. Figure 3.15 shows that the contrast of certain areas in the SR image is higher than in the BiCubic enlarged image. The visual appearance of SR images is better than BiCubic images, especially for small-size tumours, as testified by an experienced radiologist who provided the US images used in the experiments below; he stated that the overall contrast of SR images is better, and the tumour boundary is more precise.

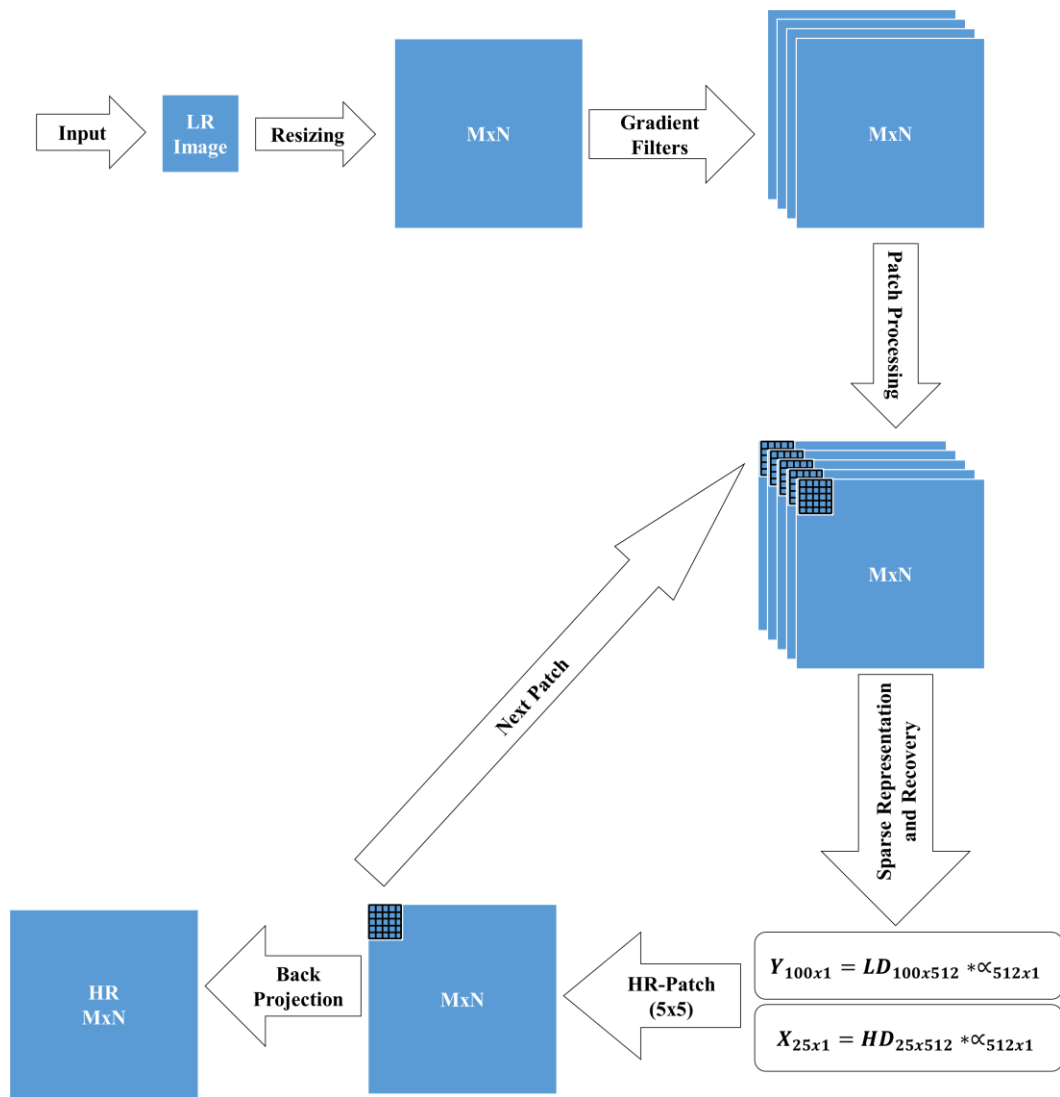


Figure 3.14 CS-based SISR algorithm [75].

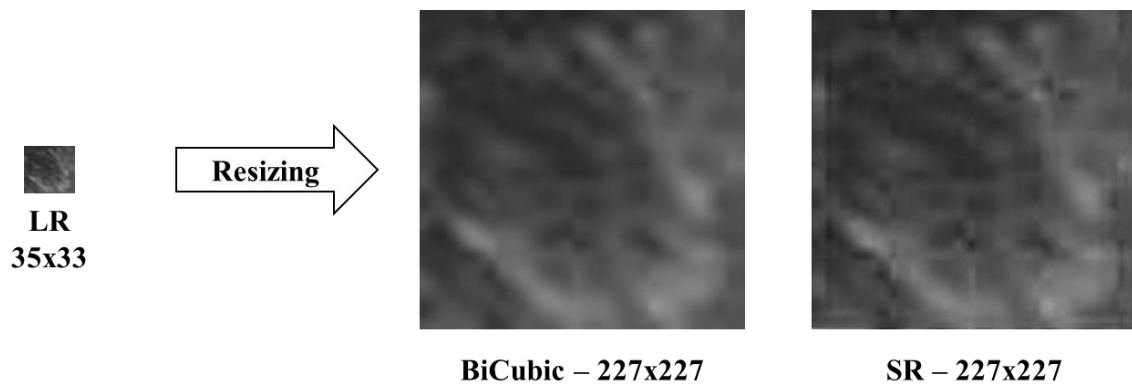


Figure 3.15 BiCubic vs. SR for resizing a LR BUS image.

### 3.2.3.3 Experimental Classification Results – BiCubic vs. SR

In this section, we present experimental results to test and compare the impact of BiCubic image resizing versus (vs.) SISR using the proposed CS-based algorithm. Although HC-ML schemes do not require image resizing and we can deal with size variation by normalising the selected HC feature, we shall present the effect of the same level RoI tumour resizing on the performance of a few such schemes. For these schemes, there is no constraint on resizing all the RoIs into a certain size as CNN architectures. The RoI tumour images are intentionally resized into 128x128 for the HC features. In this case, the tumour area post-resizing is approximately in the range of [10000, 15000] pixels which is close to the size range of the majority of the tumours. Thus, there is no need for significant RoI image upscaling, and the quality of the resized RoIs is less affected by the resizing procedure. We use SVM with a linear kernel for classification.

The Renmin dataset is the adopted dataset in these experiments. In the early stages of this research project, this was the only dataset made available to us. We conducted performance testing on this dataset, post the 2 resizing methods, for a set of DL models as well as 2 HC features. In these experiments, the tumour area is cropped using the standard method described earlier (See Figure 3.3), as the tissue padded smallest fitted bounding box to the tumour area polygonal shape. The RoI boxes were resized for the CNN models according to their architecture input size (AlexNet\_227x227), (VGG16, VGG19, ResNet18\_224x224). The experimental results for these two resizing methods are presented in Tables 3.1 and 3.2.

**Table 3.1 Performance of CNNs and HC features on Renmin dataset with BiCubic resizing.**

Renmin	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.92 <math>\pm</math> 0.03</b>	<b>0.93 <math>\pm</math> 0.05</b>	<b>0.90 <math>\pm</math> 0.04</b>	<b>0.92 <math>\pm</math> 0.03</b>	<b>0.92 <math>\pm</math> 0.03</b>
VGG16	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.04</b>	<b>0.92 <math>\pm</math> 0.06</b>	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.02</b>
VGG19	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.94 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.02</b>
ResNet18	<b>0.91 <math>\pm</math> 0.03</b>	<b>0.91 <math>\pm</math> 0.07</b>	<b>0.90 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.03</b>	<b>0.91 <math>\pm</math> 0.03</b>
HOG	<b>0.85 <math>\pm</math> 0.03</b>	<b>0.86 <math>\pm</math> 0.05</b>	<b>0.84 <math>\pm</math> 0.03</b>	<b>0.85 <math>\pm</math> 0.03</b>	<b>0.85 <math>\pm</math> 0.03</b>
ULBP	<b>0.83 <math>\pm</math> 0.04</b>	<b>0.87 <math>\pm</math> 0.04</b>	<b>0.79 <math>\pm</math> 0.07</b>	<b>0.83 <math>\pm</math> 0.03</b>	<b>0.83 <math>\pm</math> 0.04</b>

**Table 3.2 Performance of CNNs and HC features on Renmin dataset with CS-SISR resizing.**

Renmin	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.92 <math>\pm</math> 0.03</b>	<b>0.94 <math>\pm</math> 0.04</b>	<b>0.91 <math>\pm</math> 0.07</b>	<b>0.93 <math>\pm</math> 0.03</b>	<b>0.92 <math>\pm</math> 0.03</b>
VGG16	<b>0.91 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.04</b>	<b>0.90 <math>\pm</math> 0.05</b>	<b>0.91 <math>\pm</math> 0.01</b>	<b>0.91 <math>\pm</math> 0.01</b>
VGG19	<b>0.92 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.04</b>	<b>0.92 <math>\pm</math> 0.03</b>	<b>0.92 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.01</b>
ResNet18	<b>0.91 <math>\pm</math> 0.03</b>	<b>0.90 <math>\pm</math> 0.08</b>	<b>0.91 <math>\pm</math> 0.04</b>	<b>0.90 <math>\pm</math> 0.03</b>	<b>0.91 <math>\pm</math> 0.03</b>
HOG	<b>0.82 <math>\pm</math> 0.02</b>	<b>0.82 <math>\pm</math> 0.05</b>	<b>0.82 <math>\pm</math> 0.03</b>	<b>0.82 <math>\pm</math> 0.02</b>	<b>0.82 <math>\pm</math> 0.02</b>
ULBP	<b>0.85 <math>\pm</math> 0.03</b>	<b>0.85 <math>\pm</math> 0.04</b>	<b>0.85 <math>\pm</math> 0.05</b>	<b>0.85 <math>\pm</math> 0.03</b>	<b>0.85 <math>\pm</math> 0.03</b>

Overall, all DL models and HC algorithms with both image resizing methods have High-to-Excellent performances in differentiating benign from malignant cases, with DL schemes outperforming the HC feature schemes. For the DL schemes, the overall accuracy is not significantly different for the 2 resizing schemes, except that for the VGG19 scheme, the SR resizing marginally outperforms the BiCubic scheme. We note that this marginal improvement makes VGG19 with SR achieve similar accuracy to AlexNet but with a tighter standard deviation. In terms of other performance metrics (Sensitivity, Specificity, and F1-score), SR results are in marginal improvement by 1% and AlexNet achieves the best sensitivity and F1 rates among all DL schemes. Recall that sensitivity is the probability of a model predicting Malignant being truly Malignant. ResNet18 architecture performance across almost all metrics is stable using either of the resizing techniques.

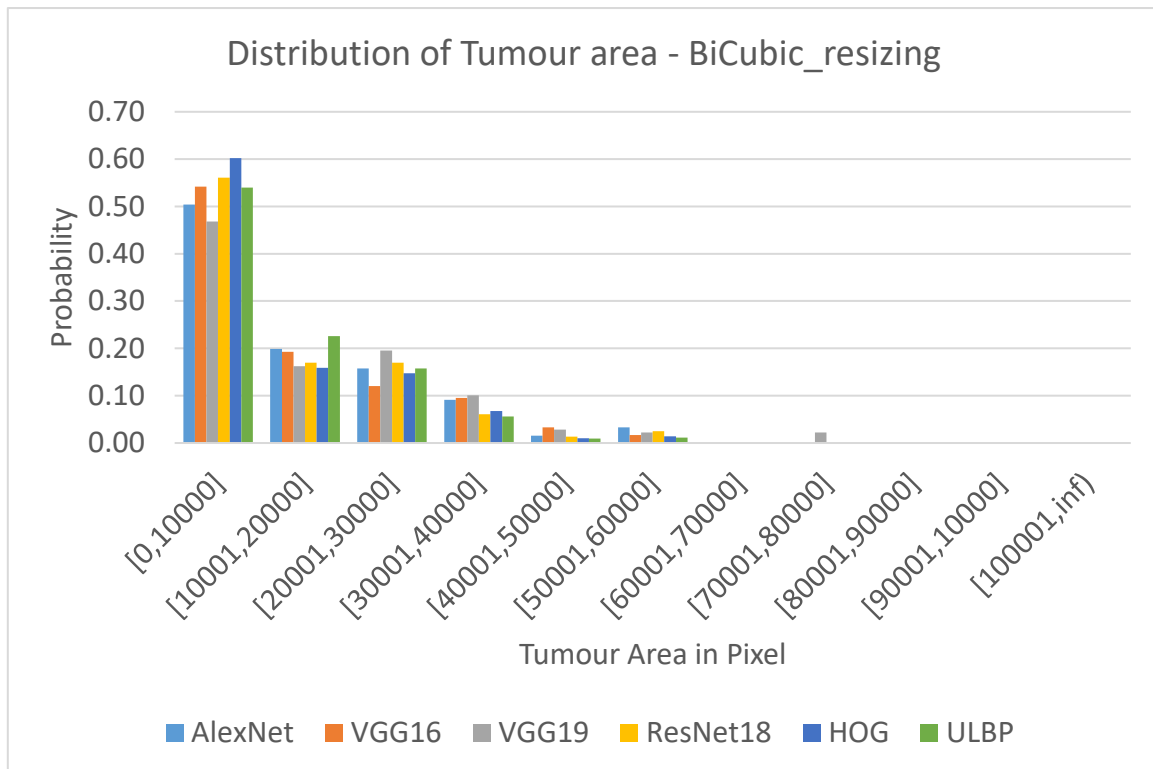
In contrast, when using SR resizing, the performance of the HOG scheme is degraded while the performance of ULBP is boosted. This may be explained by the fact that ULBP features are more linked to image texture landmarks than HOG, and SR is designed to maintain (or control degradation) the surroundings of these landmarks. Comparing sensitivity results reveals an interesting performance pattern for the two types of ML models. Except for the ResNet18, the SR resizing results in a marginal improvement (1%) for the other DL schemes, while for the HC feature scheme, the SR yields notable degradation (2%-4%). The picture is more mixed for specificity.

In summary, these experiments indicate that the performance of the various schemes is only boosted marginally by applying SR instead of BiCubic. However, a deeper analysis of these results may benefit from knowing (1) that SISR is primarily designed to upscale LR single images with improved/maintained quality and (2) in all the tumour datasets, most original RoIs are small-size images. For this, we looked at the statistics of the predicted decisions for each of the tested models and computed the distribution of the misclassified cases. Figures

3.16 and 3.17 below display the distribution of RoI tumour size of the misclassified cases for each of the above experiments.

As we explained earlier, the SR resizing technique is effective and improves the quality of the resized image when the actual image is of LR and degraded. Moreover, in Section 3.2.1, we showed that the majority of the tumours in the Renmin dataset are of LR in the range of [0, 10000] pixels. Therefore, the SR resizing technique impacts the quality of the resized versions of these images.

The probability distribution of tumour area-pixel of the misclassified cases, as shown in Figures 3.16 and 3.17 for BiCubic vs. SISR, is consistent with the marginal classification performance improvement reported in Tables 3.1 and 3.2. The figures show that by applying SR instead of BiCubic resizing, the probability of misclassifying a LR tumour image is getting lower marginally by approximately 10%. This entirely agrees with the impact of SR on LR images in comparison to HR good quality images. Moreover, the sensitivity of most of the DL schemes is improved with SR resizing, while clinically high sensitivity is more desirable compared to specificity due to the fact that misclassified malignant cases may increase fatality rate and could result in a higher cost for NHS due to more litigations.



**Figure 3.16 Distribution of tumour size for misclassified cases with BiCubic resizing.**



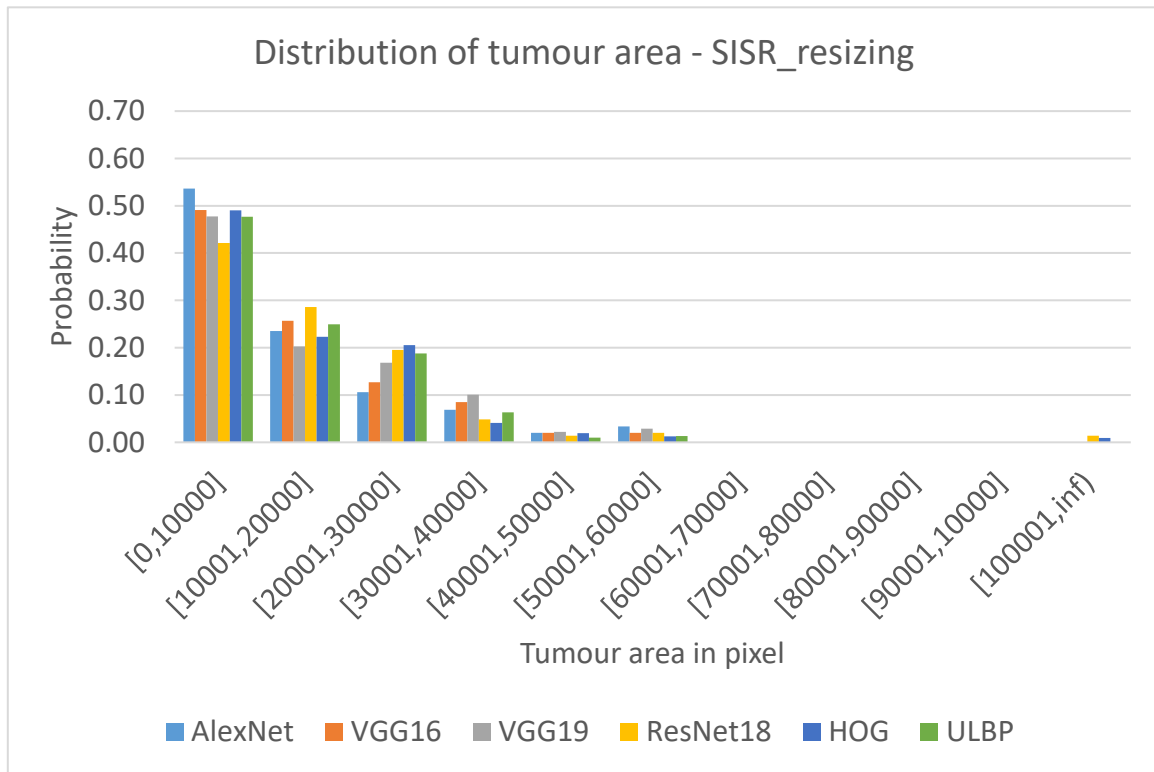


Figure 3.17 Distribution of tumour sizes for the misclassified cases with SISR resizing.

### 3.3 The Challenge of Ultrasound Image Quality Assessment

Manually analysing BUS images mainly relies on the operator's experience level. When US devices are deployed for tumour diagnosis, the operator (radiologist) holds the US probe with one hand while watching the monitor to establish a proper angle to scan the lesions and mark/measure the tumour dimensions. The probe needs to be held steadily in place long enough. Only well-trained operators can follow this procedure and determine the suitability of tumour-scanned images. This is a sensitive task, and even a minor handshake has a noticeable distorting impact on the scanned image and adversely influences the suitability for diagnostic predictions by ML schemes. Thus, it is safer to record a video of the organ/tissue that could be carefully examined to select the most suitable frame and mark the tumour border points afterwards. There are other factors influencing the suitability of US scanned images, including variation in acquisition procedures/devices as well as inter/intra observer/radiologist variation [65], [88]–[90].

Radiologist training worldwide aims to equip the participant with a globally standardised knowledge of how to detect RoI tumours using US scanning tools and how to assess image suitability. To develop automated ML schemes for US image analysis, it is very important to design reliable metrics to assess input US image suitability. In the literature, image quality

is used to reflect image suitability for input to ML analysis. This is sensible when dealing with natural images. If we adopt existing natural image quality metrics to assess US image suitability, we need to compare the sources of degrading US image contents with those of natural image contents. Natural images capture objects that the human brain is trained to recognise their geometric/structural characteristics that are invariant to size and orientation even if the image is reasonably distorted, blurred, with shadows and/or noise. These natural image-degrading effects are directly linked to the line of view of the light source, human recording skills, the sophistication of the deployed camera, and the distance at which the content is captured. US scanners use audio signals emitted on tissues within reasonable distances and generate images by the reflected signals. The structure of US scanned tissue is influenced by their dynamically changing complex environment as a result of blood (and other body liquids) flow and other factors. Note that digital image/video recording cameras in deep oceans are subject to analogous uncontrolled environments but to a rather less extent. Only well-trained radiologists may be able to recognise relevant US contents of interest (in such a dynamic environment) and distinguish image artefacts from tissue aberrations.

Notwithstanding radiologist skill requirements, the performance of ML models used to analyse US images is also influenced by natural image quality characteristics such as blurriness, shadows, poor contrast and noise. Therefore, it is sensible to consider using natural image quality metrics to assess the suitability of US tumour scan images. US images are known to be subject to a special type of noise known as speckle noise. Different US devices may cause different levels of Speckle noise.

Reference-based image quality metrics like Peak-Signal-to-Noise-Ratio (PSNR) [91], Structural Similarity Index Measure (SSIM) [92], and Universal Image Quality Index (UIQI) [93], [94] are widely used in many natural image processing/analysis applications. Given a reference image  $R$  and a target image  $T$ , both of size  $m \times n = N$ , these metrics are defined as follows:

$$PSNR(R, T) = 20 \log_{10} \left( \frac{255}{\sqrt{MSE(R, T)}} \right) \quad 3.2$$

Where the Mean Squared Error (MSE) is the average of the squared intensity differences of the  $R$  and  $T$  as:

$$MSE(R, T) = \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^n (R(i, j) - T(i, j))^2 \quad 3.3$$

SSIM is computed as follows:

$$SSIM(R, T) = L(R, T) * C(R, T) * S(R, T) \quad 3.4$$

Where

$$L(R, T) = \frac{2\mu_R\mu_T + c_1}{\mu_R^2 + \mu_T^2 + c_1} \quad 3.5$$

$$C(R, T) = \frac{2\sigma_R\sigma_T + c_2}{\sigma_R^2 + \sigma_T^2 + c_2} \quad 3.6$$

$$S(R, T) = \frac{\sigma_{RT} + c_3}{\sigma_R\sigma_T + c_3} \quad 3.7$$

For any two images,  $L(R, T)$  is a luminance comparison function that measures the proximity of mean luminance  $(\mu_R, \mu_T)$ ,  $C(R, T)$  is a contrast comparison function that measures the proximity of the standard deviations  $(\sigma_R, \sigma_T)$  of their intensity, while  $S(R, T)$  is the structure comparison function that measures the correlation coefficient between the two images, R and T. Note that  $\sigma_{RT}$  is the covariance between R and T. The positive values of the SSIM index are in [0,1]. A value of 0 means no correlation between images, and 1 means that R=T. The positive constants  $c_1$ ,  $c_2$  and  $c_3$  are used to avoid a null denominator.

PSNR and SSIM are widely employed metrics in the field of image and video processing to quantitatively evaluate the quality and similarity between original and processed ones. PSNR measures the fidelity of a reconstructed or compressed signal by computing the ratio of the maximum signal power to the mean squared error between the original and processed signals. Higher PSNR values indicate better fidelity. On the other hand, SSIM assesses perceptual quality by examining the signals' luminance, contrast, and structural information. It quantifies the similarity between signals based on their statistical properties, encompassing factors such as brightness, contrast, and structural similarity. Higher SSIM values correspond to greater perceptual similarity. Both metrics provide objective assessments of image quality, aiding in developing and optimising various image processing algorithms and techniques. They are also used to compare the effect of image transformation (such as compression), and we do not assume the presence of a transformation model that led to changing a healthy (or a benign mass) tissue into a cancerous one. We will not further discuss these metrics. Using such techniques to measure any quality distortion in an image requires a good-quality reference image.

The UIQI is another reference-based image-quality measure defined in terms of the statistical parameters of a reference image and a given image. The formula defines it:

$$UIQI(R, T) = \frac{4\sigma_{RT}\mu_R\mu_T}{(\sigma_R^2 + \sigma_T^2) * (\mu_R^2 + \mu_T^2)} \quad 3.8$$

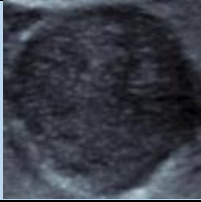
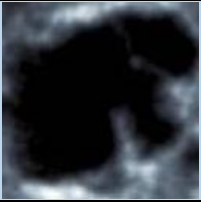
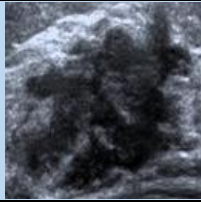
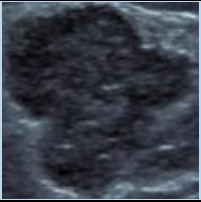
A simple manipulation and rearrangement of the above formula make UIQI closely similar to the SSIM.

$$UIQI(R, T) = \frac{\sigma_{RT}}{\sigma_R\sigma_T} * \frac{2\mu_R\mu_T}{(\mu_R^2 + \mu_T^2)} * \frac{2\sigma_R\sigma_T}{(\sigma_R^2 + \sigma_T^2)} \quad 3.9$$

The UIQI models the distortion between R and T as a product of three components: loss of correlation, luminance distortion and contrast distortion. These three factors of UIQI represent quality characterizing measures that reflect human vision system measures of distortions between R and T in terms of (1) loss of correlation, (2) luminance distortion, and (3) contrast distortion, respectively. The UIQI was modified by adding another factor called modified skewness to the other three components [94]. Table 4.3 presents the computed PSNR, SSIM, and UIQI with its factors for 4 BUS images, 2 Benign (B1, B2), and 2 Malignant (M1, M2) when each is used as a reference for the others. All images have the same size, having been resized.

The computed values for the 3 factors (Correlation, Luminance, Contrast) of UIQI for these Benign and Malignant images indicate that while no good closeness is detected between any two of the images in terms of the correlation factor, all have significant to excellent similarity with each other in terms of luminance and contrast in comparison. Yet these images are visibly distinct from each other besides being in different classes.

**Table 3.3 The computed PSNR, SSIM, UIQI and its three components for a selected 4 US images.**

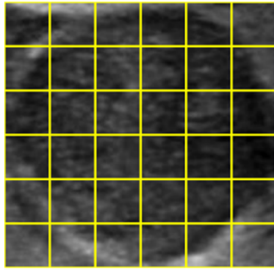
RoIs				
<b>PSNR</b>	B1	B2	M1	M2
B1	Inf	-34.38	-35.87	-30.58
B2	-34.38	Inf	-37.01	-35.00
M1	-35.87	-37.01	Inf	-35.66
M2	-30.58	-35.00	-35.66	Inf
<b>SSIM</b>	B1	B2	M1	M2
B1	1	0.02	-0.01	0.03
B2	0.02	1	0.05	-0.01
M1	-0.01	0.05	1	0.02
M2	0.03	-0.01	0.02	1
<b>Correlation</b>	B1	B2	M1	M2
B1	1	0.39	0.02	0.29
B2	0.39	1	0.3	0.26
M1	0.02	0.3	1	0.09
M2	0.29	0.26	0.09	1
<b>Luminance</b>	B1	B2	M1	M2
B1	1	0.98	0.95	0.99
B2	0.98	1	0.88	0.98
M1	0.95	0.88	1	0.95
M2	0.99	0.98	0.95	1
<b>Contrast</b>	B1	B2	M1	M2
B1	1	0.8	0.84	0.99
B2	0.8	1	0.99	0.82
M1	0.84	0.99	1	0.85
M2	0.99	0.82	0.85	1
<b>UIQI</b>	B1	B2	M1	M2
B1	1	0.31	0.02	0.28
B2	0.31	1	0.26	0.21
M1	0.02	0.26	1	0.07
M2	0.28	0.21	0.07	1

The existing reference-based quality measures are not practical in our domain as there is no standardized US image dataset of good-quality images to be used as a reference dataset. Therefore, it is difficult to define ground truth quality-labelling for this purpose. A close examination of each of these images reveals that the quality characteristics of Correlation, Luminance, and Contrast are not uniformly distributed across different areas of the same RoI. To illustrate this assertion, we split the tumour RoI for the 4 images in Table 3.3 and

divided each into equal-sized 36 (6x6) blocks. For each image, we calculate the 3 component values for cross-referencing each image block with the other 35 blocks. This results in  $630 = \frac{36 \times 35}{2}$  values for each UIQI component. We then quantize these values into 10 equal bins, as shown in Figure 3.18. For the sake of comparison with natural images, we used the same procedure on a tightly cropped passport standard face image.

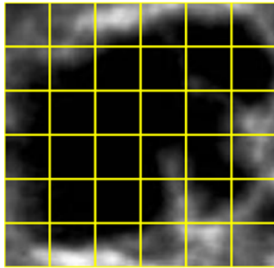
Figure 3.18 shows that the three quality components are not uniformly distributed across different areas of the same image, which is true for different tumour classes and the natural face image. For the US images, most of the cross-block correlation indices fall within the middle range bins (4-7), which is also observed for the face image. However, for both luminance and contrast, most of the cross-block indices are in the upper range of bins (8-10) for both image modalities. Comparing the US to the Natural/face image, we find that the top cross-block illumination values are 91.75% for the natural image against 43.17% for the US images. The face image is well-lit, and the other 8.2% of the cross-block bins (8 and 9) are due to the intensity differences between the eyes and the remaining facial features. However, the top contrast scores are 34.92% for the face image against 62.54% for the US images. The face image is relatively smooth away from the eye's region, and this explains why 22.86% of the cross-block contrast scores are in the low half of the bins compared to 0% for the US images.

Besides highlighting the differences between natural and US images in terms of the three quality characteristics, the above inspiring observations provide the ingredient for a no-reference IQA in US image analysis by exploiting the spatial distribution of various quality characteristics. The intended technique expands the list of 3 UIQI factors with other statistical quality measures to form a *self-reference quality feature vector* that we developed and investigated during our work for this thesis project and is presented in Chapter 5.



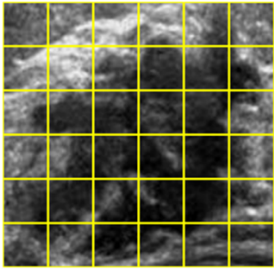
B1

Bins	1	2	3	4	5	6	7	8	9	10
Correlation	0	5	25	104	199	170	90	31	6	0
Luminance	0	0	0	2	6	12	28	51	112	419
Contrast	0	0	0	5	22	39	52	79	121	312



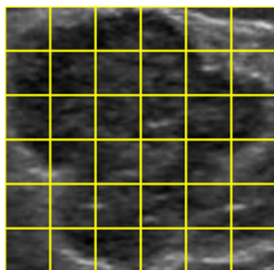
B2

Bins	1	2	3	4	5	6	7	8	9	10
Correlation	15	27	53	57	136	124	71	36	33	9
Luminance	181	59	49	29	23	28	36	32	45	148
Contrast	192	20	24	24	24	34	41	55	61	154



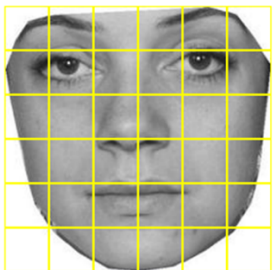
M1

Bins	1	2	3	4	5	6	7	8	9	10
Correlation	0	4	42	123	156	160	90	42	13	0
Luminance	0	0	5	14	40	37	70	80	112	272
Contrast	0	0	0	0	0	5	28	78	125	394



M2

Bins	1	2	3	4	5	6	7	8	9	10
Correlation	0	12	34	98	182	153	96	49	6	0
Luminance	0	0	0	0	0	1	7	27	119	476
Contrast	0	0	0	0	4	25	58	84	125	334



Face  
Image

Bins	1	2	3	4	5	6	7	8	9	10
Correlation	0	16	45	97	153	177	83	47	12	0
Luminance	0	0	0	0	0	0	0	6	46	578
Contrast	30	5	37	38	34	48	67	69	82	220

Figure 3.18 Display of spatial distribution of (Correlation, Luminance, and Contrast) for tumour RoIs and a face image.

Finally, we note that the various existing image quality metrics are certainly influenced by the type of noise present in images. A plethora of procedures have been developed with satisfactory/significant success for natural image denoising, depending on the noise type. The most successful approach to deal with image degradations (including noise) deals with these problems as inverse problems, similar to the discussion we had in the last section on SR. To deal with speckle noise, it has been shown that speckle noise on US images produces a different noticeable effect on different regions depending on the solidity of the regional tissue [6]. The author associated the effect level with regional intensity Skewness and Kurtosis parameters, developed an adaptive US denoising scheme, and demonstrated a significant performance improvement of their ML models applied to Ovarian cancer [6].

### **3.4 Lack of Training Samples - Solutions**

Another important performance influencing factor in DL-based US image analysis that is equally applicable to other medical image modalities is the non-availability of a sufficiently large training dataset of well-annotated and good-quality images. The availability of such a dataset guarantees an effective learning process which is crucial in training any DL architecture [36], [95]. However, designing a DL model for US tumour image analysis is a very challenging task due to the lack of publicly available large training datasets [35], [65], [66], [96], and some of the publicly available datasets like BUSI is of low quality and some of the images in the dataset have severe artefacts and annotations [61]. Unlike medical image domains, e.g., natural image analysis, one can find very large datasets such as ImageNet for designing efficient and optimal performing DL models [7], [54].

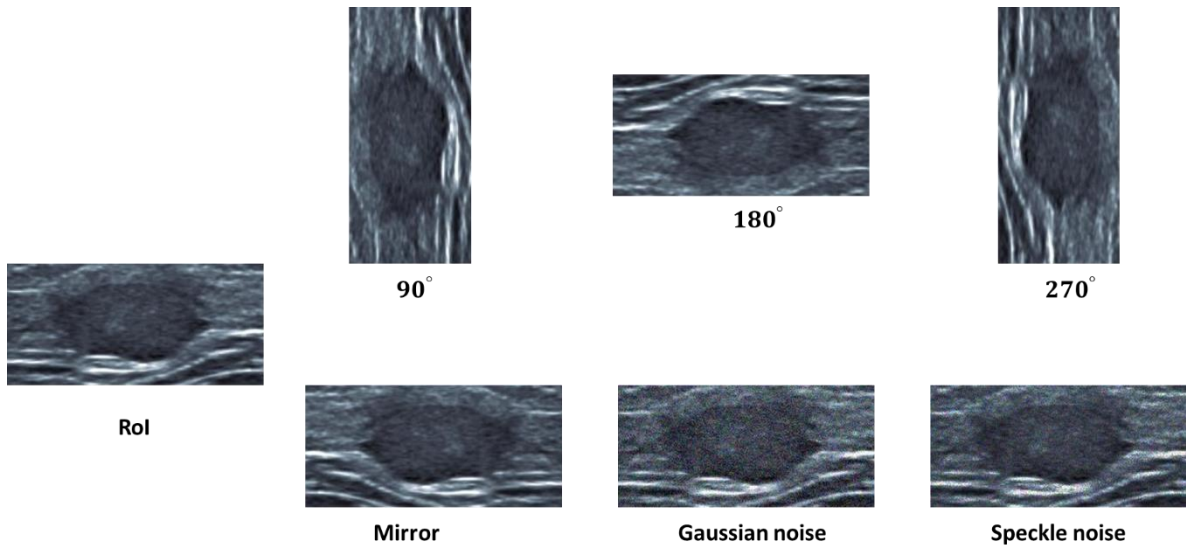
Any designed CNN architecture from scratch with the initialized parameters would not deliver a good classification performance for any given US image dataset. For the Training from Scratch-scenario, the model needs a lot of training with a large enough training dataset to efficiently learn class-discriminating from deeply encoded feature maps. Establishing such a dataset for a specific form of cancer is a lengthy, expensive, and complicated procedure that necessitates several connections and collaborations among radiologists and ML experts. Moreover, training a CNN model from scratch is computationally burdensome compared to other training protocols and HC feature methods [97]. Therefore, training CNN models with our relatively small BUS datasets is not ideal and may lead to model overfitting and biases. Although medical images, especially US images, differ significantly from natural images, some learned knowledge from natural images is transferable to the medical domain [97], [98].



Alternatively, Transfer Learning is frequently employed to address the issue of having a small non-representative training/modelling dataset [65]–[67], [89]. In this approach, instead of training from scratch, one of the state-of-the-art pre-trained optimally performing CNN models is used for transferring the learnt parameters when trained on the very large ImageNet dataset. In this approach, the convolutional layers of the pre-trained model are used for feature map extraction from the US dataset to be used to update/train the parameters of the FCLs for classification. Usually, only the parameters of the last FCL are updated [28].

Fine-tuning is a slightly different approach compared to other versions of transfer learning. In this approach, again, one of the pre-trained models is adopted; however, the parameters of all the layers of this model are updated during the added training on a subset of the intended US image dataset [28].

In general, training a CNN model from scratch is suitable when the desired dataset is sufficiently large to train such a model, and at the same time, enough computational power is committed for the training. On the other hand, transfer learning is proved to be effective in the case of having a small non-representative dataset for modelling. However, its best performance compared to the optimal performance of the pre-trained model is suboptimal. Fine-tuning is more suitable in the case of having a relatively small non-representative dataset that is quite different from the pre-trained dataset (Medical US compared to Natural Images). Therefore, Fine-tuning is the adopted CNN training method throughout our work. Researchers have developed image processing techniques to expand a small dataset of images by adding their processed versions to create a larger dataset. The main challenge is to select several image transforms that output images of the same class as the input images. Image Augmentation is a commonly used approach to address the issue of data scarcity. It is the process of generating new samples from the available training images using various known image operations and algorithms [9], [99], [100]. The classical image augmentation techniques in the medical field include simple image manipulations like rotation, horizontal and vertical flip, scaling, zooming, translating, shearing, blurring, sharpening, contrast, brightness, and noise insertion [96], [100], [101]. Figure 3.19 illustrate some image augmentation process.



**Figure 3.19** US image augmentation using rotation, mirror, and noise insertion.

In recent years, artificial image generation has been proposed to expand small image datasets into high-volume image datasets that are meant to have the same characteristics as the small dataset. DL approaches such as Generative Adversarial Networks (GAN)s are successfully used for image augmentation [102]. GANs are utilized to create synthetic US images, which help to improve the classification performance of BUS lesions [89], [103]. However, there are concerns regarding its suitability in medical image analysis, as synthetic image augmentation may further overfit the trained model and expose it to adversarial attacks [104]. The reviews conducted for the various components of this thesis project have revealed several potential image representations or RoI appending techniques. In this work, we design several novel techniques for BUS image augmentation that rely on limited available samples to create real versions of the images, similar to how clinical specialists analyse US images in their diagnostic assessments. In Chapter 6, we design three approaches for BUS image augmentation: one is based on Singular Value Decomposition (SVD) [105], the second one is based on image convolution with Hadamard-based filters [76], and the third one is specific to BUS images that uses RoI *Tumour Margin Appending* (TMA) [9].

### 3.5 Conclusion

In this chapter, we have delved into CNN performance influencing factors that arise from their architectural requirements when applied to US tumour scan image analysis. Despite the advantages that US imaging offers compared to other medical image modalities, we have uncovered significant challenges in designing DL-based US image analysis systems. These challenges encompass the variation in RoI sizes, the influence of clinician experience, the

impact of different US machines and clinical practices, the presence of inter- and intra-observer variability, the complexities of RoI cropping scenarios, and various image quality issues such as speckle noise and low contrast.

We have established that the extent of RoI size variation within our BUS datasets is a tough challenge for meeting the DL model's requirement on fixed-size input images without influencing input image quality. We demonstrated that this variation is significant and relatively dependent on the lesion class. Specifically, we observed that most benign RoIs corresponded to small-size tumours, in contrast to malignant ones. Furthermore, we have shown that the resizing procedures employed in RoI preparation have a more noticeable effect on the quality of benign cases than malignant ones.

Another significant challenge we have identified is the perceived *low quality* of US images. Unfortunately, robust and standardized IQA tools are currently unavailable for evaluating US images before inputting them into DL systems. This dearth of assessment tools hampers developing and deploying effective DL schemes in US imaging.

Moreover, the scarcity of adequately labelled training image samples remains a primary concern in the literature on DL for US image analysis. Acquiring a diverse and well-annotated dataset is a labour-intensive task, limiting the generalization and robustness of DL models. To mitigate this issue, it is imperative to develop specific and effective image augmentation techniques tailored to the medical US to expand the datasets.

In the next chapter, we will investigate the optimal cropping scenario for BUS images and explore its impact on the classification performance of DL models and HC feature schemes. We aim to further enhance our understanding of the intricacies involved in optimizing image cropping techniques for improved DL-based breast tumour classification.

## Chapter 4: Lesion Shape Cropping from US Images

The clinical procedure for using medical imaging (in this case, Ultrasound) in assisting tumour diagnoses starts by recording an US video scan of the relevant tissue/organ, and an experienced radiologist assesses the various frames of the video, identifies the frame that best contains the tumour, annotates that frame with some information/parameters related to the lesion, its boundary, and other patient-related information. Oncology clinicians, who have been trained for several years, examine the patient's US tumour selected frame with a focus on the lesion region, being the source of distinguishing features between malignant and benign masses, report their diagnoses of the case and advise the patient on the course of action. This time-consuming critical task places a high burden on health services/centres and requires highly skilled clinicians. In order to develop a ML model that can be used to support clinical teams, ideally, it is necessary to carry out all the above preparation steps, including the selection of the best US frame, segmenting the lesion and making predictions that can be examined together with results of other related medical tests. Segmentation of the lesion is either done manually by the radiologist or automatically by a tumour segmentation algorithm, and it amounts to cropping the corresponding tumour RoI. Manual tumour segmentation is time-consuming, and automatic segmentation is a challenge that we keep outside the focus of this project. To compensate for the absence of reliable segmentation, TenD radiologists provided information on the suspect masses' boundaries by marking a set of lesion boundary points sufficient to determine the location and shape of the tumour.

In this chapter, we shall consider several strategies for tumour RoI cropping using the marked lesion border points without tumour segmentation and investigate their impact on the model classification performance for DL and HC feature schemes. In section 4.1, We describe the concept of lesion cropping from a computational viewpoint and conduct a literature review of related work. Section 4.2 describes potential cropping strategies, while section 4.3 is concerned with experimental work on the performance of various DL/HC models corresponding to our chosen lesion cropping strategies. Section 4.4 is concerned with the generalisation of the developed DL models when tested on external datasets in relation to the proposed cropping strategies. Finally, in section 4.5, we use heatmap visualisation to understand the performance testing results and the impact of the tumour cropping scenarios on DL decision quality.

## 4.1 Introduction and Related Work

For computational purposes, cropping of a lesion in an US tissue/organ scan image is meant to determine its set of disease-relevant pixels. In general, lesion segmentation (automatic/manual) meets this requirement, but it is a tough challenge due to many factors, including tumour size variation (see section 3.2.1) and the fact that tumour cells do not grow uniformly in all directions, resulting in various border irregularities. Cropping strategies are expected to be inspired by the way clinicians analyse tumour US images. Experienced clinicians make their image-based diagnostic predictions by considering image features within the lesion area, including the border as well as the surrounding region, that is deemed to convey important disease-related information. However, image features extracted by traditional CAD systems often give little or no consideration to the surrounding area, and yet cropping strategies depend on the nature of information extracted from the tumour and surrounding region of the RoI.

During clinical tumour analysis, radiology experts complement known disease-relevant information (e.g., age, patient's medical history, genetic profile, and results of disease biomarkers tests) by assessing image features/information within the tumour tissue region, its border, and the immediate surrounding region. The various image regions are used to assess medically known malignancy predictors, usually referred to as signs of tumours (Cancer Signs). The lesion interior image texture information helps identify the tumour's internal *echogenicity level* and *solidity*, while the border encapsulates information about malignancy predictors such as tumour *shape* and *irregularity* levels. Image information from the periphery of the tumour primarily maintains posterior acoustic echo and lateral acoustic shadow [106]. In contrast, most existing automatic US image analysis (HC features and DL) algorithms analyse features extracted/learnt from the tumour area within the smallest bounding box with little or no consideration to the surrounding region. We shall investigate the pros and cons of mimicking the clinical approach by using features from the lesion surrounding area. Besides the challenge of determining the lesion's internal region and the border with high accuracy, we need to determine the appropriate margin of external tissue to be appended.

Very few articles in the literature address the subject of optimal RoI cropping and identifying suitable ratios for tumour margin appending. Cao et al. [65] investigated two case scenarios by feeding the whole B-mode BUS image to DL architectures vs. utilizing only the smallest RoI bounding box of the tumour area. Testing these two cases does not necessarily yield the

best cropping ratio since the entire image has too much background information, whereas the tumour RoI is more focused on the lesion site. On the other hand, some studies show that feeding an RoI to DL models with too much information from the background lowers model decision quality. This is attributed to the irrelevance of some elements of the overall image to the examined disease while contributing to the model decision [107].

Han et al. [108] studied optimal tumour margin appending for BUS tumour classification with DL, where the tumour margin is defined as the distance in pixels between the lesion boundary and the cropped rectangle bounding box. They assessed the classification performance as the margin increased from 0 to 240 pixels at different thresholds and discovered that a tumour margin of 180 pixels provides the highest classification accuracy. However, fixing the tumour margin ratio at a specific threshold for all the tumours is not ideal, especially considering the tumour size variation found in our breast datasets. For example, appending a tumour of size 20x20 with 180 pixels from the tumour periphery will add a significant amount of information from the surrounding region compared to the amount of image information present in a 20x20 pixel tumour RoI box. Therefore, it is better to append the tumour's margin with a threshold related to the tumour's actual size, and this is the approach we follow in our research. Moreover, appending tumour lesions of highly irregular border shapes may result in self-intersection [34].

Yamakawa et al. [106] investigated the optimal cropping scenario for liver tumour US image classification using a newly developed CNN architecture. They set the tumour in the centre of a square bounding box of side length  $L$ . Although the included figures in the paper display fitted ellipses, the text describes the use of a fitting circle of the tumour to determine its centre and the maximum diameter  $D$ . They determined the ideal cropping scenario by including the tumour and a portion of the surrounding region in the cropped RoI bounding box. Several ratio values have been tested, including ( $ratio = 0.1, 0.2, 0.3, \dots, 1.0, 1.1$ ) to crop the tumour with different margin appending ratios, and the classification performance has been tested at each threshold. The classification testing results demonstrate that  $ratio = 0.6$  is ideal and delivers the highest accuracy for US liver tumour classification. For our dataset of US tumour scan images, neither an ellipse nor a circle is a fair representation of tumour shapes, particularly for irregular tumours.

It is important to determine an optimal tumour margin appending ratio(s) for better lesion classification performance. In this work, besides the conventional way of tumour RoI cropping (i.e., using the smallest bounding rectangle box of the tumour polygonal shape), we study many scenarios for cropping RoIs using surrounding tissue regions. In the next

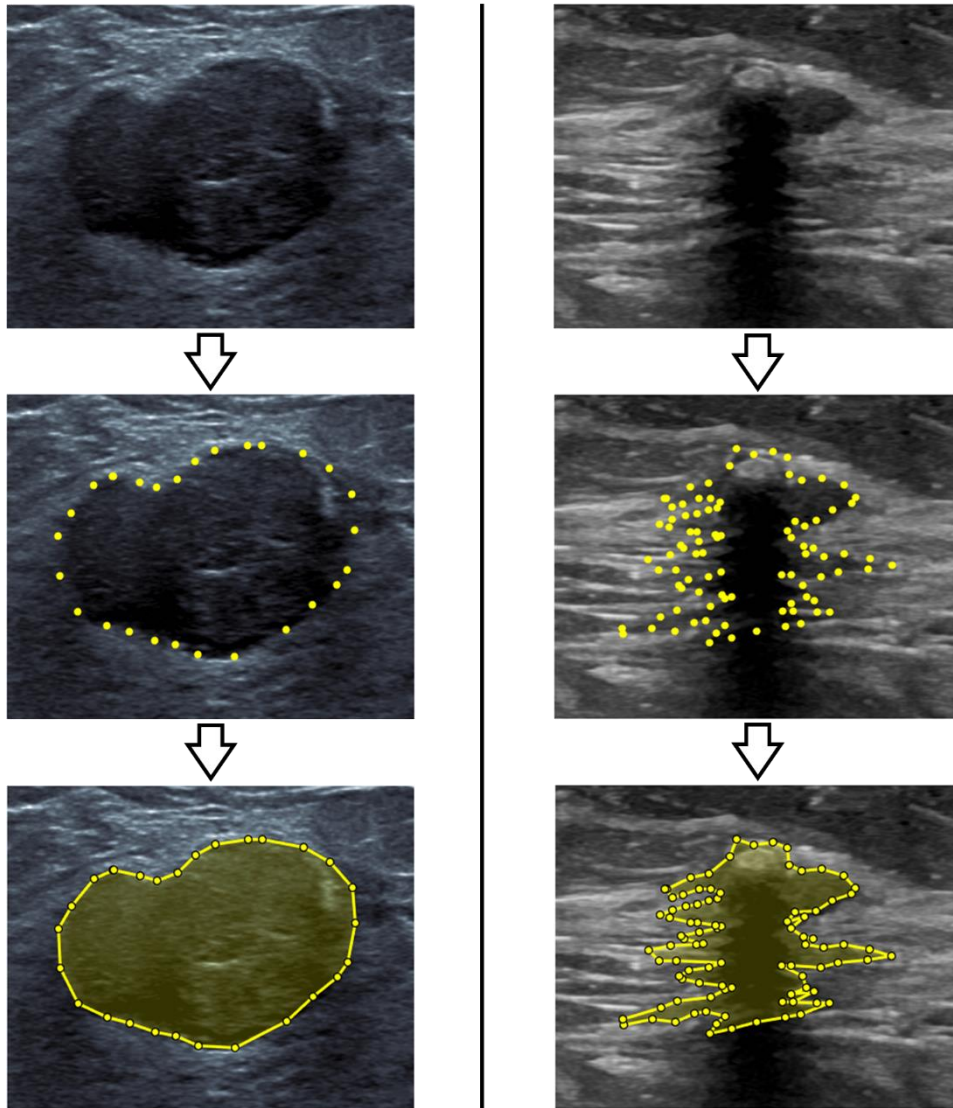
section, we shall describe various mathematical cropping strategies to propose a computationally efficient approach to implement, provide a reasonable approximation of the tumour, and is suitable for scaling/margin appending.

## 4.2 Cropping Models

Given a set of marked points,  $S$ , on the boundary of a lesion scanned by an US device, cropping of the lesion can be made depending on how we use  $S$  to determine the lesion border mathematically. Common mathematical approaches include (1) interpolating the lesion border from the set  $S$ , and (2) Curve fitting strategy, whereby one determines the border as being the best-fitted curve of the set  $S$ . Here, we shall propose an alternative simple cropping strategy by using the *Convex Hull* (CH) of set  $S$ , which includes 3 or more of the points in  $S$ . All these approaches may include external tumour tissues and/or exclude tumour tissues either due to computational errors or border point marking errors.

### 4.2.1 Lesion Cropping by Interpolations

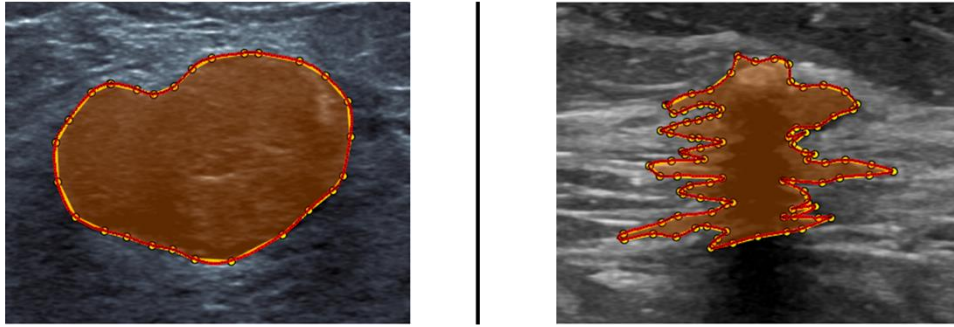
The most naïve, but easy to use, interpolation-based cropping strategy would be linear interpolation which outputs a polygonal shape whose corners are the points in  $S$ . This method would be very accurate when the tumour shape is regular. However, suppose the tumour shape is highly irregular. In that case, the tumour's polygonal shape misses lesion information from areas protruding away from its sides and/or includes non-lesion information protruding inward. Moreover, irregular lesions may present a challenge to the idea of expanding by margin appending as some parts of the circumference self-intersect after expansion resulting in multiple counting of RoI parts. Figure 4.1 below displays regular as well as irregular breast tumours with their corresponding lesion boundary points and tumour polygonal shape areas.



**Figure 4.1 Linear interpolation-based Polygonal lesion shape for (Regular vs. Irregular) lesion border.**

Non-linear interpolation of the border points provides a computationally more demanding alternative border approximation method. This approach includes using quadratic or cubic-spline curves that pass through the points of  $S$ . Generally, non-linear interpolating lesion border models the border section that passes through two or more points in  $S$  by a known polynomial equation(s) in the  $(x,y)$  values. Higher degree interpolating polynomials require more computational power but produce more winding border curves and thus are more useful for irregular lesions. For simplicity, polynomials of the same degrees are used to model all sections of the border curve, and cubic spline is the most common method. Figure 4.2, below, displays the cubic spline interpolated border of the above lesions in red colour on the tumour polygonal shape areas using linear interpolation in yellow.



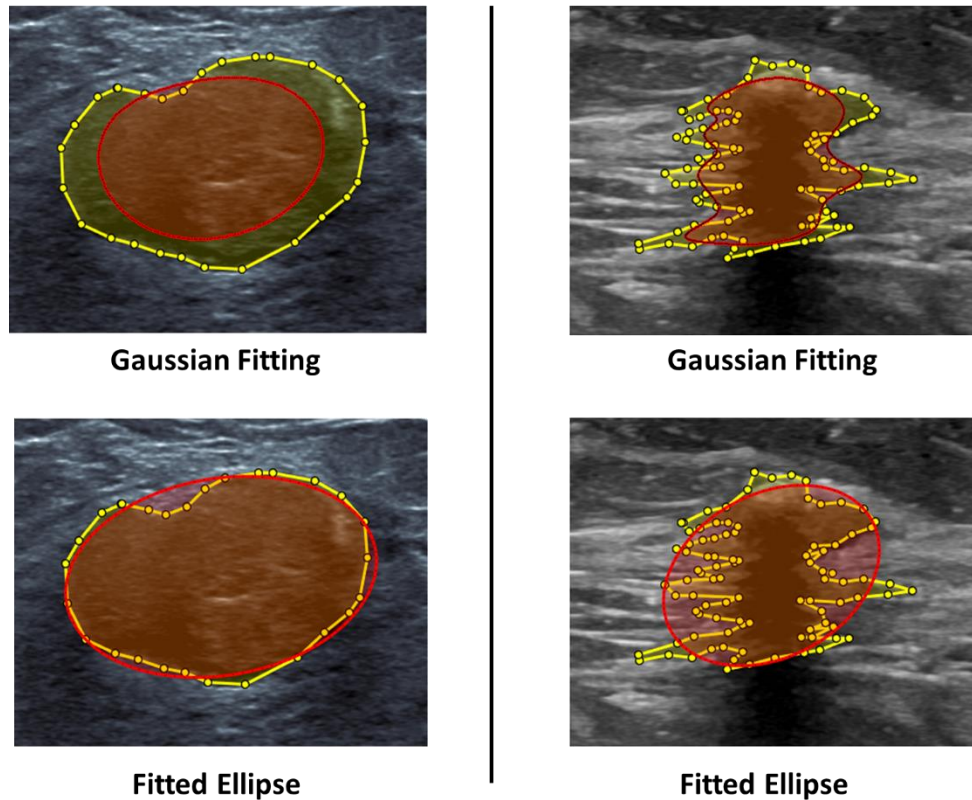


**Figure 4.2 Cubic-spline interpolation-based Polygonal lesion shape, in red colour, for (Regular vs. Irregular) lesion border.**

The assumption that different border sections have the same shape may not reflect the way tumour tissue cells grow. The curve fitting approach is a commonly used approach for approximating lesion border from the set  $S$  of radiologist-marked border points.

#### **4.2.2 Lesion Cropping by Curve Fitting**

Curve fitting strategies may seem similar to nonlinear interpolation, but the fitted curve may not pass through each of the points of  $S$ , if any. In a 2D US tumour scan, the lesion border forms a closed curve and therefore, the most commonly used curves are the closed conic sections, i.e., circles and ellipses. The fitted ellipse is widely used because most tumour tissue shapes resemble an elongated closed curve, perhaps with some irregularities and bends. The circularity measure of the actual border curve is often used as a useful indicator in the classification. However, closed piecewise curves that consist of several curves, such as parabolas or Gaussian curves, connect at their ends. Fagher Mohammed, in his PhD thesis [34], implemented several non-linear thyroid lesion border interpolations and border approximation by curve fitting (e.g. Cubic-Spline interpolations, Ellipse and Gaussian fitting approximation). Figure 4.3 below displays the Gaussian curve fitting and the fitted ellipse borders for the above breast lesions.



**Figure 4.3 Fitted Gaussian curve and fitted ellipse, in red colour, for Regular vs. Irregular lesion border.**

We observe that the Gaussian/Ellipse curves, being of regular geometry, do not provide good lesion border approximation when used for highly irregular lesion borders. However, these fitted curves have been deployed as references to measure the level of irregularity of tumour lesion border by analysing distance functions defined between the interpolated polygon points and the fitted curve [34].

### **4.2.3 Lesion Cropping by Convex Hull**

Following on from the above discussion, we conclude that neither curve fitting nor interpolation approaches can approximate the border without the possibility of including/excluding image information outside/inside the actual lesion. These undesirable effects may be due to significant variations of lesion border clarity and difficulty in choosing interpolating/fitting curves; hence, the ideal lesion cropping shape should be (1) minimising exclusion of lesion pixels, (2) maximising the inclusion of actual internal and border lesion pixels, (3) efficiency of computing the chosen shape, and (4) the ease with which the lesion margin can be expanded to facilitate margin feature extraction/learning.

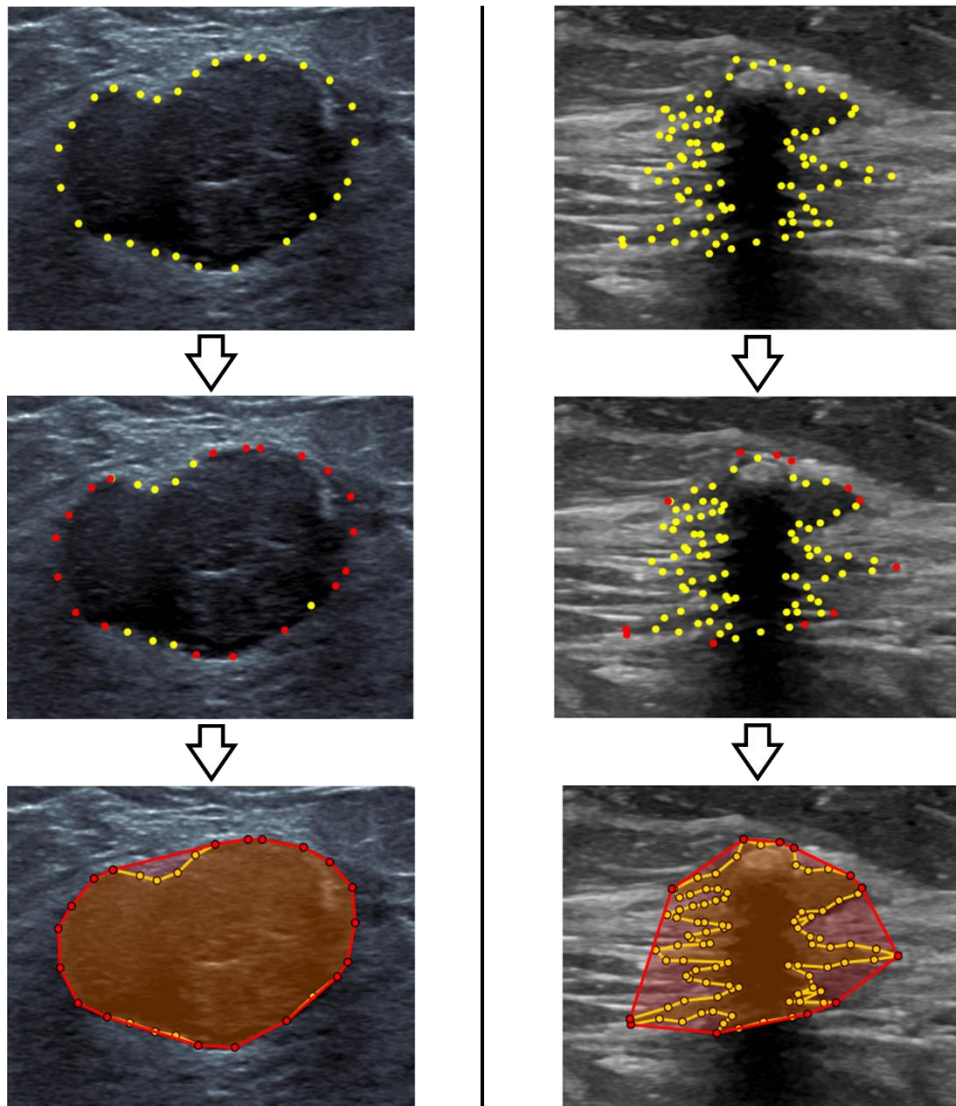
The Convex Hull (CH) of the set  $S$  of radiologist-marked lesion border points,  $CH(S)$ , is a polygonal shape with corners in  $S$  and scores well on the above 4 requirements. It covers the entire polygonal linear-interpolated tumour tissue and hence can only exclude parts of the lesion that protrude far away from the polygonal lesion border. Although it may include pixels outside the lesion tissue, the number of these pixels can be reduced by replacing all its internal pixels that are outside the polygonal lesion border with zero. Neither the CH nor the tumour polygonal shape is perfect in terms of the ideal cropping inclusion/exclusion conditions mentioned above. However, the CH shape is efficient to compute and is easy to expand by parallel translation of its sides, thereby facilitating the proportional expansion of the lesion border points. Moreover, the surrounding lesion box can be easily determined as its sides are determined by the 2 pairs of corners of  $CH(S)$  furthest away horizontally/vertically. Accordingly, we adopt  $CH(S)$  as a sensible lesion-cropping shape.

The standard RoI cropping procedure is to draw the smallest-fitted rectangular box that encloses the tumour polygon area after connecting the lesion boundary points in a polygonal form established in the order of the marked points, i.e., the linear interpolated lesion boundary. The pixel values inside the RoI rectangular box but outside the tumour polygon area are set to their original pixel values (i.e., tightly cropped tumour area with tissue padding). In this case, the classification decision is based on information mainly from the tumour area with little/no consideration of the surrounding region.

The whole B-mode image contains significant redundancy, including too much background and US annotation artefacts that may lead to poor CNN decisions. Also, the idea of margin appending the tumour using a fixed ratio without considering the lesion size, especially for small-size tumours, results in having some RoIs with too much background compared to large-size tumours. Furthermore, approximating the shape of the tumour by a fitted circle/ellipse is not quite accurate, especially when the tumour is very irregular.

The CH of a set of points in a 2D space is a geometric object/shape made up of a unique polygon linking the fewest possible points that one may go through without leaving the CH area. CH is a convex polygon with a maximum area and minimal circumference encompassing all the given points. It has a wide range of applications, including image/object detection in pattern recognition applications, e.g., see [109]–[111]. Figure 4.4 illustrates the CH of a set of lesion boundary points for the two BUS scan images. It shows that  $CH(S)$  of the lesion boundary points  $S$  covers the whole tumour area while it is a good convex

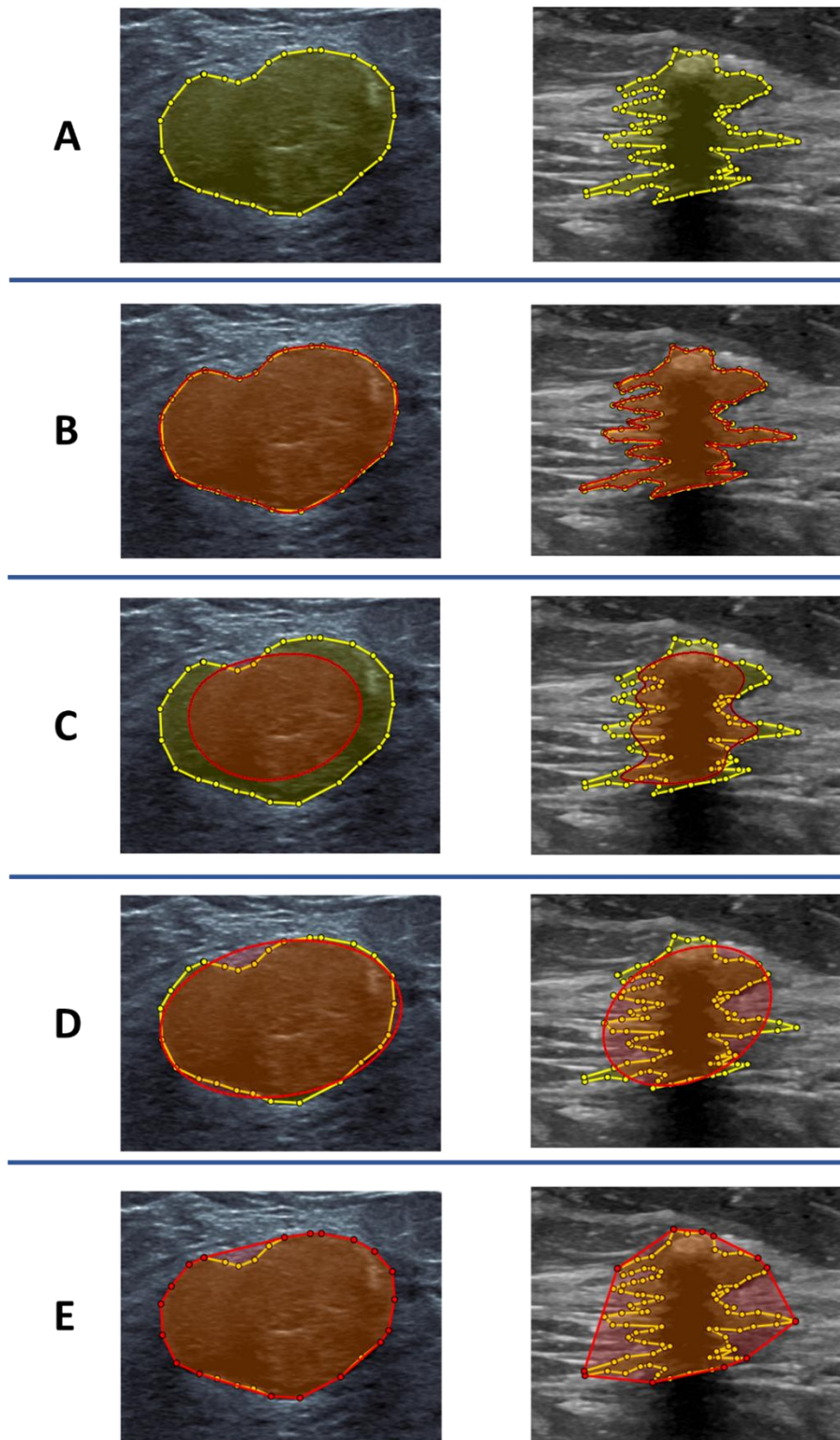
approximation of the actual shape and is suitable for expansion/margin appending. The red dots in the second row are the vertices of the determined CH.



**Figure 4.4 Steps of forming the CH polygon of a set of lesion boundary points.**

Figure 4.5, below, further illustrates the different tumour border approximations, including (A) tumour area polygon, (B) cubic spline border interpolation, (C) Gaussian curve fitting, (D) fitted ellipse, and (E) CH lesion border for the two lesions (regular vs. irregular). We note that the fitted Gaussian curve and ellipse succeed in approximating the shape of the lesions, but they most likely exclude some parts of the lesion area while the CH covers all of it. Furthermore, expanding the actual tumour polygonal area may result in self-intersecting RoI for all the schemes, while the corresponding CH solves this issue. Also, obtaining a polygonal shape inside any ratio expansion of the CH is not difficult. After expanding the CH by any ratio, it is easy to draw a polygonal shape inside the CH that is approximately

proportional to the exact tumour area by the same ratio. In this way, we can reduce the effect of the intersecting RoI problem.



**Figure 4.5** Tumour shape approximation of 2 lesions: (A) tumour area polygon, (B) cubic spline border interpolation, (C) Gaussian curve fitting, (D) fitted ellipse, and (E) CH lesion border.

Based on the above observations regarding the pros and cons of the above cropping strategies, we decided to focus on the CH scheme as our adopted strategy.

#### 4.2.4 Lesion Margin-appending Scenarios within the Convex Hull Shape

Whatever method is used to approximate/segment lesion borders in US tumour scanning images, we face the question of what values we assign to the pixels outside the approximated lesion border and inside the surrounding rectangular box. Moreover, the question extends to the region between the CH and the lesion-polygon border for our CH cropping strategy (see Figure 4.5 (E)). For the various cases, the set of these external pixels to the lesion but inside the CH is set to their original tissue pixel values, and the question becomes one about margin-appending methods. Traditionally, the entire set of pixels in the lesion surrounding area but inside the smallest rectangular bounding box is assigned to their original tissue image values. We shall call this *margin-tissue-padding* scenario, which is referred to as the *TumourT* scenario in experimental work. However, for some irregular lesion shapes, the margin is significantly larger than the actual lesion. To avoid this undesired situation, we suggest that pixels of the entire/subset lesion margin be assigned 0 value, and we call this a *margin-0-padding* scenario. In particular, we only consider 2 different 0-padding scenarios: (1) assign 0 to the entire tumour margin pixels to be referred to as the *TumourZ* scenario, and (2) assign 0 to the pixels outside the CH(S) but inside the surrounding box to be referred to as the *CHZ* scenario.

We note that clinicians examining tumour images may give more consideration to certain parts of the lesion border surrounding rather than the entire margin (e.g., regions relating to posterior acoustic echo and/or lateral acoustic shadow). This is an ideal 0-padding scenario, but we shall not adopt this scenario due to the unavailability of clinician advice, we do not have ground truth for such type RoI cropping. Instead, we shall extend our experiment by adopting the approach taken by Yamakawa et al. [106] and test the impact of different tumour margin appending ratios in the cropped RoIs on model performance. For this, we simply scale up/down the CH(S), with respect to its centroid, by a specified ratio  $s$ , to append the tumour area with different amounts of surrounding tissue. A new smallest bounding box is determined simply by the pairs of CH(S) nodes that are furthest away horizontally and vertically, respectively. Again, we have two possible ways of margin appending of the new scaled box: (1) original image pixels inside the  $s$  scaled box are assigned to their tissue values, and (2) original image pixels inside the  $s$  scaled box are assigned to 0. We denote the first padding scenario as *T* and the second padding scenario as *Z*. We selected the scaling

ratios from the set  $\{0.6, 0.8, 1.2, 1.4, 1.6, 1.8, 2.0, 2.5, 3.0, 3.5, 4.0\}$ . Figures 4.6 and 4.7 illustrate all the above-described tumour cropping scenarios, first using tissue padding followed by 0-padding. For consistency, we also include  $TumourT = CHT$ , while  $TumourZ \subseteq CHZ$ . The cropped RoIs are resized to  $227 \times 227$ , which is the AlexNet input size.

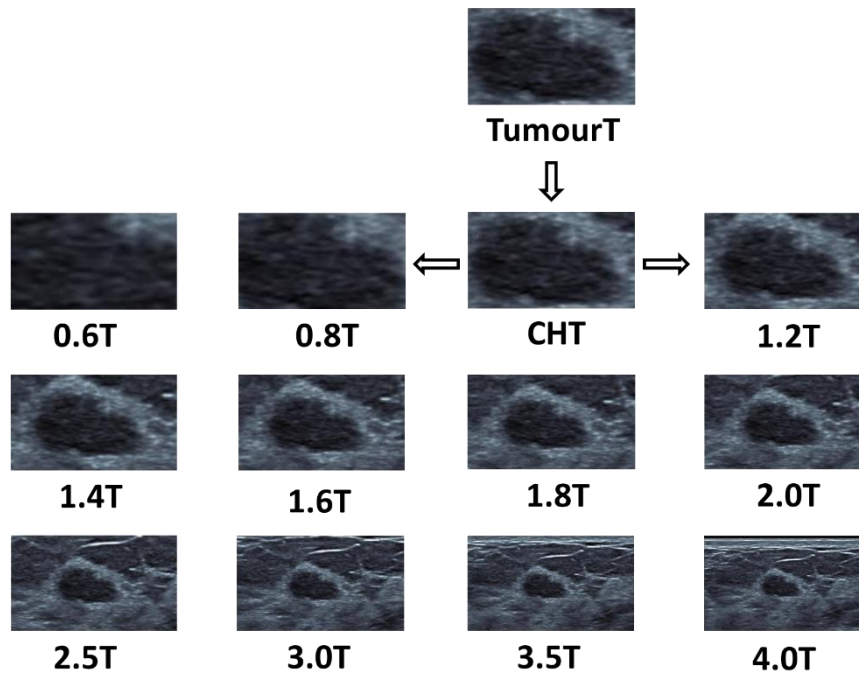
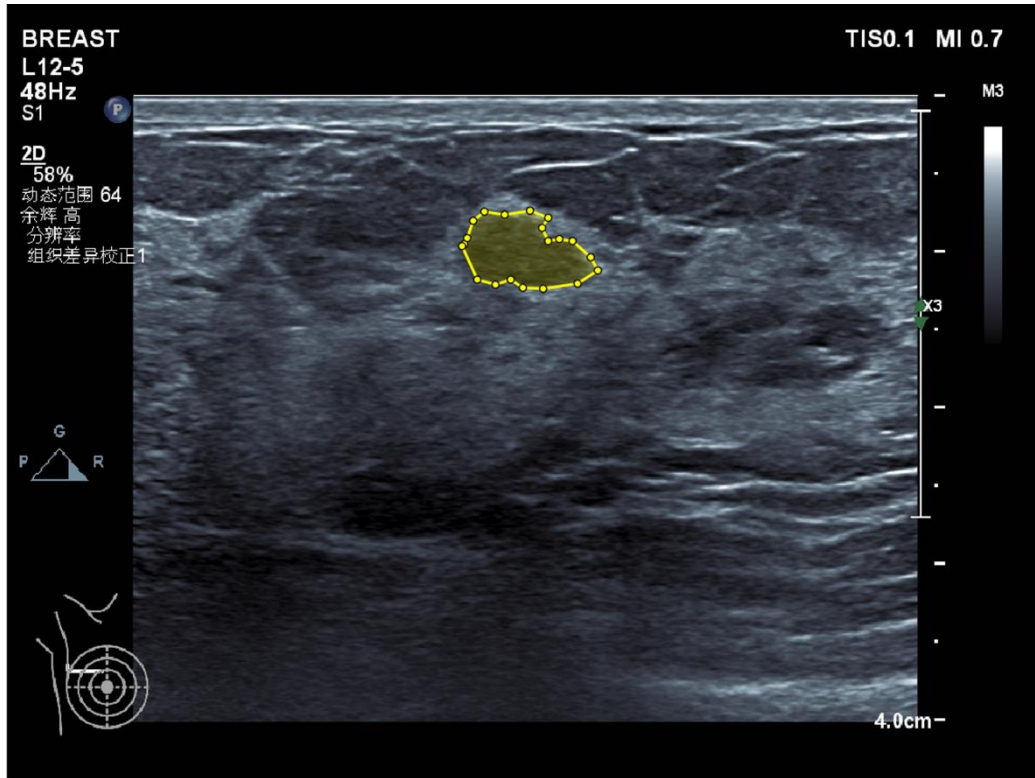


Figure 4.6 Illustrating tissue-padding tumour cropping scenarios for a breast tumour US image.

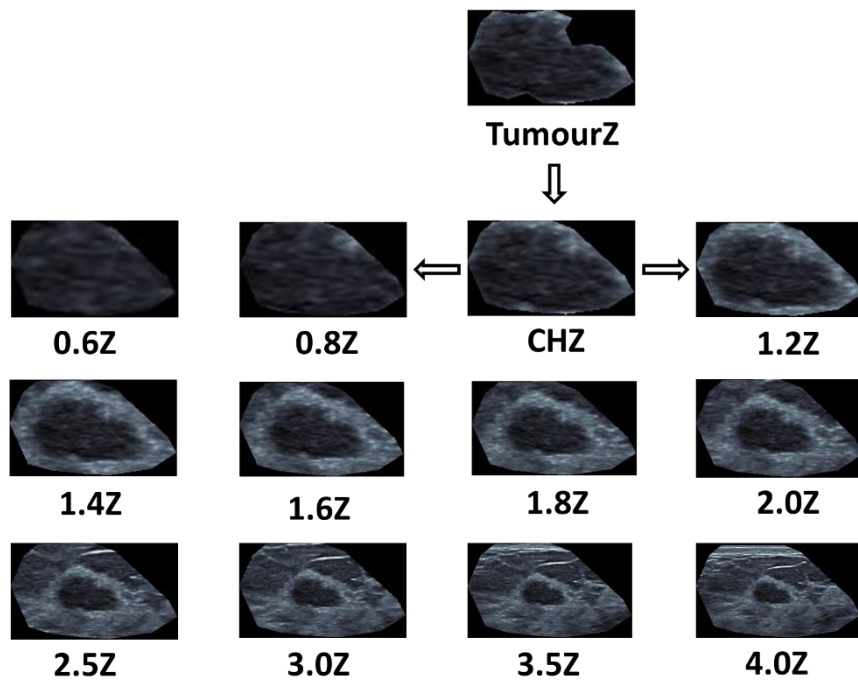
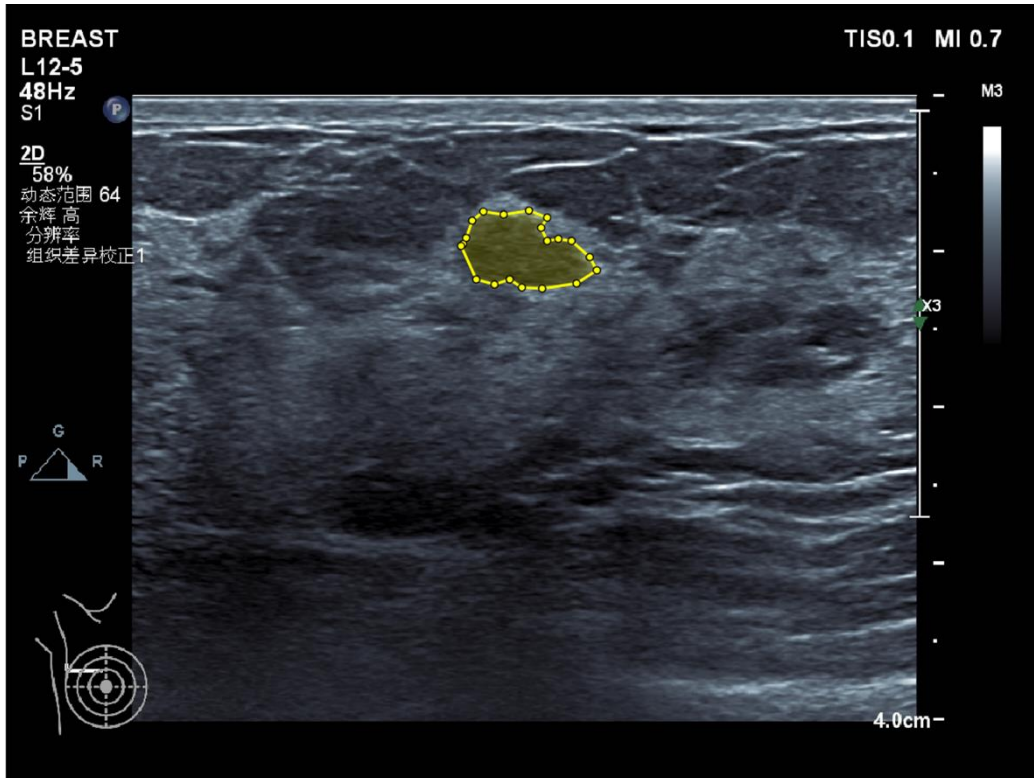


Figure 4.7 Illustrating 0-padding tumour cropping scenarios for a breast tumour US image.



### **4.3 Performance of DL/HC Models for the Convex Hull Cropping Strategies**

In this section, we shall test the performance of several CNN architectures and HC feature schemes when trained and tested on BUS lesion images with the various margin appending scenarios. We aim to determine the optimal tumour cropping ratio for model performance. In all experiments, we follow the 5-fold cross-validation protocol. We start by using the Renmin dataset due to its availability at the time of developing our tumour-cropping strategies. Later, we repeat the experiments with the Modelling dataset, which extends the Renmin dataset by datasets collected from 4 other Shanghai hospitals.

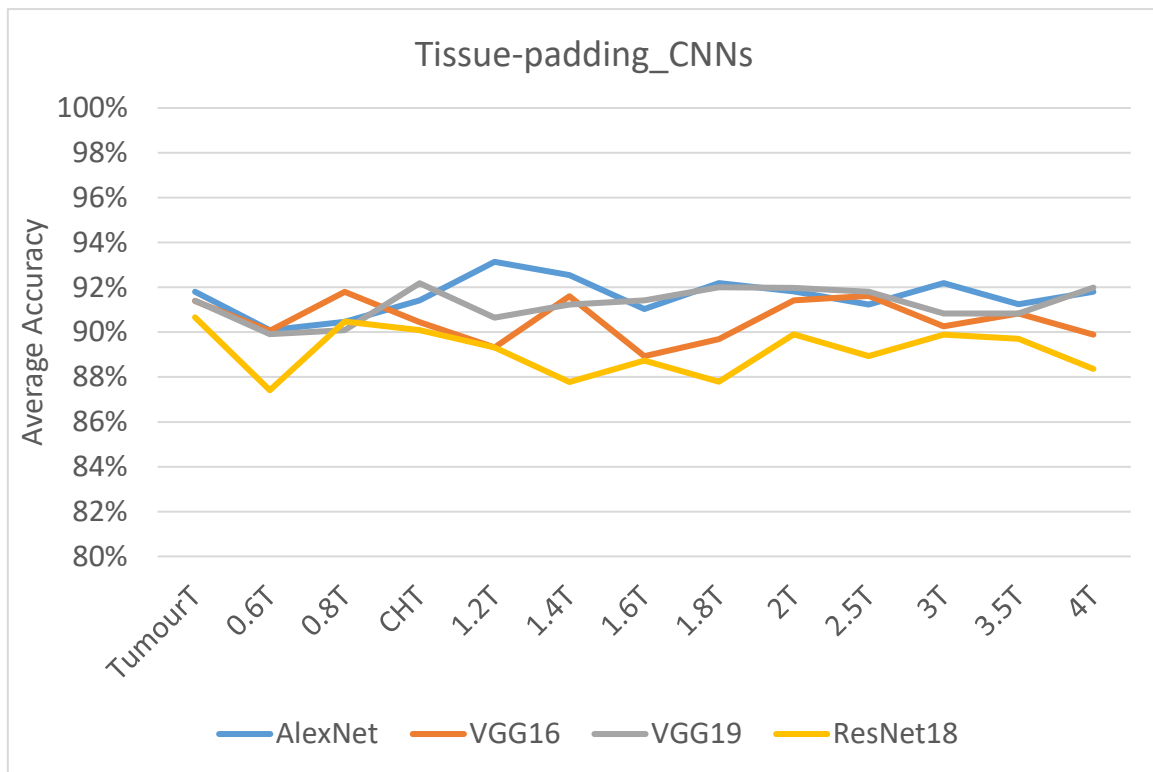
#### **4.3.1 Renmin Dataset - Performance Testing**

In this section, the impact of the CH-cropped RoIs with both sets of margin-appending strategies is tested on a variety of CNN as well as a few HC feature classification schemes. All lesion surrounding box images, post cropping and padding will be resized using the BiCubic method in accordance with the tested CNN model requirement. Moreover, the images were resized to 128x128 for the HC-feature models.

##### **4.3.1.1 Performance of CNN models**

At the time of developing our cropping and margin appending strategies, four CNN architectures were selected for our experiments: AlexNet, VGG16, VGG19, and ResNet18. All models are trained in fine-tuning mode, and the last FCL is replaced to adjust the binary classification (Benign, Malignant). The experiments do not aim to compare the performance of the various implemented CNN models but rather to estimate the range of their average accuracy for each tumour cropping scenario using the 5-fold cross-validation protocol.

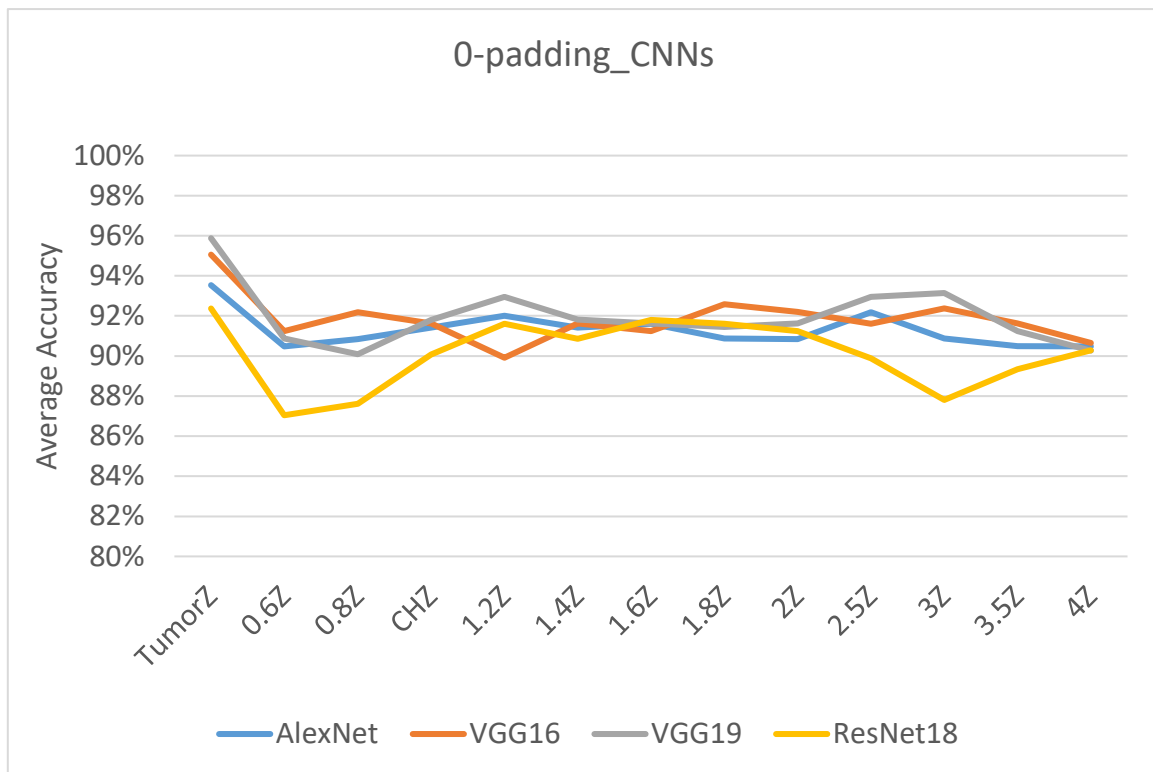
Figure 4.8 below presents the average classification accuracy for the four CNN models with the proposed TumourT and RoI CH-scaled cropping scenarios with tissue-padding.



**Figure 4.8 Performance of 4 CNN models for different cropping scenarios – Tissue padding.**

For all models, we cannot find significant differences from one cropping ratio to another. All DL models achieve accuracy in the range 91% - 92% accuracy with one or more cropping ratios. AlexNet achieves the highest accuracy of around 93% at 1.2T and 1.4T cropping ratios. The VGG16 achieves the highest accuracy of 92% (slightly outperforming the TumourT scenario) at different cropping ratios, including 0.8T, 1.4T, and 2.5T. The other two architectures perform differently at different cropping ratios with marginal increase/decrease (compared to the TumourT scenario) with the highest accuracy of no more than 92%. Overall, there is no clear pattern among the average accuracy of the four models at different cropping ratios to indicate the optimal cropping scenario. Therefore, it is hard to consider any of these cropping scenarios with tissue padding as the optimal one. However, these results indicate that the tissue of regions outside the tumour polygon should not be neglected completely.

Figure 4.9, below, presents the average classification accuracy of the same DL models but with the proposed TumourZ and RoI CH cropped scaled scenarios with 0-padding scheme.



**Figure 4.9 Performance of 4 CNN models for different cropping scenarios – Zero padding.**

Unlike the tissue padding scenario, the performance gap between the various cropping ratios for each CNN model is not negligible ( $\geq 5\%$ ). All the CNN models with the TumourZ cropping scenario outperform all the CHZ scaled ratios. VGG19 architecture achieves the highest accuracy of 96% and outperforms other models, followed by VGG16 (95% accuracy), and other models' performance is around 93%. Noticeably, the performance of all the models is the lowest at 0.6Z and 0.8Z, confirming the importance of *the regions immediately surrounding the approximated lesion border from the inside*. Except for the Resnet18 model, expanding the CHZ at ratios  $>1$  does not lead to noticeable performance differences for the other CNN models. This observation refines the last conclusion: *the regions immediately surrounding the approximated lesion border from the outside are also important*. However, the results of tissue padding, especially post scaling by ratios  $>1$ , indicate that the *inclusion of some external tumour tissue yields improved performance*. To have some understanding of this conclusion, we conducted limited experimental work to test the performance of two known HC-feature ML schemes on the images prepared according to the proposed cropping scenarios. Accordingly, this is not meant to compare the performance of these models with that of CNN models.

### 4.3.1.2 Performance of Handcrafted Feature Schemes

In order to see if these padding-related patterns of performance are due to the use of CNN models, we repeated the classification experiments with HC models of image analysis. All the images/ROIs are converted to grayscale and resized to [128, 128]. The resizing is done as a normalization for the HC feature schemes. HOG and ULBP are the selected texture features in the experiments, and the SVM with the linear kernel is the adopted classifier.

Figure 4.10 below presents the average classification performance for the selected HC texture features with different ROI cropping scenarios – tissue padding. HOG achieves an accuracy of 85% at TumourT and CHT, while its performance drops significantly to 73% and 78% at 0.6T and 0.8T respectively. Upscaling the CHT from 1.2T to 2T gradually decreases accuracy from 85% to 77%. Then, the accuracy starts to steadily increase with larger scaling ratios reaching 84% at 4T. A similar pattern of performance for all scaling ratios is achieved for ULBP; the accuracy of ULBP is 83% at TumourT and CHT. In contrast, its performance drops at 0.6T, 0.8T, and 1.2T but starts immediately to increase steadily, with larger scaling ratios reaching 83% at 4.0T. These results confirm that inside and outside lesion border contribute to improved performance, and the continuation of textures into larger border surrounding regions help regain performance lost when scaling by ratio <1.

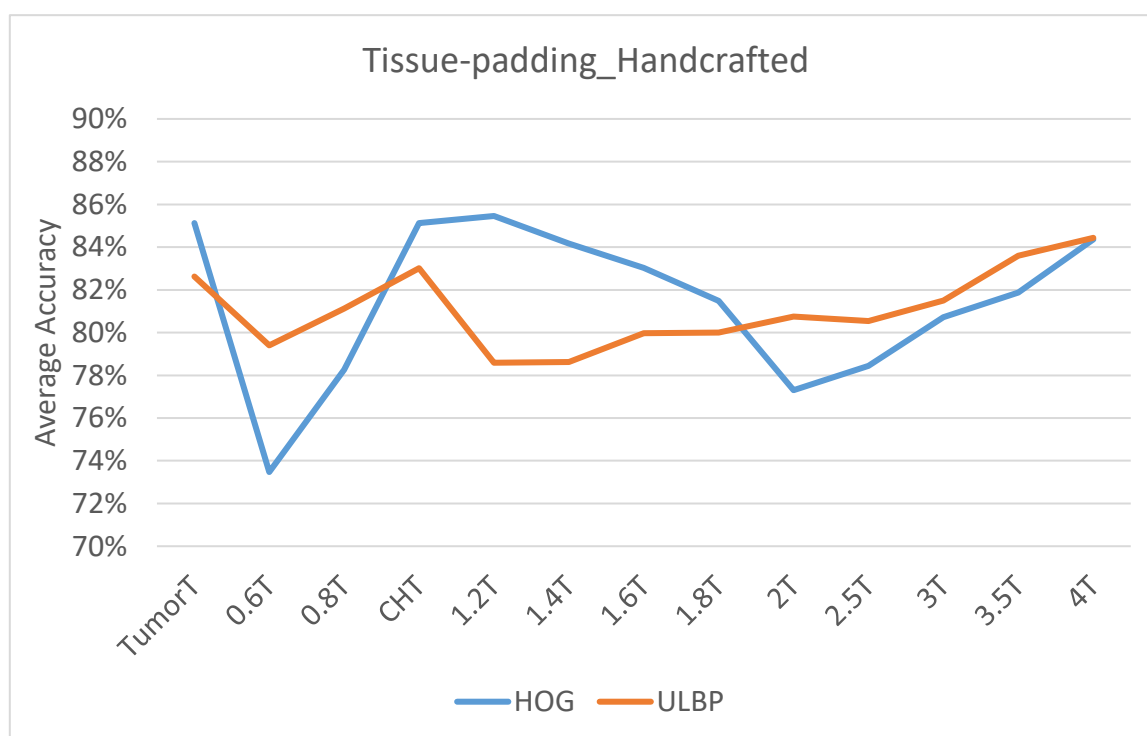
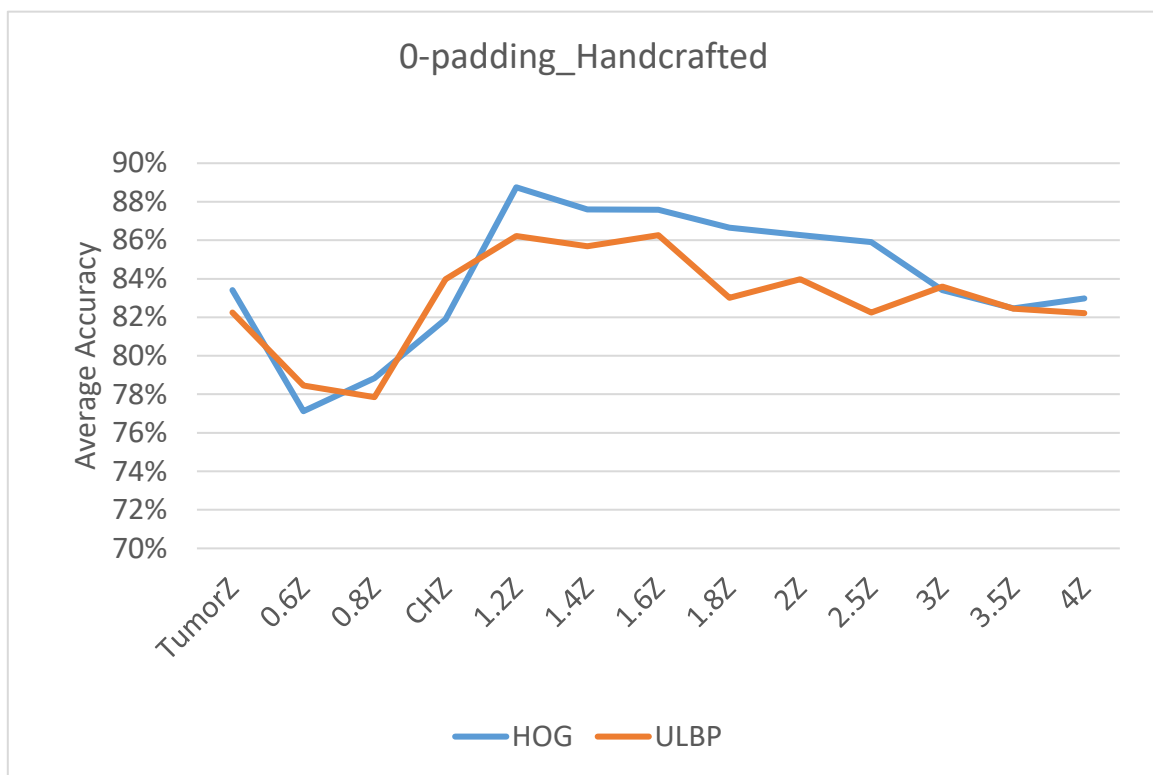


Figure 4.10 Performance of handcrafted models for cropping scenarios with Tissue padding.

Figure 4.11, below, presents the average accuracy for the same set of texture features but this time with zero-padding. There is little difference in the pattern of performance between the different extracted HC features across different scaling ratios, but this pattern is somewhat different from those in Figure 4.10 when TumourT cropping was used. For the TumourZ, both features achieve almost the same accuracy (around 83%) achieved with TumourT at and CHZ, while performance with both features drops by about 4% - 5% for the 0.6Z and 0.8Z scaling only to recover steadily with increased scaling reaching a peak of 89% with 1.2Z for HOG and a peak of 86.5% 1.6Z for ULBP. These results show that texture features outside the original CH(S) contribute to improved classification accuracy with HC feature schemes.



**Figure 4.11 Performance of handcrafted models for cropping scenarios with 0-padding.**

Comparing the results of HC feature schemes with those of CNN models may lead to different conclusions on the importance of features/tissues outside the lesion border. While the HC feature schemes achieve the highest performance at padding ratios > 1.2. For CNN BUS tumour classification, the results show that cropping scenario TumourZ achieve the highest performance leading to a simplistic conclusion that *no margin appending contributes to decision-making*. These results are influenced by the reliability of lesion-marked boundary points. However, the results of HC feature models post scaling by ratios

>1 show that it is prudent to *include some, but not all, tumour external tissue surrounding the lesion border yields improved performance*. Such a conclusion requires an in-depth analysis of the CNN models' predictions, and heatmaps visualisation may help in this respect. But before that, we extend the CNN model experiments by expanding the training dataset.

### 4.3.2 Modelling Dataset - Performance Testing

Having developed the previous cropping scenarios using the Renmin dataset, more US breast scan image samples recorded in other hospitals became available. We compiled the Modelling dataset that expands the Renmin dataset by adding US breast tumour scanned images from 4 other Shanghai hospitals. This provided the opportunity to determine the impact of having a larger dataset on the work conducted above, but only on CNN models.

In this section, all images of the lesion surrounding box were resized (post cropping and padding) using the BiCubic method in accordance with the tested CNN model requirement. Here, we shall present the results in terms of various performance measures beyond the average accuracy, but we shall present the results on the tumour surrounding box for *TumourT* and *TumourZ* appending strategy without scaling ratios as *TumourT* is the conventional RoI cropping while *TumourZ* is the optimal cropping scenario for CNN models. Table 4.1 presents the average performances of 6 CNN schemes, including AlexNet, VGG16, ResNet50, InceptionV3, Xception, and DenseNet201, trained on the Modelling dataset using the 5-fold cross-validation protocol.

**Table 4.1 Average validation performance of CNN models for *TumourT* and *TumourZ* scenarios.**

<b>TumourT</b>	<b>Accuracy</b>	<b>Sensitivity</b>	<b>Specificity</b>	<b>F1-score</b>	<b>AUC</b>
	<b>AVG ± STD</b>	<b>AVG ± STD</b>	<b>AVG ± STD</b>	<b>AVG ± STD</b>	<b>AVG ± STD</b>
<b>AlexNet</b>	<b>0.87 ± 0.02</b>	<b>0.79 ± 0.06</b>	<b>0.92 ± 0.02</b>	<b>0.82 ± 0.03</b>	<b>0.86 ± 0.03</b>
<b>VGG16</b>	<b>0.88 ± 0.01</b>	<b>0.80 ± 0.03</b>	<b>0.92 ± 0.02</b>	<b>0.83 ± 0.02</b>	<b>0.86 ± 0.01</b>
<b>ResNet50</b>	<b>0.87 ± 0.02</b>	<b>0.81 ± 0.03</b>	<b>0.91 ± 0.03</b>	<b>0.83 ± 0.03</b>	<b>0.86 ± 0.02</b>
<b>InceptionV3</b>	<b>0.85 ± 0.02</b>	<b>0.79 ± 0.05</b>	<b>0.89 ± 0.03</b>	<b>0.80 ± 0.03</b>	<b>0.84 ± 0.03</b>
<b>Xception</b>	<b>0.85 ± 0.02</b>	<b>0.80 ± 0.03</b>	<b>0.89 ± 0.03</b>	<b>0.80 ± 0.02</b>	<b>0.84 ± 0.02</b>
<b>DenseNet201</b>	<b>0.87 ± 0.01</b>	<b>0.78 ± 0.03</b>	<b>0.92 ± 0.03</b>	<b>0.82 ± 0.02</b>	<b>0.85 ± 0.01</b>
<b>TumourZ</b>	<b>Accuracy</b>	<b>Sensitivity</b>	<b>Specificity</b>	<b>F1-score</b>	<b>AUC</b>
	<b>AVG ± STD</b>	<b>AVG ± STD</b>	<b>AVG ± STD</b>	<b>AVG ± STD</b>	<b>AVG ± STD</b>
<b>AlexNet</b>	<b>0.88 ± 0.02</b>	<b>0.81 ± 0.05</b>	<b>0.93 ± 0.01</b>	<b>0.84 ± 0.03</b>	<b>0.87 ± 0.02</b>
<b>VGG16</b>	<b>0.89 ± 0.02</b>	<b>0.84 ± 0.06</b>	<b>0.92 ± 0.02</b>	<b>0.85 ± 0.03</b>	<b>0.88 ± 0.03</b>
<b>ResNet50</b>	<b>0.87 ± 0.01</b>	<b>0.84 ± 0.03</b>	<b>0.89 ± 0.01</b>	<b>0.83 ± 0.02</b>	<b>0.86 ± 0.02</b>
<b>InceptionV3</b>	<b>0.87 ± 0.02</b>	<b>0.79 ± 0.03</b>	<b>0.92 ± 0.02</b>	<b>0.82 ± 0.02</b>	<b>0.86 ± 0.02</b>
<b>Xception</b>	<b>0.87 ± 0.02</b>	<b>0.80 ± 0.05</b>	<b>0.91 ± 0.02</b>	<b>0.82 ± 0.03</b>	<b>0.86 ± 0.02</b>
<b>DenseNet201</b>	<b>0.89 ± 0.03</b>	<b>0.85 ± 0.04</b>	<b>0.91 ± 0.03</b>	<b>0.85 ± 0.04</b>	<b>0.88 ± 0.03</b>

Surprisingly, these experiments show that training with an expanded dataset does not improve the performance of the CNN models. Perhaps this reflects that the expansion did

not come from the same source but from different hospitals that might use different devices and follow different practices by different radiologists. Unfortunately, we could not expand the Renmin dataset with additional samples from the same hospital. On the other hand, these results confirmed the marginal increase in performance (around 2%) for all CNN models when images cropped with 0-padding (TumourZ) compared to the TumourT strategy. One notable observation in both cropping scenario experiments is that the gap between sensitivity and specificity is undesirably large. This means that for both cropping scenarios, more malignant tumours are misclassified than benign ones.

#### **4.4 CNN Cropped Lesion Models – Generalization Performance**

In the previous experiments, we determined the optimal cropping ratio for BUS tumour classification using DL schemes on the Renmin dataset, which is TumourZ. Here, we shall test the impact of the optimal cropping ratio on the trained DL models' capacity for generalisation onto the 2 external datasets (Test1 and BUSI) described earlier. Note that the Test1 dataset comes from a similar environment to the Modelling dataset, being collected from one of the Shanghai hospitals but distinct from those where the modelling dataset was collected. Hence, what we call generalisation performance may also be referred to as measures of model robustness.

##### **4.4.1 Generalisation of the Renmin-trained CNN Models**

Here, we used the trained CNN models on the Renmin dataset and tested their performances on the Test1 and BUSI datasets, having applied the two cropping scenarios (TumourT vs. TumourZ) on their images. The experimental work covers the 4 CNN architectures used in section 4.3.1, i.e., AlexNet, VGG16, VGG19, and ResNet18. The next 2 tables display the results of these experiments on Test1 with both (TumourT vs. TumourZ), followed by performance results on the BUSI dataset with (TumourT vs. TumourZ) scenarios, respectively.

Results in Table 4.2 show that the Renmin-trained CNN models' performance on the Test1 dataset drops significantly compared to their validation results, and the drop is much bigger for the TumourT scenario. In both cases, decisions by Resnet18 are more or less random.

**Table 4.2 Classification performance of Renmin-trained CNN models with (TumourT vs. TumourZ) for Test1 dataset.**

Test1-TumourT	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.58 <math>\pm</math> 0.04</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.32 <math>\pm</math> 0.07</b>	<b>0.64 <math>\pm</math> 0.02</b>	<b>0.66 <math>\pm</math> 0.04</b>
VGG16	<b>0.61 <math>\pm</math> 0.04</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.37 <math>\pm</math> 0.06</b>	<b>0.66 <math>\pm</math> 0.02</b>	<b>0.68 <math>\pm</math> 0.03</b>
VGG19	<b>0.66 <math>\pm</math> 0.06</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.46 <math>\pm</math> 0.10</b>	<b>0.70 <math>\pm</math> 0.04</b>	<b>0.73 <math>\pm</math> 0.05</b>
ResNet18	<b>0.57 <math>\pm</math> 0.06</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.30 <math>\pm</math> 0.10</b>	<b>0.64 <math>\pm</math> 0.03</b>	<b>0.65 <math>\pm</math> 0.05</b>
Test1-TumourZ	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.78 <math>\pm</math> 0.03</b>	<b>0.75 <math>\pm</math> 0.06</b>	<b>0.79 <math>\pm</math> 0.04</b>	<b>0.72 <math>\pm</math> 0.04</b>	<b>0.77 <math>\pm</math> 0.03</b>
VGG16	<b>0.76 <math>\pm</math> 0.03</b>	<b>0.74 <math>\pm</math> 0.10</b>	<b>0.77 <math>\pm</math> 0.06</b>	<b>0.70 <math>\pm</math> 0.04</b>	<b>0.76 <math>\pm</math> 0.03</b>
VGG19	<b>0.76 <math>\pm</math> 0.02</b>	<b>0.74 <math>\pm</math> 0.06</b>	<b>0.77 <math>\pm</math> 0.03</b>	<b>0.70 <math>\pm</math> 0.03</b>	<b>0.75 <math>\pm</math> 0.02</b>
ResNet18	<b>0.62 <math>\pm</math> 0.11</b>	<b>0.83 <math>\pm</math> 0.05</b>	<b>0.49 <math>\pm</math> 0.21</b>	<b>0.63 <math>\pm</math> 0.06</b>	<b>0.66 <math>\pm</math> 0.08</b>

Results in Table 4.3 also show that the Renmin-trained CNN models' performance on the BUSI dataset drops compared to their validation results, and the drop is less significant than those in Table 4.2. However, unlike the case of Test1 results, the sensitivity and specificity results are far from being balanced. These results again show that the average accuracy achieved with the TumourZ scenario is higher than that for the TumourT scenario. In both cases, decisions by Resnet18 are marginally better than a random process.

**Table 4.3 Classification performance of Renmin-trained CNN models with (TumourT vs. TumourZ) for BUSI dataset.**

BUSI-TumourT	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.71 <math>\pm</math> 0.04</b>	<b>0.99 <math>\pm</math> 0.01</b>	<b>0.56 <math>\pm</math> 0.06</b>	<b>0.70 <math>\pm</math> 0.03</b>	<b>0.77 <math>\pm</math> 0.03</b>
VGG16	<b>0.72 <math>\pm</math> 0.03</b>	<b>0.99 <math>\pm</math> 0.01</b>	<b>0.58 <math>\pm</math> 0.05</b>	<b>0.71 <math>\pm</math> 0.02</b>	<b>0.78 <math>\pm</math> 0.02</b>
VGG19	<b>0.75 <math>\pm</math> 0.02</b>	<b>0.99 <math>\pm</math> 0.00</b>	<b>0.63 <math>\pm</math> 0.03</b>	<b>0.73 <math>\pm</math> 0.02</b>	<b>0.81 <math>\pm</math> 0.02</b>
ResNet18	<b>0.68 <math>\pm</math> 0.05</b>	<b>0.98 <math>\pm</math> 0.01</b>	<b>0.52 <math>\pm</math> 0.08</b>	<b>0.68 <math>\pm</math> 0.03</b>	<b>0.75 <math>\pm</math> 0.03</b>
BUSI-TumourZ	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.89 <math>\pm</math> 0.03</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.84 <math>\pm</math> 0.05</b>	<b>0.87 <math>\pm</math> 0.03</b>	<b>0.92 <math>\pm</math> 0.02</b>
VGG16	<b>0.78 <math>\pm</math> 0.09</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.66 <math>\pm</math> 0.14</b>	<b>0.76 <math>\pm</math> 0.07</b>	<b>0.83 <math>\pm</math> 0.07</b>
VGG19	<b>0.82 <math>\pm</math> 0.04</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.72 <math>\pm</math> 0.06</b>	<b>0.79 <math>\pm</math> 0.04</b>	<b>0.86 <math>\pm</math> 0.03</b>
ResNet18	<b>0.66 <math>\pm</math> 0.08</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.49 <math>\pm</math> 0.13</b>	<b>0.68 <math>\pm</math> 0.05</b>	<b>0.74 <math>\pm</math> 0.06</b>

#### 4.4.2 Generalisation of the Modelling Dataset-trained CNN Models

Having trained several CNN models in fine-tuning mode using the Modelling dataset described earlier, we conducted experiments to test the trained model's generalisation ability on the 2 external datasets (Test1 and BUSI). The same CNN architectures used include AlexNet, VGG16, ResNet50, InceptionV3, Xception, and DenseNet201.



The experimental results in Table 4.4 show the average performance of the selected CNN architectures on the Test1 dataset, using the TumourT and TumourZ, respectively. For both margin appending schemes, the average accuracy of all models is above 88%. For TumourT, the highest accuracy of 94% was achieved with VGG16, but for TumourZ, the highest accuracy of 92% was achieved with Resnet50. Overall, the Sensitivity and Specificity rates achieved by all the models, and for both margin-appending strategies, are well balanced and ideally, i.e., no noticeable gap between these two metric performances. Other performance metrics, including F1-score and AUC, are reasonably high for all the CNN models. Unlike the results of the generalisation of the models trained by the Renmin for the Test1 dataset, the overall accuracy for the models with TumourZ cropping is marginally lower by a maximum of 2% compared to the case of TumourT.

**Table 4.4 Classification performance of the Modelling dataset-trained CNN models with (TumourT vs. TumourZ) for Test1 dataset.**

Test1-TumourT	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.93 <math>\pm</math> 0.04</b>	<b>0.93 <math>\pm</math> 0.03</b>	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.93 <math>\pm</math> 0.02</b>
VGG16	<b>0.94 <math>\pm</math> 0.01</b>	<b>0.93 <math>\pm</math> 0.03</b>	<b>0.94 <math>\pm</math> 0.02</b>	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.93 <math>\pm</math> 0.01</b>
ResNet50	<b>0.92 <math>\pm</math> 0.01</b>	<b>0.90 <math>\pm</math> 0.04</b>	<b>0.92 <math>\pm</math> 0.03</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.02</b>
InceptionV3	<b>0.89 <math>\pm</math> 0.01</b>	<b>0.89 <math>\pm</math> 0.05</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.86 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.02</b>
Xception	<b>0.89 <math>\pm</math> 0.01</b>	<b>0.86 <math>\pm</math> 0.04</b>	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.86 <math>\pm</math> 0.02</b>	<b>0.88 <math>\pm</math> 0.02</b>
DenseNet201	<b>0.92 <math>\pm</math> 0.01</b>	<b>0.90 <math>\pm</math> 0.04</b>	<b>0.92 <math>\pm</math> 0.01</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.02</b>
Test1-TumourZ	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.88 <math>\pm</math> 0.06</b>	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.88 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.02</b>
VGG16	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.04</b>	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.03</b>	<b>0.92 <math>\pm</math> 0.02</b>
ResNet50	<b>0.92 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.01</b>
InceptionV3	<b>0.90 <math>\pm</math> 0.01</b>	<b>0.85 <math>\pm</math> 0.06</b>	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.86 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.02</b>
Xception	<b>0.88 <math>\pm</math> 0.01</b>	<b>0.80 <math>\pm</math> 0.04</b>	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.83 <math>\pm</math> 0.01</b>	<b>0.86 <math>\pm</math> 0.01</b>
DenseNet201	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.92 <math>\pm</math> 0.04</b>	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.03</b>	<b>0.92 <math>\pm</math> 0.02</b>

Table 4.5 shows the generalisation performance of the Modelling dataset-trained CNN architectures on the external BUSI dataset for the TumourT and TumourZ appending schemes, respectively. Each and all DL models tested with TumourZ appending outperform its testing performance with TumourT appending scenario with a difference in the range (4% - 10%). Moreover, the average sensitivity and specificity rates are more balanced with TumourZ than with TumourT. For the F1 and AUC scores, the observed gaps between TumourZ and TumourT are more/less similar to those observed for the sensitivity and specificity scores.

**Table 4.5 Classification performance of the Modelling dataset-trained CNN models with (TumourT vs. TumourZ) for BUSI dataset.**

BUSI-TumourT	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.86 <math>\pm</math> 0.01</b>	<b>0.85 <math>\pm</math> 0.03</b>	<b>0.86 <math>\pm</math> 0.02</b>	<b>0.81 <math>\pm</math> 0.01</b>	<b>0.86 <math>\pm</math> 0.01</b>
VGG16	<b>0.86 <math>\pm</math> 0.01</b>	<b>0.86 <math>\pm</math> 0.04</b>	<b>0.86 <math>\pm</math> 0.03</b>	<b>0.81 <math>\pm</math> 0.01</b>	<b>0.86 <math>\pm</math> 0.01</b>
ResNet50	<b>0.83 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.01</b>	<b>0.79 <math>\pm</math> 0.02</b>	<b>0.79 <math>\pm</math> 0.01</b>	<b>0.86 <math>\pm</math> 0.01</b>
InceptionV3	<b>0.84 <math>\pm</math> 0.00</b>	<b>0.87 <math>\pm</math> 0.06</b>	<b>0.83 <math>\pm</math> 0.04</b>	<b>0.79 <math>\pm</math> 0.01</b>	<b>0.85 <math>\pm</math> 0.01</b>
Xception	<b>0.83 <math>\pm</math> 0.00</b>	<b>0.86 <math>\pm</math> 0.03</b>	<b>0.82 <math>\pm</math> 0.02</b>	<b>0.78 <math>\pm</math> 0.01</b>	<b>0.84 <math>\pm</math> 0.01</b>
DenseNet201	<b>0.83 <math>\pm</math> 0.02</b>	<b>0.86 <math>\pm</math> 0.04</b>	<b>0.82 <math>\pm</math> 0.04</b>	<b>0.78 <math>\pm</math> 0.02</b>	<b>0.84 <math>\pm</math> 0.01</b>
BUSI-TumourZ	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.90 <math>\pm</math> 0.03</b>	<b>0.92 <math>\pm</math> 0.04</b>	<b>0.90 <math>\pm</math> 0.04</b>	<b>0.87 <math>\pm</math> 0.03</b>	<b>0.91 <math>\pm</math> 0.02</b>
VGG16	<b>0.90 <math>\pm</math> 0.02</b>	<b>0.93 <math>\pm</math> 0.03</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.87 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.02</b>
ResNet50	<b>0.89 <math>\pm</math> 0.01</b>	<b>0.90 <math>\pm</math> 0.04</b>	<b>0.89 <math>\pm</math> 0.03</b>	<b>0.85 <math>\pm</math> 0.01</b>	<b>0.89 <math>\pm</math> 0.01</b>
InceptionV3	<b>0.90 <math>\pm</math> 0.00</b>	<b>0.88 <math>\pm</math> 0.07</b>	<b>0.91 <math>\pm</math> 0.03</b>	<b>0.85 <math>\pm</math> 0.01</b>	<b>0.89 <math>\pm</math> 0.02</b>
Xception	<b>0.93 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.01</b>	<b>0.93 <math>\pm</math> 0.01</b>
DenseNet201	<b>0.87 <math>\pm</math> 0.01</b>	<b>0.90 <math>\pm</math> 0.04</b>	<b>0.86 <math>\pm</math> 0.04</b>	<b>0.83 <math>\pm</math> 0.01</b>	<b>0.88 <math>\pm</math> 0.01</b>

Comparing the generalisation performances on Test1 with those achieved on BUSI, the TumourT scenario reveals a gap of (7% - 9%) in favour of Test1. At the same time, for the TumourZ scenario, the gap is not only much smaller in the range (0% - 3%), but the Xception model performs 5% more on BUSI than on Test1. These discrepancies in the generalisation performance for Test1 vs. BUSI may be partially attributed to the fact that Test1 is a dataset of unseen samples collected in another Shanghai municipality hospital, whereby one expects that their radiologist may have similar expertise.

The results in Tables 4.4 and 4.5 indicate somewhat unexpected significant differences in the generalisation performances between trained CNNs with the Modelling dataset and those trained with the Renmin dataset. What is surprising is that the images in the Renmin dataset form more than 32% of the images in the Modelling dataset. Moreover, images in the Test1 set were recorded in a Shanghai hospital, but none is included in the modelling or the Renmin Hospital datasets. The reported differences in the generalisation rates may have been impacted by the (1) variations in the deployed US devices, and (2) variations in the border marking/cropping strategies practised in the different hospitals. Note that different radiologists marked the lesion border in different hospitals. Perhaps we need to have a measure that could distinguish the contents of images recorded at Renmin Hospital from those recorded at other hospitals. One should not ignore the possibility of variation in the progression of malignancy in the images captured in the constituent hospitals. In this case, this may be due to the *shortcoming of considering tumour diagnosis as a binary classification problem*.

The experimental results in this section show that adopting the TumourZ instead of TumourT appending strategy helped to improve the generalization of the CNN models when trained with either the relatively small Renmin dataset or the expanded Modelling dataset. However, before making a hasty decision to *ignore the entire external tissue surrounding the lesion border inside the bounding box*, we need to remember that this section's experiments did not consider the cropping strategy of scaling the CH(S) by ratios  $> 1$ . Remembering the discussion we had earlier at the end of Section 4.3.1, it is indeed prudent to advise that the *inclusion of some, but not all, tumour external tissue surrounding the lesion border yields improved performance*. However, the determination of the subset(s) of the external tissue surrounding the lesion border to improve the performance of DL/HC schemes is a challenge that requires clinician advice but is outside the current scope of this thesis research work.

#### 4.5 Grad-CAM Visualization

In the above experiments, we found that the TumourZ cropping scenario is optimal for the CNN models. This may indicate that the decision made by the various CNN models pays little, if any, consideration to the tissue region outside the tumour polygon. Examining this assertion requires access to a visualisation tool of the model's decision heat maps for the two cropping scenarios that highlight the significance of tumour regions in terms of their contribution to the decisions of the various CNN models. Here, we adopt the Gradient-weighted Class Activation Mapping (Grad-CAM) visualization tool for this purpose. We use it to visually investigate the impact of TumourZ (tightly cropping the tumours with zero padding) on the CNN decision in comparison to that of the TumourT cropping scenario.

Grad-CAM is a visualization technique for DL models, specifically for CNNs. It provides a heatmap highlighting the various regions of the input image and colouring them in different shades according to the significance of contributions to the model's decision. Regions that contribute most to a particular prediction are coloured dark red. As the amount of red in the colouring of other regions decreases, their contribution to the decision gets smaller [112].

Grad-CAM has a wide range of applications in various domains; it can provide a way to understand the decisions made by black-box models, making it useful for applications where transparency and interpretability are essential. In the field of medical image analysis, Grad-CAM can be used to highlight the regions in a medical image that a DL model uses to make a diagnosis or prediction [113].

The Grad-CAM method is designed to be model-agnostic. It can be applied to any CNN architecture without modification, making it a versatile tool for visualizing and interpreting

DL models. It works by using the gradients of the target class, with respect to the feature maps in the final convolutional layer of the CNN model, to produce a coarse localization map highlighting the important regions in the input image for a given prediction. The gradients are averaged over all the feature maps and weighted by the magnitude of activation of each feature map to produce a final heatmap. The heatmap is then overlaid on the original input image to visually identify the regions that contribute the most to the target prediction. This process provides insight into the model's decision-making process by identifying the image regions in order of importance when making a prediction [112]. Ali Eskandari, in [38], investigated the Grad-CAM tool and made some modifications to provide an understanding of CNN decisions when classifying breast and thyroid lesions in US images.

Here, we use Grad-CAM to construct the heatmaps for the Xception model only when trained on the Modelling dataset. We also attempt to understand the impact of the TumourZ compared to TumourT on the model decision quality for both external datasets. We separately compute the average heatmap scores of true classified and misclassified cases for both Benign and Malignant lesions.

Test1 dataset visualized outputs are presented in Figure 4.12. It shows that for both TumourT and TumourZ cropping scenarios, the most important region is in the core of the tumour area for the true-classified benign and malignant cases. For the malignant cases, the highest scores occurred in the mid-upper tumour area, and the scores are more concentrated compared to the benign cases, while for the benign cases, the lesion's most important area is less concentrated and slightly shifts to the left side of the lesion. However, the distribution of the scores for the misclassified cases significantly differs from that for the true-classified ones; the most important areas are concentrated in small regions in different places for benign and malignant and also for TumourT vs. TumourZ. The visualization shows that moving from the TumourT to the TumourZ cropping scenario does not affect the quality of the CNN decision in general, especially for the true classified cases. The CNN model-focused area in both cases is the inner area of the tumour. Moreover, the general shape of the GRAD-CAM visualization for Benign vs. Malignant is more separable with TumourZ compared to TumourT, especially for the true classified cases.

Figure 4.13 presents the heat map visualization for the BUSI dataset. Again, a similar pattern of score distribution to the Test1 dataset is achieved across the two cropping scenarios and tumour classes for true-classified and misclassified cases.

Both figures (4.12 and 4.13) illustrate that for both scenarios, TumourT and TumourZ, the most important areas in decision-making are inside the lesion area, and the quality of the decision, in general, is unchanged, especially for the true classified cases. Moreover, the general shape of the Grad-CAM visualization for Benign vs. Malignant is more separable with TumourZ compared to TumourT. So, the outcome of the visualization from both figures for Test1 and BUSI datasets are consistent, and it shows that it is safe to use TumourZ as the optimal cropping scenario for the CNN models, which results in better model classification performance and generalization, in fact, with better CNN decision quality.

In conclusion, against all expectations, these heatmaps show that the regions outside the lesion are not seriously considered when making decisions.

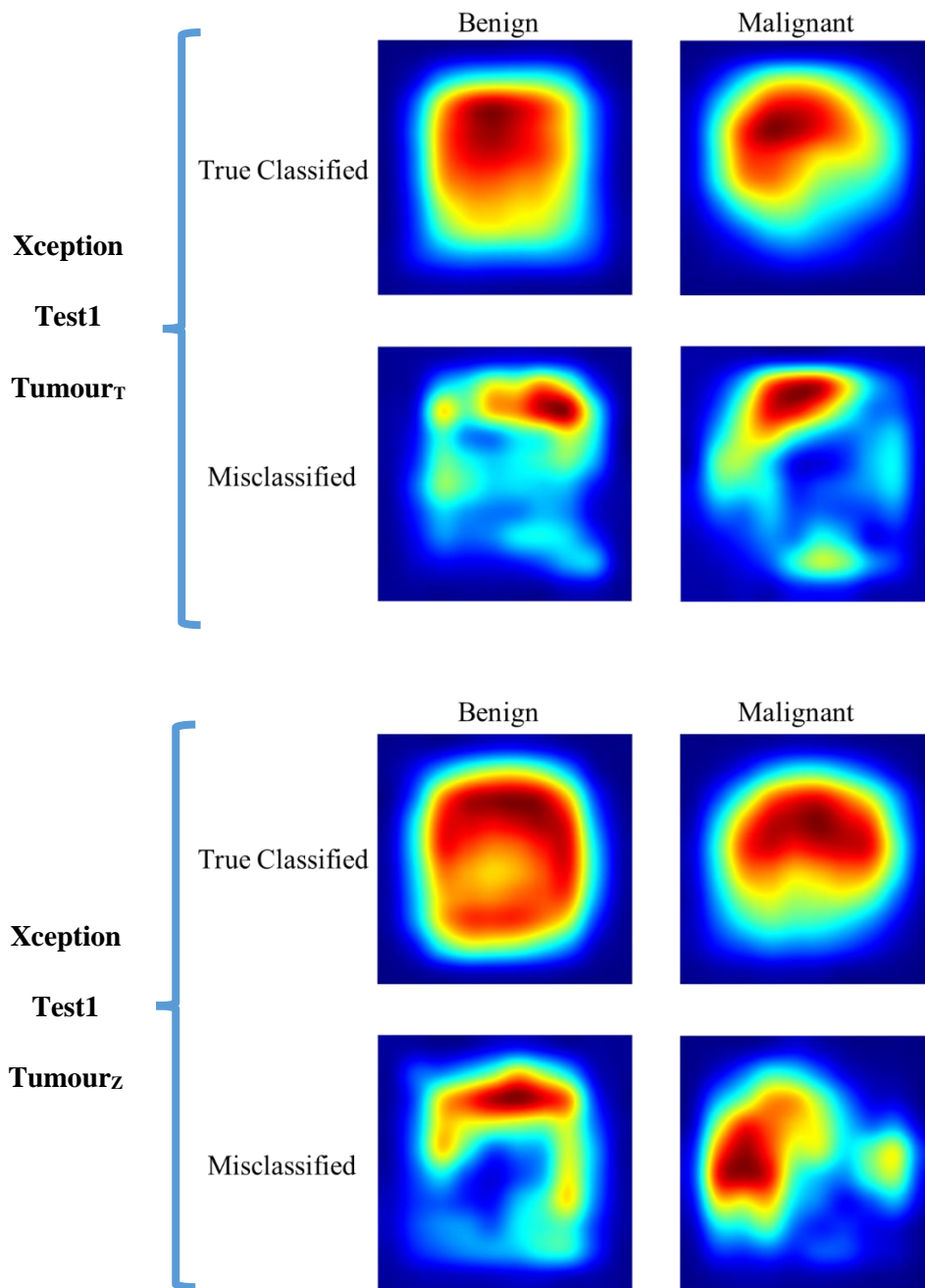


Figure 4.12 Average heatmap scores of true classified and misclassified cases for both Benign and Malignant classes by Xception model of Test1 dataset: (Tumour<sub>T</sub> vs. Tumour<sub>Z</sub>).

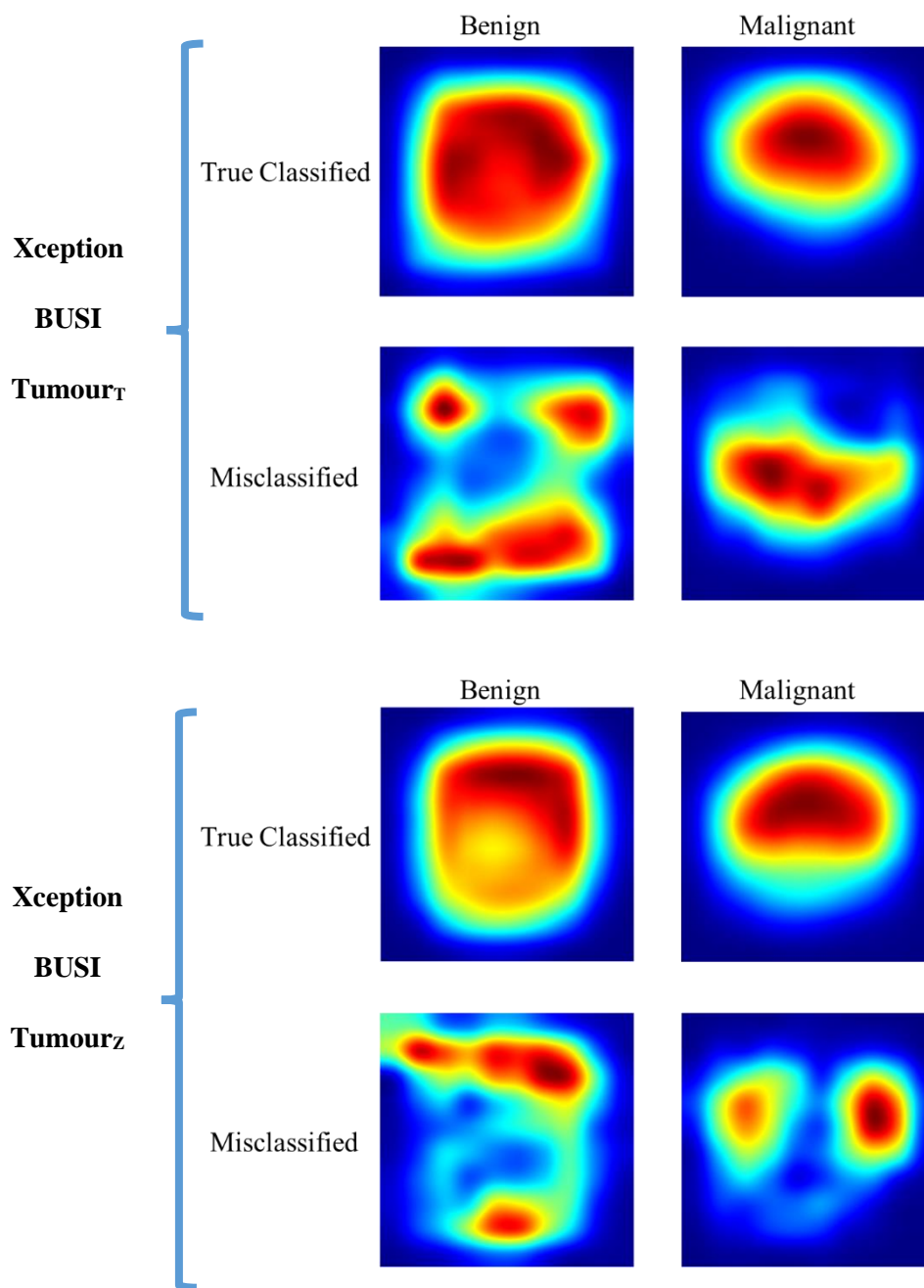


Figure 4.13 Average heatmap scores of true classified and misclassified cases for both Benign and Malignant classes by Xception model of BUSI dataset: (Tumour<sub>T</sub> vs. Tumour<sub>Z</sub>).

## 4.6 Conclusion

In this chapter, our investigations focused on the critical aspect of lesion cropping in BUS tumour images. It is a crucial pre-processing technique influencing the performance of DL image analysis schemes and their generalization capabilities into unseen external datasets. Through our analysis, we have demonstrated the superiority of CH lesion border approximation over other border approximation approaches in terms of minimizing the exclusion of lesion pixels. This approach not only exhibits computational efficiency but also lends itself to easy expansion for margin appending.

Among the proposed RoI cropping scenarios, our experiments have substantiated TumourZ as the optimal cropping scenario for CNN models. At the same time, 1.2Z emerged as the optimal tumour cropping scenario for HC feature schemes in terms of model classification performance. Moreover, the inclusion of some external tissue surrounding the lesion border, as demonstrated by the tissue-padding scenarios, has shown promise in improving model performance. However, further investigation is needed to determine the precise external tissue regions that should be included in the cropped RoI for better CNN model performance. By expanding the Renmin dataset to include images from different hospitals, we observed a reduction in model performance during the corresponding 5-fold cross-validation training of CNN models. Interestingly, this expansion significantly enhanced the CNN models' generalization capabilities on two external datasets. Furthermore, we found that TumourZ played a pivotal role in enabling CNN models trained on the Modelling dataset to overcome their reduced validation performances and exhibit superior generalization compared to the disappointing outcomes observed with Renmin-trained CNN models on the two external testing datasets.

The Grad-CAM visualizations further supported the conclusion that the TumourZ cropping scenario provides, if not superior, decision quality comparable to TumourT for CNN models. Notably, our findings indicated that DL models focus primarily on the region inside the lesion, disregarding the surrounding region in both cropping scenarios. Importantly, TumourZ resulted in improved model classification performance, generalization, and better CNN decision quality.

It is worth noting that the significant differences in generalization performances between CNN models trained on the Modelling dataset and those trained on the Renmin dataset may not be solely attributed to differences in training dataset sizes and diversity. In the next chapter, and as part of our broader investigation into performance influencing factors, we



delve into the issue of *Image Quality* to gain further insights and understanding. Overall, this chapter has shed light on the crucial role of lesion shape cropping and optimal tumour ROI cropping in DL-based BUS tumour image analysis.

# Chapter 5: A No-reference Multi-Characteristics US Image Quality Descriptor

The primary objective of this chapter's investigation is to shed light on the significant disparities observed in the DL generalization capabilities while using the Renmin dataset as the training set in contrast to the larger Modelling multi-centre dataset. At the early stages of the TenD project, a high-performing CNN model trained on data from the single medical centre (Renmin) was found to perform well-below expectation when tested on another centre's data, and the radiologists attributed this to probable dissimilarities in image *quality* between the centres. In Chapter 3, we noted that the concept of quality for US images is poorly understood due to not being rigorously investigated. Experienced radiologists learn through extensive training to pass judgement on the quality of recorded images. Still, no computational model of US image quality has so far been accepted that aligns well with experienced radiologists' assessment.

Guided by the long-established UIQI and other recent IQA schemes, this chapter aims to develop an US quality-related feature vector as a descriptor that can be employed to compare datasets from different centres. Section 5.1 reviews the reference-based UIQI and several no-reference IQA techniques, all proposed for natural images. We use a small BUS image dataset of 20 images recorded in 2 different Shanghai hospitals subjectively labelled by an experienced radiologist as good images and bad images, respectively, to demonstrate that neither UIQI nor other IQA schemes feasibly separate the images in the dataset according to the radiologist labelling. Section 5.2 lays the groundwork for proposing a no-reference multi-characteristic image quality (MCIQ) feature vector to serve as US quality descriptor and outline its construction. Section 5.3 presents the experimental findings using MCIQ to explain the disparities observed in DL generalization results in Chapter 4. Section 5.4 presents experimental work for using MCIQ for a variety of other quality-related dataset US image applications, e.g., determining the strength of MCIQ in discriminating benign from malignant masses in US breast scan images. In Section 5.5, we revisit the pilot study 20-sample dataset, summarise the limitations of MCIQ in reflecting radiologists' subjective quality assessment of US images, and discuss a potential approach to overcome these limitations. Section 5.6 is a summary of the chapter's conclusion.

## **5.1 Computer-based Measure of Ultrasound Image Quality**

In the clinical setting, the quality of US images is a key descriptor for evaluating the performance of ultrasonic imaging devices and ML analysis algorithms. Traditionally, US (and other medical) IQA is regarded as a subjective issue, and little attention was paid to developing a quantitative IQA.

### **5.1.1 Existing Work on Ultrasound Image Quality Assessment**

Normally the quality of recorded US image/video is assessed subjectively by the radiologist to select the appropriate frame prior to assessing tumour status. Radiologists have been trained for years to acquire knowledge on how to conduct US scans of different tissues/organs and select image frames of the RoI that are suitable for the clinical purpose of tumour classification. However, our literature review on automatic IQA for US images revealed no standardised method for subjective (let alone objective) IQA for this purpose. In contrast, several IQA models are known for natural images. Hemmsen et al. [114] point out that diagnostic values of US images should guide their improvement. The absence of objective IQA methods for US images is used in [114] to propose a framework for alternative subjective IQA that includes equipment and methodology for clinical IQ evaluation.

It is worth noting that the deployment of US imaging for medical diagnostics came much later than other modalities, including medical scanning modalities (e.g., X-Rays), and understandably went through several safety assessment stages of their use. In this elongated process, various medical regulatory and accrediting agencies developed regulation requirements for maintaining the high-level performance of the US devices, manufacturers set clear specifications, and it is customary to have regular tests of the performance of their devices [25], [114]. In this context, US image quality depends on the transducer (probe), the device electronics, the pre- and post-processing of transmitted and received signals, as well as the fidelity performance of the display monitors [114].

Quality assessment procedures have been proposed for B-mode and Doppler imaging systems as part of performance evaluation or quality assurance of US devices to ensure patient/operator safety in line with the manufacturer's recommendations and in compliance with regulatory and accreditation agencies [25]. The subjective quality assessment by radiologists is expected to feed into such quality assurance procedures to enable timely fault detection. Sassaroli et al. [25] describe some of these procedures as part of quality control clinical protocols for US imaging and biopsy based on computer-based methods. The various quality control tests are not possible to conduct for our task of defining US image quality

descriptors; the corresponding computer-based procedures consist of reference-based image contrast tests that compare US region images of 2 types of target objects embedded on so-called phantoms (i.e., materials that have average acoustic properties of soft tissue). These tests are based on evaluating statistical parameters of image quality, such as GMF, image contrast, CNR, and HCSR. Detailed descriptions and formulae are given in [25].

Developing objective image quality descriptors specific to the US faces a significant challenge. This is because relying heavily on existing knowledge of image quality schemes for natural images is problematic due to fundamental differences in content, including the quantity and distribution of various texture features. The remarkable success of CNN models in data/image analysis has sparked efforts to utilize these DL models for acceptable US quality assessment, particularly by expert radiologists. Recently, Zhang et al. [115] proposed a quantitative study using CNN models for the US IQA measure. Initially, a dataset of BUS images was created by degenerating a number of high-quality US images that were pre-processed and scored by 4 experienced doctors. Then, 478 US images, labelled by averaging their scores, were selected for training and testing. Afterwards, a deep CNN network and a residual network are obtained to establish the proposed IQA model. They show that the CNN-based IQA is feasible and effective, but more work is needed.

In Chapter 3, we reviewed some commonly deployed automatic reference-based IQA metrics to assess the quality of natural images. We found that the various components of the UIQI [93], [94] reflect the human brain's recognisable geometric/structural image distortion, blurring, the appearance of shadows and the presence of noise. The main challenge to directly deploying these descriptors for US images relates to identifying high-quality reference images without relying on subjective assessment. The dynamically changing micro-environment of human tissue influenced by the flow of blood (and other body liquids) means that subjective assessment can only be done by highly experienced radiology experts who are trained to recognise US contents of interest (in such a dynamic micro-environment) and distinguish artefacts from tissue aberrations.

Since the early last decade, a variety of no-reference image quality assessments (NRIQA) have been proposed but again designed specifically for natural images. NRIQA schemes can be categorised in different ways, e.g., in terms of prior awareness (or not) of distortion as being rated by human observers. The state-of-the-art of such schemes predicts natural image quality without knowledge of distortion type but relies on human opinion scores (subjective scores) to enable learning regression-based quality assessment of distorted images, see [116]. The spatial Natural Scene Statistics (NSS) model of images initially applies local

normalization of the luminance (intensities), where the local means and standard deviations are simply obtained by convolving the images with a 3x3 Gaussian filter. The NSS model is deemed to measure distortion-caused losses of image *naturalness*. The main hypothesis of the developed NRIQA schemes is that distorted images have certain latent characteristics distinguishing them from undistorted ones. These characteristics are based on so-called *Visual Words* extracted from a sufficiently large set of *pristine* and *distorted* images.

Mittal et al. [117] introduced the opinion-aware blind/referenceless image spatial quality evaluator (BRISQUE). Unlike reference-based IQA schemes, BRISQUE does not compute image fidelity (i.e., distortion-specific features). Instead, it employs the NSS (trained on features obtained from a corpus of both natural and distorted images) and relies on human judgments of the quality of these images to measure the possible distortion-caused losses of *naturalness*. Figure 1(b) in [117] demonstrates how the loss of naturalness results by mapping different natural scene images into different image domain blocks. This is similar to our observation in Chapter 3 that the non-uniformity of the spatial distribution of illumination/texture in US images distinguishes them from natural images. BRISQUE relies on a holistic measure of quality without the need for distortion-specific features. It is suitable for various image distortions, including Gaussian blur, JPEG compression, and white noise. In [118], the NIQE is, an opinion-unaware and distortion-unaware scheme, developed by the BRISQUE's authors that continued to adopt the NSS model of image naturalness, but it only uses the NSS features obtained from a corpus of natural images and removed reliance on the awareness of distortion, i.e., NIQE does not require training on large databases of human opinions. Post luminance local normalisation, the image is partitioned into  $P \times P$  patches from each of which certain NSS features are computed, and only subsets of these patches are used to train the proposed IQA scheme.

In 2015, Venkatanath N et al. [119] developed the PIQUE algorithm that generates a fine-grained block-level distortion map that mimics human behaviour. Unlike opinion-based supervised learning methods, PIQUE attempts to quantify distortion without the need for any training data. The model uses cues from the human visual system to quantify distortion. Like NIQE, it extracts NSS features from non-overlapping partitioning blocks of fixed size and labels them as uniform or nonuniform/spatially active blocks using a threshold of 10%. The spatially active blocks are scored for the 2 dominant distortions as a result of compression and the presence of noise.

For all these 3 schemes, lower scores indicate good quality, but higher scores indicate lower quality. Clear thresholds are generally unspecified by the proposed schemes but are deemed

to be application-dependent. However, in general, the PIQUE score  $>50$  indicates bad quality, and  $<50$  shows better quality. MATLAB procedures and source codes for implementing these schemes are available in the literature. It is unfeasible to use the approaches of these 3 IQA methods for US images due to many factors, including the lack of availability of large datasets of *pristine* US images and lack of awareness of the type of distortions of interest in US images. However, Dey et al. [120] recently constructed a feature vector by combining the BRISQUE, NIQE, and PIQUE quality scores with image Entropy and used it to classify breast tumours from US images. We denote this approach by *TripleIQA+Entropy*. Next, we shall present the shortcomings of using the natural image full reference UIQI and the above 3 NRIQA schemes in explaining the subjective assessment of a TenD radiologist partner of a small set of 20 BUS images recorded in 2 different Shanghai clinical centres.

### 5.1.2 Towards No-Reference Objective US IQA Descriptor

Our aim in developing US image quality descriptor was to find a quantitative scheme that can help distinguish, or not, between datasets of US images recorded at different clinical centres with the possibility, or not, of generalising CNN model performance when trained in one centre to images from other centres. Our approach, presented in the next section, was developed in several steps and benefits from the above observations and those in Chapter 3. The performance of ML models used for US image classification is also influenced by natural image quality distortion characteristics (e.g., blurriness, shadows, poor contrast, and noise). Setting aside the absence of a standardised method for subjective quality assessment of US images, our objective may benefit from testing the performance of existing full reference IQA developed for natural images on US image datasets. On the other hand, we also expect to benefit from investigating the performance of some of the known NRIQA discussed above on US image datasets. Also, in theory, one may even try to develop a CNN-based IQA scheme, as done in [115]. Our assertion, shared with [117], is that IQA model performance should correlate with expert subjective assessment. Conducting large-scale subjective studies relying on skilled radiologists is (and was) unrealistic for our research.

We initiated our investigations into testing the performance of various components of full reference UIQI as well as their product, by forming a *pilot* dataset of BUS tumour RoI images, cropped with the TumourT cropping scenario and each labelled as *Good* or *Bad* following visual examination by an experienced TenD radiology partner. The pilot dataset consists of 10 good quality images:  $\{G_1, G_2, G_3, \dots, G_{10}\}$  and 10 poor quality images

as (B\_1, B\_2, B\_3, ..., B\_10). The 3 UIQI components: loss of correlation, luminance distortion, and contrast distortion, were computed for pair (X, Y) images in the pilot dataset, and obtained scores are presented in Tables 5.1, 5.2, and 5.3, respectively.

**Table 5.1** The loss of correlation measure between pairs of Good/Bad images.

<b>Correlation (1)</b>	<b>G_1</b>	<b>G_2</b>	<b>G_3</b>	<b>G_4</b>	<b>G_5</b>	<b>G_6</b>	<b>G_7</b>	<b>G_8</b>	<b>G_9</b>	<b>G_10</b>
B_1	0.4	0.3	0.2	0.0	-0.2	0.6	0.2	0.6	0.7	-0.1
B_2	0.1	0.1	0.3	-0.1	0.0	0.6	0.1	0.4	0.4	0.0
B_3	0.0	0.2	0.2	-0.2	0.1	0.4	0.0	0.3	0.2	0.0
B_4	0.0	-0.1	0.1	-0.1	0.3	0.1	0.1	0.0	-0.1	-0.1
B_5	-0.3	-0.1	0.0	-0.1	0.3	-0.2	0.0	-0.4	-0.6	0.0
B_6	0.1	0.1	-0.1	-0.1	0.1	-0.1	-0.1	0.1	0.2	-0.2
B_7	0.1	0.1	0.0	-0.2	0.0	0.2	0.0	0.2	0.1	-0.1
B_8	0.3	0.1	0.2	-0.2	0.1	0.4	0.2	0.4	0.5	-0.1
B_9	0.5	0.1	0.0	-0.1	-0.3	0.4	0.0	0.5	0.8	-0.1
B_10	0.2	-0.1	0.2	-0.2	0.1	0.4	0.0	0.2	0.2	-0.1
<b>Correlation (2)</b>	<b>G_1</b>	<b>G_2</b>	<b>G_3</b>	<b>G_4</b>	<b>G_5</b>	<b>G_6</b>	<b>G_7</b>	<b>G_8</b>	<b>G_9</b>	<b>G_10</b>
G_1	1.0	0.0	0.0	0.0	0.0	0.3	0.1	0.4	0.5	-0.2
G_2	0.0	1.0	0.1	0.2	-0.2	0.1	0.0	0.1	0.2	0.2
G_3	0.0	0.1	1.0	0.0	0.0	0.3	0.2	0.2	0.1	0.1
G_4	0.0	0.2	0.0	1.0	-0.1	-0.1	0.0	-0.1	-0.1	0.2
G_5	0.0	-0.2	0.0	-0.1	1.0	0.0	0.2	-0.1	-0.2	-0.2
G_6	0.3	0.1	0.3	-0.1	0.0	1.0	0.2	0.6	0.5	-0.2
G_7	0.1	0.0	0.2	0.0	0.2	0.2	1.0	0.1	0.0	0.0
G_8	0.4	0.1	0.2	-0.1	-0.1	0.6	0.1	1.0	0.5	-0.1
G_9	0.5	0.2	0.1	-0.1	-0.2	0.5	0.0	0.5	1.0	-0.1
G_10	-0.2	0.2	0.1	0.2	-0.2	-0.2	0.0	-0.1	-0.1	1.0
<b>Correlation (3)</b>	<b>B_1</b>	<b>B_2</b>	<b>B_3</b>	<b>B_4</b>	<b>B_5</b>	<b>B_6</b>	<b>B_7</b>	<b>B_8</b>	<b>B_9</b>	<b>B_10</b>
B_1	1.0	0.4	0.3	-0.1	-0.5	0.0	0.2	0.6	0.7	0.2
B_2	0.4	1.0	0.5	0.2	-0.1	-0.1	0.1	0.4	0.3	0.3
B_3	0.3	0.5	1.0	0.2	0.1	0.1	0.2	0.2	0.1	0.2
B_4	-0.1	0.2	0.2	1.0	0.3	-0.1	-0.1	0.1	-0.2	0.2
B_5	-0.5	-0.1	0.1	0.3	1.0	-0.1	0.0	-0.3	-0.6	0.0
B_6	0.0	-0.1	0.1	-0.1	-0.1	1.0	0.2	0.1	0.2	-0.1
B_7	0.2	0.1	0.2	-0.1	0.0	0.2	1.0	0.1	0.1	0.1
B_8	0.6	0.4	0.2	0.1	-0.3	0.1	0.1	1.0	0.5	0.3
B_9	0.7	0.3	0.1	-0.2	-0.6	0.2	0.1	0.5	1.0	0.1
B_10	0.2	0.3	0.2	0.2	0.0	-0.1	0.1	0.3	0.1	1.0

Table 5.1 presents the results in three categories: Top: Bad images (Target) vs. Good images (Reference), Mid: Good images (Target) vs. Good images (Reference), and Bottom: Bad images (Target) vs. Bad images (Reference). The loss of correlation metric's dynamic range

is between -1 and 1, with a value of 1 indicating the highest correlation between the images. Overall, the results show a low correlation between all pairs of different images in the three categories. Therefore, no image is suitable as a reference for the loss of correlation quality metric.

**Table 5.2 The luminance distortion measure between pairs of Good/Bad images.**

<b>Luminance (1)</b>	<b>G_1</b>	<b>G_2</b>	<b>G_3</b>	<b>G_4</b>	<b>G_5</b>	<b>G_6</b>	<b>G_7</b>	<b>G_8</b>	<b>G_9</b>	<b>G_10</b>
<b>B_1</b>	1.0	0.9	1.0	1.0	1.0	0.7	1.0	0.8	1.0	1.0
<b>B_2</b>	1.0	1.0	1.0	1.0	1.0	0.6	1.0	0.7	1.0	0.9
<b>B_3</b>	1.0	0.9	1.0	1.0	1.0	0.8	1.0	0.8	0.9	1.0
<b>B_4</b>	1.0	1.0	1.0	0.9	1.0	0.6	1.0	0.6	1.0	0.9
<b>B_5</b>	1.0	1.0	1.0	0.9	0.9	0.5	1.0	0.6	1.0	0.9
<b>B_6</b>	0.9	1.0	0.9	0.9	0.9	0.5	1.0	0.5	1.0	0.8
<b>B_7</b>	1.0	1.0	1.0	0.9	1.0	0.6	1.0	0.6	1.0	0.9
<b>B_8</b>	0.9	1.0	0.9	0.8	0.9	0.5	0.9	0.5	1.0	0.8
<b>B_9</b>	1.0	0.9	1.0	1.0	1.0	0.7	1.0	0.8	1.0	1.0
<b>B_10</b>	1.0	0.9	1.0	1.0	1.0	0.8	1.0	0.8	0.9	1.0
<b>Luminance (2)</b>	<b>G_1</b>	<b>G_2</b>	<b>G_3</b>	<b>G_4</b>	<b>G_5</b>	<b>G_6</b>	<b>G_7</b>	<b>G_8</b>	<b>G_9</b>	<b>G_10</b>
<b>G_1</b>	1.0	1.0	1.0	1.0	1.0	0.7	1.0	0.7	1.0	1.0
<b>G_2</b>	1.0	1.0	1.0	0.9	1.0	0.6	1.0	0.6	1.0	0.9
<b>G_3</b>	1.0	1.0	1.0	1.0	1.0	0.7	1.0	0.7	1.0	1.0
<b>G_4</b>	1.0	0.9	1.0	1.0	1.0	0.8	1.0	0.8	0.9	1.0
<b>G_5</b>	1.0	1.0	1.0	1.0	1.0	0.7	1.0	0.8	1.0	1.0
<b>G_6</b>	0.7	0.6	0.7	0.8	0.7	1.0	0.7	1.0	0.6	0.8
<b>G_7</b>	1.0	1.0	1.0	1.0	1.0	0.7	1.0	0.7	1.0	0.9
<b>G_8</b>	0.7	0.6	0.7	0.8	0.8	1.0	0.7	1.0	0.7	0.9
<b>G_9</b>	1.0	1.0	1.0	0.9	1.0	0.6	1.0	0.7	1.0	0.9
<b>G_10</b>	1.0	0.9	1.0	1.0	1.0	0.8	0.9	0.9	0.9	1.0
<b>Luminance (3)</b>	<b>B_1</b>	<b>B_2</b>	<b>B_3</b>	<b>B_4</b>	<b>B_5</b>	<b>B_6</b>	<b>B_7</b>	<b>B_8</b>	<b>B_9</b>	<b>B_10</b>
<b>B_1</b>	1.0	1.0	1.0	0.9	0.9	0.9	0.9	0.9	1.0	1.0
<b>B_2</b>	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9
<b>B_3</b>	1.0	1.0	1.0	0.9	0.9	0.8	0.9	0.8	1.0	1.0
<b>B_4</b>	0.9	1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.9
<b>B_5</b>	0.9	1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.9
<b>B_6</b>	0.9	1.0	0.8	1.0	1.0	1.0	1.0	1.0	0.9	0.8
<b>B_7</b>	0.9	1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.9
<b>B_8</b>	0.9	1.0	0.8	1.0	1.0	1.0	1.0	1.0	0.9	0.8
<b>B_9</b>	1.0	1.0	1.0	0.9	0.9	0.9	0.9	0.9	1.0	1.0
<b>B_10</b>	1.0	0.9	1.0	0.9	0.9	0.8	0.9	0.8	1.0	1.0

Table 5.2 shows the scores of the second component of UIQI, luminance distortion, which measures the similarity between the mean luminance of two given images, X and Y. Again, Table 5.2 is organised in 3 parts as in the case of Table 5.1. This metric has a dynamic range



of [0,1], and the ideal value is 1. In the Top section, only when G\_6 or G-8 are used references, half of the Bad images score  $< 0.7$ . while in the middle table, there are a few scores  $\leq 0.7$ , and all scores in the bottom table  $\geq 0.8$ . Accordingly, luminance distortion is not a reliable full reference IQA metric for US images and is not in alignment with our radiologist's subjective assessment.

**Table 5.3 The contrast distortion measure between pairs of Good/Bad images.**

<b>Contrast (1)</b>	<b>G_1</b>	<b>G_2</b>	<b>G_3</b>	<b>G_4</b>	<b>G_5</b>	<b>G_6</b>	<b>G_7</b>	<b>G_8</b>	<b>G_9</b>	<b>G_10</b>
<b>B_1</b>	0.6	0.8	1.0	0.8	1.0	0.8	1.0	1.0	0.9	1.0
<b>B_2</b>	0.7	0.8	1.0	0.8	1.0	0.8	1.0	1.0	0.9	1.0
<b>B_3</b>	0.9	1.0	0.9	1.0	0.9	1.0	0.9	0.9	1.0	0.9
<b>B_4</b>	1.0	1.0	0.9	1.0	0.9	1.0	0.8	0.9	1.0	0.9
<b>B_5</b>	0.9	1.0	0.9	1.0	0.9	1.0	0.9	0.9	1.0	0.9
<b>B_6</b>	0.9	1.0	0.9	1.0	0.9	1.0	0.9	0.9	1.0	0.9
<b>B_7</b>	0.7	0.9	1.0	0.8	1.0	0.8	1.0	1.0	0.9	1.0
<b>B_8</b>	0.7	0.9	1.0	0.8	1.0	0.9	1.0	1.0	1.0	1.0
<b>B_9</b>	0.8	0.9	1.0	0.9	1.0	0.9	1.0	1.0	1.0	1.0
<b>B_10</b>	0.4	0.6	0.8	0.5	0.8	0.5	0.8	0.8	0.7	0.8
<b>Contrast (2)</b>	<b>G_1</b>	<b>G_2</b>	<b>G_3</b>	<b>G_4</b>	<b>G_5</b>	<b>G_6</b>	<b>G_7</b>	<b>G_8</b>	<b>G_9</b>	<b>G_10</b>
<b>G_1</b>	1.0	0.9	0.8	1.0	0.7	1.0	0.7	0.8	0.9	0.8
<b>G_2</b>	0.9	1.0	0.9	1.0	0.9	1.0	0.9	0.9	1.0	0.9
<b>G_3</b>	0.8	0.9	1.0	0.9	1.0	0.9	1.0	1.0	1.0	1.0
<b>G_4</b>	1.0	1.0	0.9	1.0	0.8	1.0	0.8	0.9	0.9	0.9
<b>G_5</b>	0.7	0.9	1.0	0.8	1.0	0.9	1.0	1.0	1.0	1.0
<b>G_6</b>	1.0	1.0	0.9	1.0	0.9	1.0	0.8	0.9	1.0	0.9
<b>G_7</b>	0.7	0.9	1.0	0.8	1.0	0.8	1.0	1.0	1.0	1.0
<b>G_8</b>	0.8	0.9	1.0	0.9	1.0	0.9	1.0	1.0	1.0	1.0
<b>G_9</b>	0.9	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0
<b>G_10</b>	0.8	0.9	1.0	0.9	1.0	0.9	1.0	1.0	1.0	1.0
<b>Contrast (3)</b>	<b>B_1</b>	<b>B_2</b>	<b>B_3</b>	<b>B_4</b>	<b>B_5</b>	<b>B_6</b>	<b>B_7</b>	<b>B_8</b>	<b>B_9</b>	<b>B_10</b>
<b>B_1</b>	1.0	1.0	0.8	0.8	0.8	0.8	1.0	1.0	1.0	0.9
<b>B_2</b>	1.0	1.0	0.8	0.8	0.8	0.8	1.0	1.0	1.0	0.9
<b>B_3</b>	0.8	0.8	1.0	1.0	1.0	1.0	0.9	0.9	0.9	0.6
<b>B_4</b>	0.8	0.8	1.0	1.0	1.0	1.0	0.8	0.9	0.9	0.5
<b>B_5</b>	0.8	0.8	1.0	1.0	1.0	1.0	0.8	0.9	0.9	0.6
<b>B_6</b>	0.8	0.8	1.0	1.0	1.0	1.0	0.9	0.9	0.9	0.6
<b>B_7</b>	1.0	1.0	0.9	0.8	0.8	0.9	1.0	1.0	1.0	0.9
<b>B_8</b>	1.0	1.0	0.9	0.9	0.9	0.9	1.0	1.0	1.0	0.8
<b>B_9</b>	1.0	1.0	0.9	0.9	0.9	0.9	1.0	1.0	1.0	0.8
<b>B_10</b>	0.9	0.9	0.6	0.5	0.6	0.6	0.9	0.8	0.8	1.0

Table 5.3 depicts the results of the contrast distortion component of UIQI. It is also organised in 3 parts, as in the case of Table 5.1. The dynamic range of this metric is [0,1], where a score nearer to 1 is a good value. Top section scores show that all, except B\_10, score high

when any of the good images are used for reference. Almost all scores in the other two tables  $> 0.7$ . Hence, contrast distortion scores do not align with our radiologist's subjective decision.

The UIQI value for a pair of images is obtained by multiplying the three quality factors mentioned earlier, as explained in Chapter 3. Table 5.4 shows the computed UIQI values organised as before into 3 sections according to the labelling pairs.

**Table 5.4 The computed UIQI measure between pairs of Good/Bad images.**

UIQI (1)	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_10
B_1	0.2	0.2	0.2	0.0	-0.2	0.3	0.2	0.4	0.6	-0.1
B_2	0.1	0.1	0.3	-0.1	0.0	0.3	0.1	0.3	0.3	0.0
B_3	0.0	0.2	0.2	-0.2	0.1	0.3	0.0	0.3	0.2	0.0
B_4	0.0	-0.1	0.1	-0.1	0.2	0.0	0.1	0.0	-0.1	-0.1
B_5	-0.3	-0.1	0.0	-0.1	0.2	-0.1	0.0	-0.2	-0.6	0.0
B_6	0.1	0.1	-0.1	-0.1	0.1	0.0	-0.1	0.0	0.2	-0.1
B_7	0.1	0.1	0.0	-0.1	0.0	0.1	0.0	0.1	0.1	-0.1
B_8	0.2	0.1	0.2	-0.1	0.1	0.2	0.2	0.2	0.5	0.0
B_9	0.4	0.1	0.0	-0.1	-0.3	0.3	0.0	0.4	0.8	-0.1
B_10	0.1	0.0	0.2	-0.1	0.1	0.2	0.0	0.1	0.1	-0.1
UIQI (2)	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_10
G_1	1.0	0.0	0.0	0.0	0.0	0.2	0.1	0.2	0.4	-0.2
G_2	0.0	1.0	0.1	0.1	-0.1	0.0	0.0	0.1	0.2	0.1
G_3	0.0	0.1	1.0	0.0	0.0	0.2	0.2	0.1	0.1	0.1
G_4	0.0	0.1	0.0	1.0	-0.1	-0.1	0.0	0.0	-0.1	0.2
G_5	0.0	-0.1	0.0	-0.1	1.0	0.0	0.2	-0.1	-0.2	-0.2
G_6	0.2	0.0	0.2	-0.1	0.0	1.0	0.1	0.5	0.3	-0.1
G_7	0.1	0.0	0.2	0.0	0.2	0.1	1.0	0.1	0.0	0.0
G_8	0.2	0.1	0.1	0.0	-0.1	0.5	0.1	1.0	0.3	-0.1
G_9	0.4	0.2	0.1	-0.1	-0.2	0.3	0.0	0.3	1.0	-0.1
G_10	-0.2	0.1	0.1	0.2	-0.2	-0.1	0.0	-0.1	-0.1	1.0
UIQI (3)	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_10
B_1	1.0	0.4	0.2	-0.1	-0.4	0.0	0.2	0.5	0.7	0.2
B_2	0.4	1.0	0.4	0.1	-0.1	-0.1	0.1	0.3	0.3	0.3
B_3	0.2	0.4	1.0	0.2	0.1	0.1	0.1	0.2	0.1	0.1
B_4	-0.1	0.1	0.2	1.0	0.3	-0.1	-0.1	0.1	-0.2	0.1
B_5	-0.4	-0.1	0.1	0.3	1.0	-0.1	0.0	-0.3	-0.5	0.0
B_6	0.0	-0.1	0.1	-0.1	-0.1	1.0	0.2	0.1	0.2	0.0
B_7	0.2	0.1	0.1	-0.1	0.0	0.2	1.0	0.1	0.1	0.1
B_8	0.5	0.3	0.2	0.1	-0.3	0.1	0.1	1.0	0.5	0.2
B_9	0.7	0.3	0.1	-0.2	-0.5	0.2	0.1	0.5	1.0	0.1
B_10	0.2	0.3	0.1	0.1	0.0	0.0	0.1	0.2	0.1	1.0

This metric has a dynamic range of  $[-1,1]$ , with 1 being the best value. All the scores for any pair of distinct images are nearer to 0, with the highest score of 0.7 only obtained twice in

the bottom table. We conclude that the UIQI is the worst full reference IQA in terms of alignment with our radiologist's subjective decision.

Before we close this section, we conducted experiments to score the quality of the 20 BUS tumour scan images in the pilot dataset using the 3 NRIQA (BRISQUE, NIQE, and PIQE) schemes. Recall that these images were recorded in different clinical centres and labelled as Good and Bad for clinical classification purposes.

**Table 5.5 The three no-reference quality scores of the images (Good and Bad).**

	<b>BRISQUE</b>	<b>NIQE</b>	<b>PIQE</b>
<b>G_1</b>	<b>38.0</b>	<b>18.9</b>	<b>48.6</b>
<b>G_2</b>	<b>37.6</b>	<b>18.9</b>	<b>20.2</b>
<b>G_3</b>	<b>42.8</b>	<b>18.9</b>	<b>26.3</b>
<b>G_4</b>	<b>43.2</b>	<b>18.9</b>	<b>60.5</b>
<b>G_5</b>	<b>36.5</b>	<b>18.9</b>	<b>29.9</b>
<b>G_6</b>	<b>43.0</b>	<b>18.9</b>	<b>53.3</b>
<b>G_7</b>	<b>29.6</b>	<b>18.9</b>	<b>33.1</b>
<b>G_8</b>	<b>41.4</b>	<b>18.9</b>	<b>44.7</b>
<b>G_9</b>	<b>46.9</b>	<b>18.9</b>	<b>51.0</b>
<b>G_10</b>	<b>42.5</b>	<b>18.9</b>	<b>37.7</b>
<b>B_1</b>	<b>47.4</b>	<b>18.9</b>	<b>79.7</b>
<b>B_2</b>	<b>48.0</b>	<b>18.9</b>	<b>80.9</b>
<b>B_3</b>	<b>21.0</b>	<b>18.9</b>	<b>31.9</b>
<b>B_4</b>	<b>16.4</b>	<b>18.9</b>	<b>33.5</b>
<b>B_5</b>	<b>16.8</b>	<b>18.9</b>	<b>35.2</b>
<b>B_6</b>	<b>15.1</b>	<b>18.9</b>	<b>31.9</b>
<b>B_7</b>	<b>30.7</b>	<b>18.9</b>	<b>15.8</b>
<b>B_8</b>	<b>53.2</b>	<b>18.9</b>	<b>75.3</b>
<b>B_9</b>	<b>45.7</b>	<b>18.9</b>	<b>75.5</b>
<b>B_10</b>	<b>46.6</b>	<b>18.9</b>	<b>38.5</b>

None of these criteria applies to our pilot cases, i.e., none is aligned with the subjective assessment of our radiology expert. These results also show that none of the 3 no-reference IQA descriptors proposed for natural images is suitable for reflecting the subjective evaluation of our radiologist in labelling the 20 US images in terms of Good and Bad.

## 5.2 No-reference Multi-Characteristic Image Quality Vector (MCIQ)

The insightful results and observations discussed in the last section make the task of US-IQA very challenging and motivated us to think of designing a kind of no-reference US quality feature vector descriptor rather than a single score descriptor. In Chapter 3, we noted that the visual examination of US images shows that neither luminance nor contrast is uniformly distributed across different parts of the same image. Such image characteristics are also observed in face images generated by morphing attacks, whereby the artefacts generated by such techniques occur in certain areas of the face (e.g., nose and eyes) [121]. To some extent, these observations are analogous to the way loss of naturalness in natural images was dealt with in [117] using the NSS model, which also illustrated this effect in a seemingly natural image constructed by mapping different natural scene images into different blocks. These effects in the case of US images may reflect variation in the scanned tissue texture and layout beside the characteristics of the US device prob (transducer).

It is well known that US images are intrinsically associated with the presence of speckle noise, which can adversely impact both image luminance and contrast, and thereby compromises the diagnostic potentials of US imaging. Empirical evidence has revealed that the impact of speckle noise on US images varies across different regions, contingent upon the firmness (i.e., solidity) of the tissue subregions [6]. Therefore, this is a plausible factor in creating the observed non-uniformity of luminance and contrast in US images. Removing speckle noise is considered an essential precondition for any tissue characterization employing US imaging [122]. To this end, an adaptive speckle US denoising strategy was formulated, which was demonstrated to markedly enhance the performance of ML models applied to Ovarian cancer [6]. Adaptiveness exploited the non-uniform spatial distribution of tissue solidity characteristics which were determined from the US scan images by elaborating conditions on block-based intensity *Skewness* and *Kurtosis* values.

The above discussion provides a plausible idea of designing a no-reference assessment image quality feature vector for US image analysis by leveraging the spatial distribution of different *quality* attributes. These quality attributes will nevertheless benefit from the Full reference IQA scheme UIQI defined for natural images. We note that Islam et al. [94] supplemented the UIQI with an additional shape factor computed by the so-called *modified skewness*. Avoiding the use of the natural image NRIQA scores of BRISQUE, NIQE, and PIQUE stems from our earlier observed differences between natural images and US images with respect to quantity and distribution of various types of texture features.

## The MCIQ Proposal

Our proposed no-reference quality feature vector scheme relies on using the three components of UIQI but is supplemented by additional statistical quality metrics determined by *Skewness* and *Kurtosis*. Accordingly, our innovative NRIQA scheme for US images outputs a *spatio-statistical* feature vector, denoted by *MCIQ*, that integrates Correlation, Luminance, Contrast, Skewness, and Kurtosis image quality measures.

To define the five components of the MCIQ vector, first, we state the five related basic statistical parameters. Let  $x = \{x_1, x_2, \dots, x_n\}$  and  $y = \{y_1, y_2, \dots, y_n\}$  be two given real-value grey-pixel sequences representing two equal-size image blocks, then:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad 5.1$$

$$\sigma_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad 5.2$$

$$\sigma_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad 5.3$$

$$s_x = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)(\sigma_x)^3} \quad 5.4$$

$$k_x = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{(n-1)(\sigma_x)^4} \quad 5.5$$

Where  $\bar{x}$  represents the mean of  $x$ ,  $\sigma_x^2$  represent its variance ( $\sigma_x$  is its standard deviation),  $\sigma_{xy}$  stands for the covariance of  $x$  and  $y$ , while  $s_x$  and  $k_x$  represent the skewness and kurtosis of  $x$ , respectively.

The MCIQ feature vector is constructed in the following steps:

**Step 1:** Partitioning the input image (a grey-scale image) into 36 same-size rectangular blocks. See Figure 5.1.

**Step 2:** compute the quality indices of each image block with respect to the other 35 blocks in terms of the 5 selected quality indices (Correlation, Luminance, Contrast, Skewness, and Kurtosis) defined by the following formulas:

$$\mathbf{Correlation}_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad 5.6$$

$$\mathbf{Luminance}_{xy} = \frac{2\bar{x}\bar{y}}{(\bar{x}^2 + \bar{y}^2)} \quad 5.7$$

$$\mathbf{Contrast}_{xy} = \frac{2\sigma_x \sigma_y}{(\sigma_x^2 + \sigma_y^2)} \quad 5.8$$

$$\mathbf{Skewness}_{xy} = \frac{2s_x s_y}{(s_x^2 + s_y^2)} \quad 5.9$$

$$\mathbf{Kurtosis}_{xy} = \frac{2k_x k_y}{(k_x^2 + k_y^2)} \quad 5.10$$

**Step 3:** Arrange each of the computed 5 quality indices in a 36×36 symmetric matrix with a unit diagonal. The upper right 630 = (36×35)/2 values above the diagonal represent the spatial distribution of the corresponding quality index between the partitioned image block pairs. We then quantize these 630 indices into 10 equal bins for each quality index and produce the 10-bin histogram vector.

**Step 4:** Finally, the 50-dimensional MCIQ feature vector is constructed by concatenating histogram vectors for 5 quality indices in the order: [Correlation, Luminance, Contrast, Skewness, Kurtosis].

Figure 5.1 below displays the steps of constructing the MCIQ feature vector for a BUS tumour RoI image.

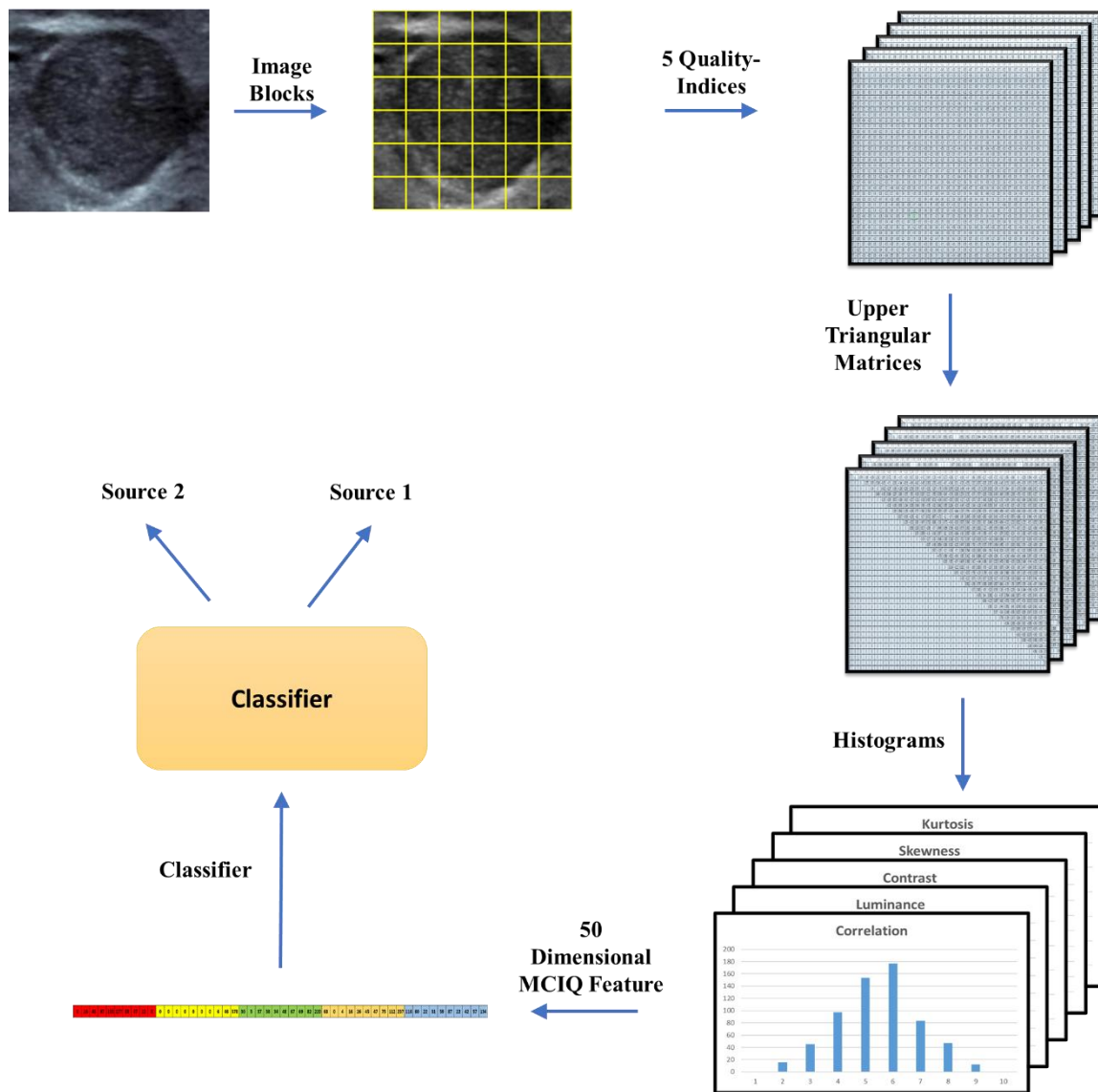


Figure 5.1 Process of building and extracting MCIQ feature vector.

In the following 2 sections, we present the results of various experiments to test the discriminating power of MCIQ between different US image datasets in relation to different purposes.

### 5.3 Explaining CNN Generalisation Results by MCIQ Feature Vectors

Our initially stated purpose of investigating US image quality descriptors was in relation to the significant disparities of DL generalization performance as a result of training with different datasets. Having developed the no-reference quality feature vector MCIQ, we conducted several experiments to test its effectiveness in distinguishing the quality descriptor of the four BUS datasets (Renmin, Modelling, Test1, and BUSI) used in the previous chapter.

To ensure that all image blocks in the MCIQ procedure contain actual tissue pixel values and no image block is affected by zero-padding, the TumourT RoI cropping scenario was selected as a standard, and the lesion RoIs were resized to 128x128. The 5-fold cross-validation training/testing protocol was adopted in all experiments. Various classifiers were utilized, including SVM with linear kernel and kNN with  $k = 1, 3,$  and  $5$ . The degree of classification performance, as determined by using the MCIQ quality components, serves as proximity between images in training and testing datasets. Low-performance rates indicate the better proximity of the training and testing datasets in terms of the MCIQ quality feature vector, and in this case, we expect reasonable generalisation performance.

### 5.3.1 MCIQ - Generalisation Association: (Renmin, Modelling) vs. Test1

The experiments in this subsection were designed to compare the training datasets, specifically Renmin and the Modelling datasets, with the Test1 dataset using the MCIQ feature vector. Recall that in the last chapter, we observed a failure to generalise the CNN models when training with the Renmin dataset and testing on the external Test1 dataset, in contrast to training with the Modelling dataset. The results of these two experiments are presented in Table 5.6 below.

Table 5.6 Quality inspection of Renmin/Modelling vs. Test1 using MCIQ.

Renmin vs. Test1	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	<b>0.69 <math>\pm</math> 0.04</b>	<b>0.59 <math>\pm</math> 0.11</b>	<b>0.74 <math>\pm</math> 0.05</b>	<b>0.58 <math>\pm</math> 0.07</b>	<b>0.67 <math>\pm</math> 0.05</b>
kNN, k=3	<b>0.70 <math>\pm</math> 0.02</b>	<b>0.55 <math>\pm</math> 0.07</b>	<b>0.79 <math>\pm</math> 0.03</b>	<b>0.57 <math>\pm</math> 0.04</b>	<b>0.67 <math>\pm</math> 0.03</b>
kNN, k=5	<b>0.70 <math>\pm</math> 0.02</b>	<b>0.55 <math>\pm</math> 0.06</b>	<b>0.79 <math>\pm</math> 0.03</b>	<b>0.58 <math>\pm</math> 0.04</b>	<b>0.67 <math>\pm</math> 0.02</b>
Linear-SVM	<b>0.73 <math>\pm</math> 0.06</b>	<b>0.52 <math>\pm</math> 0.08</b>	<b>0.86 <math>\pm</math> 0.05</b>	<b>0.59 <math>\pm</math> 0.08</b>	<b>0.69 <math>\pm</math> 0.06</b>
Modelling vs. Test1	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	<b>0.57 <math>\pm</math> 0.03</b>	<b>0.44 <math>\pm</math> 0.10</b>	<b>0.64 <math>\pm</math> 0.04</b>	<b>0.42 <math>\pm</math> 0.07</b>	<b>0.54 <math>\pm</math> 0.04</b>
kNN, k=3	<b>0.57 <math>\pm</math> 0.03</b>	<b>0.42 <math>\pm</math> 0.04</b>	<b>0.65 <math>\pm</math> 0.05</b>	<b>0.42 <math>\pm</math> 0.03</b>	<b>0.54 <math>\pm</math> 0.03</b>
kNN, k=5	<b>0.59 <math>\pm</math> 0.04</b>	<b>0.39 <math>\pm</math> 0.06</b>	<b>0.70 <math>\pm</math> 0.05</b>	<b>0.41 <math>\pm</math> 0.06</b>	<b>0.54 <math>\pm</math> 0.04</b>
Linear-SVM	<b>0.64 <math>\pm</math> 0.02</b>	<b>0.22 <math>\pm</math> 0.06</b>	<b>0.89 <math>\pm</math> 0.04</b>	<b>0.30 <math>\pm</math> 0.06</b>	<b>0.56 <math>\pm</math> 0.02</b>

These results reveal that the accuracy of kNNs and SVM classifiers in distinguishing between Renmin and Test1 datasets ranges from 69% to 73%, suggesting that these datasets differ significantly in terms of MCIQ quality components. Thus, these results explain, to a reasonable extent, the failure of CNNs model generalisation when trained on Renmin and tested on Test1. In contrast, comparing the Modelling dataset to Test1, kNNs and SVM classifiers of MCIQ descriptor exhibit lower accuracy in the range of 57% to 64%, i.e., the



Modelling dataset quality is more similar to (and less separable from) Test1. Therefore, DL models are more likely to generalize better when trained on the Modelling dataset and tested on Test1. The experimental findings in Table 5.5 are consistent with the DL generalization results presented in Chapter 4, Table 4.2 and 4.4, where DL models trained on the Modelling dataset exhibit high to excellent accuracy when tested on Test1 compared to models trained on the Renmin dataset.

### 5.3.2 MCIQ - Generalization Association: (Renmin, Modelling) vs. BUSI

We repeated similar experiments to those conducted in the previous subsection, but we used the external BUSI dataset for testing this time. Table 5.7, below, shows that the kNNs and SVM classifiers' accuracy in distinguishing between Renmin and BUSI datasets ranges from 66% to 75%, indicating that these 2 datasets differ in terms of MCIQ. Moreover, the dissimilarity level between Renmin and BUSI datasets is somewhat similar to that between Renmin and Test1. However, Tables 4.2 and 4.3, in Chapter 4, revealed that DL models trained on the Renmin dataset achieve higher generalization performance on the BUSI dataset compared to their performance on Test1. This is possibly due to other factors relating to image acquisition procedures, US machines, and the level of radiologists' experience.

In contrast, comparing Modelling and BUSI datasets, the kNNs and SVM classifiers exhibit accuracy in the range of 68% to 77%, suggesting that the Modelling dataset MCIQ differs more from BUSI than from Test1. Hence, DL models trained on the Modelling dataset are more likely to generalize slightly lower when tested on BUSI. This is consistent with the results in Chapter 4, Tables 4.4 and 4.5, which showed that training with Modelling had higher performance on the Test1 dataset than the BUSI dataset.

Table 5.7 Quality inspection of Renmin/Modelling vs. BUSI using MCIQ.

Renmin vs. BUSI	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
kNN, k=1	0.66 ± 0.03	0.67 ± 0.04	0.66 ± 0.02	0.66 ± 0.03	0.66 ± 0.03
kNN, k=3	0.67 ± 0.02	0.65 ± 0.04	0.69 ± 0.03	0.67 ± 0.03	0.67 ± 0.02
kNN, k=5	0.69 ± 0.03	0.64 ± 0.03	0.73 ± 0.04	0.67 ± 0.03	0.69 ± 0.03
Linear-SVM	0.75 ± 0.01	0.71 ± 0.04	0.79 ± 0.03	0.74 ± 0.02	0.75 ± 0.01
Modelling vs. BUSI	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
kNN, k=1	0.68 ± 0.03	0.69 ± 0.06	0.68 ± 0.06	0.68 ± 0.03	0.68 ± 0.03
kNN, k=3	0.71 ± 0.03	0.69 ± 0.06	0.72 ± 0.04	0.70 ± 0.03	0.71 ± 0.03
kNN, k=5	0.72 ± 0.03	0.69 ± 0.06	0.75 ± 0.06	0.71 ± 0.04	0.72 ± 0.03
Linear-SVM	0.77 ± 0.02	0.75 ± 0.05	0.78 ± 0.06	0.76 ± 0.02	0.77 ± 0.02

### 5.3.3 MCIQ Separability Between Images in Renmin and Modelling Datasets

We close this section by investigating the discriminating power of MCIQ for the Renmin vs. the larger Modelling datasets. This experiment was conducted in a balanced way with the expectation of providing a further explanation as to why training DL models on these two datasets separately have different generalisation behaviours when tested on external testing datasets. In line with the previous two sections, lower MCIQ performance indicates that the two datasets are closer to each other and vice versa.

Table 5.8 Quality inspection of Renmin vs. Modelling using MCIQ.

Renmin vs. Modelling	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	<b>0.38 <math>\pm</math> 0.04</b>	<b>0.35 <math>\pm</math> 0.05</b>	<b>0.40 <math>\pm</math> 0.05</b>	<b>0.36 <math>\pm</math> 0.04</b>	<b>0.38 <math>\pm</math> 0.04</b>
kNN, k=3	<b>0.47 <math>\pm</math> 0.04</b>	<b>0.47 <math>\pm</math> 0.08</b>	<b>0.48 <math>\pm</math> 0.03</b>	<b>0.47 <math>\pm</math> 0.06</b>	<b>0.47 <math>\pm</math> 0.04</b>
kNN, k=5	<b>0.51 <math>\pm</math> 0.03</b>	<b>0.54 <math>\pm</math> 0.04</b>	<b>0.48 <math>\pm</math> 0.05</b>	<b>0.52 <math>\pm</math> 0.03</b>	<b>0.51 <math>\pm</math> 0.03</b>
Linear-SVM	<b>0.60 <math>\pm</math> 0.05</b>	<b>0.51 <math>\pm</math> 0.07</b>	<b>0.70 <math>\pm</math> 0.06</b>	<b>0.56 <math>\pm</math> 0.06</b>	<b>0.60 <math>\pm</math> 0.05</b>

These results show that the two datasets are not easily separable by MCIQ. Unlike the results in Tables 5.6 and 5.7, there is a significant difference between the SVM classifier performance and that of the kNN classifier, especially when  $k = 1$ . The results of the kNN methods indicate that the MCIQ vectors for these two datasets are spread out and clustered near each other for a significant number of images. On the other hand, an accuracy of 60% obtained with a linear SVM classifier means that many of their clusters for the two datasets are separated on different sides of the SVM hyperplane. Note that the much larger size of the modelling dataset results in a wider amount of spreading of their MCIQ values.

In conclusion, the results of all experiments in this section indicate that the discrepancy in generalisation results can be attributed to the combined effects of MCIQ feature vectors as well as the big difference in the size (and diversity of clinical practises) of the two training datasets. Perhaps more investigations and refinements/extensions of MCIQ are needed.

### 5.4 Discriminating Power of MCIQ for Other Purposes

Besides the failure/success of CNN model generalisation, we shall now investigate the discriminating power of the proposed MCIQ feature vector. These investigations are concerned with questions: (1) Can MCIQ discriminate Benign from Malignant masses for samples recorded in the same centre? (2) Can MCIQ distinguish between US images and other medical image modalities used for scanning the same tissue? and (3) Can MCIQ distinguish BUS datasets from US datasets of other tissue types or Natural Images?

### 5.4.1 MCIQ - Tumour Class Association: Renmin (Benign vs. Malignant)

Here, we explore the potential of the MCIQ feature vector as a HC feature for classifying benign and malignant cases for the Renmin dataset. We also compare the results with the performance of a recent NRIQA feature developed by Dey et al. [120]. We denoted this scheme as *TripleIQA+Entropy* since it creates a quality feature vector descriptor for the classification of Benign vs. Malignant from BUS images by concatenating the IQA scores of BRISQUE, NIQE, and PIQUE plus Shannon image Entropy value. They reported a reasonably significant accuracy of (85.4% with kNN and 91.2% with SVM) trained and tested on the BUSI dataset. Tables 5.9 and 5.10 display the results of the experiments conducted for the MCIQ and TripleIQA+Entropy for the Renmin dataset, respectively.

**Table 5.9 MCIQ classification performance of Benign vs. Malignant (Renmin dataset).**

Benign vs. Malignant	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	<b>0.76 <math>\pm</math> 0.02</b>	<b>0.76 <math>\pm</math> 0.03</b>	<b>0.76 <math>\pm</math> 0.07</b>	<b>0.76 <math>\pm</math> 0.01</b>	<b>0.76 <math>\pm</math> 0.02</b>
kNN, k=3	<b>0.80 <math>\pm</math> 0.03</b>	<b>0.80 <math>\pm</math> 0.04</b>	<b>0.80 <math>\pm</math> 0.08</b>	<b>0.80 <math>\pm</math> 0.03</b>	<b>0.80 <math>\pm</math> 0.03</b>
kNN, k=5	<b>0.81 <math>\pm</math> 0.05</b>	<b>0.79 <math>\pm</math> 0.05</b>	<b>0.82 <math>\pm</math> 0.08</b>	<b>0.80 <math>\pm</math> 0.05</b>	<b>0.81 <math>\pm</math> 0.05</b>
Linear-SVM	<b>0.84 <math>\pm</math> 0.02</b>	<b>0.86 <math>\pm</math> 0.05</b>	<b>0.82 <math>\pm</math> 0.03</b>	<b>0.85 <math>\pm</math> 0.02</b>	<b>0.84 <math>\pm</math> 0.02</b>

**Table 5.10 TripleIQA+Entropy classification of Benign vs. Malignant (Renmin dataset).**

Benign vs. Malignant	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	<b>0.76 <math>\pm</math> 0.03</b>	<b>0.76 <math>\pm</math> 0.06</b>	<b>0.77 <math>\pm</math> 0.07</b>	<b>0.76 <math>\pm</math> 0.03</b>	<b>0.76 <math>\pm</math> 0.03</b>
kNN, k=3	<b>0.80 <math>\pm</math> 0.03</b>	<b>0.80 <math>\pm</math> 0.03</b>	<b>0.80 <math>\pm</math> 0.06</b>	<b>0.80 <math>\pm</math> 0.03</b>	<b>0.80 <math>\pm</math> 0.03</b>
kNN, k=5	<b>0.82 <math>\pm</math> 0.03</b>	<b>0.82 <math>\pm</math> 0.04</b>	<b>0.82 <math>\pm</math> 0.07</b>	<b>0.82 <math>\pm</math> 0.03</b>	<b>0.82 <math>\pm</math> 0.03</b>
Linear-SVM	<b>0.83 <math>\pm</math> 0.03</b>	<b>0.83 <math>\pm</math> 0.04</b>	<b>0.83 <math>\pm</math> 0.06</b>	<b>0.83 <math>\pm</math> 0.03</b>	<b>0.83 <math>\pm</math> 0.03</b>

First, the results in Table 5.9 demonstrate that the MCIQ feature vector is a reliable class-discriminating tool solely based on the spatial distribution of scores computed without reference images. Moreover, the linear SVM can achieve an accuracy of 84% with a good balance between sensitivity and specificity (86% vs. 82%). MCIQ performs as well as the TripleIQA+Entropy scheme when tested on the Renmin database (Table 5.10).

### 5.4.2 MCIQ - Tissue Type Association: Breast vs. Liver/Bladder

The objective of this section is to explore the degree of distinguishability between US scanning of different tissue types. The proximity of two tissues in terms of their quality attributes, such as breast tissues vs. liver or bladder tissues, can impact the potential success of DL model generalization if a model is trained on one tissue and tested on another. First, we briefly define the two adopted Liver and Bladder US datasets.

The Liver and Bladder US datasets comprise 420 and 177 images, respectively. These images were collected at Pudong New District Renmin Hospital in Shanghai, China, and comprise cases of both benign and malignant lesions (Liver: 268 benign and 152 malignant cases), (Bladder: 100 benign and 77 malignant cases). The images were acquired using different US machines. An experienced radiologist performed the labelling of the images, and biopsy tests were conducted to confirm the labels. Furthermore, the radiologists manually marked numerous boundary points for each lesion to enable accurate detection of the shape and location of the lesions, which serves as an alternative to automatic detection and segmentation.

Table 5.11 presents the outcomes of utilizing kNNs and SVM classifiers to differentiate breast tissues from liver and bladder tissues based on the MCIQ feature vector. The findings reveal that breast tissue can be distinguished from both liver and bladder tissues with varying levels of accuracy, ranging from 71% to 80% for Breast vs. Liver and ranging from 79% to 85% for Breast vs. Bladder. These outcomes indicate that, generally, breast tissue is more separable from Bladder tissue than from liver tissue in terms of MCIQ quality components.

**Table 5.11 MCIQ classification performance of US Breast Tissue (Renmin) vs. US Liver/Bladder Tissue.**

Breast vs. Liver	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
kNN, k=1	0.71 ± 0.02	0.64 ± 0.05	0.76 ± 0.06	0.66 ± 0.03	0.70 ± 0.02
kNN, k=3	0.75 ± 0.03	0.64 ± 0.03	0.84 ± 0.04	0.70 ± 0.03	0.74 ± 0.03
kNN, k=5	0.77 ± 0.03	0.65 ± 0.03	0.86 ± 0.05	0.71 ± 0.03	0.76 ± 0.02
Linear-SVM	0.80 ± 0.02	0.72 ± 0.03	0.87 ± 0.02	0.76 ± 0.02	0.79 ± 0.02
Breast vs. Bladder	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
kNN, k=1	0.79 ± 0.05	0.46 ± 0.12	0.91 ± 0.04	0.53 ± 0.11	0.68 ± 0.07
kNN, k=3	0.84 ± 0.02	0.47 ± 0.06	0.96 ± 0.02	0.59 ± 0.05	0.72 ± 0.03
kNN, k=5	0.83 ± 0.02	0.43 ± 0.08	0.97 ± 0.01	0.56 ± 0.08	0.70 ± 0.04
Linear-SVM	0.85 ± 0.01	0.50 ± 0.03	0.97 ± 0.02	0.63 ± 0.02	0.74 ± 0.01

### 5.4.3 MCIQ - Image Modality Association (Breast Lesion): US vs. Mammogram

Here, we explore the degree of distinguishability between different image modalities for the same tissue type, i.e., Breast Lesion: US vs. Mammogram. The proximity of two different image modalities in terms of their MCIQ quality descriptor can impact the potential success of designing generic DL models trained on a given image modality tested on another.

The widely used mammogram dataset (known as Digital Database for Screening Mammography (DDSM) [123]) is utilized to test against our BUS Renmin dataset. DDSM constitutes 2620 mammograms in total, in which 559 mammograms were randomly selected in our experiments with 302 normal cases and 257 abnormal cases. The results of our MCIQ discrimination between the two datasets are shown in Figure 5.12 below.

Table 5.12 MCIQ modality association for breast lesion: US (Renmin) vs. Mammogram (DDSM).

US vs. Mammogram	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	<b>0.85 <math>\pm</math> 0.01</b>	<b>0.87 <math>\pm</math> 0.03</b>	<b>0.84 <math>\pm</math> 0.04</b>	<b>0.86 <math>\pm</math> 0.01</b>	<b>0.85 <math>\pm</math> 0.01</b>
kNN, k=3	<b>0.88 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.03</b>	<b>0.85 <math>\pm</math> 0.04</b>	<b>0.88 <math>\pm</math> 0.02</b>	<b>0.88 <math>\pm</math> 0.02</b>
kNN, k=5	<b>0.88 <math>\pm</math> 0.01</b>	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.85 <math>\pm</math> 0.03</b>	<b>0.89 <math>\pm</math> 0.01</b>	<b>0.88 <math>\pm</math> 0.01</b>
Linear-SVM	<b>0.89 <math>\pm</math> 0.01</b>	<b>0.89 <math>\pm</math> 0.03</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.01</b>

It can be observed that the performance of the classifiers varies with different values of k and the type of classifier used. The highest accuracy achieved was  $0.89 \pm 0.01$  using the Linear SVM classifier. The results show that increasing the value of k for the kNN classifier improves the classifier's performance in terms of accuracy, sensitivity, specificity, F1-score, and AUC. However, the performance improvement was insignificant for k values greater than 3. The Linear SVM classifier consistently outperformed the kNN classifier in terms of accuracy and other performance measures, indicating that the linear SVM classifier is a more effective classifier for this particular dataset.

The results demonstrate that the spatial distribution of the MCIQ components in these two image datasets is significantly different, and any of the chosen classifiers can successfully separate them. This implies that DL models trained on US images may not be directly generalized to mammogram images without additional training.

#### 5.4.4 MCIQ Domain Association: US Breast Tissue vs. Face Images

This section aims to examine how different US tissue scan images are from Natural images in terms of the distribution of MCIQ components. To this end, we utilize the Renmin US breast dataset and a commonly used natural image dataset of face photos. These two datasets belong to distinct domains, namely, the medical and natural images domains. First, we briefly describe the face image dataset.

We merged two well-known face image databases, usually used in relation to biometric-based face recognition, to create one dataset of 169 genuine face images. The contributing Databases are (1) the AMSL face dataset (102 images), which is available online, free upon request [124], and (2) the Utrecht face dataset (67 images) [125].

Table 5.13 reports the performance of kNNs and SVM classifiers in distinguishing the Renmin dataset from the face dataset based on the MCIQ feature vector extracted from all images. The outcomes reveal that the breast RoIs can be easily distinguished from face images, achieving a classification accuracy of 100% with balanced specificity and sensitivity rated. The results demonstrate that the spatial distribution of the MCIQ components in these two different modality image datasets are significantly different, and any of the chosen classifiers can successfully separate them.

Table 5.13 MCIQ classification performance of US (Renmin) vs. Face dataset.

US vs. Face Images	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00	0.99 $\pm$ 0.01	1.00 $\pm$ 0.00	1.00 $\pm$ 0.01
kNN, k=3	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00
kNN, k=5	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00
Linear-SVM	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00

#### 5.5 Limitations of MCIQ and Potential Remedies

Any unbiased assessment of experimental results presented in the last two sections should find that while the MCIQ had a satisfactory success in partially explaining the lack of generalisation of CNN models trained with a relatively small set of US images recorded in a single clinical centre, it was more successful in distinguishing image datasets recorded for different purposes using different image modalities and/or tissues. In considering the pros and cons of the adopted approach in developing the no-reference image quality MCIQ feature vectors, instead of a single score, one has to remember the main obstacle of US image scarcity besides the absence of established knowledge on the sort of distortions (other than

Speckle noise) in US images as well as the difficulty in getting expert assessment/labelling of US image quality. Here, we shall discuss the link between MCIQ performance with that of expert quality assessment and consider ways of overcoming some of MCIQ's limitations.

### 5.5.1 MCIQ for the Pilot US dataset (Good vs. Bad)

To investigate ways of overcoming some of MCIQ's limitations, we first revisit the pilot 20 - samples BUS dataset labelled as Good/Bad to assess the level of its limitation, or otherwise, in reflecting the expert assessment. We conducted a simple experiment to determine the performance of MCIQ in separating the 20 labelled US images. The results shown in Table 5.14 confirm that the MCIQ assessment is not aligned well with the expert judgment.

Table 5.14 The pilot dataset, Good vs. Bad using MCIQ.

Good vs. Bad	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	<b>0.60 <math>\pm</math> 0.14</b>	<b>0.60 <math>\pm</math> 0.22</b>	<b>0.60 <math>\pm</math> 0.22</b>	<b>0.59 <math>\pm</math> 0.14</b>	<b>0.60 <math>\pm</math> 0.14</b>
kNN, k=3	<b>0.40 <math>\pm</math> 0.22</b>	<b>0.40 <math>\pm</math> 0.22</b>	<b>0.40 <math>\pm</math> 0.22</b>	<b>0.40 <math>\pm</math> 0.22</b>	<b>0.40 <math>\pm</math> 0.22</b>
kNN, k=5	<b>0.50 <math>\pm</math> 0.40</b>	<b>0.50 <math>\pm</math> 0.35</b>	<b>0.50 <math>\pm</math> 0.50</b>	<b>0.51 <math>\pm</math> 0.37</b>	<b>0.50 <math>\pm</math> 0.40</b>
Linear-SVM	<b>0.55 <math>\pm</math> 0.37</b>	<b>0.60 <math>\pm</math> 0.42</b>	<b>0.50 <math>\pm</math> 0.50</b>	<b>0.57 <math>\pm</math> 0.37</b>	<b>0.55 <math>\pm</math> 0.37</b>

For the sake of comparison, we repeated the above experiment by replacing the MCIQ feature vector with the 4-dimensional score vector of the TripleIQA+Entropy described in subsection 5.4.1. These results in Table 5.15 show that TripleIQA+Entropy significantly outperforms MCIQ and aligns reasonably well with the expert radiology assessment. However, the different classifiers perform differently. While the best performing TripleIQA+Entropy classifier is kNN with k=3, MCIQ with this classifier performance is the lowest, but its highest performing classifier is kNN with k=1.

Table 5.15 The pilot dataset, Good vs. Bad using TripleIQA+Entropy features.

Good vs. Bad	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	<b>0.80 <math>\pm</math> 0.11</b>	<b>0.80 <math>\pm</math> 0.27</b>	<b>0.80 <math>\pm</math> 0.27</b>	<b>0.79 <math>\pm</math> 0.14</b>	<b>0.80 <math>\pm</math> 0.11</b>
kNN, k=3	<b>0.90 <math>\pm</math> 0.14</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.80 <math>\pm</math> 0.27</b>	<b>0.92 <math>\pm</math> 0.11</b>	<b>0.90 <math>\pm</math> 0.14</b>
kNN, k=5	<b>0.70 <math>\pm</math> 0.33</b>	<b>0.80 <math>\pm</math> 0.27</b>	<b>0.60 <math>\pm</math> 0.42</b>	<b>0.74 <math>\pm</math> 0.28</b>	<b>0.70 <math>\pm</math> 0.33</b>
Linear-SVM	<b>0.70 <math>\pm</math> 0.11</b>	<b>0.70 <math>\pm</math> 0.27</b>	<b>0.70 <math>\pm</math> 0.27</b>	<b>0.69 <math>\pm</math> 0.12</b>	<b>0.70 <math>\pm</math> 0.11</b>

Next, we consider modifying the MCIQ to achieve better alignment with the radiologist assessments, even for the small pilot dataset.

### 5.5.2 Modified MCIQs for the Pilot (Good and Bad) US Dataset

The fact that MCIQ is defined by comparing the local block-wise statistical distribution of the grey-level intensities provides different possible ways of modifying MCIQ, including:

1. Extending MCIQ by adding other parameters (i.e., moments) of the grayscale local distribution.
2. Select a spatial image transformation (e.g., LBP transform), compute the transformed pixels' local statistical distribution, and construct the MCIQ-like feature vector.
3. Use an edge detection scheme and extract the slopes of linear edges in each block.

Use local statistical parameters of slopes to construct an MCIQ-like feature vector.

However, improving the performance of these schemes for the pilot small Good/Dad dataset of US images is unrealistic due to the small size of the dataset. Considering the way CNN architecture generates a sufficiently large number of smoothed versions of input images using Gaussian filters, we can extend the pilot dataset of Good/Bad images by convolving the original images with several randomly selected Gaussian filters. We created 2 pilot image datasets by convolving with 6 (and 10) 5x5 Gaussian filters, producing 120 (and 200) convolved images, respectively. For each convolved image, we computed the corresponding MCIQ feature vector. We conducted the same sets of classification experiments. Table 5.16 displays the performance of the MCIQ in distinguishing Good from Bad.

**Table 5.16 Good vs. Bad using MCIQ post 6(10) Gaussian filters convolution.**

Good vs. Bad (120)	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
kNN, k=1	0.78 ± 0.05	0.68 ± 0.07	0.88 ± 0.07	0.76 ± 0.06	0.78 ± 0.05
kNN, k=3	0.83 ± 0.09	0.78 ± 0.15	0.87 ± 0.10	0.81 ± 0.11	0.83 ± 0.09
kNN, k=5	0.83 ± 0.08	0.75 ± 0.12	0.90 ± 0.07	0.81 ± 0.09	0.83 ± 0.08
Linear-SVM	0.80 ± 0.07	0.80 ± 0.10	0.80 ± 0.14	0.80 ± 0.07	0.80 ± 0.07
Good vs. Bad (200)	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
kNN, k=1	0.81 ± 0.05	0.81 ± 0.04	0.81 ± 0.07	0.81 ± 0.05	0.81 ± 0.05
kNN, k=3	0.82 ± 0.05	0.80 ± 0.04	0.83 ± 0.10	0.81 ± 0.05	0.82 ± 0.05
kNN, k=5	0.77 ± 0.07	0.70 ± 0.09	0.83 ± 0.14	0.75 ± 0.07	0.77 ± 0.07
Linear-SVM	0.72 ± 0.06	0.76 ± 0.07	0.67 ± 0.11	0.73 ± 0.05	0.72 ± 0.06



Both experiments resulted in significantly improved performance, but increasing the number of convolution filters did not lead to better results. For the sake of comparison, we repeated the 6 Gaussian convolution filter experiment but using the 4-dimensional TripleIQA+Entropy IQA scheme. The results are shown in Table 5.17 below. Surprisingly, the performance of this scheme deteriorated significantly with all classifiers.

**Table 5.17 Bad vs. Good using TripleIQA+Entropy post 6 Gaussian filters convolution.**

Good vs. Bad (120)	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
kNN, k=1	<b>0.60 ± 0.04</b>	<b>0.57 ± 0.14</b>	<b>0.63 ± 0.10</b>	<b>0.58 ± 0.08</b>	<b>0.60 ± 0.04</b>
kNN, k=3	<b>0.65 ± 0.06</b>	<b>0.67 ± 0.12</b>	<b>0.63 ± 0.14</b>	<b>0.65 ± 0.07</b>	<b>0.65 ± 0.06</b>
kNN, k=5	<b>0.62 ± 0.07</b>	<b>0.63 ± 0.21</b>	<b>0.60 ± 0.14</b>	<b>0.61 ± 0.12</b>	<b>0.62 ± 0.07</b>
Linear-SVM	<b>0.56 ± 0.09</b>	<b>0.58 ± 0.08</b>	<b>0.53 ± 0.24</b>	<b>0.57 ± 0.03</b>	<b>0.56 ± 0.09</b>

While expanding the pilot data set by convolving with Gaussian filters improved the alignment of MCIQ with the expert quality labelling significantly, the fact that increasing the number of Gaussian filters from 6 to 10 did not make much difference raised the question about the possibility of using other types of non-smoothing filters. Based on my own previous experience with the orthonormal Hadamard matrices [76], I considered developing filters based on Hadamard matrices. The fact that convolution filters are of size  $k \times k$  with odd  $k$  values, suggests building filters using block diagonal matrices, each block of which is a Hadamard matrix. The following 6 5x5 Hadamard-based matrices, are formed by block diagonal of 1x1, 2x2 and 4x4 Hadamard matrices:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & -1 & 1 & -1 & 0 \\ 1 & 1 & -1 & -1 & 0 \\ 1 & -1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & -1 & 1 & -1 \\ 0 & 1 & 1 & -1 & -1 \\ 0 & 1 & -1 & -1 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & -1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & -1 & -1 & 0 \\ 1 & -1 & -1 & 1 & 0 \end{bmatrix}, \\
 \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

**Figure 5.2 The 6 5x5 Hadamard-based filters.**

To expand the pilot dataset via convolution with these 5x5 filters, each cropped RoI image in the 20-pilot dataset was convolved with these 6 5x5 Hadamard-based filters and extracted the MCIQ feature vector from the convolved images. We call the expanded dataset the Hadamard-pilot dataset, consisting of  $120 = (60\_Good + 60\_Bad)$  US images. The above classification experiments, with the MCIQ representations of the Hadamard-pilot dataset, were conducted, and the results are shown in Table 5.18 below. The results show the excellent alignment of this modified MCIQ with the expert radiologist quality assessment, especially by the kNN classifier with  $k=1$ , achieving near-optimal discrimination. The other kNN classifiers ( $k = 3, 5$ ) are also doing well but with a wider gap between sensitivity and specificity. These results indicate that the MCIQ feature vectors for both classes are spread out well in  $R^{50}$  but not separated from each other. The performance of MCIQ with SVM confirms this but indicates that many images in the different classes are separated by the SVM hyperplane.

**Table 5.18 Good vs. Bad using MCIQ post 6 Hadamard filter convolution.**

Good vs. Bad (120)	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	<b>0.98 <math>\pm</math> 0.04</b>	<b>0.95 <math>\pm</math> 0.07</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.97 <math>\pm</math> 0.04</b>	<b>0.98 <math>\pm</math> 0.04</b>
kNN, k=3	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.83 <math>\pm</math> 0.06</b>	<b>0.95 <math>\pm</math> 0.05</b>	<b>0.88 <math>\pm</math> 0.03</b>	<b>0.89 <math>\pm</math> 0.02</b>
kNN, k=5	<b>0.87 <math>\pm</math> 0.07</b>	<b>0.80 <math>\pm</math> 0.10</b>	<b>0.93 <math>\pm</math> 0.07</b>	<b>0.86 <math>\pm</math> 0.08</b>	<b>0.87 <math>\pm</math> 0.07</b>
Linear-SVM	<b>0.76 <math>\pm</math> 0.09</b>	<b>0.68 <math>\pm</math> 0.07</b>	<b>0.83 <math>\pm</math> 0.17</b>	<b>0.74 <math>\pm</math> 0.08</b>	<b>0.76 <math>\pm</math> 0.09</b>

To explain the improved Good vs. Bad separation, post convolution with Hadamard filters compared to those obtained with Gaussian filters, we compared the condition numbers of the two sets of 6 filters. This explanation is inspired by the research work of a fellow TenD researcher who has shown that using ill-conditioned filters in CNN models of US image analysis results in performance sensitivity and instability [126]. Table 5.19 shows that the Hadamard filters are well-conditioned, but the Gaussian filters are mostly relatively ill-conditioned. Moreover, all the Hadamard filters are almost orthogonal; hence, using them for convolving images preserves the local geometry of the images.

**Table 5.19 Comparisons of condition numbers of the Hadamard filters vs. Gaussian ones.**

Condition Number	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6
Hadamard	<b>1.4</b>	<b>1.4</b>	<b>1.4</b>	<b>2.0</b>	<b>2.0</b>	<b>2.0</b>
Gaussian	<b>20.0</b>	<b>36.1</b>	<b>10.7</b>	<b>30.2</b>	<b>12.3</b>	<b>21.2</b>

For the sake of completeness, we repeated these experiments, but instead of MCIQ, we extracted the TripleIQA+Entropy 4-dimensional feature vectors. The results shown below, in Table 5.20, show that the Hadamard filters instead of the Gaussian filters resulted in significantly improved alignment with the expert quality assessment by the kNN classifiers. However, the performance of the SVM classifiers shows that, like the MCIQ, the TripleIQA+Entropy feature vectors are spread out reasonably well in  $R^4$  but are not linearly separable by the SVM hyperplane.

**Table 5.20 Good vs. Bad using TripleIQA+Entropy post 6 Hadamard filter convolution.**

Good vs. Bad (120)	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	<b>0.78 <math>\pm</math> 0.02</b>	<b>0.73 <math>\pm</math> 0.11</b>	<b>0.82 <math>\pm</math> 0.15</b>	<b>0.76 <math>\pm</math> 0.02</b>	<b>0.78 <math>\pm</math> 0.02</b>
kNN, k=3	<b>0.71 <math>\pm</math> 0.10</b>	<b>0.72 <math>\pm</math> 0.13</b>	<b>0.70 <math>\pm</math> 0.14</b>	<b>0.71 <math>\pm</math> 0.11</b>	<b>0.71 <math>\pm</math> 0.10</b>
kNN, k=5	<b>0.71 <math>\pm</math> 0.06</b>	<b>0.77 <math>\pm</math> 0.22</b>	<b>0.65 <math>\pm</math> 0.11</b>	<b>0.71 <math>\pm</math> 0.11</b>	<b>0.71 <math>\pm</math> 0.06</b>
Linear-SVM	<b>0.48 <math>\pm</math> 0.06</b>	<b>0.50 <math>\pm</math> 0.12</b>	<b>0.47 <math>\pm</math> 0.18</b>	<b>0.49 <math>\pm</math> 0.07</b>	<b>0.48 <math>\pm</math> 0.06</b>

### 5.5.3 Performance of Modified MCIQ for Tumour Classification

The success of using Hadamard-based augmented MCIQ raises a number of points, including the use of Hadamard-based convolution as the basis for augmentation to improve the performance of CNN models. This will be investigated in the next Chapter, which is dedicated to dealing with the problem of scarcity of US images. Here, we close this chapter by conducting experiments to test the possibility of using this approach to improve the performance of the MCIQ feature vector for tumour classification. For that, we repeated the experiment of section 5.4.1 above, both post-convolution with the 6 Hadamard filters as well as with the 6 Gaussian filters. The results are shown in Table 5.21 and 5.22, respectively. Results of all kNN classifiers confirm that extracting MCIQ from the larger modified Renmin database when the cropped images are convolved with the 6 Hadamard filters significantly outperforms the cases of extracting MCIQ:

1. Only from the original Renmin cropped images, see Table 5.9,
2. From the convolved Renmin cropped RoIs using the 6 Gaussian filters, see Table 5.22.

When MCIQ were extracted from the Gaussian convolved images, the performance of all classifiers deteriorated in comparison to the results of Table 5.9. Furthermore, convolution with both types of filters resulted in lower SVM performance of the MCIQ feature vector compared to Table 5.9.

**Table 5.21 Benign vs. Malignant (Renmin) using MCIQ post 6 Hadamard filters augmentation.**

Benign vs. Malignant	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	0.92 $\pm$ 0.01	0.91 $\pm$ 0.01	0.93 $\pm$ 0.02	0.92 $\pm$ 0.01	0.92 $\pm$ 0.01
kNN, k=3	0.86 $\pm$ 0.02	0.83 $\pm$ 0.03	0.89 $\pm$ 0.02	0.86 $\pm$ 0.02	0.86 $\pm$ 0.02
kNN, k=5	0.83 $\pm$ 0.03	0.80 $\pm$ 0.04	0.87 $\pm$ 0.02	0.83 $\pm$ 0.03	0.83 $\pm$ 0.03
Linear-SVM	0.79 $\pm$ 0.01	0.81 $\pm$ 0.01	0.77 $\pm$ 0.02	0.79 $\pm$ 0.01	0.79 $\pm$ 0.01

**Table 5.22 Benign vs. Malignant (Renmin) using MCIQ post 6 Gaussian filters augmentation.**

Benign vs. Malignant	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
kNN, k=1	0.73 $\pm$ 0.01	0.70 $\pm$ 0.03	0.76 $\pm$ 0.01	0.72 $\pm$ 0.01	0.73 $\pm$ 0.01
kNN, k=3	0.75 $\pm$ 0.02	0.74 $\pm$ 0.03	0.77 $\pm$ 0.03	0.75 $\pm$ 0.02	0.75 $\pm$ 0.02
kNN, k=5	0.77 $\pm$ 0.01	0.76 $\pm$ 0.02	0.78 $\pm$ 0.03	0.76 $\pm$ 0.01	0.77 $\pm$ 0.01
Linear-SVM	0.79 $\pm$ 0.02	0.80 $\pm$ 0.02	0.78 $\pm$ 0.03	0.79 $\pm$ 0.02	0.79 $\pm$ 0.02

We close this section and Chapter 5 by making the following remark that would inspire more future work to exploit the benefits of convolution with Hadamard-based filters.

**Closing Remark:** Since convolving images with the well-conditioned and almost orthogonal Hadamard filters preserve the local geometry of the images, it is hardly surprising that they contribute positively to improving the performance of MCIQ when used for alignment with expert US-IQA or tumour classification.

## 5.6 Conclusion

Having realised the main difficulties in developing a single score IQA applicable to US images that can be used in conjunction with developing CNN models for their analysis, this Chapter adopted a simple approach that is based on comparing the various local statistical parameters of image pixel values in block-based partitioning to compute the MCIQ feature vector of a reasonably small size of 50 coordinates. This approach was motivated by observing that, unlike natural digital images, the spatial distributions of the UIQI components and speckle noise are not uniform across different parts of US images. The MCIQ feature vector encompasses the spatial distribution of 5-characteristic quality metrics: Correlation, Luminance, Contrast, Skewness, and Kurtosis. It was practically demonstrated that the MCIQ feature vector can be employed as a quality descriptor that helps partially clarify why a CNN model trained on a particular dataset may not always be generalizable to other unseen datasets. It was shown to have good tumour class dependency and can distinguish different image datasets used for various tasks (e.g., distinguishing US scans of different tissue types and distinguishing US images from other medical or natural image

modalities). We discussed the limitations of MCIQ in relation to alignment with expert radiologist assessment of quality, especially in relation to the main obstacles of US image scarcity (with standardized radiology experts) besides the absence of established knowledge on the sort of distortions (other than Speckle noise) in US images. These considerations helped develop an advanced MCIQ version that utilized the convolution by a small set of well-conditioned 5x5 Hadamard filters. In the next Chapter, we exploit these results to develop a convolution-based augmentation for improved CNN analysis of US images.

# Chapter 6: Image Augmentation for Deep Learning-based Breast Ultrasound Diagnosis

The performance of DL-CNN models for image analysis requires substantial quantities of high-quality and adequately-labelled training data to attain optimal performance. However, in medical image analysis (particularly for US tumour images), obtaining annotated images can be problematic due to the limited availability of samples, the distribution of which mimics that of the undetermined population. In Chapter 4, we demonstrated that training pre-trained CNN schemes on the reasonably large Modelling dataset of diverse samples led to robust models that can generalize well to the external testing datasets (Test1 and BUSI). However, training the same models on the smaller Renmin dataset yielded disappointingly low generalizability to the external datasets.

Image augmentation is the most commonly practised approach to address the data scarcity issue as a means of expanding the dataset by generating new images with some variation in their appearances. It is often employed to prevent model overfitting and improve generalizability into unseen samples. This Chapter is designed to review existing approaches to augment available US image datasets, propose and implement new augmentation schemes and compare their impact on the performance of pre-trained CNN models for BUS images. These schemes are not dependent on US data contents specifically, but we shall deploy them to augment the single medical centre Renmin dataset and evaluate their impact on the performance of pre-trained CNN models. We shall also develop a new US-specific approach to enlarge BUS training datasets by exploiting the benefits gained from RoI cropping using different RoI CH expanded ratios that also mitigate the challenge arising from inter-observer errors of lesion cropping [9].

In Section 6.1, we discuss the scarcity challenge in obtaining US images for DL-based lesion diagnostic tasks and its implications. We also explore existing techniques used to address this challenge. Moving on to Section 6.2, we delve into conventional image augmentation techniques, both for natural and medical images. In Section 6.3, our focus shifts to unconventional image augmentation techniques for US images, which draw on mathematical concepts beyond simple image processing-based schemes. We conduct experiments to compare the performance of these techniques with other augmentation methods. Section 6.4 introduces the Tumour Margin Appending (TMA) augmentation-like scheme, explicitly

designed for CNN-based BUS diagnostics. We present the results of experiments conducted to assess its effectiveness. Section 5.5 provides a summary of the chapter's key conclusions.

## **6.1 Data Scarcity in Ultrasound Lesion Images - Introduction.**

The challenge of data scarcity for using CNN models to analyse US lesion scan images doesn't seem to relate to the unavailability of US images when tens of thousands of hospitals/clinics worldwide daily conduct numerous US scanning of patients. Trained radiologists examine US scan images, and the patient's health status is determined and followed up by certain medical treatments/procedures. However, the digitisation of health records is lagging behind in many countries and even in those countries that embraced digital health, the process is yet to mature and lacks standardisation. In practice, there are only a few national databases and warehouses of US images that are indexed and annotated with the appropriate medical assessment in a globally agreed standardised format. The absence of standardisation can cause inconsistency in image acquisition due to many factors, including variation in image scanning procedures, diverse scanning devices, and to some extent, diversity of radiologists' training expertise. The inconsistency is often manifested by image characteristics in relation to variable resolution, contrast, and noise levels. Furthermore, the scarcity challenge is rightly compounded by ethical and privacy requirements for access to adequate databases.

One might ask, was this scarcity not as challenging for developing HC feature-based CAD models to analyse US scan images? It is certainly as serious, if not more acute, challenge but several factors may have helped, including the fact that performance expectation was not as ambitious as that from CNN models. Moreover, the emergence of several different types of image texture features that could be extracted in different transformed domains, including frequency ones, as well as the development of several classifiers, facilitated the use of multi-classifier fusion and ensemble schemes for improved performance. Indeed, at Buckingham, several such schemes were developed for diagnosing tumours from US scans of different tissues/organs, e.g., [5], [6], [22], [42], [126].

Our investigations so far indicate that the main consequences of data scarcity of US tumour images include reduced model performance, manifested by lower accuracy besides unacceptable rates of sensitivity or specificity. When trained with a limited dataset, CNN models may struggle to learn the underlying patterns and features that distinguish between malignant and benign breast lesions accurately. This limitation can result in misdiagnosis, leading to unnecessary biopsies or delayed treatment, negatively impacting patient

outcomes. Another consequence of data scarcity is model overfitting. In the absence of sufficient data, CNN models may memorize the limited dataset, resulting in poor generalization to unseen data. As a result, the CNN model's performance may appear high on the training set but fail to perform well on new, unseen data.

What makes data scarcity a more serious issue for CNN deployment is probably engrained in how CNN schemes learn discriminating image features. Although using pre-trained CNN models in fine-tuning mode on a number of US images results in learning many hidden feature patterns, but the additional training results in changing the CNN parameters to fit the training samples. However, the fact that the training US images do not form a reflective independent and identically distributed sample of the unknown population results in a lack of generalisation, as demonstrated, in Chapter 4, when we retrained several state-of-the-art pre-trained CNN models with not-so-small US images recorded in a single clinic (Renmin Hospital, Shanghai). The effects of the scarcity issue for medical image analysis using CNN models may be less drastic if trained on heterogeneous image datasets collected from different clinic centres. By training on heterogeneous datasets, DL models learn different hidden patterns of common features and patterns across a wider range of images can lead to better performance on unseen data, as we have demonstrated in Chapter 4 when we trained the state-of-the-art CNNs in fine-tuning mode on the Modelling BUS dataset collected from multiple clinics. Training a CNN model on a heterogeneous dataset can additionally help to ensure that the model is more broadly applicable to different patient populations and imaging devices. However, this may still depend on the number of available training US images.

Many data scarcity mitigating techniques have emerged for natural and medical images, the obvious of which suggest enlarging the training dataset, referred to as *Data Augmentation*. A larger training dataset provides more feature patterns for the model to learn from and could contribute to improving its ability to generalize to new unseen data as long as the additional patterns add to more diversity. This is particularly important in medical image analysis, where the underlying patterns may be subtle and difficult to learn from a small dataset. Accordingly, US data scarcity mitigating techniques benefit from focusing on more diverse feature learning rather than enlarging the dataset. These ideas may rely on choosing CNN architectures for which learning could benefit from the nature of US image content.

Several techniques have been proposed to enhance the performance of DL models in image analysis by focusing on the model's architecture itself, in addition to image augmentation techniques. Functional solutions such as dropout regularization [127], batch normalization [128], and transfer learning have been developed to enable the use of DL models on smaller



datasets [129], [130]. Dropout regularization forces the network to learn more robust features by zeroing out the activation values of randomly chosen neurons during training. Batch normalization normalizes the set of activations in a layer to subtract the batch mean from each activation and divide it by the batch standard deviation. Transfer learning involves training a network on a large dataset, such as ImageNet [7], and then using those weights as initial weights in a new classification task.

Other alternative approaches to augmentation include the Zero-shot and One-shot algorithms that have been proposed to overcome training models with extremely limited data [131], [132]. Zero-shot learning involves training a model on a set of data that is distinct from the task to which the model will eventually be applied. The model learns a relationship between the training data and the task and then applies this knowledge to classify new examples. On the other hand, one-shot learning uses a small set of examples to classify new samples by training a distance function that maps the inputs to the correct output. These methods are useful when obtaining large amounts of labelled data is difficult or impractical and can be applied in various domains such as natural language processing, computer vision, and speech recognition. Knowledge of these overfitting solutions will inform readers about other existing tools/techniques [133].

Unlike the above-mentioned techniques, **Data Augmentation** tackles the problem of overfitting and generalizability in DL by modifying the training dataset. The underlying idea is that by introducing variations and new instances into the dataset, the CNN architecture can capture previously undiscovered feature patterns that would not have been learned solely from the original dataset during training. This approach aims to enhance the model's ability to generalize and perform better on unseen data. The viability of any augmentation scheme for any dataset should not be measured solely by the size of the dataset post-augmentation. The success of any augmentation scheme for a given CNN architecture should be evaluated in terms of the performance rates (accuracy, specificity and sensitivity) when testing unseen external data. A serious factor influencing the success of an augmentation scheme for a given dataset is related to the possibility of confusing the membership of the different classes as a result of the augmentation algorithm. Note that rotating 6 may produce a 9 for character recognition. Shorten and Khohgoftaar [99] surveyed several easy-to-implement image augmentation techniques using geometric image processing transformations for natural images, referring to this problem as the *safety of post-augmentation labelling*.

Caution has been urged to explore non-augmentation methods, when possible, to alleviate the burden of US image scarcity for BUS classification using DL. Zhu et al. [105] utilized

CNNs to develop an automated system for classifying breast and thyroid lesions in US images by training with datasets of diverse US images collected from multiple clinical centres. They proposed a generic DL framework that could be applied to larger datasets collected from diverse patient populations across different centres. The study compared the performance of the proposed models with that of radiologists and analysed the relationship between correct classification outcomes and regions of input RoI images. Additionally, the authors investigated the known US characteristics shared by thyroid and breast lesions, such as shape ratio, hypo-echogenicity, and ill-defined margins. Overall, this study has significant implications for improving the early detection and treatment of breast and thyroid cancers. In the next section, we review and categorise existing image augmentation techniques, highlighting their applicability to US tumour image datasets in light of the above discussion.

## 6.2 A Review of Existing Image Augmentation Techniques

Over the years, many different image augmentation schemes have been proposed and used for different machine-learning models of computer vision tasks. There are many ways of categorising these techniques. For example, Shorten and Khohgoftaar [99] categorised existing augmentation schemes as *Data Warping* augmentations and *Oversampling* schemes that are not mutually exclusive. Here, we shall use a simple categorisation as *Conventional* (via Image processing procedures), *Synthetic* (using generative Neural Networks), and *Spectral* (using eigenvalue/singular value analysis) schemes. Not all techniques may be appropriate for every dataset or task, so it is important to carefully consider which techniques are most suitable for the problem at hand.

### 6.2.1 Conventional Augmentation Techniques

These schemes include applying geometric operations such as Rotation and Horizontal/Vertical flips [134]–[136]; noise addition of different variances into the training data [137], [138]; kernel filtering such as sharpening and blurring [139], [140]; photometric colour transforms [141], [142]. For other schemes in this category, see [99], but these are chosen for ease of implementation for US images. Some of these schemes may result in unsafe labelling, as mentioned above. In such cases, computationally expensive refinement of post-augmentation labels becomes necessary. Noise addition has different effects on natural images than on US images due to the dominance of texture features in US images. The above techniques can be combined in various ways to create a diverse set of training data for DL models. In fact, the current very generic approach to augment image datasets is

to several affine mapping of original images using a combination of geometric, colour jitters, image cropping and translation [102].

Tirindelli et al. [143] proposed a novel data augmentation approach for medical US imaging that uses well-known image processing concepts. The authors introduced a set of physics-inspired transformations, including deformation, reverb, and Signal-to-Noise Ratio, that can be applied to US B-mode images for data augmentation. These techniques are intended to align with the physics of the US and avoid generating images that radiologists may deem to be unrealistic. The proposed augmentations were used to enlarge a new US spine dataset for bone segmentation and classification. The results show that deformations & reverberation-based US augmentation slightly outperform classical augmentation but argue for further research into anatomically realistic US augmentations for training generalizable CNNs.

### **6.2.2 Synthetic Augmentation Techniques**

Several popular generative DL tools have emerged for image/data augmentation in recent years. The most popular approach is the Generative Adversarial Networks (GANs). GANs are a type of DL model consisting of two neural networks: a generator network and a discriminator network. The generator network creates synthetic data meant to resemble samples in a target domain, while the discriminator is trained to differentiate between the synthetic data and the target data. Synthetic data are created iteratively by the generator, starting with a given sample data and repeatedly computing a new data version by adding a random noise signal to it, passing it on to the discriminator network that will determine the needed adjustment, and the generator updates the currently held data until it is indistinguishable from real domain samples. GANs are highly effective at data augmentation for better training DL models with improved model performance [144]–[146].

Variational Autoencoders (VAEs) are another type of neural network architecture suitable for data augmentation. It has 2 components: an encoder and a decoder. It differs from GAN in that the encoder maps input data samples to a lower-dimensional latent space while the decoder uses the latent vector to reconstruct the original data sample. During training, VAEs optimize reconstruction loss (determining the closeness of reconstructed data from the latent space) as well as KL divergence (between the distribution of the latent space and a standard normal distribution). Once the VAE is trained, it can be used to generate new data samples by sampling from the latent space and decoding the samples into the original data domain. Overall, VAEs are powerful for data augmentation because they can generate new data

samples that are similar to the original data but with small variations, which can help improve the robustness of ML models [147].

Both GAN and VAE can be used to generate any type of data and images. Al-Dhabyani et al. [89] investigated the classification of breast masses and explored a generative approach for augmenting US images. They proposed a GAN-based image augmentation technique, DAGAN, that creates authentic, high-quality images from scratch. The effectiveness of various DL models for breast mass classification using US images was assessed in terms of accuracy post-augmentation, concluding that DAGAN has the potential to enhance the precision and efficiency of breast cancer diagnosis.

Another type of image augmentation suitable for US images is the Style Transfer technique based on neural networks to transfer one image's style onto another image's content. It employs pre-trained CNNs to extract content and style features from input images. The content features are compared between the content image and the generated image, ensuring content preservation, while the style features are represented by the Gram matrices of the feature maps and matched with those of the style image, facilitating the transfer of style. By iteratively optimizing the generated image to minimize a weighted sum of content and style losses, using techniques like gradient descent, the network creates a final output that merges the content of one image with the stylistic elements of another, enabling applications ranging from artistic rendering to visual effects. This can be useful for data augmentation because it allows us to generate new images with similar content to the original images in a dataset considered to be scarce for CNN model analysis. Overall, style transfer is a powerful data augmentation tool because it allows us to generate new images that have similar content to the original images but with different styles, which can help improve the diversity and robustness of ML models [102]. Again, this interesting approach is used mostly for natural image augmentation. However, one can envisage its use for augmentation of medical images, e.g., a style transfer network trained on CT scans and US scans of lungs could be used to transfer lung CT images onto lung US scan images.

In this work, we decided not to use generative neural network image augmentation due to several concerns. While GAN-based techniques offer a promising approach to mitigate the challenges posed by the scarcity of US medical images, their application introduces specific issues, particularly for medical imaging. Notably, the occurrence of vanishing gradients and mode collapsing has been observed, leading to limitations in generating diverse and high-quality augmentations [148], [149]. It is essential to acknowledge that the perception of image quality in US images differs from that in natural images, as established in Chapter 4.

Additionally, maintaining synchronization between the generator and discriminator in GAN neural networks can be challenging, particularly concerning the symmetry and alignment of these networks [100]. The relatively small size of available ultrasound datasets further complicates the training of intricate models. Moreover, GAN-based methods demand substantial computational resources, resulting in prolonged processing times on devices with limited capabilities. There are valid concerns regarding the reliability and credibility of GAN-based augmentation in diagnostic tasks, with potential contributions to overfitting and emphasizing the susceptibility of CNN models trained on GAN-generated images to adversarial attacks [104].

### 6.2.3 Spectral-based Augmentation Techniques

Spectral analysis of image datasets has been employed to generate a representation of natural images. This is achieved by transforming the images' coordinate system through projection onto a new vector space of the same dimension/resolution. The coordinate axes in this space are arranged based on the descending order of variances of the original image pixels away from the average of the dataset images. There are various spectral analysis methods, including Principal Component Analysis (PCA), Independent Component Analysis, and Random Projections. These methods have been used for face biometrics, dimension reduction, and CSSR (see, e.g., [75], [76], [150], [151]). These spectral image-based projections can be used for image augmentation through the generation of new images by manipulating the spectral bases. Recently, PCA has been used to propose a novel image augmentation technique based on a random permutation of coefficients of within-class most significant axis post PCA projection of an image dataset. A custom CNN was trained on the augmented surrogate images obtained from the CIFAR-10 image dataset and was shown to improve classification accuracy and ambiguity [152].

In general, the success of spectral-based augmentation schemes, such as PCA, requires sufficient diversity of the available samples and a good random population sample. Although this is the main challenge for US images, such approaches can achieve relative success. Indeed, one of our proposed innovative augmentation schemes that uses a random projection-based approach will be presented and tested in section 6.3.

A closely related approach to PCA augmentation is spectral-based augmentation that relies on each image's Singular Value Decomposition (SVD), considered as a matrix, and manipulates the SVD factors by tiny changes to form a new image. Later, in section 6.3, we give more details on the SVD augmentation and its impact on the performance of pre-trained

CNN models for BUS image analysis. The geometric methods, including flipping and rotation, alter the image geometry by mapping individual pixel values to new destinations, while SVD generates new images with similar features as the original ones but with somewhat different quality due to SVD degradation/compression.

Ahmed et al. [68] discussed the challenges of generalization in ENAS-based CNN models for breast lesion classification from US images. The paper investigates the effectiveness of various techniques, including reducing model complexity, data augmentation, and the use of unbalanced training sets to overcome generalization errors. It was shown that SVD plus geometric augmentation reduces ENAS model overfitting and performs better.

Finally, we note that appending cropped RoIs with different proportions of surrounding tissue, as discussed in Chapter 4, provides a mitigating solution for the US scarcity images by providing multiple genuine versions of the dataset images without modifying the content of the RoI regions. This is analogous to image augmentation and is consistent with the way radiologists assess US tumour images. We shall present this method in section 6.4.

### **6.3 Mathematically Inspired Augmentation Techniques for BUS**

Here, we first describe two mathematically inspired augmentation schemes that we developed and implemented as a solution to the scarcity problem of our TenD BUS datasets in relation to using pre-trained CNN models for diagnostic purposes. We shall then present the results of experimental investigations to determine the effects of these augmentation techniques as well as the conventional flip and rotation (Flip&Rot) augmentation, on the classification of BUS lesions using DL models in comparison with no augmentation.

#### **6.3.1 SVD-based Image Augmentation**

This approach is a Spectral-based augmentation scheme that, unlike PCA-based augmentation, analyses each image spectrally on its own and generates several copies of the same image. During my early study program and together with my supervisors, I contributed to the development of such a scheme that was later combined with geometric augmentation and appeared later in [68], [105].

SVD is a widely used matrix factorization technique in linear algebra, and it has important applications in signal/image processing, data analysis, and ML [76]. It is an essential tool for understanding the underlying structure of matrices in terms of the geometric profile of their columns and rows. Mathematically, given an  $m \times n$  matrix  $A$ , SVD decomposes  $A$  into the product of three matrices as follows:  $A = U \Sigma V^T$  where  $U$  is an  $m \times m$  orthogonal matrix,  $\Sigma$  is an  $m \times n$  diagonal matrix with non-negative real entries, and  $V$  is an  $n \times n$  orthogonal matrix.

The superscript T denotes the transpose of the obtained simply by turning each column into a row in the same order. The non-zero entries of the diagonal matrix  $\Sigma$  are called the singular values of A, and they are arranged in decreasing order along the diagonal, so that the first singular value represents the most important information in the matrix A. The columns of U and V are called the left and right singular vectors of A, respectively, and they form orthonormal bases for the row and column spaces of A. In other words, SVD decomposes A into a sum of rank-one matrices, where each rank-one matrix is the outer product of a left singular vector, a singular value, and a right singular vector. The larger the singular value, the more important the corresponding singular vector is in the decomposition [153].

SVD has many important properties: It provides a complete characterization of matrix A, including its rank, null space, and range. It is a unique decomposition, meaning that the left and right singular vectors and the singular values are unique up to a sign and a permutation. It is robust to noise and rounding errors, making it useful in numerical computations.

For US image augmentation, we consider each image as an  $m \times n$  matrix and apply SVD analysis of this matrix. This allows us to approximate the original matrix A by truncating the SVD at a certain rank, which can significantly reduce the dimensionality of the original matrix while preserving most of its important information. However, SVD image compression/dimension reduction affects the quality of the reconstructed image. In our work, each image in the training dataset is augmented with 4 SVD approximated images using the top (most important) 50, 40, 30, and 20 Singular values, see Figure 6.1.

Instead of invoking this compression approach, other SVD-based versions of any image can be generated by simple perturbations of some of the singular values that keeps their order along the diagonal matrix  $\Sigma$ . For example, if  $\Sigma = \text{Diag}(\sigma_1, \sigma_2, \sigma_3, \dots, \sigma_n)$ , then replacing several  $\sigma_i$ 's with  $\sigma_i + \epsilon_i$  for some small  $\epsilon_i$  values that keep the order of the new singular values and reconstructing the image, will generate as many versions as needed. However, this approach has not been implemented due to the fact that the experiments conducted are meant to be a proof of concept.

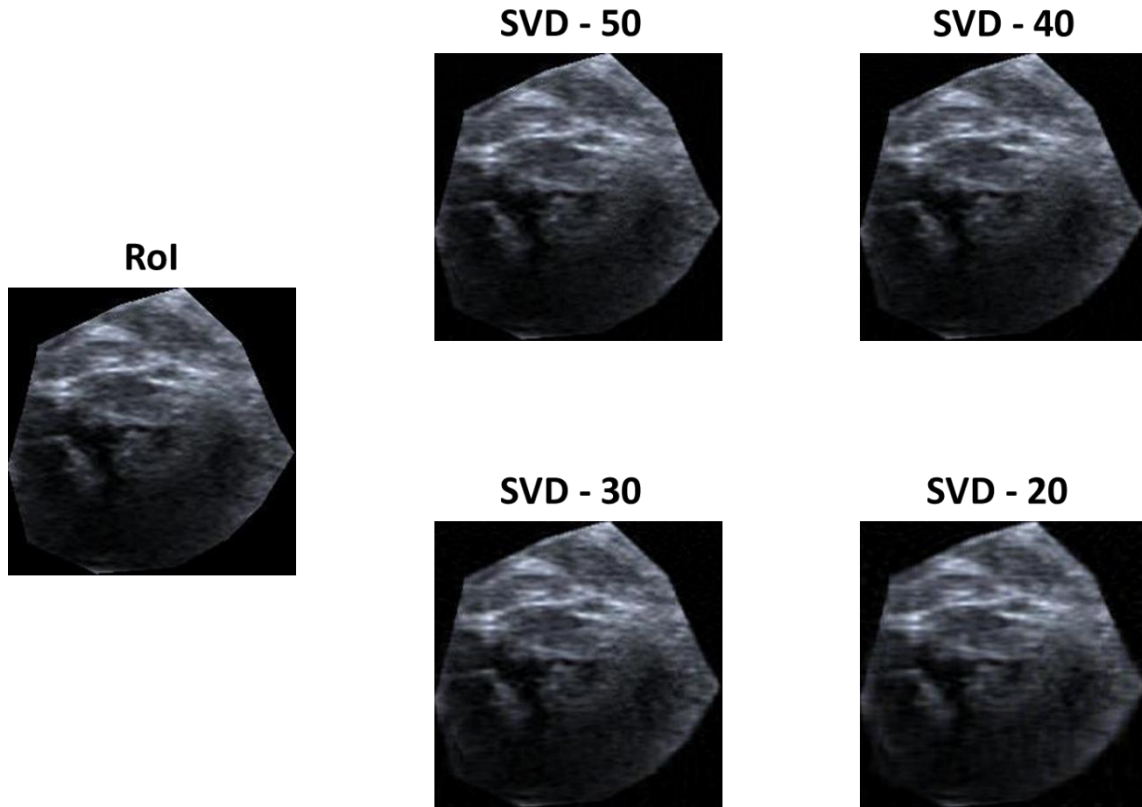


Figure 6.1 SVD-based US image augmentation via reduced number of singular values.

### 6.3.2 Hadamard-based Kernel Image Augmentation

This proposed augmentation scheme is inspired by the concept of the compressed sensing paradigm, discussed in Chapter 3, which is based on the sparse representation of an image using features formed by non-trivial linear combinations of pixels. This approach for augmentation does not benefit directly from using compressed sensing that is designed to extract a global sparse representation of the images, but instead, we propose to construct several small RIP-based  $k \times k$  matrices to transform images with such filters. Besides many other matrices that facilitate compressed sensing of images, Hadamard matrices are well-known as perfect candidate matrices that provide a rich pool for the selection of RIP matrices (see, [76], [87]), but do not directly fit the characteristics of convolution filters that are of small odd order square matrices. Hadamard matrices are usually of size  $(2^n \times 2^n)$ , and our required filters need to be  $k \times k$  matrices with a relatively small, odd number  $k$ . We simply generalise the process of modifying our MCIQ feature vectors, in Chapter 5, and construct our filters using block-diagonals of a mix of Hadamard matrices of sizes  $1 \times 1$ ,  $2 \times 2$ ,  $4 \times 4$ , ..., etc. In the last Chapter, we created the 6 ( $5 \times 5$ ) filters, displayed in Figure 5.2, by different



block-diagonals of  $1 \times 1$ , and  $(2 \times 2$  or  $4 \times 4)$  Hadamard matrices. Similarly, one can create Hadamard-based kernel filters of sizes  $3 \times 3$ ,  $7 \times 7$ , and  $11 \times 11$ .

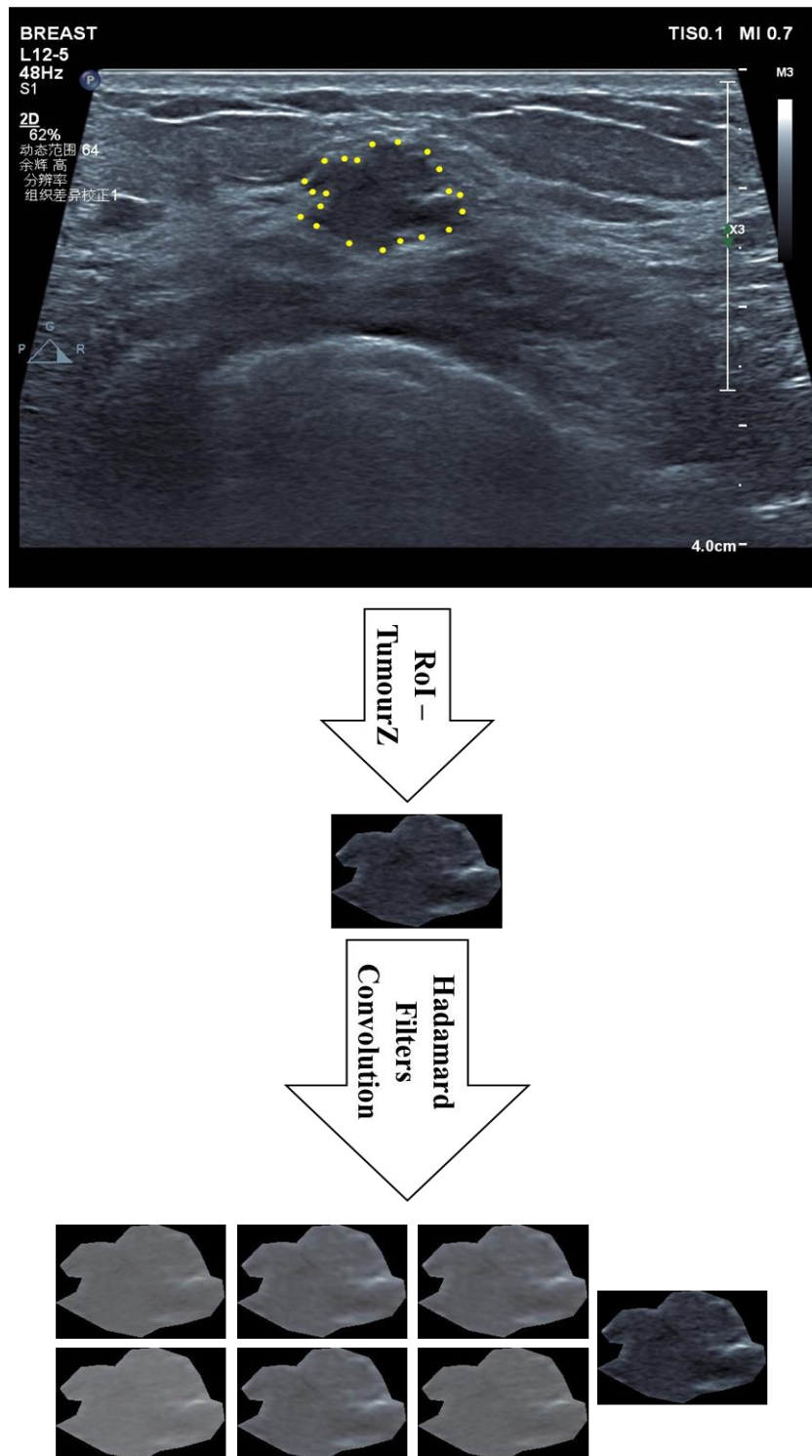


Figure 6.2 Hadamard filter US image augmentation.

Our block-diagonal method can be used for constructing many more Hadamard-based convolution filters of size beyond  $5 \times 5$ , but to demonstrate their benefits, we limit our Hadamard-based augmentation to the use of the 6 ( $5 \times 5$ ) Hadamard-based filters of Chapter 5. These filters are well-conditioned by design (condition numbers  $\leq 2$ ), see Table 5.19, and their choice for convolution-based augmentation is influenced by the fact that their action on image patches is not sensitive to tolerable perturbation of pixel values. Gaussian filters can be used, but we need to select a set of well-conditioned filters. With these 6 Hadamard-based filters, we enlarge the Renmin training dataset size by 7 folds. For each image from the training dataset, we add a new image by convolving the original tightly cropped tumour ROI bounding box with zero padding (i.e., TumourZ) with each of these 6 filters, as shown in Figure 6.2.

### 6.3.3 Augmentation Experimental Work

In this section, we present the results of our experimental work that aimed to test the performance of the same 4 state-of-the-art pre-trained CNN models on the augmented Renmin dataset of BUS tumour scan images using the above two augmentation schemes (SVD and Hadamard). For comparison, we also test the performance of the Flip&Rot conventional augmentation scheme. In all these experiments, we shall also test the performance of each of these schemes on the unseen datasets Test1 and BUSI. As before, we shall follow the 5-fold cross-validation protocol to retrain the various pre-trained CNN models on each augmented Renmin dataset in the fine-tuned version of the transfer learning mode. Although retraining these and other pre-trained CNN models on the larger multi-centre BUS Modelling dataset did not suffer from the lack of generalization to the above external datasets, we also repeated the augmentation experiments for the Modelling dataset, which led to marginal changes to their performances on the external dataset. The results are not presented here but are available in Appendix A.

We shall now present the performance of each of the above-mentioned augmentation schemes separately and discuss their generalization performance with the external datasets compared to the results obtained in Chapter 4 with no augmentation. But later, we end the section by comparing the performance of these augmentation schemes against each other, recommending means of exploiting these schemes. But before we do all these, and for the sake of comparison, below in Table 6.1, we present a replica of the performance results of the 4 pre-trained CNN models without augmentation on the external datasets.

**Table 6.1 Performance of the CNN models retrained on un-augmented Renmin dataset.**

Validation	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.94 <math>\pm</math> 0.04</b>	<b>0.94 <math>\pm</math> 0.07</b>	<b>0.93 <math>\pm</math> 0.04</b>	<b>0.93 <math>\pm</math> 0.04</b>	<b>0.94 <math>\pm</math> 0.04</b>
VGG16	<b>0.95 <math>\pm</math> 0.04</b>	<b>0.94 <math>\pm</math> 0.07</b>	<b>0.96 <math>\pm</math> 0.01</b>	<b>0.95 <math>\pm</math> 0.04</b>	<b>0.96 <math>\pm</math> 0.04</b>
VGG19	<b>0.95 <math>\pm</math> 0.04</b>	<b>0.94 <math>\pm</math> 0.06</b>	<b>0.95 <math>\pm</math> 0.03</b>	<b>0.95 <math>\pm</math> 0.04</b>	<b>0.95 <math>\pm</math> 0.07</b>
ResNet18	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.06</b>	<b>0.95 <math>\pm</math> 0.05</b>	<b>0.92 <math>\pm</math> 0.03</b>	<b>0.93 <math>\pm</math> 0.05</b>
Test1	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.78 <math>\pm</math> 0.03</b>	<b>0.75 <math>\pm</math> 0.06</b>	<b>0.79 <math>\pm</math> 0.04</b>	<b>0.72 <math>\pm</math> 0.04</b>	<b>0.77 <math>\pm</math> 0.03</b>
VGG16	<b>0.76 <math>\pm</math> 0.03</b>	<b>0.74 <math>\pm</math> 0.10</b>	<b>0.77 <math>\pm</math> 0.06</b>	<b>0.70 <math>\pm</math> 0.04</b>	<b>0.76 <math>\pm</math> 0.03</b>
VGG19	<b>0.76 <math>\pm</math> 0.02</b>	<b>0.74 <math>\pm</math> 0.06</b>	<b>0.77 <math>\pm</math> 0.03</b>	<b>0.70 <math>\pm</math> 0.03</b>	<b>0.75 <math>\pm</math> 0.02</b>
ResNet18	<b>0.62 <math>\pm</math> 0.11</b>	<b>0.83 <math>\pm</math> 0.05</b>	<b>0.49 <math>\pm</math> 0.21</b>	<b>0.63 <math>\pm</math> 0.06</b>	<b>0.66 <math>\pm</math> 0.08</b>
BUSI	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.89 <math>\pm</math> 0.03</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.84 <math>\pm</math> 0.05</b>	<b>0.87 <math>\pm</math> 0.03</b>	<b>0.92 <math>\pm</math> 0.02</b>
VGG16	<b>0.78 <math>\pm</math> 0.09</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.66 <math>\pm</math> 0.14</b>	<b>0.76 <math>\pm</math> 0.07</b>	<b>0.83 <math>\pm</math> 0.07</b>
VGG19	<b>0.82 <math>\pm</math> 0.04</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.72 <math>\pm</math> 0.06</b>	<b>0.79 <math>\pm</math> 0.04</b>	<b>0.86 <math>\pm</math> 0.03</b>
ResNet18	<b>0.66 <math>\pm</math> 0.08</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.49 <math>\pm</math> 0.13</b>	<b>0.68 <math>\pm</math> 0.05</b>	<b>0.74 <math>\pm</math> 0.06</b>

### 6.3.3.1 Performance of the Flip&Rot Augmentation Scheme

The version of the Flip&Rot augmentation implemented here enlarges the Renmin dataset by a factor of 5, whereby 4 additional images are generated from each training dataset image using rotation 90°, 180°, 270°, and a vertical flip, (see Chapter 3, Figure 3.19). Table 6.2 below displays the outcome from our experiments when the pre-trained schemes were retrained on the enlarged Renmin dataset post-augmentation with our Flip&Rot scheme.

**Table 6.2 Generalisation of pre-trained CNNs retrained with Flip&Rot -augmented Renmin dataset.**

Test1	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.78 <math>\pm</math> 0.00</b>	<b>0.71 <math>\pm</math> 0.04</b>	<b>0.82 <math>\pm</math> 0.02</b>	<b>0.71 <math>\pm</math> 0.01</b>	<b>0.76 <math>\pm</math> 0.01</b>
VGG16	<b>0.76 <math>\pm</math> 0.03</b>	<b>0.66 <math>\pm</math> 0.03</b>	<b>0.82 <math>\pm</math> 0.06</b>	<b>0.68 <math>\pm</math> 0.02</b>	<b>0.74 <math>\pm</math> 0.02</b>
VGG19	<b>0.77 <math>\pm</math> 0.02</b>	<b>0.66 <math>\pm</math> 0.04</b>	<b>0.83 <math>\pm</math> 0.04</b>	<b>0.68 <math>\pm</math> 0.02</b>	<b>0.75 <math>\pm</math> 0.01</b>
ResNet18	<b>0.70 <math>\pm</math> 0.03</b>	<b>0.75 <math>\pm</math> 0.10</b>	<b>0.67 <math>\pm</math> 0.09</b>	<b>0.66 <math>\pm</math> 0.02</b>	<b>0.71 <math>\pm</math> 0.02</b>
BUSI	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.87 <math>\pm</math> 0.01</b>	<b>0.99 <math>\pm</math> 0.00</b>	<b>0.81 <math>\pm</math> 0.02</b>	<b>0.84 <math>\pm</math> 0.01</b>	<b>0.90 <math>\pm</math> 0.01</b>
VGG16	<b>0.86 <math>\pm</math> 0.05</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.79 <math>\pm</math> 0.08</b>	<b>0.84 <math>\pm</math> 0.05</b>	<b>0.90 <math>\pm</math> 0.04</b>
VGG19	<b>0.89 <math>\pm</math> 0.03</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.84 <math>\pm</math> 0.04</b>	<b>0.87 <math>\pm</math> 0.03</b>	<b>0.92 <math>\pm</math> 0.02</b>
ResNet18	<b>0.76 <math>\pm</math> 0.07</b>	<b>1.00 <math>\pm</math> 0.01</b>	<b>0.63 <math>\pm</math> 0.11</b>	<b>0.74 <math>\pm</math> 0.06</b>	<b>0.81 <math>\pm</math> 0.05</b>

Generally, these results show that augmenting the training dataset using our Flip&Rot scheme has a slightly different effect when tested with two sets. For the Test1 dataset, the overall accuracy has only improved significantly (by about 8%) for the Resnet18 model. For

the other CNN models, the same accuracy rate was maintained or improved marginally by 1% for VGG19. Interestingly, the specificity rates for all the CNN models improved by more than a marginal increase, while the sensitivity rates deteriorated by more than marginal percentages (4% to 8%). This is somewhat not encouraging from the medical point of view in the benefits of fewer Benign cases being misclassified are lost by getting more misclassified Malignant cases.

For the BUSI dataset, except for the AlexNet model, the overall accuracy has improved significantly (by about 7% to 10%). For AlexNet, the accuracy deteriorated by 2%. Interestingly, the specificity rates for all the CNN models improved by more than a marginal increase while the sensitivity rates remained at their optimal rates achieved without augmentation. Thus, the improved accuracy is entirely the result of the improved specificity rate, i.e., fewer benign cases are misclassified while no more malignant cases are misclassified. From the medical point of view, this is a welcome improved performance, unlike the case of the Test1 dataset. These discrepancies in the medical effects of this augmentation between Test1 and BUSI datasets may be explained by the fact established in Chapter 5 (Tables 5.6 and 5.7) that the BUSI dataset is less separable from Renmin than the Test1 dataset in terms of the MCIQ quality assessment feature vector.

### **6.3.3.2 Performance of the SVD-based Image Augmentation Scheme**

First, we point out that we only implemented the SVD augmentation using the SVD-compression approach by eliminating different percentages of singular values rather than applying minor changes to some randomly selected singular values. As a result, the Renmin dataset was enlarged 5 folds. Table 6.3 below displays the outcome from our experiments when the pre-trained CNN schemes were retrained on the enlarged Renmin dataset post-augmentation with this restricted SVD augmentation scheme.

These results show that, like the case of using the Flip&Rot augmentation, augmenting the training dataset using the SVD-based scheme impacts the generalization performance differently when tested with the two external sets. For the Test1 dataset, the overall accuracy has improved for all models but only significantly (by 11%) for the Resnet18 model, which is 3% more than that with the Flip&Rot. For the other CNN models, the improvement was less than significant 2%, which is again better than the Flip&Rot. Similarly, the specificity rates for all the CNN models improved by different rates, with a minimum of 3% for VGG19 to a maximum of 19% for Resnet18. The specificity rate of AlexNet increased by a noticeable 7%, while VGG16 increased by more than the marginal rate of 5%. On the other

hand, the sensitivity rates deteriorated for AlexNet and Resnet18 by 2% and for VGG16 by 5% while the sensitivity rate improved by 2% for VGG19. From the medical point of view, these results are more encouraging than the case of Flip&Rot. Not all benefits of lower false positive misclassified benign masses have been lost completely. In fact, VGG19 also had lower false negative misclassified malignant masses.

**Table 6.3 Generalisation of pre-trained CNNs retrained with SVD-Augmented Renmin dataset.**

Test1	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
AlexNet	<b>0.80 ± 0.01</b>	<b>0.73 ± 0.04</b>	<b>0.84 ± 0.02</b>	<b>0.73 ± 0.02</b>	<b>0.78 ± 0.01</b>
VGG16	<b>0.78 ± 0.01</b>	<b>0.69 ± 0.04</b>	<b>0.84 ± 0.03</b>	<b>0.71 ± 0.02</b>	<b>0.76 ± 0.01</b>
VGG19	<b>0.78 ± 0.02</b>	<b>0.76 ± 0.02</b>	<b>0.80 ± 0.02</b>	<b>0.73 ± 0.02</b>	<b>0.78 ± 0.02</b>
ResNet18	<b>0.73 ± 0.05</b>	<b>0.81 ± 0.11</b>	<b>0.68 ± 0.13</b>	<b>0.69 ± 0.03</b>	<b>0.74 ± 0.03</b>
BUSI	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
AlexNet	<b>0.90 ± 0.02</b>	<b>0.99 ± 0.00</b>	<b>0.85 ± 0.03</b>	<b>0.87 ± 0.02</b>	<b>0.92 ± 0.01</b>
VGG16	<b>0.90 ± 0.02</b>	<b>1.00 ± 0.00</b>	<b>0.86 ± 0.03</b>	<b>0.88 ± 0.02</b>	<b>0.93 ± 0.01</b>
VGG19	<b>0.89 ± 0.03</b>	<b>1.00 ± 0.00</b>	<b>0.83 ± 0.04</b>	<b>0.86 ± 0.03</b>	<b>0.91 ± 0.02</b>
ResNet18	<b>0.79 ± 0.09</b>	<b>1.00 ± 0.00</b>	<b>0.67 ± 0.13</b>	<b>0.77 ± 0.07</b>	<b>0.84 ± 0.07</b>

For the BUSI dataset, the overall accuracy has improved for all models. Only AlexNet accuracy increased marginally by 1%; all other models improved significantly (7% for VGG19, 12% for VGG16 and 13% for Resnet18). Interestingly, the specificity rates for all except for AlexNet improved by significant rates of (20% for VGG16, 11% for VGG19, and 18% for ResNet18) but marginally by 1% for AlexNet. On the other hand, the sensitivity rates were maintained at their optimal rates achieved without augmentation except for AlexNet, which marginally deteriorated by only 1%. Thus, the improved accuracy results entirely from the improved specificity rate, i.e., fewer false positive benign misclassified masses, while no more malignant cases are misclassified. From the medical point of view, this is a welcome improved performance compared to the case of the Test1 dataset. Again, these discrepancies in the medical effects of this augmentation between Test1 and BUSI datasets may be explained by their different level of separability from Renmin in terms of the MCIQ quality assessment feature vector.

### 6.3.3.3 Performance of the Hadamard-based Image Augmentation Scheme

The Hadamard-based kernel filtering technique was used for augmentation to increase the size of the Renmin dataset. Again, we remember that implementing the Hadamard-based augmentation uses only 6 Hadamard-based 5x5 matrices, which enlarges the Renmin dataset by 7 folds. Table 6.4 presents the experimental results when the pre-trained CNN schemes were retrained on the enlarged Renmin dataset post-Hadamard-based augmentation.

**Table 6.4 Generalisation of pre-trained CNNs retrained with Hadamard-Augmented Renmin dataset.**

Test1	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.78 <math>\pm</math> 0.01</b>	<b>0.70 <math>\pm</math> 0.05</b>	<b>0.83 <math>\pm</math> 0.04</b>	<b>0.71 <math>\pm</math> 0.01</b>	<b>0.76 <math>\pm</math> 0.01</b>
VGG16	<b>0.78 <math>\pm</math> 0.02</b>	<b>0.66 <math>\pm</math> 0.02</b>	<b>0.85 <math>\pm</math> 0.02</b>	<b>0.69 <math>\pm</math> 0.02</b>	<b>0.76 <math>\pm</math> 0.02</b>
VGG19	<b>0.77 <math>\pm</math> 0.01</b>	<b>0.74 <math>\pm</math> 0.03</b>	<b>0.79 <math>\pm</math> 0.03</b>	<b>0.71 <math>\pm</math> 0.02</b>	<b>0.76 <math>\pm</math> 0.01</b>
ResNet18	<b>0.77 <math>\pm</math> 0.01</b>	<b>0.81 <math>\pm</math> 0.04</b>	<b>0.74 <math>\pm</math> 0.03</b>	<b>0.73 <math>\pm</math> 0.01</b>	<b>0.78 <math>\pm</math> 0.01</b>
BUSI	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.90 <math>\pm</math> 0.03</b>	<b>0.99 <math>\pm</math> 0.00</b>	<b>0.85 <math>\pm</math> 0.05</b>	<b>0.88 <math>\pm</math> 0.03</b>	<b>0.92 <math>\pm</math> 0.02</b>
VGG16	<b>0.92 <math>\pm</math> 0.02</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.88 <math>\pm</math> 0.03</b>	<b>0.90 <math>\pm</math> 0.02</b>	<b>0.94 <math>\pm</math> 0.01</b>
VGG19	<b>0.88 <math>\pm</math> 0.03</b>	<b>1.00 <math>\pm</math> 0.00</b>	<b>0.81 <math>\pm</math> 0.04</b>	<b>0.85 <math>\pm</math> 0.03</b>	<b>0.90 <math>\pm</math> 0.02</b>
ResNet18	<b>0.83 <math>\pm</math> 0.05</b>	<b>0.99 <math>\pm</math> 0.00</b>	<b>0.74 <math>\pm</math> 0.08</b>	<b>0.80 <math>\pm</math> 0.05</b>	<b>0.86 <math>\pm</math> 0.04</b>

These results seem to follow similar patterns of improvements/deterioration of performances on the external datasets with slight modifications. For the Test1 dataset, the overall accuracy has improved marginally (1% to 2%) for the first 3 models but is significantly improved (by 15%) for the Resnet18 model, which is 7% more than with the Flip&Rot and 4% more than that of the SVD augmentation. The specificity rates for all the CNN models improved compared to no augmentation, but the improvement for AlexNet and VGG16 are marginally better than that of the Flip&Rot and reasonably similar to SVD augmentation. For VGG19, specificity rates deteriorated by 4% (1%) when compared to Flip&Rot (SVD) augmentations. However, the specificity rate for Resnet18 has improved significantly by 25%, which is 6% to 7% more than the improvements achieved by the other 2 augmentation schemes. Interestingly, the improved specificity rates were achieved at the expense of deteriorated sensitivity rates, except for VGG19. From the medical point of view, these results are not as good as those achieved by SVD augmentation.

For the BUSI dataset, the overall accuracy has improved for all models. Only AlexNet accuracy increased marginally by 1%; all other models improved significantly (14% for VGG16, 6% for VGG19 and 17% for Resnet18). In relation to the other 2 augmentation schemes, only VGG19 is marginally outperformed when retrained with the Hadamard

augmentation by 1%. The specificity rates with Hadamard Augmentation outperform the SVD scheme by 2% for VGG16 and 7% for Resnet18. AlexNet performs equally with Hadamard and SVD augmentation, but Flip&Rot augmentation yields the lowest rate among the 3 augmentation schemes. Again, the sensitivity rates were maintained at their optimal rates achieved without augmentation except for AlexNet and Resnet, which marginally deteriorated by only 1%. Consequently, like the other augmentation schemes, the improved accuracy is the result of the improved specificity rate, i.e., fewer false positive benign misclassified masses while no/few more malignant cases are misclassified. From a medical point of view, this is a welcome improved performance compared to the Test1 dataset. Again, these discrepancies in the medical effects of this augmentation between Test1 and BUSI datasets may be explained by their different level of separability from Renmin in terms of the MCIQ quality assessment feature vector. In Chapter 5, we have shown that using the above 6 Hadamard matrices greatly improved MCIQ separability between two small datasets, with quality labels assessed by a radiologist, that were highly inseparable by MCIQ.

***Discussion:***

The experimental results of the above augmentation schemes for different pre-trained CNN models reveal that their impact on the generalisation is generally positive, with different levels of improvement ranging from modest to significant. Moreover, each scheme's improvement level varies for different unseen datasets. Although, the generalisation of the various CNN schemes without augmentation, as discussed in Chapter 4, was equally different for the 2 unseen datasets. For all the schemes, we raised the possibility of attributing the differences in their improved generalisation rates to their different level of separability from Renmin in terms of the MCIQ quality assessment feature vector. These results raise an important question about the validity of claiming that augmentation improves generalisability and makes the scheme less overfit. ***How many more unseen datasets must be tested to support such a claim?*** In the final section, we propose a potential scheme to mitigate the scarcity challenge for US images without changing the RoI image data content.

## 6.4 Mitigating BUS Data Scarcity by Margin Appending Schemes

All the above augmentation schemes as well as those in the literature, are equally applicable to any image modality and not specific to US diagnostic images. Several reasons necessitate the design of specific scarcity mitigating schemes that are suitable for using DL models for US tumour image analysis. First, while class discriminating features learnt by DL schemes are based on hidden patterns of image data within the tumour RoI, radiologists are trained to look out for the presence of certain tumour signs from various parts of RoI as well as its surrounding tissues. Secondly, the failure of generalisation performance of pre-trained CNN models for US image analysis to unseen data is often attributed to many factors, including the lack of standardised cropping procedures and labelling due to variation in radiologist experience and variation in US devices. In Chapter 4, we investigated the issue of optimal RoI cropping. We tested a few cropping schemes by extending the CH of a set of lesion boundary points marked by radiologists. The results show that certain CH expanding ratios improved accuracy.

Moreover, it is possible to have unseen data with different tumour margin ratios due to possible errors in the manual/automatic detection/segmentation. This provides the primary motivation for the proposal, in this section, of training DL algorithms with datasets that combine the various cropping ratios. This approach in enlarging the training datasets in a comparable manner to the above augmentation schemes but without manipulating the original RoI images helps reduce the impact of inaccurate cropping on trained model performances when tested on unseen datasets.

This solution aims to overcome the generalizability issues of pre-trained CNN models caused by variations in RoI cropping. In particular, we propose a novel data-sampling approach called the Tumour Margin Appending (TMA) scheme, which builds upon the previously proposed tumour cropping scenarios with Zero-padding discussed in Chapter 4. The emphasis of this technique is to retrain DL models that are resilient to different tumour-cropping ratios/methods employed at various medical centres. It also serves as a regularizer to reduce model overfitting when tested on unseen datasets obtained by an unknown tumour cropping procedure [9].



### 6.4.1 Optimal Tumour Cropping in Uncontrolled Scenario – Revisited

In Chapter 4, we discussed the issue of lesion cropping as an alternative to the challenge of automatic ROI segmentation, and to avoid expensive and error-prone manual segmentation, experienced radiologists marked a relatively small set of lesion boundary points for each US image. Linearly interpolating neighbouring boundary points results in a polygonal ROI shape, and the tightly circumscribed horizontal-vertical bounding box with zero padding is referred to as the TumourZ cropping scenario. To facilitate the inclusion of tissue data outside the polygonal ROI shape, we found that expanding this polygon at constant ratios leads to self-intersection for irregular polygonal shapes. Instead, we constructed the CH of the boundary points, resulting in what we called the CH-based cropping scenario, denoted by CHZ, after zero-padding its surrounding box. Since expanding tumour CH shape at different ratios does not suffer from self-intersection, we created several CH expanded cropping scenarios, denoted by  $\alpha Z$ , using several expansion ratios  $\alpha$  in the set  $\{0.6, 0.8, 1, 1.2, 1.4, \dots, 2, 2.5, 3, 3.5, 4\}$ .

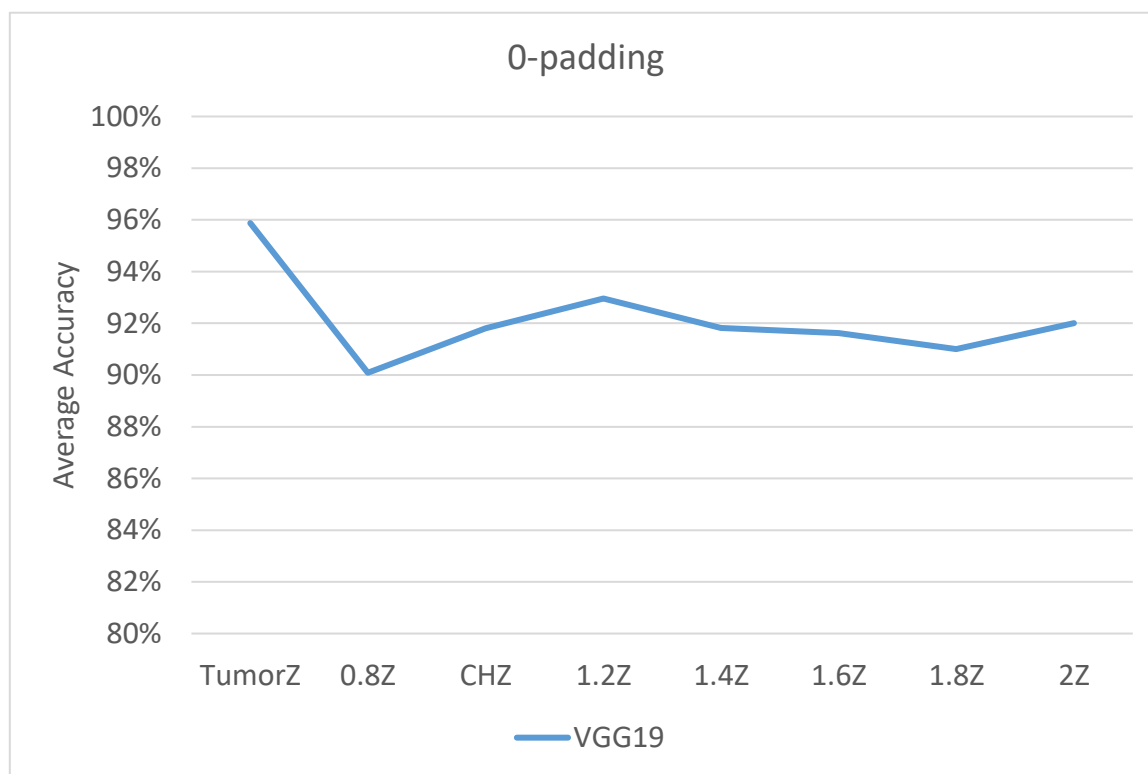


Figure 6.3 Performance of VGG19 trained and tested with cropping at the same ratio (Renmin).

After cropping at each of these ratios, we trained and tested several pre-trained CNN models in fine-tuning mode on the Renmin dataset. The averages of the 5-fold cross-validation accuracies achieved with VGG19 are displayed in Figure 6.3. The best-performing cropping scenario is TumourZ (96%), but all the others achieve accuracies in the range between 90% and 93%.

The above results show that pre-trained CNN models retrained on different single tumour margin cropping ratios and tested against the same margin appending ratio perform comparably well. In real-world applications, it may not be possible to guarantee testing will be carried out with the same standard RoI cropping ratio adopted during the creation of the CNN model. Moreover, in the absence of a reliable and efficient automatic segmentation of the tumour boundary, the cropping process becomes error-prone due to several factors. In turn, it becomes a source of lack of generalizability of models retrained with data from a standardized clinical setting. This raises the question of whether there is a cropping scenario that has an optimal performance over several cropping scenarios. For that, we conducted a new set of experiments to retrain the VGG19 model with each single cropping ratio but tested with RoIs cropped at all the proposed different ratios (uncontrolled scenario). We determined the performances of each of the corresponding models on the entire Renmin dataset. Figures 6.4 and 6.5 display subsets of these results, each displaying the performance of 3 cropping ratios. For full results, see Appendix B.

A close examination of these results shows that except for the TumourZ cropping scheme, the performance of the VGG19 model trained with any other cropping ratio maintains a reasonable performance when tested on data cropped with nearby ratios. However, the performance deteriorates when training with a single cropping ratio but testing with images cropped with far away ratios. This has also been shown to be true for other pre-trained CNN models (see [9]). These results motivate combining RoI images obtained with several cropping ratios to enlarge the training set to mitigate the US scarcity challenge.

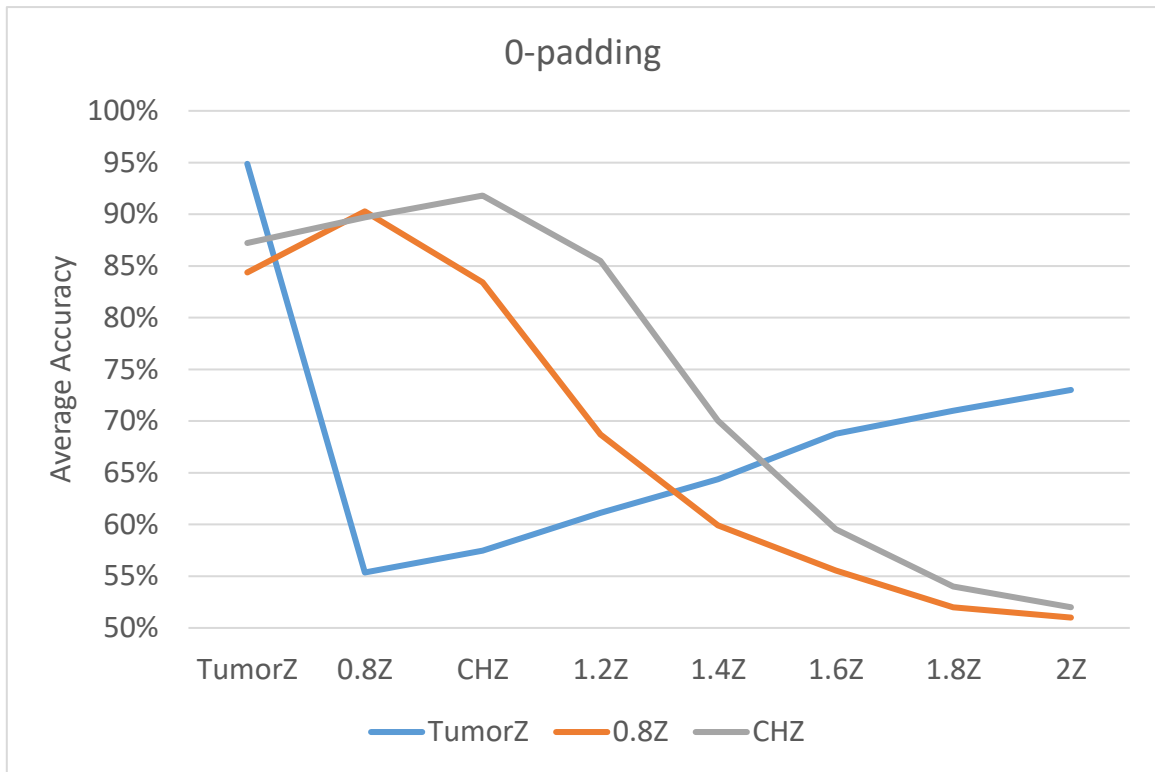


Figure 6.4 Performance of VGG19 trained with one cropping ratio and tested on all the ratios.

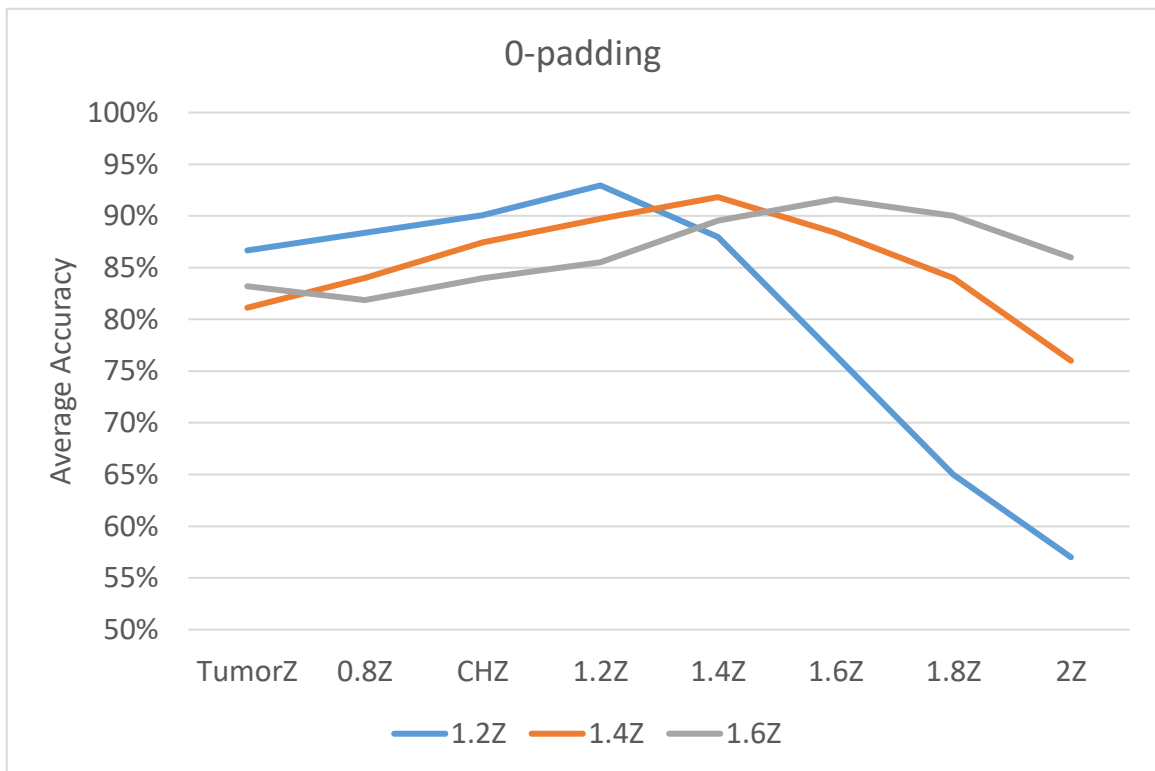


Figure 6.5 Performance of VGG19 trained with one cropping ratio and tested on all the ratios.

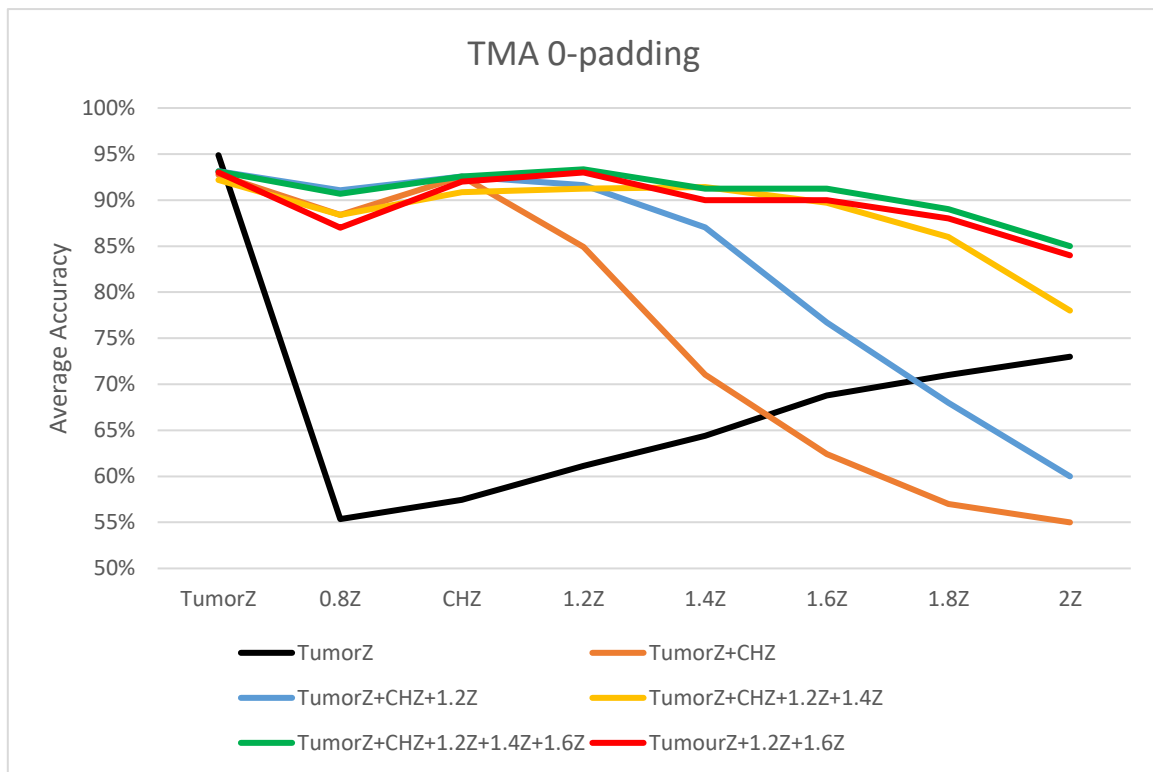
## 6.4.2 BUS Augmentation Using Tumour Margin Appending Schemes

The idea of training DL models with a single optimal margin appending ratio is impractical and cannot be guaranteed in real-world medical applications due to several factors. While it is possible to standardize the tumour margin ratio of the training dataset collected in a clinical setting that uses a controlled standard cropping scheme, it cannot be ensured for unseen datasets from different clinical centres. Moreover, the results in Figures 6.4 and 6.5 demonstrated that no single cropping scenario has optimal performance for a wide range of cropping ratios on its own. Still, several ratios exhibit locally optimal performance over a narrow range of ratios. Here, we explore the advantages of combining multiple locally optimal ratios to address the generalization issue caused by the scarcity of training datasets collected from a single clinic with a controlled RoI cropping procedure, such as the Renmin dataset. Our goal is to overcome this limitation and enhance the generalization capabilities of the models through the combination of these ratios. We propose the following US image augmentation-like 4-step process to retrain pre-trained CNN models and test their performance as follows:

1. Create versions of the training dataset with each RoI cropped at the margin appended ratios in the set  $R = \{\text{TumourZ}, 0.8Z, \text{CHZ}, 1.2Z, 1.4Z, 1.6Z, 1.8Z, 2Z\}$ .
2. Amalgamate subset A of these versions that includes TumourZ to create an enlarged dataset  $\text{Tr}(A)$ .
3. Retrain the chosen pre-trained CNN on  $\text{Tr}(A)$  in fine-tuning mode, and
4. Test the Output CNN model on the entire images in R, each scenario seperatily.

**Figure 6.6 Tumour Margin Appending steps to train/test CNN models.**

We conducted experiments by implementing this proposed process with each of the subsets:  $A = (\text{TumorZ} + \text{CHZ})$ ;  $A = (\text{TumorZ} + \text{CHZ} + 1.2Z)$ ;  $A = (\text{TumorZ} + \text{CHZ} + 1.2Z + 1.4Z)$ ;  $A = (\text{TumorZ} + \text{CHZ} + 1.2Z + 1.4Z + 1.6Z)$ ; and  $A = (\text{TumorZ} + 1.2Z + 1.6Z)$ , Figure 6.7, below, displays the results of all these experiments besides that of no augmentation.



**Figure 6.7 Performance of VGG19 trained with TMA augmented datasets.**

These results demonstrate that compared to no margin appending, all the TMA configuration schemes yield improved classification performance within a longer range of ratios than their individual components, and few of these schemes do that in a stable style. In short, TMA augmentation configurations can increase tolerance to significant RoI cropping errors that may happen for various reasons. The best-performing sample augmentation scheme was found to be (TumorZ + CHZ + 1.2Z + 1.4Z + 1.6Z), but the shorter configuration (TumorZ + 1.2Z + 1.6Z) has almost similar but marginally lower performance throughout the full range. For practical reasons, keeping the configuration at a reasonable size is recommended. However, increasing the number of augmentation ratios within the suggested range may provide sufficient data samples to build CNN models from scratch. In this case, besides enabling tolerance to cropping errors, such schemes learn additional hidden patterns of features from different surrounding regions that clinicians often rely on to generate supporting evidence for their decisions regarding signs of malignancy. However, more sophisticated schemes of inclusion of surrounding regions would be necessary to avoid relying on irrelevant or performance-decreasing information.

Determining the impact of these schemes on generalisability by testing performance on any external dataset is not necessary unless for datasets of US images that practice cropping in

significantly outside these ranges. In fact, if a dataset practices erroneous cropping at ratios with these reasonable ranges, then the failure of a CNN model to generalise to such a database cannot be attributed to this issue. The 2 external datasets, Test1 and BUSI, have been cropped in a similar way to that of the Renmin dataset.

This investigation highlights that CNN models trained using sample augmentation at multiple ratios are robust against various cropping scenarios that may occur in different hospitals. These schemes act as a regularizer to reduce some, but not all, causes of overfitting that relate to cropping errors and to some extent, remove the necessity of optimal tumour segmentation.

## 6.5 Conclusion

In this chapter, we investigated the challenge of US data scarcity, identified in Chapter 3, as an important factor influencing the performance of pre-trained CNN models when retrained with US tumour images. In the literature, this factor is usually considered a source of overfitting and image augmentation techniques are promoted as reasonably successful mitigating solutions. We found a significant number of image augmentation schemes in the literature and added two novel augmentation schemes. Most of these schemes, including the use of GAN networks, generate new image versions by image transformation/manipulation. It was shown that our simplified/limited versions of the schemes (SVD and Hadamard), along with the conventional Flip&Rot scheme, effectively enhanced the model's generalization capability to the two external datasets. However, the extent of improvement varied across different CNN architectures and differed for the two datasets. More importantly, since none of these schemes is specific to US images, it is difficult to determine which causes of lack of generalization these augmentation schemes helped mitigate. Chapter 4's investigation into determining optimal RoI tumour cropping inspired this Chapter's research. We aimed to enhance CNN learning by emulating the learning process of trainee radiologists who analyse various regions within the RoI, including its border and surrounding areas, to identify signs of tumour malignancy. We devised the TMA strategy to address the challenge of limited data and variations in RoI cropping practices. This strategy combines multiple locally optimal cropping ratios to expand scarce US training datasets and improve generalization capabilities. It does not only address the scarcity issue specific to US data but also distinguishes itself from other augmentation schemes by effectively mitigating a specific cause of low generalization to unseen datasets without the need to test on a large number of external datasets.

# Chapter 7: Conclusion and Future Research

## Challenges

Breast cancer remains a critical global health concern, demanding accurate diagnosis and early detection to enhance patients' treatment and quality of life. Significant advancements in medical sciences have helped improve patients' survival, especially for early detected cases. The acute shortage of well-trained radiologists and oncology clinicians that most healthcare systems suffer from, besides the vast cost implications, is becoming an obstacle to benefiting from the related significant advances in medicine. The considerable advances in ML algorithms for computer vision, particularly the recent exciting success of Deep Learning, are attracting significant interest as a supporting analytical tool for tumour diagnostics from medical image scans. In this thesis, we aimed to contribute to global research activities for leveraging DL techniques, particularly CNNs, for the automatic and reliable classification of breast lesions in US images. We faced many serious challenges in developing such schemes, many of which may be attributed to differences between the content and structure of US images and those of natural image datasets with which most existing CNN models have been trained. However, other challenges emanate from the critical nature of medical diagnostics compared to many computer vision classification tasks. The most critical challenges include (1) the difficulty of interpreting CNN decisions, (2) the problem of overfitting that relates to training a ML model with data from a given clinical setting and failure to generalise its performance to unseen data recorded in other clinical settings, and (3) robustness against adversarial/natural image data perturbation. Our investigations were concerned with the second challenge of overfitting and focused on issues relating to the nature/properties of BUS datasets in terms of CNN model requirements on the size of the training dataset, class sample diversity, and adequacy of input image quality. The level of adherence to these requirements influences the performance of any CNN models developed for BUS and is also strongly impacted by the specificity of US images compared to natural images. Training well-performing CNN models from scratch requires considerably large training datasets well beyond the availability of credibly labelled BUS datasets. We adopted the fine-tuning process to retrain CNN models pre-trained with natural images. Section 7.1 will present the main conclusions from the last 4 Chapters. Our investigations also generated many pilot research projects indirectly related to the thesis objectives and raised future challenges described in Section 7.2.

## 7.1 Main Conclusions

Our initial exploration of the fields of ML for image analysis revealed several US-specific factors that impact tumour diagnostics decisions. Besides the scarcity of reliably labelled US tumour scan images, these factors include considerable variation in RoI sizes within both benign and malignant masses, inter- and intra-observer variability due to differences in clinician expertise and diversity of deployed US devices that, in turn, result in different clinical practices in cropping RoIs as well as different image quality associated with the level of tolerated speckle noise and contrast level. Acquiring a diverse and well-annotated US tumour training image dataset is widely acknowledged as an essential requirement, even for retraining state-of-the-art pre-trained CNNs, for ensuring good performing models that are generalisable to unseen data and robust against adversarial/natural data tampering.

We discovered that the extent of RoI size variation within BUS datasets is significant and poses a tough challenge for meeting the fixed-size input image requirements of DL models without compromising image quality. We also found that variation of RoI size occurs in both classes, although benign cases are often smaller than most malignant cases. This discrepancy adversely affects the quality of benign cases when using conventional resizing techniques like BiCubic. To address this issue, we adopted a CSSR resizing procedure known to improve the natural image quality of degraded low-resolution. While CSSR marginally enhanced the performance of CNN models and reduced the probability of misclassifying benign cases by 10%, it also improved the perceptual quality of resized RoIs, as confirmed by an experienced radiologist. The marginality of improved performance due to using CSSR instead of BiCubic raised questions about the suitability of using human-perceived image quality metrics that are suitable for assessing the quality of natural images for evaluating the quality of US images. In fact, our literature review highlighted an anomaly between the perceived low quality of medical US images and the lack of robust and standardized IQA tools for US images. This issue needs more investigations to understand the nature of US image distortion.

To determine the impact of the significant variation of the distribution of US RoI tumour sizes on the performance of the various pre-trained CNN models, it was essential to adopt standardised RoI lesion cropping schemes. This was necessitated by the fact that only a set of RoI boundary points were labelled by the radiologist instead of time-consuming manual/automatic segmentation. The construction of the CH of the lesion border marked points, besides being computationally efficient and easy to expand, helped minimise the



exclusion of lesion pixels compared to other methods. It also facilitated several cropping schemes via parallel expansion and different scenarios of padding the region between the surrounding rectangular box and the tumour polygonal area. Tissue padding of several expanded CH schemes (e.g., the 1.2T obtained by expanding CH by 1.2 times) led to improved performance but only marginally compared to the zero padding of these schemes. Among the proposed RoI cropping/padding scenarios, the TumourZ RoI scheme emerged as the optimal choice for CNN model performance, which also performed well for HC feature schemes. From these experiments, one can conclude the reasonable potential/promise of enhanced model performance by the inclusion of some RoI external tissue pixels in the immediate surroundings of the lesion border. However, for both padding scenarios, the trained models on the Renmin dataset have low to very low generalisation when tested on the 2 unseen datasets, i.e., we have an instance of overfitting.

The training set is neither large enough, nor its samples are adequately diverse being recorded in a single clinical centre. To test if using a BUS dataset recorded in a single clinical centre accounts fully for the overfitting of the models, we repeated the experiments by training the pre-trained CNN models with the reasonably larger multi-centre Modelling dataset that includes the above-recorded BUS images, besides samples recorded in 4 additional clinical centres. The results confirmed that the TumourZ scenario maintained the validation performance but yielded superior generalization for the 2 unseen datasets. The Grad-CAM visualization tool further supported the decision quality of TumourZ in that the decisions were more reliant on the data in the tumour interior region.

Having discovered that most IQA metrics designed for natural images fail to assess medical US images effectively, we investigated methods to address this problem by developing tools tailored explicitly for BUS images. Visual examination of many typical US images revealed that the distribution of illumination and contrast across different US regions vary significantly compared to good quality natural images. We thus opted to calculate the correlation between local statistical parameters across US image blocks to design a Multi Characteristic Image Quality feature vector (MCIQ) that captures the spatial distribution of existing quality metrics. The MCIQ demonstrated good tumour class dependency and a significant ability to distinguish different image sources and datasets. Unfortunately, the MCIQ assessment did not align with the binary quality assessment of a small dataset labelled by an expert radiologist as Good and Bad images. We then developed an advanced version of MCIQ simply by utilizing image convolutions with a small set of well-conditioned Hadamard filters, which resolved the above misalignment problem. It has helped overcome

limitations arising from the scarcity of quality-standardized radiology expert-labelled US images and the limited knowledge of distortions in US images.

We finally re-investigated the impact of the scarcity of US images on the performance of the pre-trained CNN models, but instead of using the multi-centre data expansion, we explored various image augmentation techniques with the Renmin training dataset. We examined the effect of several commonly used conventional augmentations, including Flip&Rot and spectral-based augmentation using a limited version of SVD. Inspired by the success of the Hadamard-based modification of MCIQ, we utilized it for designing an effective image augmentation scheme. All these schemes are equally applicable to images of any modality. The experimental results demonstrated that these augmentation approaches improved the generalization of the various pre-trained CNN models to the 2 external datasets. However, these welcome successes are not explainable, and it is unclear what causes of overfitting these schemes help overcome. There are several reasons that necessitate the design of specific scarcity mitigating schemes that are suitable for enabling the use of DL models for US tumour image analysis. In fact, there is no guarantee that these schemes could work with other than these two external datasets. The earlier results that demonstrated the success of certain CH expanded cropping schemes did maintain good performance for nearby CH expanded cropped RoIs provided the main motivation to propose a novel data-sampling approach called the Tumour Margin Appending (TMA) scheme to enlarge small training datasets by including a combination of the various cropping ratios. This method is specific to US tumour images and distinct from other augmentation schemes. It relies on image data from the surrounding regions of the RoI, which radiologists usually use in their diagnostic decisions. The corresponding experimental results showed that TMA effectively mitigated the lack of generalization. It helped CNN models to be robust against various cropping scenarios when testing unseen samples with unknown RoI cropping schemes.

In conclusion, our research has showcased the potential of CNN models to enhance the classification of breast lesions in US images and avoid certain causes of model overfitting. By investigating factors such as RoI size normalization, resolution enhancement, optimal lesion cropping, and image augmentation, we have provided valuable insights and methodologies to improve CNN-based classification systems' accuracy and generalization capabilities. Our findings underscore the importance of addressing challenges related to RoI size variation, US-IQA, and scarcity of US images for pre-trained CNN models.

## 7.2 Future Research Challenges

The insights and methodologies presented in this thesis serve as a foundation for future research and advancements in applying DL techniques to BUS analysis. While conducting our investigations, and in light of rapid advancement in AI applications into computer vision, we also investigated several issues and topics linked to the task of using CNN for BUS but not strictly specific to the objectives of this thesis. Furthermore, several mathematical challenges naturally arose during the life of this thesis that could have more relevance to image/data analysis beyond the objectives of this thesis. These include issues with the need for size-adaptive CS tools for resizing images to appropriate fixed resolutions. The results of some of these additional investigations are not included in the previous chapters but open the way for new research projects targeting US image analysis, with more emphasis on the content and semantics of US images, not only for breast tumours. Below, we shall list these proposed projects and the obtained results, together with hints on the intended future approaches.

### 7.2.1 Hybridisation of Multiple HC and Pre-trained CNN Models

The concept of Hybrid ML models of data/image analysis relates to combining/fusing several ML models designed to solve a classification problem for improved performance or exploit the individual models' complementarity. The fusing of several HC texture features for pattern recognition has been very popular for a long time [6]. Combined deep features from a pre-trained VGG19 model with HC texture and morphological features in [35] have also been recently proposed for BUS tumour classification.

This proposed future project is designed to select a few HC texture feature classification schemes and a few pre-trained CNN models, and we investigate methods of combining them for improved classification and generalization ability, including:

1. Design a special scheme to fuse selected HC feature maps extracted post convolution layers to improve classification performance and provide visualisation tools that help interpret any CNN Model decisions.
2. Fuse the deep features of CNN models at the score level for predicting the class of US tumour image and use the HC schemes to interpret the predicted decisions.

**Figure 7.1 The proposed feature hybridisation methods.**

These suggestions benefit from 2 investigations that we carried out during the thesis project; the first was initially aimed to determine what kind of image texture features are learnt by the pre-trained CNN models, while the second was a pilot study to fuse decisions of the pre-trained CNN models using their output scores. Next, we describe these two investigations.

### 7.2.1.1 Empowering Handcrafted texture features by CNN convolution filters

Explaining CNN models' predictions is an essential reassuring element for deploying such models in medical image analysis. Identifying image statistical/textural features learnt by CNN architectures during training helps develop visualization tools for explaining their decisions. The most relevant features are those recognisable by the Human Vision System (HVS). This investigation was concerned with determining if CNN models learn HVS image texture features from BUS images.

We started by investigating the effect of the various convolution layers steps, of AlexNet architecture in fine-tuning mode, on selected statistical/textural features, including Mean, Variance, Entropy, HOG, GLCM, and ULBP, applied to BUS images. Each step of the convolution layers results in changing the various statistical/textural features depending on the entries of the layers' filters. For each of the selected textural/statistical features, we follow the same approach to create a feature map at each step of each of the convolutional layers by concatenating the extracted features from each version of the convolved image (i.e., from each channel of the extracted activation for the corresponding layer). Figure 7.2, below, illustrates the construction of these feature maps.

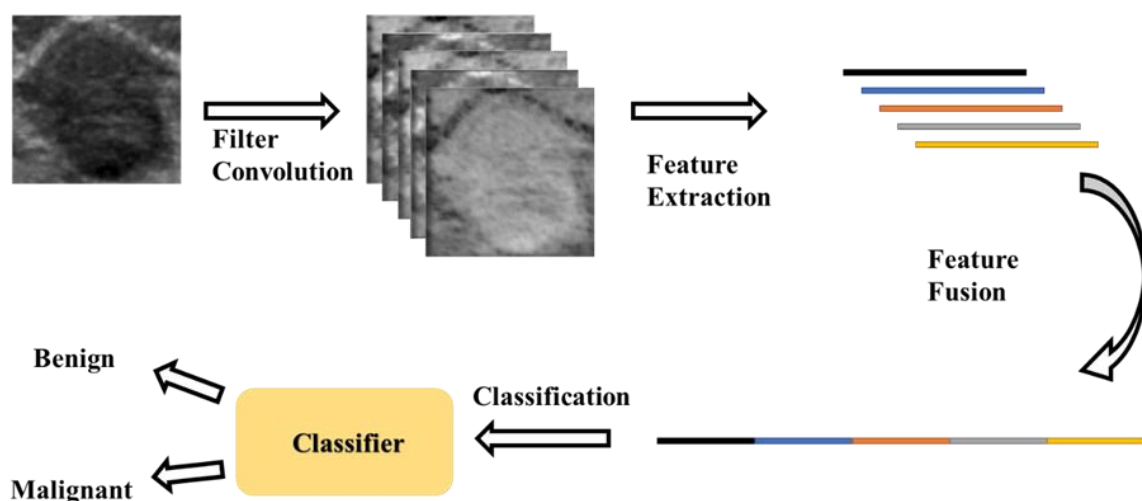


Figure 7.2 Texture Feature Augmentation with Random Filters.

We adopted the AlexNet CNN architecture due to its simplicity, but these could be repeated for any CNN architecture. At each step of each convolutional layer, the output feature maps were fed into the linear SVM classifier, and the performance of the selected HC feature maps was determined in accordance with the 5-fold cross-validation protocol. The experimental accuracies presented below are the averages of test accuracy in the 5 folds experiments.

Table 7.1, below, displays the experimental results for the 6 image statistical and textural descriptors. We also include the classification results when we used AlexNet for classification at the bottom of the table. In this case, and before passing to the FCL, each input image is represented by a feature map constructed by concatenating flattened matrices output from the last convolution layer. The Input layer is simply the performance of the selected textural/statistical descriptor extracted from the original images prior to input into the chosen CNN model.

**Table 7.1 Classification of Several HC texture features post-convolution with fine-tuned AlexNet.**

AlexNet	# Channels	Mean	Variance	Entropy	HOG	GLCM	ULBP
<b>Input Layer</b>	<b>1</b>	<b>50%</b>	<b>51%</b>	<b>49%</b>	<b>85%</b>	<b>76%</b>	<b>84%</b>
<b>conv1</b>	<b>96</b>	<b>70%</b>	<b>90%</b>	<b>91%</b>	<b>89%</b>	<b>89%</b>	<b>89%</b>
<b>relu1</b>	<b>96</b>	<b>90%</b>	<b>90%</b>	<b>91%</b>	<b>91%</b>	<b>90%</b>	<b>90%</b>
<b>norm1</b>	<b>96</b>	<b>90%</b>	<b>90%</b>	<b>91%</b>	<b>91%</b>	<b>89%</b>	<b>89%</b>
<b>pool1</b>	<b>96</b>	<b>91%</b>	<b>90%</b>	<b>91%</b>	<b>91%</b>	<b>90%</b>	<b>89%</b>
<b>conv2</b>	<b>256</b>	<b>90%</b>	<b>93%</b>	<b>90%</b>	<b>92%</b>	<b>92%</b>	<b>92%</b>
<b>relu2</b>	<b>256</b>	<b>92%</b>	<b>93%</b>	<b>91%</b>	<b>92%</b>	<b>92%</b>	<b>91%</b>
<b>norm2</b>	<b>256</b>	<b>93%</b>	<b>93%</b>	<b>90%</b>	<b>91%</b>	<b>92%</b>	<b>91%</b>
<b>pool2</b>	<b>256</b>	<b>93%</b>	<b>93%</b>	<b>89%</b>	<b>92%</b>	<b>91%</b>	<b>91%</b>
<b>conv3</b>	<b>384</b>	<b>91%</b>	<b>95%</b>	<b>92%</b>	<b>91%</b>	<b>93%</b>	<b>90%</b>
<b>relu3</b>	<b>384</b>	<b>94%</b>	<b>94%</b>	<b>91%</b>	<b>90%</b>	<b>93%</b>	<b>92%</b>
<b>conv4</b>	<b>384</b>	<b>93%</b>	<b>94%</b>	<b>91%</b>	<b>90%</b>	<b>91%</b>	<b>89%</b>
<b>relu4</b>	<b>384</b>	<b>93%</b>	<b>93%</b>	<b>92%</b>	<b>91%</b>	<b>94%</b>	<b>91%</b>
<b>conv5</b>	<b>256</b>	<b>94%</b>	<b>94%</b>	<b>93%</b>	<b>90%</b>	<b>92%</b>	<b>90%</b>
<b>relu5</b>	<b>256</b>	<b>93%</b>	<b>95%</b>	<b>94%</b>	<b>91%</b>	<b>94%</b>	<b>92%</b>
<b>pool5</b>	<b>256</b>	<b>94%</b>	<b>94%</b>	<b>92%</b>	<b>91%</b>	<b>91%</b>	<b>90%</b>
<b>AlexNet</b>		<b>93%</b>	<b>93%</b>	<b>93%</b>	<b>93%</b>	<b>93%</b>	<b>93%</b>

At the input layer, the feature vectors perform differently; the performance of the statistical ones is random, around 50%, which means these 1-dimensional feature vectors do not have any discriminating power to classify benign tumours from malignant cases. However, soon after convolving the images with the convolutions of the first layer, they all acquire good to high discriminating power. In contrast, the other texture features already have reasonably good accuracy (85% for HOG, 76% for GLCM, and 84% for ULBP) at the Input Layer.

Generally, the table shows that the discriminating power of all the HC schemes started to improve as one progresses successively through the following layers.

We found no significant difference between the performance of the features extracted from the pre-trained AlexNet's last convolution layer compared to the fine-tuned AlexNet. These results raise strong doubts if the decisions of the fine-tuned AlexNet for the BUS images are informed by the extracted HC features. This justifies our conclusion that CNN models may not learn HSV-recognised textural/statistical features.

These experiments also show that augmenting (i.e., concatenating) such HC statistical/texture features post-convolution produces higher dimensional feature maps that boost discrimination power, especially in the case of using single statistic parameters. In short, convolution empowers even the least class-discriminating HC features. The widely accepted assertion that decisions of HC statistical/texture image analysis schemes are amenable to interpretation justifies their use to interpret the decisions of the CNN models.

### **7.2.1.2 Incremental Fusion of CNN Models**

The other component of the hybridisation project is motivated by a pilot study we conducted to mimic the commonly adopted approach of fusing HC classification schemes for improved performance. There are several ways to extract and combine deep features from several CNN models. The most decision-relevant feature in CNN models is the 2-D variable scores outcome of their last FCL that encapsulates all the discriminating power. Thus, we opted for the score level fusion as the easiest to implement regardless of the individual CNN model architectures. In the pilot study, we trained the 6 well-known state-of-the-art CNN models (AlexNet, VGG16, ResNet50, InceptionV3, Xception, and DensNet201) on the modelling dataset and tested them on the 2 external testing datasets: Test1 and BUSI. We experimented with the extracted score features for 6 incrementally fused CNN models as follows:

- 1) AlexNet, (2-D)
- 2) AlexNet, VGG16- (4-D)
- 3) AlexNet, VGG16, ResNet50 - (6-D)
- 4) AlexNet, VGG16, ResNet50, InceptionV3- (8-D)
- 5) AlexNet, VGG16, ResNet50, InceptionV3, Xception - (10-D)
- 6) AlexNet, VGG16, ResNet50, InceptionV3, Xception, DensNet201 - (12-D)

In each case, the concatenated 2-D score features of the constituent CNN models are trained with the SVM cubic kernel classifier and tested on both external datasets. Table 7.2 presents

the experimental results, which demonstrate the success of this fusion approach. For both external datasets, the 4<sup>th</sup> fused DL achieves the highest classification performance of 96% for Test1 and 93% for BUSI datasets. By no means this is the only fused combination that achieves optimal performance.

**Table 7.2 Classification performance of the deep fused features with Cubic SVM.**

Test1	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
1	0.78 ± 0.26	0.85 ± 0.09	0.74 ± 0.41	0.79 ± 0.17	0.80 ± 0.21
2	0.93 ± 0.00	0.95 ± 0.00	0.91 ± 0.00	0.91 ± 0.00	0.93 ± 0.00
3	0.95 ± 0.00	0.96 ± 0.00	0.94 ± 0.00	0.93 ± 0.00	0.95 ± 0.00
4	0.96 ± 0.00	0.96 ± 0.00	0.95 ± 0.00	0.94 ± 0.00	0.96 ± 0.00
5	0.96 ± 0.00	0.96 ± 0.00	0.96 ± 0.00	0.94 ± 0.00	0.96 ± 0.00
6	0.96 ± 0.00	0.95 ± 0.00	0.96 ± 0.00	0.94 ± 0.00	0.96 ± 0.00
BUSI	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
1	0.75 ± 0.32	0.83 ± 0.19	0.70 ± 0.39	0.74 ± 0.25	0.77 ± 0.28
2	0.90 ± 0.00	0.95 ± 0.00	0.87 ± 0.00	0.87 ± 0.00	0.91 ± 0.00
3	0.91 ± 0.00	0.93 ± 0.00	0.90 ± 0.00	0.88 ± 0.00	0.92 ± 0.00
4	0.93 ± 0.00	0.93 ± 0.01	0.93 ± 0.00	0.90 ± 0.00	0.93 ± 0.00
5	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.90 ± 0.00	0.93 ± 0.00
6	0.93 ± 0.00	0.92 ± 0.00	0.93 ± 0.00	0.89 ± 0.00	0.92 ± 0.00

One of the shortcomings of the above approach is that training 4 or more large CNN models is computationally expensive. In the future, we aim to design lightweight CNN models on the augmented datasets and fuse their decision scores to achieve similar performance improvements. Moreover, this opens the door for further feature fusion and diversification using deep and HC feature schemes.

### 7.2.2 Optimal Cropping – Revisited

In Chapter 4, we identified the TumourZ cropping scenario as optimal RoI cropping/appending that helped improve the generalization of CNN models. However, this cropping scenario does not include any posterior tumour regions that encapsulate tumour class discriminating signatures. According to BI-RADS, there are breast tumour posterior signs that indicate the level of suspicion for malignancy [154]. The posterior feature describes the echogenicity (intensity) effects of the posterior area, i.e., the area immediately underneath the lesion. The posterior area may also encapsulate *enhancement*, *shadowing*, *combined echo*, or *no posterior*; see Figure 7.3.

Enhancement is detected when the posterior area appears brighter than the adjacent areas, while shadowing is manifested by the posterior area appearing darker than the adjacent areas.

In an US tumour image, we may have a combined posterior when there is more than one posterior or no posterior when the echogenicity is similar to the adjacent areas. Shadowing and enhancement are the most critical features of breast lesions as they are generally accepted as signs of highly suspicious malignancy and benignity, respectively [154], [155].

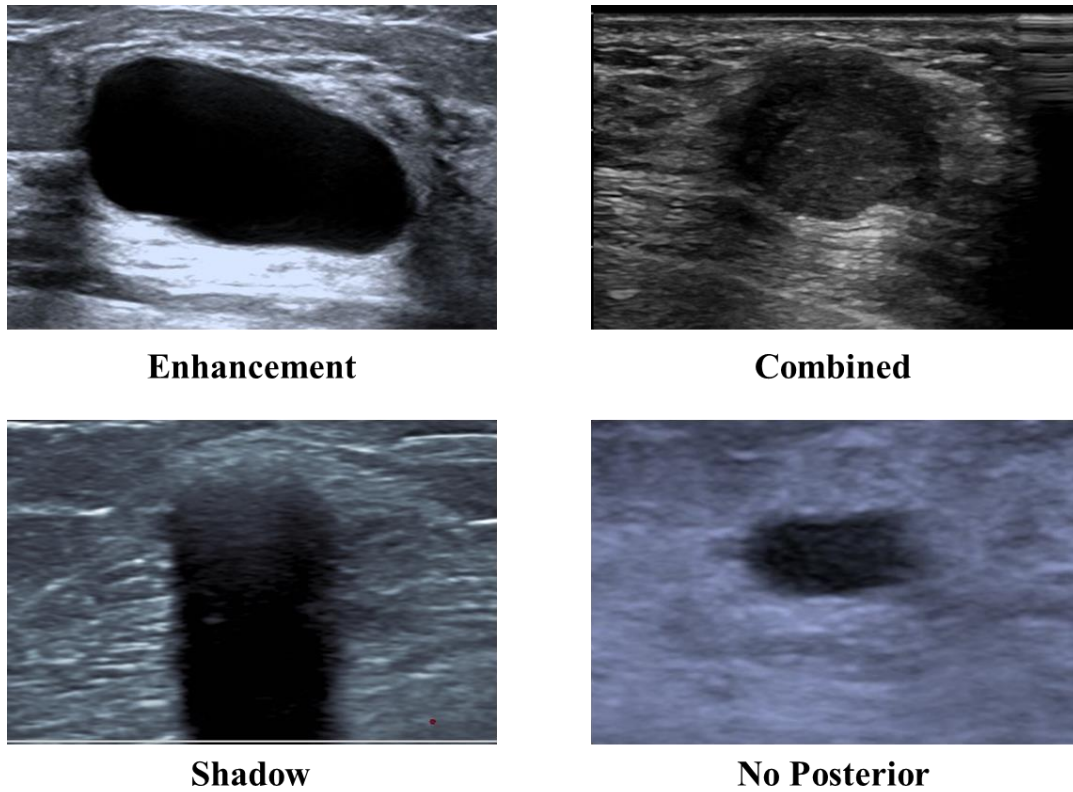


Figure 7.3 The four types of tumour posterior features

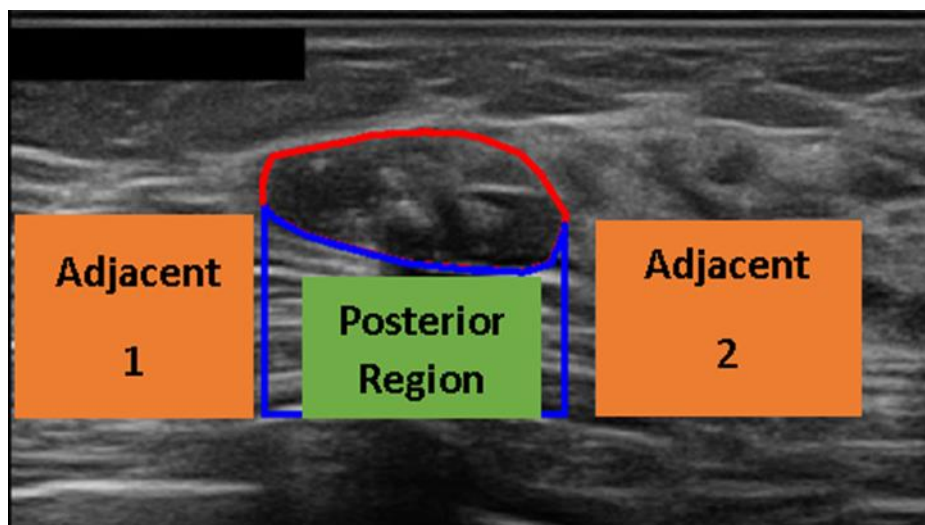


Figure 7.4 A breast tumour with its posterior region



In the future, instead of only upscaling tumour CH in all directions (i.e., appending the tumour area all around equally), we shall attempt to include the important posterior regions in the cropped RoI TumourZ to empower the class-discriminating features. However, we need RoI ground truth to do that, which an experienced radiologist can provide. In this case, the targeted posterior area can be included in the cropped RoI to complement the TumourZ scenario, see Figure 7.4.

### **7.2.3 Extending MCIQ for US-IQA Aligned with Expert Quality Labelling**

Several possible directions exist to expand the current work on the MCIQ feature vector. One such approach can be made by expanding the Hadamard-based augmentation to include different size Hadamard-based filters (7x7, 11x11, and 15x15). But another possibility is extending the 50 coordinates of MCIQ by using the block-wise distributions of other HC texture features from the original or convolved images. It is important to make this image quality more specific to US images.

# Appendix A

Here, we present the results of the experimentation we conducted, but not presented in Chapter 6. The experiments aimed at evaluating the performance of the 6 state-of-the-art pre-trained CNN models on the augmented Modelling dataset of BUS tumour scan images. We employed augmentation techniques, including Flip&Rot, SVD, and Hadamard, and assessed the models' performance on the unseen datasets Test1 and BUSI. We remind the reader that the findings from our Chapter 4 experimental analysis indicate that the retraining of the 6 pre-trained CNN models using the BUS Modelling dataset successfully mitigated concerns regarding their generalization capability to external datasets. After augmenting the Modelling dataset, we observed minor variations in their performance on these external datasets. Detailed results are presented in Tables 7.3, 7.4, and 7.5 for the Flip&Rot, SVD, and Hadamard augmentation techniques, respectively.

**Table 7.3 Generalisation of pre-trained CNNs retrained with Flip&Rot-Augmented Modelling dataset.**

Test1	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.91 <math>\pm</math> 0.01</b>	<b>0.90 <math>\pm</math> 0.06</b>	<b>0.92 <math>\pm</math> 0.03</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.02</b>
VGG16	<b>0.90 <math>\pm</math> 0.02</b>	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.04</b>	<b>0.88 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.01</b>
ResNet50	<b>0.89 <math>\pm</math> 0.01</b>	<b>0.89 <math>\pm</math> 0.03</b>	<b>0.89 <math>\pm</math> 0.01</b>	<b>0.86 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.01</b>
InceptionV3	<b>0.90 <math>\pm</math> 0.01</b>	<b>0.90 <math>\pm</math> 0.03</b>	<b>0.89 <math>\pm</math> 0.01</b>	<b>0.87 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.02</b>
Xception	<b>0.86 <math>\pm</math> 0.02</b>	<b>0.80 <math>\pm</math> 0.05</b>	<b>0.90 <math>\pm</math> 0.02</b>	<b>0.81 <math>\pm</math> 0.02</b>	<b>0.85 <math>\pm</math> 0.02</b>
DenseNet201	<b>0.88 <math>\pm</math> 0.01</b>	<b>0.86 <math>\pm</math> 0.03</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.84 <math>\pm</math> 0.01</b>	<b>0.88 <math>\pm</math> 0.01</b>
BUSI	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.89 <math>\pm</math> 0.03</b>	<b>0.94 <math>\pm</math> 0.03</b>	<b>0.86 <math>\pm</math> 0.04</b>	<b>0.86 <math>\pm</math> 0.03</b>	<b>0.90 <math>\pm</math> 0.02</b>
VGG16	<b>0.87 <math>\pm</math> 0.02</b>	<b>0.86 <math>\pm</math> 0.02</b>	<b>0.87 <math>\pm</math> 0.03</b>	<b>0.82 <math>\pm</math> 0.02</b>	<b>0.86 <math>\pm</math> 0.01</b>
ResNet50	<b>0.87 <math>\pm</math> 0.01</b>	<b>0.86 <math>\pm</math> 0.06</b>	<b>0.88 <math>\pm</math> 0.03</b>	<b>0.83 <math>\pm</math> 0.02</b>	<b>0.87 <math>\pm</math> 0.02</b>
InceptionV3	<b>0.88 <math>\pm</math> 0.01</b>	<b>0.86 <math>\pm</math> 0.03</b>	<b>0.90 <math>\pm</math> 0.01</b>	<b>0.84 <math>\pm</math> 0.02</b>	<b>0.88 <math>\pm</math> 0.02</b>
Xception	<b>0.91 <math>\pm</math> 0.01</b>	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.02</b>	<b>0.88 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.01</b>
DenseNet201	<b>0.86 <math>\pm</math> 0.02</b>	<b>0.86 <math>\pm</math> 0.04</b>	<b>0.86 <math>\pm</math> 0.04</b>	<b>0.81 <math>\pm</math> 0.03</b>	<b>0.86 <math>\pm</math> 0.02</b>

Table 7.4 Generalisation of pre-trained CNNs retrained with SVD-Augmented Modelling dataset.

Test1	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.03</b>	<b>0.94 <math>\pm</math> 0.03</b>	<b>0.90 <math>\pm</math> 0.02</b>	<b>0.92 <math>\pm</math> 0.02</b>
VGG16	<b>0.92 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.04</b>	<b>0.92 <math>\pm</math> 0.03</b>	<b>0.90 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.01</b>
ResNet50	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.92 <math>\pm</math> 0.04</b>	<b>0.93 <math>\pm</math> 0.01</b>	<b>0.90 <math>\pm</math> 0.02</b>	<b>0.92 <math>\pm</math> 0.02</b>
InceptionV3	<b>0.91 <math>\pm</math> 0.01</b>	<b>0.89 <math>\pm</math> 0.03</b>	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.88 <math>\pm</math> 0.01</b>	<b>0.91 <math>\pm</math> 0.01</b>
Xception	<b>0.85 <math>\pm</math> 0.02</b>	<b>0.75 <math>\pm</math> 0.04</b>	<b>0.91 <math>\pm</math> 0.03</b>	<b>0.79 <math>\pm</math> 0.02</b>	<b>0.83 <math>\pm</math> 0.02</b>
DenseNet201	<b>0.91 <math>\pm</math> 0.01</b>	<b>0.91 <math>\pm</math> 0.05</b>	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.02</b>
BUSI	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.91 <math>\pm</math> 0.01</b>	<b>0.95 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.88 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.01</b>
VGG16	<b>0.88 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.02</b>	<b>0.87 <math>\pm</math> 0.04</b>	<b>0.84 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.02</b>
ResNet50	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.88 <math>\pm</math> 0.02</b>	<b>0.86 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.02</b>
InceptionV3	<b>0.90 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.04</b>	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.87 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.02</b>
Xception	<b>0.93 <math>\pm</math> 0.00</b>	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.94 <math>\pm</math> 0.01</b>	<b>0.90 <math>\pm</math> 0.01</b>	<b>0.93 <math>\pm</math> 0.00</b>
DenseNet201	<b>0.88 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.04</b>	<b>0.87 <math>\pm</math> 0.05</b>	<b>0.83 <math>\pm</math> 0.02</b>	<b>0.88 <math>\pm</math> 0.01</b>

Table 7.5 Generalisation of pre-trained CNNs retrained with Hadamard-Augmented Modelling dataset.

Test1	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.93 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.94 <math>\pm</math> 0.01</b>	<b>0.91 <math>\pm</math> 0.01</b>	<b>0.93 <math>\pm</math> 0.01</b>
VGG16	<b>0.93 <math>\pm</math> 0.01</b>	<b>0.93 <math>\pm</math> 0.04</b>	<b>0.92 <math>\pm</math> 0.01</b>	<b>0.90 <math>\pm</math> 0.02</b>	<b>0.93 <math>\pm</math> 0.02</b>
ResNet50	<b>0.93 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.93 <math>\pm</math> 0.01</b>
InceptionV3	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.90 <math>\pm</math> 0.05</b>	<b>0.93 <math>\pm</math> 0.01</b>	<b>0.90 <math>\pm</math> 0.03</b>	<b>0.92 <math>\pm</math> 0.02</b>
Xception	<b>0.88 <math>\pm</math> 0.01</b>	<b>0.84 <math>\pm</math> 0.03</b>	<b>0.91 <math>\pm</math> 0.00</b>	<b>0.84 <math>\pm</math> 0.02</b>	<b>0.87 <math>\pm</math> 0.02</b>
DenseNet201	<b>0.93 <math>\pm</math> 0.01</b>	<b>0.93 <math>\pm</math> 0.05</b>	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.93 <math>\pm</math> 0.02</b>
BUSI	Accuracy	Sensitivity	Specificity	F1-score	AUC
	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD	AVG $\pm$ STD
AlexNet	<b>0.91 <math>\pm</math> 0.01</b>	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.87 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.01</b>
VGG16	<b>0.88 <math>\pm</math> 0.03</b>	<b>0.90 <math>\pm</math> 0.03</b>	<b>0.87 <math>\pm</math> 0.03</b>	<b>0.84 <math>\pm</math> 0.04</b>	<b>0.88 <math>\pm</math> 0.03</b>
ResNet50	<b>0.88 <math>\pm</math> 0.00</b>	<b>0.91 <math>\pm</math> 0.02</b>	<b>0.85 <math>\pm</math> 0.01</b>	<b>0.84 <math>\pm</math> 0.00</b>	<b>0.88 <math>\pm</math> 0.00</b>
InceptionV3	<b>0.92 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.02</b>	<b>0.92 <math>\pm</math> 0.01</b>	<b>0.88 <math>\pm</math> 0.01</b>	<b>0.92 <math>\pm</math> 0.01</b>
Xception	<b>0.90 <math>\pm</math> 0.00</b>	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.89 <math>\pm</math> 0.01</b>	<b>0.87 <math>\pm</math> 0.01</b>	<b>0.91 <math>\pm</math> 0.01</b>
DenseNet201	<b>0.88 <math>\pm</math> 0.01</b>	<b>0.86 <math>\pm</math> 0.05</b>	<b>0.89 <math>\pm</math> 0.02</b>	<b>0.84 <math>\pm</math> 0.02</b>	<b>0.88 <math>\pm</math> 0.02</b>

# Appendix B

This appendix compliments the experiments conducted in Chapter 6 on TMA schemes. Figures 7.5-8 present the full results for the experiments of retraining the VGG19 model with each single cropping ratio but tested with RoIs cropped at all the proposed different ratios. Figure 7.9, below, displays the full results of TMA augmentation scenarios besides that of no augmentation.

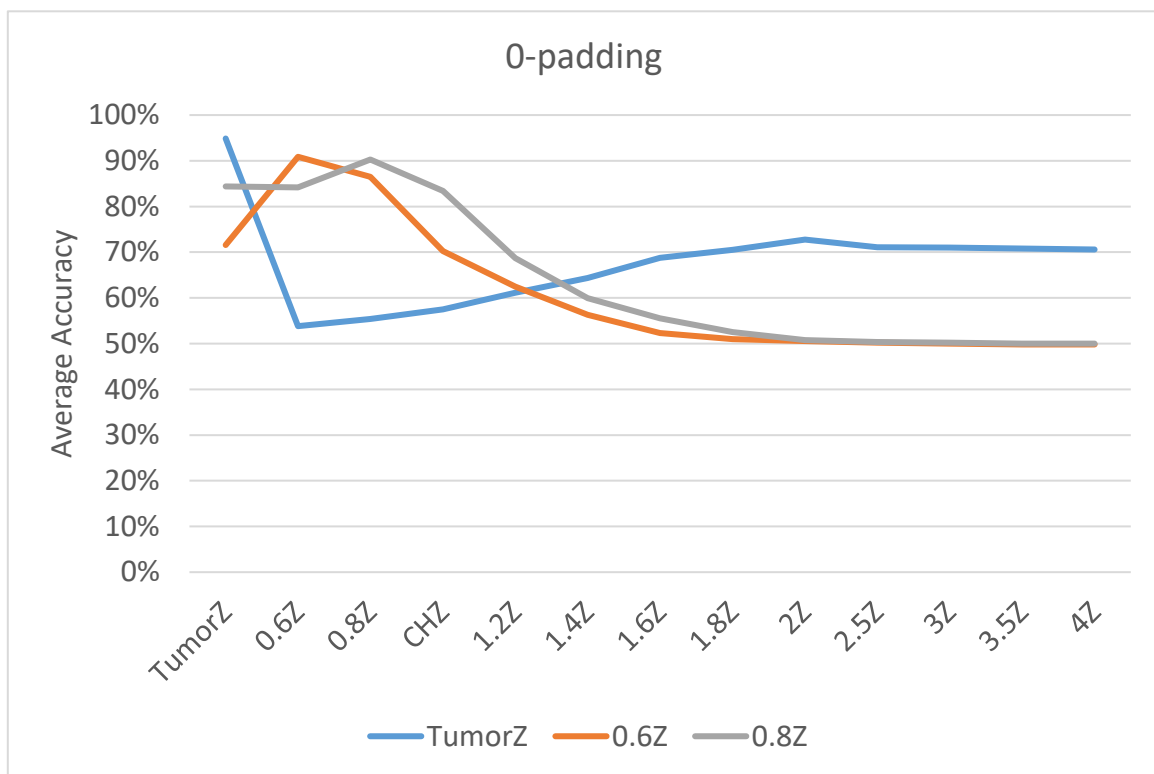


Figure 7.5 Performance of VGG19 trained with one cropping ratio and tested on all the ratios.

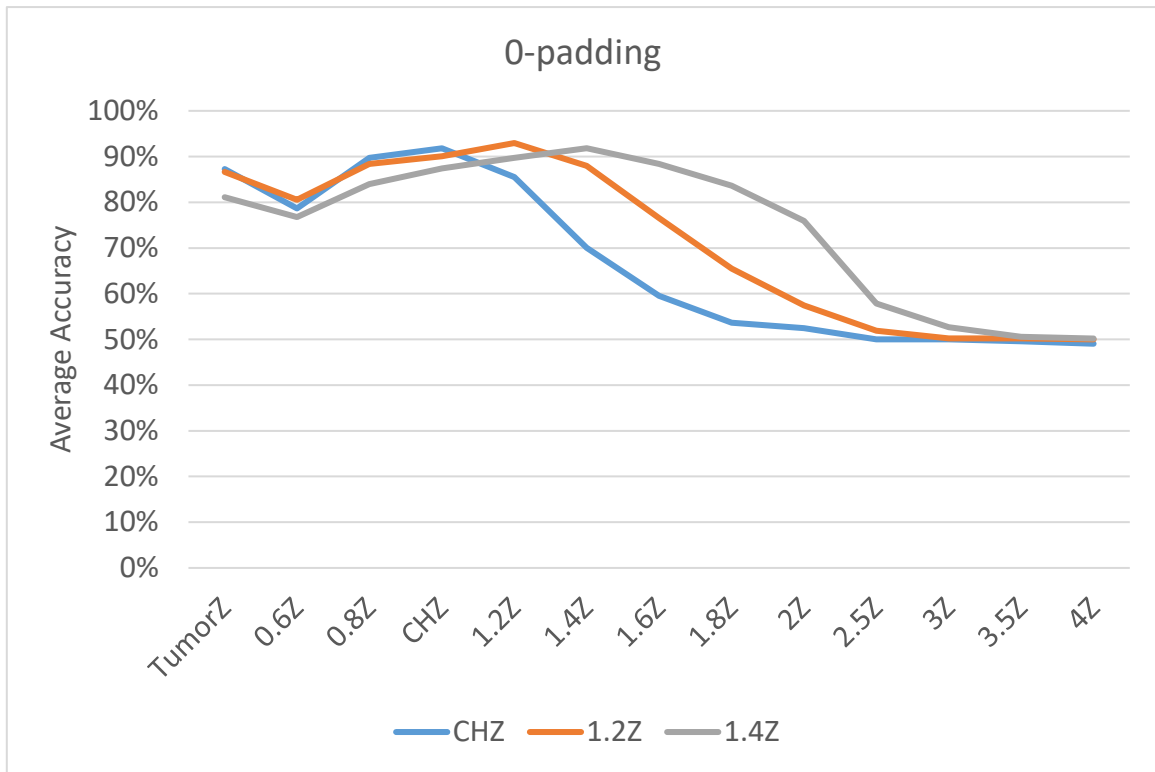


Figure 7.6 Performance of VGG19 trained with one cropping ratio and tested on all the ratios.

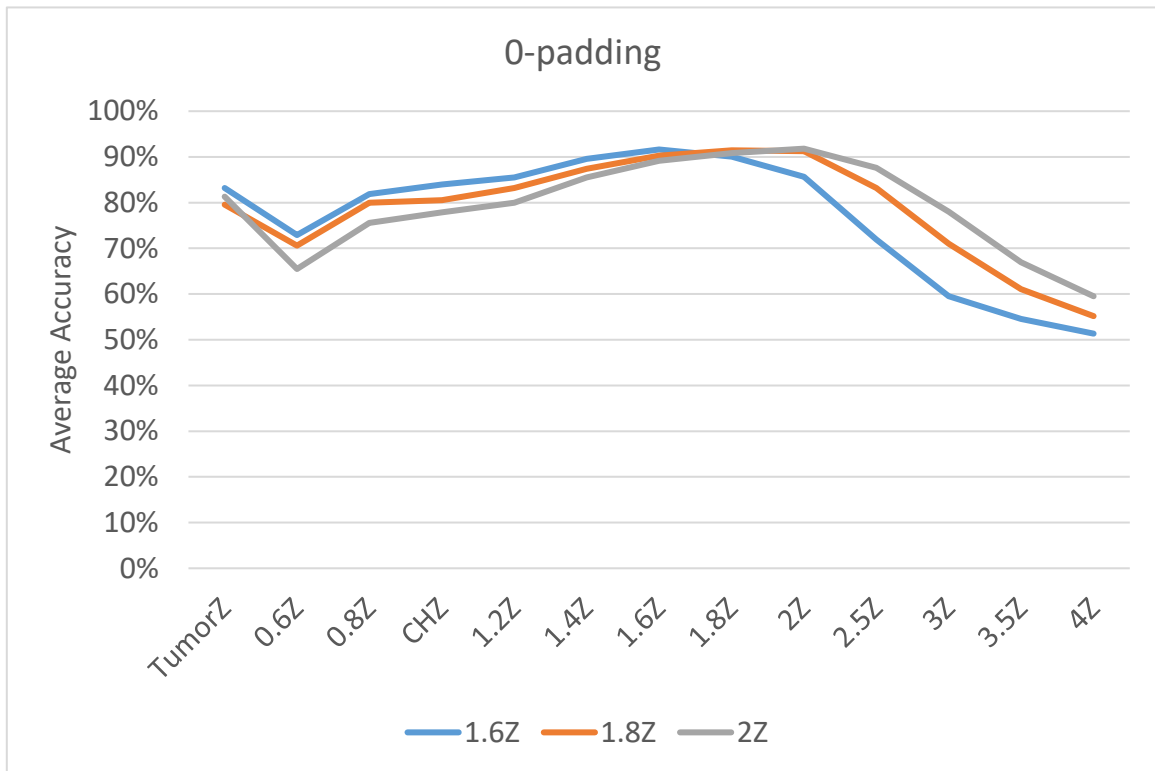


Figure 7.7 Performance of VGG19 trained with one cropping ratio and tested on all the ratios.

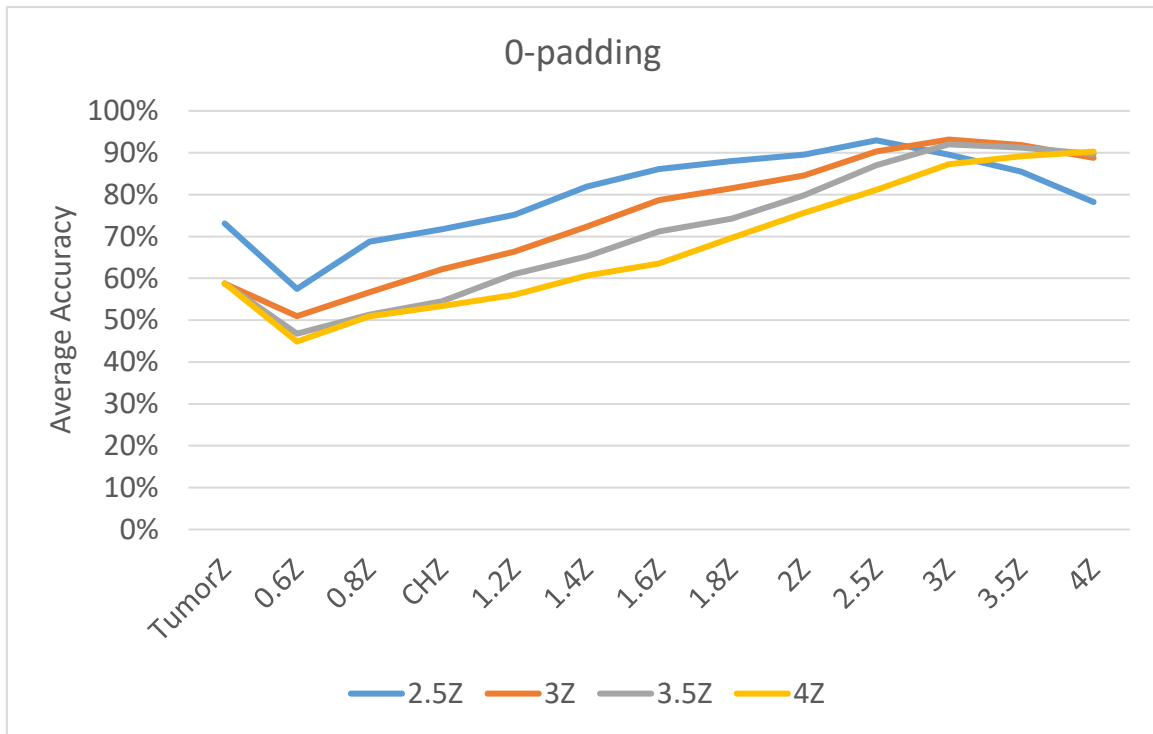


Figure 7.8 Performance of VGG19 trained with one cropping ratio and tested on all the ratios.

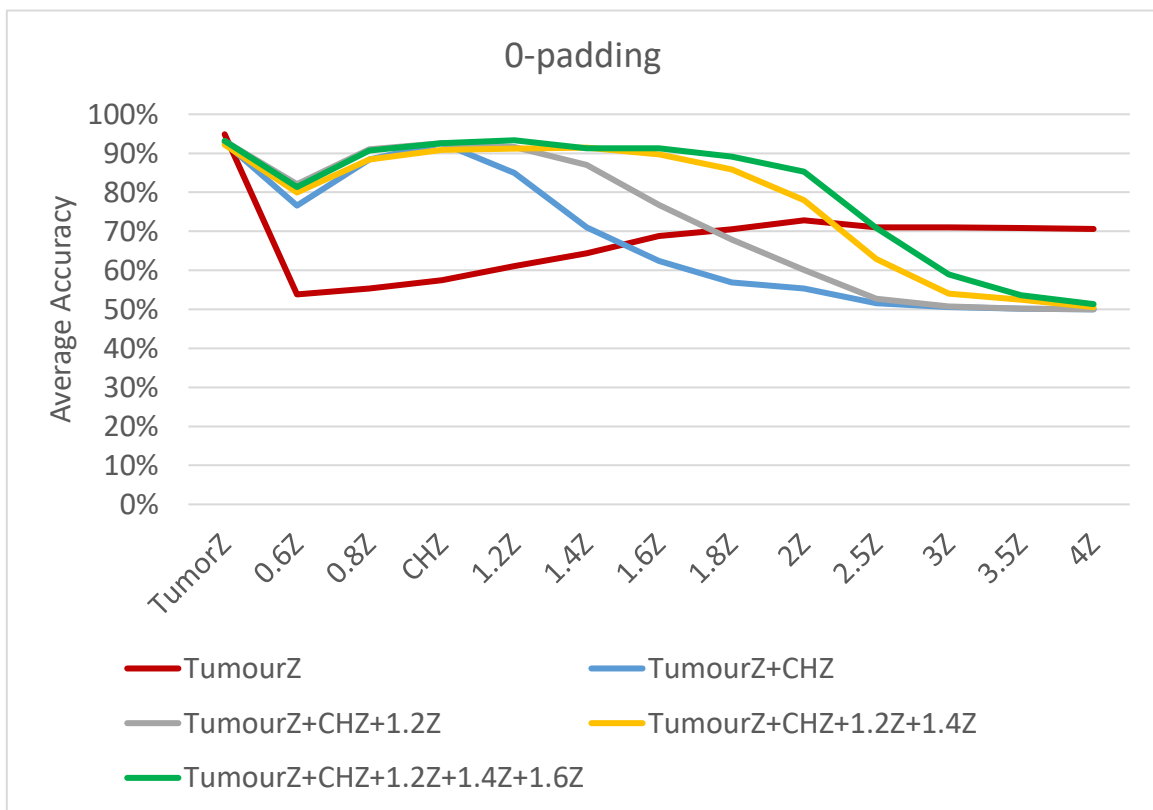


Figure 7.9 Performance of VGG19 trained with TMA augmented datasets.

# References

- [1] 'Breast Cancer Statistics | How Common Is Breast Cancer?' <https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html> (accessed May 05, 2023).
- [2] R. J. Hooley, L. M. Scoutt, and L. E. Philpotts, 'Breast Ultrasonography: State of the Art', *Radiology*, vol. 268, no. 3, pp. 642–659, Sep. 2013, doi: 10.1148/radiol.13121606.
- [3] E. J. Choi *et al.*, 'Interobserver agreement in breast ultrasound categorization in the Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness (MUST-BE) trial: results of a preliminary study', *Ultrasonography*, vol. 38, no. 2, pp. 172–180, Apr. 2019, doi: 10.14366/usg.18012.
- [4] O. Catalano *et al.*, 'Recent Advances in Ultrasound Breast Imaging: From Industry to Clinical Practice', *Diagnostics*, vol. 13, no. 5, p. 980, Mar. 2023, doi: 10.3390/diagnostics13050980.
- [5] Shan Khazendar, 'Computer-aided diagnosis of gynaecological abnormality using B-mode ultrasound images', Thesis (Doctoral), University of Buckingham, Buckingham, 2016. [Online]. Available: <http://bear.buckingham.ac.uk/351/#:~:text=http%3A//bear.buckingham.ac.uk/id/eprint/351>
- [6] Dhurgham Al-karawi, 'Texture Analysis based Machine Learning Algorithms For Ultrasound Ovarian Tumour Image Classification within Clinical Practices', PhD thesis, The University of Buckingham, Buckingham, 2019.
- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, 'ImageNet: A large-scale hierarchical image database', in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL: IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
- [8] T. Hassan, A. Alzoubi, H. Du, and S. Jassim, 'Towards optimal cropping: breast and liver tumor classification using ultrasound images', in *Multimodal Image Exploitation and Learning 2021*, S. S. Agaian, S. A. Jassim, S. P. DelMarco, and V. K. Asari, Eds., Online Only, United States: SPIE, Apr. 2021, p. 15. doi: 10.1117/12.2589038.
- [9] T. Hassan, A. AlZoubi, H. Du, and S. Jassim, 'Ultrasound image augmentation by tumor margin appending for robust deep learning based breast lesion classification', in *Multimodal Image Exploitation and Learning 2022*, S. S. Agaian, S. A. Jassim, S. P. DelMarco, and V. K. Asari, Eds., Orlando, United States: SPIE, May 2022, p. 8. doi: 10.1117/12.2618656.
- [10] 'Breast cancer'. <https://www.who.int/news-room/fact-sheets/detail/breast-cancer> (accessed Jun. 05, 2023).
- [11] A. Alexander *et al.*, 'The Impact of Breast Cancer on the Patient and the Family in Indian Perspective', *Indian J. Palliat. Care*, vol. 25, no. 1, pp. 66–72, 2019, doi: 10.4103/IJPC.IJPC\_158\_18.
- [12] F. İzci, A. S. İlgün, E. Findıklı, and V. Özmen, 'Psychiatric Symptoms and Psychosocial Problems in Patients with Breast Cancer', *J. Breast Health*, vol. 12, no. 3, pp. 94–101, Jul. 2016, doi: 10.5152/tjbh.2016.3041.
- [13] 'The impact of breast cancer awareness interventions on breast screening uptake among women in the United Kingdom: A systematic review - Natasha Anastasi, Joanne Lusher, 2019'. <https://journals.sagepub.com/doi/full/10.1177/1359105317697812> (accessed Jun. 05, 2023).
- [14] 'Breast Cancer Overview: Causes, Symptoms, Signs, Stages & Types', *Cleveland Clinic*. <https://my.clevelandclinic.org/health/diseases/3986-breast-cancer> (accessed Jun. 06, 2023).
- [15] Z. S. Lima, M. R. Ebadi, G. Amjad, and L. Younesi, 'Application of Imaging Technologies in Breast Cancer Detection: A Review Article', *Open Access Maced. J. Med. Sci.*, vol. 7, no. 5, pp. 838–848, Mar. 2019, doi: 10.3889/oamjms.2019.171.
- [16] W. A. Berg, A. I. Bandos, E. B. Mendelson, D. Lehrer, R. A. Jong, and E. D. Pisano, 'Ultrasound as the Primary Screening Test for Breast Cancer: Analysis From ACRIN 6666', *JNCI J. Natl. Cancer Inst.*, vol. 108, no. 4, p. djv367, Dec. 2015, doi: 10.1093/jnci/djv367.

- [17] 'Breast Ultrasound - an overview | ScienceDirect Topics'. <https://www.sciencedirect.com/topics/medicine-and-dentistry/breast-ultrasound> (accessed Jun. 05, 2023).
- [18] 'How do ultrasound examinations work?', *informedhealth.org*. <https://www.informedhealth.org/how-do-ultrasound-examinations-work.html> (accessed Jul. 14, 2023).
- [19] O. Catalano *et al.*, 'Recent Advances in Ultrasound Breast Imaging: From Industry to Clinical Practice', *Diagnostics*, vol. 13, no. 5, p. 980, Mar. 2023, doi: 10.3390/diagnostics13050980.
- [20] L. J. Graham *et al.*, 'Current Approaches and Challenges in Monitoring Treatment Responses in Breast Cancer', *J. Cancer*, vol. 5, no. 1, pp. 58–68, Jan. 2014, doi: 10.7150/jca.7047.
- [21] A. Y. Park and B. K. Seo, 'Up-to-date Doppler techniques for breast tumor vascularity: superb microvascular imaging and contrast-enhanced ultrasound', *Ultrasonography*, vol. 37, no. 2, pp. 98–106, Apr. 2018, doi: 10.14366/usg.17043.
- [22] T. Hassan, A. Alzoubi, H. Du, and S. Jassim, 'Towards optimal cropping: breast and liver tumor classification using ultrasound images', in *Multimodal Image Exploitation and Learning 2021*, S. S. Agaian, S. A. Jassim, S. P. DelMarco, and V. K. Asari, Eds., Online Only, United States: SPIE, Apr. 2021, p. 15. doi: 10.1117/12.2589038.
- [23] S. Gokhale, 'Ultrasound characterization of breast masses', *Indian J. Radiol. Imaging*, vol. 19, no. 3, pp. 242–247, Aug. 2009, doi: 10.4103/0971-3026.54878.
- [24] Y.-J. Mao, H.-J. Lim, M. Ni, W.-H. Yan, D. W.-C. Wong, and J. C.-W. Cheung, 'Breast Tumour Classification Using Ultrasound Elastography with Machine Learning: A Systematic Scoping Review', *Cancers*, vol. 14, no. 2, p. 367, Jan. 2022, doi: 10.3390/cancers14020367.
- [25] E. Sassaroli, C. Crake, A. Scorza, D. Kim, and M. Park, 'Image quality evaluation of ultrasound imaging systems: advanced B-modes', *J. Appl. Clin. Med. Phys.*, vol. 20, no. 3, pp. 115–124, Mar. 2019, doi: 10.1002/acm2.12544.
- [26] L. Wilkinson, V. Thomas, and N. Sharma, 'Microcalcification on mammography: approaches to interpretation and biopsy', *Br. J. Radiol.*, vol. 90, no. 1069, p. 20160594, doi: 10.1259/bjr.20160594.
- [27] 'Breast Imaging Reporting & Data System'. <https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads> (accessed Jun. 06, 2023).
- [28] Y.-W. Chen and L. C. Jain, Eds., *Deep Learning in Healthcare: Paradigms and Applications*, vol. 171. in Intelligent Systems Reference Library, vol. 171. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-32606-7.
- [29] W. Lin, K. Hasenstab, G. Moura Cunha, and A. Schwartzman, 'Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment', *Sci. Rep.*, vol. 10, no. 1, p. 20336, Nov. 2020, doi: 10.1038/s41598-020-77264-y.
- [30] H. Sellaheewa and S. A. Jassim, 'Image-Quality-Based Adaptive Face Recognition', *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 805–813, Apr. 2010, doi: 10.1109/TIM.2009.2037989.
- [31] L. Nanni, S. Ghidoni, and S. Brahmam, 'Handcrafted vs. non-handcrafted features for computer vision classification', *Pattern Recognit.*, vol. 71, pp. 158–172, Nov. 2017, doi: 10.1016/j.patcog.2017.05.025.
- [32] M. R. Zare, D. O. Alebiosu, and S. L. Lee, 'Comparison of Handcrafted Features and Deep Learning in Classification of Medical X-ray Images', in *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, Kota Kinabalu: IEEE, Mar. 2018, pp. 1–5. doi: 10.1109/INFRKM.2018.8464688.
- [33] F. Mohammad, A. Alzoubi, H. Du, and S. Jassim, 'Machine leaning assessment of border irregularity of thyroid nodules from ultrasound images', in *Multimodal Image Exploitation and Learning 2022*, S. S. Agaian, S. A. Jassim, S. P. DelMarco, and V. K. Asari, Eds., Orlando, United States: SPIE, May 2022, p. 6. doi: 10.1117/12.2618470.



- [34] Fakher Mohammad, 'Machine Learning Models for Recognizing Curve-shaped Abnormalities in Different Image Modalities', PhD thesis, The University of Buckingham, Buckingham, United Kingdom, 2022.
- [35] M. I. Daoud, S. Abdel-Rahman, T. M. Bdair, M. S. Al-Najar, F. H. Al-Hawari, and R. Alazrai, 'Breast Tumor Classification in Ultrasound Images Using Combined Deep and Handcrafted Features', *Sensors*, vol. 20, no. 23, p. 6838, Nov. 2020, doi: 10.3390/s20236838.
- [36] S. Liu *et al.*, 'Deep Learning in Medical Ultrasound Analysis: A Review', *Engineering*, vol. 5, no. 2, pp. 261–275, Apr. 2019, doi: 10.1016/j.eng.2018.11.020.
- [37] Mohammed Hussein Ahmed, 'Automatic Convolutional Neural Network Architecture Search for Breast Lesion Classification from Ultrasound Images: An ENAS Bayesian Optimization Approach', PhD thesis, The University of Buckingham, Buckingham, United Kingdom, 2022.
- [38] Ali Eskandari, 'Towards CNN Decision Understanding for Breast and Thyroid Lesions Classification from Ultrasound Images', Thesis (MSc), The University of Buckingham, Buckingham, 2022.
- [39] 'What is Deep Learning? | IBM'. <https://www.ibm.com/topics/deep-learning> (accessed Jun. 06, 2023).
- [40] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, 'Gradient-based learning applied to document recognition', *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.
- [41] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, 'A Guide to Convolutional Neural Networks for Computer Vision', *Synth. Lect. Comput. Vis.*, vol. 8, no. 1, pp. 1–207, Feb. 2018, doi: 10.2200/S00822ED1V01Y201712COV015.
- [42] Dheyaa Ahmed Ibrahim, 'Multi-level Segmentation of Gynaecological Ultrasound Images using Texture-based Trainable Models', PhD thesis, The University of Buckingham, Buckingham, United Kingdom, 2018.
- [43] Aras Asaad, 'Persistent Homology Tools for Image Analysis', Thesis (Doctoral), The University of Buckingham, Buckingham, 2020. [Online]. Available: <http://bear.buckingham.ac.uk/id/eprint/467>
- [44] A. Asaad, D. Ali, T. Majeed, and R. Rashid, 'Persistent Homology for Breast Tumor Classification Using Mammogram Scans', *Mathematics*, vol. 10, no. 21, p. 4039, Oct. 2022, doi: 10.3390/math10214039.
- [45] S. A. Jassim, 'Persistent homology features and multiple topologies for image analysis', in *Mobile Multimedia/Image Processing, Security, and Applications 2020*, S. S. Agaian, S. P. DelMarco, and V. K. Asari, Eds., Online Only, United States: SPIE, May 2020, p. 31. doi: 10.1117/12.2567052.
- [46] M. Wei *et al.*, 'Multi-feature Fusion for Ultrasound Breast Image Classification of Benign and Malignant', in *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, Xiamen, China: IEEE, Jul. 2019, pp. 474–478. doi: 10.1109/ICIVC47709.2019.8980898.
- [47] N. Dalal and B. Triggs, 'Histograms of Oriented Gradients for Human Detection', in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA: IEEE, 2005, pp. 886–893. doi: 10.1109/CVPR.2005.177.
- [48] T. Ojala, M. Pietikäinen, and D. Harwood, 'A comparative study of texture measures with classification based on featured distributions', *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996, doi: 10.1016/0031-3203(95)00067-4.
- [49] T. Ahonen, A. Hadid, and M. Pietikäinen, 'Face Recognition with Local Binary Patterns', in *Computer Vision - ECCV 2004*, T. Pajdla and J. Matas, Eds., in Lecture Notes in Computer Science, vol. 3021. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 469–481. doi: 10.1007/978-3-540-24670-1\_36.
- [50] A. Asaad and S. Jassim, 'Topological Data Analysis for Image Tampering Detection', in *Digital Forensics and Watermarking*, C. Kraetzer, Y.-Q. Shi, J. Dittmann, and H. J. Kim, Eds., in Lecture Notes in Computer Science, vol. 10431. Cham: Springer International Publishing, 2017, pp. 136–146. doi: 10.1007/978-3-319-64185-0\_11.

- [51] S. Theodoridis and K. Koutroumbas, *Pattern recognition*, 2. ed. Amsterdam: Academic Press, 2003.
- [52] 'Machine Learning - Convolution with color images', *DEV Community*, Mar. 21, 2020. <https://dev.to/sandeepbalachandran/machine-learning-convolution-with-color-images-2p41> (accessed May 05, 2023).
- [53] 'Fully Connected Deep Networks - TensorFlow for Deep Learning [Book]'. <https://www.oreilly.com/library/view/tensorflow-for-deep/9781491980446/ch04.html> (accessed Jul. 24, 2023).
- [54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet classification with deep convolutional neural networks', *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2012.
- [55] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *ArXiv14091556 Cs*, Apr. 2015, Accessed: Jun. 13, 2020. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [56] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- [57] F. Chollet, 'Xception: Deep Learning with Depthwise Separable Convolutions'. *arXiv*, Apr. 04, 2017. doi: 10.48550/arXiv.1610.02357.
- [58] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, 'Rethinking the Inception Architecture for Computer Vision', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.
- [59] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, 'Densely Connected Convolutional Networks', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.
- [60] A. Anwar, 'Difference between AlexNet, VGGNet, ResNet and Inception', *Medium*, Jan. 22, 2022. <https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaecccc96> (accessed Jul. 24, 2023).
- [61] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, 'Dataset of breast ultrasound images', *Data Brief*, vol. 28, p. 104863, Feb. 2020, doi: 10.1016/j.dib.2019.104863.
- [62] 'Receiver Operating Characteristic - an overview | ScienceDirect Topics'. <https://www.sciencedirect.com/topics/nursing-and-health-professions/receiver-operating-characteristic> (accessed Jun. 08, 2023).
- [63] M. Byra *et al.*, 'Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion', *Med. Phys.*, vol. 46, no. 2, pp. 746–755, Feb. 2019, doi: 10.1002/mp.13361.
- [64] H. Tanaka, S.-W. Chiu, T. Watanabe, S. Kaoku, and T. Yamaguchi, 'Computer-aided diagnosis system for breast ultrasound images using deep learning', *Phys. Med. Biol.*, vol. 64, no. 23, p. 235013, Dec. 2019, doi: 10.1088/1361-6560/ab5093.
- [65] Z. Cao, L. Duan, G. Yang, T. Yue, and Q. Chen, 'An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures', *BMC Med. Imaging*, vol. 19, no. 1, p. 51, Jul. 2019, doi: 10.1186/s12880-019-0349-x.
- [66] B. Zeimarani, M. G. F. Costa, N. Z. Nurani, S. R. Bianco, W. C. De Albuquerque Pereira, and C. F. F. C. Filho, 'Breast Lesion Classification in Ultrasound Images Using Deep Convolutional Neural Network', *IEEE Access*, vol. 8, pp. 133349–133359, 2020, doi: 10.1109/ACCESS.2020.3010863.
- [67] kaiwen wu, B. Xu, and Y. Wu, 'Recognition of Benign and Malignant Breast Ultrasound Images Based on Deep Transfer Learning', In Review, preprint, Jun. 2021. doi: 10.21203/rs.3.rs-249760/v1.
- [68] M. Ahmed, A. AlZoubi, and H. Du, 'Improving Generalization of ENAS-Based CNN Models for Breast Lesion Classification from Ultrasound Images', in *Medical Image Understanding and Analysis*, B. W. Papież, M. Yaqub, J. Jiao, A. I. L. Namburete, and J. A. Noble, Eds., in Lecture Notes

- in Computer Science, vol. 12722. Cham: Springer International Publishing, 2021, pp. 438–453. doi: 10.1007/978-3-030-80432-9\_33.
- [69] Mr. P. S. Parsania and Dr. P. V. Virparia, ‘A Comparative Analysis of Image Interpolation Algorithms’, *IJARCCCE*, vol. 5, no. 1, pp. 29–34, Jan. 2016, doi: 10.17148/IJARCCCE.2016.5107.
- [70] P. Miklós, ‘IMAGE INTERPOLATION TECHNIQUES’, presented at the In 2nd Siberian-Hungarian Joint Symposium on Intelligent Systems, Oct. 2004.
- [71] V. Patel and K. Mistree, ‘A Review on Different Image Interpolation Techniques for Image Enhancement’, *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, pp. 129–133, Dec. 2013.
- [72] A. Prajapati, S. Naik, and S. Mehta, ‘Evaluation of Different Image Interpolation Algorithms’, *Int. J. Comput. Appl.*, vol. 58, no. 12, pp. 6–12, Nov. 2012, doi: 10.5120/9332-3638.
- [73] W. Burger and M. J. Burge, *Digital Image Processing: An Algorithmic Introduction Using Java*. in Texts in Computer Science. London: Springer London, 2016. doi: 10.1007/978-1-4471-6684-9.
- [74] C. Suresh, S. Singh, R. Saini, and A. K. Saini, ‘A Comparative Analysis of Image Scaling Algorithms’, *Int. J. Image Graph. Signal Process.*, vol. 5, no. 5, pp. 55–62, Apr. 2013, doi: 10.5815/ijigsp.2013.05.07.
- [75] N. Al-Hassan, ‘Mathematically inspired approaches to face recognition in uncontrolled conditions: super resolution and compressive sensing’, doctoral, University of Buckingham, 2014. Accessed: Jun. 30, 2020. [Online]. Available: <http://bear.buckingham.ac.uk/6/>
- [76] T. M. Hassan, ‘Data-independent vs. data-dependent dimension reduction for pattern recognition in high dimensional spaces’, masters, University of Buckingham, 2017. Accessed: Feb. 25, 2018. [Online]. Available: <http://bear.buckingham.ac.uk/199/>
- [77] A. Bovik, Ed., *The essential guide to image processing*, 2. ed. Amsterdam: Academic Pr, 2009.
- [78] M. Elad and A. Feuer, ‘Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images’, *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1646–1658, Dec. 1997, doi: 10.1109/83.650118.
- [79] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang, ‘Super-resolution image reconstruction: a technical overview’, *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, May 2003, doi: 10.1109/MSP.2003.1203207.
- [80] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, ‘Image Super-Resolution by Neural Texture Transfer’, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, Jun. 2019, pp. 7974–7983. doi: 10.1109/CVPR.2019.00817.
- [81] C. Dong, C. C. Loy, K. He, and X. Tang, ‘Image Super-Resolution Using Deep Convolutional Networks’, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016, doi: 10.1109/TPAMI.2015.2439281.
- [82] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, ‘Loss Functions for Image Restoration With Neural Networks’, *IEEE Trans. Comput. Imaging*, vol. 3, no. 1, pp. 47–57, Mar. 2017, doi: 10.1109/TCI.2016.2644865.
- [83] J. Johnson, A. Alahi, and L. Fei-Fei, ‘Perceptual Losses for Real-Time Style Transfer and Super-Resolution’, *ArXiv160308155 Cs*, Mar. 2016, Accessed: Jun. 30, 2020. [Online]. Available: <http://arxiv.org/abs/1603.08155>
- [84] C. Ledig *et al.*, ‘Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network’, *ArXiv160904802 Cs Stat*, May 2017, Accessed: Jun. 30, 2020. [Online]. Available: <http://arxiv.org/abs/1609.04802>
- [85] Jianchao Yang, J. Wright, T. S. Huang, and Yi Ma, ‘Image Super-Resolution Via Sparse Representation’, *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010, doi: 10.1109/TIP.2010.2050625.
- [86] Jianchao Yang, J. Wright, T. Huang, and Yi Ma, ‘Image super-resolution as sparse representation of raw image patches’, in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA: IEEE, Jun. 2008, pp. 1–8. doi: 10.1109/CVPR.2008.4587647.
- [87] S. S. Agaian, Ed., *Hadamard transforms*. Bellingham, Wash: SPIE, 2011.

- [88] M. H. Yap *et al.*, 'Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks', *IEEE J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1218–1226, Jul. 2018, doi: 10.1109/JBHI.2017.2731873.
- [89] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, 'Deep Learning Approaches for Data Augmentation and Classification of Breast Masses using Ultrasound Images', *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, 2019, doi: 10.14569/IJACSA.2019.0100579.
- [90] S. Misra *et al.*, 'Bi-Modal Transfer Learning for Classifying Breast Cancers via Combined B-Mode and Ultrasound Strain Imaging', *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 69, no. 1, pp. 222–232, Jan. 2022, doi: 10.1109/TUFFC.2021.3119251.
- [91] A. Horé and D. Ziou, 'Image Quality Metrics: PSNR vs. SSIM', in *2010 20th International Conference on Pattern Recognition*, Aug. 2010, pp. 2366–2369. doi: 10.1109/ICPR.2010.579.
- [92] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, 'Image Quality Assessment: From Error Visibility to Structural Similarity', *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.
- [93] Zhou Wang and A. C. Bovik, 'A universal image quality index', *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002, doi: 10.1109/97.995823.
- [94] S. Md. R. Islam, X. Huang, and K. Le, 'A Novel Image Quality Index for Image Quality Assessment', in *Neural Information Processing*, M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil, Eds., in *Lecture Notes in Computer Science*, vol. 8228. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 549–556. doi: 10.1007/978-3-642-42051-1\_68.
- [95] G. Litjens *et al.*, 'A survey on deep learning in medical image analysis', *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.
- [96] A. Hijab, M. A. Rushdi, M. M. Gomaa, and A. Eldeib, 'Breast Cancer Classification in Ultrasound Images using Transfer Learning', in *2019 Fifth International Conference on Advances in Biomedical Engineering (ICABME)*, Tripoli, Lebanon: IEEE, Oct. 2019, pp. 1–4. doi: 10.1109/ICABME47164.2019.8940291.
- [97] N. Tajbakhsh *et al.*, 'Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?', *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016, doi: 10.1109/TMI.2016.2535302.
- [98] G. Wang *et al.*, 'Interactive Medical Image Segmentation Using Deep Learning With Image-Specific Fine Tuning', *IEEE Trans. Med. Imaging*, vol. 37, no. 7, pp. 1562–1573, Jul. 2018, doi: 10.1109/TMI.2018.2791721.
- [99] C. Shorten and T. M. Khoshgoftaar, 'A survey on Image Data Augmentation for Deep Learning', *J. Big Data*, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.
- [100] E. Goceri, 'Medical image data augmentation: techniques, comparisons and interpretations', *Artif. Intell. Rev.*, Mar. 2023, doi: 10.1007/s10462-023-10453-z.
- [101] G. Ayana, K. Dese, and S. Choe, 'Transfer Learning in Breast Cancer Diagnoses via Ultrasound Imaging', *Cancers*, vol. 13, no. 4, p. 738, Feb. 2021, doi: 10.3390/cancers13040738.
- [102] L. Perez and J. Wang, 'The Effectiveness of Data Augmentation in Image Classification using Deep Learning', 2017, doi: 10.48550/ARXIV.1712.04621.
- [103] T. Pang, J. H. D. Wong, W. L. Ng, and C. S. Chan, 'Semi-supervised GAN-based Radiomics Model for Data Augmentation in Breast Ultrasound Mass Classification', *Comput. Methods Programs Biomed.*, vol. 203, p. 106018, May 2021, doi: 10.1016/j.cmpb.2021.106018.
- [104] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, 'Synthetic data in machine learning for medicine and healthcare', *Nat. Biomed. Eng.*, vol. 5, no. 6, pp. 493–497, Jun. 2021, doi: 10.1038/s41551-021-00751-8.
- [105] Y.-C. Zhu *et al.*, 'A generic deep learning framework to classify thyroid and breast lesions in ultrasound images', *Ultrasonics*, vol. 110, p. 106300, Feb. 2021, doi: 10.1016/j.ultras.2020.106300.

- [106] M. Yamakawa, T. Shiina, N. Nishida, and M. Kudo, 'Optimal cropping for input images used in a convolutional neural network for ultrasonic diagnosis of liver tumors', *Jpn. J. Appl. Phys.*, vol. Volume 59, Apr. 2020, doi: 10.35848/1347-4065/ab80dd.
- [107] T. Majeed, R. Rashid, D. Ali, and A. Asaad, 'Issues associated with deploying CNN transfer learning to detect COVID-19 from chest X-rays', *Phys. Eng. Sci. Med.*, vol. 43, no. 4, pp. 1289–1303, Dec. 2020, doi: 10.1007/s13246-020-00934-8.
- [108] S. Han *et al.*, 'A deep learning framework for supporting the classification of breast lesions in ultrasound images', *Phys. Med. Biol.*, vol. 62, no. 19, pp. 7714–7728, Sep. 2017, doi: 10.1088/1361-6560/aa82ec.
- [109] S. Liu-Yu and M. Thonnat, 'Using apparent boundary and convex hull for the shape characterization of foraminifera images', in *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol. IV. Conference D: Architectures for Vision and Pattern Recognition*, The Hague, Netherlands: IEEE Comput. Soc. Press, 1992, pp. 569–572. doi: 10.1109/ICPR.1992.202051.
- [110] I. Pitas, *Digital image processing algorithms and applications*. New York: Wiley, 2000.
- [111] G. T. Toussaint, 'Computational Geometric Problems in Pattern Recognition', in *Pattern Recognition Theory and Applications*, J. Kittler, K. S. Fu, and L.-F. Pau, Eds., Dordrecht: Springer Netherlands, 1982, pp. 73–91. doi: 10.1007/978-94-009-7772-3\_7.
- [112] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization', in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.
- [113] H. Panwar, P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, 'A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images', *Chaos Solitons Fractals*, vol. 140, p. 110190, Nov. 2020, doi: 10.1016/j.chaos.2020.110190.
- [114] M. C. Hemmsen, M. M. Petersen, S. I. Nikolov, M. B. Nielsen, and J. A. Jensen, 'Ultrasound image quality assessment: a framework for evaluation of clinical image quality', presented at the SPIE Medical Imaging, J. D'hooge and S. A. McAleavey, Eds., San Diego, California, USA, Mar. 2010, p. 76290C. doi: 10.1117/12.840664.
- [115] S. Zhang, Y. Wang, J. Jiang, J. Dong, W. Yi, and W. Hou, 'CNN-Based Medical Ultrasound Image Quality Assessment', *Complexity*, vol. 2021, pp. 1–9, Jul. 2021, doi: 10.1155/2021/9938367.
- [116] A. Mittal, G. S. Muralidhar, J. Ghosh, and A. C. Bovik, 'Blind Image Quality Assessment Without Human Training Using Latent Quality Factors', *IEEE Signal Process. Lett.*, vol. 19, no. 2, pp. 75–78, Feb. 2012, doi: 10.1109/LSP.2011.2179293.
- [117] A. Mittal, A. K. Moorthy, and A. C. Bovik, 'No-Reference Image Quality Assessment in the Spatial Domain', *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012, doi: 10.1109/TIP.2012.2214050.
- [118] A. Mittal, R. Soundararajan, and A. C. Bovik, 'Making a "Completely Blind" Image Quality Analyzer', *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013, doi: 10.1109/LSP.2012.2227726.
- [119] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, S. S. Channappayya, and S. S. Medasani, 'Blind image quality evaluation using perception based features', in *2015 Twenty First National Conference on Communications (NCC)*, Mumbai, India: IEEE, Feb. 2015, pp. 1–6. doi: 10.1109/NCC.2015.7084843.
- [120] R. Dey, D. Bhattacharjee, C. Kollmann, and O. Krejcar, 'Classification of Breast Tumor from Ultrasound Images Using No-Reference Image Quality Assessment', in *Proceedings of International Conference on Frontiers in Computing and Systems*, S. Basu, D. K. Kole, A. K. Maji, D. Plewczynski, and D. Bhattacharjee, Eds., in Lecture Notes in Networks and Systems, vol. 404. Singapore: Springer Nature Singapore, 2023, pp. 341–349. doi: 10.1007/978-981-19-0105-8\_33.

- [121] B. Fu, N. Spiller, C. Chen, and N. Damer, 'The effect of face morphing on face image quality', in *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany: IEEE, Sep. 2021, pp. 1–5. doi: 10.1109/BIOSIG52210.2021.9548302.
- [122] O. V. Michailovich and A. Tannenbaum, 'Despeckling of medical ultrasound images', *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 53, no. 1, pp. 64–78, Jan. 2006, doi: 10.1109/TUFFC.2006.1588392.
- [123] M. D. Heath, K. Bowyer, D. Kopans, and R. H. Moore, 'THE DIGITAL DATABASE FOR SCREENING MAMMOGRAPHY', 2007. Accessed: Apr. 17, 2023. [Online]. Available: <https://www.semanticscholar.org/paper/THE-DIGITAL-DATABASE-FOR-SCREENING-MAMMOGRAPHY-Heath-Bowyer/ff2218b349f89026ffaaccdf807228fa497c04bd>
- [124] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, and J. Dittmann, 'Extended *StirTrace* benchmarking of biometric and forensic qualities of morphed face images', *IET Biom.*, vol. 7, no. 4, pp. 325–332, Jul. 2018, doi: 10.1049/iet-bmt.2017.0147.
- [125] A. Makrushin, T. Neubert, and J. Dittmann, 'Humans Vs. Algorithms: Assessment of Security Risks Posed by Facial Morphing to Identity Verification at Border Control', in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Prague, Czech Republic: SCITEPRESS - Science and Technology Publications, 2019, pp. 513–520. doi: 10.5220/0007378905130520.
- [126] J. Ghafari, H. Du, and S. A. Jassim, 'Sensitivity and stability of pretrained CNN filters', in *Multimodal Image Exploitation and Learning 2021*, S. S. Agaian, S. A. Jassim, S. P. DelMarco, and V. K. Asari, Eds., Online Only, United States: SPIE, Apr. 2021, p. 10. doi: 10.1117/12.2589521.
- [127] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 'Dropout: a simple way to prevent neural networks from overfitting', *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
- [128] S. Ioffe and C. Szegedy, 'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift', in *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, Jun. 2015, pp. 448–456. Accessed: Apr. 10, 2023. [Online]. Available: <https://proceedings.mlr.press/v37/ioffe15.html>
- [129] K. Weiss, T. M. Khoshgoftaar, and D. Wang, 'A survey of transfer learning', *J. Big Data*, vol. 3, no. 1, p. 9, Dec. 2016, doi: 10.1186/s40537-016-0043-6.
- [130] Ling Shao, Fan Zhu, and Xuelong Li, 'Transfer Learning for Visual Categorization: A Survey', *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015, doi: 10.1109/TNNLS.2014.2330900.
- [131] M. Palatucci, D. Pomerleau, G. Hinton, and T. M. Mitchell, 'Zero-shot learning with semantic output codes', in *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, in NIPS'09. Red Hook, NY, USA: Curran Associates Inc., Dec. 2009, pp. 1410–1418.
- [132] G. R. Koch, 'Siamese Neural Networks for One-Shot Image Recognition', 2015. Accessed: Apr. 10, 2023. [Online]. Available: <https://www.semanticscholar.org/paper/Siamese-Neural-Networks-for-One-Shot-Image-Koch/f216444d4f2959b4520c61d20003fa30a199670a>
- [133] J. Kukačka, V. Golkov, and D. Cremers, 'Regularization for Deep Learning: A Taxonomy', 2017, doi: 10.48550/ARXIV.1710.10686.
- [134] B. A. Jonsson *et al.*, 'Brain age prediction using deep learning uncovers associated sequence variants', *Nat. Commun.*, vol. 10, no. 1, p. 5409, Nov. 2019, doi: 10.1038/s41467-019-13163-9.
- [135] M. Ueda *et al.*, 'An Age Estimation Method Using 3D-CNN From Brain MRI Images', in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Venice, Italy: IEEE, Apr. 2019, pp. 380–383. doi: 10.1109/ISBI.2019.8759392.
- [136] R. Hu, G. Ruan, S. Xiang, M. Huang, Q. Liang, and J. Li, 'Automated Diagnosis of COVID-19 Using Deep Learning and Data Augmentation on Chest CT', *Health Informatics*, preprint, Apr. 2020. doi: 10.1101/2020.04.24.20078998.

- [137] A. R. Khan, S. Khan, M. Harouni, R. Abbasi, S. Iqbal, and Z. Mehmood, 'Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification', *Microsc. Res. Tech.*, vol. 84, no. 7, pp. 1389–1399, Jul. 2021, doi: 10.1002/jemt.23694.
- [138] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, 'Explainable COVID-19 Detection Using Chest CT Scans and Deep Learning', *Sensors*, vol. 21, no. 2, Art. no. 2, Jan. 2021, doi: 10.3390/s21020455.
- [139] M. Sajjad, S. Khan, K. Muhammad, W. Wu, A. Ullah, and S. W. Baik, 'Multi-grade brain tumor classification using deep CNN with extensive data augmentation', *J. Comput. Sci.*, vol. 30, pp. 174–182, Jan. 2019, doi: 10.1016/j.jocs.2018.12.003.
- [140] G. Kang, X. Dong, L. Zheng, and Y. Yang, 'PatchShuffle Regularization'. arXiv, Jul. 22, 2017. doi: 10.48550/arXiv.1707.07103.
- [141] A. H. Ornek and M. Ceylan, 'Comparison of Traditional Transformations for Data Augmentation in Deep Learning of Medical Thermography', in *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, Budapest, Hungary: IEEE, Jul. 2019, pp. 191–194. doi: 10.1109/TSP.2019.8769068.
- [142] E. K. Kim, H. Lee, J. Y. Kim, and S. Kim, 'Data Augmentation Method by Applying Color Perturbation of Inverse PSNR and Geometric Transformations for Object Recognition Based on Deep Learning', *Appl. Sci.*, vol. 10, no. 11, Art. no. 11, Jan. 2020, doi: 10.3390/app10113755.
- [143] M. Tirindelli, C. Eilers, W. Simson, M. Paschali, M. F. Azampour, and N. Navab, 'Rethinking Ultrasound Augmentation: A Physics-Inspired Approach', in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 690–700. doi: 10.1007/978-3-030-87237-3\_66.
- [144] D. Ma, D. Lu, K. Popuri, L. Wang, M. F. Beg, and Alzheimer's Disease Neuroimaging Initiative, 'Differential Diagnosis of Frontotemporal Dementia, Alzheimer's Disease, and Normal Aging Using a Multi-Scale Multi-Type Feature Generative Adversarial Deep Neural Network on Structural Magnetic Resonance Images', *Front. Neurosci.*, vol. 14, p. 853, 2020, doi: 10.3389/fnins.2020.00853.
- [145] C. Muramatsu *et al.*, 'Improving breast mass classification by shared data with domain transformation using a generative adversarial network', *Comput. Biol. Med.*, vol. 119, p. 103698, Apr. 2020, doi: 10.1016/j.combiomed.2020.103698.
- [146] Q. Wang, X. Zhang, W. Chen, K. Wang, and X. Zhang, 'Class-Aware Multi-window Adversarial Lung Nodule Synthesis Conditioned on Semantic Features', in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 589–598. doi: 10.1007/978-3-030-59725-2\_57.
- [147] M. Pesteie, P. Abolmaesumi, and R. N. Rohling, 'Adaptive Augmentation of Medical Data Using Independently Conditional Variational Auto-Encoders', *IEEE Trans. Med. Imaging*, vol. 38, no. 12, pp. 2807–2820, Dec. 2019, doi: 10.1109/TMI.2019.2914656.
- [148] X. Yi, E. Walia, and P. Babyn, 'Generative adversarial network in medical imaging: A review', *Med. Image Anal.*, vol. 58, p. 101552, Dec. 2019, doi: 10.1016/j.media.2019.101552.
- [149] K. Zhang, 'On Mode Collapse in Generative Adversarial Networks', in *Artificial Neural Networks and Machine Learning – ICANN 2021*, I. Farkaš, P. Masulli, S. Otte, and S. Wermter, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 563–574. doi: 10.1007/978-3-030-86340-1\_45.
- [150] HARIN SELLAHEWA, 'WAVELET-BASED AUTOMATIC FACE RECOGNITION FOR CONSTRAINED DEVICES', PhD thesis, The University of Buckingham, Buckingham, United Kingdom, 2006.
- [151] HISHAM AL-ASSAM, 'ENTROPY EVALUATION AND SECURITY MEASURES FOR RELIABLE SINGLE/MULTI-FACTOR BIOMETRIC AUTHENTICATION AND BIOMETRIC KEYS', PhD thesis, The University of Buckingham, Buckingham, United Kingdom, 2013.

- [152] O. O. Abayomi-Alli, R. Damaševičius, M. Wieczorek, and M. Woźniak, 'Data Augmentation Using Principal Component Resampling for Image Recognition by Deep Learning', in *Artificial Intelligence and Soft Computing*, L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, and J. M. Zurada, Eds., in Lecture Notes in Computer Science, vol. 12416. Cham: Springer International Publishing, 2020, pp. 39–48. doi: 10.1007/978-3-030-61534-5\_4.
- [153] V. Klema and A. Laub, 'The singular value decomposition: Its computation and some applications', *IEEE Trans. Autom. Control*, vol. 25, no. 2, pp. 164–176, Apr. 1980, doi: 10.1109/TAC.1980.1102314.
- [154] M. Qiao, Y. Hu, Y. Guo, Y. Wang, and J. Yu, 'Breast Tumor Classification Based on a Computerized Breast Imaging Reporting and Data System Feature System: Breast Tumor Classification by a Digital BI-RADS System', *J. Ultrasound Med.*, vol. 37, no. 2, pp. 403–415, Feb. 2018, doi: 10.1002/jum.14350.
- [155] J. Shan, S. K. Alam, B. Garra, Y. Zhang, and T. Ahmed, 'Computer-Aided Diagnosis for Breast Ultrasound Using Computerized BI-RADS Features and Machine Learning Methods', *Ultrasound Med. Biol.*, vol. 42, no. 4, pp. 980–988, Apr. 2016, doi: 10.1016/j.ultrasmedbio.2015.11.016.