Breast Lesion Detection in Ultrasound Images Using Deep Neural Networks:

Clustering Based Approach for False Positive Reduction

Thesis submitted for the degree of

Doctor of Philosophy in Computing to the School of Computing, University of Buckingham

Anu Bose

1903165
School of Computing
University of Buckingham

Submission Date: June 2023

Abstract

Breast cancer is one of the most common forms of cancer. Popular imaging modalities used for breast cancer screening include Mammograms, Ultrasound (US), and Magnetic Resonance Imaging (MRI). US is a widely adopted modality due to its relative affordability, portability and higher patient safety. Early detection of lesion(s) is crucial to ensure a high survival rate and minimise adverse effects on the body. Currently, we face a global crisis in the number of experienced radiologists available per patient. Therefore, automating lesion detection, with Artificial intelligence (AI) acting as a secondary opinion, can assist radiologists in faster diagnosis. In recent years, deep-learning (DL) based object detection methods have become popular in Computer-Aided Diagnosis (CAD) systems due to their ability in extracting high level, abstract features resulting in their higher generalisation capability and applicability in real-life operations.

Compared to object detection in natural images, lesion detection in US images is a challenging task due to the inherent characteristics of these images. Due to these challenges and a lack of large-scale US datasets, the number of DL-based lesion detection methods developed for US images is relatively lower compared to object detection methods developed for natural images. Thus, it is common practice to modify an existing object detector originally designed for natural images for lesion detection in US images. One such popularly adapted detector is Faster R-CNN (FRCNN). Limited attention has been given to adapted FRCNN for breast lesion detection in US images. The adaptation results in a relatively high detection rate along with a high number of false positive (FP) detections that degrade the overall performance. Such high FPs may mystify radiologists in reading and interpreting the US images and lead to unnecessary additional checks and biopsies. Reducing FPs in breast US images still remains an open investigation area which provides us the motivation for this study. Up to the point reported in this thesis, no work has been specifically developed to adequately address the issue of FPs in DL-based detection methods for breast lesion detection in US images.

The aim of this research is to create a novel and effective DL-based method for detecting breast lesions from 2D US images. The research starts by investigating the effectiveness of FRCNN for breast lesion detection using large datasets of US images collected from different medical centres and machine makers. The research then provides the first solution to address the issue of FP detections by searching

and identifying the optimal training and architectural hyperparameters of this powerful network. The adapted FRCNN model outperformed the original FRCNN through a significant reduction in FPs and small negative impact on the number of correct detections. Additionally, the adapted model also surpassed several existing detectors developed for natural images as well as those adapted for breast lesion detection in US images. Furthermore, this research develops a new U-Detect method. U-Detect is a clustering-based approach that combines unsupervised learning technique and the adapted FRCNN to reduce the FP detections. Two variants of the U-Detect method are developed: U-Detect-Base and U-Detect-RPN models. Both U-Detect models outperform original and adapted FRCNN models through considerable reduction in FPs resulting in its higher precision. Additionally, U-Detect-RPN detected higher number of lesions than the adapted FRCNN model.

Inspired by the domain knowledge of breast lesion characteristics, we further enhanced the architecture of U-Detect by developing a new classification-based approach (U-DetectH) that uses a fusion of textural and morphological handcrafted features to improve the classification scores in U-Detect and ultimately reduce the FP detections. Two variants of U-Detect-H are developed: U-DetectH-Base and U-DetectH-RPN models. The research concludes that on multiple datasets comprising a combined total of 3119 US images, U-DetectH-Base outperforms original FRCNN with 5.49% to 32.83% higher precision and a small drop of 0.27% to 10.02% in recall. This significantly higher precision is due to a 31.86% to 77.07% reduction in FPs. The work presented in this thesis provides an approach for scientists to design a robust object detection model for other cancer types as well as other medical modalities.

Acknowledgements

Firstly, I would like to thank my supervisors, Dr. Alaa AlZoubi, Dr. Tuan Nguyen and Prof. Hongbo Du. I am eternally grateful for your guidance, expertise, and mentorship. I would like to express my gratitude to Dr. Alaa for always taking the time to help me grow professionally and personally. His wisdom, constant enthusiasm and words of encouragement have truly inspired me throughout this journey. I would also like to express my gratitude to TenD AI Medical Technologies Ltd. for their funding this research and providing the resources for this work. I am thankful for my labmates for their continued support and friendship. I have thoroughly enjoyed our lively discussions and the time we spent together. I would also like to express my gratitude towards the University of Buckingham for the resources, opportunities and support provided to me throughout my time here. The administrative and support staff have been absolutely wonderful.

I would like to express my sincere thanks to Prasad sir and his family for being by my side ever since I first arrived in England. I am truly grateful to have him as a personal mentor, guiding me right from my initial days. His unwavering support not only helped me settle in but also continued to be a source of strength as I embarked on my PhD journey. I will always cherish the kindness and guidance he provided and continues to provide. To my friends in Buckingham, Manchester, and Mumbai, your love and encouragement have been a constant source of strength, no matter the time or place. Rishi, you have been a delightful friend, making this PhD experience truly memorable. Our time together will forever hold a special place in my heart. I want to extend my heartfelt appreciation to Michael for his incredible support throughout my journey. His patience, understanding, and willingness to listen have been a tremendous source of comfort. Michael, thank you for welcoming me into your family and making England feel like a second home. To my lovely sister, Arti, your remarkable ability to find humor in challenging moments has been a blessing. Your laughter and lightheartedness have brought joy during difficult times, reminding me to embrace life's lighter side. Your unwavering support and love have been my constant strength. I am truly blessed to have you as my sister. Finally, my dear mother, Salina, and father, Bose, to whom I owe everything. I wouldn't be here without your unconditional love, support, and blessings. I dedicate this thesis to you both, with a heart full of gratitude and love.

List of Publications

[1] Bose A., Nguyen T., Du H., AlZoubi A. (2022) Faster R-CNN Hyperparameter Selection for Breast Lesion Detection in 2D Ultrasound Images. In: Jansen T., Jensen R., Mac Parthaláin N., Lin CM. (eds) Advances in Computational Intelligence Systems. UKCI 2021. Advances in Intelligent Systems and Computing, vol 1409. Springer, Cham.

[2] Bose A., Nguyen T., Du H., AlZoubi A. U-Detect: Lesion Detection in Ultrasound Breast Images. In: IEEE Transactions on Medical Imaging. Under preparation for submission.

Abbreviations

AP	Average Precision
AI	Artifical Intelligence
AUC	Area Under Curve
BIC	Bayesian Information Criteria
BI-RADS	Breast Imaging Reporting and Data System
CAD	Computer Aided Diagnosis
CARL	Classification Aware Regression Loss
CE	Cross Entropy
CIOU	Complete Intersection-over-Union
CNN	Convolutional Neural Network
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCE-MRI	Dynamic Contrast-Enhanced Magnetic Resonance Imaging
DIOU	Distance Intersection-over-Union
DL	Deep Learning
DWR	Depth-to-Width ratio
ENAS	Efficient Network Architecture Search
FC	Fully Connected
FCN	Fully Convolutional Network
FFDM	Full-Field Digital Mammography
FLDA	Fisher's Linear Discriminant Analysis
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
FPS	Frames per second
FRCNN	Faster Region-based Convolutional Neural Network
GAP	Global Average Pooling
GARPN	Guided-Anchoring Region Proposal Network
GIOU	Generalised Intersection-over-Union
GT	Ground Truth
HOG	Histogram of Oriented Gradients

IRV2	Inception-ResNet-v2
KPCA	Kernel Principle Component Analysis
LBP	Local Binary Pattern
IOU	Intersection-Over-Union
LDA	Linear Discriminant Analysis
mAP	Mean Average Precision
ML	Machine Learning
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
MRMR	Maximum Relevance - Minimum Redundancy
MS-COCO	Microsoft Common Objects in Context
NAS	Network Architecture Search
NMS	Non-Maximal Suppression
OHEM	Online Hard Example Mining
PASCAL VOC	Pattern Analysis, Statistical Modeling, and Computational Learning Visual Object Classes
PCA	Principle Component Analysis
PCLBP	Phase Congruency Local Binary Pattern
PISA	Prime Sample Attention
R-CNN	Region-based Convolutional Neural Network
RL	Reinforcement Learning
RLBP	Rotation-Invariant Local Binary Pattern
ROC	Receiver Operating Characteristic
ROI	Region-of-Interest
RPN	Region Proposal Network
S-OHEM	Stratified Online Hard Example Mining
SIFT	Scale-Invariant Feature Transform
SPP	Spatial Pyramid Pooling
SSD	Single Shot MultiBox Detector
SVM	Support Vector Machine
t-SNE	t-Distributed Stochastic Neighbor Embedding
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TS	Training Set
ULBP	Uniform Local Binary Pattern
YOLO	You Only Look Once

Declaration of Originality

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis entitled "Breast Lesion Detection in Ultrasound Images using Deep Neural Networks: Clustering Based Approach for False Positive Reduction" are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration.

Anu Bose

Date: 28/06/2023

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is intended to present new deep learning solutions for tumours detection in 2D ultrasound images, in particular, breast lesion detection based on an effective and efficient deep convolutional neural network. This chapter presents an overview of the research and the layout of the thesis. First, the problem statement and research motivation is detailed in Section 1.1. Next, Section 1.2 details the aim and objectives of this research to address the problem. This is followed by Section 1.3 which presents an overview of the research methodology. Section 1.4 lists the contributions made by this work. Section 1.5 details the ethics. Finally, Section 1.6 provides an overview of the structure of this thesis.

1.1 Research Motivation and Problem Statement

Cancer is a serious disease causing the second highest number of fatalities globally [7]. Breast cancer is one of the most common forms of cancer, constituting 11.6% of the global cancer cases [8]. Fundamentally, breast cancer is an uncontrolled growth of cells that eventually leads to formation of lesion(s) [9]. A lesion can be categorised as benign or malignant. Benign lesions are not fatal and do not spread to any other region of the body whereas malignant lesions pose a fatal threat and can spread from its region of origin (also known as primary site) to other parts of the body, either through the bloodstream or through lymph nodes (lymphatic system). If the cancerous cells travel through the bloodstream, distant organs can be affected whereas if the cells travel through the lymphatic system, lymph nodes are likely to be affected [10]. Based on the amount of disruption and spread, American Cancer Society categorises malignant lesions as *local*, *regional* or *distant* [11]. A local lesion

is one that is contained within the breast. A regional lesion is one that has spread to the surrounding tissues/organs and a distant lesion is one that has spread even further to distant tissues/organs. A patient with a local lesion has a 99% chance of survival. However, the survival rate drops to 86% if the lesion progresses to the regional stage. The survival rate drops drastically to 27% if the lesion advances further to the distant stage [11]. Thus, early detection and diagnosis of the lesion is of vital importance.

The first stage of screening for breast cancer is a clinical examination, after which it is common practice to utilise imaging modalities to generate a scanned image of the region. Widely used imaging modalities include Ultrasound (US), Mammography, Magnetic Resonance Imaging (MRI), etc. Modalities such as mammograms and MRI expose the patient to harmful radiation. Mammograms use X-rays for generating a scan of the region while MRI uses magnetic and radio waves. On the other hand, US machines do not expose the patients to harmful radiations as they use sound waves for generating a scan of the region. Additionally, US machines are portable and relatively affordable than other modalities. Unlike other modalities, the process of generating US scans is simple and does not typically invoke anxiety or stress in the patient. Furthermore, US is more effective in screening abnormal regions of the breast that may not appear clearly in mammograms images. For dense breasts, US scans are much more reliable than mammograms [12, 13, 14]. For these reasons, US imaging is currently gaining popularity, especially in developing countries [15]. An important point to mention here is that it is common to use multiple modalities to achieve a more comprehensive view of the complex or difficult-to-assess lesions.

After the scan is generated, a radiologist diagnoses and reports the lesion on the basis of a widely used Breast Imaging Reporting and Data System (BI-RADS) [16]. This system labels lesions on the basis of its type and severity starting from 0 (indicating a need for further examination), 1 (indicating a high certainty of the lesion being non-cancerous) to 6 (indicating lesion proven malignant through biopsy). After this, at least one other radiologist performs a blind diagnosis to confirm the initial reading and ensure that no lesion is missed. As mentioned previously, it is crucial to detect lesion(s) at the earliest possible stage. Unfortunately, we currently face a concerning shortfall of radiologists worldwide, thereby delaying timely diagnosis and treatments of millions of cancer patients globally [17].

In Europe alone, there are only 13 radiologists available per 100,000 patients as of 2022 [18]. The

UK has an acutely low number of 8.5 radiologists per 100,000 [18]. This rate is even lower for developing countries. For instance, Malaysia has only 30 radiologists per million patients [19]. Globally, as well as in the UK, this number has increased at a very slow pace over the last few years, with projections estimating similar slow growth in the following years. Along with the slow growth, the demand continues to grow at a much faster pace due to an increase in cancer patients worldwide [20]. Additionally, the UK also has an ageing population further increasing the requirement [17]. Currently, the UK faces a 29% deficiency of clinical radiologists, predicted to increase further to 39% by 2026 [17]. Furthermore, there are discrepancies in the number of radiologists in various regions of a country. For example, in the UK, London does not face a lack of radiologists while North and West Wales have a considerable lack of 54% [17].

Multiple solutions have been proposed and utilised to address this issue, one of which is the use of Artificial Intelligence (AI) in Computer-Aided-Diagnosis (CAD) systems to assist radiologists in their decision. Use of AI for automated detection of breast lesion(s) in an US image can help radiologists in various ways such as: perform initial scanning to identify suspected cases; support faster confirmation of the diagnosis; and aid in avoiding missed detection/diagnosis. This helps in utilising the radiologists' time more efficiently through faster decisions on detections as well as added assistance in challenging cases [21].

Automating object detection is an established field of research. Traditional methods in this field extract handcrafted features to detect object(s) in an image. Since these methods extract low-level features such as edges, contours, etc., models trained on these features have poor generalisation capabilities. Also, the development of these features requires significant input from the developer and is limited to their knowledge of the domain. In recent years, deep learning (DL) based detectors have become popular as they address the important drawbacks of traditional methods thereby marking a significant developmental stage in this field. DL networks have higher generalisation capability as they extract features of higher abstraction and require little involvement of the developer for selection and extraction of these features. Compared to object detection in natural images, the speed of innovation and development in the field of breast lesion detection in US images is much slower due to the difficulties posed by the inherent characteristics of the US images such as poor resolution, unclear boundary of the lesion, similarity in the texture of the lesion (object) and other tissues (background

region), significant variation in images generated from different machines. Additionally, lack of publicly available dataset and the difficulty and challenges involved in collection of these images further slow the progress in this field.

Therefore, it is common practice to adapt detectors developed for natural images for detection of breast lesions in US images. One such DL-based detector popularly adapted for not only breast lesion detection in US images but also for detection of other lesions in images from different modalities is the Faster R-CNN (FRCNN) network. It is a 2-stage detector where the first stage acts as a coarse detector while the second stage acts as a finer detector, filtering through the output of the first stage. When utilised in its original configuration for breast lesion detection in US images, FRCNN correctly detects a high number of lesions. However, the overall performance of the model drops due to the high number of false positive (FP) detections also output by the model. Successful adaptation of this network for breast lesion detection in US images include either modifications of the modelling hyperparameters and/or the network architecture.

Although the adapted FRCNN models show high performance, they have the following drawbacks. Firstly, these methods do not provide an experimental evaluation of the impact of each modification on the overall performance. As these methods modify several modelling hyperparameters and/or network architecture, understanding the influence of individual modifications can provide valuable insights and facilitate further adaptations by researchers. Secondly, some methods use small to medium sized datasets. Due to large variation in US images collected from different sources, one of the important prerequisites of a reliable detector in this field is high generalisation. Therefore, when developed on small to medium sized datasets, it is hard to gauge the generalisation capabilities of their modifications on datasets collected from different hospitals and generated using different US machines.

In addition to modifications of existing detectors developed for natural images, several novel 2-stage detectors designed for breast lesion detection in US images have also been proposed. However, despite the high performance of these methods, a prominent and recurring issue in all these models is the FP detections. The issue of FP detections is a significant concern not only in the field of breast lesion detection in US images but also in detectors developed for other lesions in images of various modalities. While methods developed for breast lesion detection in US images introduce certain

modifications such as modification in the network training loss to improve the overall performance and reduce FPs, there are no dedicated approaches specifically designed to reduce FPs in DL-based breast lesion detectors developed for US images. In this research, we aim to address these limitations in the current literature by developing a new and effective method for breast lesion detection in US images.

1.2 Aim and Objectives

The aim of this research is to develop a novel and effective solutions for breast lesion detection from static 2D US images. The main research objectives are summarised as follows:

1. To review, understand and acquire knowledge on state-of-the-art object detection methods developed for natural and medical image, focusing particularly on methods developed for breast lesion detection in US images;

2. To evaluate the effectiveness of existing deep and reinforcement learning networks in detecting breast lesions from US images acquired from clinical settings;

3. To adapt Faster R-CNN network for breast lesion detection in 2D US images with the goal to reduce the false positive detections while incurring minimal negative impact on the number of correct detections;

4. Improving the number of correct detections as well as reducing the number of false positive reduction in the adapted Faster R-CNN model through improving its classification accuracy;

5. To develop a novel method that combines Faster R-CNN network and region proposal clustering for effective reduction of challenging and persistent false positive detections with negligible effect on the number of correct detections;

6. To embed medical domain knowledge to improve the classification accuracy of the region proposal to further reduce the false positive detections;

7. To evaluate and compare the proposed methods against various state-of-the-art techniques of breast lesion detection using large datasets collected from different hospitals.

The outcome of this research is an added value to the Computer-Aided Diagnosis systems. It provides detection models for processing and identifying lesions in US images collected in clinical settings. Such detection models supports other tasks such as lesion classification and segmentation. But it is worth bearing in mind that this research is only concerned with lesion detection in 2D static image, not lesion detection from video nor lesion contour segmentation. In addition, although the presented solutions may well be applicable to different cancer types (e.g. thyroid and prostate), the scope of this research project is mainly focused on breast lesion detection.

1.3 Research Methodology: An Overview

The research in this thesis follows the approach of investigation to evaluations to comparisons to major development. Empirical evidence from experiments and analysis are used to support the creation and development of new and novel solutions and models. A blend of both deductive and inductive reasoning based on sound understanding is practised throughout the research. This section presents a brief overview of our general methodology.

This work aims to improve reliability of breast lesion detection methods in 2D static US images through reduction of persistent and challenging FP detections while maintaining a high number of detected lesions. To accomplish this, the work is developed in an iterative fashion where the outcome of each stage is used for the development of next. First, we start by evaluating the state-of-the-art object detection techniques including FRCNN as one of the most powerful approaches for object detection. After selecting FRCNN as best performing method for breast lesion detection, several network hyperparameters have been examined for reducing the FPs detection and the overall detection performance. We modified FRCNN by searching for optimal network hyperparameters which results in a new network called adapted FRCNN. Such adaptations provide in-depth understanding of the two stages of object detection techniques, limitations, and ultimately areas of improvement. After the adaptation of FRCNN, we propose a new detection technique called U-Detect that uses FRCNN as a base detection network and unsupervised learning to reduce the FPs detection. Finally, inspired by the domain knowledge of breast lesion characteristics, we modify U-Detect by introducing U-DetectH that uses a set of handcrafted texture and morphological features to improve the overall detection accuracy and reduce FPs detections in particular.

In this research, the evaluation plays a major role for validating and testing alternative solutions as well as evaluating performance of various networks. Therefore, the data quality is crucial. To ensure the soundness of the research outcomes, the collected US images are of clinically acceptable quality, acquired using different medical centres and several US machine makers, covering a large range of breast lesions of benign and malignant of different sizes. Experimental protocols as well as performance metrics used in this research will be further explained in later chapters.

1.4 Contributions

The contributions of this research can be summarized as follows:

1. A good understanding and critical evaluation of various existing deep learning methods for breast lesion detection in 2D US images;. The thesis provides a comprehensive analysis of Faster R-CNN network performance limitations through a systematic evaluation.

2. Adapt Faster R-CNN hyperparameters for creating effective models for detecting breast lesions from ultrasound images. This involves exploring the effects of Faster R-CNN hyperparameters (anchor boxes, base network's training samples' selection, training and test proposals, training loss) in reducing the false positives in breast lesion detection.

3. A novel detection method (U-Detect) that uses the x-means clustering and trainable features to reduce the false positive detection of the region proposal.

4. A novel detection method (U-DetectH) that uses the x-means clustering and both trainable and handcrafted features to reduce the false positive detection of the region proposal.

5. A new region proposal candidate merging method to reduce the overlapping of the false positive detections.

6. A decision fusion method to refine the classification score of region proposals to further improve the single false positive detections cases as well as the number of correct detections.

7. Evidence of the overall effectiveness of the U-Detect and U-DetectH methods through extensive analysis and experiments using US breast images collected from different hospitals and clinical settings.

8. Compare U-Detect and U-DetectH methods against the state-of-the-art detection methods designed for breast lesion detection and objects detection in natural images.

1.5 Ethics

This project is part of TenD Buckingham Research and Development Centre (TBRDC). TBRDC is a collaboration partnership between TenD AI Medical Technologies Ltd and University of Buckingham. All images used for this project were collected by TenD AI Medical Technologies Ltd, thus acting as the third party provider for this work. These images were collected from various hospitals in China in agreement with TenD AI Medical Technologies Ltd. All images were anonymized by TenD AI Medical Technologies Ltd. Collected data consists of the images and their respective labels. Nature of the tumour as revealed through pathology test and tumour location marked by an experienced radiologist. The images are securely stored on the local share point created by TBRDC with limited access only to the researches involved in this research. No participants were recruited in Buckingham for the purpose of this project.This research was granted ethics approval by the Research and Ethnics Committee of the School of Computing, University of Buckingham before start of this project.

1.6 Thesis Structure

The remainder of this thesis is organised as follows. **Chapter 2** describes background concepts and methods relevant to this research. **Chapter 3** presents review of the literature in breast lesion detection in 2D US images as well as detection of different types of lesions in images from different screening modalities. Experimental setup used for this research including datasets and evaluation metrics is described in **Chapter 4**. **Chapter 5** presents the adaption of FRCNN for breast lesion detection in US images. **Chapter 6** presents a novel idea of proposing a new detection method based on combining FRCNN and x-means clustering. The investigation results from the previous chapter which provides the foundation for the proposed idea in this chapter. **Chapter 7** proposes a classification approach based on handcrafted features for improving candidate selection in the clustering method through improving the final detections. Discussion on the important concepts and findings in this research is presented in **Chapter 8**. **Chapter 9** concludes the thesis and describes the future work.

Chapter 2

Background

This chapter describes fundamental concepts and methods relevant to this research. Section 2.1 provides an overview of breast cancer including a description of lesions, various modalities used for breast cancer screening with particular focus on US. Section 2.2 describes concepts relating to the automation of breast lesion detection ,including methods for extraction of features from an image, dimensionality reduction methods for the extracted features to machine learning (ML) methods that utilise the extracted features for classification and object detection tasks. In Section 2.3, deep learning networks that overcome the drawbacks of traditional methods are described. Furthermore, a detailed description of Faster RCNN (FRCNN) is provided in this section. Finally, this chapter is summarised in Section 2.4.

2.1 BI-RADS for Breast Cancer Screening

BI-RADS was developed by the American College of Radiology to report breast US, mammogram, and MRI scans. In this reporting system, a scan is assigned a BI-RADS score from 0 to 6. Each score is associated with a predefined diagnosis. For example, if a scan is assigned a BI-RADS score of 1, then it indicates no lesion was found in the scanned region. Each BI-RADS score and its associated diagnosis is detailed in Table 2.1. To categorise a lesion in an US image using BI-RADS, various aspects of tumour are considered such as shape of the mass, its orientation with respect to the skin surface, its margin, echo pattern of the mass and its posterior region, presence of calcifications in and around the lesion, etc. Detailed list of these features is shown in Table 2.2.

BI-RADS Score	Diagnosis
0	Need additional imaging or prior exam
1	Negative (Essentially 0% cancer likelihood)
2	Benign (Essentially 0% cancer likelihood)
3	Probably Benign (>0% and ≤2% cancer likelihood)
4	4a. Low suspicion for malignancy (>2% to ≤10%)
	4b. Moderate suspicion for malignancy (>10% to ≤ 50%)
	4c. High suspicion for malignancy (>50% to <95%)
5	High suspicion of malignancy (≤95% cancer likelihood)
6	Biopsy proven

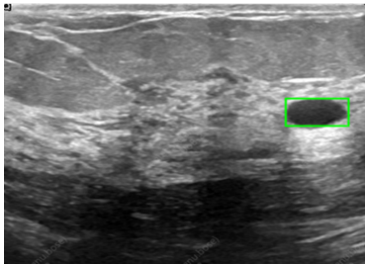
Table 2.1: BI-RADS scoring for ultrasound scans [6].

Characteristic	Types
Breast Composition	1. Homogeneous - fat
	2. Homogeneous - fibroglandular
	3. Heterogeneous
Mass	Shape: Oval, round, irregular
	Margin: Circumscribed or not circumscribed (indistinct, angular, microlobulated, spiculated)
	Orientation: Parallel or not parallel
	Echo Pattern:Anechoic, hyperechoic, complex cystic, solid hypoechoic,isoechoic,heterogeneous
	Posterior Features: No features, enchancement, shadowing, combined pattern
Calcifications	In mass, outside mass, intraductal
Associated Features	Architectural distortion, duct changes, skin thickening, skin retraction, edema, vascularity (absent, internal, rim), elasticity.
Special cases	Simple cyst, clustered microcysts, complicated cyst, mass in or on skin, foreign body (including implants), intramammary lymph node, AVM, Mondor disease, postsurgical fluid collection, fat necrosis

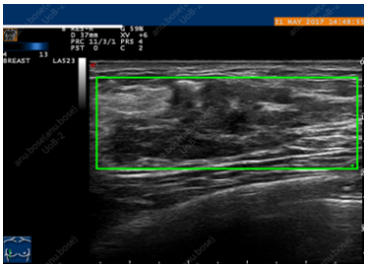
Table 2.2: Ultrasound lexicon [6].

For example, margin is useful in identification of the lesion type and its consecutive BI-RADS score. A benign tumour typically has a clear and distinct margin as shown in Figure 2.1a. On the other hand, a malignant tumour typically has an unclear and irregular margin as shown in Figure 2.1b. Another typical sign of malignancy is the orientation of the lesion. A benign lesion is usually parallel to the skin or appear 'wider-than-tall' in the US image as shown in Figure 2.1a whereas malignant lesions are generally perpendicular to the skin or they appear 'taller-than-wide' in the US image as shown in Figure 2.1b. It is important to highlight here that no single feature is a definite sign of the type of cancer. As previously mentioned, to ensure low negative impact on the body and high survival

rate, it is of vital importance that the lesion is detected at the earliest possible stage. However, we currently face a serious lack of trained radiologists. Thus, automating lesion detection can assist radiologists in faster diagnosis by acting as secondary opinion. This automation is achieved using machine learning methods for image analysis. The following section provides a detailed explanation of this field, focusing on relevant techniques and approaches.



(a) Benign lesion



(b) Malignant lesion

Figure 2.1: Benign and malignant lesions in US images (Green box indicates the lesion).

2.2 Feature Extraction, Dimension Reduction and Machine Learning for Image Analysis

Image analysis is a well-established field of research. An image is fundamentally a matrix of pixel values. Thus, image analysis is the study of the nature of the image pixels to extract useful information i.e. features which are then utilised to train ML models to perform tasks such as object detection. We first describe the commonly used feature extraction methods in Section 2.2.1. The extracted features generally have a high dimension which increases computation time of the ML model and may negatively impact the overall performance of the model. Over the years, several methods have been developed to efficiently reduce the dimension of the extracted features without losing integral information. Popular methods of dimension reduction are described in Section 2.2.2. Using these features, ML models are trained to perform tasks such as classification or object detection. The three types of ML models and their utilisation for the various tasks is described in Section 2.2.3.

2.2.1 Feature Extraction

Feature extraction is an important aspect of image analysis. Features can be defined as patterns in an image. Traditional machine learning methods utilise handcrafted methods to extract features

for image processing applications. However, these traditional methods have important drawbacks which are addressed by a relatively new approach called deep learning. This section describes both handcrafted and deep learning-based feature extraction methods.

2.2.1.1 Handcrafted Features

Handcrafted methods extract low-level features such as edges of an image. These methods rely on the knowledge of the developer for extraction of important, distinguishing features. Generally, each method extracts either global features like contrast of the image or local features such as edges and shapes. To ensure that important textural and morphological features in the image are captured, it is common practice to use multiple techniques together. Some of the popular handcrafted feature extraction methods include HOG [22], LBP [23], ULBP [24], SIFT [25], GLCM [26], filters such as Sobel filter, etc. For analysis of breast lesions in US images, HOG, ULBP and GLCM are commonly used. Therefore, further details of these methods are provided in this section. HOG and ULBP capture local features whereas GLCM captures global features.

Histogram of Oriented Gradients (HOG) is a simple and efficient method for capturing local textural features in an image, focusing on the shape of objects. This is achieved by extraction of local features such as edges and contours which contribute in the identification of the overall shape of objects in the image. HOG captures the degree (magnitude) and direction (gradient) of change in pixel intensities with respect to x-axis and y-axis for each pixel in an image and represents this information in a histogram. One of the popular bin sizes for the histogram is 20 degrees which produces a 9×1 dimensional histogram if the angles are unsigned ($\theta \in [0, 180]$) or a 180×1 dimensional histogram if the angles are signed ($\theta \in [0, 360]$). HOG features are invariant to geometric transformations and relatively robust to changes in illumination. Due to these characteristics, HOG is widely used for feature extraction.

Local Binary Pattern (LBP) [23] is another simple and widely-used method of feature extraction which captures local textural features using compact descriptors. LBP computation begins by creating a binary pattern for a cell of predefined number of neighbouring pixels using the following thresholding operation: if any neighbouring pixels have equal or greater intensity than the centre pixel, then they are assigned a value of 1, otherwise they are assigned 0. Each value in this pattern is then

multiplied with its corresponding binomial factor to generate the LBP value for that cell. LBP feature vector for an image is a histogram of LBP values computed for each cell. Despite its advantages, LBP suffers from high dimensionality and is negatively impacted by rotation i.e., the LBP of the original image would not be the same if the image was rotated. To overcome these limitations, several adaptations of LBP have been proposed over the years. One of the effective and simple adaptations is Uniform LBP (ULBP) [24].

ULBP captures the frequency of uniform binary patterns in an image. As this textural information varies for images of different classes, capturing it aids in accurate image classification. A binary pattern is categorised as 'uniform' if it contains up to 2 transitions in bit value. All other patterns with more than 2 transitions are labelled as non-uniform. Binomial factors of only the uniform patterns are recorded, while for all non-uniform patterns, P+1 value is recorded where P is the number of neighbouring pixels. ULBP can be recorded as signed or unsigned. Signed ULBP is similar to rotation-invariant LBP (RLBP) [27] where only the minimum value for the binary pattern of the cell is considered whereas unsigned ULBP is the same as LBP. Unsigned ULBP feature vector has a dimension of $P + 2$ of which $P + 1$ bins are reserved for the uniform patterns and 1 bin for all non-uniform patterns. Signed ULBP feature vector has a dimension of $(P \times (P - 1)) + 3$ of which $(P(P - 1)) + 2$ bins are reserved for uniform patterns and 1 bin for all non-uniform patterns. Thus, use of ULBP reduces the size of the feature vector. For $P = 8$, LBP of an image is 256×1 dimension whereas unsigned ULBP is 10×1 and signed ULBP is 59×1. Therefore, ULBP effectively combines the advantages of LBP and RLBP while significantly reducing the feature vector dimension. Due to these advantages, ULBP is one of the common variations of LBP used.

Gray Level Co-occurrence Matrix (GLCM) [26] is another popular method used to extract global textural features unlike HOG or LBP (and its variants) where local textural features are extracted. Specifically, GLCM extracts second-order textural features by capturing the relationship between a pair of pixels. During GLCM computation, the pixel being analysed is referred to as 'reference' pixel and its neighbouring pixel is referred to as 'neighbour' pixel. Distance between the reference and neighbour pixels is called 'offset'. One of the popular methods used to describe a neighbour pixel is using its position with respect to the reference pixel. To extract GLCM features, a GLCM matrix is constructed for single reference-neighbour pixel relationship.

The GLCM matrix is a square matrix whose dimension can be based either on the range of pixel values in an image or the quantization level of the image. The GLCM matrix is constructed by registering the frequency of every pair of pixels with the predefined reference-neighbour relationship in the image. The generated GLCM matrix is then normalised. For each GLCM matrix, a number of features are computed which form the GLCM feature vector of the image. The original GLCM paper extracts 14 features from a single GLCM matrix. Some of the popular features computed for breast lesions are contrast, energy, correlation and entropy.

2.2.1.2 Deep Learning Based Feature Extraction

Although traditional methods of feature extraction have various advantages, they suffer from important drawbacks. These methods rely on the developer to determine the optimal setup (such as cell size) to extract useful features. Also, as these methods extract low-level features such as edges, shapes, contrast, etc., they have low generalisation capabilities. In recent years, deep learning (DL) networks have become more prominent for feature extraction as these networks overcome the drawbacks of the traditional methods of feature extraction. This section provides further explanation of DL networks, their utilisation for feature extraction, and their effectiveness in overcoming the drawbacks of traditional feature extraction methods.

Perceptron, also referred to as neuron, is the building block of all DL architectures. It is a decision-making unit, which takes several binary inputs and outputs a binary value [28]. For complex tasks, a network of neurons called neural network or multi-layer perceptron (MLP) is used. As the complexity of the task increases, the number of neuron layers also increases. Neural networks with a considerable number of perceptron layers are called deep neural networks. For two or higher dimensional inputs such as images, convolutional neural networks (CNNs) are more effective than MLPs. Irrespective of the network structure, each neuron in the network requires optimization of its parameters to produce the desired output. This optimization is performed automatically during model training using a popularly used technique called backpropagation. Backpropagation involves the use of training loss which is the measure of the difference between the expected output and the output generated by the model. Based on this training loss, weights and other learnable parameters of the network are updated so as to generate the desired output i.e. minimise the training loss. Thus, the network automatically learns

the features to be extracted without any involvement of the developer.

In order to extract features from a CNN, it needs to be trained for its parameters to be optimised. Generally, CNN is trained for image classification. For this, two to three perceptron layers (also called fully connected (FC) layers) are added after the last convolution layer. Output of the CNN layer is vectorised and served as the input of the FC layers. The final FC layer is the output layer and contains the exact number of neurons as the number of classes in the dataset. However, the FC layers are prone to overfitting. Regularisation methods such as dropout are used to address this issue. A recent method to address the overfitting issue of FC layers is the use of global average pooling (GAP) layer. The GAP layer is generally placed before the FC layers or, in some networks, completely replaces the FC layers (except the FC layer acting as the output). The GAP layer computes the average of each feature map, output by the last convolution layer, and generates vectorised averages for further processing. This layer has no learnable parameters as it only computes the average, thereby completely removing the overfitting issue. Thus, it acts as a 'structural regulariser'.

After the CNN model is trained, it can be used to extract features of images for a range of applications. Initial layers of the CNN extract low-level features while the deeper layers extract higher-level, abstract features. Due to the higher quality of features, DL networks have higher generalisation capabilities than traditional methods. Thus, DL networks are more suitable for real-life applications than traditional methods [2]. Equally important is the lower degree of involvement of the developer in individually selecting features to be extracted. These characteristics render the DL based networks versatile and easily adaptable to different domains without requiring significant changes to the network unlike handcrafted features which require modifications with change in dataset or domain. Although theoretical knowledge of DL networks has been established for many decades, it was only recently that these methods became more popular. This was due to the availability of higher computational capacity. The first CNN to successfully use DL was AlexNet [29]. Features extracted from AlexNet were used for classification of natural RGB images. This was a significant breakthrough in the field of image processing. Some of the recent state-of-the-art CNNs include ResNet [30], Inception-ResNet-v2[31], ResNeXt [32], etc.

2.2.2 Dimension Reduction

Features extracted from an image generally have large dimensions. Use of high dimensional feature vectors increases computation time of the ML model. Furthermore, in higher dimensional feature vectors, the data becomes sparse. The impact of noisy features in such a feature vector is generally larger. Additionally, finding patterns using such high-dimensional feature vectors is a challenging task resulting in poorer overall performance of the model. This issue caused by the high dimensionality of the feature vectors is referred to as the 'curse of dimensionality'. Several techniques have been developed to reduce the negative impacts of the high feature dimensionality, one of which is dimensionality reduction methods. Dimension reduction methods can be classified into the following two categories as shown in Figure 2.2: feature selection and feature projection. Feature selection methods reduce dimension through retention of only important, distinguishing features and removal of redundant and noisy features. On the other hand, feature projection methods, as the name suggests, project the feature vector to a lower-dimensional space where the new dimensions are a linear or non-linear combination of the original features.

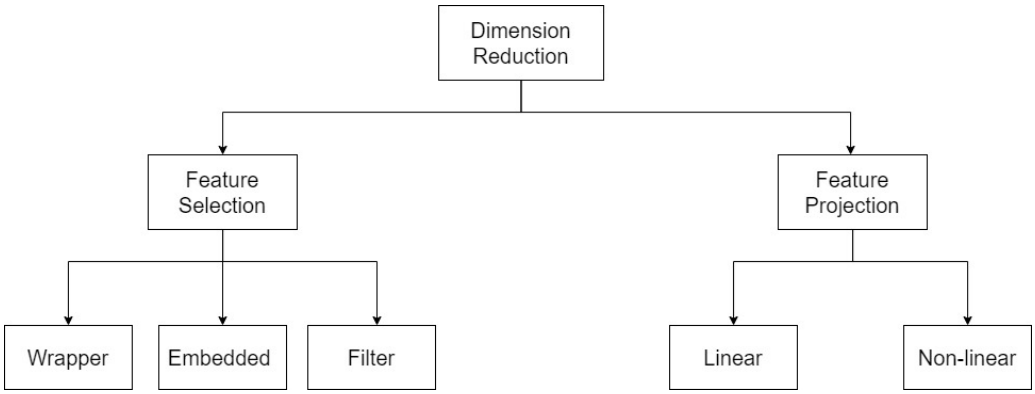


Figure 2.2: Types of dimension reduction methods.

Feature selection methods can be further divided into three categories; wrapper, embedded and filter methods. Wrapper methods reduce dimension by evaluating various sets of features and retaining only the best performing set. Examples of this method include forward and backward selection methods. Both these methods are computationally heavy and time-consuming as it requires training and testing models for every feature set. Filter methods compute each features' importance using statistical methods such as correlation, information gain, chi-square, etc. and redundant and/or irrel-

evant features are removed. These methods are faster than wrapper methods since it does not require training and testing multiple models. Embedded methods combine the filter and wrapper methods where a small set of important features are first evaluated. New features are added iteratively in the order of their importance and evaluated. Best performing set is then selected as optimal. Some examples of embedded methods include Lasso L1 and tree-based methods. Feature projection methods can be categorised as linear and non-linear methods based on the method used for creation of the new dimension. Linear methods project the feature vectors onto a new dimension created from a linear combination of the original features. Some popular examples of this method include PCA [33], LDA [34] and SVD [35]. Non-linear methods are applicable for non-linear data and examples include Kernel PCA (KPCA) [36], t-SNE [37] and auto-encoder [38]. PCA and KPCA are commonly used and are therefore explained in further detail.

Principal Component Analysis (PCA) is a linear feature projection method which not only reduces the dimension of the feature vectors (input data) but also helps in interpretability of high-dimensional data. Essentially, PCA projects the original data into a new space where the new dimensions are a linear combination of the original dimensions. The process begins by standardising the input feature vectors due to the sensitivity of PCA to variance. Standardisation is typically done by subtracting each dimension with its mean. Next, a covariance matrix is generated. While variance measures the spread of the data in a single dimension, covariance measures the spread between two dimensions. Covariance between two dimensions provides information on the change of each dimension with respect to each other. For example, consider a 2D dataset where the dimensions represent height and width. Here, variance of each dimension represents change in height or width with respect to its own respective mean. On the other hand, covariance represents change in height with respect to width or vice versa. Covariance of a dimension with itself is its variance.

Next, eigenvectors and eigenvalues are computed for the covariance matrix. Eigenvectors represent the direction of maximal variance in the dataset. They are unit-length vectors as they only indicate the *direction* of maximal variance. The eigenvectors are fundamentally a linear combination of the original dimension. Eigenvalues show the degree of variance captured by the eigenvectors i.e., the amount of information captured. Following this computation, all eigenvectors are reordered in descending order of their eigenvalues. Eigenvector with the highest eigenvalue captures the largest amount of

information. First Principal Component (PC) is the eigenvector with the largest eigenvalue. For dimension reduction, PCs with a certain amount of cumulative variance (sum of eigenvalues) are selected. For example, in a 3D data, if the cumulative variance of first 2 PCs is 99% (meaning sum of eigenvalues of the two PCs is 99%), then the third PC can be dropped for dimension reduction with only 1% loss in information. Cumulative variance is also referred to as cumulative information (CI). Furthermore, visualisation of the data in various combinations of the PCs provide useful information on the nature of the dataset. The original dataset can be transformed from the PC space using Equation 2.1.

$$(OriginalData)^T = (FeatureVector) \times (NewData) + (OriginalMean) \tag{2.1}$$

PCA is a simple and popularly used method of dimension reduction. However, PCA relies on the data being linearly separable, limiting its use in non-linear data. To overcome this limitation, **Kernel PCA** (KPCA) was proposed. KPCA first transforms the n-dimensional original data into $n+1$ dimension. This transformation makes the data linearly separable. After this, standard PCA steps are applied. Instead of computing individual values for the new dimension(s), kernel trick is used. Some of the commonly used kernels include linear, polynomial, gaussian, gaussian RBF, etc. These kernels are defined in Equation 2.2. In application of breast lesion in US images, gaussian and gaussian RBF are most common.

$$\begin{aligned} Polynomial &\implies K(\bar{x}_i, \bar{x}_j) = (r + \bar{x}_i \cdot \bar{x_j})^n \\ RBF(Gaussian) &\implies K(\bar{x}_i, \bar{x}_j) = exp(\frac{-||\bar{x}_i - \bar{x}_j||^2}{2\sigma^2}) \\ Sigmoid &\implies K(\bar{x}_i, \bar{x}_j) = tanh(\sigma \bar{x}_i \bar{x}_j + r) \end{aligned} \tag{2.2}$$

2.2.3 Machine Learning for Image Analysis

Once the appropriate features are extracted, ML techniques are used to develop models that utilise these features for a variety of tasks such as image classification, object detection, object segmentation, etc. ML methods can be broadly classified into three main types based on the dataset used, namely, supervised, unsupervised and reinforcement learning based methods. Remainder of this section provides further details on these ML methods.

2.2.3.1 Supervised Machine Learning

Supervised learning methods require labelled dataset to train the model. Owing to the model's reliance on the labels provided by a 'supervisor', this type of ML is referred to as supervised ML. Here, the ML model can be trained to either classify unseen test data into one of the predefined classes (discrete output such as assigning a category or object class) or predict an outcome/trend (continuous output such as weather prediction). Some of the popular supervised ML methods include linear regression, logistic regression, decision tree, random forest, neural network, SVM, etc. Linear regression essentially identifies the relationship between dependent and independent variables. Here, the outcome is continuous. Logistic regression is similar to linear regression but its outcome is discrete.

Decision tree, as the name suggests, is a tree-like model consisting of a parent node, internal nodes and leaf nodes. The internal nodes represent individual features of the data and leaf nodes represent the output classes. Final categorisation of the data is achieved with the help of conditional statements to direct the flow between the internal nodes, providing a visual representation of the decision-making process. This is typically easier to apply on smaller datasets. Random forests is a collection of decision trees, designed to reduce variance thereby increasing the accuracy of the final decision. When used for continuous output, it is referred to as 'regression trees'. K-nearest neighbour is a relatively simpler, non-parametric method of supervised learning. Here, the classification of a data sample is based on its closeness to predefined 'k' nearest neighbours; the class assigned to this data sample is the most common class of its 'k' nearest neighbours. In this research, SVM and SoftMax are utilised. Both SVM and Softmax are popularly used in traditional and DL based networks. Thus, the remainder of this section provides a deeper explanation of these methods.

Support Vector Machine (SVM) is used for classification and regression tasks in various application fields including image analysis. In this research, SVM is utilised for classification tasks. Thus, the following discussion of this method is in the context of its use for classification. Due to its simple and sophisticated design, it is still in popular use since its inception in 1995 [39]. Originally, SVM was designed for linear data. SVM separates an n-dimensional data using a $(n-1)$ dimensional optimal hyperplane. Once this hyperplane is generated for training data, any unseen sample can be classified in the appropriate class depending on its position with respect to the hyperplane.

The output of a trained SVM is a *score* which represents the distance of that data from the decision boundary. Based on the nature of the dataset, SVM can be either hard-margin or soft-margin based. In hard-margin based SVM, data samples belonging to separate classes have no overlap with each other; samples from each class lie in their distinct region. However, in many real-life scenarios, there is no clear separation between the classes of the data; a small minority of samples from one class lie in the other class. In such cases, soft-margin method is used where the aim is to separate the *majority* of the data of individual classes, allowing small errors or overlaps.

Considerable majority of real-life datasets are nonlinear in nature. Since they are not linearly separable, the traditional linear SVM cannot be applied. To address this issue, a non-linear SVM model was developed. Here, the dataset is mapped onto a higher dimension where the dataset is linearly separable. Thus, after this mapping, the standard SVM technique can be applied. For the transformation, kernel trick is used. Instead of computing values of the new dimension for each data sample, kernel trick directly provides the dot product of the transformed sample pairs to be used. Commonly used kernels are defined in Equation 2.2 in Section 2.2.2. SVM is in popular use due to its robustness and applicability in high dimensional data.

Apart from SVM, **SoftMax** is also popularly used as a classifier, especially in DL applications. Unlike SVM, SoftMax considers all input values in relation to one another. While SVM outputs uncalibrated values as the classification scores, SoftMax outputs probabilities for each class which add up to 1. Thus, the SoftMax provides easier interpretation of the classification output. As an example, consider a DL network with SoftMax as the classifier. Here, features of the final layer (typically an FC layer) are processed through a SoftMax layer i.e., a layer of neurons with SoftMax as the activation function. The number of neurons in the SoftMax layer is set to the total number of classes in the dataset. For training a model with SoftMax, cross-entropy loss is used. As the output predictions add up to $[0, 1]$, SoftMax encourages assignment of higher scores to the most likely class and lower scores to all other classes. Due to its advantages, it is commonly used in DL networks for a range of applications including image processing applications such as object detection, image classification, etc.

2.2.3.2 Unsupervised Machine Learning

Unsupervised ML is used when the ground truth label is unknown. These methods are typically used to find patterns or trends in the data in order to develop a deeper understanding of the same. One of the popular methods of unsupervised ML is cluster analysis or clustering. Here, the dataset is grouped into clusters where all data samples in a cluster are *similar* to each other. Equally, samples in one cluster are dissimilar to those in all other clusters. Similarity between samples is measured using a predefined distance metric. Clusters are typically represented by a centroid. This technique is useful in statistical data analysis applied in many fields including pattern recognition and image analysis. Clustering methods can be classified into five categories based on the technique and/or principle used as shown in Figure 2.3. These categories include distribution-based, density-based, fuzzy, hierarchical and partition-based clustering methods.

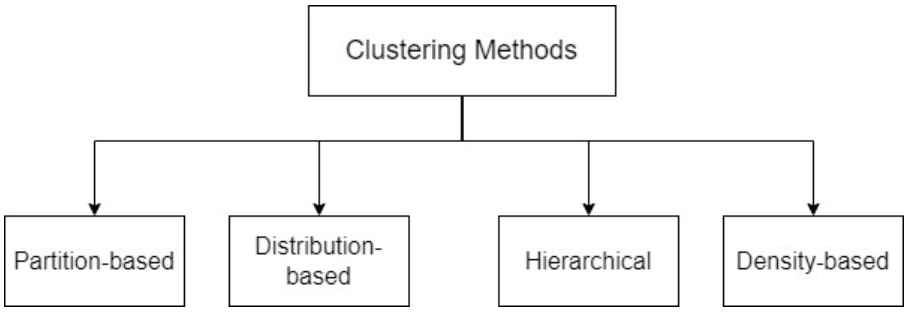


Figure 2.3: Categories of clustering methods.

In distribution-based methods, the clusters are formed based on the assumption that they follow a certain standard distribution such as normal/gaussian distribution. Density-based methods cluster the data on the basis of density such that densely packed regions or elements close to each other are grouped into one cluster. A popular density-based method is DBSCAN [40]. Fuzzy methods allow all samples a degree of membership to every centroid such that an element can be partially a part of multiple clusters. Methods that allow membership to multiple centroids are also referred to as soft-clustering methods. Hierarchical methods group data samples in a hierarchy of clusters. This hierarchy is usually visualised using a dendrogram. Hierarchical methods are either agglomerative or divisive. Agglomerative methods use a bottom-up approach where the process begins by assigning all samples to individual clusters. On the other hand, divisive methods follow a top-down method starting by grouping all data samples in a single cluster and dividing the cluster further in each iteration

till a predefined condition is met.

Partition-based methods divide or 'partition' the data into a predefined number of clusters based on a distance metric. One of the popular examples of this method as well as a widely adopted clustering technique is k-means clustering [41] and its adaptations. In k-means clustering method, after the dataset is divided into a predefined number of clusters,each cluster is represented by a centroid which is the mean of all samples in that cluster. The aim of the algorithm is to minimise the distance between the samples belonging to one cluster while increasing the distance between samples of different clusters. This distance can be computed using any of the standard metrics based on the application. Euclidean distance is the most commonly used distance metric. The process of clustering data samples using k-means clustering method involves the following steps beginning with standardisation of the dataset such that it has a mean 0 and standard deviation 1. As k-means clustering relies on distance between data samples, standardisation of the dataset ensures a proportional impact of the individual features of each sample on the distance measurement. After standardisation, k number of elements are randomly selected as centroids. In the next step, all elements are assigned to one of the k clusters. This step is referred to as the cluster assignment step. In this step, the distance of all samples from each k centroids is computed. Each sample is then assigned to the cluster of the closest centroid.

After cluster assignment, new centroids are computed for each cluster. This new centroid is the mean of the cluster. This step of computing a new centroid is referred to as the centroid update step. The process continues with cluster assignment and centroid update steps performed iteratively until the centroids in two consecutive steps remain constant (convergence) or the maximum number of iterations is reached. Thus, k-means clustering is a simple and computationally light clustering method which can be scaled to cluster large datasets with considerable feature size. However, this method has an important drawback relating to the selection of initial centroids (seeds). Poor selection of seeds can lead to poor overall clusters and/or high computation time due to larger number of iterations required for convergence.

Since its inception, many methods have modified k-means to overcome its shortcomings or further improve its overall performance. One such method, called K-means++ clustering [42], addresses the issue of sensitivity to seed selection in k-means clustering by replacing the randomised selection with

an intelligent selection mechanism. After data standardisation, K-means++ clustering begins with the selection of a single random data sample as the first centroid. Next, the distance of all samples from this centroid is computed. The sample furthest away from this centroid is selected as the next centroid. This process continues till all k centroids are selected. After this, the cluster assignment and centroid update steps of k-means clustering are followed. Although this process of seed selection is comparatively more time-consuming, the quality of clusters formed using these seeds is higher and the overall computation time is reduced due to faster convergence.

Both k-means and k-means++ require a predefined k. However, in some datasets, the optimal number of clusters is unknown. In such datasets, the optimal k can be found using a manual or automatic approach. One of the manual methods to find optimal k is using the elbow method. Here, a range of k values are applied to the dataset. After clustering, for each k, the mean of the squared distance of each sample in a cluster to its centroid is computed. This value is also referred to as 'distortion'. The trend in change in the distortion value with respect to change in k depends on the distance metric used. For instance, if distortion is measured in terms of Euclidean distance, as k increases, the number of samples in a cluster decreases which results in lower distortion. Evaluated values of k and their respective distortion values are plotted in a graph with k in the x-axis. The value of k where distortion stabilises is referred to as the 'elbow'. The k value at the elbow is chosen as the optimal since it provides a good balance between distortion and number of clusters.

Besides manual approach, automation is also used to determine optimal k. One such automation method is x-means clustering [1]. X-means clustering is an adaptation of k-means clustering where instead of a predefined value of k, a range of k values are provided and the algorithm finds the optimal k in this predefined range. Assuming the range of k is $[k_{min}, k_{max}]$, the first step here is to cluster the samples for k_{min} using the k-means clustering method. In the next step, each centroid is further divided into two child centroids and k-means is performed in the individual clusters using the new child centroids. To decide whether the parent centroid or the child centroids are the optimal, Bayesian Information Criteria (BIC) as defined by Kass and Wasserman (1995) is used as the metric. BIC is defined in Equation 2.3.

$$BIC = (\sum_{n=1}^{k} R_n logR - R_n logR_n - \frac{R_n M}{2} log(2\pi\hat{\sigma}^2) - \frac{1}{2}M(R_n - 1)) - C \cdot k \cdot logR(M+1) \quad (2.3)$$

Here, k is the number of clusters, R_n is the number of proposals in n_{th} cluster, R is the total number of proposals in all clusters, M is the size of the feature vector and C is the weight assigned to the penalty term which is set to a default value of 0.5. Variance $\hat{\sigma}$ is defined in Equation 2.4.

$$\hat{\sigma}^2 = \frac{1}{(R-k)M} \sum_{n=1}^{k} abs(x_i - \mu_i)^2 \quad (2.4)$$

The first term in Equation 2.3 calculates how well the elements fit in the clusters using log-likelihood and the second term penalises high numbers of clusters. Thus, a balance between good fit and number of clusters is achieved. If the BIC of the child centroids is greater than that of the parent centroid, then the child centroids are preserved and the parent centroid is discarded, and vice versa. An example of this merging and division of clusters is shown in Figure 2.4. Here, the initial three clusters are individually divided into two clusters each as shown in Figure 2.4a. In clusters where the BIC score of the child clusters is greater than that of the parent cluster, the child clusters are retained whereas in the opposite case, the parent is maintained as shown in Figure 2.4b.

(a) BIC Computation (b) Final Clusters

Figure 2.4: X-means clustering using BIC [1].

This process continues till the total number of clusters remains unchanged for two consecutive iterations or till k_{max} is reached. Thus, x-means clustering automates the selection of optimal k from a range of k values thereby addressing the drawback of k-means and k-means++ clustering methods.

2.2.3.3 Reinforcement Learning Based Machine Learning

Reinforcement learning (RL) based ML methods consists of a goal-oriented agent trained using positive and negative reinforcements provided in the form of a scalar feedback loop. The goal of the agent is to find the optimal path for completion of the assigned task. For instance, some RL detectors are trained with the goal to find objects in an image using cost-effective paths (least number of steps). Unlike supervised learning, RL methods do not depend on an external supervisor to provide the expected output for each step. Instead, the agent learns based on its own past experience which consists of past actions and their respective rewards or punishments. RL is also different from unsupervised learning in that it does not find patterns in the input training data. Unlike unsupervised learning, RL relies on a feedback signal to learn the optimal path in order to reach its predefined goal.

2.3 Object Detection

Object detection is a well-established field of image analysis. Traditional methods of object detection consist of the following four stages: image preprocessing, image segmentation, feature extraction and classification. In the first stage, the image is preprocessed using various filtering and image enhancement to remove noisy data which would otherwise negatively impact the model performance. In the second stage, the image is segmented in order to extract regions of the image that potentially contain an object. Commonly used segmentation methods include thresholding, edge detection techniques, pixel clustering, etc. In stage three, the features of the selected regions are extracted. The traditional methods commonly use handcrafted methods of feature extraction. Finally, the extracted features are classified into one of the object classes or as background.

Designing each stage requires considerable knowledge of the corresponding fields. Recent years have witnessed the development of a considerable number of deep-learning (DL) based detectors due to the following reasons. First, traditional methods rely on the use of handcrafted feature extraction methods. As described in Section 2.2.1.2, deep CNN features extract higher-level data resulting in higher generalisation capabilities. Furthermore, these networks are more adaptable for applications across various domains in comparison to traditional methods. Additionally, DL-based detectors combine the three stages (stages two to four) of object detection in a single model.

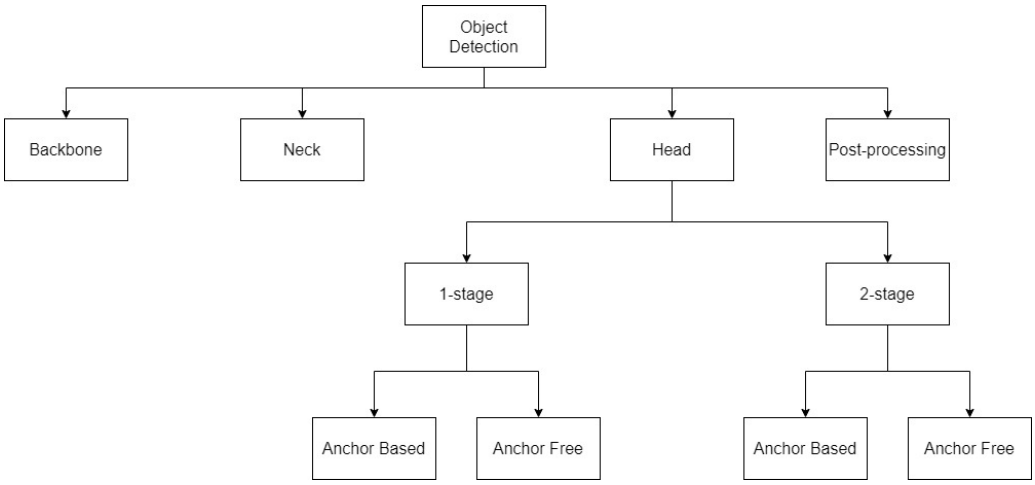


Figure 2.5: Components of an object detection network.

A DL-based object detector contains four main components as shown in Figure 2.5. The backbone of a detector is a pretrained classification network. Certain layers (depending on the detector) and the last FC layers of these networks are modified for object detection. The neck of the object detector refers to additional layers or networks introduced to improve overall performance. Head of the object detection network represents crucial components responsible for generating the final output of the detector. Finally, all object detection networks utilise a post-processing mechanism to remove redundant detections.

Depending on the number of stages in 'head', object detection networks can be classified as 1-stage or 2-stage as shown in Figure 2.6. 1-stage detectors generate output detections in a single forward pass of the image through the entire network. On the other hand, 2-stage detectors pass the input image through two stages of the network; first stage acts as a coarse detector and the second stage filters through the output of the first stage acting as finer detector. 1-stage detectors are generally computationally faster than 2-stage detectors due to smaller network size. However, 1-stage detectors typically have poorer localisation capabilities than 2-stage detectors, especially for smaller and unusually shaped objects. Popular 1-stage detectors include SSD [43] and YOLO family of detectors [44, 45, 46, 47, 48]. Popular 2-stage detectors include the RCNN family of detectors [49, 50, 51, 52].

YOLO [44], a 1-stage detector, consists of 24 convolution layers, followed by 2 FC layers. For detection, the image is first divided into a number of $S \times S$ grids [44]. After processing through the

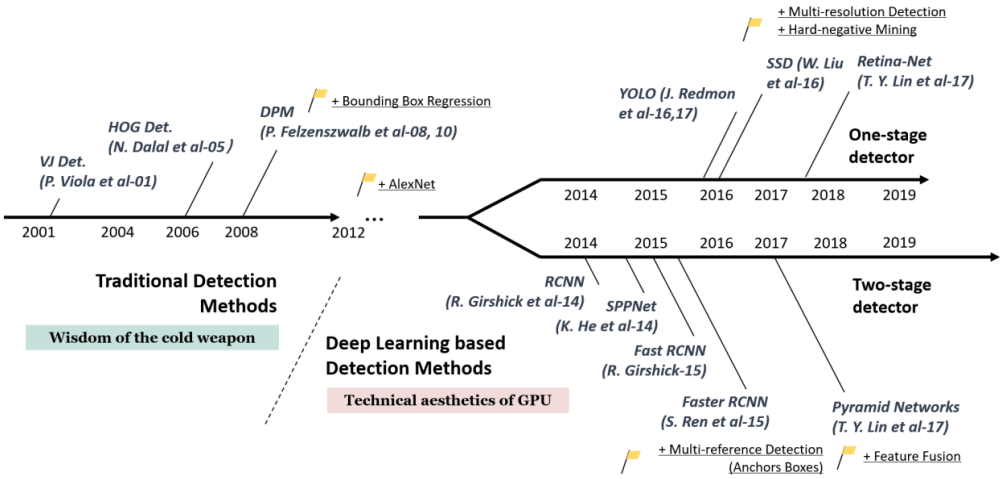


Figure 2.6: Object detection timeline [2].

network, following two outputs are generated for each grid - classification scores and bounding boxes. If the centre of the object is suspected to fall in a grid, a high classification score is assigned to the class of the potential object. On the other hand, if the grid is suspected to cover the background region, a high classification score is assigned to the background class. Furthermore, for each grid that potentially contains an object, two bounding boxes are generated in order to capture the object of each class. A post-processing mechanism called Non-Maximal Suppression (NMS) [53] is applied to remove redundant boxes and the final output detections are generated. Thus, generation of bounding boxes and their classification is performed in a single stage. Various updates have since been developed to overcome these drawbacks. These updates are described in Chapter 3.

Single Shot MultiBox Detector [43] is another popular 1-stage detector that surpasses YOLO network in its ability to localise objects, especially ones with small dimensions. SSD uses pretrained VGG16 classification model as its backbone. Feature maps output by each convolutional layer of the VGG16 model are divided into a specific number of regions referred to as *locations*. The number of locations varies for feature maps of each convolutional layer. An important distinction between SSD and YOLO is the use of predefined bounding boxes commonly referred to as reference or anchor boxes. In particular, SSD uses 4 anchor boxes in each location. Aspect ratio and size of these boxes varies with the feature maps used.

For larger objects, features maps of deeper layers prove useful whereas for smaller objects, feature

maps of initial layers prove useful. Anchor boxes for each layer were designed to take advantage of this characteristic. These anchor boxes are plotted on each location and during model training/testing. The bounding box regression branch computes the transformation required to ensure the anchor box tightly fits a potential object. These transformed bounding boxes are then passed through a classifier which assigns classification scores for each class. These boxes are then processed through NMS to remove redundant boxes and output final detections of the model. Two versions of SSD were developed depending on the input image size, namely, SSD500 (image size of 500×500) and SSD300 (image size of 300×300). Compared to YOLO, SSD has better localisation capabilities and was also experimentally proven better for detection of small objects. Furthermore, SSD is also computationally faster than YOLO.

It is important to note here that a large majority of object detectors are developed for natural (non-medical) images, including YOLO and SSD. This is due to the publicly available large datasets of natural images. For example, some of the commonly used, publicly available datasets of natural images include ImageNet [54], PASCAL VOC [55] and MS-COCO [56]. These datasets contain millions of images, with thousands in a large variety of classes. However, similar large datasets of medical images are not publicly available. Furthermore, collection of medical images and labelling of the objects such lesions by experienced professionals is a tedious and time-consuming process. Therefore, the development of detectors for medical images is comparatively slower. For this reason as well as adaptability of DL networks, it is common practice to modify object detectors developed for natural images for object detection in medical images. Specifically in breast lesion detection in US images, FRCNN is commonly adapted. Therefore, FRCNN is described in further detail in the following section.

2.3.1 FRCNN

FRCNN is a 2-stage detector. The architecture of FRCNN is shown in Figure 2.7. VGG16[57], pretrained for classification of natural images (ImageNet dataset) is used as the backbone network to extract features of the input image. Extracted features are then input to the first stage of this detector which is referred to as Region Proposal Network (RPN). RPN generates the first set of (coarse) detections, referred to as proposals. After removing redundant proposals using NMS, features of the remaining proposals are passed through to the second stage of the detector, referred to as base net-

work, for further processing. Output of the base network is then processed through NMS to remove redundant boxes and remaining proposals are output as the final detections of the model.

As VGG16 is a classification network, it undergoes the following two modifications to adapt it for object detection: 1. ROI pooling layer inserted after the last convolution layer and 2. Replacement of classification layer with two new branches (classification branch and bounding box regression branch). The FC layers of the VGG16 network require a constant input size of $7 \times 7 \times 512$. ROI pooling layer is used to ensure this requirement is met. The ROI pooling layer has two inputs. First is the feature map from the last convolution layer of VGG16 and second is the coordinates of proposals generated by the RPN. Using the proposal coordinates, features of each proposal are extracted from the feature map. These features then undergo a 7×7 maxpooling operation. Therefore, irrespective of the proposal size, the output of the ROI pooling layer is maintained at $7 \times 7 \times 512$. Secondly, the final FC layers of the VGG16 network are replaced in the base network. The remainder of this section provides further details of both stages of the FRCNN network.

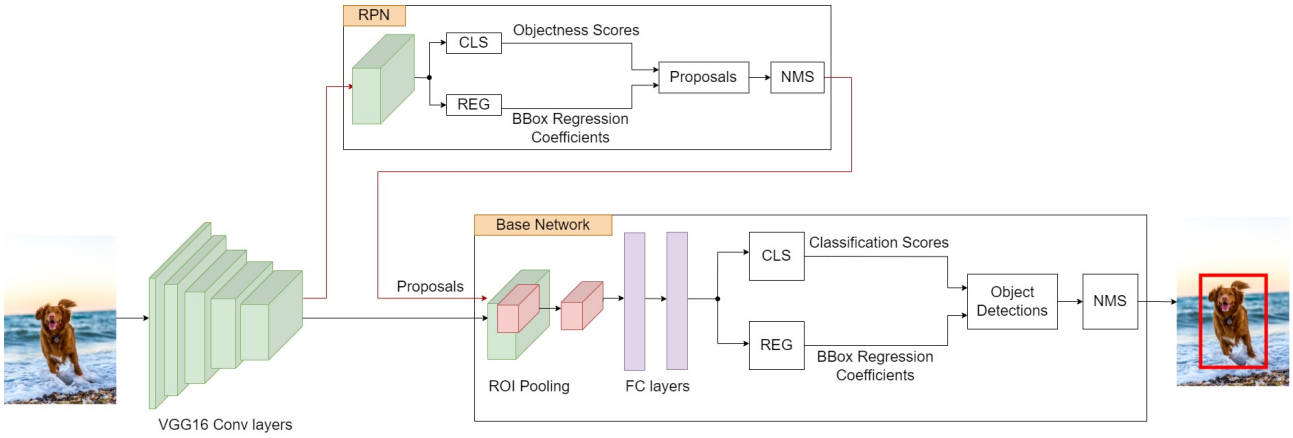


Figure 2.7: Faster R-CNN architecture using VGG16 as backbone.

2.3.1.1 Region Proposal Network (RPN)

RPN, first stage of the FRCNN network, is a shallow, fully convolutional network which uses the output of the last convolutional layer of the VGG16 model as its input to generate proposals. RPN uses predefined anchor boxes to generate proposals. The size and aspect ratios of the anchor boxes are selected so as to accommodate different object shapes and sizes. In this case, 9 anchor boxes of

scales $\{8, 16, 32\}$ and aspect ratios $\{1:1, 1:2, 2:1\}$ are used. These 9 anchor boxes are plotted across each spatial location on the input feature map. If the input feature map is size $W \times H$, then a total of $9 \times W \times H$ anchor boxes are plotted.

In the RPN, the input feature map is processed through a 3×3 convolution layer. After this, the network splits into two branches: classification branch and bounding box regression branch. Classification branch is responsible for identifying the likelihood of an object being present in each anchor box. Each anchor box is classified as foreground or background. Classification scores assigned by the RPN are also referred to as objectness scores. Given that 9 anchor boxes are used and classification scores assigned to two object classes (foreground and background), the output of the classification branch has a dimension of $W \times H \times 18$. On the other hand, the bounding box regression branch is responsible for identifying the transformation to be applied on each anchor box for it to tightly bound a potential object. Every anchor box is described using 4 coordinates (x, y, w, h) where (x, y) are coordinates of the centre point and (w, h) are the width and height of the anchor box. Thus, output of the bounding box regression branch is the changes to be applied to each of these 4 coordinates for every anchor box to tightly fit a potential object. Therefore, with 9 anchor boxes, output of the bounding box regression branch has a $W \times H \times 36$ dimension. Both classification and regression branches consist of 1×1 convolution layers. Thus, owing to the fully-convolutional nature for the RPN network and the use of the ROI pooling layer, FRCNN network does not require a predefined dimension of the input image.

For an image of size 600×1000, the feature map output by the last convolution layer of VGG16 has 38×63 dimension. As anchor boxes are plotted on every spatial location of this feature map, a total of 21546 anchor boxes are generated. Anchor boxes that are cross-boundary are eliminated leaving around 6000 boxes. These 6000 boxes are processed through the RPN to generate proposals. NMS is applied to remove redundant proposals. NMS consists of two main steps. In the first step, proposals with objectness score below the predefined threshold of 0.3 are discarded. In the next step, the highest scoring proposal is selected and moved to the final output of the network and all proposals with 70% or higher overlap with this proposal are removed. The same process is applied to all remaining proposals till every proposal is either discarded or moved to the final output. After NMS is applied, only 2000 top-scoring proposals are sent through to the base network during model training. During model testing, only the top 300 proposals are used.

RPN training samples are selected from the generated anchor boxes. Anchor boxes with overlap of 70% or higher with ground truth box(es) are considered as positive training samples and those with overlap of 30% or lower are considered as negative training samples. Remaining anchor boxes are not used for training. A minibatch of size 256 with a $1:1$ ratio of positive and negative samples is used. In the event of lack of positive training samples, the minibatch is padded with additional negative samples so as to ensure constant size. Training loss of each minibatch is defined as in Equation 2.5.

$$L_{rpn}(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i*) + \lambda \frac{1}{N_{reg}} \sum_i P_i^* L_{reg}(t_i, t_i^*) \tag{2.5}$$

where p_i is the output objectness score for the i_{th} anchor box in the minibatch and t_i is output transformation of the four coordinates for the same anchor box. p_i^* is the ground truth label which is set to 1 for positive training samples and 0 for negative training samples. Ground truth transformation is represented by t_i^* which is computed only for positive samples. N_{cls} and N_{reg} are the number of training samples in each minibatch for classification and bounding box regression task, respectively. Since N_{cls} is larger than N_{reg}, λ is set at 10 to ensure proportional weights are assigned to both losses. L_{cls} is log loss and L_{reg} smooth L1 loss.

2.3.1.2 Base Network

Base network is the second stage of the FRCNN network. Proposals generated by the RPN and processed through the ROI pooling layer are passed through to the base network for further processing. As previously mentioned, the final FC layer of the VGG16 network is replaced with a classification branch and bounding box regression branch of the base network. Both these branches consist of FC layers. Classification branch is responsible for classifying every proposal into one of the object classes or as background and the bounding box regression branch computes the offsets required to ensure that the proposals tightly bound potential objects.

Base network is trained using RPN generated proposals. Compared to RPN, the base network's training samples are selected with a broader overlap range. Specifically, proposals with an overlap of 50% or higher with the ground truth box(es) are considered as positive samples whereas those

with 10% to 50% overlap are considered as negative samples. Proposals with no overlap with the ground truth are used for hard-negative mining. A minibatch is formed using 128 samples of which 25% are positive and remaining are negative. Training loss L_{base} for every minibatch is the same as L_{rpn} defined in Equation 2.5. However, in L_{base}, λ is set to 1. The FRCNN model can be trained in multiple ways. The most efficient mechanism is the end-to-end training where both RPN and base network are trained simultaneously. Loss of the whole model is the combined loss of RPN and base network.

2.4 Summary

This chapter provided an insight into the medical and computational background relevant to this research. In particular, an overview of breast cancer, types of lesions and need for early detection was highlighted. The current lack of radiologists and the usefulness of automating lesion detection in addressing this issue was also discussed. After this, relevant computational concepts were detailed. Specifically, popular methods of feature extraction, dimensionality reduction and ML techniques utilising the extracted features for various applications were presented. Furthermore, an overview of object detection methods along with detailed explanation of important detectors was provided. Therefore, this chapter presented the fundamental concepts of this research. In the following chapter, state-of-the-art object detection methods for breast lesion detection as well as FP reduction techniques are presented.

Chapter 3

Literature Review

This chapter presents an extensive review of the existing research in the field of object detection in natural images with greater focus on breast lesion detection in 2D US images and lesion detection in images generated from a range of modalities. In each section, along with a thorough review of the existing detection methods, we also highlight the various methods of false positive (FP) reduction employed in these domains. The aim of this chapter is to provide a comprehensive exploration of this research field and highlight the advancements and limitations in the context of FP reduction techniques in each domain.

Section 3.1 details object detection methods developed for natural images along with details on techniques developed used to reduce FPs in these networks. Following this, Section 3.2 presents breast lesion detection methods for 2D US images which includes methods that adapt detection networks developed for natural images as well as novel methods developed for high performance and FP reduction. Section 3.3 provides a review of methods developed for detection of a variety of lesions in images generated from various other modalities along with FP reduction methods. This chapter is summarised in Section 3.4.

3.1 Object Detection in Natural Images

Object detection in natural images is a well-established field of research. Traditional methods of object detection rely on handcrafted methods for feature extraction. As described in Chapter 2, these methods have a common drawback of low generalisation capabilities due to the extraction of low-

level features. DL networks overcome the drawbacks of the handcrafted methods through extraction of features with higher levels of abstraction which results in the higher generalisation capabilities of these networks. Due to their high generalisation capabilities, DL networks are commonly used for real-life applications. Compared to handcrafted methods based models, DL networks are easier to adapt to different datasets and domains. In 2012, a classification method called AlexNet [29] revolutionised the field by successful utilisation of deep CNN to achieve remarkable performance. This network inspired the subsequent advancement in this field. Since 2012, DL networks have been used not only for image classification but in various fields of image analysis including but not limited to object detection, image segmentation, image recognition, etc. In the remainder of this section, the current state-of-the-art object detections methods developed for natural images are presented after which methods developed to refine these detectors including FP reduction are detailed.

3.1.1 DL-Based Detectors

Object detection networks consist of the following four parts: backbone, head, neck and post-processing mechanism. The function of each part is described in Section 2.3 in Chapter 2. Three publicly available datasets commonly used for development of DL networks for object detection in natural images are ImageNet, PASCAL VOC and MS-COCO. ImageNet is a much larger dataset than the PASCAL VOC and MS-COCO. However, ImageNet does have exhaustive annotations of the objects. Thus, it is common practice to first train the backbone network for classification using ImageNet dataset. Through transfer learning, this trained network is then adapted for object detection through transfer learning using PASCAL VOC or MS-COCO dataset.

Depending on the number of stages in the *head*, object detection networks can be categorised as 1-stage or 2-stage detectors. Some of the popular 1-stage detectors in natural images are YOLO family of detectors [44, 45, 46, 47, 48], SSD [43] and reinforcement based detectors such as [58, 5, 59, 60]. The popular 2-stage detectors are the R-CNN family [49, 50, 51] and reinforcement learning detectors [61, 62]. 1-stage detectors have an advantage of higher speed due to lesser computation cost. However, they share a major limitation in its localisation abilities and overall generalisation, especially of smaller objects and unusually shaped objects. On the other hand, 2-stage detectors require higher computation but have better generalisation along with better detection of small, unusually shaped objects due to the two stages of detection. Detailed architecture of YOLO, SSD and FRCNN is pro-

vided in Section 2.3 of Chapter 2.

Object detection methods developed before R-CNN [49] were mostly based on handcrafted features such as SIFT [25] and HOG [22]. These methods follow an exhaustive search and are not robust due to use of low-level features. R-CNN generates region proposals (regions of the image that potentially contains an object) using selective search and AlexNet for feature extraction as well as classification and bounding box regression. By combining traditional and DL methods, R-CNN outperforms its predecessors by a large margin. R-CNN made two important contributions to this field of research. First, R-CNN demonstrated a successful adaptation of a network trained for classification (AlexNet) for the task of detection. Secondly, their method of fine-tuning a deep CNN network, originally trained on a large dataset using a smaller dataset, paved the road for applications where the dataset available is considerably small. The network is easily expandable by using different region proposal generation methods and/or use of a deeper feature extraction CNN.

However, R-CNN has important drawbacks. Firstly, training R-CNN is a multi-stage process; first the region proposal generation method is tested, then the deep CNN is trained and fine-tuned, and finally linear SVM is trained for each object class. Bounding box regression is trained separately. This makes training the R-CNN model a very time-consuming process. Secondly, for every image, around 2000 proposals are generated. Features of each proposal are individually extracted for further training. This makes the training process computationally expensive. Fast R-CNN [50] reduces computation time by using the ROI pooling layer for an efficient extraction of proposal features. However, both Fast R-CNN and R-CNN, are limited by the performance of the traditional method used for region proposal generation. Poor performance of this stage inevitably has a negative impact on the performance of the whole network. Furthermore, as the proposal generation stage is a traditional method which requires manual adaptation, it cannot be automatically improved during the training of the CNN stages.

FRCNN [51] addresses the common drawbacks of R-CNN and Fast R-CNN. In FRCNN, a trainable, DL based method is used for region proposal generation. This trainable network is called Region Proposal Network or RPN. Further details of FRCNN architecture and training process is provided in Section 2.3.1 in Chapter 2. Introduction of a trainable head (RPN) improved mAP to 78.9% in comparison to Fast R-CNN. Additionally, time required for testing an image was reduced by 250

times in FRCNN in comparison to R-CNN and 25 times in comparison to Fast R-CNN. A relatively recent development in this field is the use of Network Architecture Search (NAS) [63]. NAS is a method to automatically develop CNN architectures based on predefined conditions for the provided dataset. In the field of object detection, NAS is typically used to design the backbone network. One such network is EffecientDet [64] that uses EfficientNet [65] which was developed using NAS as its backbone. However, NAS is outside the scope of this research.

3.1.2 Refinement of DL-Based Detectors

Apart from development of novel object detection networks, several methods have been proposed that focus solely on the refinement of existing detectors. Specifically, these methods concentrate on addressing drawbacks of one or more components of the object detection network in order to improve the overall performance of the detector. Remainder of this section explores these methods in further detail. Novel classification networks commonly investigate their use as the backbone in existing detectors. For instance, He et al. [30] evaluate their proposed ResNet classification networks as the backbone of FRCNN and show higher overall performance of this FRCNN model in comparison to the FRCNN model using other classification networks such as VGG16. Methods such as Spatial Pyramid Pooling network (SPPNet) [66] and Feature Pyramid Network (FPN) [67] improve proposal generation in the neck of existing detectors by introducing additional layers that improve the quality of features extracted.

SPPNet uses an SPP layer to improve overall detection performance. This layer is inserted after the last convolution layer of the backbone network. SPP layer is essentially maxpooling operation on the input feature maps individually using a range of window sizes. Output of this layer is sent to the FC layers. Thus, use of this layer allows the detector to accept input images of varying sizes. R-CNN with SPP layer had mAP of 59.2% which was 0.7% higher than R-CNN without SPP layer. Also, this layer processed all 2000 proposals simultaneously which improved overall speed by 24 to 64 times depending on the number of maxpooling levels used. On the other hand, the FPN network concatenates feature maps of several convolutional layers of the backbone network in a 'bottom-up fashion'. Feature maps from each concatenation level are individually used for proposal generation which is unlike the traditional method of using a single feature map from the last convolution layer of the backbone network. Thus, by combining low-level features with higher-level abstract features, the

FPN network improves detection of objects of varying sizes. FPN outperforms other methods such as image pyramid [68] where the same image is input in different scales or feature pyramid [67] where the feature maps from different layers are used directly, without the concatenation with feature maps from other layers. Using FPN in FRCNN (with ResNet50 as the backbone) improved mAP of the FRCNN model by 3.8 points from 53.1% to 56.9% in MS-COCO dataset.

Apart from the introduction of additional neck architecture, several methods have been developed for improving overall performance of the head itself. This is done in one of two ways; change in the architecture of the head or refinement of anchor boxes. Anchor boxes are predefined boxes used in the head of a detector. They are described in further detail in Section 2.3 of Chapter 2. Methods such as AttractioNet [69] and Cascade RPN [70] improve performance through architectural changes. AttractioNet uses multiple stages of heads, each trained to improve the detection of the previous stage. This head architecture is referred to as 'iterative RPN' [69]. The use of this iterative process improved the quality of region proposals which resulted in higher detection performance. On the other hand, this iterative process can lead to the model overfitting the training set. Furthermore, using multiple iterations of the RPN leads to misalignment of anchor boxes which results in lower quality of output detection. A potential improvement strategy is to use deformable convolution [71].

Cascade RPN addresses the issues of the iterative RPN improving the quality and connections of the multiple RPN stages used. In essence, it replaces the iterative flow of RPN with a cascaded one. With only a single anchor box at each location, the first stage of Cascade RPN consists of a dilated convolution layer on which the predefined anchor boxes are plotted and regressed. Feature maps from the dilated convolution layer along with the regressed anchor boxes are then sent to the second stage. Second stage consists of an adaptive convolution layer, a novel convolution method introduced in this work, followed by classification and further bounding box regression. Cascade RPN used in FRCNN led to mAP of 40.6% as opposed to FRCNN using traditional RPN which has a mAP of 36.9%. This method also outperformed other iterative RPN networks.

Another common method of detector head improvement is through an improvement of anchor boxes. Choice of anchor boxes plays an important role in overall performance of the detector. Instead of manual computation of anchor boxes as per the dataset and domain, methods have been

developed to either improve predefined anchor boxes or automatically estimate anchor boxes during detector training. Zhong et al. [72] improve anchor boxes during training by introducing a new branch attached to the bounding box regression branch which learns the amount of change to be made to the width and height of the predefined anchor boxes. FRCNN with mAP 76.4 was outperformed by this method with a mAP of 80.69. Other methods such as [73] and [74] develop bags of anchor boxes during training and an optimal bag is selected depending on the input image. However, these methods require an initial predefined set of anchor boxes. A recent method called GARPN (Guided Anchoring RPN) [75] automatically estimates anchor boxes from scratch.

GARPN is a shallow network that replaces traditional RPN. It estimates anchor boxes for every location. These estimated anchor boxes and the adapted feature map are then processed in the same manner as traditional RPN (described in Section 2.3.1 of Chapter 2). GARPN consists of two branches; location prediction and shape prediction. Output of the location prediction branch is a segmentation map. A score threshold of 0.1 defines active regions in this map. Active regions represent areas on the image which are most likely to contain an object. To avoid using inactive regions, all further convolutions in the detector are replaced with masked convolution layers. Shape prediction branch predicts height and width of potential object in every location. Only one anchor box is predicted on every location of the input feature map. Owing the large variation in size of the anchor boxes, deformable convolution [71] is used to adapt the input feature map. Only anchor boxes in the active regions are considered. GARPN improves mAP of FRCNN by 2.7 points in COCO 2017 dataset. Furthermore, when a pretrained FRCNN model was fine-tuned using GARPN proposals, its performance improved by 2.3 points mAP proving the higher quality of proposals generated by this method. Recent developments have been made in detecting objects without anchor boxes [76, 77, 78, 79, 80, 81]. In general, this is an emerging field and these methods have a common drawback of inability to achieve high performance in complex scenarios.

A common post-processing mechanism used in a significant proportion of existing detectors is Non-Maximal Suppression (NMS) [53]. Further details of this method is provided in Section 2.3.1 in Chapter 2. Several works such as SoftNMS [82] and MaxPool NMS [83] have been proposed to improve the performance of NMS. Unlike NMS, SoftNMS avoids discarding low scoring boxes in a single thresholding operation. Here, after the highest scoring box is selected, scores of boxes with high

overlap with this box are reduced in proportion to the overlap. So, a box that is distant from the highest scoring box would have a lower reduction in its score whereas one that is closer would have a higher reduction in its score. After this process is performed for all boxes, the lowest scoring boxes are discarded. MaxPooling NMS approaches this issue from a different perspective. Here, output of the classification branch undergoes a maxpooling operation in order to retain only those areas of the output that have higher probability of containing an image. This is especially helpful in reducing FPs. To compute overlap between boxes in the post-processing mechanism of the detector, Intersection-over-Union (defined in Section 4.2 in Chapter 4) is commonly used. To improve this metric, works such as [84, 85, 85] have been proposed.

In general, detectors have an imbalance in the number of positive and negative training samples. Due to the excessive amount of negative samples, the probability of the selection of hard negative samples is low. This is addressed using hard negative mining where such hard negative samples are specifically chosen to improve classification accuracy. Online Hard Negative Mining (OHEM) [86] and its successor Stratified OHEM (S-OHEM) [87] are popular methods used for this task where hard negative samples are chosen based on the confidence score loss. Negative samples with the largest loss are used for training. Methods such as Libra-RCNN [88] use IOU for selecting training samples, instead of training loss like traditional hard sample mining techniques. Thus, a balanced set of training samples are selected including hard negative samples. This method removes the need for classifying all training losses first to generate confidence score loss for identification of hard samples.

Various training losses have also been proposed for improving class imbalance by utilisation of weights in the loss or ranking samples. Focal loss [89] increases weights of hard negative samples in the loss function which enforces updates for such samples and has proportional reduction in weights of easier negative samples. Thus, hard negative samples have a larger impact on training. PISA [90], CARL [90], DR loss [91] and AP loss [92] uses ranking of samples to address class imbalance. One of the recent methods propose PISA and CARL loss [90] to improve classification accuracy through linking classification and bounding box regression branches of detection networks. Finally, in some works, search space is reduced to increase probability of using only hard negative samples for training. Recent methods such as Single-shot Refinement Network [93] and Enriched Refinement Network [94] filter out negative anchor boxes to increase probability of using hard negative samples during training.

3.2 Breast Lesion Detection in Ultrasound Images

Breast lesion detection in US images is a field that has been evolving for decades. In this section, we first present the traditional methods used for breast lesion detection in US images after which DL-based object detection methods are discussed. In both sections, techniques used for FP reduction are also highlighted.

3.2.1 Traditional Methods

Traditional method of breast lesion detection in US images involves four stages, namely, pre-processing, proposal generation, feature extraction and classification. In the pre-processing stage, the quality of the US image is improved which includes removal of speckle noise common in US images. This is commonly achieved using methods such as contrast-enhancement, histogram equalisation, filtering, etc. After pre-processing, proposals are generated using a variety of techniques such as using segmentation. In the next stage, features of these regions are extracted and used for their classification in the final stage. Given the higher relevance of the last two stages (feature extraction and classification) to this work, they are discussed in further depth. Many works have been proposed that focus purely on these two stages. Therefore, these methods are evaluated as a classifier (not as a detector) using popular evaluation methods that include area under ROC (receiver operating curve) curve and/or accuracy, usually accompanied with sensitivity and specificity.

Types of Handcrafted Features

Handcrafted features are selected for extraction of features characteristic of breast lesions in US images. These features can be classified into two main types: morphological and textural. Morphological features provide information relating to the shape and contour of the lesion [95]. Some of the common morphological features include lesion solidity, convexity, compactness, elongation, form factor, roundness, area. Aspect ratio, sometimes as referred to as orientation or depth-to-width ratio (DWR), is one the most common morphological features. Some morphological features such as computation of roundness, convexity, elongation require segmentation of the lesion for their extraction. Textural features provide information on the underlying pattern between pixels in an image. These are usually first order or second order features. First order features provide an overview of the image. Some examples include contrast, entropy, homogeneity, etc. of the whole image. Second-order features provide

information regarding the relationship between pairs of pixels in an image. Commonly used methods to extract textural features include GLCM, HOG, LBP and ULBP. Theoretical background and the extraction process of these methods have been described in Section 2.2.1 of Chapter 2.

Classification using a single feature type

Several traditional methods rely solely on one type of feature (either morphological or textural) for description of the lesion and its subsequent classification. For example, Chang et al. [96] use only morphological features (form factor, roundness, aspect ratio, convexity, solidity and extent) were used. Using these features, 210 images were classified by SVM with an accuracy of 90.95%. Sehgal et al. [97] uses only margin features such as margin sharpness, echogenicity, etc. along with age of the patient to classify lesions as benign or malignant which achieved an AUC of 0.87 on a dataset of 58 images. Although these methods have shown high reliable performance, the dataset is too small to appreciate their model's generalisation capabilities.

Likewise, following methods only use textural features. Gomez et al. [98] use 22 features from 240 GLCM matrices to describe the texture of lesions. MRMR was used for reducing the dimension of the final feature vector to 17×1. With the help of MRMR, contrast and correlation were found to provide the most important, discriminating features of the lesion. Their model achieved a high AUC of 0.87 over 436 images showing that use of GLCM alone can help classify lesions as benign or malignant. The model proposed by Chen et al. [99] uses autocorrelation, a second-order textural feature, to characterise the lesions. On a large dataset of 1020 images, this method had a high classification accuracy of 96.47%, outperforming three experienced radiologists.

Here, the difference between radiologists and the model was highest in precision caused due to difference in number of FPs. Radiologists had an average of 43.13% precision whereas the model had 81.4%. Furthermore, 2 out of the 3 radiologists had 6 FNs whereas the model and one of the radiologists missed only 1 lesion. Abdel et al. [100] studied the impact of five different textural features individually, namely, GLCM, ULBP, phase congruency-based LBP (PCLBP), HOG and pattern lacunarity spectrum (PLS). It was found that HOG had the highest AUC of 0.989 on a dataset of 59 images. This was followed by LBP and PCLBP with AUC 0.95 and 0.923, respectively.

Methods such as [101, 101, 102, 103, 104, 105] use a multiple textural features for higher classification accuracy. For example, Wei et al. [101] use GLCM, HOG and LBP for classification of 600 images. Individually, HOG had highest precision of 79.63%, GLCM had the highest sensitivity of 82.61% and LBP had the highest F1-score of 0.8. However, the combination of all these features outperformed the individual features with highest F-score of 0.839 along with highest sensitivity, specificity, and precision. Benaouali et al. [102] use HOG and LBP to classify 780 images. Here too the combination of the features had highest accuracy of 96% as well as higher sensitivity and specificity than individual features.

Classification using a combination of textural and morphological features

With the increased complexity in the US images, it has become more popular to use a combination of both morphological and textural features as a variety of lesion characteristics are captured in the combined feature vector which in turn improves the overall performance of the model. For instance, Alvarenga et al. [106] use GLCM and complexity curve features to classify 152 images. Various combination of the combined feature vector was analysed, and the best performing feature vector, with an accuracy of 84.2%, contained GLCM's contrast, correlation, standard deviation, the contrast, and maximum value of transition of the internal region of the lesion (excluding the margin). Similarly, Wu et al. [107] use solidity (morphological feature) and autocovariance (textural feature) for classification of 210 images which achieved an accuracy of 92.86%.

Wei et al. [108] use shape related features such as compactness and textural features using GLCM, HOG and LBP. PCA was applied only on the textural feature vector due to its significantly larger dimension. The proposed model consisted of two separate classifiers, one for the morphological features (Naïve Bayes classifier) and another for textural features (SVM). The final classification score output was a weighted sum of both scores assigned by classifiers (90% weight assigned to the SVM). On a dataset of 448 images, this model had highest accuracy of 87.78% which is higher than that of three other works that used only a single feature and two other works that used a single classifier for both morphological and textural features.

Similarly, Wei et al. [109] also use GLCM, HOG and LBP as textural features and a different range of morphological features such as an ellipse's direct least square fitting. On a large dataset

of 1061 images, the combined features outperformed both the individual morphological and textural features with an accuracy of 87.32%. Other works such as [110, 111, 112, 113, 114, 115] use a range of textural features such as Tamura features, posterior acoustic attenuation, echo pattern, etc along with morphological features like aspect ratio, normalized residual value (NRV), Hu-moments, etc. In many of these works, the final feature vector has a large dimension which can adversely impact the performance. Commonly used dimension reduction methods include PCA [108, 116, 117] whereas others use feature selection such as stepwise logistic regression, RFE, etc [104, 106, 118]. The choice of classifier is popularly SVM [108, 109, 110, 101, 119, 107] and neural network [103, 120, 99, 105]. Some works compare various classifiers such as LDA, random forest, decision trees with SVM. Generally SVM outperforms majority of the evaluated classifiers [111, 114, 121, 102].

Deep CNN for feature extraction

Due to the drawbacks of handcrafted features and the advantages of DL methods as described in Section 2.2.1 of Chapter 2, DL networks have been used for feature extraction in multiple recent works. An important requirement for DL methods is a large, diverse training set. However, such datasets of US images are not publicly available, and their collection is quite a challenging and time-consuming process. This issue is addressed by employing transfer learning on models pre-trained on large (publicly available) datasets of natural images. One of the early works such as one by Huynh et al. [122] prove the effectiveness of this strategy. In particular, Huynh et al. [122] experimentally prove that the DL-based classifier trained using transfer learning outperformed handcrafted features based classifiers. Other works [123, 124] show the higher classification accuracy of the model after transfer learning compared to the model trained from scratch using breast lesion dataset. With these developments, use of DL networks for feature extraction gained popularity.

Han et al. [125] successfully utilise GoogleNet [126], a deep classification network developed and pretrained for object detection in natural images, for classification of breast lesions in US images. This work also showed that use of 180 margin pixels improved the classification accuracy of the network. Their final model has a high accuracy of 91.23% on a large dataset of 4254 benign and 3154 malignant lesions. Some methods further utilise the generalisation capabilities of the DL architectures to classify lesions in US images of different modes such as B-mode, Doppler mode, etc. For example, Bressem et al. [127] investigate ResNet18 (pretrained for classification of natural images) for classification of four

modes of US images. This model had a reliably high accuracy of 95.43%.

Combination of learnable and handcrafted features

Despite the drawbacks of handcrafted features, they provide critical information that might be lost in deep CNN networks especially for small lesions. Therefore, a combination of DL and handcrafted features are widely used to leverage advantages of both methods. For example, Peng et al. [128] utilise shape-related features (morphological) and DL features (textural) to classify breast lesions. It was found experimentally that use of this combined feature vector had higher accuracy than use of either handcrafted or DL features alone. Similarly, Antropova et al. [129] combine morphological and textural features related to size and shape of the lesion with DL features to provide high classification accuracy of 0.9. Similar performance improvement was also found in DCE-MRI and FFDM images.

3.2.2 Deep-Learning Based Detectors

Although these traditional methods described previously provide reliable accuracy, there is still a reliance on outside models or sources for the ROI. In recent years, DL models combine the three stages of traditional detection methods (proposal generation, feature extraction and classification) in a single model. Furthermore, DL based detectors eliminate the requirement for manual feature engineering that is required in traditional handcrafted methods. Compared to classification of breast US images, breast lesion detection is a less explored field due to the difficult and time-consuming process of annotation by experienced radiologists.

Nonetheless, important development has taken place in this field. One of the pioneering works by Yap et al. [130] investigated the impact of using three DL networks, namely Patch-based LeNet, UNet and FCN-AlexNet, and compared their performance to that of three traditional methods using handcrafted features. This investigation was conducted using two datasets, datasets A (generated in 2001) and dataset B (generated in 2012), containing 306 and 163 images each. Overall, DL-based methods outperformed the traditional methods. FCN-AlexNet had the best performance in various test settings while LeNet had the lowest number of FPs per images (FPI). A more recent work by Cao et al. [131] evaluated multiple 1-stage and 2-stage object detection networks that were originally designed for natural images for breast lesion detection in US images. The methods analysed include FRCNN with ZFNet and VGG16 as backbone, YOLO, YOLOv3, SSD300 and SSD500 with ZFNET

and VGG16 backbone. These models were trained on 860 images and tested on 183 images, all resized to 256×256. SSD300 with ZFNet had the best performance with an F-measure of 79.38%.

Use of ZFNet led to better detection of benign lesions whereas VGG16 was better suited for malignant lesions. As ZFNet is shallower than VGG16, the features extracted by this network are of comparatively lower levels of abstraction. Thus, the number of benign lesions detected by ZFNet is greater. However, it underfits the more challenging malignant lesions due to the lower-level features extracted. On the other hand, VGG16 extracts higher-level features which enables it to outperform ZFNet in detection of the malignant lesions but it underfits benign lesions leading to poor performance in the same. Therefore, due to higher number of benign lesions in the test set, SSD300 with ZFNET outperforms SSD300 with VGG16. Although the 1-stage detectors evaluated in this work are faster and have higher performance, their performance for detection of challenging lesions are lower. Furthermore, the dataset used in this work is small, limiting the generalizability of the results.

3.2.2.1 Adaptation of FRCNN

In recent years, FRCNN has been widely adapted for breast lesion detection in US images. These methods adapt the network either through modifying the network architecture or modelling hyper-parameters or both. In one such work, Zhang et al. [4] adapt FRCNN (pretrained VGG16 as the backbone) by modifying, both, the network architecture and modelling hyperparameters with the aim to improve the overall performance of the network for breast lesion detection in US images, focusing particularly on improving detection of small lesions. Also, the final output of the network classifies the detected lesion into one of three BI-RADS categories, namely category 2 (benign), category 3 (likely benign) and malignant (category 4 to 6), in order to further assist the radiologists in their diagnosis.

Their work was conducted on a large dataset of size 3103 images collected from a hospital and 150 images crawled from the internet using "Scrapy+Selenium+Phantomjs" framework [4]. After augmentation, the training set consisted of 6000 images and the test set consisted of 1200 images. The boundary black marker region around the scan image was removed in both training and test sets. Labelling was done so as to include 10% of the margin around the lesion in the GT box to utilise the critical information contained in this region. In terms of network architecture, they made two major changes. Firstly, convolution layers 2 and 4 of the base VGG16 network were fused so as to introduce

lower-level features such as edges from convolution layer 2 to the higher-level features of convolution layer 4. Secondly, an additional RPN was utilised. The additional RPN uses the feature maps from the fused convolution layer 4 while the standard RPN uses feature maps from convolution layer 5.

In terms of the modelling hyperparameters, test proposals were reduced to 200 in line with lower number of objects in an US image and aspect ratio of anchor boxes were modified to 1:1, 3:2, 2:1 in line with the elliptical/circular shape typical of a breast lesion. During model testing, every image was tested in its original size, zoomed at 0.6 and 1.5 and mirrored. The final output was the fusion of results of all 4 images using a voting mechanism. The proposed model outperformed original FRCNN in all categories of classes as well as overall performance. The proposed model had an overall mAP of 0.913 surpassing original FRCNN with mAP 0.861. However, the modified version had a slightly slower speed of 4.11 FPS compared to that of the original at 4.68 FPS.

When compared to YOLOv3, a 1-stage detector, their model outperformed by 0.071 overall mAP. But, YOLOv3 had a higher speed of 15.6 FPS. Compared to YOLOv3, detection of smaller lesions was 8.2% better in their modified version of FRCNN along with higher IOU of the output detections. This is evidence of the higher accuracy of 2-stage detectors in detection of challenging cases found in US images. However, this method has two main drawbacks. Their images were all high-quality (high resolution), including ones crawled from the internet. Secondly, these images were collected from a single hospital. Use of images from multiple hospitals helps analyse the generalisation capabilities of a network since these images would have significant differences.

In a similar, more recent work, Yap et al. [132] adapted FRCNN for breast lesion detection in US images through modifications of modelling hyperparameters. This work proposed a novel 2-tier transfer learning method which is suitable for small datasets. As the issue of small datasets is prevalent in this field, their proposed 2-tier transfer learning step is an important contribution. In this work, two datasets were used. Dataset A contained 306 images from 2001 whereas datasets B contained 163 images from 2012. Dataset B had relatively higher resolution images than dataset A. Inception-ResNet-v2 is used as the FRCNN backbone in this work. Following modelling hyperparameters were modified: anchor boxes updated to “64px, 128px and 256px” [132] and aspect ratio changed to 0.5,1, and 1.5 [132], number of proposals reduced to 100 and NMS score threshold increased to 0.9 from

the original value of 0.3. When trained and tested on the datasets individually, their proposed model had the highest accuracy of 0.8892 and 0.831 in datasets A and B, respectively, outperforming a FCN-AlexNet model. Compared to dataset A, all models had higher FPs and lower performance in the new dataset B. This along with the small size of both datasets point towards poor generalisation capability of the network.

3.2.2.2 Novel Methods

Apart from the adaptations of the FRCNN network, several works have proposed novel 2-stage detection networks for breast lesion detection in 2D US images. These methods were primarily designed for overall high detection performance. Additionally, they employ one of the following two common techniques to specifically ensure low FP detection. First strategy is the generation of a segmentation map that is used as reference for removal of FPs in the output of the detector. Second technique is the use of a separate network for further classification of the output of the detector.

One of such notable works in this area is the novel 2-stage detector proposed by Huang et al. [133] to detect breast lesions in 2D US images and classify them in one of five BI-RADS categories (3, 4A, 4B, 4C and 5). Their detector consists of two sub-networks, namely, ROI-CNN and G-CNN. The input image is first processed through the ROI-CNN, which is the first stage of their detection network. ROI-CNN outputs a segmentation map indicating potential lesion locations referred to as regions of interest (ROI). Based on the segmented map, the background regions of the input US image are masked, leaving only the potential lesion regions. This masked image is then sent to the second stage, which is the G-CNN network, for further classification of the potential lesion regions into one of the five BI-RADS categories or background. ROI-CNN is a fully convolutional network (FCN) based on VGG16 with new layers added to efficiently concatenate lower-layer features with those of the higher-layer whereas G-CNN is an 18-layer encoder network. They also use refinement mechanisms based on C-V level sets to ensure high quality of the output of ROI-CNN.

In this work, a large dataset of 2238 images, resized to 228x228, was utilised. Dataset augmentation was performed separately for the two stages of the network. The proposed model had an average accuracy of 0.934 in the other four categories whereas category 4B it dropped to 0.735. This drop in performance for lesions of 4B BI-RADS category was attributed to the relatively smaller number of

these lesions in the dataset as well as their challenging nature. The two main drawbacks here is the lack of comparison to other 2-stage detectors and no information on the computation time. Although this performance is reliably high, the use of two large networks might render this network computationally slower.

Tao et al. [134] address this issue of large network size by proposing a novel 1-stage model that combines the two stages (ROI generation and classification). In particular, both stages use a common feature extraction network. This model consists of three sub-networks, base-net, seg-net, and cls-net [134]. Base-net generates feature maps for the input US image which is used by seg-net to generate a segmentation output containing ROI. Output of the seg-net is sent through to this cls-net which classifies the ROI regions as benign or malignant. Feature maps from four layers of the base-net are used to create cls-net. The first feature map extracted from base-net is concatenated with the output of seg-net. Thus, only the ROI regions output by seg-net are processed in cls-net. Anchor boxes are used at five stages in cls-net. The size and aspect ratio of these anchor boxes varies with respect to the receptive field of that layer as well as the lesion size and aspect ratio. A total of 4753 anchor boxes were used.

This work uses a large dataset of 2280 images. Overall, their proposed network method outperformed SSD with highest F-measure of 90.78% and a small drop in computation speed. SSD had a detection time of 0.016s whereas their proposed model had a speed of 0.111s. Additionally, they experimentally prove the reduction in FPs due the use of cls-net. To address the lack of large annotated US datasets (where the GT lesion is provided by experience radiologists), numerous works propose the use semi-supervised learning where the model is trained with a combination of annotated and unannotated images [135, 136].

In summary, despite the advantage of lower computation time, 1-stage detectors have poorer performance than 2-stage detectors in detection of breast lesions in US images, especially challenging cases. FRCNN is a popular method used for this application. This network was originally designed for object detection in natural images. To adapt this network for breast lesion detection in US images, two main categories of changes are made. First, the network architecture is adapted to improve detection of breast lesions such as the introduction of a secondary RPN network. Second, the modelling

hyperparameters are adapted, most common being anchor boxes and number of test proposals. Most methods perform both types of modifications for improving overall performance as well as reduction of FPs. These modifications show reliable performance of the FRCNN network after these modifications, outperforming 1-stage detectors like YOLOv2 and SSD. However, they have two major drawbacks. Firstly, the experimental evaluation of the proposed modifications are not provided. Thus, the impact of individual modifications on the overall performance and/or FP reduction is unknown. Secondly, certain methods use small to medium sized datasets for their evaluation which raises concerns regarding the generalisation capabilities of these networks on large datasets, collected from different hospitals and US machines.

We address these drawbacks in Chapter 5 through a comprehensive evaluation of the modelling hyperparameters of the FRCNN network when used for breast lesion detection on our large dataset of US images collected from multiple hospitals in different countries. Based on this investigation, we design an adapted FRCNN model which outperforms the original FRCNN through considerable reduction in FPs. Both modified FRCNN methods as well as novel detectors developed specifically for breast lesion detection in US images do not specifically address the prevalent issue of FP detections. We address this gap through our novel U-Detect and U-DetectH methods proposed in Chapters 6 and 7, respectively.

3.3 Lesion Detection of Other Cancer Types on Various Modalities

The four stages of traditional methods of lesion detection were previously described in Section 3.2. This section focuses on the feature extraction and classification methods used for lesion detection in images generated by various modalities. As seen with methods developed for breast lesion detection, due to the increased complexity of images, it is common practice to extract a combination of morphological and textural features for lesion classification. For instance, Byra et al. [137] extract 22 morphological and textural features to classify breast lesions in mammograms using SVM. These features include 5 shape related features, 3 features measuring sharpness of the edge and 14 GLCM texture features. Performance of the SVM classifier was compared to that of novel strict two-surface proximal (S2SP) classifiers. Both these classifiers had a high classification accuracy of around 0.95 when trained with features selected using FLDA. Similarly Al-Dhabyani et al [124] use 20 handcrafted

features including 12 based on shape, 8 on wavelet local maxima were used to classify thyroid lesions using SVM. The SVM had higher classification accuracy of 0.96 in comparison to a probabilistic neural network which had a comparatively lower accuracy of 0.91. Several works adapt the standard feature extraction methods in order to improve classification accuracy. For example, Han et al. [138] successfully adapted GLCM for improved characterisation of breast lesions in mammograms. Use of their improved GLCM resulted in overall higher classification accuracy along with requiring less computation time .

As described in the Section 2.2.1 of Chapter 2, DL methods have gained popularity due to their advantages over handcrafted features. Han et al. [125] propose a CNN model to classify breast lesions. In their work, the higher classification accuracy of the DL based model in comparison to the traditional methods is experimentally proven. Zhu et al. [139] propose generic deep CNN models developed for classification of breast and thyroid lesions in US images. These networks were pretrained on natural images and transfer learning was performed to adapt the network for this domain. With the same parameters used for thyroid and breast classification networks, both models had high classification accuracy of 86.5% in the thyroid dataset and 89% in breast dataset. Furthermore, it was also experimentally shown that the use of the classification model trained for classifying thyroid lesions was also effective in the classifying breast lesions with high sensitivity and specificity. Their models also outperformed three radiologists.

Many works use a combination of handcrafted and DL features to combine their advantages and achieve high classification accuracy. For instance, Ciritis et al. [140] use HOG and LBP along with features extracted by a VGG-F model to classify thyroid lesions in US images. A novel method of feature voting was used for fusion of these features. Compared to models using features extracted from a single method, their model had the highest classification accuracy of 0.931 especially in challenging malignant lesions on a dataset of 1037 images. In another such method proposed by Zhuang et al. [141], handcrafted features relating to shape, size and texture of the breast lesion were utilised along with features extracted from VGG19. This method had high classification accuracy in detection of breast lesions in images from three modalities, US, DCE-MRI and FFDM.

In recent years, DL based detection networks that combine the three stages of proposal generation,

feature extraction and classification have become more popular owing to their better generalisation capabilities and robustness as described in Section 3.2.2. FRCNN is one of the popular methods used in detection of lesions in images from different modalities. Like with breast lesion detection in US images, here too the original FRCNN network is modified to adapt it for the lesion detection. For example, Zhuang et al. [142] use deformable convolution layers in the backbone network of the FR-CNN model to improve detection of different sizes and shapes of lesions. Additionally, to specifically improve detection of small lesions, a multi-scale FPN, designed using NAS, was also utilised. Use of NAS-FPN increased FP per image (FPI) in one dataset in comparison to that of the original FRCNN (from 0.0277 to 0.0327 in the best case). But there was an increase in true positive rate (TPR) and a notable improvement in detection of small lesions. An et al. use the FRCNN model with Inception v2 as the backbone network for detection of carotid plaque in US images. Several modelling hyper-parameters were also modified. This adapted FRCNN model had mAP of 58.62% and high average precision of 91.02%. The aim of this method was to be one of the pioneer research in this domain and provide fundamental knowledge for further development in this domain.

A recent work by Liu et al. [143] on thyroid lesion detection in ultrasound images, FRCNN was used with ResNet50 backbone along with FPN. Here, anchor boxes were adapted for each level of the FPN output in accordance with size and aspect ratio of the lesions in their dataset. Furthermore, to improve classification performance of the base network, they introduced a cost matrix applied to the classification loss of the base network. This cost matrix heavily penalised misclassification of benign and malignant lesions as background and had a smaller penalty for misclassification of benign lesions as malignant. These modifications improved mAP from 0.938 of original FRCNN with FPN to mAP of 0.947 of their proposed model. Additionally, they utilised a modified ZFNet for further classification of the final detections of the modified FRCNN model to further improve the classification of lesions. Their proposed modified ZFNet model outperformed 5 radiologists in sensitivity, specificity as well as accuracy. Similarly, FRCNN was modified for lesion tracking in US images by Igarashi et al. [144]. Here, the final output was taken from the RPN and the base network's classification and bounding box regression branches were removed. This was done to reduce computation time and provide overall stability in the detections between consecutive frames. Although these modifications reduced overall F-measure, localisation was improved along with reduction in instability between frames and computation time.

Li et al. [145] proposed a FRCNN model, modified for thyroid lesion detection in ultrasound images. Here, ZFNet is used as the base network. To improve detection of small lesions, feature maps of layers 3 and 5 were concatenated. The scaling ratio for this concatenation is learnt during training. The concatenated feature map is then input to RPN. Choice of the layers to concatenate was done through an extensive combination of up to 3 layers where the selected layers had the best performance. A 'spatially constrained layer' [66] was used to stabilise the results. Original FRCNN with ZFNet backbone had a TPR of 0.868 which was improved to 0.935 using their modifications. There was also a reduction in FPR from 0.289 to 0.185 as well as false negative rate (FNR) from 0.132 to 0.07.

Ribli et al. [146] adapt FRCNN for breast lesion detection in MRI scans without architectural changes. Here, the lower threshold for RPN's positive sample selection was dropped from 0.7 to 0.5. This was done due to the small size of lesions in MRI scans in comparison to the whole image. This threshold helped increase the number of positive samples otherwise missed with the original threshold. Also, the NMS threshold was dropped to 0.1 from 0.3 in the original FRCNN configuration. Their modified FRCNN model had a high AUC of 0.95. Similarly, Akselrod et al. [147] modified FRCNN for breast lesion detection in mammogram images. Modifications here involved resizing the input image to 4000×3000 and preprocessing this image to remove background and majority of the normal tissue. This preprocessed image is then divided into grids. FRCNN processes individual grids for lesion detection. Result is the concatenation of all grids to form the original (preprocessed) image. This method had an average precision of 0.72. Although it did not outperform the original FRCNN, their work adds valuable information in modification of this network.

Adaptation of Mask R-CNN has also gained popularity in this domain. An et al. [148] use Mask R-CNN with FPN and soft-NMS was used for detection of breast lesions in CBIS-DDSM dataset. The final mAP of this modified model of 0.66 was higher than that of the original Mask R-CNN model by 0.55. Similarly, Abdolali et al. [149] also modified Mask R-CNN for detection of thyroid lesions in US images. On the other hand, 1-stage methods, with or without modifications, were also evaluated for lesion detection. For instance, Xie et al. [150] use SSD for thyroid nodule detection in US images. On the other hand, Wu et al. [151] use an ensemble of 1-stage and 2-stage detectors for kidney abnormalities detection. Their detection model consisted of three parts. First, the US image is

selected in the first stage which is then passed to the second stage where the kidney is detected in the US image and the last stage classified the detected kidney as normal or abnormal. First stage is based on CNN and the second stage is an ensemble of three detectors (SSD, RetinaNet and RefineDet). The final output is an aggregation of the outputs of these three models. Final classification stage is a novel CNN proposed in this work. Their ensemble method had the highest TP of 98% and classification accuracy of 94.67% on a dataset of 3772 abdominal US images.

3.4 Summary

This chapter presented an extensive review of the state-of-the-art object detection methods developed for object detection in natural images as well as lesion detection in medical images with emphasis on breast lesion detection in US images. In addition, we highlighted the methods developed for FP reduction in all domains. In the review, the popularity of the FRCNN model for not only breast lesion detection in US images but also lesion detection in medical images was evident. We also highlight the following important gaps in the literature that are addressed in our work. Firstly, the methods that adapt FRCNN model for breast lesion detection in US images provide insufficient experimental evaluation of their modification and in some works, small to medium sized dataset is used. We address these drawbacks in Chapter 5 where we study the impact of several modelling hyperparameters on our large dataset of US images. The adapted FRCNN model designed through this investigation successfully outperforms the original FRCNN model due to significantly lower number of FP detections. Secondly, although novel methods of breast lesion detection in US images have been developed, no method specifically focuses on FP reduction which is a common and important issue in this domain. FP detections can result in unnecessary and painful checks including biopsy. Therefore, in Chapters 6 and 7, we propose novel U-Detect and U-DetectH methods to address this gap.

Chapter 4

Dataset and Experimental Setup

This chapter provides general preparation for the rest of the chapters in this thesis. The chapter details the datasets and the experimental setup used in this research. Section 4.1 describes the ultrasound images datasets used in this research including its collection, annotation and exploration. Section 4.2 details the experimental setup used for training and testing all models and presents the evaluation metrics used to measure or analyse the performance of all models.

4.1 Breast Ultrasound Images Datasets: Collection, Exploration and Annotation

A total of five datasets of 2D breast US images* have been collected and used in this study. For simplicity purposes, the datasets are named as dataset A, B, C, D, and E. Datasets A, B, D, and E were provided by our collaborator TenD AI Medical Technologies Ltd., Shanghai, China. Images in these datasets were collected from multiple hospitals and generated using various US machine makers including Siemens Oxana 2, Siemens S3000, Toshiba Apolio 500, GE Logic E9, and Philips Epic 7. All images as well as hospital names were anonymised by TenD AI Medical Technologies Ltd. The lesion delineation (lesion boundary points or Region-of-Interest (ROI)) were annotated by experienced radiologists of 10-20 years of experience manually with the assistance of a MATLAB software tool provided by Zhu et al. [139]. While detecting the type of the lesion (benign or malignant) is outside the scope of this study, a pathology report that confirms the histopathological assessment of tissue samples obtained via biopsy or surgery was used to confirm the nature of each lesion.

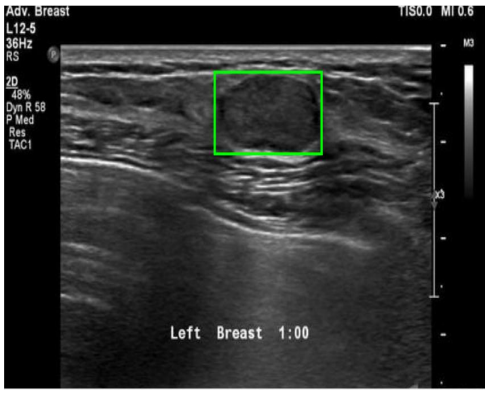
*The term *image* refers to breast ultrasound image unless specified otherwise.

ROI coordinates provided by the radiologists are used to construct ground truth (GT) boxes. First, from the ROI coordinates of a lesion, the minimum and maximum values of x and y-coordinates ($x_{min}, x_{max}, y_{min}, y_{max}$) are determined. These values represent the top-left corner (x_{min}, y_{min}) and bottom-right corner (x_{max}, y_{max}) of the GT box. Each lesion is then described using the x- and y-coordinates of the top-left corner, width ($x_{max} - x_{min} + 1$) and height ($y_{max} - y_{min} + 1$) of the GT box. Dataset C is a publicly available dataset [152]. ROI in this dataset was provided in the form of a mask image for each lesion. ROI coordinates were then extracted using this mask image. GT boxes were then constructed using these ROI coordinates using the aforementioned steps.

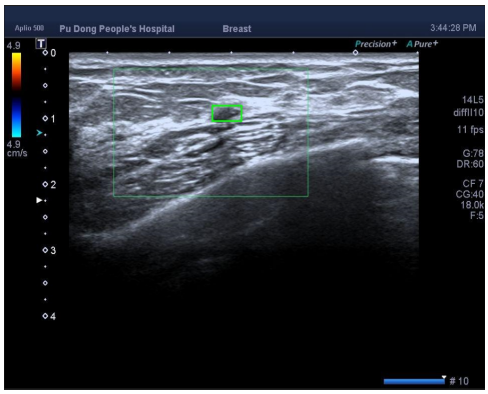
When generated, US images include a black boundary around the scanned region. This boundary region is referred to as black marker region and it contains information such as the settings of the US machine, region being scanned, date and time, etc. Except dataset C, images in all datasets contain the black marker region. It is worth mentioning that no preprocessing (e.g. image normalisation and filtering) was conducted on any of the datasets except for removal of corrupt images that contained artefacts occluding the lesion characteristics. To comply with the clinical research common practice, images from one hospital (or medical centre) was used for creating the detection models, and images independently sampled from other hospitals for external testing. Dataset A is used as modelling dataset while datasets B-E are used solely as external unseen test sets. Further detailed description of each dataset and its characteristics are provided in the remainder of this section.

Dataset A

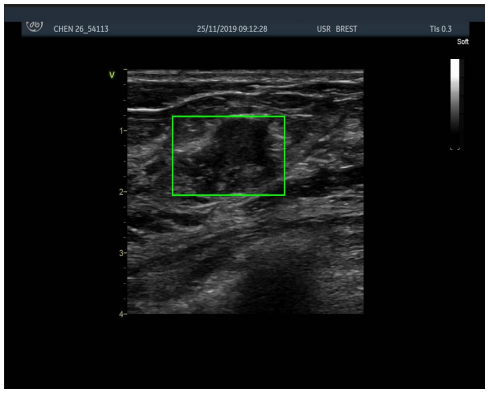
This dataset consists of 1733 images in total, comprising 1070 benign and 663 malignant cases, collected from two hospitals. All images in this dataset contain one lesion. Figure 4.1 shows samples of different lesions from this dataset. Figure 4.1a shows a simple benign lesion with clear boundary; Figure 4.1b shows a very small benign lesion, generally hard to detect due to its size; Figure 4.1c shows challenging benign lesions where the lesion boundary is unclear and irregular and its texture is closely similar to that of the background normal tissue; Figure 4.1d illustrates the taller-than-wide nature of malignant lesions; Figure 4.1e shows large malignant lesions with very unclear boundaries; and finally Figure 4.1f shows image sample that contains Doppler. Images with Colour Doppler were observed mostly in malignant lesions. Samples in Figure 4.1 show that dataset A contains a large variety of images and lesions. Therefore, it was selected as modelling dataset.



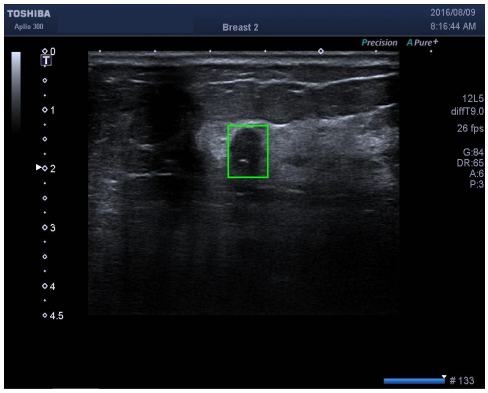
(a)



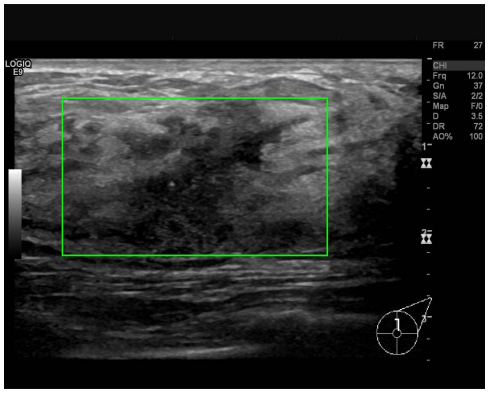
(b)



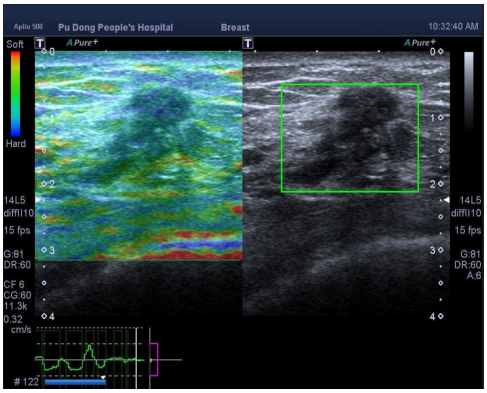
(c)



(d)



(e)



(f)

Figure 4.1: Sample images from dataset A (Green box: ground truth box encompassing the lesion).

Dataset B

Dataset B contains 150 images (100 benign and 50 malignant cases), collected from one of the same hospitals as dataset A. Every image contains one lesion. Figure 4.2 shows sample benign and malignant images cases from this dataset. Figure 4.2a and 4.2b show simple benign and malignant lesions from this dataset; Figure 4.2c shows a challenging benign case with texture similar to that of the background normal tissue and Figure 4.2d shows a challenging malignant case where the boundary of the lesion is unclear and irregular.

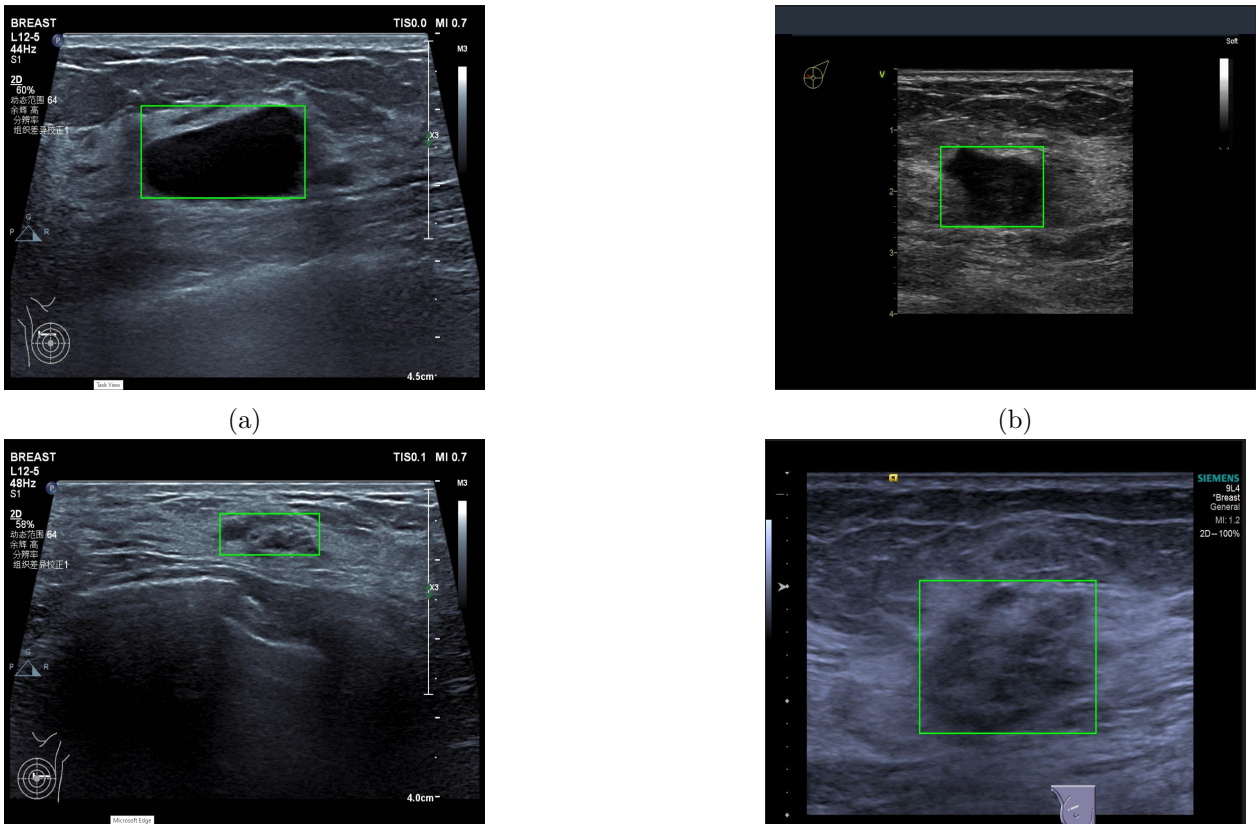


Figure 4.2: Sample images from dataset B (Green box: ground truth box encompassing the lesion).

Dataset C

Dataset C is a publicly available dataset [152]. It consists of 349 benign and 177 malignant, totaling to 509 images. Two machines were used to generate these images; LOGIQ E9 and LOGIQ E9 Agile with ML16-15 D matrix linear probe at 1-5MHz frequency. All images were collected from Baheya

Hospital for Early Detection & treatment of Women's Cancer, Cairo, Egypt [152]. All ROIs were provided by radiologists from the same hospital. These images contain 1 to 3 lesions per image. Here, corrupt images where the lesion was covered with artefacts were removed.

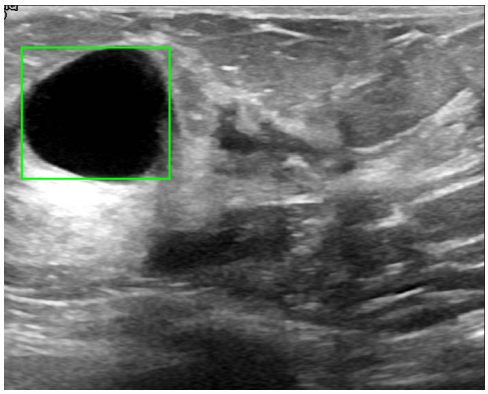
Figure 4.3 shows samples of benign and malignant cases from this dataset. Dataset C has a lot more challenging cases in comparison to other datasets. A considerable number of benign and malignant lesions have unclear boundaries or very similar texture as the background tissue. Figure 4.3a shows a case with single benign lesion; Figure 4.3b shows an example of multiple benign lesions; Figures 4.3c and 4.3d show challenging benign lesions that have unclear boundary or textures similar to that of the background; Figure 4.3e shows a benign case with irregular boundary; Figure 4.3f shows a case with single malignant lesion; Figure 4.3g shows a malignant lesion with unclear boundary with taller-than-wide characteristic; and finally Figure4.3h shows a malignant lesion with texture similar to the background normal tissue.

Dataset D

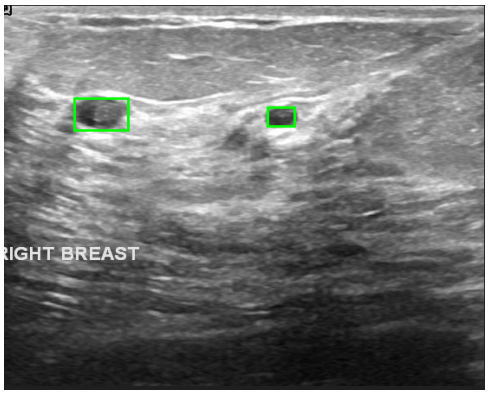
Dataset D comprises 383 benign and 170 malignant cases, a total of 553 images, collected from a single hospital. All images contain one lesion. Figure 4.4 shows samples of benign and malignant lesions from this dataset. Figures 4.4a and 4.4b show cases of benign and malignant lesions from this dataset, respectively. Figures 4.4c and 4.4d show challenging benign and malignant cases. The majority of the images in this dataset contain an additional doppler scan of the lesion such as the ones illustrated in Figures 4.4e, 4.4f 4.4g, 4.4h, 4.4e, 4.4f, 4.4g and 4.4h.

Dataset E

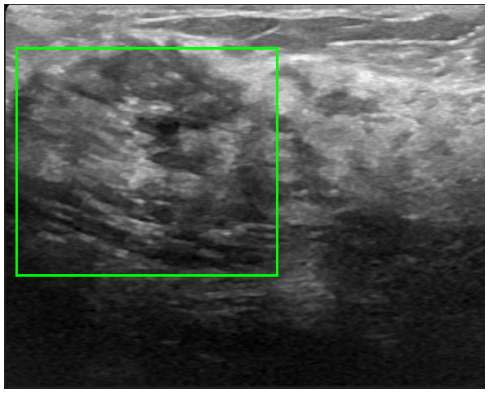
Dataset E consists of 168 images in total, of which 72 are benign and 96 are malignant cases, collected from a single hospital. These images contain one lesion. It is the only dataset where the number of malignant lesions is slightly higher than that of the benign lesions. Figure 4.7 shows sample images from this dataset. Figures 4.5a and 4.5b show simple benign and malignant lesions from this dataset; Figure 4.5c shows a taller-than-wide malignant lesion and Figure 4.5d shows a challenging cases of malignant lesions where the boundary is unclear.



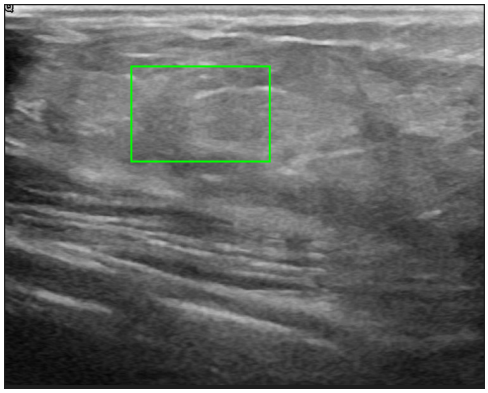
(a)



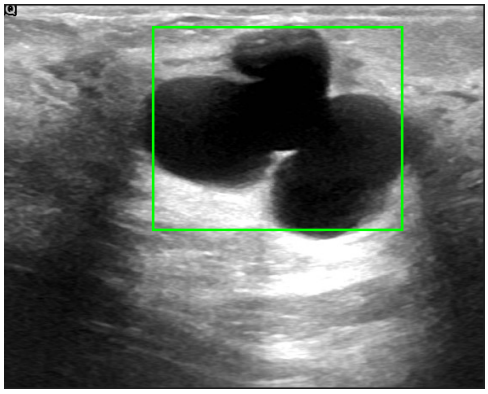
(b)



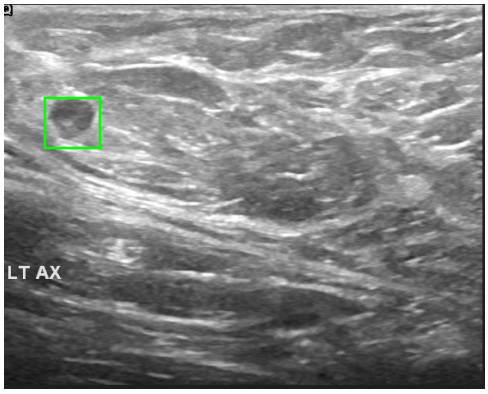
(c)



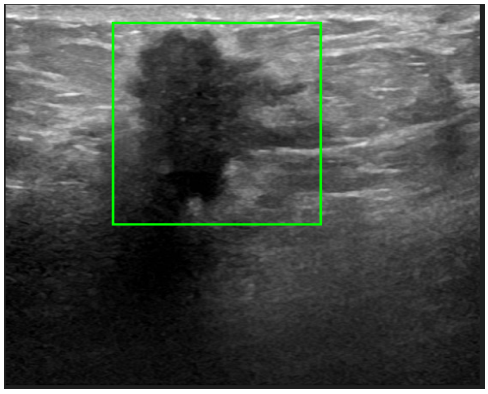
(d)



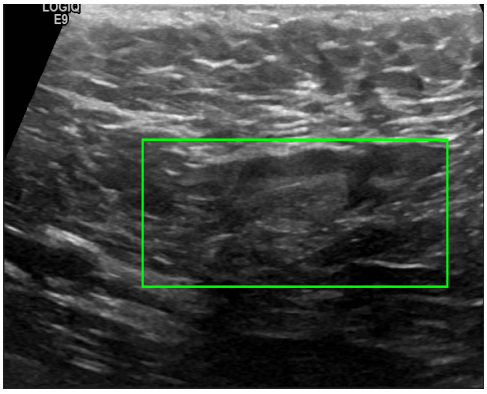
(e)



(f)

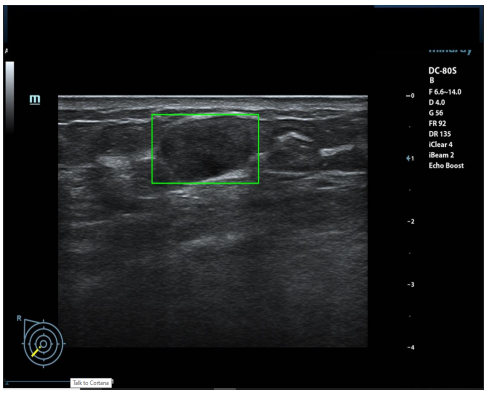


(g)

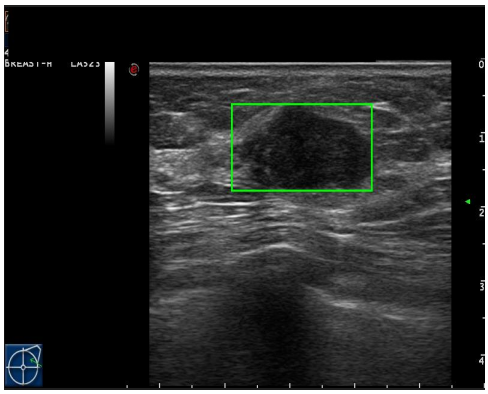


(h)

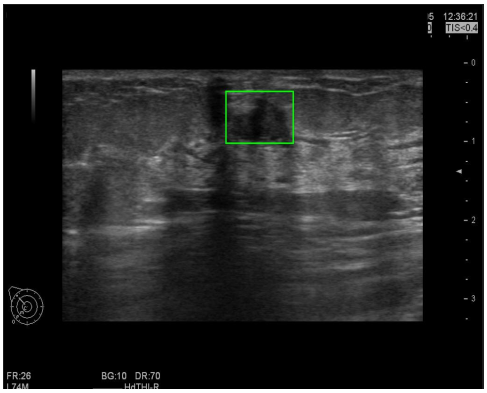
Figure 4.3: Sample images from dataset C (Green box: ground truth box encompassing the lesion).



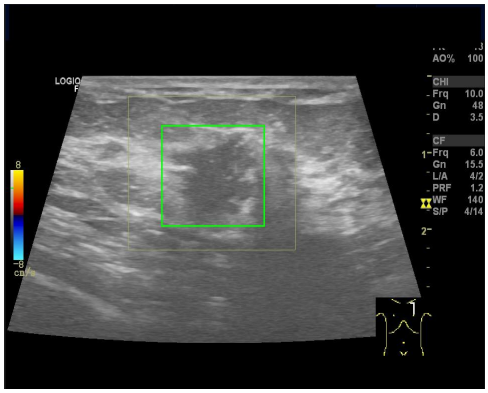
(a)



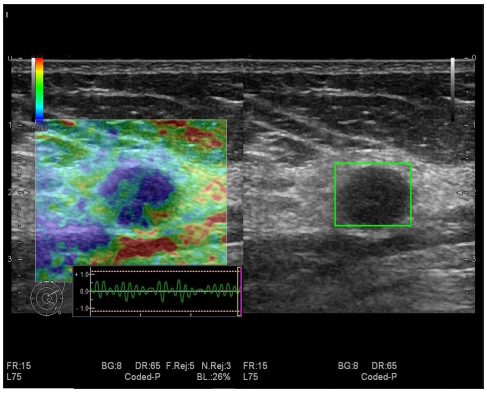
(b)



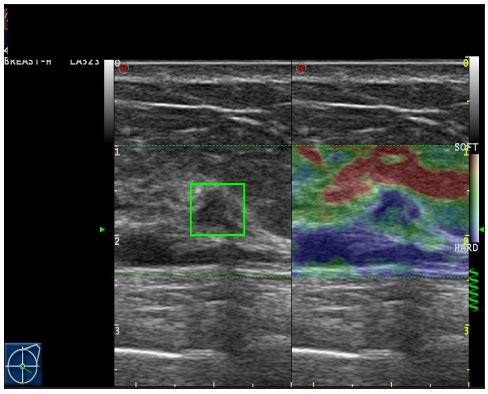
(c)



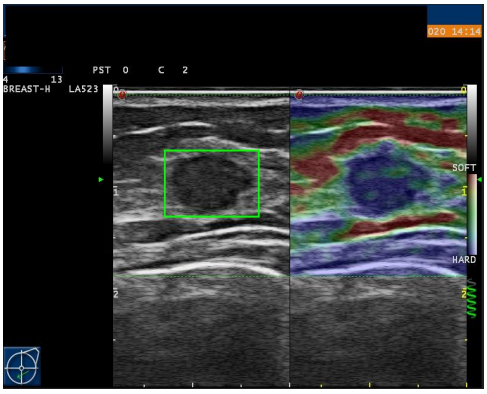
(d)



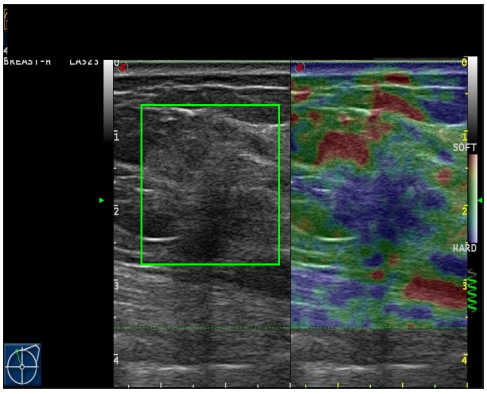
(e)



(f)

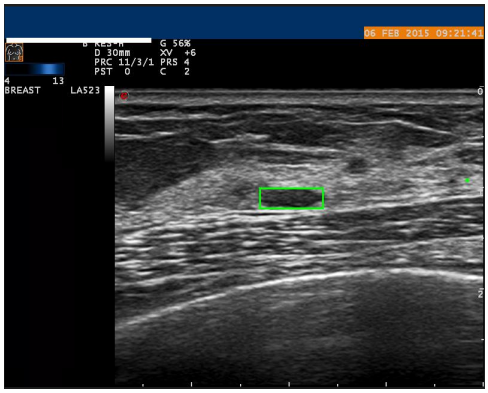


(g)

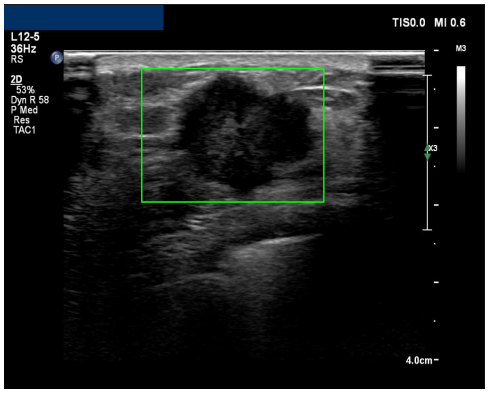


(h)

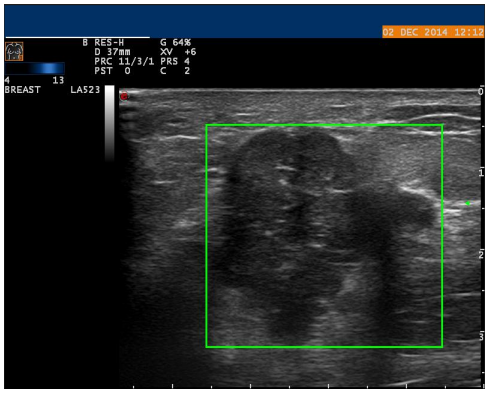
Figure 4.4: Sample images from dataset D (Green box: ground truth box encompassing the lesion).



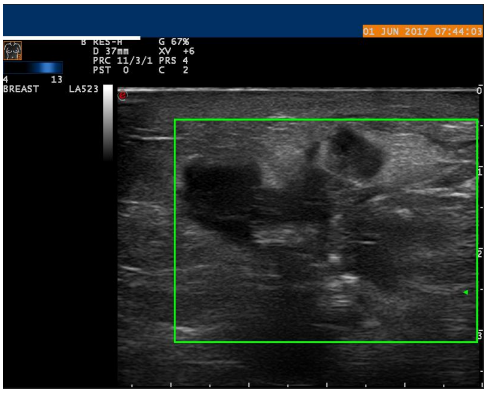
(a)



(b)



(c)



(d)

Figure 4.5: Sample images from dataset E (Green box: ground truth box encompassing the lesion).

Lesion Characteristics

One of the main variations observed across different datasets is the size* of the image lesion. Differences in the lesion size appearance in the image refers to several factors such as the physical size of the lesion and the way the image is acquired (zoom in or out). Such variations are common and our study sets no constraints on the image acquisitions. On the other hand, all the detection models analysed in this research are anchor-based. As discussed in Chapter 3, majority of the lesion detection methods adapt detectors developed for object detection in natural images for their application. The predefined anchor boxes in these detectors are based on natural images. In order to adapt them for breast lesion detection in US images, an understanding of the lesion size ($height \times width$) and aspect ratio (shape) is crucial.

Lesion size is the area ($height \times width$) of the GT box tightly covering the lesion. For readability and ease of visualisation, all lesion sizes are divided by 1000. Thus, a lesion of size 50 implies that the actual size of the lesion is 50,000 pixels. Table 4.1 shows the overall number of lesions in each dataset in different ranges of lesion sizes. As seen in Table 4.1, the vast majority of the lesions have small sizes in the range of < 50. The number of lesions in the higher size ranges is comparatively lower.

Lesion Size (in 1000s)	**Overall**				
	A	B	C	D	E
< 50	1227	127	372	440	99
$[50, 100)$	288	15	99	84	43
$[100, 200)$	182	7	45	28	26
≥ 200	36	1	10	1	0
Total	1733	150	526	553	168

Table 4.1: Number of lesions in different size range in all datasets.

To develop a deeper understanding, the distribution of benign and malignant lesions in these size ranges is analysed. Figure 4.6 shows the distribution of benign and malignant lesions in all datasets. In general, malignant lesions tend to be larger than benign lesions. In all datasets, the proportion of benign lesions with size < 50 is much lower in malignant lesions. Overall, around 89% of the benign lesions have size < 50 whereas only 48% of malignant lesions fall in this size range. On the other hand, only 2.6% of benign lesions are large (size ≥ 100) whereas 20.5% of the malignant lesions fall in this size range. Dataset B has the highest percent of small benign and malignant lesions. Dataset A has

*Size here refers to number of pixels.

the largest percent of large malignant lesions (around 28%). Dataset D has the second highest number of small lesions. This is due to the nature of the scans in this dataset. Dataset D contains images with the normal (B-mode) scan as well as doppler scan, side-by-side. Thus, even if a lesion is large enough to cover the whole scan region, the overall size of the lesion (pixels) is small. Such a size difference in benign and malignant lesions in US images is due to the following two reasons. First, generally, benign lesions are smaller than malignant lesions. Secondly, it is common practice for clinicians to capture enlarged images of malignant lesions for a clear visual of this lesion to ensure correct diagnosis.

Furthermore, all datasets (with the exception of dataset E) contain a larger number of benign lesions than malignant lesions. Therefore, the overall distribution of the lesion sizes of these dataset follows a distribution pattern closer to that of benign lesions. On the other hand, as dataset E contains a larger number of malignant lesions, the size distribution of all lesions in this dataset has a similar distribution as that of malignant lesions, with the lowest proportion of lesions being in the size range of < 50. Designing anchor boxes and for deeper understanding of the performance of the detection models as well as their development, aspect ratio is an important measurement. Aspect ratio is the ratio of width and height of the GT box tightly covering the lesion. Table 4.2 presents the mean aspect ratio of benign and malignant lesions in all size ranges. Overall, benign lesions have larger aspect ratio than malignant lesions in the same size range. This is due to the nature of these lesions as described in Section 2.1 in Chapter 2. Benign lesions are typically parallel to the surface of the skin. As a result, the benign lesions generally appear 'wider-than-tall' in the US image (large aspect ratio). On the other hand, malignant lesions are generally perpendicular to the skin surface and appear as 'taller-than-wide' in the US image. Thus, malignant lesions generally have smaller aspect ratio than benign lesions.

Lesion Size	**Benign Lesions A.R.**					**Malignant Lesions A.R.**				
(in 1000s)	**A**	**B**	**C**	**D**	**E**	**A**	**B**	**C**	**D**	**E**
< 50	1.73	1.66	1.84	1.81	1.72	1.47	1.61	1.44	1.54	1.53
$[50, 100)$	1.76	2.42	1.91	1.75	1.96	1.58	1.66	1.41	1.47	1.70
$[100, 200)$	1.73	1.53	1.64	2.01	2.75	1.68	1.35	1.51	1.34	1.63
≥ 200	1.48	-	1.66	2.23	-	1.60	1.37	1.34	-	-

Table 4.2: Aspect Ratio (A.R.) of benign and malignant lesions in different size range in all datasets ('-' indicates no lesion in that size range was present).

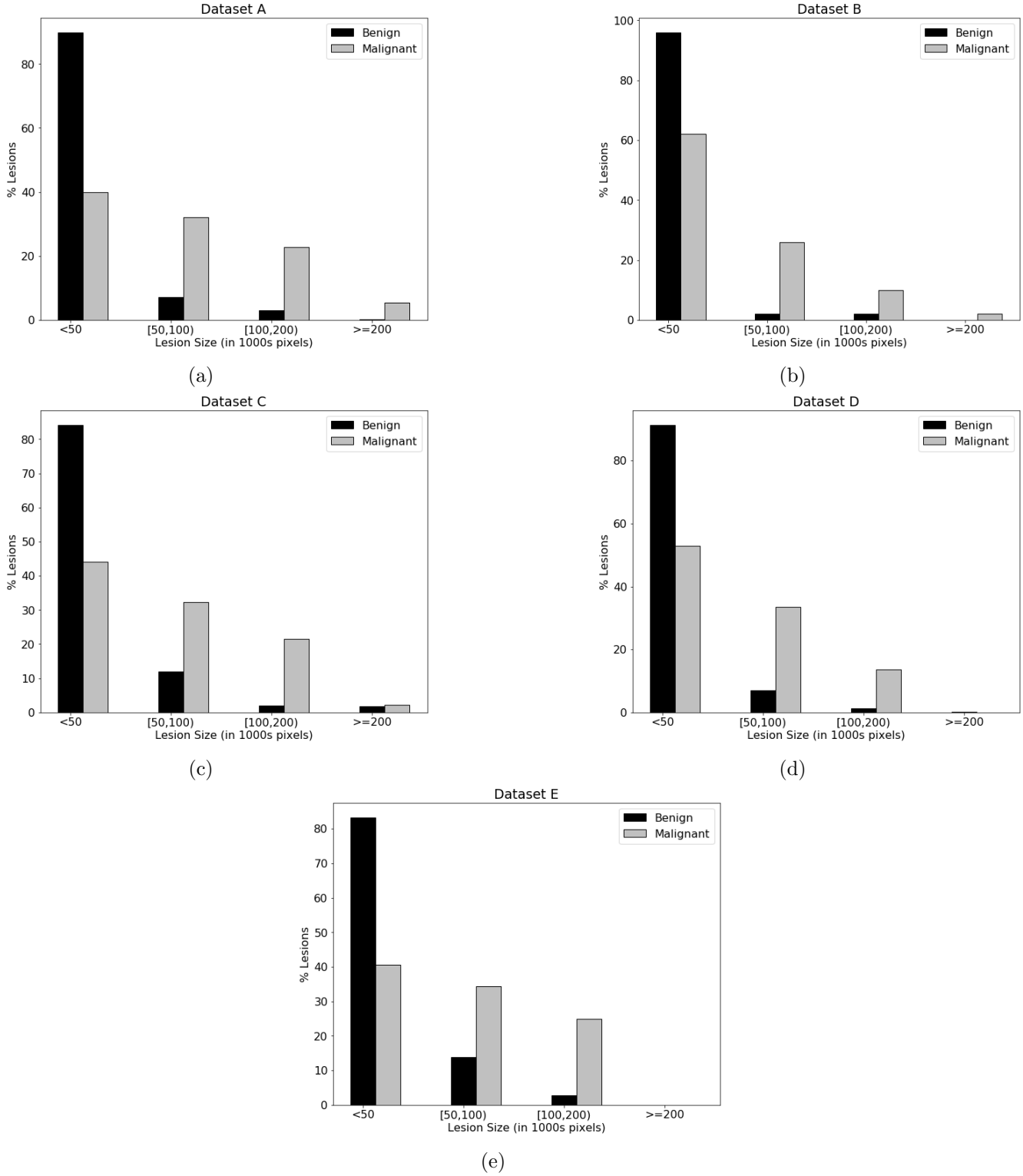


Figure 4.6: Distribution of benign and malignant lesion sizes in all datasets.

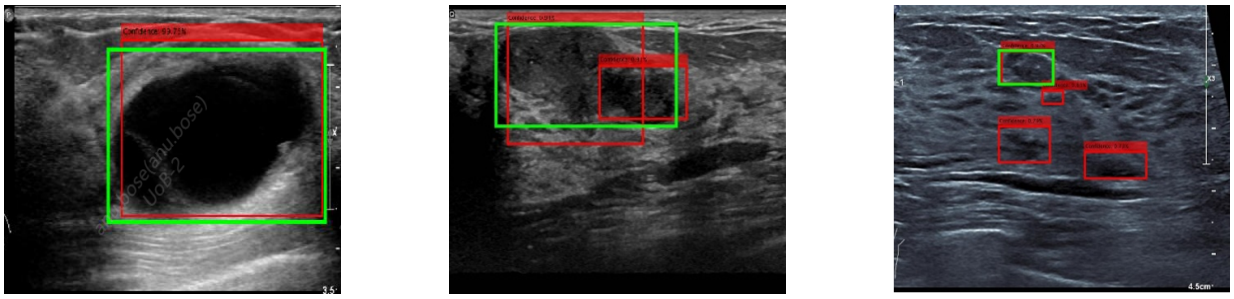
In summary, this section presented the characteristics of lesions in our datasets. Particularly, size and aspect ratios of the lesions were studied and presented as this provides critical information for development and evaluation of detection models. Overall, the characteristics of the lesions in US images reflect their biological characteristics. Benign lesions are generally smaller and have larger aspect ratio than malignant lesions. Irrespective of the type of the lesion, larger lesions have smaller aspect ratio. As seen throughout this section, the variety of lesions used in this research is large.

4.2 Experimental Setup and Evaluation Metrics

This section details the experimental platform, protocols and evaluation metrics used in this research. All implementations are conducted on Intel(R) Xenon(R) Gold 6230R CPU with 64 bit OS and NVidia GeForce. MATLAB 2020b is used to compose the experimental scripts. To determine the detection accuracy, a 5-fold cross validation protocol is used. In each iteration, we split the US images in dataset A into training and testing sets at a ratio of 80% to 20%. The 80:20 ratio was selected to strike a balance between using a diverse set of images for model training while reserving a sufficient number for thorough testing. The medical research testing protocol for external tests is followed by evaluating the models that have been internally tested during the cross-validation. Following a detailed description of evaluation matrices:

Detection Models: All detection models are evaluated using standard evaluation metrics of this field. First, the quality of all output detections is computed using Intersection-over-Union (IOU). IOU measures the degree of overlap between output box(es) and GT box(es) covering the lesion(s). IOU is the ratio of the area of intersection between two both to the area of their union. Based on IOU, the output detections are categorised as either true positive (TP) or false positive (FP). If a detection has $IOU \geq 0.5$ with the GT box, then it is considered as a TP detection. Thus, a TP detection covers at least 50% of the lesion. An example of this is shown in Figure 4.7a. On the other hand, if a detection has $IOU < 0.5$ with the GT box, it is considered a FP detection. The IOU threshold of 0.5 for categorisation of detections as TP or FP is a standard threshold used for evaluation of detection models across all domains.

FP detections are further categorised as additional boxes or low IOU FPs depending on its IOU



(a) (b) (c)

Figure 4.7: Types of output detections. (a) TP detection (b) TP detection with overlapping FP detection. (c) TP detection with FP detections in background scan region (additional boxes).

with GT box. Low IOU FPs have IOU in the range of $(0, 0.5)$ whereas additional boxes have no overlap with GT (IOU 0). Additional boxes typically cover background regions that have lesion-like texture. Figures 4.7b and 4.7cprovide sample cases of low IOU FPs and additional boxes, respectively. If the model generates multiple outputs for a single lesion, the output detection with highest confidence score is considered as the *main* detection and all other detections are categorised as FP by default. Depending on the IOU of the *main* detection with GT, it is categorised as TP or FP. If the image contains > 1 lesion and the number of output detections exceeds the number of lesions, then the same steps are followed. False negative (FN) cases are those where the model completely misses the lesion (no bounding box was generated for the images).

After categorisation of all output detections, performance of the model is summarised using precision, recall and F-measure defined in Equations 4.1, 4.2 and 4.3, respectively. Precision represents the proportion of correct detections (TP) from all the output detections generated by the model. A high precision value indicates that a large majority of the output detections are TPs whereas a low precision value indicates a considerable portion of the output detections are FPs. Recall represents the proportion of objects (lesions) in the test set that were correctly detected by the model as TPs. Thus, a high recall indicates that the majority of the lesions in the test set were correctly detected by the model (TP) whereas a low recall indicates that a significant portion of the lesions were missed by the model (FN). F-measure summarises the performance of the model to a single value. Both precision and recall have equal impact on the final F-measure of the model.

$$Precision = \frac{TP}{(TP + FP)} \tag{4.1}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{4.2}$$

$$F - measure = 2 \times \frac{Precision * Recall}{Precision + Recall} \tag{4.3}$$

An important characteristic of a detection model is its computation time. In this work, computation time is the time required by a detector to fully process a single input image and generate the output. As mentioned in Section 4.1, all models are trained on 5 folds of the dataset A. Of the five trained models, the model with highest F-measure is selected as the best model.

Cluster analysis and evaluation:

In this research, k-means++ and x-means clustering are used. K-means++ clustering, used in Chapter 5 to cluster GT boxes, requires a predefined k which is the number of clusters to be formed. To determine the optimal k, the elbow method is used. The elbow method is described in detail in Section 2.2.3.2 Chapter 2. As an overview, distortion is the mean of the squared distance between centroids and all samples in their respective cluster. The elbow method involves manually selecting an optimal k from a range of k values on the basis of change in distortion. In our application, inspired by YOLO-v2 [45], distance between two boxes is measured in terms of IOU as shown in Equation 4.4. Thus, distortion is also measured in terms of IOU as shown in Equation 4.5.

$$d(box1, box2) = 1 - IOU(box1, box2) \tag{4.4}$$

$$distortion = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} d(centroid_i, box_j)}{k} \tag{4.5}$$

Here, n represents the total number of samples in the i^{th} cluster. Thus, higher distortion indicates better representation of the cluster by its centroid and vice versa. We refer to distortion as mean IOU. Therefore, low values k have low mean IOU and vice versa. A graph is generated with k values plotted on the x-axis and their respective mean IOU (distortion) values in the y-axis. In this graph, the value of k where the mean IOU (distortion) value stabilises is selected as optimal k. Thus, a good balance between number of clusters and representation of the dataset is achieved. X-means clustering is used in Chapters 6 and 7 to cluster test proposals generated by the RPN. This method of clustering

requires a predefined range of k from k_{min} to k_{max}. The optimal k is automatically determined from this range using Bayesian Information Criteria (BIC) described in Section 2.2.3.2 Chapter 2.

Classification Models: Images used for training and testing classification models are assigned GT labels. The images are referred to as positive or negative samples based on the assigned GT labels. First step in measuring the performance of a classification model is to categorise the test images as TP, FP, FN or true negative (TN) based on labels output by the model. If a positive or negative sample is correctly labelled as positive or negative by the model, then the image is categorised as TP or TN, respectively. On the other hand, if the positive or negative sample is incorrectly labelled as negative or positive by the model, then the image is categorised as FP or FN, respectively. After the test images are categorised, the model's performance is measured using accuracy, specificity, precision, recall and F-measure.

Accuracy of a classifier is defined in Equation 4.6. It represents the proportion of test images that were correctly labelled by the classifier. Specificity, defined in Equation 4.7, represents the proportion of negative samples that were correctly classified by the model. Precision, recall and F-measure are defined in Equations 4.1, 4.2 and 4.3, respectively. Precision of a classification model indicates the proportion of true (GT) positive samples out of all the samples that the model labelled as positive. Thus, 80% precision of a model indicates that 80% of the true positive samples were correctly labelled by the model. Recall of a classification model indicates the proportion of positive samples correctly identified by the model. Thus, a classification model with 80% recall indicates that 80% of the positive samples were correctly labelled positive by the model. Finally, F-measure summarises precision and recall values into a single performance value.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{4.6}$$

$$Specificty = \frac{TN}{TN + FP} \tag{4.7}$$

Chapter 5

Adaptation of FRCNN for Breast Lesion Detection in 2D Ultrasound Images

FRCNN is a deep-learning based 2-stage detector developed for object detection in natural images. As detailed in Chapter 3, FRCNN is popularly adapted not only for breast lesion detection in US images but also for detection of other types of lesions in images generated from different modalities. These works adapt FRCNN by modifying the modelling hyperparameters and/or the network architecture. Although the FRCNN network modified in these works have shown high performance for breast lesion detection in US images, they have two drawbacks: insufficient experimental evaluation of the modifications and/or the use of a small to medium sized dataset.

Generally, FRCNN is adapted through modification of several modelling hyperparameters along with the network architecture. However, the impact of the individual modifications on the overall performance is not supported with sufficient experimental evaluation. Furthermore, a considerable portion of these works use a small to medium sized dataset. Since US images collected from different sources (US machines, hospitals, etc.) have significant variations, the generalisation capability of a lesion detection model is a crucial characteristic. In methods that use of small to medium sized dataset, it is difficult to gauge the generalisation capabilities of the adapted FRCNN model proposed.

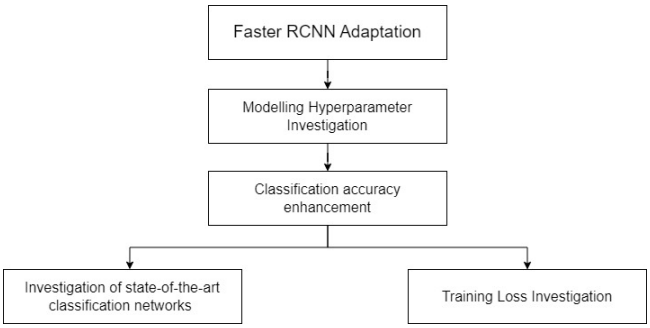


Figure 5.1: Stages involved in the adaptation of FRCNN for breast lesion detection in US images.

We address these common drawbacks in this chapter through an investigation of the impact of several modelling hyperparameters from both stages of the FRCNN model using our large dataset of US images. Through this investigation, optimal values of each hyperparameter is determined and used to design our adapted FRCNN model. The investigated hyperparameters and development of the adapted FRCNN model are described in Section 5.1. To reduce the FP detections of adapted FRCNN, two methods to improve the classification accuracy of both stages of the network were investigated, namely, evaluation of state-of-the-art classification networks as the backbone network of the adapted FRCNN model and impact of various training losses. Section 5.2 contains the details of these investigations. Section 5.3 presents the experimental results of all investigations. Section 5.4 discusses findings relevant to the work presented in this chapter. Section 5.5 contains the summary of the work presented in this chapter along with an overview of the main contributions and findings.

5.1 FRCNN Investigation

This section describes the investigation of the modelling hyperparameters of the FRCNN network and its adaptation for breast lesion detection in US images. The architecture of the FRCNN network along with training details is comprehensively explained in Section 2.3.1 of Chapter 2. Section 5.1.1 details the investigated modelling hyperparameters from both stages of the network and Section 5.1.2 describes the development of an adapted FRCNN model.

5.1.1 Optimal Modelling Hyperparameters Selection

This section describes the evaluated FRCNN modelling hyperparameters. First, the evaluated modelling hyperparameter from the first stage (RPN) is described. This is followed by a description of the

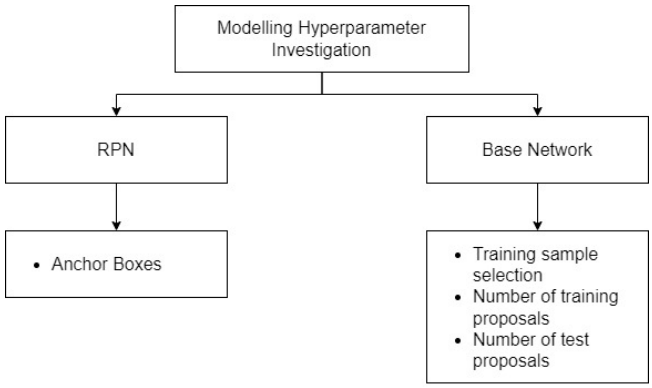


Figure 5.2: Investigated FRCNN modelling hyperparameters.

hyperparameters investigated in the second stage (base network) of the network. Figure 5.2 illustrates the investigated hyperparameters.

5.1.1.1 Region Proposal Network (RPN) Hyperparameters

RPN is the first stage of the FRCNN network. It relies on a predefined set of anchor boxes for generation of proposals which are then passed to the base network for further processing. Anchor boxes are, fundamentally, the first set of predictions of potential object size which is refined by the RPN and the base network. Poor selection of anchor boxes negatively impacts the quality of proposals generated by the RPN. As the base network is trained using RPN-generated proposals, the low quality of proposals has a negative influence on the base network's training. Furthermore, the base network relies on the test proposals generated by the RPN during model testing. Therefore, low quality of these proposals would severely limit the detection performance of the base network and, consequently, the whole network. Therefore, selection of the anchor boxes has a significant influence on the overall training and performance of the model. Thus, their impact is studied in further detail.

We investigated four sets of anchor boxes. First, the default anchor boxes of the FRCNN model, with scales {8, 16, 32} and aspect ratios {1:1, 1:2, 2:1}, are studied. Since the size of breast lesions is relatively smaller in comparison to objects in natural images, anchor boxes of scale 1 and aspect ratio {1:1, 1:2, 2:1}, are studied. These anchor boxes are referred to as fundamental anchor boxes. To include a larger variation of potential lesion sizes, the original and fundamental anchor boxes are combined to form a new set of anchor boxes referred to as combined anchor boxes [153].

The aforementioned three sets of anchor boxes are manually estimated based on our knowledge of the objects (lesions). To automate the anchor box estimation process, we used k-means++ clustering [154]. Anchor boxes estimated using this method are referred to as k-means++ anchor boxes [153]. This method of automatic estimation is inspired by YOLOv2 [45] where k-means clustering is used to automatically estimate anchor boxes for object detection in natural images. To compute k-means++ anchor boxes, all GT boxes from the modelling dataset (dataset A) were used. The GT boxes are defined by their heights and widths. In other words, the input to the k-means++ clustering is the height and width of all GT boxes. K-means++ requires a predefined value of k. Here, k centroids represent the k estimated anchor boxes. Optimal k is manually determined using the elbow method described in Section 4.2 of Chapter 4. For this, $k = [1, 30]$ with increments of 1 were evaluated.

Optimal k selected using the elbow method provides a good balance between generalisation and computation cost is achieved. Higher k values a relatively high mean IOU i.e. a very good representation of the boxes in the dataset. However, the model trained with these anchor boxes will overfit on the modelling dataset and will not generalise well on other datasets. Also, the computation cost increases with an increase in the number of anchor boxes as the RPN is required to process through a higher number of anchor boxes. Alternatively, low k values have a relatively low mean IOU, i.e. inadequate representation of the dataset, but the computation cost will be low due to lower number of anchor boxes. Selection of optimal k for this dataset is presented in Section 5.3.2. To the best of our knowledge, no other work in this field has utilised this method of automated anchor box estimation for breast lesion detection in US images.

5.1.1.2 Base Network Hyperparameters

This section details the investigated modelling hyperparameters of the base network which is the second stage of the FRCNN model. This includes selection of samples for training the base network, number of training proposals, and number of test proposals.

Training Samples: Base network is trained with proposals generated by the RPN. Selection of training samples directly influences the classification accuracy of the base network, in turn impacting the number of FPs generated by the model. In the original configuration, proposals with $IOU \geq 0.5$ with GT are considered as positive training samples and those with $0.1 \leq IOU < 0.5$ are considered

as negative training samples. Proposals with IOU 0 are used for hard-example mining. Since these IOU thresholds were selected for object detection in natural images, the impact of a range of these thresholds when used for breast lesion detection in US images is studied.

Positive samples train the network to correctly identify objects (lesions) whereas negative samples train the model to correctly identify background regions. The original FRCNN uses a broad range for selection of the training samples. However, when used for breast lesion detection in US images, this range proved ineffective as experimentally shown in Section 5.3.1 due to the textural similarity between background regions and lesions. We hypothesise that use of a smaller IOU range for selection of both positive and negative samples would improve the network's ability in differentiating lesions from background regions. Therefore, following IOU thresholds were evaluated for selection of positive training samples: $[0.6, 1]$, $[0.7, 1]$ and $[0.8, 1]$ with negative sample selection maintained at the default IOU threshold of $[0.1, 0.5)$. The original FRCNN model is trained with IOU 0 samples for hard example mining. However, when utilised for breast lesion detection in US images, it was experimentally found that the performance of the network improved without the use of hard-example mining as shown in Section 5.3.1. Thus, for this evaluation of negative sample selection, hard example mining was not used. Following IOU threshold ranges were evaluated for negative sample selection with the positive sample selection threshold maintained at its default values of $[0.5, 1]$: $[0, 0.2)$, $[0, 0.3)$, and $[0, 0.4)$.

Number of Training Proposals: In the original FRCNN model, the base network is trained with 2000 RPN-generated training proposals. This high number was selected according to the larger number of objects typically found in a natural image. In comparison, the average number of objects present in one breast US image is much lower. Therefore, lower values of these training proposals, specifically 300 and 1000, were studied.

Number of Test Proposals : During model testing, RPN in the original FRCNN model forwards its top 300 test proposals to the base network for further refinement. This large number is set for the same reason as the large number for training proposals, i.e., the comparatively higher number of objects found in a typical natural image. Since the average number of objects is much lower in US images, a lower number of test proposals, specifically 100, was investigated.

5.1.2 Adapted FRCNN Development

Throughout the investigation of modelling hyperparameters described in Section 5.1.1, optimal values for each hyperparameter were selected. This selection was done on the basis of overall performance on a smaller modelling dataset referred to as dataset A-small which is a subset of dataset A. Dataset A-small is described in further detail in Section 5.3. Selected optimal values are presented in Section 5.3.3. The FRCNN model designed with the selected optimal values is referred to as adapted FRCNN.

5.2 Classification Accuracy Enhancement

FPs output by the adapted FRCNN model are caused due to the following two drawbacks: incorrect classification of FP proposals by both stages of the FRCNN model and incorrect retention of FP proposals by NMS (post-processing method used in both stages of FRCNN). This chapter addresses the issue of poor classification accuracy of the RPN and base network. Particularly, the following two strategies were evaluated. Firstly, multiple state-of-the-art classification networks as the backbone network of the adapted FRCNN model are evaluated. Secondly, training losses aimed at reducing hard samples responsible for FP detections are investigated. In particular, these losses are utilised in the RPN of the adapted FRCNN model in order to improve the overall quality of proposals passed to the base network which would in turn reduce the number of FPs generated by the model. Both these methods are described in this section.

5.2.1 Backbone Networks

The default backbone network in the FRCNN model is VGG16 which is also used in adapted FRCNN. Both RPN and base network rely on the features extracted by this backbone network for their tasks of classification and bounding box regression. Additionally, the accuracy of the backbone network has a strong influence on the classification accuracy of the base network. Therefore, various classifiers are evaluated as the backbone of the adapted FRCNN model. These include ResNet50 [30], ResNet101 [30], Inception-ResNet-v2 [31] and Inception-v3[155]. These selected networks have high accuracy in classification of natural images and have also proven to have similar high classification and detection performance in medical images. When utilised in the adapted FRCNN model, these networks are pretrained on ImageNet dataset for classification of natural images.

5.2.2 Network Training Loss

RPN and base network contain a classification and bounding box regression branch. FRCNN uses binary cross-entropy (CE) for classification task and smooth-L1 loss for bounding box regression task in both these stages. Final loss of the network is the sum of all four losses (classification and regression loss of the RPN and base network). Training loss of classification and regression branches are completely independent of each other which causes the following scenario: a high IOU proposal that is correctly assigned a high classification score is transformed into a FP due to the regression branch output transforming it to a low IOU detection ($IOU < 0.5$). The output detection in this scenario is a low IOU FP with high classification score. Furthermore, loss of all training samples are assigned equal weights, without more focus given to hard samples.

To address these issues, PISA and CARL losses [90] are evaluated. PISA loss is used in the classification branch whereas CARL loss is used in the regression branch. In essence, PISA loss improves classification accuracy by assigning larger weight to training loss of hard samples and smaller weight to that of the easy samples. Since FPs are hard samples that are misclassified, PISA loss is used to directly address these cases. On the other hand, CARL loss is used in the regression branch. It combines the classification and regression branches by introducing the classification score assigned to the training sample in the loss of the regression branch such that poor transformation of a high scoring classification output is assigned higher weight. For instance, if a TP proposal that is assigned a high classification score is converted to a low IOU FP proposal after application of the regression output, then the regression loss of this sample is assigned higher weight than a correct transformation of the same sample. In this manner, CARL loss addresses FP detections caused due to the disconnect between classification and regression branches.

Table 5.1 shows the evaluated combinations of losses. Additionally, PISA negative is studied to improve the classification of hard negative samples that result in FPs. Two variations of PISA negative are also evaluated. These variations are based on the changes in γ value in the Equation 5.1 [90] for weight computation.

$$w_i = ((1 - \beta)u_i + \beta))^{\gamma} \tag{5.1}$$

Loss	Classification Loss	Regression Loss
Default	Cross Entropy	Smooth L1
PISA + Smooth L1	PISA	Smooth L1
CE + Smooth L1	Cross Entropy	Smooth L1
PISA + CARL	PISA	CARL

Table 5.1: Investigated training losses for RPN of adapted FRCNN model.

Here, β is a bias value responsible for determining the smallest weight assigned to training loss of samples, u_i is the assigned rank of the i^{th} sample, γ determines the degree of importance given to a sample and w_i is the weight assigned to the loss of the sample. In this investigation, γ is increased to 1 and 1.5 to increase the importance given to hard negative samples i.e. FP proposals. These training losses are only investigated in the RPN of the adapted FRCNN model in order to improve the quality of training and test proposals. Higher quality of training proposals positively impact the classification accuracy of the base network. During model testing, the reduction of FP proposals sent through to the base network would in turn reduce the number of FPs generated by the model.

5.3 Experimental Results

This section contains experimental results of the investigations described in Sections 5.1 and 5.2. First, performance of the original FRCNN model is presented in Section 5.3.1. This is followed by the investigation of the modelling hyperparameters in Section 5.3.2. FRCNN models trained for the study of the modelling hyperparameters in Section 5.3.2 are trained with default values of all other modelling hyperparameters with the exception of the modelling hyperparameter being studied. After the investigation of the modelling hyperparameters and the selection of their optimal values, Section 5.3.3 details the performance of the adapted FRCNN model which is designed using the selected optimal values. Sections 5.3.4 and 5.3.5 present the impact on the classification accuracy of the adapted FRCNN model with change in the backbone network and training loss, respectively. Performance of the adapted FRCNN model is then compared to that of several state-of-the-art detection methods in Section 5.3.6.

FRCNN models in Sections 5.3.1, 5.3.2 and 5.3.3 are trained on 5 folds of dataset A-small. This dataset is a subset of the dataset A described in Section 4.1 of Chapter 4. Dataset A-small consists of 524 images (262 benign and 262 malignant cases). Performance of the models in these datasets was

used for selection of optimal hyperparameters and design of the adapted FRCNN model. Furthermore, in these sections, datasets B and C are used unseen test sets. Reproducibility of the selected optimal values on larger dataset (dataset A) is presented in Section 8.1 in Chapter 8. FRCNN models in the following Sections 5.3.4, 5.3.5 and 5.3.6 are trained on 5 folds of dataset A while datasets B, C, D and E are used as unseen external test sets as described in Section 4.1 of Chapter 2. Unless otherwise specified, all performance values are an average over 5-folds.

All FRCNN models are trained in an end-to-end fashion using the following common modelling hyperparameters: weights are initialised using normal distribution; models are trained for a total of 10 epochs with the initial learning rate set to 0.001 which drops by a factor of 0.1 at the 7^{th} epoch; weight decay is set to 0.0005; momentum of 0.9 is used; minibatch size of RPN and base network is set to 256 and 128, respectively; stochastic gradient descent is used as the optimization algorithm for model training. If a FRCNN model is trained or tested with modelling hyperparameters differing from that of the original FRCNN, it is highlighted in this section.

5.3.1 Original FRCNN for Breast Lesion Detection in US Images

Table 5.2 shows the performance of the original FRCNN model on datasets A-small, B and C. Overall, this model has high recall due to the high number of detected lesions. On the other hand, due to the high number of FPs, its precision is considerably low. Thus, despite its high recall, the F-measure of this model is low due to its low precision.

Model	**Dataset A-small**			**Dataset B**			**Dataset C**		
	P	R	F	P	R	F	P	R	F
Original	55.60	98.98	70.89	54.51	99.09	70.25	42.84	86.08	57.04

Table 5.2: Precision (P), Recall(R) and F-measure (F) of original FRCNN model.

The correct detections of this model (TPs) generally have high IOU, typically in the range of $[0.7, 0.8)$, as shown in Figure 5.4. Generally, the output detections of the original FRCNN model contain one high IOU TP along with at least one FP. The FPs of this model can be divided into two main categories; low IOU FPs (FPs with IOU $(0, 0.5)$ and additional boxes (FPs with IOU 0 or covering small regions of a large lesion with low IOU). Figure 5.3 illustrates both these FP cases. Figure 5.5 shows the IOU of FPs with GT lesion(s) in all three datasets. As seen in this figure, out of the

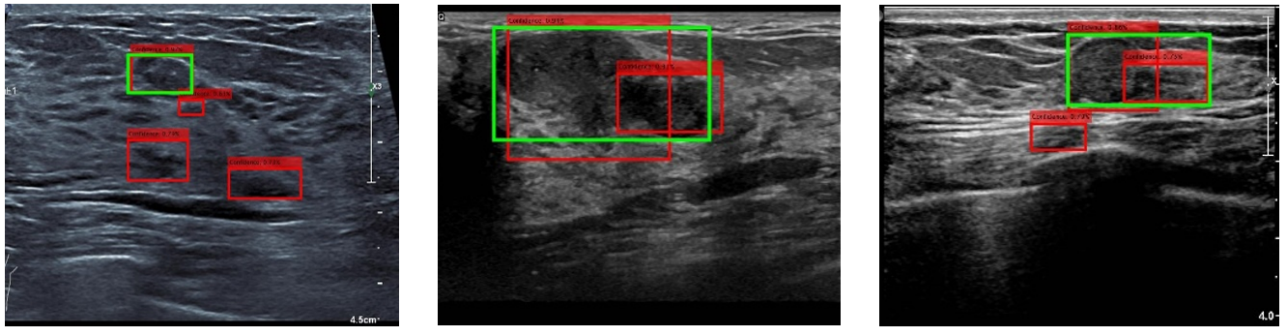


Figure 5.3: FPs generated by original FRCNN (Green box: Ground truth, red boxes: output detections).

total FPs generated by the original FRCNN model, around 80% are additional boxes (FP detections with $IOU0$). Typically, these additional boxes are scored lower than the main TP detection with rare exceptions.

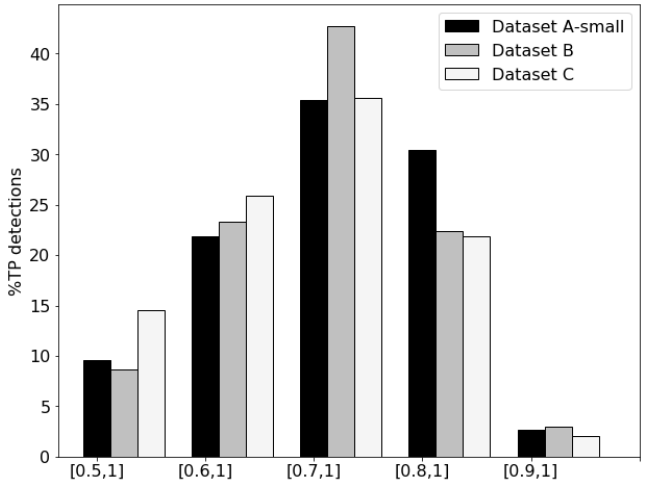


Figure 5.4: IOU distribution of TP detections of original FRCNN.

Performance of the original FRCNN model is relatively lower in dataset C due to the nature of images in this dataset. Ths dataset contains a considerable number of challenging lesions that have background-like texture, an example of which is shown in Figure 5.6. These lesions are missed by the original FRCNN model resulting in its relatively lower recall in this dataset. Also, like other datasets, the low precision here is also due to the high number of FPs, specifically additional boxes. The images in this dataset are commonly populated with lesion-like regions in the background normal region of the scan. Examples of such images are shown in Figure 5.6. Additionally, in some images, one GT

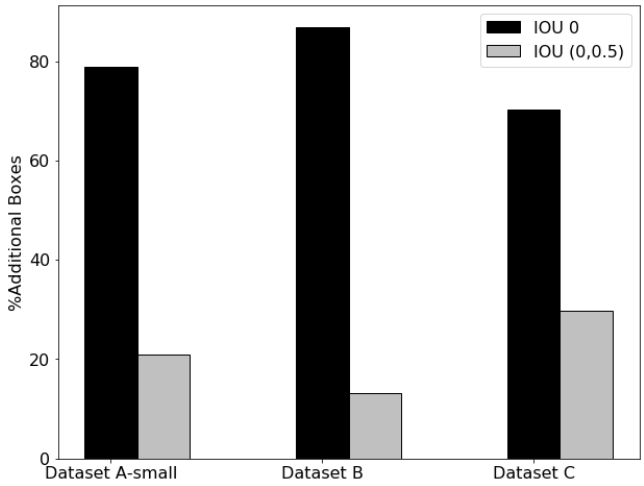


Figure 5.5: IOU distribution of additional boxes (FPs) with ground truth in original FRCNN.

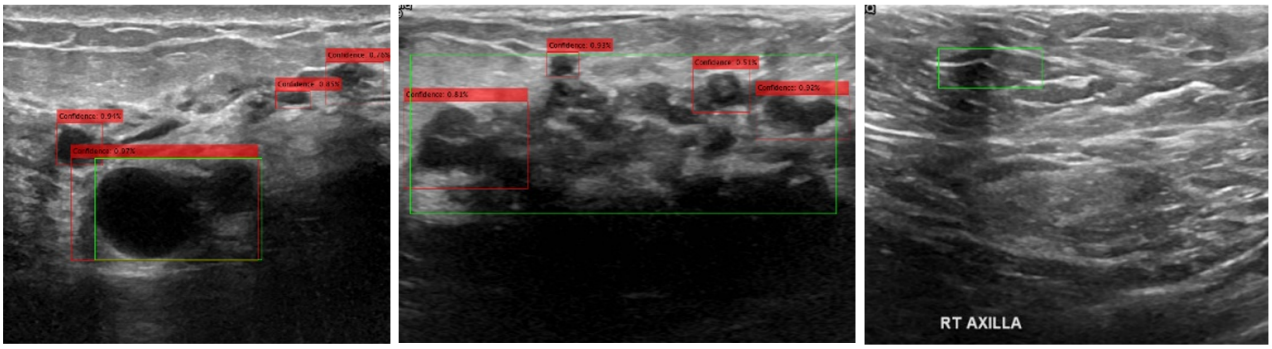


Figure 5.6: Original FRCNN performance in dataset C (a) FPs in lesion-like regions of the background scan area (b) FP due to multiple detections in large lesion (c) Lesion with background like texture missed by the model (FN). Green boxes: Ground truth. Red boxes: output detections.

box contains two lesion-like regions as shown in Figure 5.7. In these cases, the original FRCNN model detects the two separate lesion-like regions as individual lesions, thereby increasing the number of FP detections in this dataset.

5.3.2 Modelling Hyperparameters Selection

In this section, performance of various modelling hyperparameters of the RPN and base network as well as selection of their optimal values is presented.

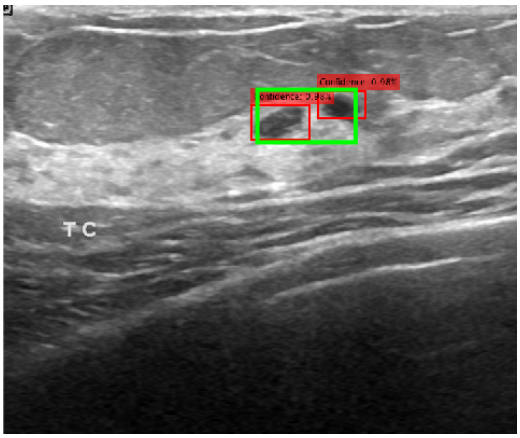


Figure 5.7: Detections by adapted FRCNN (red boxes) for lesion-like regions in ground truth box (in green) from dataset C.

5.3.2.1 RPN: Anchor Box Selection

This section presents the impact of four sets of anchor boxes on the overall performance. Besides the anchor boxes, all modelling hyperparameters of the FRCNN models were maintained at their default values. The evaluated anchor boxes include original, fundamental, combined and k-means++ anchor boxes. Original, fundamental and combined anchor boxes were selected manually whereas k-means++ anchor boxes were automatically estimated using GT boxes of modelling dataset (dataset A-small). As described in Section 5.1.1, k values in the range of $[1, 30]$ with increment of 1 were evaluated. The elbow method is used to select optimal k from this range and mean IOU is the distortion measured for each evaluated k. Detailed explanation of the elbow method and mean IOU is provided in Section 2.2.3.2 Chapter 2. The distribution of mean IOU versus a range of k values is shown in Figure 5.8. In this figure, 'number of anchors' represent k values. For low k values, the mean IOU is low which increases drastically as the k increases. After $k = 5$, the change in mean IOU is relatively smaller. At $k = 5$, the mean IOU is 0.75. As the k increases, the mean IOU increases further in the range of 0.8 to 0.9. Thus, based on the elbow method, $k = 5$ is selected as optimal as it provides a good balance between generalisation and computation cost.

Table 5.3 shows the performance of all anchor boxes on dataset A-small and unseen test sets (datasets B and C). Figure 5.9 shows the change in the number of TP, FP and FN with the change in anchor boxes in datasets A-small and B. Change in TP, FP and FN in dataset C with change in anchor boxes is shown in Figure A.1 in Appendix A. For brevity, FRCNN models are referred to by

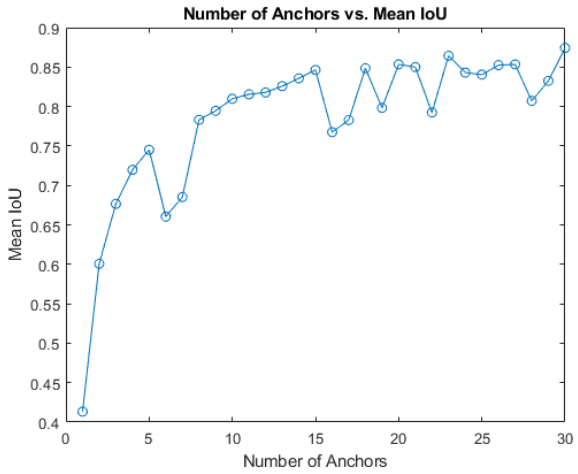


Figure 5.8: Mean IOU for various number of anchor boxes: Dataset A-small.

their anchor boxes. Overall, the fundamental anchor boxes have the highest F-measure, outperforming all anchor boxes including original. This high F-measure is due to its high precision with a small drop in recall. On the other hand, combined and k-means++ anchor boxes have high recall. But due to their overall low precision, the F-measure of these models drops below that of the original anchor boxes.

Dataset	**Anchor Box**	**Precision**	**Recall**	**F-measure**
A-small	Original	55.60	98.98	70.89
	Fundamental	**77.55**	96.32	**85.77**
	Combined	50.19	99.57	66.27
	K-means++	45.78	**99.78**	62.60
B	Original	54.51	99.09	70.25
	Fundamental	**77.63**	93.12	**84.66**
	Combined	42.90	**99.69**	64.41
	K-means++	45.51	99.54	62.35
C	Original	39.28	87.91	54.07
	Fundamental	**52.56**	82.89	**64.30**
	Combined	34.65	**89.51**	49.72
	K-means++	30.78	88.84	45.51

Table 5.3: Performance of various anchor boxes (selected optimal value in bold).

Fundamental anchor boxes have comparatively lowest number of TPs. However, due to the significantly lower number of FPs, fundamental anchor boxes have around 17.66% to 29.99% higher precision in comparison to all other anchor boxes. Compared to the original anchor boxes, fundamental anchor boxes have 13.27% to 21.95% higher precision. On the other hand, due to the slightly higher number

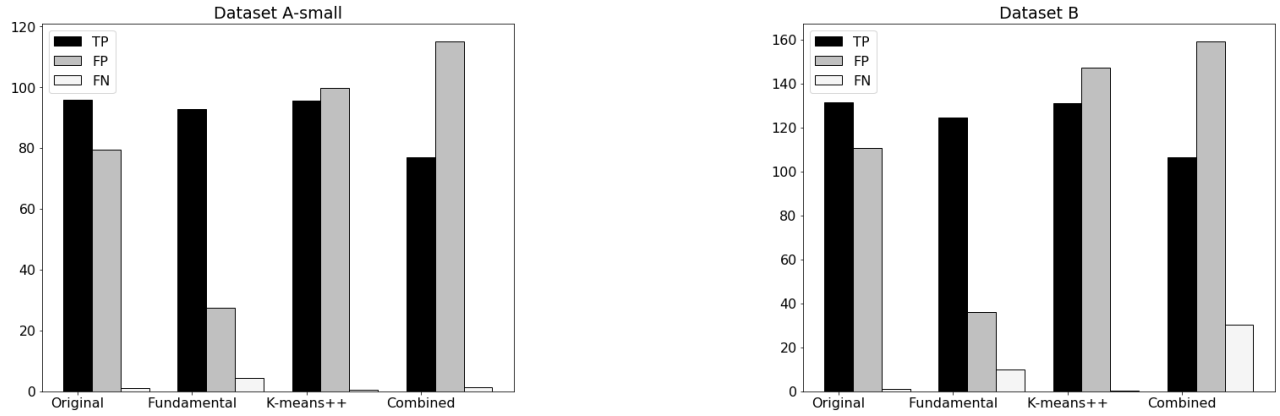


Figure 5.9: FRCNN performance with all anchor boxes.

of missed lesions (lower TP and higher FN), fundamental anchor boxes have 3.12% to 6.32% lower recall compared to other models. As the overall drop in recall was much smaller than the improvement in precision, the fundamental anchor boxes have 14.53% to 19.18% higher F-measure than all other anchor boxes. Compared to the original anchor boxes, the F-measure of the fundamental anchor boxes is 10.23% to 14.88% higher.

Combined and K-means++ anchor boxes have the highest recall due to the lowest number of missed lesions i.e. highest TPs and lowest FNs. With combined anchor boxes, this is due to the higher number of anchor boxes that fit lesions of varying sizes and shapes whereas with k-means++ anchor boxes, the high recall is due to the use of anchor boxes designed for breast lesions from GT boxes. These anchor boxes have around 1.21% to 2.96% higher recall than the other two anchor boxes. However, these anchor boxes have low precision due to the high number of FPs. Compared to the original anchor boxes, combined anchor boxes have 4.63% to 11.61% lower precision whereas k-means++ anchor boxes have 8.50% to 9.82% lower precision. Given this large drop in precision, the F-measure of these anchor boxes is also lower than that of the original anchor boxes. Specifically, combined anchor boxes have 4.91% to 8.01% and k-means++ anchor boxes have 10.52% to 11.71% lower F-measure that the original anchor boxes.

In summary, use of a small number of anchor boxes designed for breast lesions led to a lower number of FPs along with a small increase in missed lesions as there are fewer reference boxes to fit a variety of lesions with different sizes and aspect ratios. On the other hand, use of larger numbers

of anchor boxes such as combined anchor boxes improved the network's classification accuracy for challenging lesions but also caused incorrect classification of lesion-like background regions resulting in high number of FPs. Thus, the selection of the anchor boxes should be done so as to establish a good balance between the number of correct detections and FPs. Based on the performance on the modelling dataset, fundamental anchor boxes are selected as optimal.

5.3.2.2 Base Network

This section presents the evaluation of following modelling hyperparameters of the base network: training sample selection, number of training and test proposals. An important point to note here is that only the mentioned hyperparameters are changed and all other hyperparameters are maintained at their default values.

Training Sample Selection

This section details the impact of base network's training sample selection on the overall performance of the FRCNN model. All hyperparameters, except the IOU range of the positive/negative training samples, are set to the default values used in the original FRCNN model. First, performance of the FRCNN model with change in positive samples selection is presented which is followed by the performance change due to change in negative samples selection.

Positive Sample Selection: Table 5.4 shows the change in performance with various investigated thresholds for positive training samples selection. The change in number of TP, FP and FN for these thresholds in datasets A-small and B is shown in Figure 5.10. Figure A.2 in Appendix A shows this change in TP, FP and FN in dataset C. Positive samples help the classifier in differentiating between lesions and background regions. With a smaller IOU threshold range for selection of positive samples (where the lower threshold is increased and the higher threshold held constant at 1), the percentage of *lesion* present in the positive training samples is higher. This leads to an improvement in the classification of lesions as well as reduction in incorrect classification of background regions with lesion-like texture.

Thus, as seen in Figure 5.10, as the IOU range decreases, the number of FPs also decreases. However, use of a small IOU range also leads to a higher sensitivity of the model to the proportion

of background region in the proposals. Therefore, proposals containing lesions as well as background regions are incorrectly classified as background. Additionally, challenging lesions with background-like texture are also incorrectly classified as background. Thus, as the range becomes smaller, the number of missed lesions increases. After a certain point, the sensitivity of the model increases to the point where all proposals are classified as background.

Dataset	**Positive Training Sample Threshold**	**Precision**	**Recall**	**F-measure**
A-small	[0.5,1]	55.60	**98.98**	70.89
	[0.6,1]	64.82	98.55	77.60
	[0.7,1]	**81.77**	84.88	**82.96**
	[0.8,1]	0	0	0
B	[0.5,1]	54.51	**99.09**	70.25
	[0.6,1]	61.28	96.48	74.73
	[0.7,1]	**82.30**	76.58	**79.13**
	[0.8,1]	0	0	0
C	[0.5,1]	39.28	**87.91**	54.07
	[0.6,1]	45.63	81.59	**58.19**
	[0.7,1]	**60.48**	53.82	56.72
	[0.8,1]	0	0	0

Table 5.4: Performance of positive training sample selection thresholds (selected optimal value in bold).

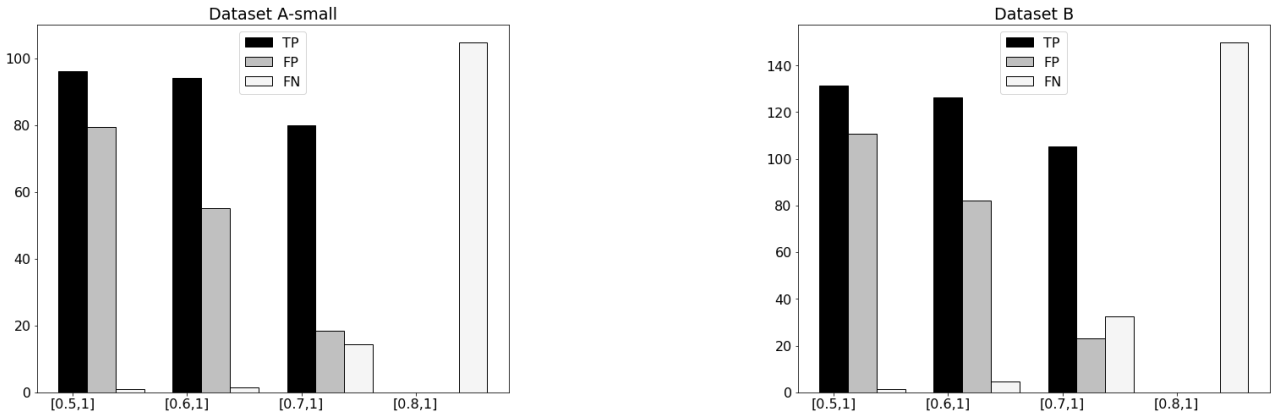


Figure 5.10: FRCNN performance with variations in base network's positive training samples.

This is seen in the case of the model trained with an IOU range of $[0.8, 1]$ for positive sample selection. This model is trained to only accept a maximum of 20% background region in the proposal for it

to be classified as lesion. Although the RPN generates high IOU proposals, the number of proposals with $IOU \geq 0.8$ is small. Thus, this threshold of $[0.8, 1]$ has no correct detections as all proposals are classified as background. To summarise, as the IOU range for positive samples becomes smaller, there is a reduction in FPs with a subsequent increase in missed lesions. Of the investigated IOU ranges, $[0.7, 1]$ has the highest F-measure. Compared to the model trained with default IOU range of $[0.5, 1]$, the model trained with IOU range $[0.7, 1]$ has 26.17% higher precision and 14.1% lower recall resulting in 12.07% higher F-measure. Based on the performance on the modelling dataset, $[0.7, 1]$ is selected as optimal since it has the highest F-measure.

In dataset C, $[0.7, 1]$ threshold has a larger drop in recall than increase in precision. This is because of the challenging lesions in this dataset. As shown in Figure 5.6, a considerable number of lesions in this dataset have background-like texture. With a smaller threshold range, a larger number of these challenging lesions are classified as background leading to an increase in missed lesions. In this dataset, threshold of $[0.6, 1]$ has the highest F-measure since it is comparatively less sensitive to background regions and background-like lesions than the model trained with $[0.7, 1]$ range. It is important to note that, in this dataset, the $[0.7, 1]$ range has second-highest F-measure of 56.72%, very close to 58.19% F-measure of $[0.6, 1]$ range based model.

Negative Sample Selection: Table 5.5 shows the change in performance with the change in negative sample selection thresholds. Figure 5.11 shows the change in the number of TP, FP and FN for the investigated negative samples thresholds in datasets A-small and B. Figure A.3 in Appendix A shows this change in dataset C. Negative training samples influence the base network's accuracy in classification of background regions in a similar manner as the positive samples influence classification of lesions. The original FRCNN range of $[0, 0.5)$ (samples with IOU 0 being used for hard example mining) is quite broad. With this range, these samples contain up to 50% of the lesion. Due to the similarity in texture of lesion and background regions, such a range causes incorrect classification. Consequently, use of a smaller range (with the upper threshold reduced to smaller value and lower threshold maintained at 0) improves the base network's classification accuracy since these negative training samples contain a much larger proportion of the background region than the lesion. Therefore, the number of FPs is lower as the range becomes smaller.

Dataset	Negative Training Sample Threshold	Precision	Recall	F-measure
A-small	**[0,0.2)**	**71.65**	**98.98**	**82.95**
	[0,0.3)	71.49	98.57	82.45
	[0,0.4)	71.49	98.81	82.67
	[0,0.5)	70.80	98.57	81.98
	[0.1,0.5)	55.60	**98.98**	70.89
B	**[0,0.2)**	**72.82**	95.26	**82.52**
	[0,0.3)	71.86	95.95	82.06
	[0,0.4)	70.87	95.57	81.32
	[0,0.5)	72.35	96.45	82.51
	[0.1,0.5)	54.51	**99.09**	70.25
C	**[0,0.2)**	40.42	88.00	55.37
	[0,0.3)	**41.85**	87.76	**56.50**
	[0,0.4)	40.41	87.68	55.27
	[0,0.5)	41.32	87.17	55.87
	[0.1,0.5)	39.28	**87.91**	54.07

Table 5.5: Performance of negative training sample selection thresholds (selected optimal value in bold).

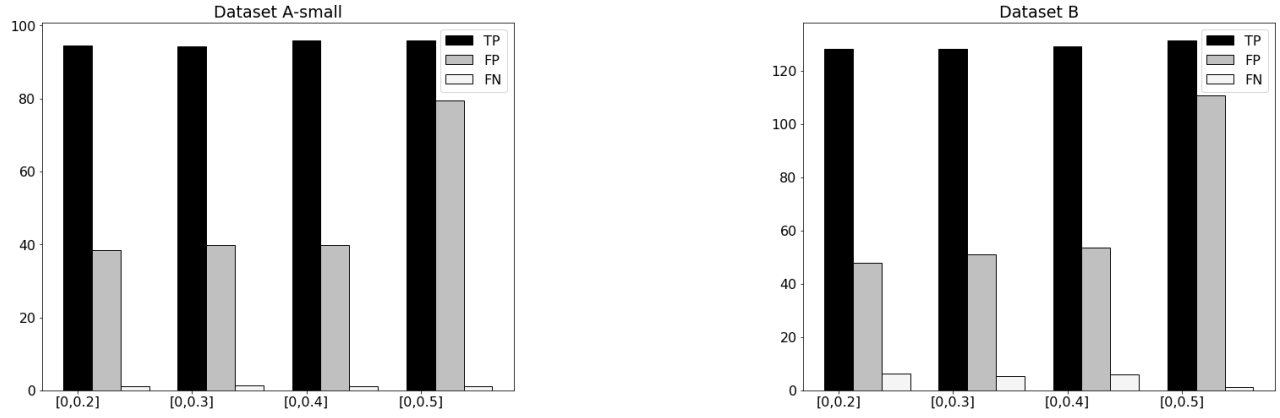


Figure 5.11: FRCNN performance with variations in base network's negative training samples

On the other hand, such a small range causes incorrect classification of lesions with background-like texture leading to an increase in missed lesions. However, the increase in missed lesions is lesser than the reduction in FPs. Therefore, as the range becomes smaller, the precision of the model increases as shown in Table 5.11. Also, recall changes to a comparatively smaller degre. Therefore, the overall performance of the model improves with the use of lower thresholds for selection of negative samples. It is important to note here that the degree of reduction in FP with change in the negative samples IOU range is smaller in comparison to that of positive samples. Based on the

best performance of $[0, 0.2)$ range on the modelling dataset with 16.05% higher precision than the original FRCNN threshold and no change in recall, it is selected as the optimal negative threshold. Similar performance is seen in dataset B. However, in dataset C, $[0, 0.3)$ range has higher precision and F-measure than the selected optimal threshold of $[0, 0.2)$. The reason for this is similar to what was seen for positive sample selection. Due to the nature of the lesions and the US images in this dataset, a slightly broader range helps reduce the FPs more than a smaller range such as $[0, 0.2)$. It is important to highlight here that the overall performance of $[0, 0.2)$ is only slightly lower than $[0, 0.3)$.

Selection of Number of Training Proposals

Table 5.6 shows the impact of the number of training proposals on the overall performance. Figure 5.12 shows the change in the number of TP, FP and FN with the change in number of training proposals in datasets A-small and B. Figure A.5 in Appendix A shows the change in TP, FP and FN with change in the number of training proposals in dataset C. In these models, all hyperparameters with the exception of the number of training proposals are set to their default value. The original FRCNN model uses 2000 training proposals that were set in accordance to the high number of objects in the natural images. However, when used for breast lesion detection in US images, use of 2000 training proposals led to a low performance due to the overall lower number of objects (lesions) in an average US image. When reduced to 1000 training proposals, classification accuracy of the model improves resulting in lower number of FPs. This is also accompanied with a small drop in the number of correctly detected lesions. But this drop in the correct detections is much smaller than the drop in the number of FPs.

Dataset	**Training Proposals**	**Precision**	**Recall**	**F-measure**
A-small	300	49.49	**99.18**	65.56
	1000	**58.88**	98.79	**73.43**
	2000	55.60	98.98	70.89
B	300	49.04	**100.00**	65.80
	1000	53.96	98.91	69.75
	2000	**54.51**	99.09	**70.25**
C	300	40.28	73.71	52.00
	1000	**40.90**	86.81	**55.52**
	2000	39.28	**87.91**	54.07

Table 5.6: Impact of various number of training proposals (selected optimal value in bold).

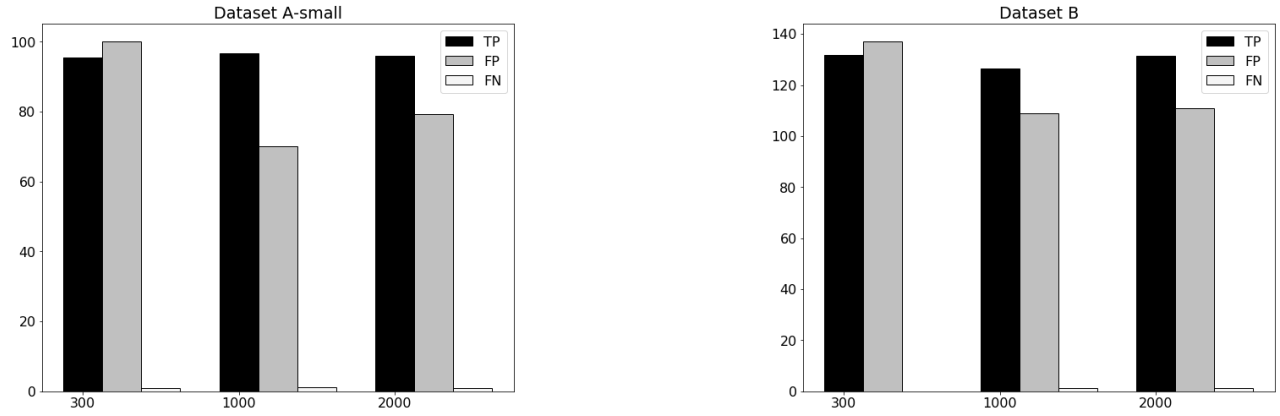


Figure 5.12: FRCNN performance with variations in number of training proposals.

When the number of training proposals is dropped even lower to 300, the model's classification accuracy drops due to underfitting. Given the challenging nature of the breast US lesions, using a small number of samples limits the variation in the training samples thus limiting its generalisation capabilities. This causes poorer classification accuracy of the background regions resulting in higher number of FPs. In the modelling dataset, use of 1000 proposals improved the precision by 3.28% than the default 2000 training proposals of the original FRCNN model with only a small drop of 0.19% in recall. Thus, due to higher precision, the 1000 proposals model has 2.54% higher F-measure than the original FRCNN model. Use of 300 training proposals led to highest recall of 99.18% which is 0.19% higher than that of the original model. However, due to a large drop in precision (6.11% lower than the original model), it has 5.33% lower F-measure than the original model. Based on the performance on the modelling dataset, 1000 training proposals is considered as optimal. In dataset B, use of 1000 training proposals had the second highest F-measure due to slightly lower precision than the original model. In dataset C, the negative impact of small variation in the training set when using 300 training proposals is amplified as evidenced through the lowest recall of the 300 training proposals based model.

Selection of Number of Test Proposals

This section details the impact of changing the number of test proposals on the overall performance of the FRCNN model. Here, all FRCNN models are trained with default modelling hyperparameter values. Table 5.7 shows the impact of 100 and 300 test proposals on the performance of original FR-CNN trained for breast lesion detection in US images. Figure 5.13 shows the change in the number of TP, FP and FN with the change in test proposals in datasets A-small and B. Figure A.5 in Appendix

A shows the change in TP, FP and FN with change in the number of test proposals in dataset C. Default number of test proposals in the original FRCNN is 300 which is in line with the number of objects present in an average natural image. However, the average number of objects (lesions) in an average US image is lesser in comparison. Therefore, when FRCNN is used for breast lesion detection, out of the 300 test proposals generated by the RPN, a considerable portion are FPs which are passed through to the base network. These FP proposals are also incorrectly classified by the base network as lesions resulting in FP detections in the final output of the model.

Dataset	Training Proposals	Precision	Recall	F-measure
A-small	**100**	**59.95**	98.80	**74.2**5
	300	55.60	**98.98**	70.89
B	**100**	**58.68**	98.33	**73.44**
	300	54.51	**99.09**	70.25
C	**100**	**42.84**	86.08	**57.04**
	300	39.28	**87.91**	54.07

Table 5.7: Impact of number of test proposals on the performance of FRCNN (selected optimal value in bold).

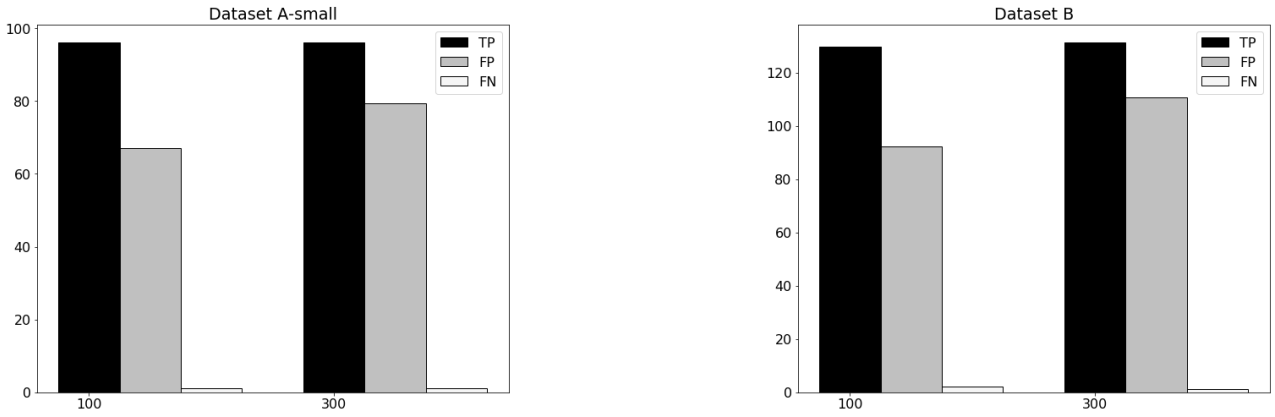


Figure 5.13: FRCNN performance with variations in number of test proposals.

The RPN typically assigns low classification scores to FP proposals. The higher scoring proposals generally consist of high IOU TP proposals. Therefore, when the number of test proposals is reduced to 100, only the 100 top-scoring (mostly TP) test proposals are sent through to the base network. Thus, as the majority of the FP proposals are discarded at the RPN and not passed through to the base network, the total number of FP detections in the final output of the model is also reduced.

On the other hand, due to relatively poorer classification accuracy of the RPN, proposals containing challenging lesions with background-like texture are also scored low. When all 300 test proposals are sent through to the base network, they are correctly classified as lesions by the base network due to its higher classification accuracy. However, when only the top 100 proposals are selected, these challenging lesions are discarded at the RPN and thus missed by the network. Therefore, the number of missed lesions (lower TP and higher FN) is higher when 100 test proposals are used. However, the drop in the number of missed lesions is much smaller than the reduction in FPs as shown in Figure 5.13. Therefore, the model using 100 test proposals has 3.56% to 4.35% higher precision and 0.18% to 1.83% lower recall than the model using 300 test proposals. Thus, use of 100 test proposals resulted in 2.97% to 3.36% higher F-measure than the default 300 test proposals. Based on its performance in the modelling dataset, 100 test proposals is considered as optimal.

5.3.3 Breast Lesion Detection with Adapted FRCNN

Based on the investigation presented in Section 5.3.2, following optimal values of the evaluated hyperparameters were selected to design the adapted FRCNN model:

- Fundamental Anchor Boxes with scale 1 and aspect ratio $\{1:1, 1:2, 2:1\}$
- $[0.7, 1]$ threshold for positive sample selection for base network training
- $[0, 0.2)$ threshold for negative sample selection for base network training
- 1000 training proposals
- 100 test proposals

As previously mentioned, these optimal values were selected using dataset A-small as the modelling dataset. Reproducibility of these values, using a larger dataset (dataset A) as the modelling dataset, is presented in Section 8.1 of Chapter 8. Table 5.8 shows the performance of the original and adapted FRCNN models in all datasets. Overall, the adapted FRCNN model outperforms the original FRCNN with higher precision and relatively lower recall. In the modelling dataset, the adapted FRCNN model has 19.85% higher precision with only 4.62% drop in recall leading to a 12.75% raise in F-measure. Similar performance change is also seen in the external test sets. Furthermore, irrespective of the anchor boxes, use of the selected optimal values of hyperparameters leads to an improvement

in performance of the model. For instance, FRCNN models using original and k-means++ anchor boxes and optimal values for training sample selection, number of training and test proposals show higher performance in all datasets. Figures 5.14 and 5.15 show performance of these models in datasets A-small and B. Figures A.7 and A.6 in Appendix A show the performance of these models in dataset C.

Dataset	**Original FRCNN**			**Adapted FRCNN**		
	P	**R**	**F**	**P**	**R**	**F**
A-small	55.60	**98.98**	70.89	**75.45**	94.36	**83.64**
B	54.51	**99.09**	70.25	**76.39**	85.37	**80.79**
C	39.28	**87.91**	54.07	**44.70**	70.63	**54.70**

Table 5.8: Precision (P), Recall(R) and F-measure (F) of original and adapted FRCNN.

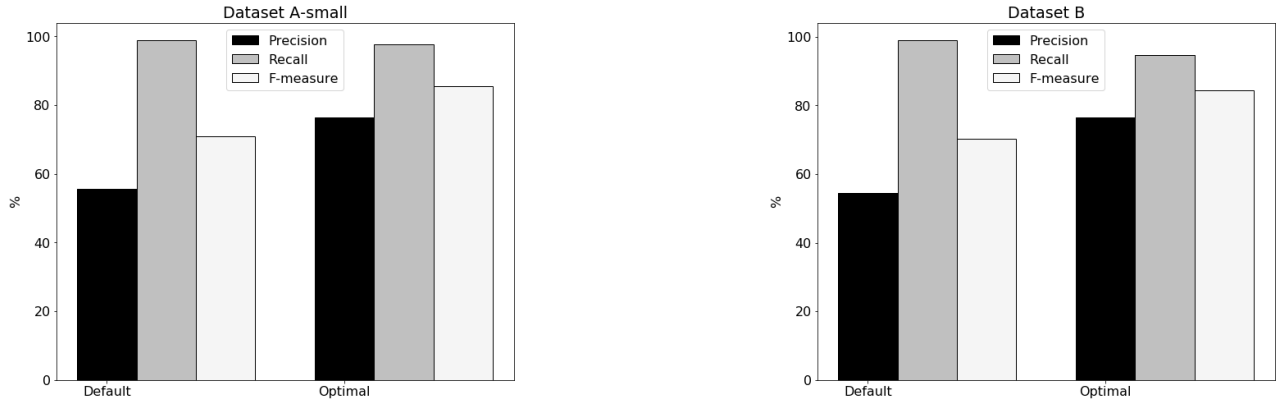


Figure 5.14: Original anchor boxes with default and optimal modelling hyperparameters.

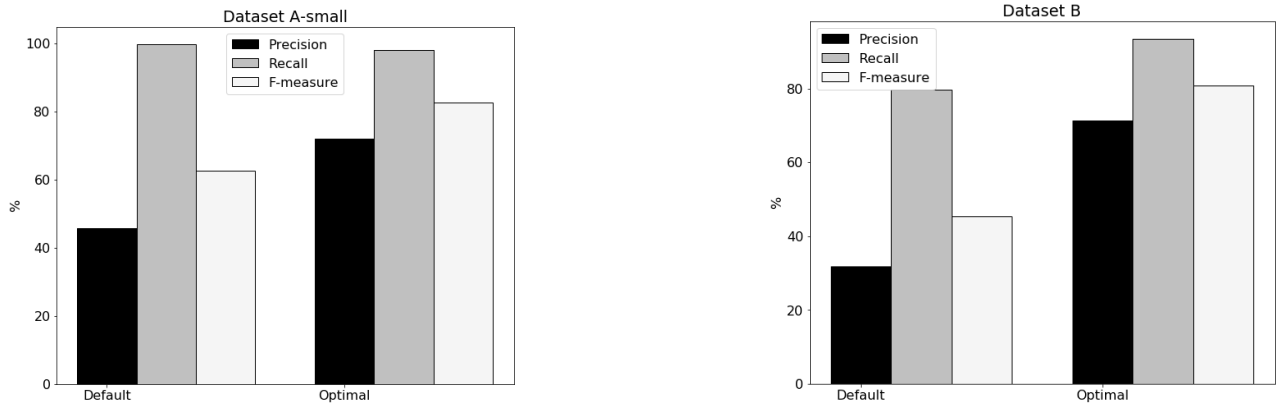


Figure 5.15: K-means++ anchor boxes with default and optimal modelling hyperparameters.

Figure 5.18 shows the number of TP, FP and FN of the original and adapted FRCNN models in

datasets A-small and B. Figure A.9 in Appendix A shows the number of TP, FP and FN in original and adapted FRCNN models in dataset C. As seen in this figure, the adapted FRCNN model has a considerably lower number of FPs than the original FRCNN which is responsible for its higher precision. Overall, the reduction in total FPs is largely due to the reduction in additional boxes. Figure 5.19 shows sample cases of FP reduction in the adapted FRCNN model. However, compared to the original FRCNN model, the adapted FRCNN model has a higher number of missed lesions (lower TPs and higher FNs) resulting in its lower recall. Figure 5.20 shows an example of lesions that were detected by the original model but missed by the adapted FRCNN model. Additionally, TP detections of the original FRCNN model have overall higher IOU than those of the adapted FRCNN model as shown in Figure 5.16 (Figure A.8 in Appendix A shows the IOU distribution of TP detections of original and adapted FRCNN models in dataset C). Figure 5.17 shows sample TPs detected by the original and adapted FRCNN model. Overall, due to considerably lower number of FPs and a small drop in the number of missed lesions, the adapted FRCNN model has higher precision and F-measure than the original FRCNN.

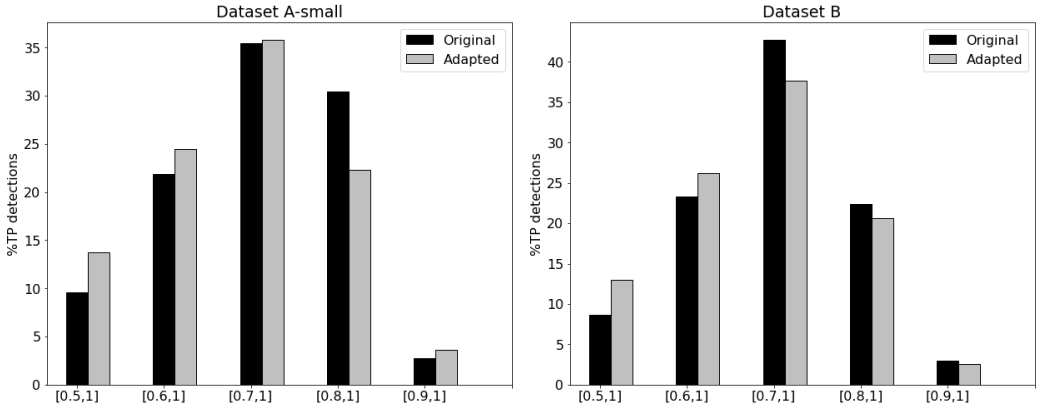


Figure 5.16: IOU distribution of TPs in original and adapted FRCNN in all datasets.

Both original and adapted FRCNN models have poorer performance in dataset C. This is because of the challenging images in this dataset. Like other datasets, the adapted FRCNN model outperformed the original FRCNN model in this dataset. Both models have high quality TP cases as shown in Figure 5.16(c). Sample TP detections of single and multiple lesions from this dataset are shown in Figures 5.21 and 5.22. However, the number of TPs detected by either model is relatively lower than the other datasets as this dataset contains large number of lesions with texture very close to that of

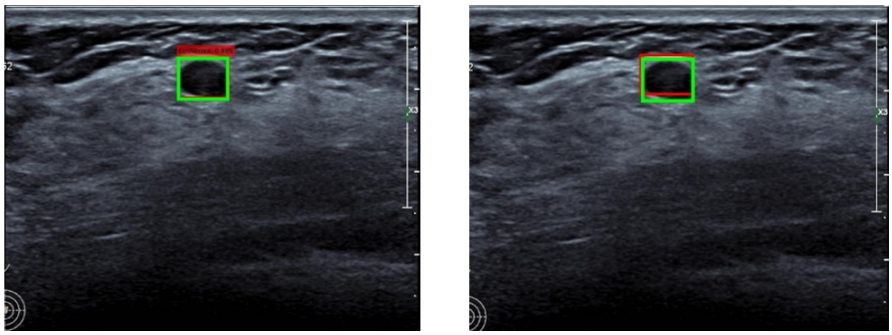


Figure 5.17: Sample TP in original FRCNN (left) and adapted FRCNN (right) in Datasets A and B. Green box: ground truth and red box: output boxes (Black marker region removed for confidentiality reason).

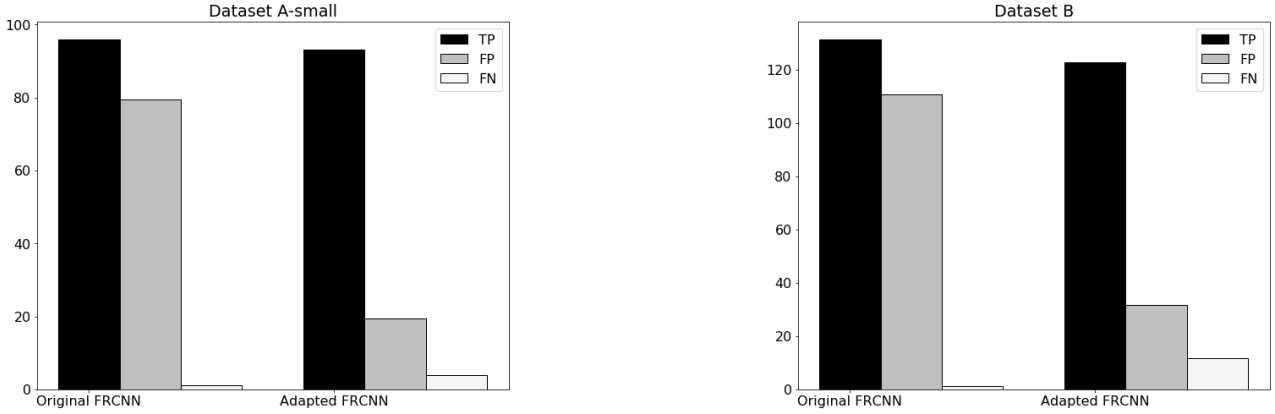


Figure 5.18: Number of TP, FP and FN detections in original and adapted FRCNN models in datasets A-small and B.

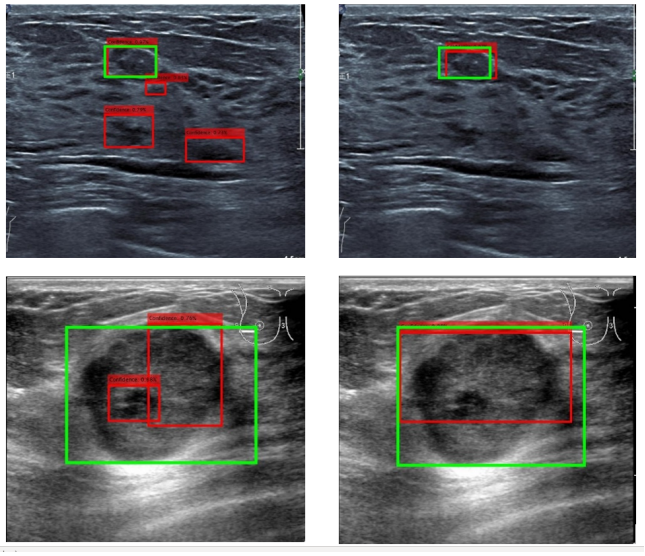


Figure 5.19: Sample FPs in Original Faster R-CNN (left) eliminated in Optimal Faster R-CNN (right) in Datasets A and B. Green box represents ground truth and red box represents output boxes (Black marker region removed for confidentiality reason).

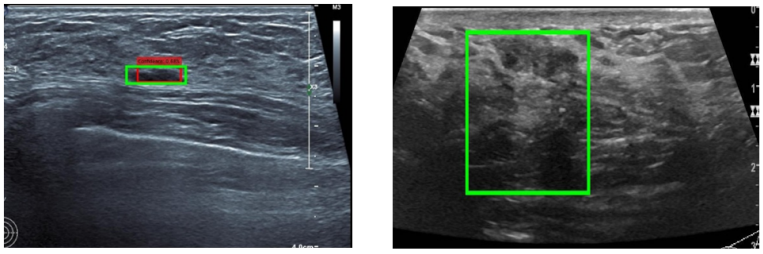


Figure 5.20: Lesion missed by adapted FRCNN but detected by original FRCNN (left) and lesion missed by both original and adapted FRCNN (right). Green box: ground truth and red box: output boxes. (Black marker region removed for confidentiality reason).

the background region as shown in Figure 5.26.

As seen with other datasets, adapted FRCNN successfully reduces FP detections in this dataset. An example of this is given in Figure 5.24. However, due to the nature of the images in this dataset, the number of FPs output by the adapted FRCNN is objectively high. Figure 5.23 shows an example case where the adapted FRCNN model reduced the number of additional boxes (FPs) in comparison to the original FRCNN model, but did not eliminate them entirely. In GT boxes that contain two distinct lesion-like regions as shown in Figure 5.25, both models detect these regions as two separate lesions instead of covering the entire GT box with a single detection thereby resulting in FPs.

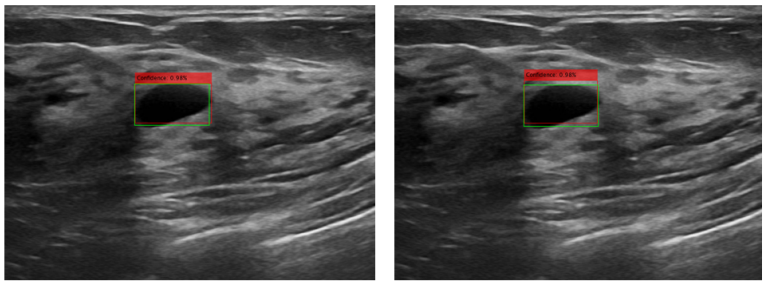


Figure 5.21: TP (single lesion) in original FRCNN (left) and adapted FRCNN (right) in Dataset C. Green box: ground truth and red box: output boxes).

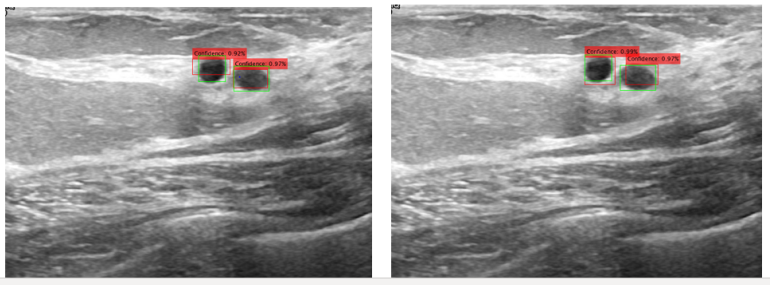


Figure 5.22: TP (multiple lesions) in original FRCNN (left) and adapted FRCNN (right) in Dataset C. (Green box: ground truth and red box: output boxes).

Therefore, the important drawbacks in literature are successfully addressed through an investigation of the impact of individual hyperparameters on the overall performance on a large dataset. The adapted FRCNN designed through this investigation outperformed the original FRCNN through a substantial reduction in FP detections therefore successfully adapting the model for breast lesion detection in US images. This work has been published to aid the researchers' understanding of the

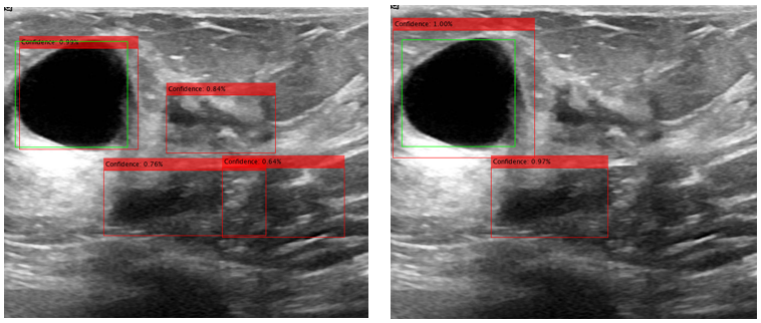


Figure 5.23: FPs in lesion-like regions of the background region in original FRCNN (left) which were reduced in adapted FRCNN (right) in Dataset C. Green box: ground truth and red box: output boxes.

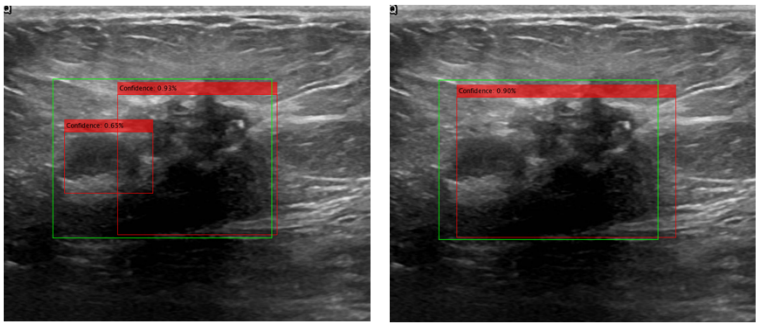


Figure 5.24: Multiple FPs in large lesions generated by original FRCNN (left) replaced by single TP detection by adapted FRCNN (right) in Dataset C. Green box: ground truth and red box: output boxes.

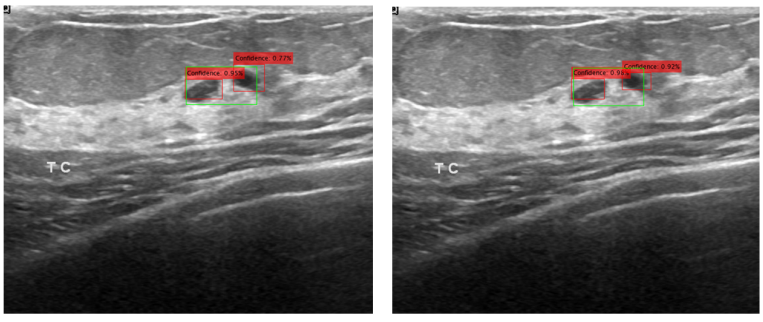


Figure 5.25: Multiple FPs in small lesion generated by original FRCNN (left) replaced by single detection by adapted FRCNN (right) in Dataset C. Green box: ground truth and red box: output boxes.

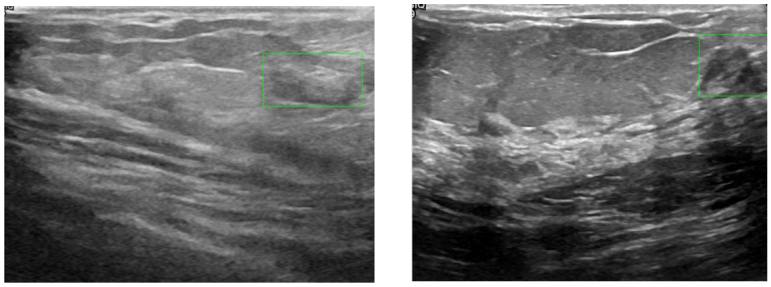


Figure 5.26: FN in original and adapted FRCNN in Dataset C. Green box: ground truth.

impact of these hyperparameters and to appropriately adapt the FRCNN model for their application.

Despite the high performance of the adapted FRCNN model, the number of FPs generated by this model is relatively significant. The common FP detections of this model are caused due to two main reasons. Firstly, proposals covering lesion-like background regions are incorrectly assigned high classification scores by both stages of the model. The common type of FP detections caused due to poor classification accuracy of the model is additional boxes . Secondly, the post-processing mechanism used in the model (both at the RPN and base network) incorrectly filters out high IOU TP proposals and retains FP proposals. Common FP issue cases caused due to NMS are single FP and multiple FP detections.

In case of single FP detections, the model outputs a single FP detection with either no overlap with the GT (IOU 0) or low IOU with GT (IOU < 0.5). Figure 5.27 shows sample single low IOU FP detection output by the adapted FRCNN model. Multiple FP detections are cases where the model generates multiple FP detections covering background regions (additional boxes) and/or small regions of a large lesion (low IOU FP). Multiple FP detections are occasionally accompanied by TP detection of the lesion. Figure 5.23 shows examples of multiple FP detections output by the adapted FRCNN model. We address the low classification accuracy of the adapted FRCNN model in this chapter in Sections 5.3.4 and 5.3.5. We propose novel U-Detect and U-DetectH methods in Chapters 6 and 7, respectively, to reduce FP detections resulting from improper filtering of proposals by the NMS.

5.3.4 Backbone Network Selection for Adapted FRCNN

In this section, FPs caused due to poor classification accuracy are addressed through an investigation of various state-of-the-art classification models as the backbone of the adapted FRCNN model. The

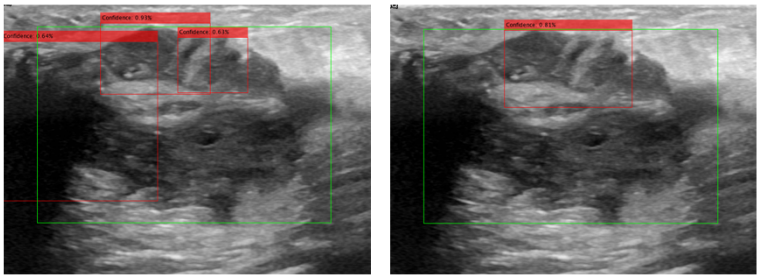


Figure 5.27: Multiple FPs generated by original FRCNN model (left) reduced to single FP in the adapted FRCNN model (right)

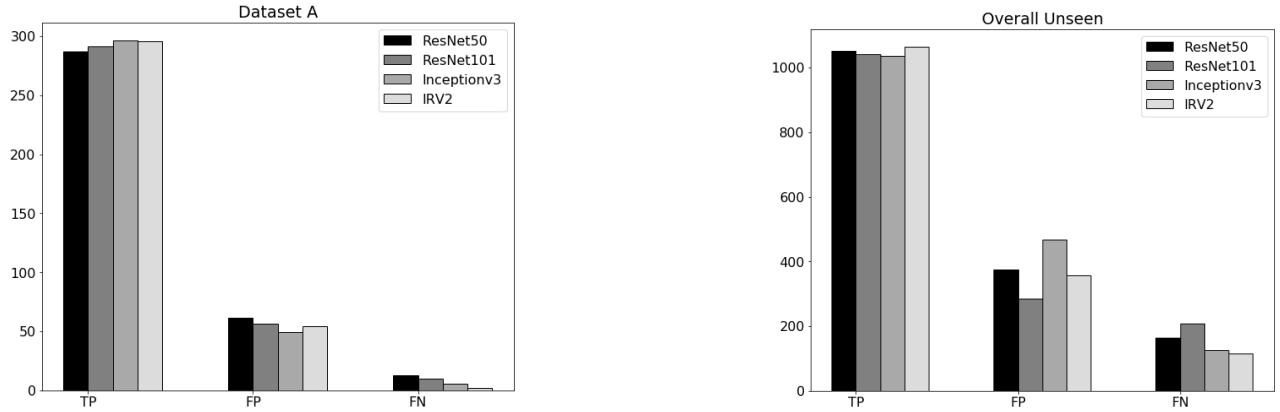
models presented in Sections 5.3.1 and 5.3.2 were trained on dataset A-small. At this point, a larger dataset (described in the Section 4.1) was made available for this research. Therefore, all analysis and experiments from this point on are conducted on this larger dataset. Particularly, dataset A is used as modelling dataset and datasets B to E are used as unseen external test sets as described in Section 4.1 Chapter 4. Comparison of the performance of selected optimal values as well as adapted FRCNN using dataset A as modelling and datasets B to E as external test sets is described in Chapter 8.

The adapted FRCNN model presented thus far uses VGG16 as the backbone which is the default backbone network of FRCNN. Table 5.9 shows the performance of the state-of-the-art classification networks used as backbone in the adapted FRCNN. Figure 5.28 shows the change in the number of TP, FP and FN in datasets A and all unseen test sets (datasets B to E), respectively. Overall, Inception-ResNet-v2 (IRV2) has the best performance. The IRV2 model has the highest number of detected lesions in comparison to all other models whereas VGG16 has the lowest. In terms of FPs, ResNet101 has the lowest number of both types of FPs, closely followed by IRV2. Inception-v3 has the highest number of FPs. Due to the highest number of detected lesions as well as low number of FPs, IRV2 has the highest precision and recall overall which results in its highest F-measure.

The IRV2 model also has the lowest number of FNs. Thus, this model has the highest recall given its high TP and low FN. The highest FN was found in the VGG16 model. In the overall external test sets, ResNet101 has a higher number of FNs than ResNet50. This higher FNs is caused due to high FNs in dataset C alone; in all other external datasets (datasets B, D and E), ResNet101 has a lower number of FNs than ResNet50. Dataset C contains a significant number of challenging small lesions. Critical textural information of such lesions are lost in the deep layers of ResNet101. In networks

Dataset	Adapted FRCNN Model	Precision	Recall	F-measure
A	VGG16	78.44	94.49	85.67
	ResNet50	82.40	95.67	88.50
	ResNet101	83.74	96.74	89.77
	Inception-v3	**85.71**	98.19	**91.53**
	Inception-ResNet-v2	84.56	**99.39**	91.36
B	VGG16	85.72	94.42	89.83
	ResNet50	87.70	96.32	91.80
	ResNet101	90.83	97.69	94.13
	Inception-v3	90.65	98.85	94.53
	Inception-ResNet-v2	**92.04**	**99.28**	**95.51**
C	VGG16	58.41	69.51	63.34
	ResNet50	63.40	**83.34**	71.89
	ResNet101	**71.54**	73.44	71.93
	Inception-v3	66.49	81.25	68.68
	Inception-ResNet-v2	58.69	82.95	**72.90**
D	VGG16	76.38	78.50	77.40
	ResNet50	77.34	83.40	80.10
	ResNet101	**78.33**	84.57	81.30
	Inception-v3	69.08	88.65	77.58
	Inception-ResNet-v2	74.55	**93.14**	**82.69**
E	VGG16	79.69	95.13	86.70
	ResNet50	86.29	97.65	91.60
	ResNet101	**89.44**	97.54	93.30
	Inception-v3	84.44	**99.21**	91.22
	Inception-ResNet-v2	88.89	99.08	**93.68**

Table 5.9: Performance of various backbone networks in adapted FRCNN.



(a) (b)

Figure 5.28: Number of TP, FP and FN for various backbone networks in adapted FRCNN model.

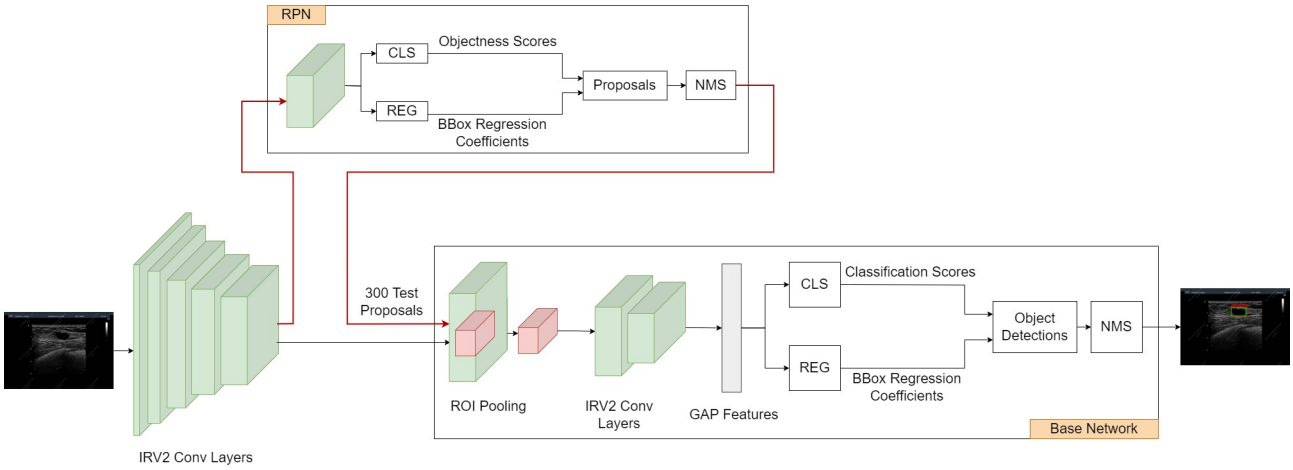
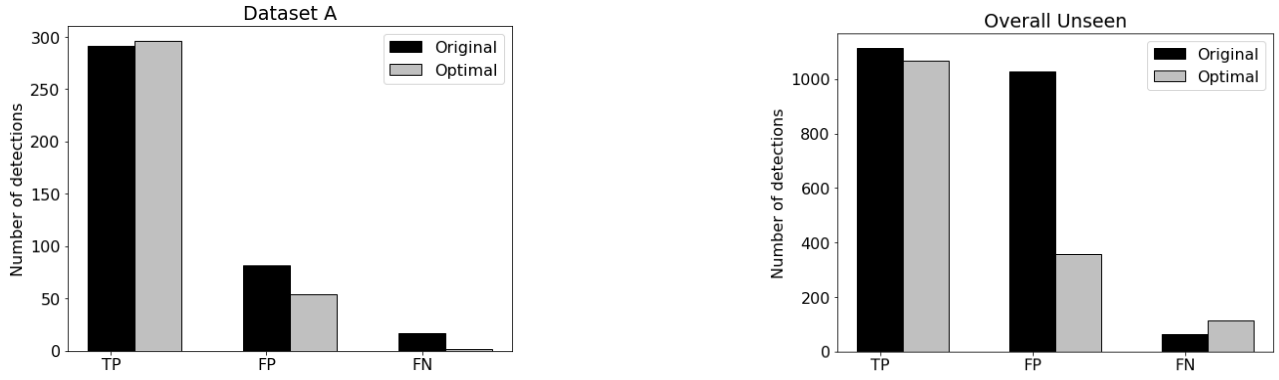


Figure 5.29: Architecture of adapted FRCNN with IRV2 backbone.

such as IRV2, the combination of inception and residual blocks ensures that such features are not lost with increased network depth. This high performance of IRV2 is due to its high classification accuracy resulting from the use of a combination of inception and residual blocks that helps preserve the important textural features in an image in deeper layers of the network. Compared to VGG16, IRV2 is more capable in correctly classifying proposals containing lesions as well as those containing background lesion-like regions. Therefore, the adapted model with IRV2 has 6.12% and 6.32% higher precision as well as 4.9% and 4.86% higher recall in modelling and overall external test sets, respectively. Thus, the adapted FRCNN model with IRV2 backbone has 5.69% and 5.68% higher F-measure than VGG16 in modelling and overall external test sets, respectively. Architecture of the adapted FRCNN model with IRV2 backbone is shown in Figure 5.29.

The change in number of TP, FP and FN of original FRCNN with VGG16 backbone and adapted FRCNN model with IRV2 backbone is shown in Figure 5.30. In comparison to the original FRCNN, the IRV2 model has 6.11% to 22.94% higher precision due to lower number of FPs due to higher number of missed lesions (lower number of TPs and higher number of FNs). Due to the comparatively higher precision and a relatively small drop in recall, the adapted FRCNN model using IRV2 has 5.04% to 14.78% higher in F-measure than the original FRCNN model in modelling and overall unseen test sets, respectively.



(a) (b)

Figure 5.30: Number of TP, FP and FN of original and adapted FRCNN models.

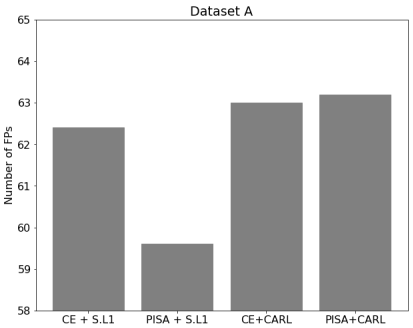
5.3.5 Training Loss Selection for RPN of Adapted FRCNN Model

This section details the impact of PISA and CARL training loss on the classification accuracy of the adapted FRCNN model. This study is performed on the adapted FRCNN model with ResNet50 as the backbone. Table 5.10 shows the performance of all losses. Figures 5.31, 5.32 and 5.33 show the change in the number of TP, FP and FN with PISA+CARL loss variations. The default training loss used in FRCNN is CE + Smooth L1 (see Table 5.9). Overall, use of PISA + Smooth L1 has the highest performance due to its high recall and relatively smaller drop in precision.

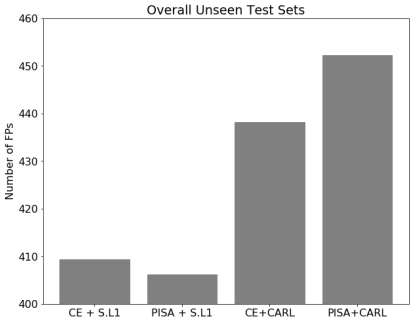
(a) (b)

Figure 5.31: TP of adapted FRCNN model using various training losses.

Like PISA + Smooth L1, CE + CARL and PISA + CARL losses also have higher recall than the default loss. This is due to the lower number of FNs despite lower number of TPs compared to the default loss. In particular, PISA + Smooth L1 loss has 0.19% and 0.16% higher recall than the default loss in modelling and overall external test sets, respectively. However, these losses have a

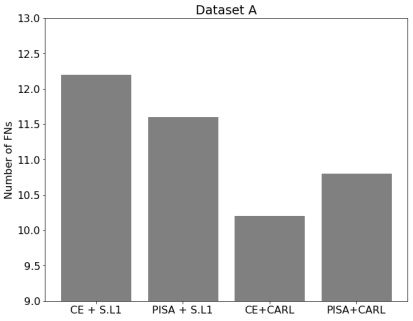


(a)

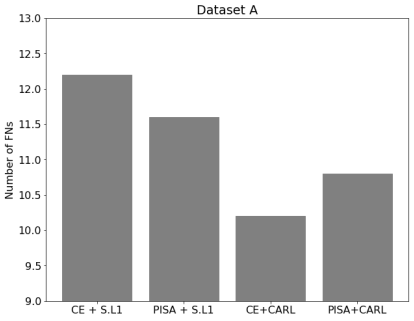


(b)

Figure 5.32: FP of adapted FRCNN model using various training losses.



(a)



(b)

Figure 5.33: FN of adapted FRCNN model using various training losses.

Dataset	Training Loss	Precision	Recall	F-measure
A	PISA + Smooth L1	**82.96**	96.14	**89.04**
	CE + CARL	82.18	**96.61**	88.79
	PISA + CARL	82.07	96.40	88.64
Overall External Test Sets	PISA + Smooth L1	**72.46**	88.07	**79.49**
	CE + CARL	70.96	88.07	78.56
	PISA + CARL	70.27	**88.33**	78.24
B	PISA + Smooth L1	86.82	97.48	91.84
	CE + CARL	**87.15**	**97.53**	**92.04**
	PISA + CARL	86.09	97.16	90.91
C	PISA + Smooth L1	**60.54**	**85.62**	**70.88**
	CE + CARL	59.84	85.22	70.23
	PISA + CARL	58.84	85.48	69.62
D	PISA + Smooth L1	**77.62**	84.58	**80.92**
	CE + CARL	75.03	84.28	79.85
	PISA + CARL	74.69	**85.30**	79.61
E	PISA + Smooth L1	**86.93**	97.80	**92.03**
	CE + CARL	84.22	97.22	90.76
	PISA + CARL	84.93	**98.30**	91.11

Table 5.10: Impact of PISA and CARL on adapted FRCNN (ResNet50 backbone).

higher number of low IOU FPs than the default loss as they convert the TP detections of the default loss to low IOU FP detections. Also, CE + CARL and PISA + CARL losses have a higher number of additional boxes. Thus, due to lower number of TPs and higher number of FPs, both CE + CARL and PISA + CARL losses have lower precision than the default loss.

On the other hand, PISA + CARL loss has the lowest number of additional boxes, including in comparison to the default loss. Thus, despite a higher number of low IOU FPs, PISA + Smooth L1 loss has a lower number of total FPs than the default loss. In the modelling dataset, PISA + Smooth L1 loss has 0.64% higher precision than the default loss due to the lower number of FPs (despite the lower number of TPs). However, in overall external datasets, the drop in the number of TPs was larger than the drop in FPs. Therefore, here, the PISA + Smooth L1 loss has 0.08% lower precision than the default losses. Thus, due to the relatively higher recall, PISA + Smooth L1 loss has 0.47% and 0.11% higher F-measure than the default loss, also outperforming CE + CARL and PISA + CARL losses.

In summary, the three PISA and CARL losses improved the classification accuracy for challenging

lesions. However, use of CARL loss in the regression branch had a negative impact on the classification accuracy of lesion-like background regions. Although PISA + Smooth L1 loss improved the overall performance of the adapted FRCNN model, the degree of improvement was marginal. Therefore, this loss is evaluated further with the aim to reduce FPs without any negative impact on the recall. In particular, PISA negative loss that is responsible for the weight assigned to training loss of hard negative samples (FPs) is studied further and presented in the remainder of this section.

PISA negative loss: Table 5.11 shows the performance of PISA negative and its variants, PISA gamma 1 and PISA gamma 1.5. An important point to note here is that these PISA negative losses are used in the classification branch and smooth L1 loss used in the regression branch. This analysis is performed on a single fold. Figures 5.34, 5.35 and 5.36 show the change of in the number of TP, FP and FN with the use of PISA negative and its variants. Overall, PISA negative has the highest performance, outperforming all losses including PISA + Smooth L1 (see Table 5.10) and default loss (see Table 5.9). This high performance of PISA negative loss is due to its high recall.

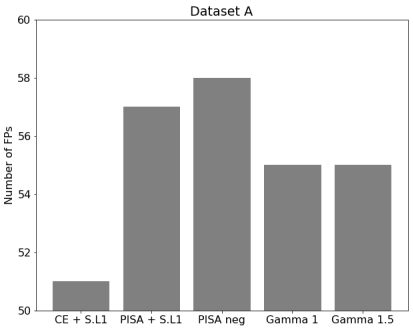
(a) (b)

Figure 5.34: TP of adapted FRCNN model using PISA negative training losses and its variants.

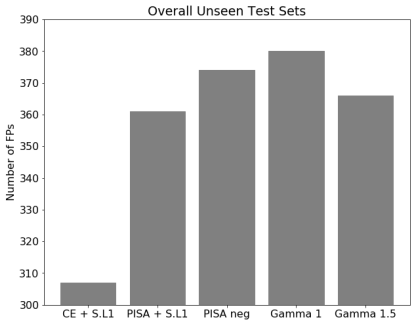
Compared to PISA + Smooth L1, PISA negative has a higher number of detected lesions (higher TPs and lower FNs). Thus, PISA negative has 0.07% and 1.34% higher recall in modelling and overall external test sets, respectively. However, the number of FPs in PISA negative loss is higher than that of PISA + Smooth L1. Thus, despite the higher number of TPs, PISA negative has 0.05% and 0.24% lower precision than the PISA + Smooth L1 loss in modelling and overall external test sets, respectively. Since the increase in recall is higher than the drop in precision, PISA negative loss has 0.43% higher F-measure than that of PISA + Smooth L1 in overall external tests. In the modelling dataset,

Dataset	Training Loss	Precision	Recall	F-measure
A	Default	**84.91**	93.79	89.13
	PISA + Smooth L1	83.33	95.00	**88.79**
	PISA negative	83.29	95.07	**88.79**
	PISA negative 1	84.06	**96.35**	89.78
	PISA negative 1.5	83.97	94.74	89.03
Overall External Test Sets	Default	**77.09**	83.78	80.30
	PISA + Smooth L1	74.25	86.25	79.80
	PISA negative	74.01	**87.58**	**80.23**
	PISA negative 1	73.83	87.15	79.94
	PISA negative 1.5	74.33	87.10	80.21
B	Default	86.39	94.07	90.07
	PISA + Smooth L1	85.62	**97.76**	91.29
	PISA negative	87.16	95.56	**95.17**
	PISA negative 1	**87.33**	97.04	91.93
	PISA negative 1.5	86.75	**97.76**	91.93
C	Default	**67.51**	83.60	74.70
	PISA + Smooth L1	63.37	82.77	71.88
	PISA negative	63.19	83.93	**72.10**
	PISA negative 1	64.92	**86.46**	74.16
	PISA negative 1.5	64.71	85.56	73.68
D	Default	**81.47**	77.56	79.47
	PISA + Smooth L1	77.71	**82.60**	80.08
	PISA negative	78.48	81.43	**81.83**
	PISA negative 1	74.44	81.76	79.01
	PISA negative 1.5	77.80	82.16	79.92
E	Default	88.02	95.45	91.59
	PISA + Smooth L1	**90.96**	97.42	**94.08**
	PISA negative	86.55	**98.01**	91.93
	PISA negative 1	85.88	97.99	91.54
	PISA negative 1.5	86.55	**98.01**	91.93

Table 5.11: Impact of PISA negative and its variation in the RPN of adapted FRCNN (ResNet50 backbone).

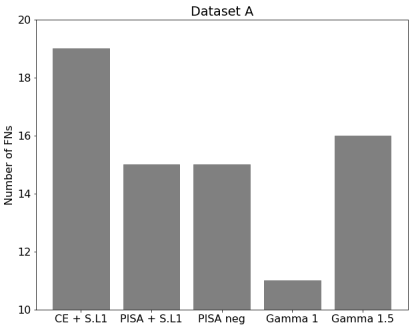


(a)

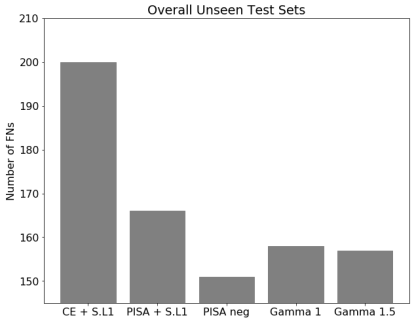


(b)

Figure 5.35: FP of adapted FRCNN model using PISA negative training losses and its variants.



(a)



(b)

Figure 5.36: FN of adapted FRCNN model using PISA negative training losses and its variants.

both losses have the same F-measure as the change in precision and recall is similar. Therefore, use of PISA negative improves classification of background-like lesions but at the expense of incorrect classification of lesion-like background regions. Increasing γ (defined in Equation 5.1 in Section 5.2.2) of PISA negative loss to 1 (PISA gamma 1 loss) led to a drop in performance in all metrics. However, increasing it further to 1.5 (PISA gamma 1.5 loss) improved precision through reduction in low IOU FPs. But along with the reduction in FPs, there was also a drop in the number of correctly detected lesions leading to lower recall. Thus, increasing the importance assigned to hard negative samples (γ) led to reduction of FPs at the expense of correctly detected lesions.

In summary, overall, all investigated losses had a lower number of missed lesions in comparison to the default losses. Only PISA + Smooth L1 loss had a lower number of FPs than the default loss. Therefore, improving training loss assigned to the classification branch improved the overall classification accuracy of the model for challenging lesions. Using only PISA negative loss and its variants led to a further reduction in FNs while negatively impacting FPs. Irrespective of the impact, change in the overall performance of the adapted FRCNN model with the various losses is negligible.

5.3.6 Comparison with State-of-the-Art Object Detection Methods

In this section, the adapted FRCNN with IRV2 is compared to object detections methods developed for natural images as well as those developed for breast lesion detection in 2D US images. For natural images' methods, both 1-stage and 2-stage detectors were evaluated. YOLOv2, and SSD are the aforementioned 1-stage detectors whereas Mask R-CNN is the 2-stage detector. Two breast lesion detectors [3, 4] were also evaluated. Overall, adapted FRCNN outperforms all evaluated detectors. Remainder of this section is organised as follows: first, performance of the adapted FRCNN is compared to that of the detectors developed for natural images in Section 5.3.6.1 followed by its comparison to the breast lesion detectors in Section 5.3.6.2. Finally, computation time for all evaluated models is described.

5.3.6.1 Object Detectors Developed for Natural Images

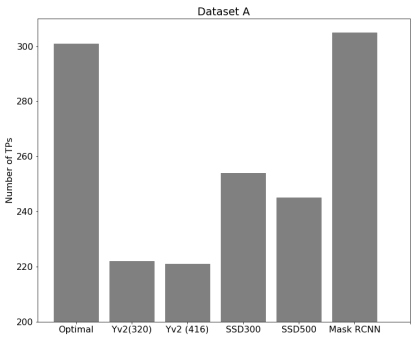
Table 5.12 shows the performance of all the object detectors developed for natural images in comparison with the original and adapted FRCNN. Figures 5.37, 5.38 and 5.39 show the number of TP, FP and FN of these detectors. Overall, the adapted FRCNN model has the highest performance.

Analysis of the performance of the individual detectors and their comparison to the adapted FRCNN model is as follows.

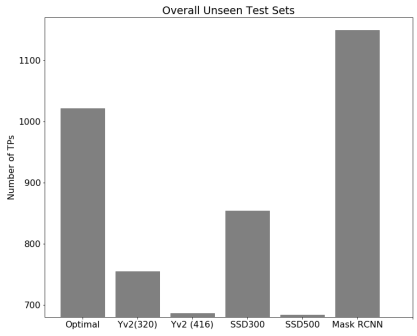
Dataset	Model	Precision	Recall	F-measure
A	Original FRCNN	78.45	96.23	86.32
	Adapted FRCNN	**84.56**	**99.39**	**91.36**
	Mask R-CNN	37.70	97.44	54.37
	YOLOv2 (320)	73.03	70.48	71.73
	YOLOv2 (416)	77.54	70.38	73.79
	SSD300	80.89	82.20	81.54
	SSD500	81.67	78.03	79.80
B	Original FRCNN	73.34	**99.85**	84.50
	Adapted FRCNN	**92.04**	99.28	**95.51**
	Mask R-CNN	41.64	98.56	58.55
	YOLOv2 (320)	82.79	69.99	75.66
	YOLOv2 (416)	82.93	73.91	78.16
	SSD300	82.98	86.03	84.48
	SSD500	83.06	76.87	79.84
C	Original FRCNN	35.60	90.20	51.02
	Adapted FRCNN	**58.69**	82.95	**72.90**
	Mask R-CNN	18.37	**94.02**	30.74
	YOLOv2 (320)	44.02	59.68	50.67
	YOLOv2 (416)	51.41	57.56	54.31
	SSD300	58.12	64.73	61.25
	SSD500	48.50	47.62	48.05
D	Original FRCNN	64.80	**95.58**	77.22
	Adapted FRCNN	**74.55**	93.14	**82.69**
	Mask R-CNN	37.54	88.14	52.66
	YOLOv2 (320)	64.79	53.00	58.31
	YOLOv2 (416)	75.68	42.58	54.50
	SSD300	84.13	59.66	69.81
	SSD500	85.04	54.10	66.13
E	Original FRCNN	70.92	**99.47**	82.35
	Adapted FRCNN	**88.89**	99.08	**93.68**
	Mask R-CNN	21.90	98.74	35.84
	YOLOv2 (320)	75.61	79.49	77.50
	YOLOv2 (416)	87.10	67.50	76.06
	SSD300	77.25	86.00	81.39
	SSD500	76.92	52.98	62.75

Table 5.12: Performance of adapted FRCNN in comparison to state-of-the-art detectors developed for object detection in natural images.

Of the evaluated YOLOv2 models, YOLOv2 (416) has better overall performance than YOLOv2

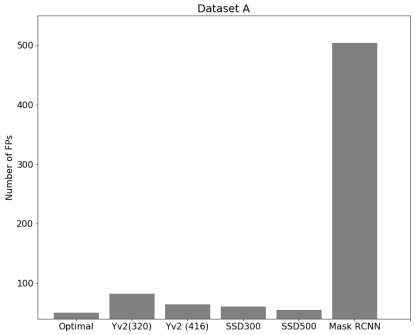


(a)

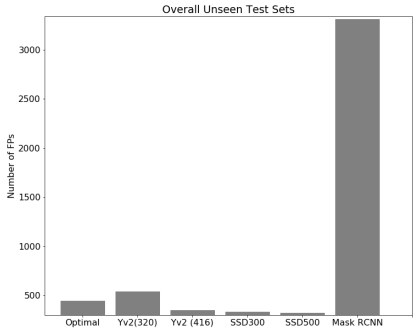


(b)

Figure 5.37: TP of detectors developed for object detection in natural images. Yv2(20): YOLOv2 (320), Yv2(416): YOLOv2(416).

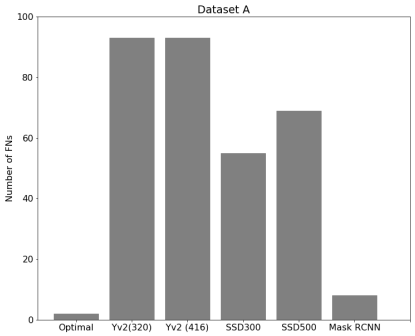


(a)

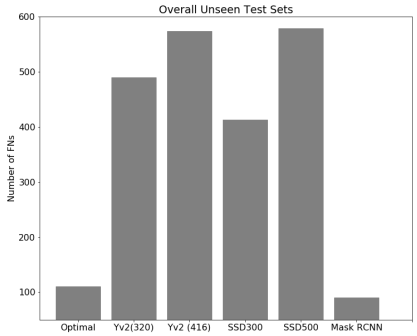


(b)

Figure 5.38: FP of detectors developed for object detection in natural images. Yv2(20): YOLOv2 (320), Yv2(416): YOLOv2(416).



(a)



(b)

Figure 5.39: FN of detectors developed for object detection in natural images. Yv2(20): YOLOv2 (320), Yv2(416): YOLOv2(416).

(320) due to its higher precision despite its comparatively lower recall. The higher precision of YOLOv2(416) is due to lower number of FPs whereas its lower recall is due to higher number of missed lesions (lower TPs and higher FNs). Adapted FRCNN model outperforms both these variations of YOLOv2 due to higher number of correct detections and lower number of FPs as seen in Figures 5.37, 5.38 and 5.39. Compared to YOLOv2 (320), the adapted FRCNN model has 18.88% higher F-measure in the modelling dataset and 21.04% higher in unseen test sets. Compared to YOLOv2 (416), adapted FRCNN model has 16.82% and 20.66% higher F-measure in modelling and unseen test sets, respectively.

SSD300 has 2.39% and 0.39% lower precision than the adapted FRCNN model in modelling and overall external test sets, respectively. This is due to the comparatively higher number of TPs and lower number of FPs produced by the adapted FRCNN model. In all datasets, SSD300 has the smallest number of low IOU FPs. SSD uses a variety of anchor boxes designed specifically for each feature extraction layer. Given the high number of anchor boxes covering various aspect ratios and scales, lesions were detected with higher IOU TP detection resulting in reduction of low IOU FPs. However, due to the use of the high number of anchor boxes, the number of additional boxes generated by SSD models is high, resulting in a high number of total FPs generated by this model.

The only exception to this is dataset D where SSD300 has a lower number of, both, low IOU FPs and additional boxes than the adapted FRCNN model.Therefore, despite the drop in TPs, in dataset D, SSD300 has 11.37% higher precision than the adapted FRCNN model. In terms of recall, compared to SSD300, the adapted FRCNN model has 17.19% to 23.04% higher recall in modelling and unseen test sets, respectively, given the higher number of TPs and lower number of FNs of the adapted FRCNN model. Therefore, due to its higher precision and recall, the adapted FRCNN model has 9.07% higher F-measure in the modelling dataset and 10.78% higher in overall unseen test sets than SSD300. In dataset D, despite its higher precision, SSD300 has 33.54% lower recall than the adapted FRCNN model resulting in its overall lower F-measure.

Similar performance is seen in SSD500. Compared to adapted FRCNN, this model has 1.61% and 4.3% lower precision in modelling and unseen test sets, respectively. In terms of recall, the adapted FRCNN model has 21.36% and 36.32% higher recall than the SSD500. Therefore, due to higher pre-

cision and recall, the F-measure of the adapted FRCNN model is 10.81% and 20.1% higher than the SSD500 model in modelling and overall unseen test sets, respectively. Compared to SSD300, SSD500 had a higher number of missed lesions with higher FN along with lower TP. Thus, a larger drop in recall in comparison to the adapted FRCNN model was observed with this model. To summarise, use of SSD helped reduce low IOU FPs due to the design of anchor boxes but the use of only a single stage led to a drop in correct detections.

In comparison to mask RCNN, a 2-stage detector, adapted FRCNN has 0.6% to 5.06 % higher recall on all datasets except dataset C. In dataset C, mask R-CNN has 12.61% higher recall than the adapted FRCNN model. In all other datasets, its recall is very close to that of the adapted FRCNN model. Overall, mask R-CNN has a higher number of TPs than the adapted FRCNN model. This is because mask R-CNN uses an ROI-align layer which improves the quality of features extracted for each proposal in comparison to the ROI-pooling layer used in FRCNN models. This resulted in better processing (classification and regression) of high IOU proposals which leads to higher number of TPs. But due to the higher number of FNs, it has comparatively lower recall than the adapted FRCNN model.

Also, this model produces a considerably high number of FPs. Therefore, its precision is 45.58% and 46.7% lower than that of the adapted FRCNN model in modelling and unseen test sets, respectively. Due to this large drop in precision, F-measure of the mask R-CNN model is 36.24% and 40.12% lower than the adapted FRCNN model in modelling and overall unseen tests, respectively. Additionally, we also investigated the performance of Hierarchical Reinforcement Learning [5] model for breast lesion detection in US images. This model was designed for object detection in natural images. When utilised for breast lesion detection in our dataset, its performance was significantly low. Detailed description of the model and its performance is presented in Section 8.6 in Chapter 8.6.

In summary, 2-stage detectors have the highest recall. In comparison, 1-stage detectors face difficulty in detection of lesions leading to their relatively lower recall. This drawback of 1-stage detectors was especially seen in datasets C and D that contain a high number of challenging lesions. On the other hand, 1-stage detectors have comparable precision to that of the adapted FRCNN model, and in some cases, higher precision than the adapted FRCNN model due to lower number of FPs.

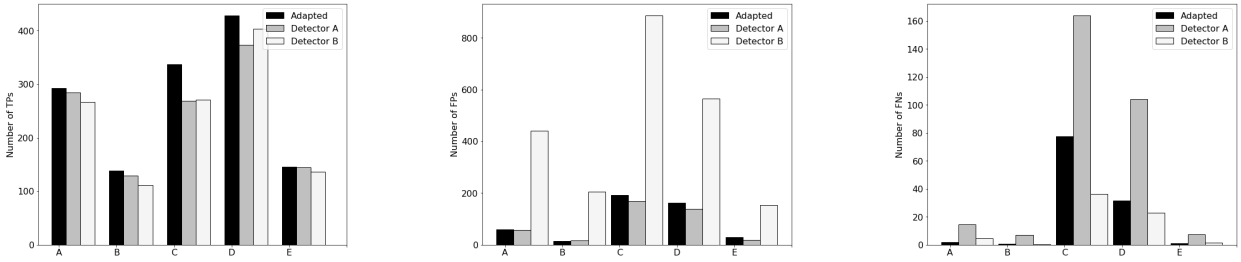
5.3.6.2 Breast Lesion Detectors

Table 5.13 compares the performance of the adapted FRCNN model to that of two breast lesion detection methods [3, 4]. Both these methods modify the FRCNN model for breast lesion detection in US images through network and modelling hyperparameters adaptation. First, the detector proposed in [3], referred to as detector A, is evaluated. Detector A undergoes modifications of the several modelling hyperparameters including anchor boxes, NMS score threshold, and number of proposals. Apart from this, modifications are also made to the input image and training mechanism. The backbone network used here is IRV2. Detailed description of this detector is provided in Section 3.2 of Chapter 3.

Dataset	**Model**	**Precision**	**Recall**	**F-measure**
A	Adapted FRCNN	**84.56**	**99.39**	**91.36**
	Detector A [3]	83.27	95.09	88.79
	Detector B [4]	38.11	98.23	54.73
Overall External Test Sets	Adapted FRCNN	**74.93**	90.32	**81.90**
	Detector A [3]	72.70	76.37	74.40
	Detector B [4]	33.75	**93.78**	49.63
B	Adapted FRCNN	**92.04**	99.28	**95.51**
	Detector A [3]	87.92	94.57	91.09
	Detector B [4]	35.63	**99.64**	52.41
C	Adapted FRCNN	58.69	82.95	**72.90**
	Detector A [3]	**61.35**	62.23	61.33
	Detector B [4]	23.52	**88.19**	37.07
D	Adapted FRCNN	**74.55**	93.14	**82.69**
	Detector A [3]	73.10	78.15	75.45
	Detector B [4]	41.83	**94.69**	57.90
E	Adapted FRCNN	**88.89**	**99.08**	**93.68**
	Detector A [3]	88.36	94.85	91.48
	Detector B [4]	47.25	98.83	63.79

Table 5.13: Performance of adapted FRCNN in comparison to breast lesion detectors.

Compared to detector A, the adapted FRCNN model has 1.82% to 6.04% higher F-measure. Overall, this method has precision comparable to that of the adapted FRCNN model, but the lower F-measure is due to larger drop in recall. Specifically, the adapted FRCNN model has 4.30% to 14.07% higher recall than this detector with only 0.24% lower precision. The comparison of the number of TP, FP and FN of the adapted FRCNN model and detector A is shown in Figure 5.40. This detector has



(a) TP (b) FP (c) FN

Figure 5.40: Number of TP, FP and FN of adapted FRCNN (IRV2 base) and breast lesion detection methods. Detector A: Method [3], Detector B: Method [4].

3.4% to 14.1% lower number of FPs (low IOU FPs and additional boxes) than the adapted FRCNN model. However, the overall lower precision of this model in comparison to adapted FRCNN is due to 3.07% to 12.7% lower number of TPs. In datasets D and E, due to lower number of FPs, detector A has 0.34% and 4.66% higher precision than the adapted FRCNN model despite the lower number of TPs in these datasets. In all other datasets, the precision of detector A is lower than that of the adapted FRCNN.

Along with a lower number of TPs, this detector also has a higher number of FNs. For instance, in dataset A, the number of FNs increased from 1.8 in the adapted FRCNN model to 14.4 in detector A and in overall unseen test sets, an increase from 111 FNs in the adapted FRCNN model to 282.8 in detector A. Due to the higher number of missed lesions, detector A has lower recall than the adapted FRCNN. The largest reduction in recall using this detector was seen in datasets C and D. Recall of dataset C and D is 19.18% and 15.05% lower than that of the adapted FRCNN model, respectively. In their published work, no detailed evaluation on the impact of applied modifications on the performance of the network is presented. However, our investigation shows that challenging lesions, more commonly found in datasets C and D, are typically scored lower than 0.9. Therefore, use of a high NMS score threshold of 0.9 resulted in these lesions being missed leading to a drop in the model's recall. On the other hand, use of the high NMS score threshold reduced FPs as the majority of the FPs are scored lower than 0.9.

Method proposed in [4] is referred to as detector B. In detector B the FRCNN network architecture and modelling hyperparameters were modified to improve detection of small lesions as well as overall performance of the FRCNN model. Here, an additional RPN is introduced along with concatenation

of convolution layers 2 and 4 of the base VGG16 network. Other adaptations include change in anchor boxes , number of test proposals and the training/testing mechanism. These changes are explained in detail in Section 3.2 in Chapter 3. In comparison to the adapted FRCNN model, this method has an overall 35.88% to 30.81% lower F-measure in modelling and unseen test sets, respectively. Although detector B has 1.16% to 3.34% higher recall than the adapted FRCNN model, the overall F-measure of this model is lower than the adapted FRCNN model due to its considerably lower precision. Compared to adapted FRCNN , this model has 45.17% to 38.71% lower precision.

Figure 5.40 compares the number of TP, FP and FN of detector B with that of the adapted FRCNN model. Compared to the adapted FRCNN model, this detector has a lower number of TPs as well as a higher number of FPs resulting in its lower precision. Here, the lower TPs do not indicate missed lesions as these were detected but with low IOU FPs. Apart from a higher number of low IOU FPs, detector B also has a higher number of additional boxes in comparison to the adapted FRCNN model. Compared to adapted FRCNN, detector B has a lower number of FNs. Thus, despite the lower number of TPs, the recall of detector B is higher than that of the adapted FRCNN model. The largest difference in recall of this detector and adapted FRCNN is seen in dataset C that contains very challenging lesions. Here, recall of detector B is 6.78% higher than that of the adapted FRCNN model. Like detector A, the published work on detector B [4] does not specify the impact of individual modifications. However, through our investigation of this detector, the following analyses were drawn. First, the high number of anchor boxes along with use of two RPN results in a high number of FPs. But due to the same characteristics and the improved quality of textural features resulting from concatenation of VGG16 layers, the number of lesions detected by detector B was high.

Computation Time: Table 5.14 shows the minimum, maximum and average computation time of all models. The computation time, as described in Section 4.2 Chapter 4, is time required by the detector to process a single input image. over all datasets for all these networks. The time presented in Table 5.14 is the average over all datasets (datasets A to E). 1-stage detectors are naturally faster than the 2-stage networks due to smaller sized networks. Of this, YOLOv2 (320) is the fastest with an average computation time of 0.03 seconds. Mask R-CNN is faster than the adapted FRCNN model with an average computation time of 0.29 seconds. Adapted FRCNN model had the highest computation time of 0.39 seconds (average) due to the use of a deeper base network. Detector A

has an average computation time of 0.34 seconds comparable to that of the adapted FRCNN model. On the other hand, detector B is faster than both detector A and the adapted FRCNN model with average computation time of 0.24 seconds.

Model	**Computation Time (sec)**		
	Average	**Minimum**	**Maximum**
Original FRCNN	0.16	0.09	3.27
Adapted FRCNN	0.39	0.37	1.95
YOLOv2 (320)	0.03	0.03	0.36
YOLOv2 (416)	0.03	0.03	0.37
SSD300	0.04	0.03	0.32
SSD500	0.16	0.14	0.73
Mask R-CNN	0.30	0.25	1.50
Detector A [3]	0.34	0.32	1.64
Detector B [4]	0.24	0.10	2.25

Table 5.14: Computation time (in seconds) of all evalutated state-of-the-art detectors in comparison to original and adapted FRCNN.

5.4 Discussion

This chapter addressed the current gap in the literature through an investigation of the modelling hyperparameters of the FRCNN network on a large dataset of US images. Through the investigation an adapted FRCNN was developed which outperformed the original FRCNN model as well as other state-of-the-art detectors. Although this investigation and adaptation of the FRCNN model was conducted for breast lesion detection in US images, it can also prove useful for detection of other types of lesions in US images with minor modifications. For instance, thyroid lesions in US images have been shown to have similar characteristics as breast lesions in US images [139]. However, the average number of thyroid lesions in a typical US image is comparatively higher than that of breast lesion US images. Therefore, increasing the number of proposals would prove beneficial.

Apart from the modelling hyperparameters evaluated in this chapter, an investigation of RPN training samples and NMS score threshold were also conducted. RPN uses samples with IOU $[0.7, 1]$ as positive and those with IOU $[0, 0.3]$ as negative. Broadening this range such as $[0.5, 1]$ for positive samples and $(0, 0.5]$ for negative samples led to an overall drop in performance. When trained with a higher NMS score threshold of 0.9 instead of 0.3, no lesions were detected. Since the margin of

the lesion carries important information, the impact of introducing this region in the training of the adapted FRCNN model was studied. Here, the adapted FRCNN model with IRV2 was trained with GT boxes that covered 2%, 8% and 16% of the surrounding margin region. It was found that use of 2% margin was the best performing one out of the three investigated values. However, it performed poorly in comparison to the model trained without any margin in the GT. Additional impact of GARPN [75] and fusion of the base network's convolutional layers were also investigated to improve the classification accuracy of the adapted FRCNN model in order to reduce FPs. Description and performance of these methods is provided in further detail in Section 8.2 of Chapter 8.

The evaluation metrics used in this work are commonly utilised for lesion detection. However, these metrics do not address some of the intricacies of this field. For instance, many of the low IOU FP detections by the FRCNN models are centred on the lesion; these detections cover the lesion plus margin area surrounding the lesion. Although these are classified as FP due to their low IOU , they successfully detect the lesion albeit with larger margin area. Over the years, methods have been proposed to overcome this issue. One such method uses the centre of the detections as the guide for their classification as TP or FP. Here, if the centre of the output detection lies inside the GT box, it is classified as TP, otherwise as FP. However, as shown in [3], this technique fails to correctly evaluate the model as there is a higher acceptance of very low IOU detections as TPs. Development of a standard evaluation method for lesion detection is a growing field of research as improved evaluation metrics help in gaining a deeper understanding of the model's performance.

5.5 Summary

This chapter presented the investigation and the adaptation of FRCNN for breast lesion detection in 2D US images. FRCNN is a 2-stage detector that was originally designed for object detection in natural images. It is popularly used for lesion detection in medical images through modification of its modelling hyperparameters and/or its network architecture. However, these methods have two important drawbacks, namely, insufficient data on the evaluation of the modifications and use of a small to medium sized dataset. We addressed these drawbacks in this chapter through a thorough investigation of the important modelling hyperparameters of this network using our large dataset of US images collected from real-life clinical settings.

The adapted FRCNN, designed through the investigation of the hyperparameters, outperformed the original FRCNN due to a reduction of 28% to 61% in FPs resulting in 5% to 21% higher precision and a small negative impact of 0.27% to 9% in recall. The most common FPs in this adapted FRCNN model were caused due to low classification accuracy of the RPN and base network and improper filtering of proposals by the nms at both stages of the network. In this chapter, two methods to improve the classification accuracy of the adapted FRCNN were presented. These include investigation of state-of-the-art classification networks as backbone of the adapted FRCNN model and evaluation of several training losses.

Following is the summary of the main findings and contributions made in this chapter:

- Investigation of the impact of various modelling hyperparameters of the FRCNN model on the overall performance for breast lesion detection in US images.
- Successful reduction of FPs in FRCNN models trained with optimal values of the investigated hyperparameters and using any set of anchor boxes was presented.
- Irrespective of the anchor boxes used, use of the optimal values of the other investigated hyperparameters successfully improved the performance of the model in comparison to its counterpart trained on default values of the hyperparameters. This improved performance was achieved due to a considerable reduction in FP cases.
- The cause for FPs in the adapted FRCNN model was identified as the incorrect classification of FP proposals at both stages of the network and incorrect retention of FP proposals after filtering through NMS in both stages.
- Following state-of-the-art classification networks were evaluated to improve the classification accuracy of the adapted FRCNN model: ResNet50, ResNet101, Inception-v3 and Inception-ResNet-v2 (IRV2). Of these , IRV2 had the best overall performance due to its high classification accuracy which not only reduced FPs but also increased the number of correct detections of the adapted FRCNN model.
- PISA and CARL losses were evaluated to specifically improve the classification accuracy of the RPn thereby improving the quality of proposals as well as overall performance of the model.

- All investigated losses improved the classification accuracy of the model for challenging lesions. Overall, use of PISA loss in the classification branch had the highest performance largely due to the increase in detected lesions and small reduction of FPs. However, this improvement was small . Use of CARL and other variants of PISAloss had a negative impact on the number of FPs.

- The adapted FRCNN model with IRV2 as the backbone was compared to several state-of-the-art object detectors including YOLOv2, SSD and mask R-CNN which were designed for object detection in natural images and two breast lesion detection methods [3, 4] developed for US images. The adapted FRCNN model outperformed all evaluated detectors.

In summary, this chapter addressed an important gap in the literature and proposed an adapted FRCNN that successfully modified the FRCNN model for breast lesion detection in US images. Use of IRV2 as the backbone network successfully improved the overall performance of the adapted FRCNN model by addressing the issue of classification accuracy in both stages of FRCNN. In the following chapter, a novel U-Detect method is proposed in order to reduce FP detections caused by NMS.

Chapter 6

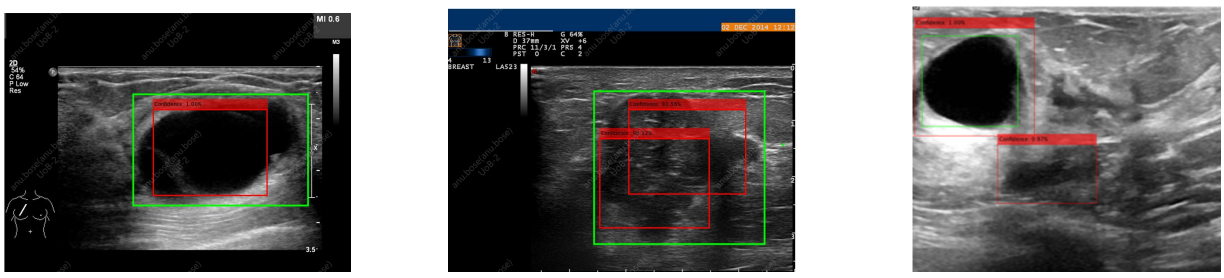
U-Detect: A Clustering-Based Approach Using Learned Features

Chapter 5 presented our successful adaptation of the FRCNN method for detecting breast lesions from US images with performance higher than the existing detection techniques. However, despite the adaptation of the network hyperparameters, the false positive (FP) detection remains an issue. The aim of this chapter is to present a novel method to reduce FP detections while maintaining acceptable level of true positive (TP) detections.

Both stages of the FRCNN network – Region Proposal Network (RPN) and base network – generate a high number of proposals for a single input image. It is critical to remove the redundant and poor-quality (low IOU) proposals for reliable performance and lower computation cost. The post-processing method employed in FRCNN is called Non-Maximal Suppression (NMS) which is a simple and computationally light technique. It is applied to the proposals, first, when they are generated at the RPN and, second, after their processing at the base network. NMS uses classification scores and IOU for filtering out proposals. In particular, first, all proposals with classification scores < 0.3 are removed. From the remaining proposals, the proposal with highest classification score is moved to the final output and any proposals with $IOU \geq 0.7$ with this highest-scoring proposal are discarded as redundant. This process continues until all proposals are either moved to the final output or discarded. Despite its advantages, NMS has major limitations as it does not consider textural information in identifying redundant proposals. The improper grouping of proposals as redundant

based on their overlap with the highest scoring proposal causes the common issue cases found in the adapted FRCNN presented in Chapter 5.

In particular, our work presented in Chapter 5 Section 5.3.3 shows that NMS produces two types of FPs: single FPs and multiple FPs. Single FPs are those detection cases where the FRCNN model produces a single detection which is either a low IOU FP, i.e., IOU $(0, 0.5)$ or an additional box covering the background region with $IOU = 0$. On the other hand, multiple FPs is a scenario where the model's produces more than one FP (with or without a TP detection) which includes low IOU FPs with IOU $(0, 0.5)$ that cover small regions of a lesion, FPs with IOU 0, or combination of all/some of these FP types. Figure 6.1 illustrates different types of FPs.



(a) Single FP (b) Multiple FPs (c) Multiple FPs

Figure 6.1: Issue cases of adapted FRCNN model caused by NMS.

In the majority of the single FP output scenarios, the adapted FRCNN model produces at least one TP proposal at the RPN and base network as shown in Chapter 5 Section 5.3.3. However, due to the higher score of the FP proposal, it is selected as the final output and because of the high IOU of the TP proposal with this FP proposal, the TP proposal is incorrectly discarded as redundant. Similarly, in case of multiple FPs output, the FP proposals are filtered through the NMS due to their high classification score and improper grouping as shown in Chapter 5 Section 5.3.3.

This chapter presents a new method called U-Detect that combines unsupervised learning technique and FRCNN to reduce the FP detections. In particular, we propose a clustering approach to reduce the FP detections caused by the limitations of NMS. We hypothesise that clustering proposals with similar learnable texture features leads to identification of distinctive group of detections of one object (lesion), while the candidate selection from each group leads to the true detection of the object.

The proposed approach can be summarised as follows. First, proposals in the RPN or base network of FRCNN are represented using learnable features extracted from convolutional neural network layers. Then, the x-means clustering method is used to group texturally similar proposals. After clusters are formed, a single best proposal is selected from each cluster, referred to as the candidate of that cluster, and remaining proposals in that cluster are considered redundant and discarded. As the number of clusters generated is typically high, the selected candidates tend to have high overlap with each other. Therefore, these candidates are processed through NMS to remove spatially redundant proposals. Since the candidates have been selected through the clustering mechanism, processing them through NMS at this stage does not have the same negative impact on the overall performance as seen in the adapted FRCNN model. Finally, from the remaining candidates, if any candidates have an overlap of over 20% with each other, they are merged into one box through a novel candidate merging method. Output of the candidate merging method is the final output of the model. This approach is the first piece of research work to use unsupervised learning techniques to reduce the FP detections in FRCNN.

This chapter is composed of the following sections. Section 6.1 provides a detailed description of our approach which includes features extraction and representation of the proposals, dimensionality reduction, proposal clustering, candidate selection and candidate merging method. The performance of the proposed approach is presented in Section 6.2. Finally, the chapter concludes with discussion in Section 6.3 and a summary in Section 6.4.

6.1 U-Detect for False Positive Reduction

We propose a novel method to reduce FP detections. The main approach of the proposed solution is to use unsupervised learning to cluster proposals in the RPN or base network on the basis of textural similarity. This is done to overcome the issue cases (single and multiple FPs) resulting from improper filtering through NMS. U-Detect method is designed for the test-stage of a pretrained detector (FR-CNN) thereby adding no additional computational cost to the model training. The proposed method consists of five main phases, namely, learnable feature extraction, dimensionality reduction, proposal clustering, candidate selection, and candidate merging. Figure 6.2 shows our U-Detect method.

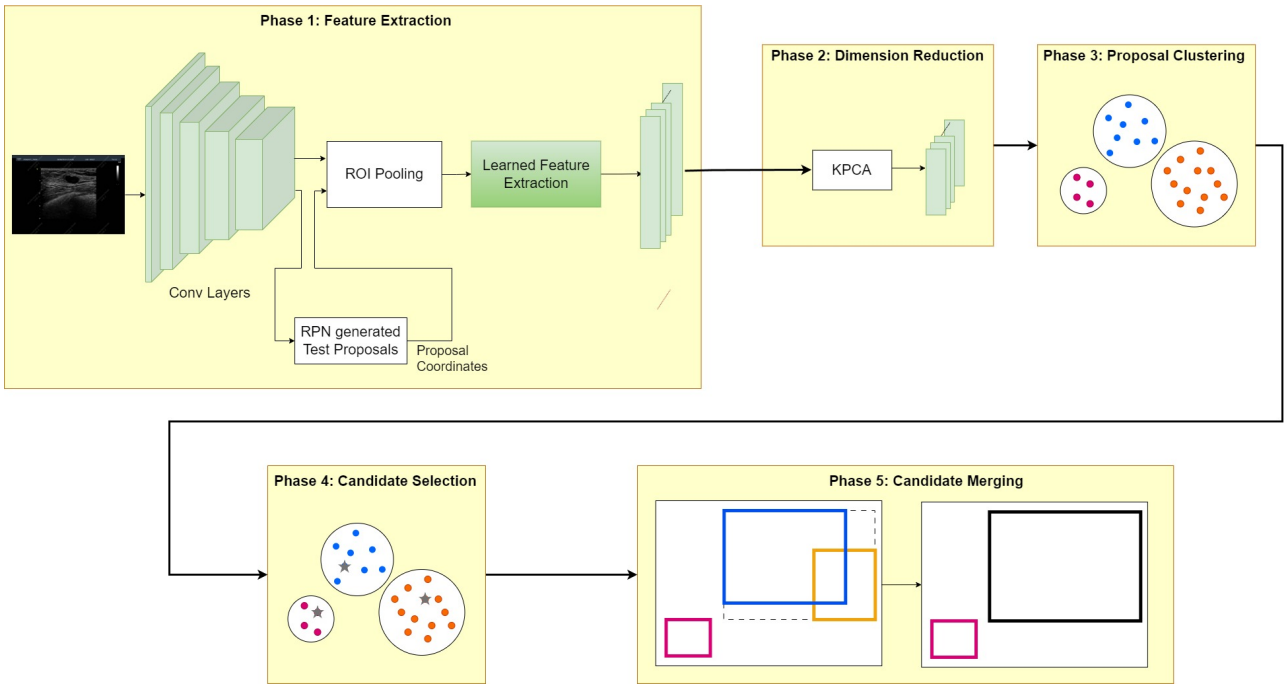


Figure 6.2: The proposed U-Detect method.

We use our U-Detect method to process proposals generated by the RPN as well as proposals processed by the base network of the pretrained adapted FRCNN model with IRV2 base from Chapter 5. Depending on the extraction of learned features, two U-Detect methods are developed, namely, U-Detect-RPN and U-Detect-Base. Figures 6.3 and 6.4 illustrate U-Detect-RPN and U-Detect-Base models, respectively, highlighting the process from phase one to phase six in both methods. Sections 6.1.1 to 6.1.5 describe each phase of the U-Detect method in further detail.

6.1.1 Learned Feature Extraction

Proposals generated by FRCNN RPN and base-network represent image regions with potential presence of lesions. Each proposal has its own texture characteristics which can be extracted and used to identify the correct detection. This section provides details of our approach of learned features extraction used for description of proposals. Features extraction of both U-Detect-RPN and U-Detect-Base are described as follows.

6.1.1.1 RPN Features in U-Detect-RPN

In FRCNN, the quality of RPN test proposals sent through to the base network has a considerable impact on the overall performance of the detection model. During model testing, features of the input image are extracted by the convolution layers of the base network, referred to as conv-features, which are then input to two branches: RPN and ROI pooling layer as illustrated in Figure 6.3. RPN uses the conv-features to generate proposals. After classification, an average of 300 test proposals per image is generated. ROI pooling layer extracts features of each test proposal from the conv-features which is then sent to the base network for further processing. The extracted features of each proposal have a dimension of $17 \times 17 \times 1088$ where 1088 is the number of channels in the described convolution layer.

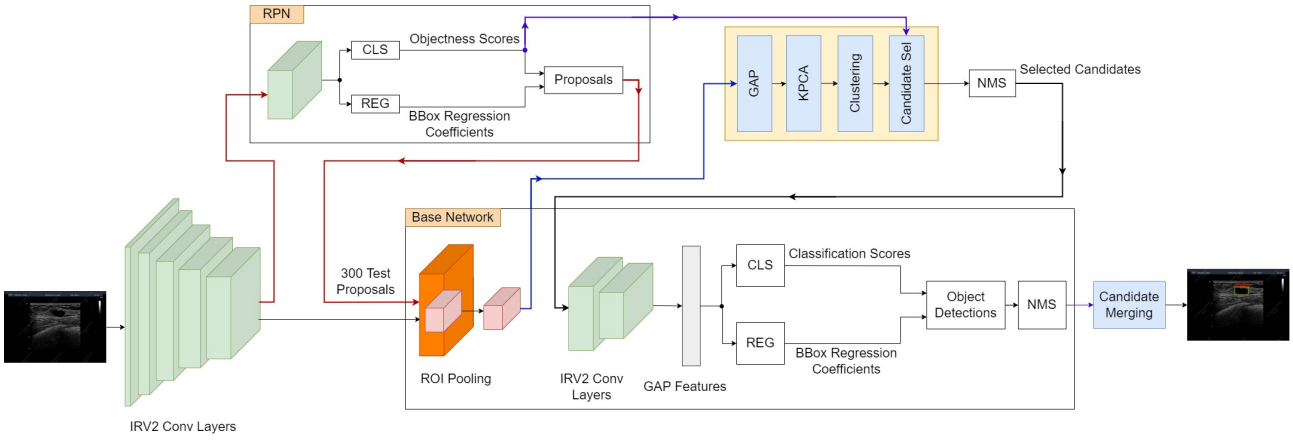


Figure 6.3: Proposed U-Detect-RPN method.

To use the features output by the ROI pooling layer, for defining each proposal, every feature map needs to be flattened and concatenated to form an extremely large feature vector of size $(17 \times 17 \times 1088) \times 1$. Such a feature vector would undoubtedly increase computation time. Furthermore, part of these feature maps contain little to no textural information or redundant information. Thus, to effectively reduce the dimension without losing vital textural information of each feature map, global average pooling (GAP) is utilised. Essentially, GAP replaces every feature map with the average of all its individual features. So, the impact of redundant and irrelevant feature maps is reduced while important information from remaining feature maps is preserved. Detailed description of GAP layer is provided in Section 2.2.1.2 Chapter 2. After applying GAP, the dimensionality of the feature vector is reduced from $(17 \times 17 \times 1088) \times 1$ to 1088×1. This 1088×1 feature vector is referred to as

RPN-GAP features and is used to describe proposals in U-Detect-RPN.

6.1.1.2 Base Network Features in U-Detect-Base

ROI pooling layer outputs features of each RPN-generated proposal as shown in Figure 6.4. These features are then processed through convolution layers and finally through a GAP layer embedded in the base network. Output feature vectors from this GAP layer, referred to as Base-GAP feature vectors, are used for classification and bounding box regression of the test proposals by the base network. The Base-GAP feature vector has higher abstraction and contains more textural information than the RPN-GAP feature vector. Therefore, in U-Detect-Base, each proposal is described using its Base-GAP feature vector, of size 1536×1.

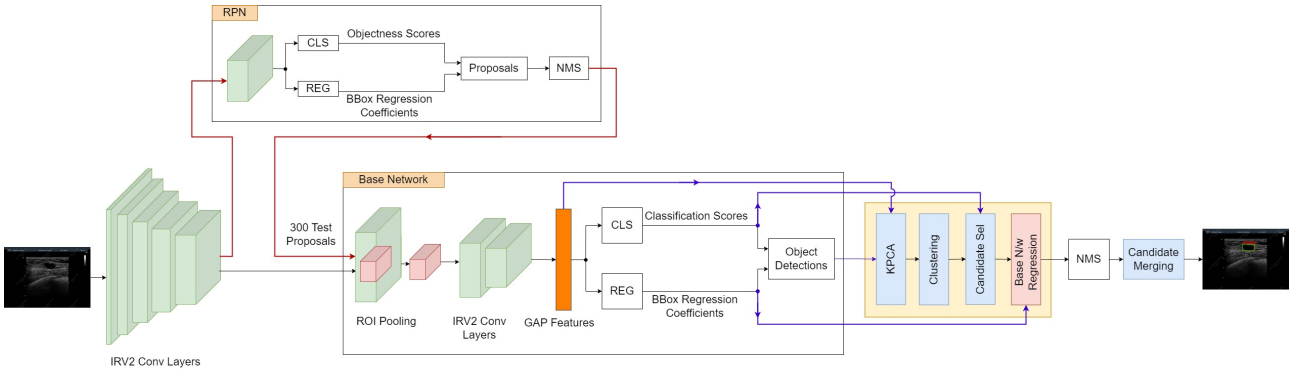


Figure 6.4: Proposed U-Detect-Base method.

6.1.2 Dimensionality Reduction

Given the large size of the RPN-GAP and Base-GAP feature vectors presented in Section 6.1.1, the Principal Component Analysis (PCA) and Kernel Principal Component Analysis (KPCA) are used and compared to reduce the dimensionality of the extracted feature vectors in phase two of the U-Detect method. Detailed explanation of these dimensionality reduction methods is provided in Section 2.2.2 Chapter 2.

6.1.2.1 Principal Component Analysis (PCA)

PCA, a commonly used dimension reduction technique, is employed to reduce dimensions of both RPN-GAP and Base-GAP features. To find the optimal number of principal components (PCs) for each feature vector, a range of PCs were investigated. This range was determined using a preliminary test

conducted to identify the number of PCs required to capture 95% to 99% of the textural information contained in the entire feature vector i.e., the number of PCs with cumulative information (CI) of 95% to 99%. Based on this, for both RPN-GAP and Base-GAP feature vectors, 5, 10, 15, and 20 PCs were evaluated and the optimal number of PCs was determined experimentally as illustrated in sections 6.2.1.2 and 6.2.2.2.

6.1.2.2 Kernel Principal Component Analysis (KPCA)

CNN features are nonlinear due to the nonlinear nature of the input data (image) as well as the multiple non-linear activation functions used in the convolutional neural network. KPCA essentially adapts PCA for non-linear data. Thus, use of KPCA to reduce dimension of RPN-GAP and Base-GAP features was also investigated. Like with PCA, an initial test was conducted to find the number of kernel PCs required to contain 95% to 99% CI. Based on the findings of this test, the following number of kernel PCs were evaluated for both feature vectors: 5, 10, 15, and 20. The optimal number of kernel PCs was selected experimentally as shown in sections 6.2.1.2 and 6.2.2.2.

6.1.3 Proposal Clustering

Using extracted learned features, proposals are grouped using x-means clustering method [1] in phase three of the U-Detect method. X-means clustering is an adaptation of k-means clustering where the requirement of predefining number of clusters (k) to be formed is eliminated. Instead, a range of potential k values are input to the method and an optimal k is found iteratively during model testing. In particular, the range of k is set to $[1, max\_prop]$ where max_prop represents the maximum number of proposals generated for that image. Typically, max_prop is 300 but in some images with small lesions, it drops to around 296 or 297. Distance measure is a core component used by clustering algorithms to group similar data points into the same clusters. In our method, the cosine similarity method is used to measure the distance between the feature vectors representing the texture information of the proposals. Equations 6.1 and 6.2 define the cosine similarity method.

$$dist(A, B) = 1 - cos(A, B) \tag{6.1}$$

where

$$cos(A, B) = \frac{(A \cdot B)}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{6.2}$$

Here, A and B represent feature vectors of two proposals. A small distance between two proposals therefore indicates their high textural similarity and vice versa. Thus, proposals that are texturally similar, such as high IOU TP proposals, are clustered together separate from other proposals such as low IOU FP. Another core component of x-means algorithm is the Bayesian information criterion (BIC). BIC is the metric used to determine optimal k. BIC is defined in Equation 2.3 in Section 4.2 Chapter 4. BIC score is assigned such that a balance between good fit and total number of clusters is measured. In particular, BIC uses log-likelihood to determine fit of the clusters in the data and a penalty term to penalise high numbers of clusters. Penalty is impacted by dimension of the feature vector M, number of cluster k, number of proposals in the cluster R_n, total number of proposals and a constant C. C is set to a default value of 0.5. As all values with the values of C are dependent on the image and variable k, we investigate the impact of the constant C in its impact on the overall number of clusters in U-Detect models. Specifically, C is optimised for both U-Detect-Base and U-Detect-RPN using a range of values between 0 and 20.

6.1.4 Candidate Selection

In a typical clustering task, the centroid of a cluster is selected as its candidate. However, centroids are a *mean* representation of their cluster. As the aim of this method is to select the *best* proposal from a cluster, centroid is not used for candidate selection. Generally, both RPN and base network assign high confidence scores to high IOU proposals and low scores to low IOU proposals. Therefore, to ensure that the objective of only filtering out the best proposals is met, RPN and base network's classification scores are used for candidate selection in U-Detect-RPN and U-Detect-Base, respectively. Specifically, after proposals are clustered using learned features, from each cluster, the proposal with highest RPN/base network classification score is selected as the candidate of that cluster in phase four of the U-Detect method.

It is worth noting that selected candidates tend to have high overlaps with each other especially when clustered with low BIC penalty. Thus, they are processed through NMS to remove low-scoring and redundant candidates. Since these candidates have been identified through the clustering method,

passing them through NMS does not cause the issue cases that NMS is otherwise prone to cause. In U-Detect-RPN, selected candidates are processed through the NMS before passing through to the base network, after which they are passed to the final phase as shown in Figure 6.3. On the other hand, in U-Detect-Base, the selected candidates pass through the NMS and then directly move to the final phase as shown in Figure 6.4.

6.1.5 Candidates Merging

As illustrated in Figure 6.1b, the multiple overlapping FPs are typically found in large lesions where the FPs cover small regions of the lesion. These boxes are not removed through NMS due to their lower overlap with each other and relatively high classification score. Candidate merging method is specifically developed to reduce the number of multiple overlapping FPs. In particular, candidates (output detections of the model) with $IOU \geq 0.2$ are merged to form single detection in the final phase of the U-Detect method.

6.2 Experimental Results and Analysis

This section presents the experimental results of the proposed U-Detect method. Adapted FRCNN with IRV2 base from Section 5.3.4 of Chapter 5 is used as the pretrained detector. Same modelling dataset (A) and external datasets (B, C, D, and E) presented in Section 4.1 of Chapter 4 are used to evaluate the U-Detect. All experimental results are the average of 5-folds unless specified otherwise. First, performance of U-Detect-RPN is described in Section 6.2.1 which is followed by Section 6.2.2 detailing the performance of U-Detect-Base method. Development of both models begins with a base U-Detect model using the entire feature vector (RPN-GAP or Base-GAP), default C of 0.5 and without candidate merging. Every phase is adapted or added in individual investigation to study its impact. Therefore, the output of each investigation/adaptation is the input to the next step.

6.2.1 U-Detect-RPN

This section reports the evaluation results of U-Detect-RPN. First, performance of U-Detect-RPN with entire RPN-GAP features used for proposal description is detailed. This is followed by a description of the impact of using PCA and KPCA for dimensionality reduction of the RPN-GAP feature vector. Next, analysis of parameter C in BIC penalty for building a better balance between quality and

number of clusters is detailed. All these models do not include phase five (candidate merging). The impact of this final phase is studied separately and presented in the final part of this section.

6.2.1.1 RPN-GAP Feature Vector

Table 6.1 shows the performance of the U-Detect-RPN using entire RPN-GAP features in comparison to that of adapted FRCNN. U-Detect-RPN outperforms adapted FRCNN with 0.27% to 1.52% higher precision in modelling and unseen test sets, respectively, with a comparatively small drop of 0.5% in recall in the overall unseen test sets (there was no change in recall of modelling dataset). This improvement in performance is due to increase in TP detections along with reduction in FPs.

Dataset	**Model**	**Precision**	**Recall**	**F-measure**
A	Adapted FRCNN	83.18	**99.39**	90.55
	U-Detect-RPN	**83.45**	**99.39**	**90.71**
Overall External Test Sets	Adapted FRCNN	70.69	**90.58**	79.33
	U-Detect-RPN	**72.21**	90.08	**80.11**
B	Adapted FRCNN	90.16	**99.42**	94.53
	U-Detect-RPN	**91.47**	**99.42**	**95.26**
C	Adapted FRCNN	60.15	**81.55**	68.97
	U-Detect-RPN	**61.64**	80.23	**69.48**
D	Adapted FRCNN	72.58	**93.20**	81.46
	U-Detect-RPN	**73.35**	93.08	**81.92**
E	Adapted FRCNN	82.56	99.20	89.74
	U-Detect-RPN	**85.76**	**99.21**	**91.84**

Table 6.1: Performance of U-Detect-RPN and adapted FRCNN models.

U-Detect-RPN overcomes the drawbacks of NMS in all datasets through effective filtering of proposals, ensuring that TP proposals are not overshadowed by FP ones. FP proposals with $IOU = 0$, covering lesion-like background regions are clustered with texturally similar proposals. The selected candidates from such clusters typically have high overlap with TP candidates. Due to the higher score of the TP candidates, these FP candidates are removed in the NMS phase. Without U-Detect-RPN, these FP candidates are output by the adapted FRCNN as additional boxes. Candidates from clusters containing low IOU FP proposals are also removed in a similar manner. Thus, compared to the adapted FRCNN model, U-Detect-RPN produces a higher number of TPs along with a lower number of both types of FP detections, namely, low IOU FPs and additional boxes.

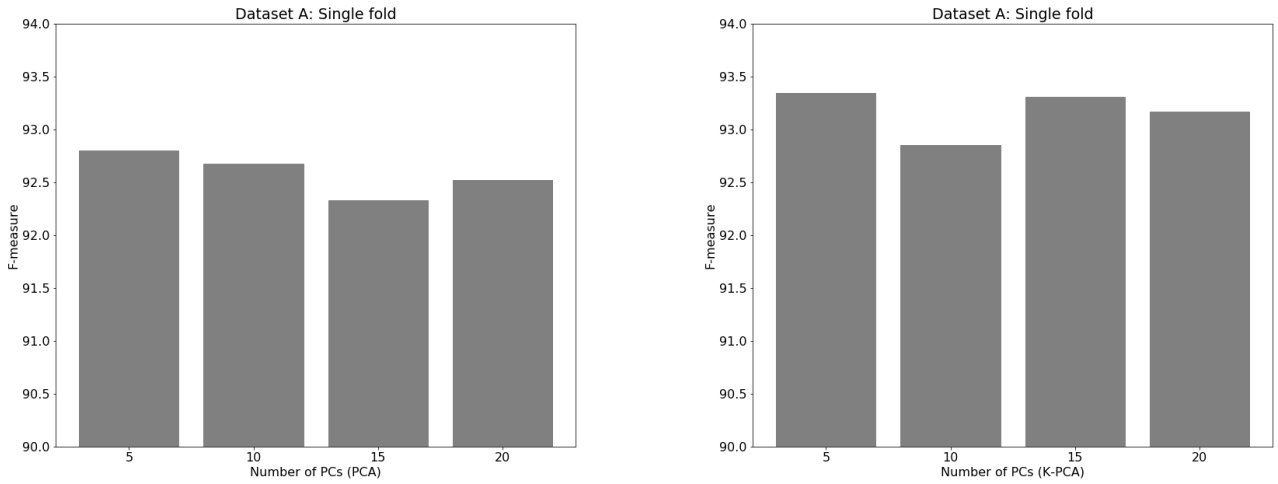
However, the U-Detect-RPN model has a higher number of FNs in overall unseen tests compared to the adapted FRCNN model. Due to the relatively poorer classification accuracy of the RPN, challenging lesions with background-like texture are typically scored low. In adapted FRCNN where all RPN-generated proposals are sent through to the base network, proposals containing such lesions are correctly classified by the base network owing to its higher classification accuracy. But, in U-Detect-RPN, due to the textural similarity between lesion and background, TP proposals covering the lesion are clustered with those covering background regions (FP proposals). Owing to the lower score assigned to the TP proposal, it is not selected as the candidate. Since no TP proposal is sent through to the base network, it is missed completely by the model. Additionally, lesions missed by adapted FRCNN are also missed by U-Detect-RPN. In these common FN cases, all proposals covering the lesion are assigned very low classification scores (i.e., labelled as background). As the classification scores assigned to proposals are not altered by U-Detect models, despite good clustering of the proposals, these lesions are also missed by U-Detect-RPN. An example of such missed lesions common to both models and their clusters is shown in Figure 6.16.

6.2.1.2 Dimension Reduction of RPN-GAP Feature Vector

The previous section demonstrated the performance of U-Detect with full RPN-GAP features (1088×1). This section reports the evaluation results of using PCA and KPCA to reduce the dimension of RPN-GAP feature vector. From this point on, for the sake of brevity, all U-Detect-RPN models are referred to by the feature vectors used for proposal description. For instance, RPN-GAP model refers to the U-Detect-RPN model that uses RPN-GAP features for description of proposals. Both these models do not include candidate merging method.

To identify the optimal number of PCs, the performance of U-Detect-RPN was evaluated using 5, 10, 15 and 20 PCs using a single fold of modelling dataset. F-measure was used to identify the optimal number of PCs. Similar to PCA analysis, the number of kernel PCs was evaluated using the same single fold of modelling dataset and number of optimal kernel PCs was identified based on the performance in this fold. Figure 6.5 shows the performance of all tested PCA and KPCA components. Overall, as the number of components increased, performance of the model dropped. Based on the performance on the modelling dataset, 5 PCs and 5 kernel PCs were selected as optimal. Table 6.2 shows the impact of reducing dimension of RPN-GAP feature vector of size 1088×1 to 5×1 using

PCA and KPCA. The U-Detect-RPN model using 5 PCs of RPN-GAP feature vector for proposal description is referred to as RPN-GAP-PCA model and the model using 5 kernel PCs is referred to as RPN-GAP-KPCA. Overall, both RPN-GAP-PCA and RPN-GAP-KPCA outperform RPN-GAP model and adapted FRCNN model (see Table 6.1 for performance of RPN-GAP and adapted FRCNN models). The RPN-GAP-KPCA model has the highest F-measure in all datasets.



(a) PCA (b) KPCA

Figure 6.5: Evaluation of number of PCs in PCA and KPCA applied to RPN-GAP feature vector.

Dataset	U-Detect-RPN Model	Precision	Recall	F-measure
A	RPN-GAP-PCA	**84.34**	97.21	90.31
	RPN-GAP-KPCA	84.08	**99.26**	**91.02**
Overall External Test Sets	RPN-GAP-PCA	**75.54**	88.35	81.38
	RPN-GAP-KPCA	74.85	**90.37**	**81.84**
B	RPN-GAP-PCA	91.18	95.97	93.47
	RPN-GAP-KPCA	**92.51**	**99.43**	**95.83**
C	RPN-GAP-PCA	**66.57**	78.97	71.94
	RPN-GAP-KPCA	65.84	**80.59**	**72.24**
D	RPN-GAP-PCA	**76.16**	90.81	82.74
	RPN-GAP-KPCA	74.73	**93.14**	**82.81**
E	RPN-GAP-PCA	86.13	98.65	91.85
	RPN-GAP-KPCA	**86.64**	**99.21**	**92.38**

Table 6.2: Impact of dimension reduction of RPN-GAP features using PCA and KPCA on the performance of U-Detect-RPN.

The RPN-GAP-PCA model outperforms RPN-GAP with 0.01% to 2.46% higher F-measure in the majority of the datasets (datasets C to E). In datasets A and B, the F-measure of this model is

lower than that of RPN-GAP by 0.4% and 1.79%, respectively. In general, the RPN-GAP-PCA model has 0.89% to 3.23% higher precision in modelling dataset and overall external test sets, respectively. Largest improvement of 4.93% was seen in dataset C. However, this model has 2.18% and 1.76% lower recall than the RPN-GAP model in modelling and overall external test sets, respectively. In datasets A and B, due to larger drop in recall than improvement in precision, RPN-GAP-PCA model has lower F-measure than RPN-GAP model. On the other hand, in datasets C to E, the drop in recall was not significant leading to higher F-measure of the RPN-GAP-PCA model. Similar performance was seen in comparison to the adapted FRCNN. In particular, the RPN-GAP-PCA model outperforms the adapted FRCNN model in datasets C, E, and E with 1.28% to 2.97% higher F-measure due to higher precision and a relatively smaller drop in recall. In datasets A and B, the RPN-GAP-PCA model has 0.24% to 1.06% lower F-measure due to larger drop in recall than the gain in precision.

On the other hand, RPN-GAP-KPCA has 0.31% and 1.72% higher F-measure than the RPN-GAP model in modelling and overall external test sets, respectively, of which the largest improvement is seen in dataset C (2.76% higher F-measure). This higher F-measure of the RPN-GAP-KPCA model is due to its higher precision as well as recall. Specifically, the RPN-GAP-KPCA model has 0.63% to 4.2% higher precision and 0.01% to 0.36% higher recall in modelling and overall external test sets. Only exception to this is modelling datasets where this model has 0.13% lower recall than the RPN-GAP model. In comparison to the adapted FRCNN model, the RPN-GAP-KPCA model has higher F-measure (0.47% higher in modelling dataset and 2.53% in overall unseen test sets) due to higher precision in the range of 0.9% to 5.69% with relatively smaller drop of 0.06% to 0.96% in recall. In datasets B and D, along with higher precision, recall was also higher. In comparison to RPN-GAP-PCA model, the RPN-GAP-KPCA model has 0.71% to 0.46% higher F-measure in modelling and unseen test sets, respectively, which is due to average of 2.05% higher recall over all datasets and a small drop of 0.26% to 0.66% in precision.

The higher precision of both RPN-GAP-PCA and RPN-GAP-KPCA models in comparison to the adapted FRCNN and RPN GAP models is due to the reduction of FPs (both low IOU FPs and additional boxes). PCA and KPCA condense the whole RPN-GAP feature vector such that the impact of noisy, redundant information in the original feature vector is reduced without loss of critical information. This leads to improved clustering that results in overall improvement in performance.

Additionally, due to the superior quality of the KPCA feature vector, the RPN-GAP-KPCA model has a generally higher number of TPs than all models in modelling and external test sets. In this case, use of KPCA feature vectors further improves the quality of clusters formed in comparison to the RPN-GAP and RPN-GAP-KPCA models resulting in selection of higher IOU candidates for further processing.

However, despite the higher number of correct detections, RPN-GAP-KPCA has lower precision than the RPN-GAP-PCA model due to the higher number of additional boxes. While the superior quality of the KPCA feature vector helps reduce low IOU FPs due to improved clustering, the same quality of the KPCA feature vector also causes an increase in the number of clusters formed in lesion-like regions of the background, with the majority of the proposals covering these regions in single-element clusters (clusters containing a single proposal). Therefore, these proposals are sent through to the base network, leading to additional boxes in the output of the model thereby increasing the total number of FPs generated by this model.

However, both RPN-GAP-PCA and RPN-GAP-KPCA models have higher FN than the RPN-GAP and adapted FRCNN. Specifically, RPN-GAP-PCA has the highest number of missed lesions (low TPs and high FNs) leading to its lowest recall in all datasets. RPN-GAP-KPCA has higher recall than the RPN-GAP model due to the relatively higher TPs despite the lower number of FNs. Most lesions missed by the RPN-GAP-PCA and RPN-GAP-KPCA models are small and challenging with background-like texture. Due to condensation of already weak textural information contained in RPN-GAP feature vectors for such lesions, proposals covering these lesions are clustered with those covering background regions. Furthermore, FNs missed in the RPN-GAP model are also missed here due to the relatively poor classification accuracy of the RPN as explained in the previous section. In terms of computation time, reducing the dimension from RPN-GAP in RPN-GAP-PCA and RPN-GAP-KPCA led to a reduction in computation time from 16.52 seconds for the U-Detect model using RPN-GAP features to 3.96 seconds in U-Detect-model using RPN-GAP-KPCA model.

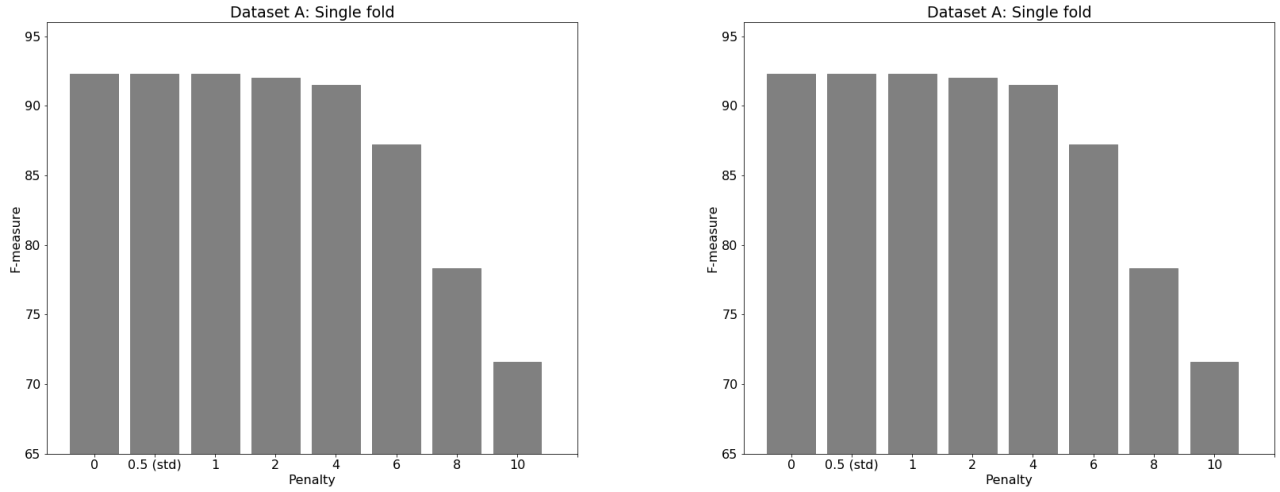
In summary, reducing dimension of RPN-GAP feature vector from 1088×1 to 5×1 using either PCA or KPCA led to an overall improvement in performance along with reduction in computation time. However, in both RPN-GAP-PCA and RPN-GAP-KPCA, a significant number of clusters contain

single proposals, followed by a considerable portion containing a very small number of proposals. Due to this, majority of the proposals through to NMS in phase five without going through the intelligent filtering of U-Detect method, which results in small to no change in issue cases caused by NMS. This high number of clusters is caused due to the low default penalty used in the BIC metric of x-means clustering. Following section evaluates the BIC metric of x-means.

6.2.1.3 X-means Penalty Evaluation for RPN-GAP

This section investigates the impact of C in the BIC penalty term on the overall performance of RPN-GAP-PCA and RPN-GAP-KPCA. Both these models use the optimal number of components (5 PCs and 5 kernel PCs) selected in the previous section. Also, these models do not include candidate merging (phase five of the U-Detect method). First, the penalty term in the RPN-GAP-PCA model is evaluated which is followed by its evaluation in RPN-GAP-KPCA. Figures 6.6a and 6.6b show the impact of $C = \{0, 0.5(\text{default}), 1, 2, 4, 6, 8, 10\}$ on a single fold of modelling dataset. In both RPN-GAP-PCA and RPN-GAP-KPCA, higher C values result in lower recall and F-measure with comparatively small change in precision. As the penalty increases, the number of clusters decreases. Candidate selection depends on the classification score assigned to the proposals by the RPN. Due to the poorer classification accuracy of the RPN, lower IOU proposals are selected as candidates, subsequently discarding high IOU proposals. This results in an increase in missed lesions leading to lower recall and lower F-measure. Based on this performance, for RPN-GAP-PCA, $C = 1$ is selected as the optimal. Although the F-measure for RPN-GAP-PCA models with $C = 0$ and $C = 0.5$ is the same as $C = 1$, $C = 1$ is selected as optimal since it provides a better balance between number of clusters and performance. Similarly, for RPN-GAP-KPCA, $C = 1$ is selected as optimal.

Table 6.3 shows the impact of this optimal penalty in RPN-GAP-PCA and RPN-GAP-KPCA models, respectively (see Table 6.2 for performance of these models with default C). In the RPN-GAP-PCA model, increasing C to selected optimal value of 1 leads to an improvement in the model's performance. In comparison to the default C, use of optimal C increased F-measure by 0.38% to 0.23% over modelling and unseen datasets, respectively, which was due to an improvement of 0.55% to 0.37% in precision and 0.16% to 0.3% in recall. However, in the RPN-GAP-KPCA model, use of optimal C results in no change to a small drop in overall performance. Specifically, in datasets A, B, and E, increasing C leads to no change in performance whereas in datasets C and D,higher C results



(a) PCA (b) KPCA

Figure 6.6: Impact of C in to RPN-GAP-PCA and RPN-GAP-KPCA models.

in a small drop of 0.045% precision and 0.015% recall.

Dataset	**Optimal C in U-Detect-RPN**	**Precision**	**Recall**	**F-measure**
A	RPN-GAP-PCA	**84.89**	97.37	90.69
	RPN-GAP-KPCA	84.08	**99.26**	**91.02**
Overall External Test Sets	RPN-GAP-PCA	**75.91**	88.40	81.61
	RPN-GAP-KPCA	74.82	**90.36**	**81.82**
B	RPN-GAP-PCA	91.67	97.28	94.37
	RPN-GAP-KPCA	**92.51**	**99.43**	**95.83**
C	RPN-GAP-PCA	**67.05**	78.76	72.11
	RPN-GAP-KPCA	65.79	**80.57**	**72.19**
D	RPN-GAP-PCA	**76.41**	90.86	**82.91**
	RPN-GAP-KPCA	74.69	**93.13**	82.79
E	RPN-GAP-PCA	86.36	98.26	91.81
	RPN-GAP-KPCA	**86.64**	**99.21**	**92.38**

Table 6.3: Performance of optimal C in U-Detect-RPN model using RPN-GAP-PCA feature vector.

Optimal C has a comparatively lesser number of clusters, especially single-element clusters, than the default C. Reducing total number of clusters ensures that FPs are clustered with texturally similar background regions unlike default C where the high number of single-element clusters restricts the reduction of FPs as majority of the proposals are sent through to the base network without filtering. Also, the improved clustering limits the overshadowing of TP proposals by FP proposals in single-

element and small clusters. Therefore, due to the higher number of TPs and lower number of FPs, optimal C has higher precision in modelling and overall unseen test sets, respectively. However, as previously mentioned, with the reduced number of clusters and poorer classification accuracy of the RPN, the optimal C models had a higher number of missed lesions.

RPN-GAP-KPCA model with optimal C has poorer performance in datasets C and D as these datasets contain a higher number of challenging lesions that have high textural similarity with the background region. Due to low classification scores assigned to proposals covering these lesions as well as reduction in the number of clusters with higher C, these challenging lesions were missed which resulted in lower recall and F-measure. Due to higher overall F-measure, the RPN-GAP-KPCA model with default penalty is selected as optimal and used in all RPN-GAP-KPCA models mentioned from hereon. Likewise, due to the superior performance of optimal C in RPN-GAP-PCA model, this penalty is used in all RPN-GAP-PCA models referred from this point on.

6.2.1.4 Candidate Merging Method Application

This section reports the performance of the candidate merging method in RPN-GAP-PCA and RPN-GAP-KPCA models with optimal number of PCs/kernel PCs and optimal C. Table 6.4 shows the performance of these RPN-GAP-PCA and RPN-GAP-KPCA model with application of candidates merging method (see Table 6.3 for performance of these models without candidate merging method). Both models use optimal C. Overall, the candidate merging method improved the overall performance of both models due to an improvement in precision and, in many cases, recall. The higher precision is due to a reduction of FPs along with an increase in TPs. This is because, typically, the multiple FPs covering small sections of a lesion are merged into a single box which covers the entire lesion with high IOU $IOU > 0.5$ thereby converting the FP detections to a single TP detection. In some cases where the merged box has $IOU < 0.5$, the total number of FPs is still reduced from multiple FPs to a single low IOU FP. Occasionally, one of the overlapping boxes is a TP detection. Merging this TP detection with the overlapping FPs leads to a drop in TP along with a drop in total FPs. But these scenarios are relatively rare.

Lastly, there is no change in FNs as the candidate merging method does not influence any other aspects of the U-Detect-RPN method including clustering of proposals and classification score assigned

to proposals. Therefore, majority datasets had an improvement in recall with the use of merging due to an increase in TPs and no change in FNs. However, in datasets where the TPs are converted to low IOU FPs, there is a small drop in recall. Due to the best performance of RPN-GAP-KPCA model with candidate merging method (along with optimal kernel PCs and C) as well as better suitability of KPCA for dimension reduction, this model is selected as optimal U-Detect-RPN model. All references from hereon to the U-Detect-RPN model use this optimal setup.

Compared to adapted FRCNN (see Table 6.1), the U-Detect-RPN modelhas 1.33% to 8.42% higher precision due to 8.78% to 38.16% lower number of FPs and a relatively small drop of 0.12% to 0.21% in recall. In datasets B and E, the U-Detect-RPN model has 0.01% higher recall than the adapted FRCNN model. Thus, the overall F-measure is 0.7% to 5.05% higher. The improvement in precision and recall is due to improved filtering of proposals which restricted overshadowing of TP proposals as well as improved filtering of FP proposals. Single-fold performance of the U-Detect-RPN model is provided in Section B.1 in Appendix B. Therefore, the proposed U-Detect-RPN method successfully overcomes drawbacks of NMS.

Dataset	**U-Detect-RPN Models with Candidate Merging**	**Precision**	**Recall**	**F-measure**
A	RPN-GAP-PCA + merge	**84.59**	96.99	90.35
	RPN-GAP-KPCA + merge	84.51	**99.26**	**91.26**
Overall External Test Sets	RPN-GAP-PCA + merge	**77.17**	88.67	82.52
	RPN-GAP-KPCA + merge	76.53	**90.38**	**82.88**
B	RPN-GAP-PCA + merge	93.07	95.85	94.43
	RPN-GAP-KPCA + merge	**93.73**	**99.43**	**96.50**
C	RPN-GAP-PCA + merge	**68.57**	**80.81**	**74.02**
	RPN-GAP-KPCA + merge	**68.57**	**80.81**	**74.02**
D	RPN-GAP-PCA + merge	**77.20**	90.21	**83.09**
	RPN-GAP-KPCA + merge	75.10	**93.08**	83.02
E	RPN-GAP-PCA + merge	88.79	98.39	93.33
	RPN-GAP-KPCA + merge	**89.93**	**99.21**	**94.33**

Table 6.4: Impact of candidate merging method on U-Detect-RPN models using RPN-GAP-PCA and RPN-GAP-KPCA feature vectors.

6.2.2 U-Detect-Base

This section presents experimental results of the U-Detect-Base model. Similar to U-Detect-RPN, the feature vector selection, dimensionality reduction, x-means penalty, and candidates merging are

evaluated. This section starts with a description of the performance of the U-Detect-Base model with the entire Base-GAP feature vector for proposal description. Next, an evaluation of PCA and KPCA for reducing dimension of the Base-GAP feature vector is presented. This is followed by an analysis of C in x-means penalty.As with evaluation of various phases of the U-Detect-RPN model, all U-Detect-Base models presented in these sections do not use the candidate merging method (phase six) except in the final section where the impact of the candidate merging method is exclusively studied.

6.2.2.1 Base-GAP Feature Vector

Table 6.5 shows the performance of the U-Detect-Base model with Base-GAP feature vector. Compared to adapted FRCNN (see Table 6.1), U-Detect-Base model has a small improvement of 0.06% to 0.19% in the overall F-measure due to higher precision with no to small change in recall. This improvement in precision is a consequence of reduction in FPs with the use of the U-Detect method. As with U-Detect-RPN, FNs of adapted FRCNN are also missed in the U-Detect-Base model as the U-Detect models do not modify classification score assigned to proposals. Despite the overall improvement in performance of the U-Detect-Base model, the high dimension of this feature vector increases computation time and hinders optimal clustering. Thus, in the following section, use of PCA and KPCA for reducing its dimension is investigated.

Dataset	**U-Detect-Base**		
	Precision	**Recall**	**F-measure**
A	83.27	99.39	90.61
Overall External Test Sets	70.94	90.57	79.49
B	90.28	99.42	94.60
C	60.38	81.53	69.13
D	72.77	93.19	81.59
E	82.85	99.20	89.93

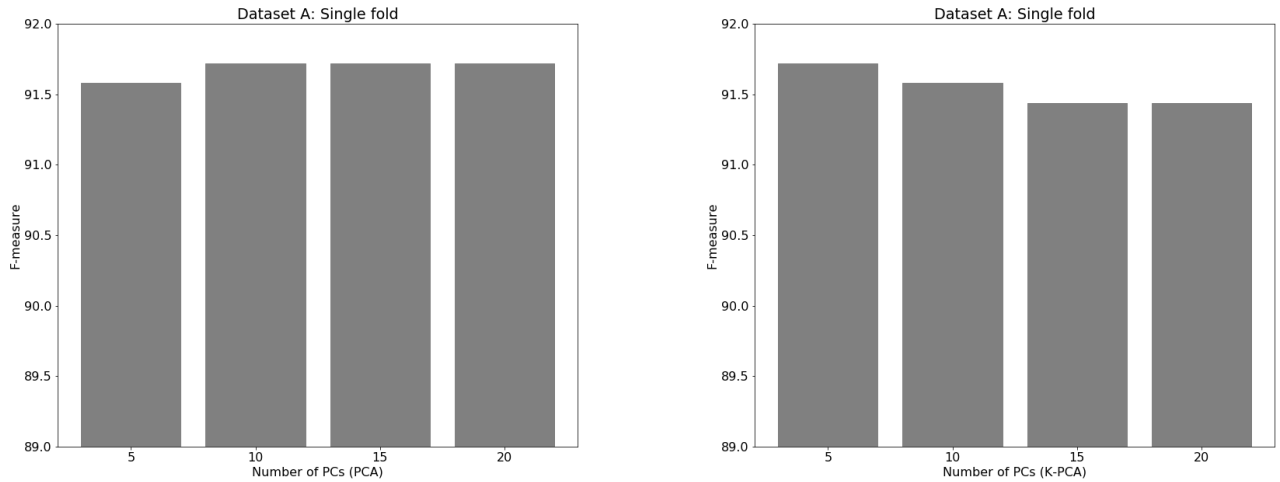
Table 6.5: Performance of U-Detect-Base model using Base-GAP feature vector for proposal description.

6.2.2.2 Dimension Reduction of Base Network Feature Vector

This section presents the impact of reduction in dimension of base-GAP feature vector using PCA and KPCA. All U-Detect-Base models are referred to by the feature vector used for description of proposals for the sake of brevity. For example, the Base-GAP model refers to the U-Detect-Base

model using base-GAP feature vector for proposal description.

Figures 6.7a and 6.7b show the performance of 5, 10, 15 and 20 PCs and kernel PCs on the same fold of modelling dataset as used in Section 6.2.1.2. As illustrated in Figure 6.7a, 5 PCs has the lowest performance and an increase to 10 PCs improves the performance after which the performance is unchanged with increase in PCs. The lower F-measure of the 5 PCs is due to higher number of FP in comparison to the other three models (10 PCs, 15 PCs and 20 PCs). As the information contained in 5 PCs is lower, proposals are not clustered correctly leading to incorrect filtering out of FP detections. Use of 15 PCs provides a good balance between small feature size and performance. Thus, 15 PCs is selected as optimal. The U-Detect-Base model using 15 PCs is referred to as Base-GAP-PCA. On the other hand, the lowest number of kernel PCs (5 PCs) has the best performance as shown in Figure 6.7b. Increasing the number of kernel PCs leads to an increase in FPs with no change in FN resulting in a drop in precision and F-measure with higher number of kernel PCs (recall is unchanged). This is because of an increase in the number of clusters with higher kernel PCs. Based on this performance, 5 PCs is selected as optimal and the U-Detect-Base model using this feature vector is referred to as Base-GAP-KPCA.



(a) PCA (b) KPCA

Figure 6.7: Evaluation of PCA and KPCA components of Base-GAP feature vector.

Table 6.6 shows the performance of Base-GAP-PCA and Base-GAP-KPCA. Both models outperform Base-GAP (see Table 6.5) and adapted FRCNN models (see Tables 6.1). Use of PCA has

a similar impact on the overall performance as seen in Section 6.2.1.2. Base-GAP-PCA model has higher F-measure than Base-GAP model; 0.16% in modelling dataset and 1.01% in overall unseen test sets. This is due to improvement of 0.42% to 1.82% in precision along with relatively smaller drop in recall of 0.2% to 0.49% in modelling and unseen test sets, respectively. Similarly, Base-GAP-KPCA outperforms the Base-GAP model. Here, the F-measure improved by 0.18% to 0.9% over modelling and external test sets, respectively which is the result of an increase in precision (0.41% to 1.62%) with a comparatively smaller drop in recall (0.14% to 0.31%). Also, both Base-GAP-PCA and Base-GAP-KPCA models outperforms the adapted FRCNN model with higher precision and a relatively small drop in recall.

Dataset	**U-Detect-Base Model**	**Precision**	**Recall**	**F-measure**
A	Base-GAP-PCA	**83.69**	99.19	90.77
	Base-GAP-KPCA	83.68	**99.25**	**90.79**
Overall External	Base-GAP-PCA	**72.76**	90.08	**80.50**
Test Sets	Base-GAP-KPCA	72.56	**90.26**	80.39
B	Base-GAP-PCA	**90.60**	**99.28**	**94.71**
	Base-GAP-KPCA	90.48	99.14	94.57
C	Base-GAP-PCA	**61.92**	80.60	69.81
	Base-GAP-KPCA	61.74	**80.98**	**69.86**
D	Base-GAP-PCA	**75.72**	92.85	**83.31**
	Base-GAP-KPCA	75.13	**93.03**	83.01
E	Base-GAP-PCA	83.38	**99.20**	**90.22**
	Base-GAP-KPCA	**83.41**	99.07	90.17

Table 6.6: Impact of dimension reduction using PCA and KPCA on Base-GAP features on performance of U-Detect-Base model.

Impact of reducing the dimension of the base-GAP feature vector using PCA and KPCA is similar to that seen in U-Detect-RPN described in Section 6.1.2. Like U-Detect-RPN, using PCA and KPCA feature vectors led to an overall reduction in FPs leading to the higher precision of Base-GAP-PCA and Base-GAP-KPCA in comparison to the Base-GAP and adapted FRCNN models. Likewise, due the reduced textural information in the PCA and KPCA feature vectors, these models had comparatively higher number of FNs. Also, due to the higher quality of the KPCA feature vector, the Base-GAP-KPCA model has higher TPs than the Base-GAP-PCA model along with higher FPs. Therefore, Base-GAP-PCA has 0.01% to 0.2% higher precision than Base-GAP-KPCA but 0.06% lower recall. Due to the larger drop in precision, Base-GAP-PCA has 0.11% higher F-measure overall. In the

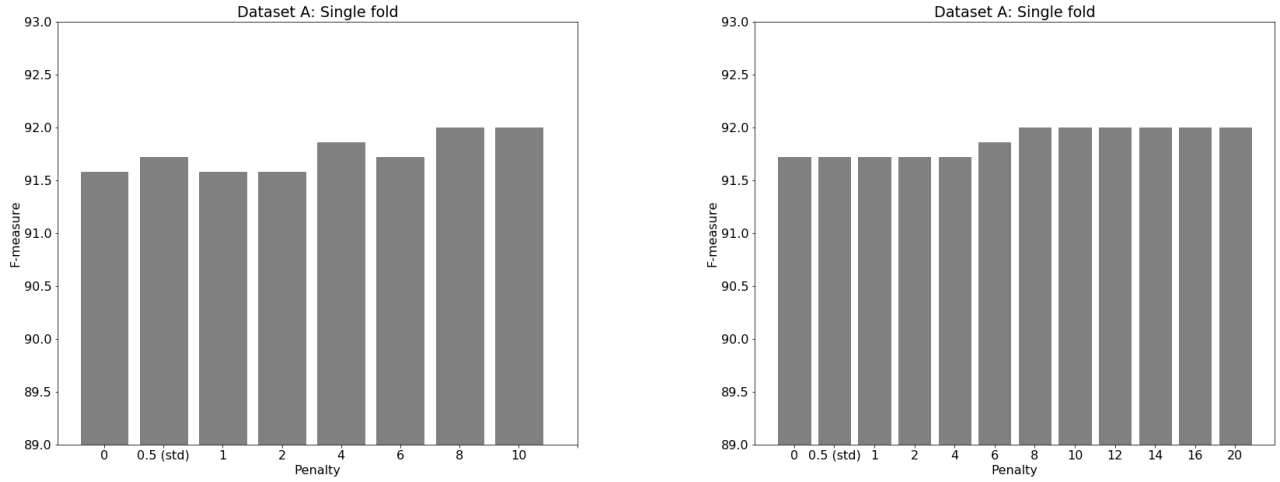
following section, the impact of variation in BIC penalty in both these models is investigated.

6.2.2.3 X-means Penalty Selection

This section evaluates the impact of C on the performance of Base-GAP-PCA and Base-GAP-KPCA models. Both models use the optimal number of PCs and kernel PCs identified in the previous section. Figure 6.8a shows the impact of $C = \{0, 0.5(\text{default}), 1, 2, 4, 6, 8, 10\}$ in the Base-GAP-PCA model. Figure 6.8b shows the impact of $C = \{0, 0.5(\text{default}), 1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$ on the performance of Base-GAP-KPCA model. Both these analyses were conducted on the same fold of modelling dataset as used in Section 6.2.1.3 for U-Detect-RPN.

Overall, an increase in C improves F-measure due to reduction in FPs in a similar manner as described in Section 6.2.1.3. As C increases, the number of clusters, especially single-element clusters. This reduction in the number of clusters with higher C value is shown in Figure 6.9. Figure 6.10 illustrates the reduction in the high number of clusters when C is increased. Higher C values promote better clustering of proposals which in turn reduces the number of FPs with negligible impact on the number of correct detections. An example of reduction in FPs with higher penalty is shown in Figure 6.11. Thus, increasing the penalty leads to increase in precision with recall remaining relatively constant. Smaller range of C was investigated for Base-GAP-PCA as values higher than 10 led to an undesired reduction in the number of clusters. Based on the performance on the modelling dataset, $C = 10$ is selected as optimal for Base-GAP-PCA and $C = 14$ for Base-GAP-KPCA. In U-Detect-RPN, higher penalty led to lower F-measure and the selected optimal value of C is much lower than optimal value of C selected in U-Detect-Base. As previously mentioned, higher C values reduce the number of clusters i.e. larger number of proposals in each cluster. Due to the comparatively higher classification accuracy of the base network, better quality of candidates (proposals with high IOU) were correctly selected as candidates from these large clusters. This results in further reduction in FPs without significant negative impact on correct detections.

Table 6.7 shows the performance of optimal C in Base-GAP-PCA and Base-GAP-KPCA models, respectively (see Table 6.6 for performance of these models with default C). In Base-GAP-PCA, changing C from default value of 0.5 to the selected optimal value of 10 led to an increase of 0.22% to 2.31% in F-measure owing to its higher precision. As seen with U-Detect-RPN in Section 6.2.1.3,



(a) PCA

(b) KPCA

Figure 6.8: Evaluation of C in U-Detect-Base model using Base-GAP-PCA and Base-GAP-KPCA feature vectors.

optimal C value in Base-GAP-PCA leads to a reduction in FPs as well as number of correct detections. This resulted in its higher precision of 0.37% to 4.1%. On the other hand, recall of the Base-GAP-PCA model had a relatively smaller drop of 0.38% in overall unseen test sets (no change in modelling dataset). Similarly, compared to adapted FRCNN, Base-GAP-PCA model with optimal C has 0.44% to 2.94% higher F-measure due to 0.88% to 5.46% improvement in precision and 0.2% to 0.89% drop in recall in modelling and unseen test sets, respectively.

Dataset	**Optimal C in U-Detect-Base**	**Precision**	**Recall**	**F-measure**
A	Base-GAP-PCA	84.06	**99.19**	90.99
	Base-GAP-KPCA	**84.41**	99.19	**91.19**
Overall External Test Sets	Base-GAP-PCA	75.98	**89.70**	82.24
	Base-GAP-KPCA	**77.15**	89.69	**82.94**
B	Base-GAP-PCA	91.52	**99.14**	95.16
	Base-GAP-KPCA	**92.12**	**99.14**	**95.49**
C	Base-GAP-PCA	66.02	79.82	72.12
	Base-GAP-KPCA	**68.02**	**79.83**	**73.38**
D	Base-GAP-PCA	78.08	92.70	84.68
	Base-GAP-KPCA	**78.29**	**92.72**	**84.82**
E	Base-GAP-PCA	85.39	**99.07**	91.50
	Base-GAP-KPCA	**87.00**	**99.07**	**92.56**

Table 6.7: Impact of the optimal C in U-Detect-Base model using Base-GAP-PCA feature vector.

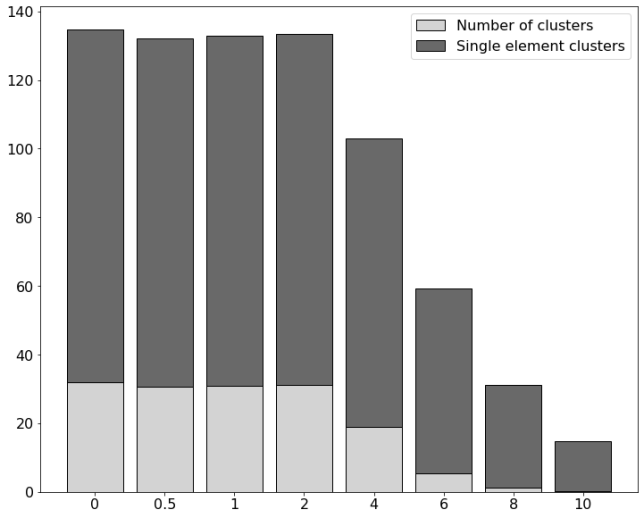
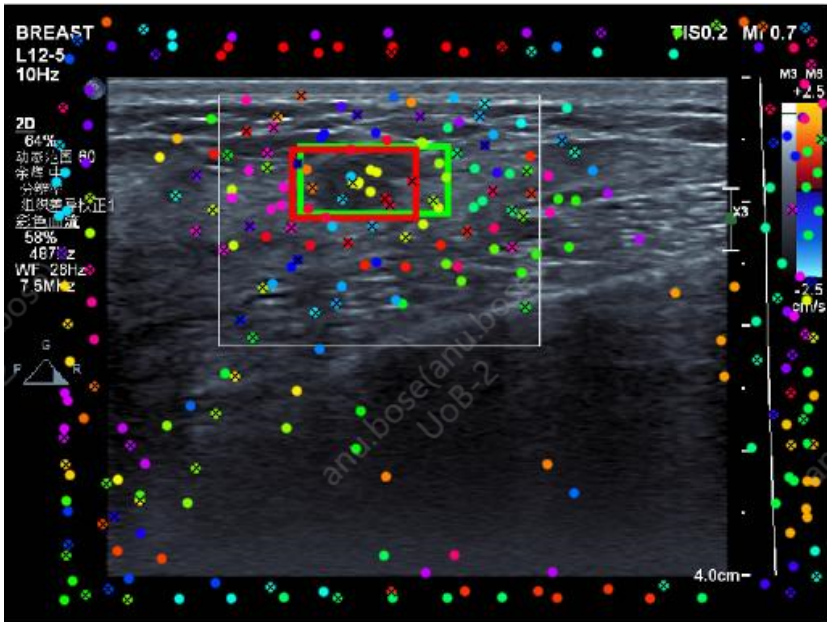


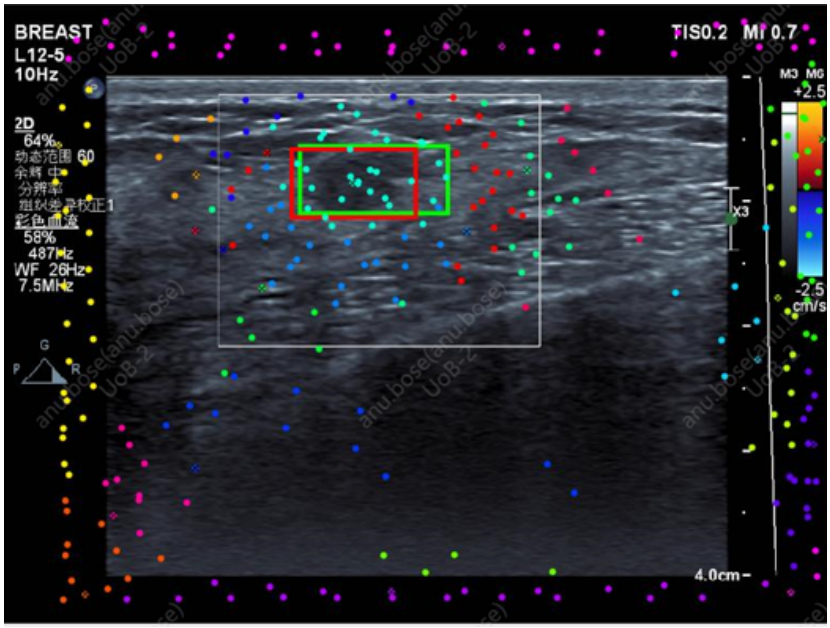
Figure 6.9: Impact of change in C on total number of clusters.

Use of optimal C in Base-GAP-KPCA improved precision by 0.73% to 4.59% with only a smaller drop of 0.06% to 0.57% in recall leading to an overall improvement of F-measure in comparison to the Base-GAP model. Likewise, the Base-GAP-KPCA model with optimal C also outperforms adapted FRCNN with 1.23% to 6.63% higher precision and only 0.2% to 0.9% drop in recall leading to an overall higher F-measure of 0.64% to 3.64% (see Table 6.1 for performance of adapted FRCNN). Here, the optimal C Base-GAP-KPCA model has a lower number of FPs than adapted FRCNN. An example of the reduction in one of the challenging additional boxes (FP boxes with $IOU = 0$) of adapted FRCNN from dataset C in the Base-GAP-KPCA model (optimal C) is shown in Figure 6.13. Majority of the TP cases from the adapted FRCNN model are retained. Figure 6.12 shows clusters of TP cases common to both models. Multiple lesions are also detected in the Base-GAP-KPCA model as shown in Figure 6.14.

The small increase in missed lesions compared to the adapted FRCNN is due to improper clustering. In these cases, proposals covering lesion and lesion-like background regions are clustered together. Here, due to the higher score of the proposal covering lesion-like background regions, it is selected as the candidate from the cluster leading to a missed lesion. An example of this is shown in Figure 6.15. Apart from this, lesions missed by the adapted FRCNN model (FNs) are also missed by this model due to the low classification scores assigned to all proposals covering the lesion. An example of such missed lesions common to both models and their clusters is shown in Figure 6.16.

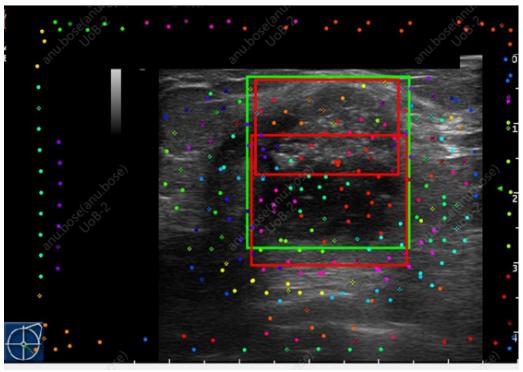


(a) $C = 0$

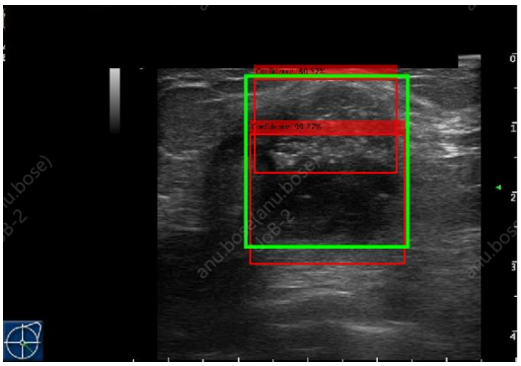


(b) $C = 10$

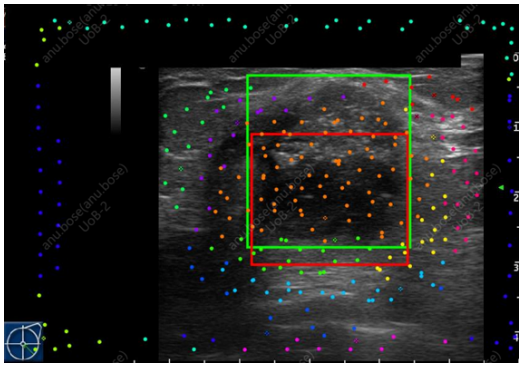
Figure 6.10: Illustration of the impact of C on number of clusters. Each dot represents the centre point of a proposal and each colour represents one cluster. Dots marked with a cross ('x') were selected candidates from their respective clusters.



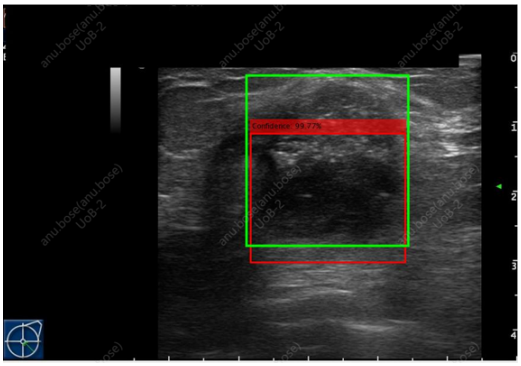
(a) $C = 0$ cluster



(b) $C = 0$ output

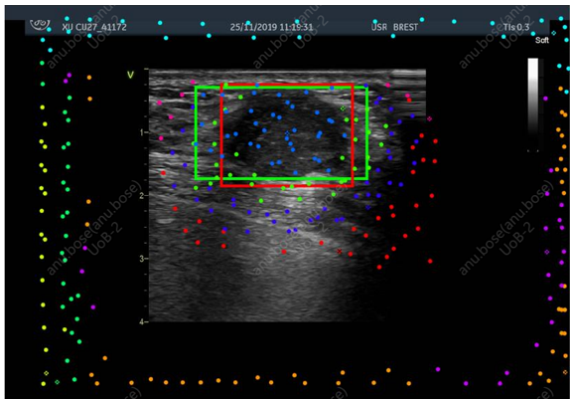


(c) $C = 10$ cluster

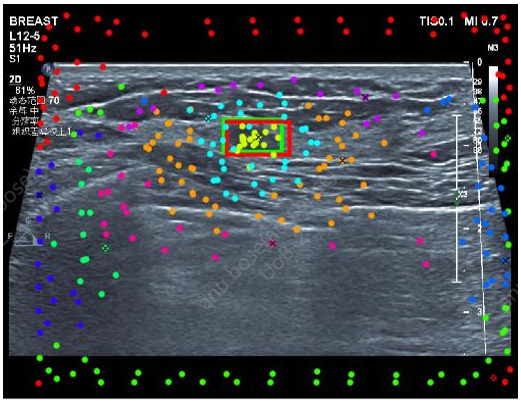


(d) $C = 10$ output

Figure 6.11: Sample reduction of overlapping low IOU FP with higher C.

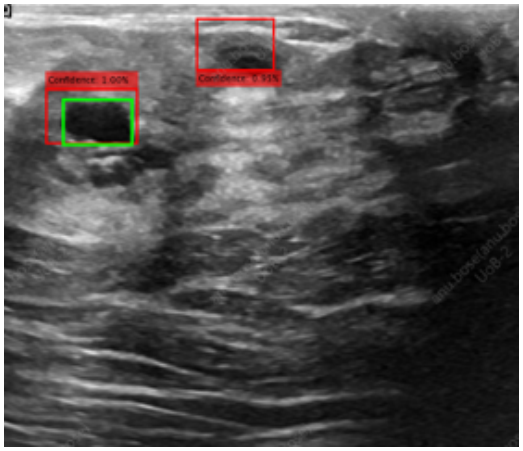


(a)

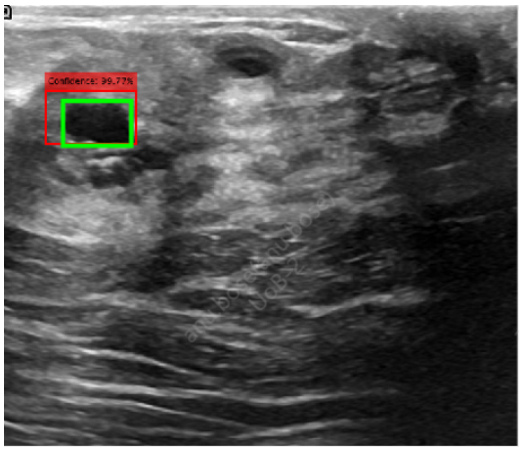


(b)

Figure 6.12: Clusters generated in TP detections of U-Detect-Base model using Base-GAP-KPCA feature vector.

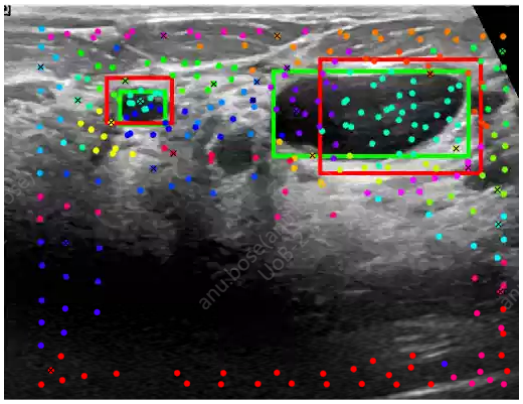


(a) Adapted FRCNN output

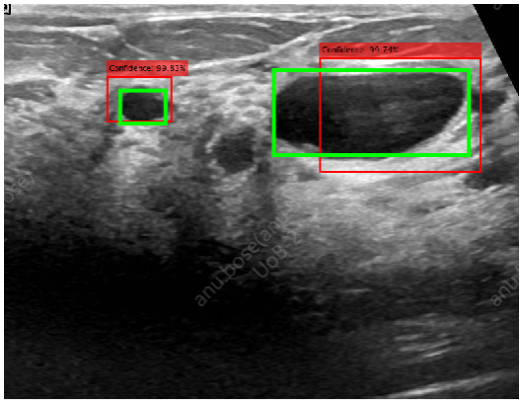


(b) Base-GAP+KPCA Output

Figure 6.13: Additional box reduction in U-Detect-Base model using Base-GAP-KPCA feature vector.

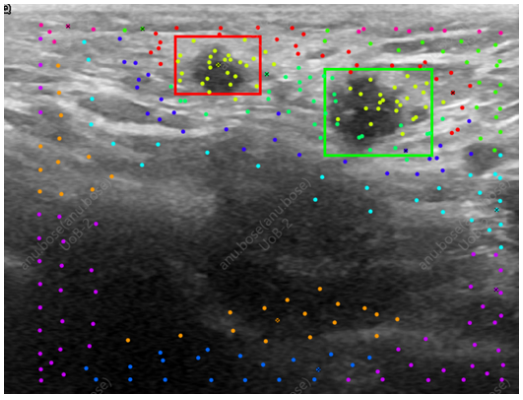


(a) Clusters

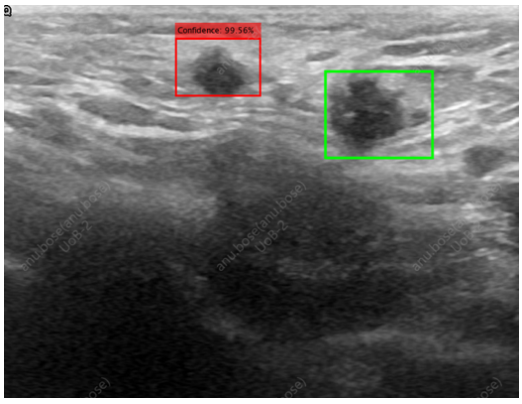


(b) Output

Figure 6.14: Detection of multiple lesions by U-Detect-Base model using Base-GAP-KPCA feature vector.

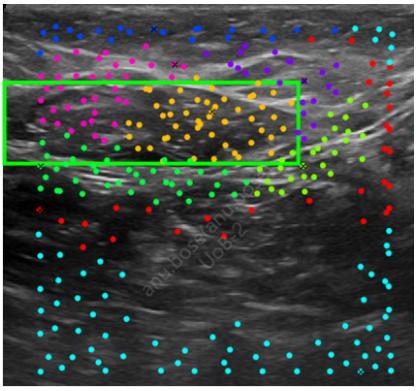


(a) Clusters

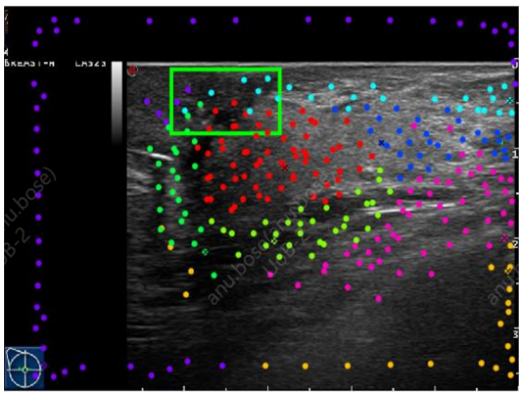


(b) Output Detections

Figure 6.15: Increased missed lesion in U-Detect-Base model using Base-GAP-KPCA feature vector.



(a)



(b)

Figure 6.16: Clusters generated in FN cases of U-Detect-Base model using Base-GAP-KPCA feature vector.

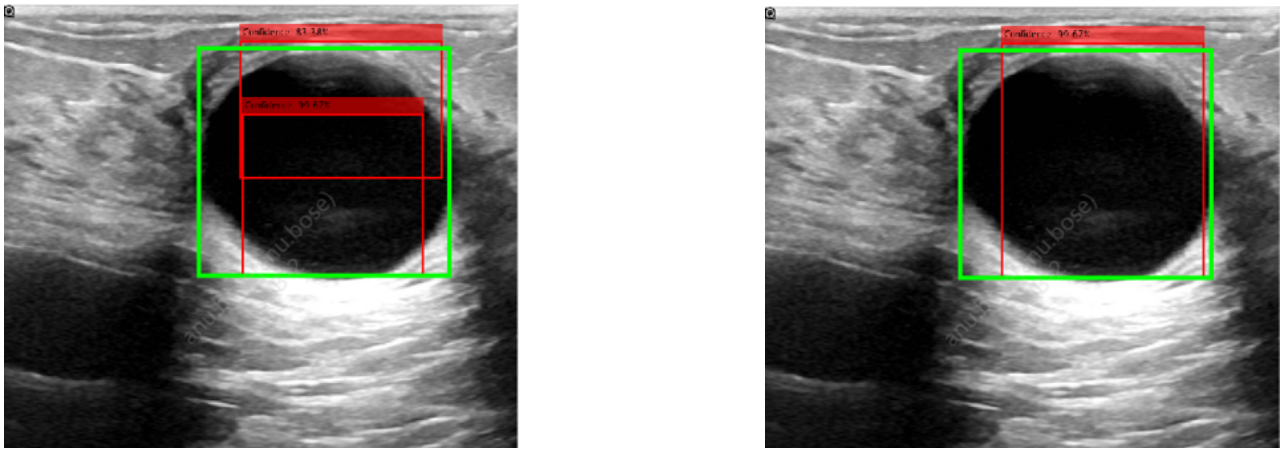
6.2.2.4 Candidate Merging Method Application

This section details the impact of candidate merging method (phase five) on the performance of Base-GAP-PCA and Base-GAP-KPCA models. Note that both these models use the optimal number of PCs/kernel PCs and their respective optimal C (see Table 6.7 for performance of these models without candidate merging method). Table 6.8 shows the impact of the candidate merging method in these models.Irrespective of the model, use of the merging mechanism led to a reduction of FPs thereby improving overall precision of the model. Furthermore, some of the merged boxes cover the lesion with $IOU \geq 0.5$ resulting in an increase in TPs. An example of such a TP merged box is as shown in Figure 6.17.

Base-GAP-KPCA outperforms Base-GAP-PCA model with and without merging. Thus, due to the best performance as well as applicability of the Base-GAP-KPCA model with optimal number of PCs, C and candidate merging method, it is considered as the optimal setup for U-Detect-Base. All mentions of U-Detect-Base models from hereon refer to this model. The U-Detect-Base model outperforms adapted FRCNN with 1.18% to 7.87% higher precision resulting from 6.08% to 38.89% lower number of FPs and only a small drop of 0.13% to 1.72% in recall. Single-fold performance of the U-Detect-RPN model is provided in Section B.2 in Appendix B. Therefore, using the clustering network to filter the final detections of the base network is effective in addressing drawbacks of NMS while causing minimal negative impact on the number of correct detections.

Dataset	U-Detect-Base Models with Candidate Merging	Precision	Recall	F-measure
A	Base-GAP-PCA + merge	84.22	**99.25**	91.11
	Base-GAP-KPCA + merge	**84.41**	99.19	**91.19**
Overall External Test Sets	Base-GAP-PCA + merge	76.94	**89.74**	82.84
	Base-GAP-KPCA + merge	**77.15**	89.69	**82.94**
B	Base-GAP-PCA + merge	91.87	**99.14**	95.35
	Base-GAP-KPCA + merge	**92.12**	**99.14**	**95.49**
C	Base-GAP-PCA + merge	67.25	**79.96**	72.97
	Base-GAP-KPCA + merge	**68.02**	79.83	**73.38**
D	Base-GAP-PCA + merge	**78.53**	92.70	**84.96**
	Base-GAP-KPCA + merge	78.29	**92.72**	84.82
E	Base-GAP-PCA + merge	**87.28**	**99.07**	**92.72**
	Base-GAP-KPCA + merge	87.00	**99.07**	92.56

Table 6.8: Impact of candidate merging method on U-Detect-Base models using Base-GAP-PCA and Base-GAP-KPCA feature vectors.



(a) Without merging

(b) With merging

Figure 6.17: Impact of candidate merging method on reduction of overlapping FPs.

6.2.3 U-Detect-RPN vs U-Detect-Base

This section compares the performance of the two U-Detect-RPN (see Table 6.4) and U-Detect-Base (see Table 6.8). U-Detect-Base has a comparatively lower performance than U-Detect-RPN due to lower precision and recall. In both models, the base network has higher classification accuracy than the RPN as shown in Table 6.9. However, in U-Detect-RPN, the quality of proposals sent through to the base network is superior leading to a better overall performance. On the other hand, U-Detect-Base uses comparatively poorer proposals due to their improper filtering by NMS. Therefore, despite

no change in the classification accuracy, U-Detect-Base has a comparatively lower F-measure than U-Detect-RPN.

IOU	**Dataset A**		**Dataset C**		**Dataset D**	
	RPN	Base	RPN	Base	RPN	Base
0	18.21	17.61	35.85	33.25	22.81	23.90
(0,0.5)	74.53	71.38	56.22	53.93	69.74	64.21
[0.5,1]	7.26	11.01	7.92	12.82	7.44	11.88

Table 6.9: Percentage of proposals in the different IOU ranges after classification and bounding box regression by RPN and base network of the adapted FRCNN model.

Overall, the U-Detect-RPN model outperforms the U-Detect-Base model with 0.1% and 1.77% higher F-measure. However, in dataset D, U-Detect-Base had 1.8% higher F-measure due to 3.19% higher precision resulting from lower number of FPs and 0.36% lower recall. Irrespective of the location, both U-Detect-RPN and U-Detect-Base outperform adapted FRCNN through reduction of NMS issue cases including single and multiple FPs with minimal negative impact on number of correct detections. Therefore both models have higher precision than the adapted FRCNN. An important distinction to note here is that along with higher precision, U-Detect-RPN also has higher recall than the adapted FRCNN model due to higher number of TPs. In terms of computation time, U-Detect-Base is faster. Average computation time of U-Detect-Base is 0.96 seconds whereas that for U-Detect-RPN is 3.96 seconds. In comparison to computation time of adapted FRCNN (0.39 seconds), U-Detect-Base adds only an additional 0.57 seconds.

Despite the improvement to the overall number of FPs, two types of FPs still remain in both these models, namely, single low IOU FPs and FP+FN cases. Single low IOU FPs are cases where the output of the model is a single output detection with $IOU < 0.5$. Secondly, FP+FN are scenarios where the single output detection covers lesion-like background region and the lesion is completely missed thereby creating a FP and a FN. Additionally, since the classification scores assigned to proposals is not altered by the clustering mechanism, there is no change in FNs in comparison to the adapted FRCNN model despite high quality clustering.

These common issues cases are caused due to poor classification scores assigned to proposals. In these cases, there exists a high IOU, TP detection in the cluster but due to its poorer classification

score, it is not selected as the candidate of the cluster and discarded as redundant proposals. In single low IOU FP output of U-Detect-Base, the cluster containing this low IOU FP also contains at least one high IOU proposal with an average IOU of 55% and average rank of 4.56 in the cluster on the basis of its classification score. Thus, despite the presence of TP, the low IOU FP proposal is selected. In the FN+FP scenario, best proposals in the cluster have an average IOU of 0.48 and an average rank of 7.6. Although here the best proposal is not a TP, it still detects the lesion instead of the selected candidate which is a FP box with IOU 0 with the GT lesion. Finally, in FN cases, the cluster covering the lesion contains proposals with an average of IOU 0.52 but as these proposals are labelled as background, they are discarded and the lesion is missed.

6.3 Discussion

This chapter successfully addressed the issue cases of the adapted FRCNN model caused by improper filtering of the proposals by NMS using a novel U-Detect method. This section discusses investigations relevant to the proposed U-Detect method. In phase three of this method, candidates from each cluster are selected based on their classification score. However, the common practice in clustering methods is to use the centroid of a cluster as its candidate since it provides a good representation of that cluster. Therefore, it is worth investigating the impact of using centroid as the candidate selection method is also studied in the U-Detect method. Specifically, the impact of the centroid method of candidate selection was studied in U-Detect-RPN where proposals are described using RPN-GAP features and default C was used. Distance between proposals is measured using cosine similarity as described in Section 6.1.3. Also, the candidate merging method was not applied to the final candidates. An important point to note here is that centroid is typically not a proposal as it is the mean of any given set of proposals in a cluster. Since we require the candidate to be one of the proposals, the selected candidate in this case is the proposal *closest* to the centroid, where *closeness* is measured in terms of textural similarity. Equation 6.3 defines this method of candidate selection.

$$candidate_j = min(dist(p_{ij}, centroid_j)) \qquad for \;\; i \in [1,n],\; j \in [1,m]$$

where (6.3)

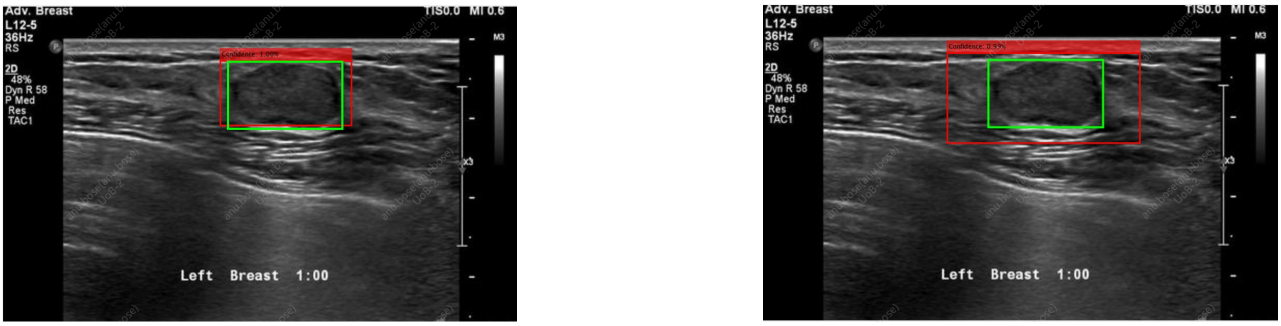
$$dist(p_{ij}, centroid_j) = 1 - cos(p_{ij}, centroid_j) \quad for\; i \in [1,n],\; j \in [1,m]$$

Here, p_{ij} represents i^{th} proposal in j^{th} cluster, n is the total number of proposals in that cluster and

m is the total number of clusters. Thus, the proposal that is most texturally similar to the centroid is selected as the candidate for that cluster. Table 6.10 shows the performance of the centroid method for candidate selection in comparison RPN score. The centroid method achieved lower performance compared to the adapted FRCNN (see Table 6.1) with 0.23% to 0.84% lower F-measure. This drop in performance is due to higher number of FPs along with reduction in the number of correctly detected lesions. Here, as the selected candidate is the best mean representation of all proposals in the cluster, lower IOU boxes are selected as candidates in spite of the presence of high IOU boxes in the cluster, thereby converting TP detections of the adapted FRCNN model to low IOU FPs. An example of this is shown in Figure 6.18.

Dataset	**U-Detect-RPN Model**	**Precision**	**Recall**	**F-measure**
A	RPN-GAP (Centroid)	82.81	**99.39**	90.34
	RPN-GAP (RPN Score)	**83.45**	**99.39**	**90.71**
Overall External	RPN-GAP (Centroid)	70.08	89.80	78.66
Test Sets	RPN-GAP (RPN Score)	**72.21**	**90.08**	**80.11**

Table 6.10: Comparison to candidate selection methods in U-Detect-RPN model using RPN-GAP features.



(a) Adapted FRCNN

(b) RPN-GAP (centroid)

Figure 6.18: Increase in low IOU FP reduction using centroid for candidate selection in U-Detect-RPN model.

However, compared to the adapted FRCNN, this centroid-based U-Detect-RPN model had an overall lower number of additional boxes due to the effectiveness of the U-Detect method. Here, FP proposals covering background lesion-like regions are clustered together. Their centroid is either removed during NMS due to high overlap with other candidates or due to the low confidence score assigned by the base network. In spite of this drop in additional boxes, total FPs of the centroid

model was still higher than the adapted FRCNN due to the high number of low IOU FPs. Along with the drop in TPs through conversion to low IOU FPs, there was also an increase in FNs. Some of the challenging lesions with background-like texture are clustered with proposals covering background due to similarity in texture. As majority proposals in this cluster cover the background region, centroid also covers the background region. Thus, the proposal containing the lesion is discarded leading to a missed lesion. In other cases, candidates contain significant percentage of margin leading to assignment of low classification score by the base network.

Compared to the model using RPN classification score for candidate selection, the model using centroid method of candidate selection had 0.41% to 1.84% lower F-measure. Unlike the centroid method, use of RPN score (typically) ensures that the proposal with highest IOU with the lesion is selected as the candidate which led to reduction in FPs as well as an increase in number of detected lesions. Number of FNs of both methods is almost the same. While challenging lesions in centroid-based method are missed due to selection of background-covering proposal as candidate, in RPN-score-based model lesions are missed due to poor classification score assigned by the RPN. Both models have common FNs with adapted FRCNN model as the U-Detect models do not modify the classification score of proposals.

Initially, tests were conducted using k-means clustering. This required predefining the number of clusters. Although the performance was reliably high, using predefined k limited any appropriate adaptation required for an individual image. Use of k-means also limits application of the U-Detect method in detectors of other types of lesions where the average number of lesions is higher. Therefore, x-means clustering is used to ensure that the appropriate number of clusters is used for each image without relying on a single predefined value making the network more adaptable to other detectors and datasets/domains. X-means clustering [1] employs BIC as the metric to merge or break clusters. The BIC equation detailed in the original x-means paper uses incorrect maximum likelihood estimate (MLE) for variance [156]. When used in this format, all proposals were placed in a single cluster. In this work, the correct form of this equation as detailed in [156] is used.

6.4 Summary

This chapter presented a new approach based on unsupervised learning for reducing false positive detection for breast lesion detection in ultrasound images. In particular, a new method called U-Detect was proposed and consists of learnable features extraction, dimensionality reduction, proposals clustering using x-means clustering method, candidates selection, and candidates merging. U-Detect is only used during the test-stage of a pretrained detector thereby adding no additional computational cost to model training. The chapter consists of two main parts. The first part introduced a new method called U-Detect-RPN which uses features extracted from the last convolution layer of inception-resnet-B block as a texture information to reduce the FPs.

The second part presented another variant of U-Detect called U-Detect-Base which uses features extracted from the last convolution layer of inception-resnet-B block as a texture information to reduce the FPs. Using the datasets detailed in Section 4.1, both U-Detect-RPN and U-Detect-Base models outperformed the adapted FRCNN by reducing FPs by 8.45% to 47.53% without significant negative impact of 0.07% to 0.98% on recall. In some cases, U-Detect-RPN improved the total number of correct detections along with the reduction of FPs. Of the two models, U-Detect-RPN outperformed U-Detect-Base. This is because of the higher quality of proposals sent to the base network in U-Detect-RPN which improved the overall performance. However, U-Detect-RPN had higher computation time while U- Detect-Base had computation time comparable to that of the adapted FRCNN.

The main findings of this chapter can be summarised as follows:

- The FPs issue was found to be due to the poor classification accuracy of the network and the post-processing method, NMS, used to remove redundant proposals.
- The study in this chapter demonstrated that the U-Detect method reduces the FPs and increases the performance of detecting breast lesions in ultrasound images.
- The hypothesis of clustering proposals with similar learnable texture features (RPN or Base) for FPs reduction was validated.
- U-Detect-RPN outperformed U-Detect-Base and produced high quality proposals sent to the base network which improved the overall performance.

Out of the remaining FPs in both U-Detect-RPN and U-Detect-Base, single low IOU FPs were the most common. This is because of the poor classification accuracy. In particular, the cluster containing these single low IOU FP also contain high IOU detections. However, due to the lower score assigned to the high IOU detections, the single low IOU FP in that cluster is selected as the candidate. In some cases, additional boxes (or detections) are also clustered with proposals covering the lesion with high IOU. But due to the higher classification score of the additional box, it is selected as the candidate leading to a missed lesion and a FP. In the following chapter, these common issue cases are addressed with the help of handcrafted features designed to improve the classification score providing an adaption of the U-Detect method.

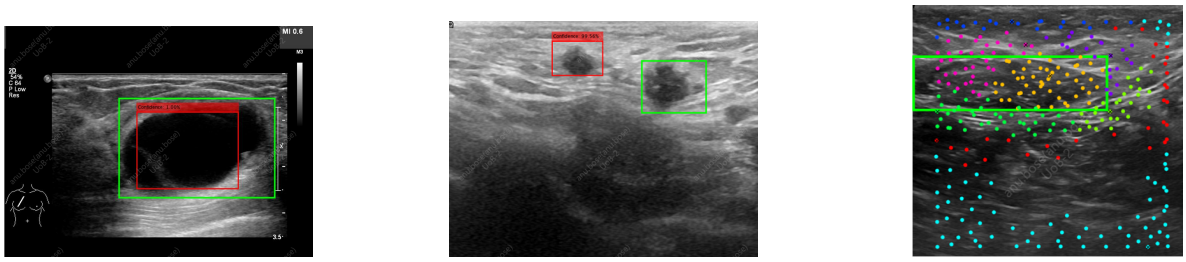
Chapter 7

U-DetectH: A Classification-based Approach using Handcrafted Features for FP Reduction

In Chapter 6, a novel method (U-Detect) was presented which successfully reduced FP detections of adapted FRCN. Based on the learnable features used by the U-Detect method, two networks called U-Detect-RPN and U-Detect-Base were proposed. Both U-Detect-RPN and U-Detect-Base outperformed adapted FRCNN through successful reduction of FPs. Additionally, U-Detect-RPN also had a higher number of correct detections than the adapted FRCNN model in datasets B and E. The performance of U-Detect based learned features in FPs reduction provides motivation to further improve the proposals selection using engineered features. Therefore, the aim of this chapter is to introduce a new set of features to improve the proposals selection performance and ultimately reduce FP detections.

U-Detect models have the following common types of issue cases: single low IOU FP, FP + FN, and FN, in that order of commonality, as discussed in Section 6.2.3 of Chapter 6. Single low IOU FPs are cases where the model outputs a single detection covering the lesion with $IOU < 0.5$. In FP + FN, the output is a single detection that covers lesion-like regions in the background (FP) thereby completely missing the lesion (FN). Figure 7.1 illustrates these issue cases. As presented in Section 6.2.3 of Chapter 6, the cause of the cases is incorrect assignment of classification scores to the proposals. In the majority of these cases, there exists at least one high-IOU TP proposal in the

same cluster as the output issue case. However, the TP proposal is assigned a lower score than the FP proposal. Therefore, from this cluster, the FP proposal is selected as the candidate while the high-IOU proposal is discarded as redundant. In case of FN, all proposals covering the lesion are assigned low scores, classifying them as background. Therefore, despite the good quality of clusters, all proposals, including the high-IOU TPs, are discarded.



(a) Single FP (b) FN+FP (c) FN

Figure 7.1: Common issue cases of U-Detect models.

In this chapter, these cases are addressed through improving the classification score assigned to the proposals with the help of handcrafted features. We hypothesise that the improvement in classification scores would lead to selection of higher IOU TP candidates, consequently reducing the number of FPs and missed detections where the definition and selection of the handcrafted features are inspired by the domain knowledge of the lesion characteristics.

The proposed approach is referred to as U-DetectH. It builds on the U-Detect method of Chapter 6 and differs from it in two phases. In phase one, while U-Detect extracts only learned features for each proposal, U-DetectH extracts learned as well as handcrafted features. Second, a new phase is added where the classification score assigned to each proposal is modified using SVM model(s) trained on the extracted handcrafted features. Remainder of the phases of both methods (U-Detect and U-DetectH) are the same. Therefore, this chapter focuses mainly on phases novel to U-DetectH. First, Section 7.1 details the proposed U-DetectH method. Next, Section 7.2 presents the performance and analysis of U-DetectH and several handcrafted features. Section 7.3 discusses key points related to this work and Section 7.4 provides a summary of the key findings from this chapter.

7.1 U-DetectH for False Positive Reduction

In this chapter, we propose a new approach of adapting the U-Detect method to reduce FPs using handcrafted features. The handcrafted features have been identified and selected to extract important texture and morphological features of breast lesions that would help in distinguishing proposals containing lesions from those containing background regions. These features were determined with reference to the characteristics of breast lesions in US images used by radiologists to assign BI-RADS score as detailed in Section 2.1 of Chapter 2. An SVM classifier is trained using these handcrafted features to improve the classification score assigned to the proposals by the RPN/base network. This, in turn, helps reduce FPs through selection of high IOU TP proposals.

Figure 7.2 provides an overview of the proposed U-DetectH method. U-DetectH consists of the following six phases: feature extraction, dimensionality reduction, proposal clustering, classification score update, candidate selection, and candidate merging. In phase one, learned and handcrafted features of every proposal are extracted. In phase two, dimension of both learned and handcrafted feature vectors is reduced. These dimensionally reduced learned feature vectors are then used in the following phase (phase three) to cluster proposals based on their texture using x-means clustering. In the fourth phase, classification decision fusion of SVM trained on handcrafted features and RPN/base network is performed which outputs a new, updated score for every proposal. Using this updated score, candidates from each cluster are selected in phase five.

Classification decision fusion in this manner is used for the following two reasons. First, the quality of the clusters formed using learned features is reliable as evidenced through the high performance of the final U-Detect-RPN and U-Detect-Base models in Chapter 6. Second, as described in Section 6.2.3 of Chapter 6, the common FP detections of U-Detect models (single low IOU FPs and FP+FNs) are caused due to incorrect candidate selection which is a result of low classification score assigned to high IOU proposals in comparison to FP proposal present in the same cluster. Likewise, FNs of the U-Detect models are caused due to low classification score assigned to all proposals containing the lesion. Therefore, through this decision fusion, classification scores assigned to high IOU proposals can be improved which would in turn lead to their selection as candidates from their respective clusters. Through such an improvement in classification score and candidate selection, the issue cases

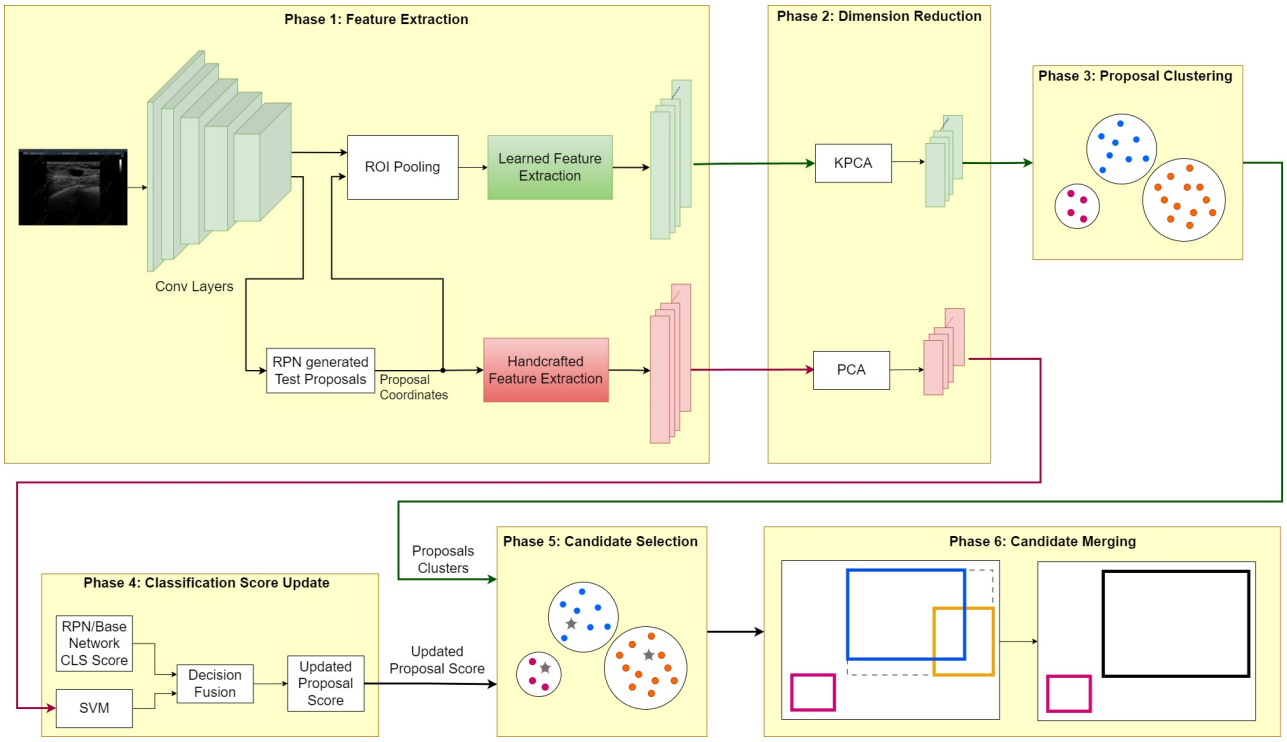


Figure 7.2: Overview of U-DetectH method.

of U-Detect models can be addressed. After candidates are selected in phase five using the updated score, they are passed through phase six where they are filtered through the NMS without causing a negative impact on the output. The filtered candidates are then processed through the candidate merging method presented in Section 6.1.5 of Chapter 6. Based on the location of the U-DetectH method, two networks are derived, namely, U-DetectH-RPN and U-DetectH-Base. Figures 7.3 and 7.4 show the proposed U-DetectH-RPN and U-DetectH-Base, respectively.

Given the similarity in phases of both U-Detect and U-DetectH, the remainder of this section describes investigation of the phases specific to U-DetectH, i.e., handcrafted feature extraction in phase one, dimension reduction of handcrafted features in phase two and classification decision fusion in phase four. First, Section 7.1.1 details the selected handcrafted features. Section 7.1.2 describes dimensionality reduction technique used for the handcrafted features. Finally, Section 7.1.4 describes the decision fusion method used.

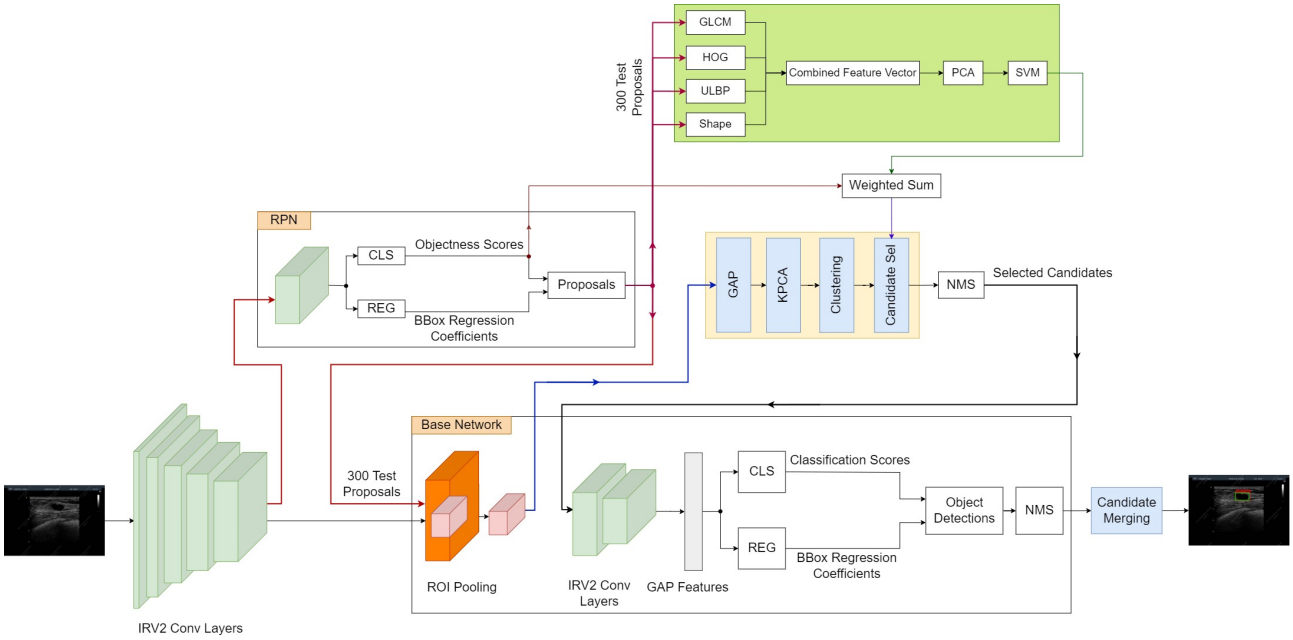


Figure 7.3: Overview of U-DetectH-RPN method.

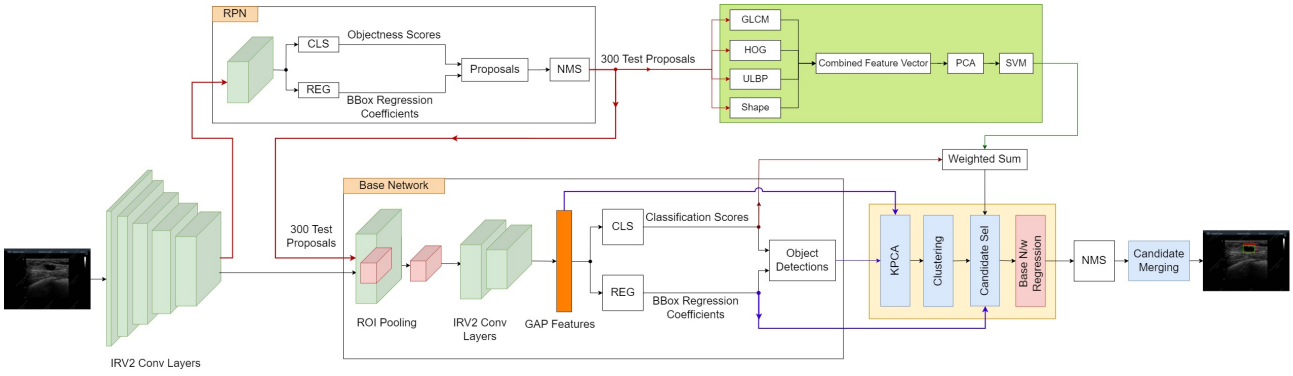
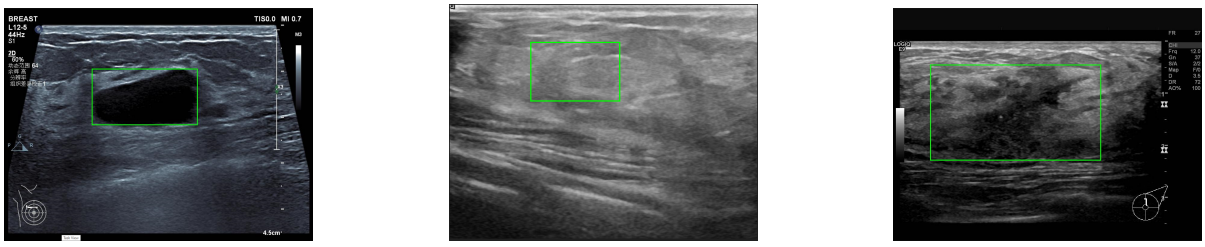


Figure 7.4: Overview of U-DetectH-Base method.

7.1.1 Extraction of Handcrafted Features

Inspired by the way the radiologists read and report the breast lesion characteristics such as echogenicity, margin and shape, a set of discriminating features are defined to improve the classification scores assigned to the proposals. As illustrated in Chapter 2 Section 2.1, echogenicity of a lesion is its property to reflect US waves in reference to its surrounding tissues. This is an important textural characteristic of the lesion used for its BI-RADS classification. Depending on the nature of the lesion, its echogenicity varies. For instance, fluid-filled lesions appear dark in an US image as fluids absorb the US waves instead of reflecting them. These are usually benign lesions/cysts as shown in Figure 7.5a. On the other hand, solid lesions have isoechoic texture as solid mass have high reflectivity. An example of an isoechoic lesion is shown in Figure 7.5b.



(a) Anechoic (b) Isoechoic (c) Hyperechoic

Figure 7.5: Example of lesions with different echogenicities.

Also, some fluid-filled lesions contain solid calcifications which appear as bright spots in an otherwise dark lesion as seen in figures 7.5b and 7.5c. These calcifications also provide important information useful for the classification of the lesion. Lesions may also have mixed echogenicity (non-uniform texture) depending on its composition. For instance, hyperechoic lesions (usually malignant) contain fat and therefore have bright grey regions in the lesion as shown in Figure 7.5c. Margin of a lesion contains important textural information pertaining to the type of lesion. Furthermore, proposals containing lesions generally have a distinctive aspect ratio as lesions are typically elliptical in shape. Such characteristics can be used to distinguish lesions from normal breast tissue.

We propose to extract HOG, GLCM, ULBP and aspect ratio (shape) features to represent each proposal where the aspect ratio feature captures the orientation of the lesion, GLCM captures global textural features relating to the lesions' echogenicity and HOG and ULBP capture local textural

features such as edges and contours caused by the change in contrast with varying echogenicity of the lesion and the background tissue as well as calcifications found inside the lesion. Then, these features are used with SVM to recognise the proposal as foreground (lesion) or background. Theoretical background of HOG, GLCM and ULBP is provided in Section 2.2.1 Chapter 2. Remainder of this section presents the extraction of the proposed handcrafted features.

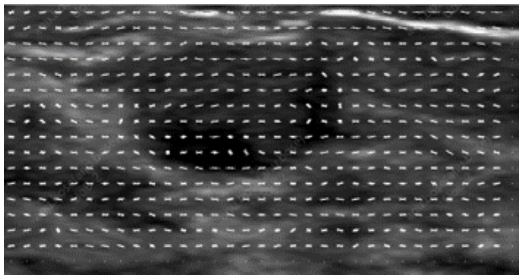
7.1.1.1 Gray-level Co-variance Matrix (GLCM)

GLCM is a second-order textural feature which provides global textural information. In particular, GLCM provides textural information in relation to the echogenicity of the region of the image contained in the proposal. For each proposal, 16 GLCM matrices are constructed for four distances (1, 2, 3, 4) and four angles (0, 45, 90, 135). These values have been commonly used in the literature for breast lesion classification as discussed in Chapter 3. From each of these 16 GLCM matrices, four textural metrics are computed, namely, contrast, correlation, entropy, and energy, which provide details with respect to echogenicity. These textural metrics are defined in Section 2.2.1 of Chapter2. Therefore, each proposal is described using 64 textural metrics computed for the 16 GLCM matrices. Thus, the dimension of the GLCM feature vector is 64×1.

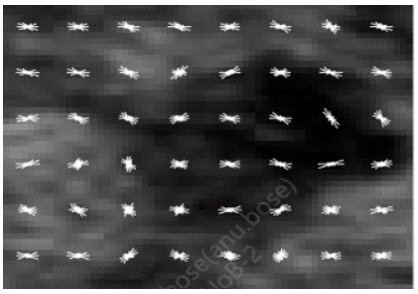
7.1.1.2 Histogram of Oriented Gradients (HOG)

HOG extracts local textural features in an image such as edges and contours. Such local features help in identification of the lesion. As proposals have varied sizes, using the traditional approach of extracting HOG would lead to an inconsistent feature size of across proposals generated for an image. For example, consider two proposals, $P1$ of size 139×264 and $P2$ of size 50×65. As shown in Figure 7.6, use of the traditional method of predefining the number of pixels in a cell (cell size) generates HOG feature vectors of different dimensions for each proposal. Additionally, an optimal cell size for $P1$ works poorly in $P2$ and vice versa. Also, predefined cell size designed for the average proposal size fails for proposals of smaller dimensions as shown in Figure 7.6f.

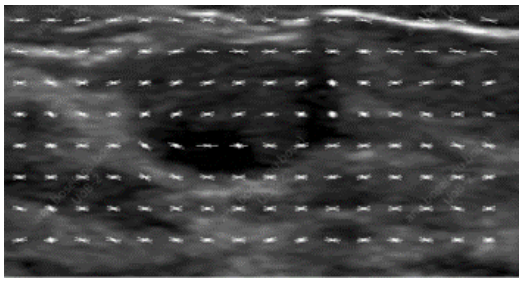
To overcome this issue and maintain a constant feature vector size, the total *number of cells* is predefined instead of cell size. Thus, every image is divided into the same number of cells with the only difference being in the number of pixels in each cell. Using this approach also adapts HOG for each individual proposal. We defined the number of cells to 4×4 and bin size to 9. With block



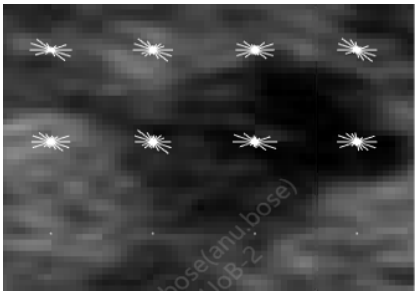
(a) Cell size 8×8. Feature size = 4608



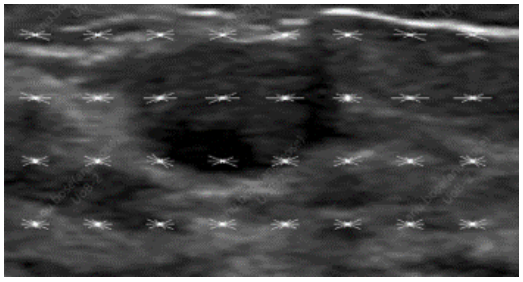
(b) Cell size 8×8. Feature size = 432



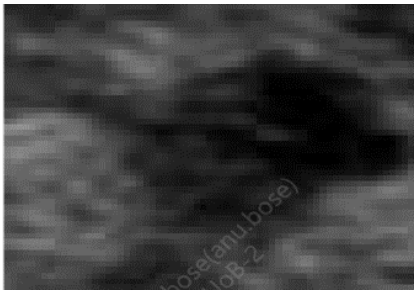
(c) Cell size 16×16. Feature size = 1152



(d) Cell size 16×16. Feature size = 72



(e) Cell size 32×32. Feature size = 288



(f) Cell size 32×32. Feature size = NA

Figure 7.6: Impact of varying cell size in two sample proposals $P1$ of size 139×264 (left column) and $P2$ of size 50×65 (right column).

size set to 2×2 cells and no overlap between blocks, the output feature vector is maintained at a constant size of 144×1. These parameters have been commonly used in HOG feature extraction, with the exception of the number of cells. Number of cells is determined so as to ensure a balance between quality of features and dimension of the feature vector; higher number of cells would improve the textural information captured in the feature vector but it greatly increases the dimension of the feature vector and vice versa.

7.1.1.3 Uniform Local Binary Pattern (ULBP)

Compared to LBP, ULBP is invariant to rotation as well as the dimension of the feature vector is considerably smaller. ULBP provides local textural information such as edges which aid in the classification of lesions. Here, the window size is adapted to the size of the input proposal. Thus, 59 ULBP features are extracted for the entire proposal. This feature vector is a histogram of binary patterns. Thus, each feature in the ULBP feature vector is a histogram bin for a binary pattern(s). Of the 59 bins, 58 bins are reserved for 58 uniform patterns and 1 bin for all non-uniform patterns. Thus, the feature vector has a constant size of 59×1, irrespective of the proposal size.

7.1.1.4 Aspect Ratio

Lesions have a characteristic elliptical shape which is captured by its unique aspect ratio or shape. Shape here refers to the ratio of width and height of the proposal. This single feature is not used on its own for classification but in combination with the above-mentioned textural features.

7.1.1.5 Feature Fusion

We produce a new feature vector that combines all textural and morphological features, namely, GLCM, HOG, ULBP, and shape. This results in a feature vector of size 268×1. This feature vector is referred to as the combined feature vector. We hypothesise that combining all textural and morphological features, capturing global and local textural features of the proposals along with morphological features, can improve the proposal selection and ultimately the lesion detection. Table 7.1 provides a summary of the extracted features.

Feature Extraction Method	Type of Feature	Feature Vector Size
GLCM	Global texture	64×1
ULBP	Global texture	59×1
HOG	Local texture	144×1
Shape	Morphological	1×1
Combined	Textural and Morphological	268×1

Table 7.1: Handcrafted features used in U-DetectH models.

7.1.2 Dimension Reduction of Handcrafted Feature Vector

Given the large size of the combined features in Section 7.1.1, the Principal Component Analysis (PCA) is used to reduce the dimensionality of the extracted feature vectors. In particular, PCA is applied to the 268×1 handcrafted feature vector extracted from each proposal. This dimensionally reduced feature vector is then used for classification of each proposal using SVM. The optimal number of principal components (PCs) were estimated using the same method presented in Section 6.1.2.1 in Chapter 6. An initial test was conducted to determine the number of PCs that capture 99% cumulative information (CI). The number of PCs capturing 99% CI ranged from 25 to 41. Based on this, the number of PCs to be evaluated was determined. The optimal number of PCs was estimated experimentally as presented in Section 7.2.2.

7.1.3 Proposals Classification

After extraction of the handcrafted features and their dimension reduction using PCA, the proposals are classified using an SVM model. This SVM model is trained independently using proposals generated for images from modelling dataset (dataset A) by the RPN of adapted FRCNN. Details of the selection of SVM training samples and its modelling hyperparameters is provided in Section 7.2.1. Thus, SVM outputs classification scores for each proposal in phase four. At this point, each proposal has two classification scores – one assigned by the RPN/base network and the second assigned by the SVM. After this, a classification decision fusion of the RPN/base network score and the SVM score is performed which outputs an updated score for each proposal.

7.1.4 Classification Decision Fusion

This section describes the method of decision fusion between the classification score assigned by the RPN/base network and the SVM model. In particular, weighted sum is used to generate an updated

proposal score as a fusion between the two classifiers as shown in the following Equation 7.1.4.

$$\begin{aligned} RPN: \quad & score_{new}(rpn) = w_{svm} \times score_{svm} + w_{rpn} \times score_{rpn} \\ Base\ Network: \quad & score_{new}(base) = w_{svm} \times score_{svm} + w_{base} \times score_{base} \end{aligned} \tag{7.1}$$

Here, $score_{new}(rpn)$ and $score_{new}(base)$ represent the updated score in U-DetectH-RPN and U-DetectH-Base, respectively. $score_{svm}$, $score_{rpn}$ and $score_{base}$ are the classification scores assigned by SVM, RPN and base network, respectively. w_{svm}, w_{rpn} and w_{base} represent weights assigned to SVM, RPN and base network, respectively. All weights lie in the range $[0, 1]$ and $w_{svm} + w_{rpn}(or\ w_{base}) = 1$. Optimal weights were identified experimentally as $w_{svm} = 0.1$ and $w_{rpn} = w_{base} = 0.9$. Experimental evaluation of various weights and identification of the aforementioned weights as optimal is discussed in Section 8.5 of Chapter 8.

7.2 Experimental Results and Analysis

This section presents the experimental results and analysis of the U-DetectH method. As this method uses an SVM classifier trained on the combined feature vector, first a description of the training set used to train this classifier is provided in Section 7.2.1. Both U-DetectH models (U-DetectH-RPN and U-DetectH-Base) use U-Detect models from Chapter 6 as the base. The pretrained detector here is the adapted FRCNN model using IRV2 as the backbone. All models are evaluated on dataset A as modelling dataset and datasets B to E as external test sets as described in Section 4.1 of Chapter 4. All results shown in this section are average of 5-folds unless otherwise specified. The performance of U-DetectH-Base and U-DetectH-RPN methods are presented in Sections 7.2.2 and 7.2.3, respectively.

7.2.1 SVM Training Dataset Generation

This section details the selection of samples used to train SVM models on combined feature vectors. As previously mentioned, all U-DetectH models use pretrained adapted FRCNN model with IRV2 as the base detector. The 5-folds splits of the modelling dataset (dataset A) used for training adapted FRCNN model are used to generate training and testing sets for the SVM model. In particular, training images from each fold of the modelling dataset are used to generate SVM training samples for that fold. Thus, all U-Detect and U-DetectH models are tested on the same splits of the modelling dataset. Likewise, as with U-Detect models, datasets B to E are also used as unseen external test

sets for the U-DetectH models. For training the SVM model, proposals generated by the RPN of the adapted FRCNN (IRV2 base) for the training images are considered. Table 7.2 shows the number of proposals in a range of IOU (with GT box) values, generated for one-fold of the training set. As the IOU increases, the total number of proposals decreases. This is due to the presence of only one, generally small to medium sized, lesion in an average image in the modelling dataset.

IOU	**Total Samples**	**Selected Training Samples**	**Sample Class**
0	164428	292	Negative (Background)
$(0, 0.1)$	98511	292	
$[0.1, 0.2)$	55761	292	
$[0.2, 0.3)$	30011	292	
$[0.3, 0.4)$	17157	292	
$[0.4, 0.5)$	9292	292	
$[0.5, 0.6)$	5264	292	
$[0.6, 0.7)$	2941	292	
$[0.7, 0.8)$	1498	1498	Positive (Lesion)
$[0.8, 0.9)$	750	750	
$[0.9, 1]$	86	86	

Table 7.2: SVM training sample selection.

The RPN and base network have high classification accuracy for very high IOU proposals (IOU $[0.7, 1]$) and very low IOU proposals (IOU $[0, 0.3)$). For example, over 99% of proposals with IOU 0, 97% of proposals with IOU $(0, 0.2)$ and around 87% with IOU $[0.2, 0.3)$ had classification scores in the range $[0, 0.1)$. However, their classification accuracy drops for proposals with IOU $[0.3, 0.6]$. For instance, Figures 7.7 and 7.8 show the range of base network's classification scores assigned to proposals with IOU $[0.4, 0.5)$ and IOU $[0.5, 0.6)$, respectively. 46.71% proposals with IOU $[0.4, 0.5)$ were correctly scored < 0.3. However, 34.14% of these low IOU FP proposals were incorrectly assigned very high scores in the range of $[0.9, 1]$.

Likewise, TP proposals with IOU in the range $[0.5, 0.6)$, base network incorrectly assigned 25.61% of these with score < 0.3. Similar classification score distribution was seen with the RPN. When such TP and FP proposals are clustered together, the FP proposal is incorrectly selected as the candidate from that cluster due to its higher classification score resulting in the issue cases of U-Detect models (single low IOU FP and FP+FN). Thus, the aim of the SVM model is to increase the classification score assigned to high IOU proposals as well as reduce that assigned to lower IOU, FP proposals.

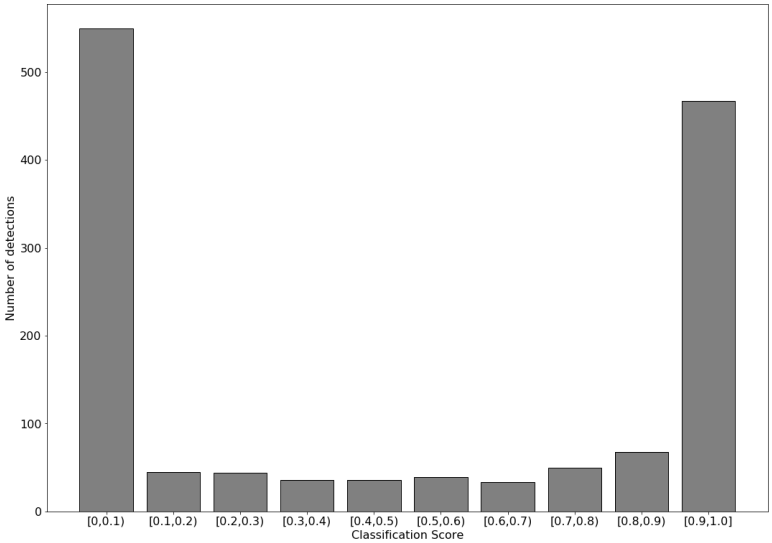


Figure 7.7: Classification scores assigned by the base network of adapted FRCNN model (IRV2 backbone) to proposals with IOU $[0.4, 0.5)$ with the GT box.

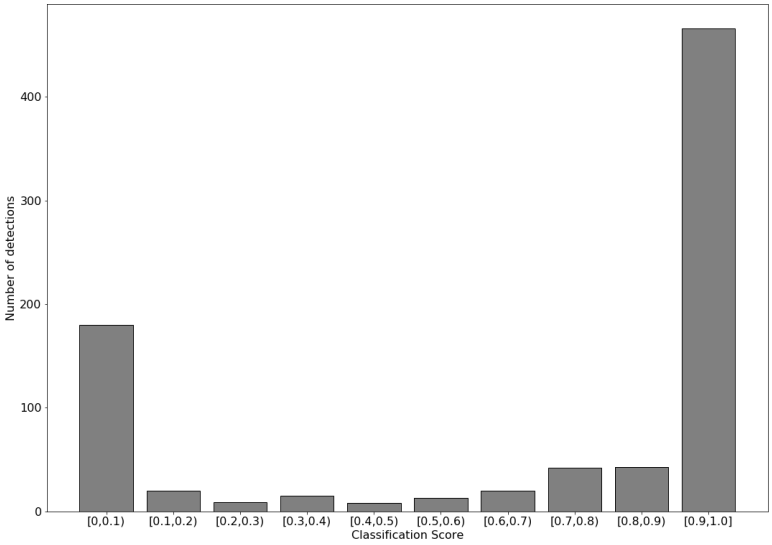


Figure 7.8: Classification scores assigned by the base network of adapted FRCNN model (IRV2 backbone) to proposals with IOU $[0.5, 0.6)$ with the GT box.

Based on this, all proposals with $IOU \geq 0.7$ are selected as positive (lesion class) training samples. In the example of the fold described in Table 7.2, total number of positive samples equals 2334.

To balance the number of positive and negative samples, prevent the model from becoming biased towards one class as well as ensure variety in the negative samples, 292 proposals from each IOU bin mentioned in Table 7.2 (up to IOU 0.7) were randomly selected as negative (background class) samples, which in this fold totals to 2336 negative samples. All SVM models use gaussian RBF kernels. Input dataset to all SVM models is standardised to its weighted mean and standard deviation. To test the classification accuracy of the trained SVM classifier, proposals generated by the RPN of the adapted FRCNN for all images in the test set of the modelling dataset as well as external test sets (B to E) are used. Classification accuracy of the SVM models (average of 5 folds) is presented in the Table 7.3. Additionally, three other training sets were evaluated. Their performance is discussed in depth in Section 8.3 Chapter 8.

Dataset	**Accuracy**	**Precision**	**Recall**	**F-measure**	**Specificity**
A	95.38	7.01	84.85	12.95	95.42
B (benign)	96.10	5.45	91.43	10.29	96.11
B (malignant)	94.09	4.96	74.14	9.30	94.18
C	95.28	6.19	73.37	11.41	95.38
D	96.32	8.33	69.72	14.89	96.44
E	86.79	1.87	78.39	3.66	86.82

Table 7.3: Classification accuracy of SVM model trained on combined feature vector.

7.2.2 U-DetectH-Base

This section presents the performance of U-DetectH-Base in three main parts. In part 1, performance of the U-DetectH-Base with SVM model trained on the entire combined feature vector (without dimensionality reduction) is discussed. Part 2 presents the impact of dimensionality reduction of the combined feature vector using PCA on the overall performance. Here, the SVM model is trained on the same dataset described in Section 7.2.1 but using the dimensionally reduced feature vector. To understand the influence of individual features, performance of U-DetectH-Base using SVM models trained on these features was studied and presented in the part 3 of this section. In particular, three additional SVM models were trained using HOG, GLCM and ULBP feature vectors individually, using the same training set described in Section 7.2.1. Classification decision fusion of these SVM models

and the base network in the U-DetectH-Base model is performed using weighted sum described in Section 7.1.4.

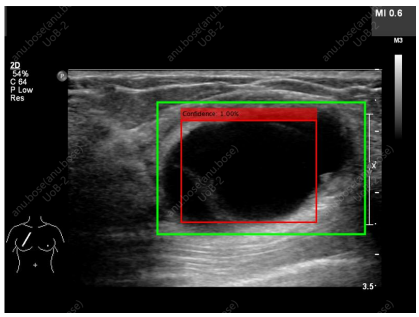
PART 1: U-DetectH-Base with Single SVM Model

This section details the performance of the U-DetectH-Base model using the SVM model trained on combined feature vector. For brevity, this U-DetectH-Base model is referred to as combined-SVM model. Table 7.4 shows the performance of combined-SVM model. Although the combined-SVM model has slightly lower performance in the modelling dataset in comparison to the U-Detect-Base, in the unseen external datasets, it has 0.23% higher precision as well as 0.12% higher recall (see Table 6.8 for performance of U-Detect-Base). This higher performance of the combined SVM-model is due to higher number of correct detections and lower number of FPs, overall. Compared to U-Detect-Base model, combined-SVM has 0.14% to 1.24% higher number of TPs and 1.45% to 16.67% lower number of FNs as well as 1.69% to 9.17% lower number of FPs. The exception to this improvement is in modelling dataset where the combined-SVM model has lower number of TPs due to 0.55% lower TPs and 2.96% higher FPs than the U-Detect-Base model (no change in the number of FNs).

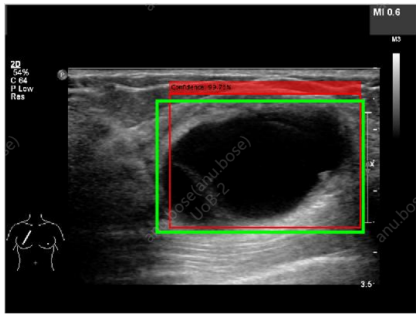
Dataset	**Combined-SVM**		
	Precision	**Recall**	**F-measure**
A	83.95	97.78	90.91
Overall External Test Sets	77.36	89.81	83.11
B	92.26	99.28	95.63
C	68.43	80.17	73.76
D	77.98	92.58	84.59
E	88.17	99.20	93.33

Table 7.4: Performance of U-DetectH-Base using combined-SVM model.

Figure 7.9 shows an example of reduction of single low FP, one of the common issue cases of U-Detect-Base, by the combined-SVM model. Apart from reduction of FPs and FNs, there was also a reduction in FP+FN cases (one of the common issue cases of U-Detect-Base) where the model outputs a single detection that covers the background region while completely missing the lesion. Such cases were reduced in the combined-SVM model due to improvement of scores assigned to proposals leading to a subsequent selection of the high-IOU proposal from the cluster instead of IOU 0 FP proposal. An example of this is shown in Figure 7.10.

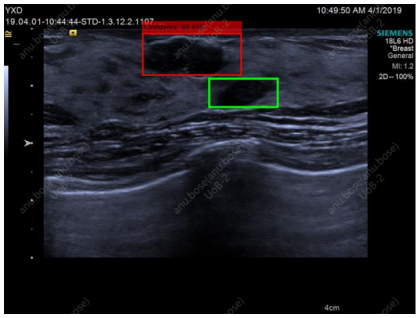


(a) U-Detect-Base

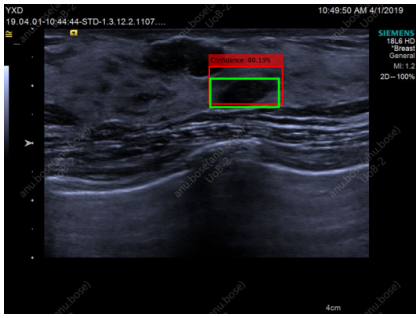


(b) Combined-SVM

Figure 7.9: Reduction of single low IOU FPs using U-DetectH-Base with combined-SVM model.



(a) U-Detect-Base

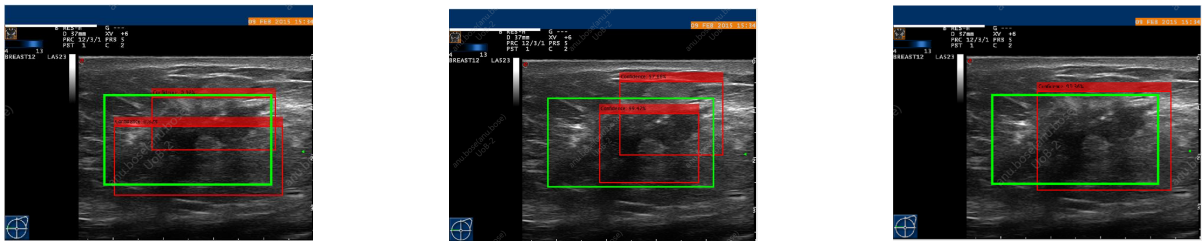


(b) Combined-SVM

Figure 7.10: Reduction of FP + FN in U-DetectH-Base with combined-SVM model.

The combined-SVM model also outperforms the adapted FRCNN and, consequently, original FRCNN (see Tables 6.1 and 5.12 for performance of adapted FRCNN and original FRCNN models, respectively). This is due to its considerably lower number of FPs and relatively small drop in FNs. Figure 7.11 shows an example of reduction of FP detection, found in both original and adapted FRCNN model, by the combined-SVM model. Specifically, compared to adapted FRCNN, combined-SVM model has 6.08% to 38.89% lower FPs leading to 0.77% to 8.28% higher precision. The drop in recall was relatively lower (0.77% to 0.62%) resulting in 0.36% to 4.79% higher F-measure than the adapted FRCNN. Likewise, compared to the original FRCNN, combined-SVM model has 31.86% to 77.07% lower FPs. Therefore, the combined-SVM model has 5.49% to 32.83% higher precision. With only a drop of 0.27% to 10.02% in recall, combined-SVM model has 4.59% to 22.74% higher F-measure than the original FRCNN model. The original FRCNN model maintains lower FNs than adapted FRCNN models, U-Detect and combined-SVM models. An example of such a case is shown in Figure 7.12a where the lesion is detected by the original FRCNN model but missed by adapted

and combined-SVM model. But some challenging lesions continue to be missed by all models. Figure 7.12b shows an example of such a challenging lesion. Single-fold performance of the combined-SVM model is presented in Section C.2 in Appendix C.



(a) Original FRCNN (b) Adapted FRCNN (c) Combined-SVM

Figure 7.11: FP reduction by U-DetectH-Base model with combined-SVM model in comparison to original and adapted FRCNN models.

(a) Original FRCNN (b) All models

Figure 7.12: FN examples. 7.12a: Lesion detected by original FRCNN but missed by other models including adapted FRCNN, U-Detect and U-DetectH-Base with combined-SVM. 7.12b: Missed by all models.

In summary, combined-SVM model addresses the common issue cases (single low IOU FPs, FP+FN, and FNs) of U-Detect-Base model by improving classification scores assigned to proposals by the base network. However, computation time of this model is relatively high in comparison to that of the U-Detect model. For a single image, combined-SVM requires an average of 10.6 seconds to process whereas the U-Detect-Base model requires 0.96 seconds. This increase in computation time is due to the additional time required for extraction of all features (HOG, GLCM, ULBP and shape). Additionally, we also investigated the impact of using a feature selection method of dimensionality reduction on the overall performance. In particular, we investigate the Maximum-Relevance and Minimum-Redundancy (MRMR) [157] method for the dimension reduction of the combined-SVM.

The performance of this model is discussed in Section 8.4 of Chapter 8.

PART 2: Dimension Reduction of Handcrafted Features in U-DetectH-Base

Part 1 demonstrated the performance of U-DetectH-Base with no dimension reduction of handcrafted features. This section details the impact of dimension reduction on the overall performance. First, the optimal number of PCs is determined. Here, 25, 30, 35, 40 PCs were investigated based on the range of components with 99% CI as described in Section 7.1.2. Optimal number of PCs is determined based on the performance of these PCs on a single fold of modelling dataset (same fold used in Sections 6.2.1.2 and 6.2.2.2 in Chapter 6). Figure 7.13 shows the impact of these PCs on the overall performance. As the number of PCs increases, the overall performance improves. Here, increasing the number of PCs led to retention of useful information that improves the overall classification performance of the SVM models and, in turn, the performance of the U-DetectH-Base models. Based on this performance, 40 PCs are selected as optimal. Classification performance of the SVM model trained on 40 PCs of combined feature vector is detailed in Table C.4 Appendix C.

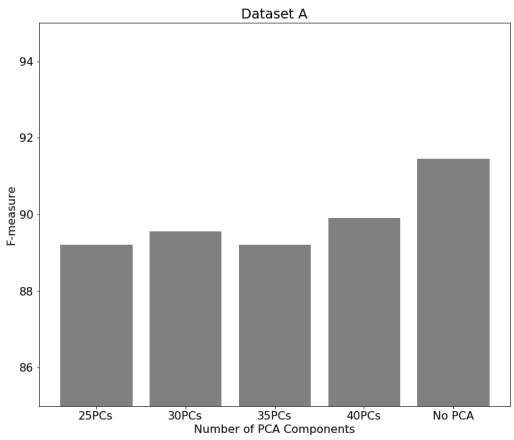


Figure 7.13: Impact of change in PCs on performance of U-DetectH-Base model using combined-SVM on a single fold of dataset A.

Table 7.5 compares the performance of U-DetectH-Base model using SVM model trained on the whole combined feature vector (combined-SVM) with that of the U-DetectH-Base model using SVM model trained on dimensionally reduced feature vector (combined-PCA-SVM). The combined-PCA-SVM model performs poorly in comparison to the combined-SVM model in all datasets. The combined-PCA-SVM model has 2.6% to 13.02% lower precision along with 0.03% to 2.53% lower recall resulting

in 1.54% to 8.01% lower F-measure than the combined-SVM model.

Dataset	U-DetectH-base Model	Precision	Recall	F-measure
A	Combined-SVM	**84.97**	98.99	**91.45**
	Combined-PCA-SVM	82.37	**98.96**	89.91
Overall External Test Sets	Combined-SVM	**77.60**	**89.80**	**83.25**
	Combined-PCA-SVM	68.21	88.91	77.20
B	Combined-SVM	**90.67**	**99.27**	**94.77**
	Combined-PCA-SVM	78.67	99.16	87.73
C	Combined-SVM	**67.23**	**78.02**	**72.23**
	Combined-PCA-SVM	57.54	75.49	65.30
D	Combined-SVM	**80.00**	94.02	**86.44**
	Combined-PCA-SVM	72.73	**94.12**	82.05
E	Combined-SVM	**86.98**	**100.00**	**93.04**
	Combined-PCA-SVM	73.96	**100.00**	85.03

Table 7.5: Impact of PCA on the performance of combined-SVM model in U-DetectH-Base: Single Fold.

This drop in the performance of the model with the use of PCA is due to reduction in TPs and increase in FPs. In general, FNs had small to no change. Use of PCA feature vectors led to poorer classification of background regions causing an increase in additional boxes (FPs). Furthermore, due to the reduced textural information in these feature vectors, lesions are detected with lower IOU leading to an increase in low IOU FPs along with reduction in TPs. However, the dimensionally reduced feature vectors did not cause a significant, if any, change to the number of detected lesions (FNs) showing that although these features are capable of detecting lesions, they contain insufficient textural information for correct classification of background regions with lesion-like texture and detection of lesions with high IOU. As the combined-PCA-SVM model has poorer performance, it is not used for further investigations.

PART 3: U-DetectH-Base with Single Feature-Based SVM Models

This section details the performance of U-DetectH-Base with single feature-based SVM models, namely, GLCM-, HOG- and ULBP-SVM. These SVM models were trained using the same dataset used for combined-SVM model training (described in Section 7.2.1). The classification performance of GLCM-SVM, HOG-SVM and ULBP-SVM is detailed in Section C.1 of Appendix C in Tables C.1, C.2 and C.3, respectively.Table 7.6 shows the performance of these models in U-DetectH-Base and

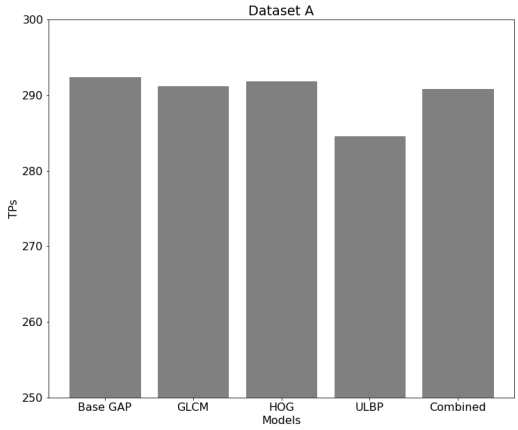
Figure 7.14 shows the number of TP, FP, and FN of these models in comparison to U-Detect-Base and combined-SVM. Of these models, GLCM-SVM is the best performing, followed by HOG-SVM and finally ULBP-SVM. However, these models do not outperform U-Detect-Base and combined-SVM.

Dataset	**U-DetectH-Base Model**	**Precision**	**Recall**	**F-measure**
A	GLCM-SVM	84.02	99.12	90.91
	ULBP-SVM	82.16	99.17	89.83
	HOG-SVM	**84.18**	**99.18**	**91.05**
Overall External Test Sets	GLCM-SVM	**74.64**	**87.79**	**80.68**
	ULBP-SVM	71.06	87.57	78.45
	HOG-SVM	73.83	87.73	80.18
B	GLCM-SVM	**91.42**	**98.70**	**94.92**
	ULBP-SVM	87.13	98.63	92.52
	HOG-SVM	89.55	98.67	93.88
C	GLCM-SVM	61.12	79.87	69.22
	ULBP-SVM	58.29	80.46	67.56
	HOG-SVM	**62.25**	**80.72**	**70.25**
D	GLCM-SVM	**78.74**	**87.06**	**82.67**
	ULBP-SVM	75.80	86.59	80.82
	HOG-SVM	76.80	86.80	81.47
E	GLCM-SVM	84.45	**99.45**	91.32
	ULBP-SVM	81.59	99.42	89.62
	HOG-SVM	**85.99**	**99.45**	**92.22**

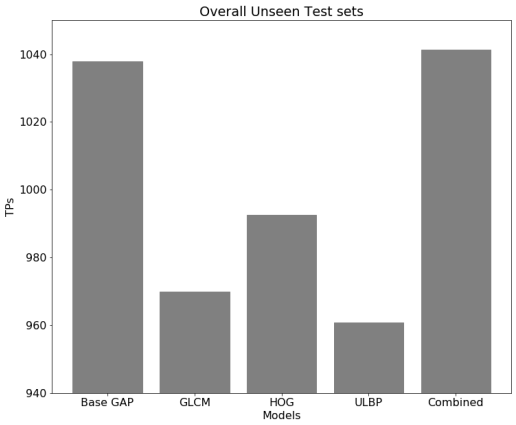
Table 7.6: Performance of U-DetectH-Base using single-feature based SVM models.

TP Of the three models, HOG-SVM has the highest number of TPs. Compared to global features extracted by GLCM, HOG extracts local textural features which provides greater textural information that aids correct identification of high IOU proposals. These proposals typically contain challenging small lesions that require such intricate, local textural details contained in HOG feature vectors for their correct identification. An important point to note here is that GLCM-SVM has a comparable number of TPs as HOG-SVM and it does not *miss* these small lesions but detects them with lower IOU ($IOU < 0.5$). An example of this is shown in Figure 7.15. Although ULBP also captures local textural features, ULBP-SVM completely misses these lesions as the degree of information captured is insufficient.

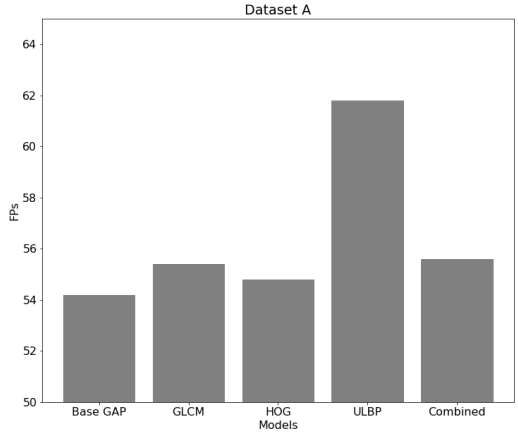
FP GLCM-SVM has the lowest number of, both, additional boxes and low IOU FPs (despite conversion of TPs to low IOU FPs) in comparison to HOG- and ULBP-SVM. This was closely followed by HOG-SVM. ULBP-SVM has the highest number of FPs. High number of FPs in



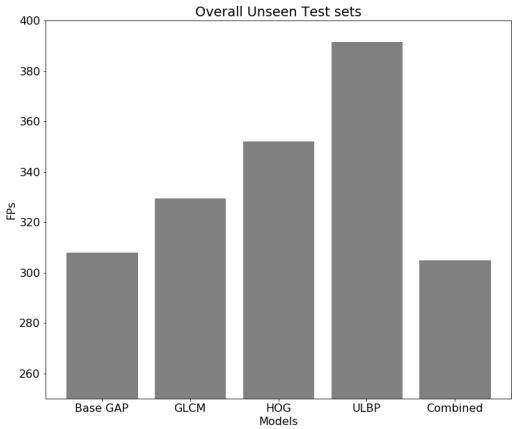
(a) TPs: Dataset A



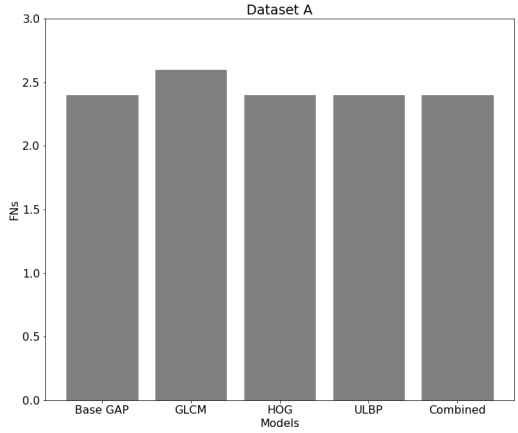
(b) TPs: External Test Sets



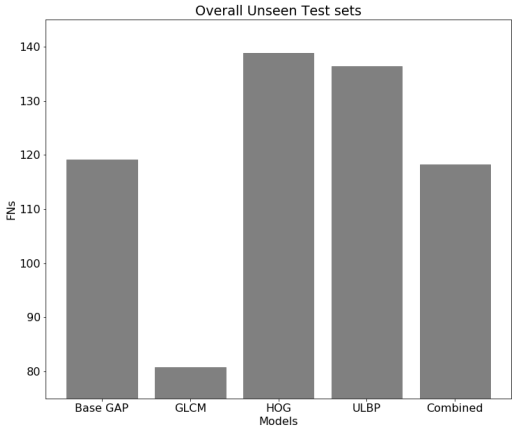
(c) FPs: Dataset A



(d) FPs: External Test Sets

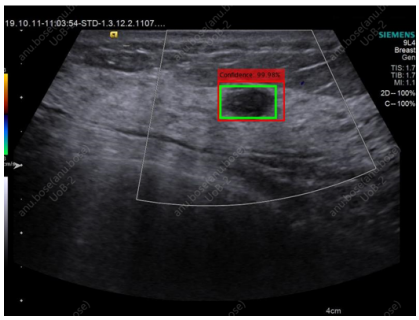


(e) FNs: Dataset A

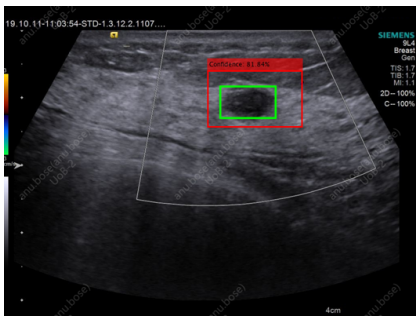


(f) FNs: External Test Sets

Figure 7.14: Number of TP, FP and FN of U-DetectH-Base using single feature-based SVM models and U-Detect-Base (base-GAP) model.



(a) U-Detect-Base model



(b) GLCM-SVM model

Figure 7.15: Low IOU FP in U-DetectH-Base with GLCM-SVM model.

HOG-SVM is due to the textural similarities between proposals containing lesions and those containing lesion-like background regions. Thus, use of local textural features alone improves TP detections at the expense of FPs (except ULBP where the number of the opposite is true due to insufficient textural information contained in the feature vector).

FN GLCM-SVM also has the lowest number of FNs compared to HOG- and ULBP-SVM. HOG-SVM has the same number of FNs as GLCM-SVM except in dataset C where the number of FNs is higher. This dataset contains lesions that have high textural similarity with the background region. As HOG-SVM relies on extraction of distinct local textures such as edges and contours for correct classification, it misses such lesions, whereas GLCM-SVM correctly classifies these cases since it only uses the global texture of the entire proposal. ULBP-SVM has the highest number of FNs.

In summary, local textural features extracted using HOG helps identify high IOU proposals but falls short when the lesion has a very challenging texture (isoechoic) and unclear boundary. On the other hand, global features extracted using GLCM have a higher number of detected lesions (including isoechoic lesions) but with small, challenging lesions, the lower IOU proposal is selected. Finally, ULBP features of the entire proposal does not contain sufficient local textural information for reliable classification performance.

Given the lowest number of FPs and high number of correct detections, the GLCM-SVM model has the highest precision and F-measure among the three single feature SVM models. HOG-SVM has the highest recall due to the highest number of correct detections (highest TPs and second-highest FNs). However, in comparison to U-Detect-Base, all three single-feature based SVM models have

lower performance due to relatively lower number of correct detections as well as higher FPs. This is because these SVM models have lower classification accuracy than the base network. Therefore, a decision fusion of their scores results in a poorer score assigned to the proposals. This results in incorrect selection of candidates which negatively impacts the overall performance of the model. It is important to note here that GLCM-SVM model made a positive impact on the scores of lesions that were otherwise missed by the U-Detect-Base, leading to a lower number of FNs than the U-Detect-Base model.

Combined-SVM model (see Table 7.4) outperforms all models including single feature-based models due to higher number of TPs and lower number of FPs (additional boxes and low IOU FPs). Use of HOG along with ULBP in the combined feature vector helps identify high IOU proposals while GLCM helps detect challenging lesions as well as successfully differentiate between proposals covering lesion and lesion-like regions. The combined-SVM model has lower FNs than HOG-SVM and ULBP-SVM. However, compared to GLCM-SVM, it has slightly higher FNs. Use of GLCM in the combined feature vector helps reduce FNs compared to the U-Detect-Base model (see Table 6.8 for performance of U-Detect-Base model) but due to the negative impact of HOG and ULBP, it has higher FNs than GLCM. Also, like combined-SVM, GLCM-SVM also reduces FP+FN cases. Therefore, a combination of all textural and morphological features incorporates their respective advantageous characteristics with relatively smaller degree of their negative counterparts.

All single-feature based U-DetectH models have higher computation time in comparison to the U-Detect-Base model due to the added time required for extraction of their respective handcrafted features of each test proposal.The U-Detect-Base model has an average computation time of 0.96 seconds. Of the three single features, ULBP extraction requires the least amount of time making the ULBP-SVM model fastest among the SVM models with an average computation time of 2.33 seconds. This is closely followed by HOG-SVM whose average computation time is 3.17 seconds. Extraction of GLCM of each proposal requires the highest computation time resulting in the high computation time of 8.42 seconds by the GLCM-SVM model. As all features are extracted in the combined-SVM, its computation time is highest (10.6 seconds). This computation cost can be significantly reduced through improving the functions used for feature extraction step, especially for extraction of GLCM features since it is the largest contributing factor in the high computation time of combined-SVM.

7.2.3 U-DetectH-RPN

This section details the performance of U-DetectH-RPN. The U-DetectH-RPN model builds on the U-Detect-RPN model from Chapter 6 and uses the same dataset split (dataset A as modelling dataset and datasets B to E as unseen test sets) used in Chapter 6 and for U-DetectH-Base models. Here, dataset A is used as the modelling dataset and datasets B to E are used as external test sets. Table 7.7 presents the performance of all U-DetectH-RPN models. Like with U-DetectH-Base, all U-DetectH-RPN models are referred to by the SVM model used. Here, of all U-DetectH-RPN models, the combined-SVM model has the best performance overall. However, the combined-SVM model has lower performance in comparison to U-Detect-RPN.

The SVM models perform in a similar manner as observed in U-DetectH-Base described in Section 7.2.2. Considering single feature-based models only (GLCM, HOG and ULBP), HOG-SVM has the highest number of TPs overall, closely followed by GLCM-SVM. GLCM-SVM has the lowest number of FPs and FNs. ULBP-SVM has the lowest performance overall. Finally, the combination of both textural and morphological features in the combined-SVM model outperformed all single feature-based models. Combined-SVM has highest TPs and lowest FPs. FNs of the combined-SVM model is comparable to that of the GLCM-SVM model. Thus, the combined-SVM model has the highest precision and recall. Although the combined-SVM model outperformed U-Detect-RPN by 2.33% in F-measure over modelling dataset, it has 3.77% lower F-measure in external test sets except dataset D. The largest drop in performance by the combined-SVM model is seen in dataset C where combined-SVM had 7.72% lower F-measure than the U-Detect-RPN model. In dataset D, both GLCM and combined-SVM models outperform U-Detect-RPN by 0.16% and 0.38% in F-measure, respectively.

Unlike U-DetectH-Base, U-DetectH-RPN does not outperform its U-Detect counterpart (U-Detect-RPN). This is due to the difference in classification accuracy of the RPN and the base network. The SVM models have poorer classification accuracy than the RPN and base network. Therefore, due to the higher classification accuracy of the base network, the negative impact on the number of correct detections with the use of SVM models was limited. However, due to the comparatively poorer classification accuracy of the RPN, the lower classification accuracy of the SVM models led to a reduction in the number of detected lesions. Since it was found in Section 7.2.2 that reducing dimension of feature vectors leads to poorer performance, investigation of PCA based models for SVM models in

Dataset	Model	Precision	Recall	F-measure
A	U-Detect-RPN	84.08	**99.26**	91.02
	U-DetectH-RPN with:			
	GLCM-SVM	88.25	98.72	93.19
	ULBP-SVM	87.97	98.71	93.03
	HOG-SVM	**88.29**	99.04	**93.35**
	Combined-SVM	**88.29**	99.04	**93.35**
Overall External Test Sets	U-Detect-RPN	**74.85**	**90.37**	**81.84**
	U-DetectH-RPN with:			
	GLCM-SVM	67.90	90.12	77.45
	ULBP-SVM	67.80	90.20	77.42
	HOG-SVM	68.12	90.24	77.63
	Combined-SVM	68.75	90.31	78.07
B	U-Detect-RPN	**92.51**	**99.43**	**95.83**
	U-DetectH-RPN with:			
	GLCM-SVM	86.62	99.27	92.52
	ULBP-SVM	87.18	99.27	92.83
	HOG-SVM	87.74	99.27	93.15
	Combined-SVM	87.18	99.27	92.83
C	U-Detect-RPN	**65.84**	**80.59**	**72.24**
	U-DetectH-RPN with:			
	GLCM-SVM	53.94	77.37	63.57
	ULBP-SVM	53.45	77.37	63.23
	HOG-SVM	54.18	77.60	63.81
	Combined-SVM	55.15	77.72	64.52
D	U-Detect-RPN	**74.73**	93.14	82.81
	U-DetectH-RPN with:			
	GLCM-SVM	73.38	94.84	82.74
	ULBP-SVM	73.62	95.04	82.97
	HOG-SVM	73.33	95.03	82.78
	Combined-SVM	73.96	**95.06**	**83.19**
E	U-Detect-RPN	**86.64**	99.21	**92.38**
	U-DetectH-RPN with:			
	GLCM-SVM	75.13	**100.00**	85.80
	ULBP-SVM	75.13	**100.00**	85.80
	HOG-SVM	76.06	**100.00**	86.41
	Combined-SVM	76.04	**100.00**	86.39

Table 7.7: Performance of U-DetectH-RPN models.

U-DetectH-RPN was not conducted.

U-Detect-RPN does not outperform U-DetectH-Base with combined-SVM model (see Tables 6.4 and 7.4 for 5-fold average performance of U-Detect-RPN and combined-SVM, respectively). In par-

ticular, the U-DetectH-Base model has 0.89% and 2.75% higher precision than the U-Detect-RPN in modelling and overall external test sets, respectively. This was due to the considerably lower number of FPs. Specifically, the U-DetectH-Base model has 6.81% to 14.68% lower FPs in modelling and unseen external test sets, respectively. Also, the U-DetectH-Base model has 0.27% to 0.57% lower recall than the U-Detect-RPN model. Despite the lower recall, the U-DetectH-Base model has 0.43% to 1.41% higher F-measure than U-Detect-RPN due to its higher precision. The largest change was seen in dataset D where the U-DetectH-Base model had 5.27% higher precision due to 25.68% lower FPs along with 0.88% higher recall due to higher number of TPs.

7.3 Discussion

This chapter presented and evaluated our methods U-DetectH-Base and U-DetectH-RPN for accurate breast lesion detection in US images. We discuss three main investigations in this section. First, performance of multi-model classification decision fusion using all SVM models (GLCM-, HOG-, ULBP- and combined-SVM) and base network classifier is discussed. As shown in Sections 7.2.2 and 7.2.3, each feature vector has its unique as well as common advantageous characteristics. Therefore, to further strengthen these characteristics, decision fusion of all five classifiers was investigated. Second, the use of statistical moments for dimension reduction was investigated and discussed in this section. In particular, four central moments of the ULBP feature vector were extracted and used for proposal description. As central moments extract important characteristics of a feature vector, this investigation was conducted to study its effectiveness in dimension reduction. Third, the performance of a fusion of HOG and GLCM features on the overall performance was investigated. Individually, HOG-SVM and GLCM-SVM had a reliably high performance. However, these did not outperform U-DetectH using combined-SVM or U-Detect models. Therefore, this fusion of HOG and GLCM features was studied. The evaluation of the impact of such a fusion on the overall performance is discussed in this section.

Multi-model Classification: To strengthen advantageous characteristics of each feature vector, multi-model classification decision fusion using all four SVM models (HOG-, GLCM-, ULBP- and combined-SVM) and the base network of the U-DetectH-Base model is performed. In this model, all handcrafted features are extracted for each proposal along with the extraction of learned features in phase one. After dimension reduction of learned features in phase two, the proposals are clustered

using the dimensionally reduced learned features in phase three. In phase four, all proposals are classified using the extracted handcrafted features and their respective SVM models. At this point, each proposal is assigned with five classification scores (four scores from the SVM models and one score by the base network). Through decision fusion of these five scores, the classification score of each proposal is updated in this phase. The updated score is then used for candidate selection in phase five and selected candidates are processed through NMS and candidate merging in phase six. Two decision fusion methods were evaluated, namely, weighted sum, and majority voting.

For weighted sum (described in Section 7.1.4), following weights are assigned to each of the five classifiers based on their classification accuracy: $w_{base} = 0.4, w_{combined} = 0.3, w_{glcm} = 0.2, w_{hog} = 0.1, w_{ulbp} = 0.1$. For majority voting, if a sample is labelled as background by all SVMs (i.e., scored < 0.5) then its score is updated with the sum of scores of all SVMs, else the base network's classification score is maintained as the new score. Therefore, the aim of using this decision fusion method is to utilise the SVM models to filter out the FP proposals without impacting the TP proposals generated by the model.

Table 7.8 shows the performance of two decision fusion methods. For concision, these U-DetectH-Base models are referred to by means of the decision fusion mechanism used. For example, if the model uses weighted sum as the decision fusion mechanism, then it is referred to as weighted-sum based model or simply, weighted-sum. Overall, of the two decision-fusion methods, weighted-sum has the higher performance. However, these methods do not outperform single-SVM models (see Tables 7.6 and 7.4 for performance of single-SVM models) and U-Detect-Base model (see Table 6.8 for U-Detect-Base performance). Weighted-sum based model has a higher number of correct detections than the majority-voting model. In particular, the weighted-sum model has 203 and 746.6 higher number of TPs than the majority-voting model in modelling and external test sets, respectively. In terms of FNs, the weighted-sum model has 243.8 and 1002.4 lesser FNs in modelling and unseen external test sets, respectively. This led to 61.13% to 91.74% higher recall of the weighted-sum model in comparison to the majority-voting model.

On the other hand, the majority-voting based model has 42.8 and 304 lesser FPs than the weighted-sum model in modelling and unseen external test sets, respectively. This resulted in 0.23% to 11.28%

Dataset	U-DetectH-Base Model	Precision	Recall	F-measure
A	Weighted Sum	83.23	**99.24**	**90.52**
	Majority Voting	**83.46**	26.00	38.34
Overall External Test Sets	Weighted Sum	73.26	**87.41**	**79.70**
	Majority Voting	**79.61**	16.28	26.73
B	Weighted Sum	89.53	**98.52**	**93.81**
	Majority Voting	**93.75**	29.55	44.75
C	Weighted Sum	62.04	**81.53**	**70.45**
	Majority Voting	**73.32**	20.39	31.42
D	Weighted Sum	76.09	**85.49**	**80.51**
	Majority Voting	**79.47**	11.53	19.87
E	Weighted Sum	84.78	**99.16**	**91.40**
	Majority Voting	**94.79**	7.42	13.63

Table 7.8: Performance of multi-model classification decision fusion in U-DetectH-Base.

higher precision of the majority-voting model. However, due to considerably lower recall, the majority-voting based model has 39.03% to 77.77% lower F-measure than the weighted-sum based model. In comparison to U-Detect-Base and U-DetectH-Base with GLCM-, HOG- and combined-SVM models, both these models have lower performance. Weighted-sum based model outperforms U-DetectH-Base model using ULBP-SVM with 0.69% to 2.89% higher F-measure.

All SVM models have their individual useful characteristics such as high number of correct detections by the GLCM-SVM model as well as disadvantageous characteristics such as high missed lesions of the ULBP-SVM model. Therefore, when used in the weighted-sum model, their useful unique characteristics are subdued due to the influence of other SVM models, and their common disadvantageous characteristics are highlighted. Also, the base network has overall higher classification accuracy than the SVM models. In this weighted-sum model, the influence of the base network is reduced and that of the SVM models is increased. Therefore, the weighted-sum model has relatively higher number of missed lesions as well as higher number of FPs than the U-Detect-Base. Since SVM models are trained to classify proposals with $IOU < 0.7$ as negative (background), TP proposals with IOU $[0.5, 07)$ are assigned low scores by most of these models. Therefore, the score of such proposals is reduced in the majority-voting based model resulting in incorrect candidate selection which leads to a high number of missed lesions despite their correct classification by the base network. On the other hand, this same characteristic of the SVM models helps reduce FPs. However, the number of missed lesions is too high for this model to be viable.

Computation cost of decision fusion models is higher than that of the U-DetectH-Base model using any SVM models including combined-SVM. For combined-SVM, computation time is 10.6 seconds whereas for decision fusion models it is 10.72 seconds. Like with the combined-SVM model, the high computation time is largely attributed to the time required for extraction of all handcrafted features for 300 test proposals. Slight increase in the time with decision fusion models is due to the additional time used for classification of all proposals using all four SVM models. As with combined-SVM model, improvement in the feature extraction function will improve the overall computation time of all these models.

ULBP-Moment Feature Vector: The ULBP feature vector is a histogram of uniform patterns of dimension 59×1. To reduce the dimension of this feature vector while retaining only key textural information, a new feature vector called ULBP moment was investigated. This feature vector consists of the four central moments of the ULBP histogram. Central moments are defined using Equation 7.2.

$$m_k = \frac{1}{N} \sum_{i=1}^{N} (\pi_i - \bar{\pi})^k \tag{7.2}$$

Here, k is the order of the central moment, N is length of the feature vector which in this case is 59×1 (size of ULBP histogram) and π_i is the i^{th} element in the feature vector. First central moment π_1 is the mean of the feature vector. Since here the moments are all central, $\pi_1 = 0$ for all feature vectors. Therefore, π_1 is not used and only $k \in [2, 5]$ are used as the feature vector.

An SVM model was trained using this ULBP moment feature vector using the same training set as all other SVM models as described in Section 7.2.1 of this chapter. Classification performance of this SVM model is detailed in Table C.5 in Section C.1 Appendix C. To study the impact of this SVM, it was employed in U-DetectH-Base. For brevity, this model is referred to as ULBP-M-SVM. Decision fusion was achieved through weighted-sum where the ULBP-M-SVM model was assigned weight of 0.1 and the base network 0.9. Table 7.9 shows the performance of the ULBP-M-SVM model. Overall, this model had lower F-measure than the ULBP-SVM in all datasets (see Table 7.6 for performance of ULBP-SVM model). This was due to lower number of correct detections and higher FPs.

ULBP-M-SVM had 2.2% to 0.68% lower F-measure than ULBP-SVM in modelling and external test sets, respectively. However, in comparison to the PCA-based ULBP-SVM model, this ULBP-M-SVM performed better in terms of precision and F-measure. ULBP-SVM had the poorest performance in comparison to HOG-, GLCM- and combined-SVM models since it captures only the global textural features, insufficient for correct classification of breast lesions. Reducing this feature vector further led to further drop in performance.

Dataset	**ULBP-M-SVM**		
	Precision	**Recall**	**F-measure**
A	82.95	98.97	90.25
Overall External Test Sets	72.38	88.88	79.79
B	84.00	99.21	90.97
C	58.89	75.55	66.19
D	78.00	93.26	84.95
E	81.07	100.00	89.54

Table 7.9: Performance of U-DetectH-Base with ULBP-M-SVM model.

HOG and GLCM Feature Fusion: A fusion of HOG and GLCM feature vectors was studied as, individually, these feature vectors have a relatively high performance. An SVM model is trained on this fused feature vector using the same training set described in Section 7.2.1 and classification performance of this SVM model is presented in Table C.6 in Section C.1 Appendix C. The (HOG+GLCM)-SVM model is used in the U-DetectH-Base model. This U-DetectH-Base model is referred to as (HOG+GLCM)-SVM for brevity. Decision fusion between the classification score assigned by the SVM model and the base network is achieved using the weighted sum method described in Section 7.1.4. Optimal weights of $w_{svm} = 0.1$ and $w_{base} = 0.9$ were found experimentally.

Table 7.10 shows the performance of this SVM model in U-DetectH-Base. (HOG+GLCM)-SVM model does not outperform the combined-SVM (see Table 7.4 for performance of combined-SVM model). However, it combines the characteristics of both GLCM and HOG leading to an overall higher F-measure than both GLCM-SVM and HOG-SVM (See Table 7.6 for performance of HOG-SVM and GLCM-SVM). Here, the number of TPs were higher than either model along with a lower number of FPs which led to its higher precision. However, there was a small increase in the number of FNs. But due to the larger increase in TPs, recall of the (HOG+GLCM)-SVM model was also

higher than both models. Thus, combination of these two features improved TPs of GLCM as well as reduced FPs and FNs of HOG model.

Dataset	**(HOG+GLCM)-SVM**		
	Precision	**Recall**	**F-measure**
A	86.42	99.01	92.28
Overall External Test Sets	77.19	89.63	82.95
B	89.33	99.26	94.04
C	65.09	76.84	70.84
D	80.36	94.24	86.75
E	89.35	100.00	94.38

Table 7.10: Performance of (HOG + GLCM)-SVM model in U-DetectH-Base model: Single Fold.

7.4 Summary

This chapter presented a novel approach for breast lesion detection in US images. In particular, a new method (U-DetectH) to address common issue cases of the U-Detect method, namely, single low IOU FPs, FP+FN and FN was presented and evaluated. These issue cases were caused due to incorrect classification scores assigned to proposals which led to poor quality (low IOU) of candidates selected from each cluster. Therefore, handcrafted features that capture lesion characteristics were used to train SVM classifier which was used in conjunction with the network's classifier in order to improve the classification scores. Particularly, methods to extract textural features relating to echogenicity, margin, calcification, and aspect ratio were selected.

Total of five features were investigated, namely, HOG, GLCM, ULBP, aspect ratio (shape) and a combined feature vector which is a fusion of all features. The proposed U-DetectH consists of six phases of which three phases are common with the U-Detect method. This chapter focused on the phases unique to the U-DetectH method (phase one, two and four). Depending on the type of learned features used, two methods called U-DetectH-Base and U-DetectH-RPN were developed. Similar to U-Detect, the U-DetectH method is also only used during model testing. The datasets used for the development and testing of these methods are the same as those used for U-Detect models. Dataset A was used as modelling dataset and datasets B to E were used as external test sets. U-DetectH-Base with SVM model trained on combined feature vector outperformed U-Detect method with 2.41% higher precision along with 0.14% higher recall than U-Detect-Base. The improvement in both pre-

cision and recall is reduction of issue cases of U-detect method as well as an increase in number of correct detections.

Consequently, U-DetectH-Base also outperformed original and adapted FRCNN models. This model had 5.49% to 25.37% higher precision than the original FRCNN which is a result of 30% to 70% lower FP detections along with a small drop of 1.55% to 4.89% in recall. Compared to the adapted FRCNN, this U-DetectH method has 14.72% higher precision and only 0.37% lower recall. On the other hand, U-DetectH-RPN did not outperform U-Detect-RPN due to the lower classification accuracy of the RPN. Overall, computation time of the U-DetectH models is higher than that of U-Detect models due to the additional time required for extraction of all five handcrafted features for an average of 300 proposals generated for each image. The main contributing factor for the high computation time is the extraction of GLCM features. Therefore, improving extraction of this feature vector would considerably reduce computation time of the U-DetectH models.

Following is a summary of the main findings of this chapter:

- The investigation presented in this chapter shows the effectiveness of using handcrafted features in improving the classification accuracy of the base network for reduction of FPs as well as increasing the number of detected lesions.
- U-DetectH-Base model outperformed U-Detect-RPN due 6.81% to 14.68% lower FPs. This resulted in 0.89% and 2.75% higher precision of the U-DetectH-Base model. Despite 0.27% to 0.57% lower recall, the U-DetectH-Base model had 0.43% to 1.41% higher F-measure than U-Detect-RPN.
- HOG features were useful in detection of high-IOU proposals. However, it also caused incorrect classification of lesion-like background regions leading to an increase in FPs.
- GLCM features proved useful in the detection of challenging lesions. Since the extracted features are global, the GLCM-SVM model successfully detected challenging lesions with unclear boundaries. However, the HOG-SVM model struggled to detect such lesions as it relies on strong local textural features.
- The ULBP-SVM model had the poorest performance due to insufficient textural information

contained in this feature vector for correct classification of the proposals.

- Combination of both textural and morphological features had the best performing SVM model due to the increased information contained in the feature vector.
- Reducing the dimension of combined feature vector using PCA led to an overall drop in performance of U-DetectH models.
- All SVM models were trained to identify high IOU proposals as positive samples. Due to the generally poorer classification accuracy of the RPN, use of the U-DetectH method had a negative impact on the overall performance.
- U-DetectH models using combined feature based-SVM model had higher computation time than the U-Detect models due to the added time required for extraction of all handcrafted features (HOG, GLCM, ULBP and shape) for every test proposal. In particular, extraction of GLCM features required the highest computation time.

In conclusion, the best performing model of this chapter, U-DetectH-Base with combined-SVM, outperforms all both U-Detect-RPN and U-Detect-Base. Use of a combination of textural and morphological features helped not only in reduction of FPs but also increase in number of correct detections. Although the computation cost of this model is higher than U-Detect-Base, improvement in the feature extraction process can address this issue and provide a better balance between computation cost and performance.

Chapter 8

Discussion

In this thesis, we presented our work on automating breast lesion detection in 2D US images using deep CNN. In particular, we focused on FP reduction through first an extensive study of a popularly used FRCNN network and successfully adapting it for breast lesion detection in US images using our large and varied dataset collected from real-life clinical settings. Our adapted FRCNN model outperformed the original FRCNN through a significant reduction of FP detections and a small drop in the number of correct detections. We then presented two novel methods, U-Detect and U-DetectH, that further reduce the FP detections of our adapted FRCNN model. In this chapter, we present important evaluations relevant to the work presented in this thesis.

In Chapter 5, we used dataset A-small for selection of optimal values of investigated modelling hyperparameters of the FRCNN model. In Section 5.1.1, we present the performance of the same modelling hyperparameters when tested with dataset A used for modelling, highlighting the reproducibility of the selected optimal values with the different modelling datasets. After development of the adapted FRCNN model, we investigated training losses and various state-of-the-art classification networks as the backbone of adapted FRCNN in order to improve the classification accuracy of this model for reduction of FP detections. In Section 8.1, we present two other methods that were also investigated in order to improve classification accuracy of the FRCNN network, namely, impact of GARPN [75] and fusion of convolution layers of the backbone network.

Our novel U-Detect method, presented in Chapter 6, successfully reduced FP detections and outperformed both original and adapted FRCNN. The FP detections of this method were caused due to

selection of lower IOU FP detections from clusters containing high IOU TP detections. We addressed these cases in Chapter 7 where we presented a novel U-DetectH method that utilises handcrafted features to improve the classification scores assigned to high IOU TP and, conversely, reduce the scores of low IOU FP detections. For this, an SVM model was trained on the selected handcrafted features. The training set selected for this SVM model is described in Section 7.2.1 of Chapter 7. Other training sets were also evaluated for the SVM model and are described in this chapter in Section 8.3.

Dimensionality reduction of the handcrafted features using PCA was also investigated in Chapter 7. PCA is a 'feature projection' based method commonly used for reducing dimension of handcrafted features. Thus, use of a 'feature selection' type of dimension reduction method, specifically MRMR [157], is investigated to study its impact on the overall performance. This investigation is presented in Section 8.4. Furthermore, in U-DetectH, classification score assigned to each proposal is updated through a decision fusion of the scores assigned by the RPN/base network and the SVM model. Weighted sum method was used for the decision fusion. Section 8.5 provides experimental evaluation of the investigated weights of RPN/base network and SVM as well as justifies the selection of $w_{svm} = 0.1$ and $w_{rpn/base} = 0.9$ as optimal. All detection methods investigated in this research are based on supervised learning. We also investigated the performance of a reinforcement learning (RL) based network for breast lesion detection in our dataset of US images. This study is presented in Section 8.6.

8.1 FRCNN Hyperparameter Reproducibility

In Chapter 5, various modelling hyperparameters of the FRCNN model were investigated to study their impact on the overall performance. A range of values of these hyperparameters were evaluated and the values with the best performance (highest F-measure) on the modelling dataset (dataset A-small) were selected as optimal. An adapted FRCNN model was then designed using the selected optimal values. As shown in Section 5.3.2 of Chapter 5, adapted FRCNN model outperformed the original FRCNN model through considerable reduction in FPs and small drop in number of correct detections. With the availability of a larger dataset, the same investigation of the modelling hyperparameters (detailed in Section 5.1.1) was performed with dataset A used as modelling dataset to study the reproducibility of the optimal values. Optimal values were then selected based on the performance

over this modelling dataset. This section details the difference and similarities in the optimal values selected based on the performance on dataset A-small and A.

Irrespective of the modelling dataset used, anchor boxes and number of test proposals maintain the same optimal values. On the other hand, the number of samples required to train the base network and optimal thresholds for selection of these samples differ when dataset A is used as the modelling dataset. The optimal values of the number of training samples and the training sample IOU threshold varies in the larger modelling dataset due to the increase in variation of the training samples. Dataset A-small is a subset of dataset A as described in Section 5.3. While dataset A-small consists of 524 images collected from one hospital, dataset A consists of 1733 images collected from two hospitals. Such an increase in the variation of the training set requires a higher number of training samples and broader IOU range for selection of training samples in order to utilise the variety of samples in the larger dataset. However, performance of the optimal values selected from dataset A-small have second highest performance in dataset A, generally very close to that of the highest performing value. A deeper explanation of the individual hyperparameters on the new dataset is as follows.

The optimal number of training proposals increased from 300 to 1000 when the dataset was changed from A-small to A. However, an important point to note is that although 1000 proposals had the highest F-measure in dataset A, performance 300 training proposals had the second highest F-measure, very close to that of 1000 proposals. Based on their performance on dataset A-small, $[0.7, 1]$ threshold range was selected as the optimal IOU range for selection of positive training samples. When evaluated using dataset A, the optimal threshold for positive samples was experimentally found to be $[0.5, 1]$. As mentioned in Section 5.3.2.2 Chapter 5, positive training samples have a direct impact on the model's accuracy in classification of high IOU TP proposals as well FP proposals containing background lesion-like regions. When a smaller IOU range is used for selection of positive training samples, the model becomes sensitive to background regions in proposals. Furthermore, challenging lesions with background-like texture are incorrectly classified as background by the base network. Also, for such lesions, the number of high IOU proposals generated by the RPN is relatively lower.

Due to large variation in the lesions and images in dataset A, the variation in the background region is also higher. Thus, when evaluated on this dataset, the optimal threshold for positive samples

has a larger IOU range to include more variation of the background region as well as ensure correct classification of the challenging lesions. It is important to note here that the difference in performance of models trained with thresholds $[0.5, 1]$ and $[0.7, 1]$ is small. The model trained with an IOU range of $[0.7, 1]$ for positive sample selection had the second-highest F-measure which was only 0.07% lower than that of the best performing model trained with optimal IOU range of $[0.5, 1]$. Optimal thresholds for negative samples, selected based on their performance on dataset A-small (modelling dataset), was $[0, 0.2)$ which changed to $[0, 0.4)$ when trained on dataset A. Like with positive threshold, the larger variation in the background region requires a broader range in the threshold for negative samples. But the difference in performance of the optimal values selected based on the two datasets is small.

An adapted FRCNN model is trained on the optimal values selected based on performance on dataset A. Table 8.1 shows the performance of this adapted FRCNN model in comparison to that trained with optimal values selected based on performance on dataset A–small and original FRCNN model. Both adapted FRCNN models outperform the original FRCNN through effective reduction in FPs with small negative impact on the number of detected lesions. Adapted FRCNN trained on values selected from dataset A has around 2.58% to 4.42% higher F-measure in modelling and overall unseen test sets, respectively. Thus, the small drop in performance of the adapted FRCNN based on dataset A-small over our large datasets of 3119 images (of which 4 datasets were unseen external test sets collected from multiple hospitals and generated using various US machines) is evidence of the generalisation capability of the selected optimal values.

Dataset	**Original**			**Adapted (Dataset A)**			**Adapted (Dataset A-small)**		
	P	R	F	P	R	F	P	R	F
A	78.45	**96.23**	86.32	**81.71**	95.94	**88.25**	78.44	94.49	85.67
Overall External Test Sets	51.99	**94.71**	67.12	**72.40**	87.47	**79.19**	70.90	79.09	74.77

Table 8.1: Precision (P), Recall(R) and F-measure (F) of Original and Optimal Faster R-CNN using dataset A-small and dataset A (Average of 5-folds).

This investigation highlights the impact of the dataset on the selection of optimal hyperparameters. As this study is performed on two datasets, the following important points can be drawn: 1.) Despite the large variation on these datasets, there was no change to the optimal values of anchor boxes and number of test proposals proving their generalisation capabilities. 2.) Although the optimal values

for number and IOU range for base network's training sample selection varies with the larger dataset, the change in performance of the individual hyperparameters as well as the overall models trained with their respective optimal values is small. This is evidence of the overall generalisation capability of these hyperparameter values.

8.2 Improving Classification Accuracy of Adapted FRCNN Model

As discussed in Section 5.3.3 Chapter 5, the adapted FRCNN outperformed the original FRCNN due to lower number of FPs. However, the number of FPs in the adapted FRCNN was still considerable. One of the reasons for these FPs was poor classification of FP proposals. Thus, evaluation of various training losses and classification networks as the backbone network were studied and presented in Sections 5.3.5 and 5.3.4 in Chapter 5. Besides these strategies, two other techniques were also investigated, namely, Guided-Anchoring based RPN (GARPN) [75] and fusion of convolution layers of the backbone network. GARPN modifies the first stage (RPN) of the FRCNN model in order to improve the quality of proposals which would in turn improve the overall performance of the model including lower number of FPs. The second investigated technique is the fusion of convolutional layers of the backbone network to improve the quality of features used by both stages of the adapted FRCNN model and thereby improve the quality of proposals and output detections of the model. The remainder of this section describes both these techniques and their impact on the overall performance in further detail.

GARPN: GARPN was developed for object detection in natural images. It replaces the default RPN network of the FRCNN model to improve the quality of proposals generated by the RPN. Thus, lower number of FP proposals and higher number of TP proposals are sent through to the base network for its training and during test time thereby reducing FPs in the final output and improving the overall performance of the model. In particular, the classification and regression branches of the RPN are replaced with two new branches, namely, shape prediction branch and location prediction branch. The shape prediction branch predicts the shape (width and height) of a potential object on every point of the feature map whereas the location prediction branch generates a probability map which contains the probability of the presence of an object at every location of the input convolutional feature map. The output of both these branches is then combined such that only shapes generated at high proba-

bility locations are sent through for further processing. The architecture of this network is described in further in Section 3.1.2 in Chapter 3.

The GARPN model was trained and tested on five folds of dataset C. For comparison, the adapted FRCNN model was also trained and tested on the same folds of dataset C. Overall, GARPN had 17.34% higher precision and 23.06% lower recall. The high precision of this model is due to the lower number of FPs generated. This is due to the use of the location prediction branch which filters out FP proposals at the RPN. As these FP proposals are not passed to the base network, the number of FP detections output by this model is lower. However, the GARPN model has a comparatively lower recall than the adapted FRCNN model which is a result of higher number of missed lesions. Due to the lower classification accuracy of the RPN in the GARPN model, the high IOU proposals covering challenging lesions are not passed through to the base network resulting in missed lesions. Thus, owing to the lower recall, the overall F-measure of the GARPN model is lower than that of the adapted FRCNN despite its higher precision.

Layer Fusion: To improve the classification accuracy of the adapted FRCNN model, fusion of convolutional layers of the base network was investigated. Initial convolutional layers of a network extract low-level features such as edges and contours. As the layer depth increases, the features extract become more abstract. In deeper layers, the resolution of the low-level features reduces, especially for objects of small size. Fusion of initial convolutional layers with the deeper convolutional layers introduces the low-level features with the abstract features. Thus, the fused features maps hold higher textural information. As classification accuracy of both stages of the FRCNN model relies on the quality of the input features, impact of layer fusion on the reduction of FPs was studied.

This study was conducted on adapted FRCNN with ResNet50 as its backbone network. The ResNet50 network is divided into five convolutional blocks. Feature map output by convolution block 4 (conv4) is used by the RPN for proposal generation. After proposals are generated by the RPN, ROI pooling is performed on the same feature map for further processing of the proposals by the base network. For this study, the output feature map of convolutional block 1 (conv1) is fused with the output feature map of conv4. The fused feature map is then used by both RPN and base network. Overall, the adapted FRCNN model using fused layers had higher recall but lower precision than

the adapted FRCNN model that does not use layer fusion. Due to the larger drop in precision than the improvement in recall, the F-measure of this fused model was lower than that of the adapted FRCNN. The recall of the model was higher due to higher number of detected lesions (higher TP and lower FN). However, due to the higher number of FPs generated by this model, its precision was lower.

Furthermore, as the number of fused layers increased, a similar performance change was noted (higher recall and lower precision) due to an increase in detected lesions along with an increase in FPs. The introduction of the low-level features improved the detection of small lesions whose features are lost in deeper layers. However, this fusion also caused incorrect classification of FP proposals covering background lesion-like regions both at the RPN as well as base network leading to an increase in FP detections. This is because of similarity in lower-level textural features (such as edges and contours) of the background lesion-like regions and the lesion.

8.3 Training Samples Selection for SVM Models in U-DetectH

U-Detect models had the following common issue cases: single low IOU FP, FP+FN (where a single detection is output by the model which covers background region and completely misses the lesion resulting in a FP and FN) and FNs. The U-DetectH model proposed in Chapter 7 addressed these issue cases with the help of handcrafted features that were selected with reference to characteristics of lesions in US images used by radiologists to assign its BI-RADS score. In particular, an SVM model trained on a fusion of handcrafted features was used to improve classification score assigned to the proposals by RPN/base network. The aim of the SVM model was to improve the scores assigned to high IOU proposals and reduce the scores assigned to low IOU FP proposals thus ensuring the selection of TP proposals as candidates resulting in the reduction of FP detections of the model.

For training the SVM model, proposals generated by the RPN of the adapted FRCNN model for images in the same split of the modelling dataset (dataset A) were considered. Of these, proposals with $IOU[0.7, 1]$ were used as positive samples and 292 proposals, randomly selected from seven bins of $IOU[0, 0.7)$ (detailed in Table 8.2), were used as negative samples. This training set (TS) is referred to as optimal TS and used for SVM training in Chapter 7. Besides this training set, three other TSs, derived from the same proposals, were also evaluated. These training sets are referred to as TS1, TS2

and TS3. This section details the performance of the SVM and U-DetectH models trained with the TS1, TS2 and TS3 training sets.

Table 8.2 shows the selected training samples in the four training sets. In TS1 and TS2, the positive samples are the same as the ones used in optimal TS. They differ from optimal TS in the selection of negative samples only. On the other hand, TS3 differs from optimal TS in both positive and negative sample selection. In TS1, only proposals with IOU $[0.6, 0.7)$ are considered as negative and those with *IOU* $[0.7, 1]$ are considered as positive samples. This was to focus purely on the common issue cases of single low IOU and FP+FN found in U-Detect models. The SVM model trained using TS1, in general, had the highest specificity but overall poor performance due to the limited variation in negative samples. Detailed performance of combined-SVM model trained using TS1 is presented in Table C.7 in Section C.1 Appendix C. Therefore, in TS2, the range for negative samples was increased to $[0.5, 0.7)$ to introduce more hard negative samples. This increase in negative samples' variation improved recall of the SVM model along with an increase in FPs and drop in TNs i.e. drop in accuracy, specificity, precision and F-measure. Performance of combined-SVM model trained using TS2 set is detailed in Table C.8 of Section C.1 Appendix C. When the range of negative samples was increased further so as to include hard as well as easy negative samples in the optimal TS training set, these metrics were balanced as evidenced through its performance detailed in Table 7.3 in Section 7.2.1 of Chapter 7.

<table>
<tr><th>Training Sample IOU</th><th>Optimal TS</th><th>TS1</th><th>TS2</th><th>TS3</th></tr>
<tr><td>0</td><td rowspan="8">Negative</td><td>-</td><td>-</td><td rowspan="8">Negative</td></tr>
<tr><td>$(0, 0.1)$</td><td>-</td><td>-</td></tr>
<tr><td>$[0.1, 0.2)$</td><td>-</td><td>-</td></tr>
<tr><td>$[0.2, 0.3)$</td><td>-</td><td>-</td></tr>
<tr><td>$[0.3, 0.4)$</td><td>-</td><td>-</td></tr>
<tr><td>$[0.4, 0.5)$</td><td>-</td><td></td></tr>
<tr><td>$[0.5, 0.6)$</td><td></td><td rowspan="2">Negative</td></tr>
<tr><td>$[0.6, 0.7)$</td><td>Negative</td></tr>
<tr><td>$[0.7, 0.8)$</td><td colspan="3" rowspan="3">Positive</td><td rowspan="4">Positive</td></tr>
<tr><td>$[0.8, 0.9)$</td></tr>
<tr><td>$[0.9, 1)$</td></tr>
<tr><td>1</td><td>-</td><td>-</td><td>-</td></tr>
</table>

Table 8.2: Evaluated SVM training sets. *Negative*: Negative training samples. *Positive*: Positive training samples.

In TS3, proposals with $IOU[0.7, 1]$ were selected as positive samples. Along with this, ground truth boxes i.e., $IOU = 1$ samples, were used as positive samples. RPN rarely generates proposals with absolute overlap of $IOU = 1$ with the lesion. Thus, introduction of these GT boxes was evaluated to study its impact on the overall performance. To balance the number of negative samples, 640 proposals from each IOU bin in range of $[0.4, 0.7)$, 632 proposals from IOU bins in the range of $[0.3, 0.4)$ and 292 proposals from IOU bins in the range of $[0, 0.3)$ were selected. Higher number of proposals from higher IOU bins were selected to introduce more challenging negative samples in comparison to easier negative samples with lower IOU. Classification accuracy of this model is detailed in Table C.10 in Section C.1 Appendix C.

An SVM model trained using GLCM features and TS3 samples is utilised in the U-DetectH-Base model. Decision fusion was performed using weighted-sum method where $w_{svm} = 0.1$ and $w_{base} = 0.9$ (same weights used for all SVM models in Chapter 7). This U-DetectH-Base model is referred to as GLCM-new for brevity. Table 8.3 shows the performance of the GLCM-new model. The performance of the GLCM-new model is compared to that of the U-DetectH-Base model using GLCM-SVM (referred to as GLCM-SVM model). Performance of the GLCM-SVM model was presented in Table 7.6 in Section 7.2.2 Chapter 7.

The GLCM-new model had a small improvement in the overall detection performance due to lower number of FPs and FNs. However, the number of TPs in the GLCM-new model was lower than that of GLCM-SVM. Thus, use of GT boxes in the training of the SVM model improved classification accuracy of the model. This in turn improved the classification score assigned to TP proposals covering challenging lesions as well as FP proposals covering lesion-like background regions (additional boxes). On the other hand, multiple lesions were detected with low IOU FP proposals by the GLCM-new model. This is because of the higher classification score assigned to lower IOU FP proposals in comparison to TP proposals present in the cluster. Thus, use of TS3 improves the classification of challenging lesions and additional boxes but at the expense of low IOU FP.

Dataset	GLCM-new		
	Precision	**Recall**	**F-measure**
A	85.60	99.07	91.85
Overall External Test Sets	76.34	89.75	82.51

Table 8.3: Impact of new training set on GLCM-SVM in U-DetectH-Base model: Single Fold.

8.4 Dimension Reduction using MRMR in U-DetectH Models

In Chapter 7, dimension reduction of the handcrafted features using PCA in the U-DetectH model was investigated. Use of PCA was experimentally found to result in a lower performance overall. This was due to poorer classification of FP proposals covering background regions as well as low IOU FP proposals. However, the change in number of FNs did not undergo significant change. Thus, the dimensionally reduced feature vector contained insufficient textural information for correct classification of FP proposals only. PCA is a feature projection method of dimension reduction. In order to study the impact of the feature selection method for dimension reduction, MRMR (Maximum Relevance - Minimum Redundancy)[157] was evaluated.

While PCA projects the feature vector into another space for dimension reduction, MRMR selects a predefined number of components (mC) from the feature vector where the number of components is less than the dimension of the feature vector. In particular, MRMR ranks all elements of the feature vector in terms of its importance, ensuring low redundancy between the selected top elements. If $mC = 10$, then the top 10 ranking elements are selected. The impact of MRMR is studied on the U-DetectH-Base model. Particularly, Base-GAP features and handcrafted features of each proposal are extracted in phase one. In phase two, KPCA is used to reduce dimension of Base-GAP features and MRMR is used to reduce the dimension of handcrafted features. Proposals are clustered using the dimensionally-reduced Base-GAP features in phase three. In phase four, using the dimensionally reduced handcrafted features, the proposals are classified by a pretrained SVM model. Weighted sum is used for decision fusion of the classification scores of base network and SVM model in phase four, using the same optimal weights ($w_{base} = 0.9$ and $w_{svm} = 0.1$) as used by U-DetectH models. In phase five, candidates from each cluster are selected based on the updated proposal score and in phase six, the selected candidates are processed through NMS and candidate merging method.

Like PCA, the optimal number of components (mC) needs to be predefined. Thus, an analysis of the number of mCs required to capture 99% of the textural information was conducted. This study was performed on a single fold of modelling dataset (dataset A) using the combined feature vector extracted for each proposal. Combined feature vector is a fusion of HOG, GLCM, ULBP and shape. This fold is the same as the one used for PCA and KPCA investigations in Chapters 6 and 7. Based on this investigation, the range of mCs to be evaluated was determined. Following mCs were evaluated : $85, 95, 105, 115, 125$.

Figure 8.1 shows the performance of the investigated mCs. Overall, the overall performance dropped with the increase in the number of mCs. Higher number of mCs had lower TPs and higher FPs. Thus, 85 mCs was selected as optimal based on its highest F-measure. Using the optimal mCs, an SVM model was trained on the same training set detailed in Section 7.2.1. Classification performance of this SVM model is presented in Table C.10 in Section C.1 Appendix C. U-DetectH-Base model with MRMR-based combined-SVM models using optimal mCs was trained on a single fol of dataset A (modelling dataset); same fold used in Section 7.2.2. Table 8.4 shows the performance of this U-DetectH-Base with MRMR-based combined-SVM. For brevity, U-DetectH-Base models are referred to by the feature vector used. Thus, combined-SVM, combined-PCA-SVM and combined-MRMR-SVM refer to U-DetectH-Base models using no dimension reduction on the combined feature vector, using PCA for dimension reduction and using MRMR for dimension reduction, respectively. Performance of combined-SVM and combined-PCA-SVM models are presented in Table 7.5.

Dataset	**Combined-MRMR**		
	Precision	**Recall**	**F-measure**
A	85.55	99.00	91.78
Overall External Test Sets	78.17	89.91	83.63

Table 8.4: Performance of combined-MRMR in U-DetectH-Base model: Single fold.

Overall, the combined-MRMR-SVM model has considerably higher F-measure than combined-PCA-SVM and a marginally higher performance in comparison to combined-SVM model. Compared to combined-PCA-SVM, combined-MRMR-SVM has a higher number of TPs as well as lower number of FPs. Thus, combined-MRMR-SVM has 2.22% to 7.62% higher F-measure than combined-PCA-SVM model due to 3.76% to 12.42% higher precision with 0.05% to 2.12% higher recall in modelling

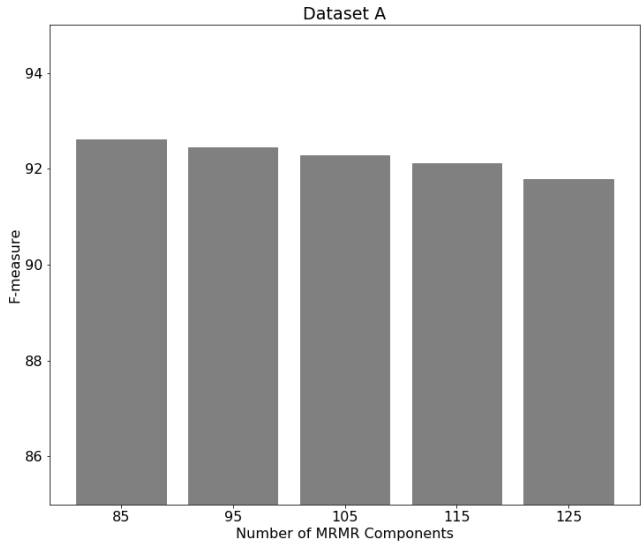


Figure 8.1: Impact of various MRMR components on performance of all combined feature vector in U-DetectH-Base model: Single fold of dataset A.

and unseen external test sets, respectively. Compared to combine-SVM, combined-MRMR-SVM had 0.51% to 1.16% higher precision and 0.01% to 0.42% higher recall resulting in 0.33% to 0.48% higher F-measure. However, in dataset E, both combined-MRMR-SVM and combined-SVM models had the same performance. Furthermore, in dataset B, combined-MRMR-SVM had 0.67% lower precision, 0.01% lower recall and 0.36% lower F-measure than combined-SVM model.

Use of MRMR improved the performance through a successful removal of redundant and noisy features. However, the improvement in performance is small. Additionally, MRMR also provides useful conceptual information regarding important, distinguishing features, a characteristic that is absent in PCA. For instance, through investigation of various mCs of GLCM feature vector, it was found that contrast and correlation were consistently ranked higher. Also, features extracted from GLCM matrices computed for (-1,0), (0,2), (-4,4) and (-4,-4) pixel relationships were consistently used in the MRMR feature vector. This indicates that the mentioned features and pixel relationship provide the most important textural information that helps distinguish proposals containing lesions from those containing background region.

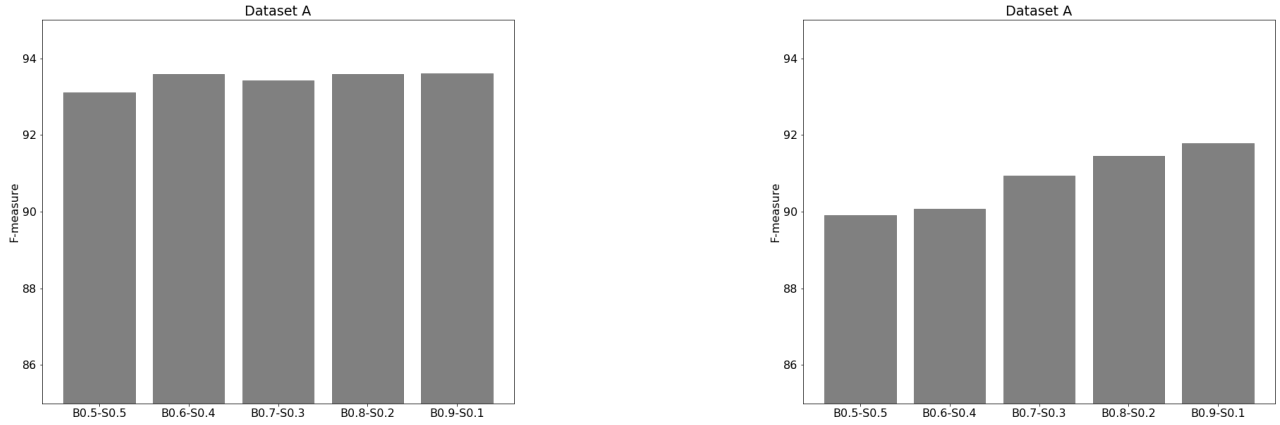
8.5 Weighted Sum Analysis in U-DetectH Models

The U-DetectH models proposed in Chapter 7 use an SVM model in conjunction with RPN/base network to update the classification scores assigned to proposals. The decision fusion method used for updating the proposal score is weighted sum. The optimal weights of $w_{rpn} = w_{base} = 0.9$ and $w_{svm} = 0.1$ were found experimentally and used in all U-DetectH models. The experimental evaluation of all weights investigated as well as the selection of aforementioned weights as optimal is presented in this section.

Weights in the range of $(0, 1)$ with increments of 0.1 were investigated with the condition that $w_{rpn/base} + w_{svm} = 1$. Table 8.5 details the weights were investigated on a single fold of modelling dataset (dataset A). The same fold is used for evaluation of weights for all U-DetectH models. For illustration, consider U-DetectH-Base models using GLCM-SVM and HOG-SVM. Figure 8.2 shows the impact of the investigated weights on the overall performance of U-DetectH-Base models. In both models, higher the w_{base}, higher is the F-measure of the model. Such a performance trend was seen in all U-DetectH models. This is because the RPN/base network have overall higher classification accuracy in comparison to SVM models. Furthermore, these models are trained to assign lower scores to TP proposals with IOU in the range $[0.5, 0.7)$. Thus, with lower weight assigned to the RPN/base network, the number of detected lesions is lower leading to overall lower F-measure. Based on the highest F-measure over a single fold of modelling dataset, $w_{rpn} = w_{base} = 0.9$ and $w_{svm} = 0.1$ are selected as optimal and used in all U-DetectH models using one SVM model in the network.

Test	w_{svm}	w_{rpn}/w_{base}
1	0.1	0.9
2	0.2	0.8
3	0.3	0.7
4	0.4	0.6
5	0.5	0.5

Table 8.5: Investigated weights for weighted sum decision fusion of a single SVM model and RPN/base network in U-DetectH models.



(a) GLCM (b) HOG

Figure 8.2: Change in F-measure with change in weights assigned to the base network, GLCM-SVM and HOG-SVM in U-DetectH-Base model. Base network is represented using B whereas both GLCM-SVM and HOG-SVM models are represented using S.

8.6 Reinforcement Learning (RL) for Breast Lesion Detection

As part of the investigation and study of the various detection methods developed for breast lesion detection, a RL based method was also investigated. As described in Section 2.2.3.3 in Chapter 2, RL differs from both supervised and unsupervised learning. It is an emerging field of study. When used for detection, RL agents automatically learn the detection during model training using a feedback signal which is a scalar value. We studied the performance of a hierarchical RL network [5] developed for object detection in natural images. We refer to this network as HRL-US. Here, VGG16, pretrained for classification of natural images, is used for feature extraction. A deep Q-network is used as the agent. Figure 8.3 shows the possible steps that the agent can take for successful object detection. Step 0, i.e., the default starting state, is a bounding box covering the entire image. This bounding box is then divided into five bounding boxes covering five quarters; four quarters in each standard quadrant and one quarter covering the central region. The agent then selects one of these quarters that potentially contains the lesion (object) or it can decide to terminate the search and output the current bounding box. If the search is not terminated, then the selected quarter is divided into further five quarters and so on. The agent is permitted a maximum of 10 steps to detect an object in the image.

We trained the HRL-US model on 5-folds of dataset A-small. For this investigation, we also trained the HRL network for object detection in natural images using MS-COCO dataset in order to compare

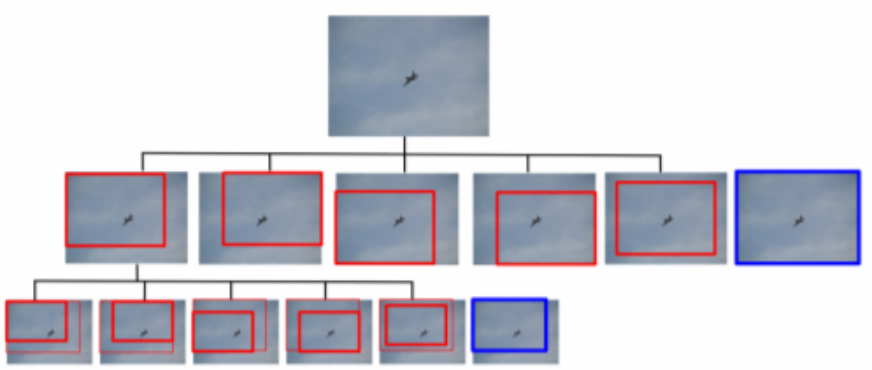


Figure 8.3: Action steps of the RL agent in HRL network [5]. Blue bounding boxes represent selected quarter from previous action step and red bounding boxes represent five possible quarters that the agent can select in that step.

the performance of this dataset on natural and US images. This network is referred to as HRL-natural. Both models are trained using the default hyperparameters detailed in [5]. The performance of these models was evaluated using precision-recall curve. Based on this curve, average precision (AP) was calculated for each model. HRL-natural had a higher performance than HRL-US. HRL-natural had an AP of 11.64. In HRL-natural, the agent terminated the search in two to three steps for the majority of the detections. Output detections with IOU $[0.1, 0.6]$ required two to six steps. On the other hand, AP of HRL-US was much lower than either of the natural images' models at 0.11. Also, the agent required an overall higher number of steps in HRL-US. In particular, in the majority (57.26%) of the US images, the agent used all 10 steps. In a significant proportion of these images, the lesion was completely missed. In 9.17% of the images, the agent terminated the search in step 2. Of all the output detections of HRL-US models, 61.64% cases had no overlap with the lesion (additional boxes), 37.79% had up to 50% overlap ($0 < IOU < 0.5$) (low IOU FP) and only 0.57% were TP detections with $IOU > 0.5$.

HRL performs poorly due to the inherent challenging nature of US images as well as poor designs of steps allocated to the agent. Nature of the agent's steps prevents correct detection of small lesions as the agent cannot change width or height of the bounding box. For the same reason, unless at least one of the quarters tightly fits the lesion, majority detections have low IOU. Additionally, incorrect selection in one critical step leads to missed lesions as the agent is not allowed to retrace its path and

modify its selection in this critical step. This performance can therefore be improved through two main improvements. First, improvement in the quality of features extracted for better representation of the current step. Secondly, improving the steps to allow greater flexibility to the agent. In summary, RL is a new, emerging field with promising results. However, one of its major requirements of large and varied dataset for optimal training of the agent limits its development in this field of breast lesion detection in US images. Therefore, this field of detection networks requires further development for successful application in medical images.

Chapter 9

Conclusion and Future Work

Automating breast lesion detection is an important field of research. Given the high average ratio of patients to available experienced radiologists, such an automation can assist radiologists in their diagnosis process and facilitate faster, accurate reading. In the field of lesion detection in US images, a fundamental issue that challenges the research community is false positive (FP) detections. This thesis presented methods for breast lesion detection in US images using deep convolutional neural networks and unsupervised learning. In particular, this thesis aims to address the issue of FPs in breast lesion detection in 2D US images. This chapter serves as the conclusion for the whole thesis and provides key contributions and findings from this research. It also outlines the possible future investigations and potential developmental direction of this research.

9.1 Summary of the Thesis

The main aim of this research is to develop and design a method for breast lesion detection in US images. Figure 9.1 illustrates the key components of this research. Component (A) is the adaptation of the FRCNN model for breast lesion detection in US images. First step in this adaptation was the study of various modelling hyperparameters and their impact on overall performance. Through this study, optimal values for each of the investigated modelling hyperparameters was determined. The FRCNN model trained using these optimal values, referred to as adapted FRCNN, outperformed original FRCNN through a significant reduction in FPs along with a small drop in the number of correctly detected lesions. Furthermore, the adapted FRCNN model outperformed several state-of-the-art detectors developed for object detection in natural images as well as those adapted for breast

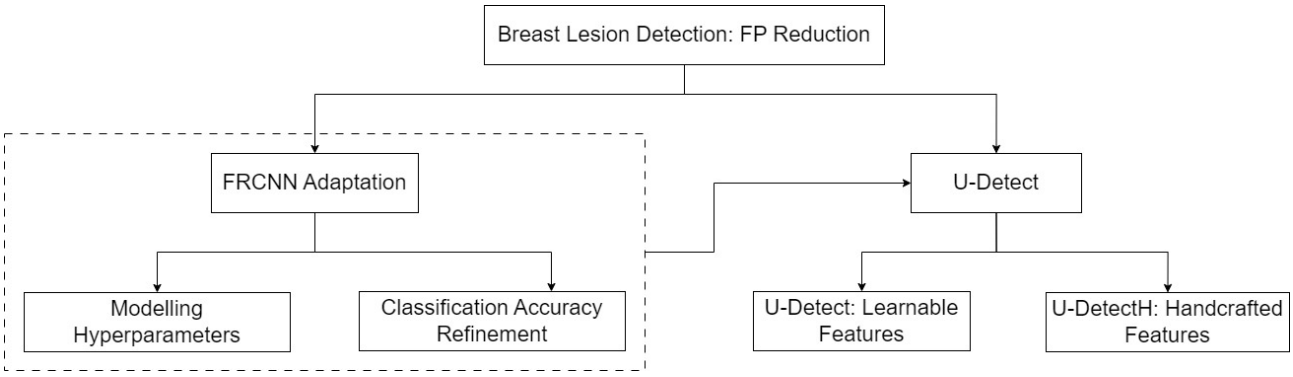


Figure 9.1: Key research components.

lesion detection in US images. Further modifications of the adapted FRCNN model were performed to improve its classification accuracy which would in turn reduce FP detections. These modifications included an investigation of various state-of-the-art classification networks as the backbone as well as different training losses. Through this investigation, the adapted FRCNN model was further modified. Specifically, the IRV2 network outperformed the default VGG16 as the backbone of the model. It not only had a lower number of FPs but also a higher number of correctly detected lesions. All investigated models improved the number of correct detections. However, the overall number of FPs were higher with the use of these losses.

The next main component (B) of this research is the novel U-Detect method which combines unsupervised learning with the adapted FRCNN model for reduction of FPs for breast lesion detection in US images. This model utilises learned features of the adapted FRCNN model to cluster proposals based on their textural similarity using x-means clustering method and selects highest scoring proposal from each cluster. Depending on the features extracted, the following two networks were developed: U-Detect-RPN and U-Detect-Base. Both these networks outperform original as well as adapted FRCNN models through considerably higher precision as a result of lower number of FPs. Additionally, U-Detect-RPN also improves the number of correct detections made by the adapted FRCNN resulting in not only higher precision but also higher recall than the adapted FRCNN model.

The third component of this work (C) is the novel U-DetectH method which utilises textural and morphological handcrafted features, selected based on domain knowledge, in order to improve the classification scores assigned by the adapted FRCNN model. This method builds on the U-

Detect method to improve the quality of proposals (improve). Following two U-DetectH models were derived: U-DetectH-RPN and U-DetectH-Base. Both these models outperformed the original and adapted FRCNN models due to a large reduction in FPs with a relatively smaller drop in the number of correct detections.

9.2 Main Achievements

The main achievements of this thesis starts from comprehensive investigations in Chapter 5 that evaluates the effectiveness of FRCNN and several state-of-the-art object detectors for breast lesion detection in 2D US images. We presented an extensive analysis of the FRCNN network and its limitations. We conducted a, first of its kind, systematic study of several modelling hyperparameters from both stages of the FRCNN network for breast lesion detection on our large dataset of US images. Optimal values of the evaluated modelling hyperparameters were selected during this study. The adapted FRCNN model designed using the selected optimal values outperformed the original FRCNN model through considerable reduction of 28% to 61% in FPs with a relatively small drop in the number of detection lesions. This study addresses an important gap in the current literature and has therefore been published to aid researchers in their understanding of the impact of these hyperparameters on the overall performance as well as for the adaptation of the FRCNN model for lesion detection in US images. Furthermore, when compared to multiple state-of-the-art detectors developed not only for object detection in natural images but also those adapted for breast lesion detection in US images, the adapted FRCNN model had the highest performance.

The adapted FRCNN model was modified further in Chapter 5 to improve its classification accuracy for FP reduction through an investigation of various training losses and state-of-the-art classification networks as backbone of the adapted FRCNN model. The evaluated losses improved the number of correct detections. Use of PISA loss in the classification branch and the default smooth L1 loss in the bounding box regression branch reduced the number of FPs albeit by a very small degree. On the other hand, use of deeper backbone networks led to a larger improvement in the classification accuracy as well as overall performance of the adapted FRCNN model. Of the evaluated classification networks, adapted FRCNN model using Inception-ResNet-v2 as the backbone had the best performance, outperforming adapted FRCNN using VGG16 (default backbone network of the FRCNN model) through

not only lower number of FPs but also higher number of correct detections. However, the number of FPs in the adapted FRCNN model was still significant.

Therefore, in Chapter 6, we proposed a novel U-Detect method that combines unsupervised learning with adapted FRCNN developed for FP reduction. The U-Detect method essentially intelligently filters proposals so that high IOU proposals at the RPN or base network are retained and low IOU FP proposals are discarded. This is achieved by texture-based clustering of the proposals and selection of high IOU proposals from each cluster with the help of their classification scores. We also proposed a novel candidate merging method used in the final phase of this method that reduces the number of overlapping FPs. Based on the learned features extracted to describe the proposals, two networks were derived, namely, U-Detect-RPN and U-Detect-Base. Both, U-Detect-RPN and U-Detect-Base, outperformed original and adapted FRCNN. Compared to adapted FRCNN, the U-Detect-RPN model has 8.78% to 47.53% lower number of FPs where the U-Detect-Base model has 8.45% to 32.72% lower number of FPs. Thus, the U-Detect methods successfully address the FP issue cases of the adapted FRCNN model. The U-Detect method is only used during model testing. Therefore, it adds no additional computation cost to model training. U-Detect-RPN has a high computation cost of 3.96 seconds but the U-Detect-Base model has a much lower computation cost of 0.96 seconds which is very close to that of adapted FRCNN model(0.57 seconds).

We presented the common issue cases of the U-Detect models. These issue cases were caused due to incorrect selection of proposals from the cluster containing high IOU TP proposals. Thus, in Chapter 7 we addressed these cases through a novel classification-based U-DetectH method. The U-DetectH model uses a fusion of textural and morphological handcrafted features to improve the classification scores assigned to high IOU proposals and reduce the scores assigned to lower IOU FP proposals. Selection of these features was inspired by the characteristics of breast lesions in US images studied by radiologists to assign the lesions' BI-RADS score. Particularly, HOG, GLCM, ULBP and shape features were used to train an SVM model.

The classification scores assigned to proposals were updated through a decision fusion of RPN/base network classification scores and SVM classification scores using weighted sum method. Based on the extraction of learned features, U-DetectH-RPN and U-DetectH-Base models were developed. Both

U-DetectH models outperformed original and adapted FRCNN models. Compared to U-Detect-Base, U-DetectH-Base had 0.13% to 4.59% higher F-measure due to its lower number of FPs as well as higher number of TPs. Furthermore, U-DetectH-Base also outperformed U-Detect-RPN. Despite the lower recall, the U-DetectH-Base model has 0.43% to 1.41% higher F-measure than U-Detect-RPN due to its higher precision. However, U-DetectH-RPN did not outperform U-Detect-RPN due to the relatively lower classification accuracy.

U-DetectH-Base had 6.08% to 38.89% lesser FP detections than the adapted FRCNN model which led to 0.77% to 8.28% higher precision. Although it had 0.77% to 0.62% drop in recall, the overall F-measure was 0.36% to 4.79% higher due to the comparatively higher precision. Similarly, compared to the original FRCNN, the U-DetectH-Base model had 31.86% to 77.07% lower number of FPs resulting in 5.49% to 32.83% higher precision. Due to a relatively small drop in the number of detected lesions, recall of the U-DetectH-Base model is 0.27% to 10% lower than the original FRCNN model. Thus, owing to the higher precision than the original FRCNN model, U-DetectH-Base had 4.59% to 22.74% higher F-measure. Additionally, we also presented the impact of individual features (HOG, GLCM and ULBP) on the performance of both U-Detect models. Due to the additional time required for extraction of all handcrafted features, U-DetectH models had higher computation time in comparison to their U-Detect counterparts.

In summary, in this thesis, we address important gaps in the literature. First, we address the drawbacks of existing methods to modify the FRCNN model for breast lesion detection in US images through an investigation of the impact of FRCNN modelling hyperparameters on a large US dataset. FP detection is a prevalent issue not only in breast lesion detection for US images but also for lesion detection in US images. Although several works have adapted the FRCNN model or proposed novel detectors, the issue of FP detections has not been adequately addressed. In this work, we bridged this gap through development of novel U-Detect and U-DetectH models that successfully reduce FPs. This work has been developed on a large and varied dataset of US images collected from a variety of US machines from a range of hospitals based in different countries. The effectiveness of our adaptation and novel methods (U-Detect and U-DetectH) on such a dataset strongly suggests high generalisation capability.

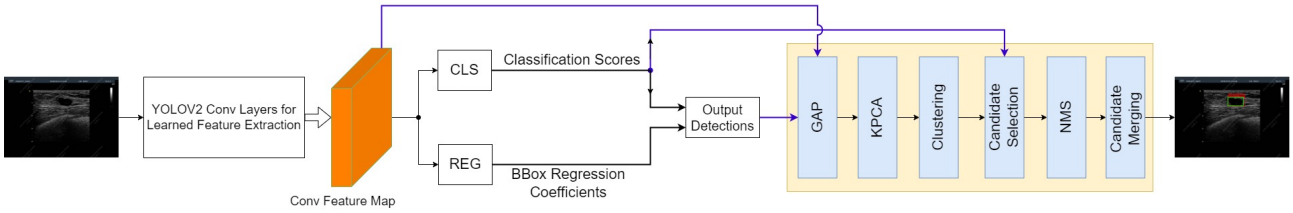


Figure 9.2: U-Detect method in YOLOv2 detector.

9.3 Future Work

The research presented in this highlighted several important future works that complement the investigation of adapting our U-Detect and U-DetectH methods.

Evaluation of U-Detect and U-DetectH methods for detection of other types of lesions: Both U-Detect and U-DetectH were developed for breast lesion detection in US images. However, a recent work [139] argues that breast and thyroid lesions have similar characteristics in US images. Thus, a promising direction of this work is an investigation of U-Detect and U-DetectH methods in detection of other types of lesions. Such a study could provide valuable information in terms of the generalisation capabilities of the proposed methods.

Evaluation of U-Detect and U-DetectH methods with different base detectors: Both U-Detect and U-DetectH methods are used during the model's test-stage. Therefore, irrespective of the nature of the detector, both these methods can be utilised. This investigation would prove insightful in understanding the adaptability of the U-Detect and U-DetectH methods. For example, consider the U-Detect method used for FP reduction in the YOLOv2 network as shown in Figure 9.2. Output of the YOLOv2 network is the classification score assigned to all *cells* in the image and the bounding box regression for all anchor boxes assigned to each cell. Therefore, the proposals here are bounding boxes that have been classified as an object and transformed as per the output of the bounding box regression branch. In phase 1 of the U-Detect method, these proposals are defined using the learned features output by the last convolution layer of the YOLOv2 backbone network. If the backbone network used in YOLOv2 is DarkNet-19 and the input image is 416×416, then the dimension of the learned features for each proposal is $13 \times 13 \times 1536$.

Given the large dimension, in phase 2, GAP can be first applied to reduce the feature vector to 1536×1 dimension, followed by KPCA which further reduces it to a feature of size 5×1. These feature vectors are used to cluster proposals using x-means clustering in phase 3 of the U-Detect method. Candidates from each cluster are selected in phase 4 based on the classification score assigned to each proposal. The candidates then are processed through NMS to remove redundant boxes. In phase 5, candidate merging method is applied to remove overlapping FPs.

Furthermore, U-DetectH uses handcrafted features selected to extract important features of breast lesions in US images. In recent work [139], the similarity between breast and thyroid lesions in US images was highlighted. Therefore, evaluation of the U-DetectH in detection of other lesions could provide valuable information in terms of the generalisation capabilities of the U-DetectH model for detection of other lesions in US images. Considering the above example of the trained YOLOv2 detector for thyroid lesion detection in US images, Figure 9.3 illustrates the use of U-DetectH in this network. As the combined feature vector is proven to be most effective, all four handcrafted features (HOG, GLCM, ULBP and shape) are extracted for each proposal. Decision fusion (weighted sum) of the SVM model trained separately and the YOLOv2 classifier provides the updated classification score for all proposals. After proposals are clustered using the learned features, candidates are selected on the basis of the updated classification score. Remaining phases are the same as the U-Detect method.

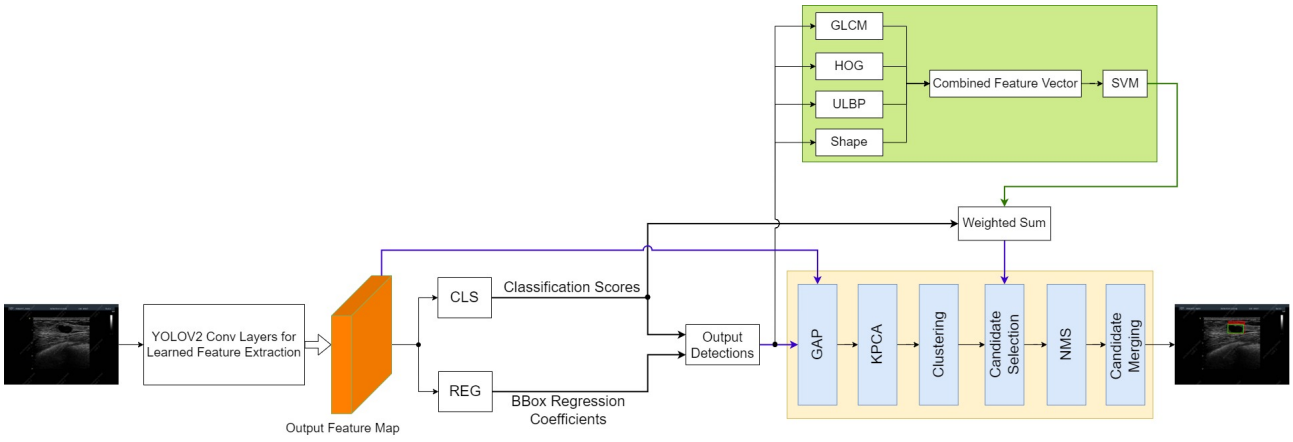


Figure 9.3: U-DetectH method in YOLOv2 detector.

Evaluation of U-Detect and U-DetectH methods for object detection in natural images: To further evaluate the effectiveness and generalisation capabilities of the U-Detect and U-DetectH

methods, an important further work is the investigation of this method for object detection in natural images. This would provide valuable understanding of the U-Detect method in this domain.

Reduction of computation time of the U-DetectH models: As discussed in Chapter 7, the high computation time of the U-DetectH models is caused due to the high computation time required for extraction of GLCM features of the test proposals. We used the inbuilt MATLAB function for GLCM features extraction. Thus, to improve the overall speed of the U-DetectH model, the GLCM feature extraction process will be improved through development of a new function for this process as part of the future work. The MATLAB function uses several checks and computation that are not relevant or required in this work. The new function will focus purely on the generation of GLCM matrices and features required for this work to reduce the overall computation time.

Automation of Network Optimization: In this research, manual optimization is performed since the aim of this work was to study the impact of the various hyperparameters involved. With the larger dataset now available, automating this optimization is an important future work. A promising direction is using ENAS (Efficient Network Architecture Search). ENAS is a growing field of research and has proven successful in multiple image processing tasks for medical images such as image classification [158] and segmentation [159]. Specifically, use of ENAS for hyperparameter optimization will be studied. Furthermore, one of the recent works in this field, NAS-FPN [160], has successfully used NAS for automatic development of an optimal FPN network that concatenates layers of the base classification network for improving the features extracted by the model and in turn the overall performance for object detection in natural images. Study of such a network will also be performed to further improve the classification and bounding box regression accuracy of the adapted FRCNN model for breast lesion detection in US images. The impact of U-Detect and U-DetectH on this adapted FRCNN model will also be studied.

Bibliography

[1] D. Pelleg, A. W. Moore, *et al.*, "X-means: Extending k-means with efficient estimation of the number of clusters.," in *Icml*, vol. 1, pp. 727–734, 2000.

[2] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[3] M. H. Yap, M. Goyal, F. Osman, R. Marti, E. Denton, A. Juette, and R. Zwiggelaar, "Breast ultrasound region of interest detection and lesion localisation," *Artificial Intelligence in Medicine*, vol. 107, p. 101880, 2020.

[4] Z. Zhang, X. Zhang, X. Lin, L. Dong, S. Zhang, X. Zhang, D. Sun, and K. Yuan, "Ultrasonic diagnosis of breast nodules using modified faster r-cnn," *Ultrasonic Imaging*, vol. 41, no. 6, pp. 353–367, 2019.

[5] M. Bellver, X. Giró-i Nieto, F. Marqués, and J. Torres, "Hierarchical object detection with deep reinforcement learning," *arXiv preprint arXiv:1611.03718*, 2016.

[6] H. Zonderland and R. Smithuis, "Bi-rads for mammography and ultrasound 2013," 2014. Available at `https://radiologyassistant.nl/breast/bi-rads/bi-rads-for-mammography-and-ultrasound-2013`, Accessed on 19.01.2022.

[7] World Health Organization, "Latest global cancer data: Cancer burden rises to 19.3 million new cases and 10.0 million cancer deaths in 2020." Available at `https://www.iarc.who.int/wp-content/uploads/2020/12/pr292_E.pdf`, Accessed on 15.04.2020.

[8] World Health Organization, "Who cancer report 2020 global profile." Available at `https://www.paho.org/hq/index.php?option=com_docman&view=download&category_`

slug=4-cancer-country-profiles-2020&alias=51561-global-cancer-profile-2020&Itemid=270&lang=en, Accessed on 15.04.2020.

[9] C. R. UK, "Breast cancer," Sep 2019. Available at https://www.cancerresearchuk.org/about-cancer/breast-cancer/about, Accessed on 20.02.2020.

[10] American Cancer Society, "What is cancer? — cancer basics," 2022. Available at https://www.cancer.org/treatment/understanding-your-diagnosis/what-is-cancer.html, Accessed on 25.05.2022.

[11] American Cancer Society, "Cancer facts and figures 2020." Available at https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf, Accessed: 15.04.2020.

[12] R. Sood, A. F. Rositch, D. Shakoor, E. Ambinder, K.-L. Pool, E. Pollack, D. J. Mollura, L. A. Mullen, and S. C. Harvey, "Ultrasound for breast cancer detection globally: A systematic review and meta-analysis," *Journal of Global Oncology*, no. 5, pp. 1–17, 2019.

[13] T. M. Kolb, J. Lichy, and J. H. Newhouse, "Comparison of the performance of screening mammography, physical examination, and breast us and evaluation of factors that influence them: an analysis of 27,825 patient evaluations," *Radiology*, vol. 225, no. 1, pp. 165–175, 2002.

[14] D. Thigpen, A. Kappler, and R. Brem, "The role of ultrasound in screening dense breasts—a review of the literature and practical solutions for implementation," *Diagnostics*, vol. 8, no. 1, p. 20, 2018.

[15] American Cancer Society, "Tests to find out if breast cancer has spread," January 2022. Available at https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/tests-to-find-out-if-breast-cancer-has-spread.html, Accessed on 19.01.2022.

[16] American College of Radiology, "Breast imaging reporting data system (bi-rads®)." Available at https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads#Ultrasound, Accessed on 16.08.2022.

[17] Cancer Research UK, "New reports show staff shortages continue to hold back cancer care." Available online at https://news.cancerresearchuk.org/2022/06/09/new-reports-show-staff-shortages-continue-to-hold-back-cancer-care/, 2022. Accessed on 19.01.2022.

[18] G. Plimmer and S. Neville, "Nhs sends x-rays abroad amid acute uk shortage of radiologists."

[19] Royal College of Radiologists (RCR), "Clinical radiology workforce census 2023." https://www.rcr.ac.uk/sites/default/files/documents/rcr_clinical_radiology_workforce_census_2023.pdf, 2023.

[20] N. Azamjah, Y. Soltan-Zadeh, and F. Zayeri, "Global trend of breast cancer mortality rate: A 25-year study," *Asian Pacific journal of cancer prevention: APJCP*, vol. 20, no. 7, p. 2015, 2019.

[21] Y. Gu, W. Xu, B. Lin, X. An, J. Tian, H. Ran, W. Ren, C. Chang, J. Yuan, C. Kang, *et al.*, "Deep learning based on ultrasound images assists breast lesion diagnosis in china: a multicenter diagnostic study," *Insights into Imaging*, vol. 13, no. 1, p. 124, 2022.

[22] 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)*, vol. 1, pp. 886–893, IEEE, 2005.

[23] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

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

[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[26] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[27] M. Pietikäinen, T. Ojala, and Z. Xu, "Rotation-invariant texture classification using feature distributions," *Pattern recognition*, vol. 33, no. 1, pp. 43–52, 2000.

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

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

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

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[32] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1492–1500, IEEE, 2017.

[33] I. Jolliffe, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[34] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[35] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Johns Hopkins University Press, 4th ed., 2012.

[36] B. Schölkopf, A. Smola, and K.-R. Muller, "Kernel principal component analysis," in *International Conference on Artificial Neural Networks (ICANN)*, pp. 583–588, Springer, 1997.

[37] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[38] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

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

[40] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD*, vol. 96, no. 34, pp. 226–231, 1996.

[41] J. MacQueen, "Classification and analysis of multivariate observations," in *5th Berkeley Symp. Math. Statist. Probability*, pp. 281–297, 1967.

[42] S. Vassilvitskii and D. Arthur, "k-means++: The advantages of careful seeding," 2006.

[43] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

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

[45] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

[46] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[47] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[48] B. A. . L. H. Y. M. Wang, C. Y., "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022.

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

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

[51] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[52] G. G. D. P. He, K. and R. Girshick, "Mask r-cnn," 2017.

[53] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR)*, pp. 850–855, IEEE, 2006.

[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

[55] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision (IJCV)*, vol. 88, no. 2, pp. 303–338, 2010.

[56] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

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

[58] J. C. Caicedo and S. Lazebnik, "Active object localization with deep reinforcement learning," in *Proceedings of the IEEE international conference on computer vision*, pp. 2488–2496, 2015.

[59] Z. Jie, X. Liang, J. Feng, X. Jin, W. Lu, and S. Yan, "Tree-structured reinforcement learning for sequential object localization," in *Advances in Neural Information Processing Systems*, pp. 127–135, 2016.

[60] J. König, S. Malberg, M. Martens, S. Niehaus, A. Krohn-Grimberghe, and A. Ramaswamy, "Multi-stage reinforcement learning for object detection," in *Science and Information Conference*, pp. 178–191, Springer, 2019.

[61] A. Pirinen and C. Sminchisescu, "Deep reinforcement learning of region proposal networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6945–6954, 2018.

[62] B. Uzkent, C. Yeh, and S. Ermon, "Efficient object detection in large images using deep reinforcement learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1824–1833, 2020.

[63] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.

[64] M. Tan and R. Chen, "Efficientdet: Scalable and efficient object detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10781–10790, 2020.

[65] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6105–6114, 2019.

[66] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[67] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[68] E. H. Adelson and C. H. Anderson, "The pyramid as a structure for efficient computation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 504–507, IEEE, 1984.

[69] S. Gidaris and N. Komodakis, "Attend refine repeat: Active box proposal generation via in-out localization," *arXiv preprint arXiv:1606.04446*, 2016.

[70] T. Vu, H. Jang, T. X. Pham, and C. D. Yoo, "Cascade rpn: Delving into high-quality region proposal network with adaptive convolution," *arXiv preprint arXiv:1909.06720*, 2019.

[71] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 764–773, 2017.

[72] Y. Zhong, J. Wang, J. Peng, and L. Zhang, "Anchor box optimization for object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1286–1294, 2020.

[73] X. Zhang, F. Wan, C. Liu, X. Ji, and Q. Ye, "Learning to match anchors for visual object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[74] W. Ke, T. Zhang, Z. Huang, Q. Ye, J. Liu, and D. Huang, "Multiple anchor learning for visual object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10206–10215, 2020.

[75] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2965–2974, 2019.

[76] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9657–9666, 2019.

[77] Z. Yang, Y. Xu, H. Xue, Z. Zhang, R. Urtasun, L. Wang, S. Lin, and H. Hu, "Dense reppoints: Representing visual objects with dense point sets," *arXiv preprint arXiv:1912.11473*, vol. 2, 2019.

[78] Y. Chen, Z. Zhang, Y. Cao, L. Wang, S. Lin, and H. Hu, "Reppoints v2: Verification meets regression for object detection," *arXiv preprint arXiv:2007.08508*, 2020.

[79] J. Cao, H. Cholakkal, R. M. Anwer, F. S. Khan, Y. Pang, and L. Shao, "D2det: Towards high quality object detection and instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11485–11494, 2020.

[80] S. Lan, Z. Ren, Y. Wu, L. S. Davis, and G. Hua, "Saccadenet: A fast and accurate object detector," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10397–10406, 2020.

[81] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "Foveabox: Beyound anchor-based object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 7389–7398, 2020.

[82] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, pp. 5561–5569, 2017.

[83] L. Cai, B. Zhao, Z. Wang, J. Lin, C. S. Foo, M. S. Aly, and V. Chandrasekhar, "Maxpoolnms: getting rid of nms bottlenecks in two-stage object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9356–9364, 2019.

[84] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12993–13000, 2020.

[85] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2019.

[86] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016.

[87] M. Li, Z. Zhang, H. Yu, X. Chen, and D. Li, "S-ohem: stratified online hard example mining for object detection," in *CCF Chinese Conference on Computer Vision*, pp. 166–177, Springer, 2017.

[88] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 821–830, 2019.

[89] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[90] Y. Cao, K. Chen, C. C. Loy, and D. Lin, "Prime sample attention in object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11583–11591, 2020.

[91] Q. Qian, L. Chen, H. Li, and R. Jin, "Dr loss: Improving object detection by distributional ranking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12164–12172, 2020.

[92] K. Chen, W. Lin, J. See, J. Wang, J. Zou, *et al.*, "Ap-loss for accurate one-stage object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[93] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4203–4212, 2018.

[94] J. Nie, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Enriched feature guided refinement network for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9537–9546, 2019.

[95] H. Behnam, F. S. Zakeri, and N. Ahmadinejad, "Breast mass classification on sonographic images on the basis of shape analysis," *Journal of Medical Ultrasonics*, vol. 37, no. 4, pp. 181–186, 2010.

[96] R.-F. Chang, W.-J. Wu, W. K. Moon, and D.-R. Chen, "Automatic ultrasound segmentation and morphology based diagnosis of solid breast tumors," *Breast cancer research and treatment*, vol. 89, no. 2, pp. 179–185, 2005.

[97] C. M. Sehgal, T. W. Cary, S. A. Kangas, S. P. Weinstein, S. M. Schultz, P. H. Arger, and E. F. Conant, "Computer-based margin analysis of breast sonography for differentiating malignant and benign masses," *Journal of ultrasound in medicine*, vol. 23, no. 9, pp. 1201–1209, 2004.

[98] W. Gómez, W. C. A. Pereira, and A. F. C. Infantosi, "Analysis of co-occurrence texture statistics as a function of gray-level quantization for classifying breast ultrasound," *IEEE transactions on medical imaging*, vol. 31, no. 10, pp. 1889–1899, 2012.

[99] D.-R. Chen, R.-F. Chang, Y.-L. Huang, Y.-H. Chou, C.-M. Tiu, and P.-P. Tsai, "Texture analysis of breast tumors on sonograms," in *Seminars in Ultrasound, CT and MRI*, vol. 21, pp. 308–316, Elsevier, 2000.

[100] M. Abdel-Nasser, J. Melendez, A. Moreno, O. A. Omer, and D. Puig, "Breast tumor classification in ultrasound images using texture analysis and super-resolution methods," *Engineering Applications of Artificial Intelligence*, vol. 59, pp. 84–92, 2017.

[101] M. Wei, X. Wu, J. Zhu, P. Liu, Y. Luo, L. Zheng, and Y. Du, "Multi-feature fusion for ultrasound breast image classification of benign and malignant," in *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, pp. 474–478, IEEE, 2019.

[102] M. Benaouali, M. Bentoumi, M. Touati, A. T. Ahmed, and M. Mimi, "Segmentation and classification of benign and malignant breast tumors via texture characterization from ultrasound images," in *2022 7th International Conference on Image and Signal Processing and their Applications (ISPA)*, pp. 1–4, IEEE, 2022.

[103] M. Rahmawaty, H. A. Nugroho, Y. Triyani, I. Ardiyanto, and I. Soesanti, "Classification of breast ultrasound images based on texture analysis," in *2016 1st International Conference on Biomedical Engineering (IBIOMED)*, pp. 1–6, IEEE, 2016.

[104] R. Sivaramakrishna, K. A. Powell, M. L. Lieber, W. A. Chilcote, and R. Shekhar, "Texture analysis of lesions in breast ultrasound images," *Computerized medical imaging and graphics*, vol. 26, no. 5, pp. 303–307, 2002.

[105] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, "Trainable model based on new uniform lbp feature to identify the risk of the breast cancer," in *2019 international conference on advanced science and engineering (ICOASE)*, pp. 106–111, IEEE, 2019.

[106] A. V. Alvarenga, W. C. Pereira, A. F. C. Infantosi, and C. M. Azevedo, "Complexity curve and grey level co-occurrence matrix in the texture evaluation of breast tumor on ultrasound images," *Medical physics*, vol. 34, no. 2, pp. 379–387, 2007.

[107] W.-J. Wu and W. K. Moon, "Ultrasound breast tumor image computer-aided diagnosis with texture and morphological features," *Academic radiology*, vol. 15, no. 7, pp. 873–880, 2008.

[108] M. Wei, Y. Du, X. Wu, Q. Su, J. Zhu, L. Zheng, G. Lv, and J. Zhuang, "A benign and malignant breast tumor classification method via efficiently combining texture and morphological features on ultrasound images," *Computational and Mathematical Methods in Medicine*, vol. 2020, 2020.

[109] M. Wei, Y. Du, X. Wu, and J. Zhu, "Automatic classification of benign and malignant breast tumors in ultrasound image with texture and morphological features," in *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, pp. 126–130, IEEE, 2019.

[110] K. M. Prabusankarlal, P. Thirumoorthy, and R. Manavalan, "Assessment of combined textural and morphological features for diagnosis of breast masses in ultrasound," *Human-centric Computing and Information Sciences*, vol. 5, no. 1, pp. 1–17, 2015.

[111] M. H. Yap, E. A. Edirisinghe, and H. E. Bez, "A comparative study in ultrasound breast imaging classification," in *Medical Imaging 2009: Image Processing*, vol. 7259, pp. 598–608, SPIE, 2009.

[112] H. Nemat, H. Fehri, N. Ahmadinejad, A. F. Frangi, and A. Gooya, "Classification of breast lesions in ultrasonography using sparse logistic regression and morphology-based texture features," *Medical physics*, vol. 45, no. 9, pp. 4112–4124, 2018.

[113] K. Horsch, M. L. Giger, L. A. Venta, and C. J. Vyborny, "Computerized diagnosis of breast lesions on ultrasound," *Medical physics*, vol. 29, no. 2, pp. 157–164, 2002.

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

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

[116] Y.-L. Huang, S.-H. Lin, and D.-R. Chen, "Computer-aided diagnosis applied to 3-d us of solid breast nodules by using principal component analysis and image retrieval," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pp. 1802–1805, IEEE, 2006.

[117] Y. D. Chun, S. Y. Seo, and N. C. Kim, "Image retrieval using bdip and bvlc moments," *IEEE transactions on circuits and systems for video technology*, vol. 13, no. 9, pp. 951–957, 2003.

[118] B. S. Garra, B. H. Krasner, S. C. Horii, S. Ascher, S. K. Mun, and R. K. Zeman, "Improving the distinction between benign and malignant breast lesions: the value of sonographic texture analysis," *Ultrasonic imaging*, vol. 15, no. 4, pp. 267–285, 1993.

[119] L. Cai, X. Wang, Y. Wang, Y. Guo, J. Yu, and Y. Wang, "Robust phase-based texture descriptor for classification of breast ultrasound images," *Biomedical engineering online*, vol. 14, no. 1, pp. 1–21, 2015.

[120] D. Q. Zeebaree, A. M. Abdulazeez, D. A. Zebari, H. Haron, and H. N. A. Hamed, "Multi-level fusion in ultrasound for cancer detection based on uniform lbp features," *Computers, Materials & Continua*, vol. 66, no. 3, pp. 3363–3382, 2021.

[121] Y.-L. Huang, K.-L. Wang, and D.-R. Chen, "Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines," *Neural Computing & Applications*, vol. 15, no. 2, pp. 164–169, 2006.

[122] B. Huynh, K. Drukker, and M. Giger, "Mo-de-207b-06: Computer-aided diagnosis of breast ultrasound images using transfer learning from deep convolutional neural networks," *Medical physics*, vol. 43, no. 6Part30, pp. 3705–3705, 2016.

[123] 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)*, pp. 1–4, IEEE, 2019.

[124] W. Al-Dhabyani, M. Gomaa, H. Khaled, and F. Aly, "Deep learning approaches for data augmentation and classification of breast masses using ultrasound images," *Int. J. Adv. Comput. Sci. Appl*, vol. 10, no. 5, pp. 1–11, 2019.

[125] S. Han, H.-K. Kang, J.-Y. Jeong, M.-H. Park, W. Kim, W.-C. Bang, and Y.-K. Seong, "A deep learning framework for supporting the classification of breast lesions in ultrasound images," *Physics in Medicine & Biology*, vol. 62, no. 19, p. 7714, 2017.

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

[127] K. K. Bressem, L. C. Adams, C. Erxleben, B. Hamm, S. M. Niehues, and J. L. Vahldiek, "Comparing different deep learning architectures for classification of chest radiographs," *Scientific reports*, vol. 10, no. 1, pp. 1–16, 2020.

[128] J. Peng, C. Bao, C. Hu, X. Wang, W. Jian, and W. Liu, "Automated mammographic mass detection using deformable convolution and multiscale features," *Medical & biological engineering & computing*, vol. 58, no. 7, pp. 1405–1417, 2020.

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

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

[131] 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 medical imaging*, vol. 19, no. 1, pp. 1–9, 2019.

[132] M. H. Yap, M. Goyal, F. Osman, R. Martí, E. Denton, A. Juette, and R. Zwiggelaar, "Breast ultrasound region of interest detection and lesion localisation," *Artificial Intelligence in Medicine*, vol. 107, p. 101880, 2020.

[133] Y. Huang, L. Han, H. Dou, H. Luo, Z. Yuan, Q. Liu, J. Zhang, and G. Yin, "Two-stage cnns for computerized bi-rads categorization in breast ultrasound images," *Biomedical engineering online*, vol. 18, no. 1, pp. 1–18, 2019.

[134] C. Tao, K. Chen, L. Han, Y. Peng, C. Li, Z. Hua, and J. Lin, "New one-step model of breast tumor locating based on deep learning," *Journal of X-ray Science and Technology*, vol. 27, no. 5, pp. 839–856, 2019.

[135] S. Y. Shin, S. Lee, I. D. Yun, S. M. Kim, and K. M. Lee, "Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images," *IEEE transactions on medical imaging*, vol. 38, no. 3, pp. 762–774, 2018.

[136] Y. Gao, B. Liu, Y. Zhu, L. Chen, M. Tan, X. Xiao, G. Yu, and Y. Guo, "Detection and recognition of ultrasound breast nodules based on semi-supervised deep learning: a powerful alternative strategy," *Quantitative Imaging in Medicine and Surgery*, vol. 11, no. 6, p. 2265, 2021.

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

[138] V. Punithavathi and D. Devakumari, "Detection of breast lesion using improved glcm feature based extraction in mammogram images," *published by SSRN eLibrary*, 2020.

[139] Y.-C. Zhu, A. AlZoubi, S. Jassim, Q. Jiang, Y. Zhang, Y.-B. Wang, X.-D. Ye, and D. Hongbo, "A generic deep learning framework to classify thyroid and breast lesions in ultrasound images," *Ultrasonics*, vol. 110, p. 106300, 2021.

[140] A. Ciritsis, C. Rossi, M. Eberhard, M. Marcon, A. S. Becker, and A. Boss, "Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making," *European radiology*, vol. 29, no. 10, pp. 5458–5468, 2019.

[141] Z. Zhuang, Y. Kang, A. N. Joseph Raj, Y. Yuan, W. Ding, and S. Qiu, "Breast ultrasound lesion classification based on image decomposition and transfer learning," *Medical Physics*, vol. 47, no. 12, pp. 6257–6269, 2020.

[142] Z. Zhuang, Z. Yang, S. Zhuang, A. N. Joseph Raj, Y. Yuan, and R. Nersisson, "Multi-features-based automated breast tumor diagnosis using ultrasound image and support vector machine," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[143] T. Liu, Q. Guo, C. Lian, X. Ren, S. Liang, J. Yu, L. Niu, W. Sun, and D. Shen, "Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks," *Medical image analysis*, vol. 58, p. 101555, 2019.

[144] R. Igarashi, K. Tomita, Y. Nishiyama, Y. Shigenari, and N. Koizumi, "Lesion tracking method using cnn for non-invasive ultrasound theranostic system," in *2019 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, pp. 228–234, IEEE, 2019.

[145] H. Li, J. Weng, Y. Shi, W. Gu, Y. Mao, Y. Wang, W. Liu, and J. Zhang, "An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[146] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Scientific reports*, vol. 8, no. 1, pp. 1–7, 2018.

[147] A. Akselrod-Ballin, L. Karlinsky, S. Alpert, S. Hasoul, R. Ben-Ari, and E. Barkan, "A region based convolutional network for tumor detection and classification in breast mammography," in *Deep learning and data labeling for medical applications*, pp. 197–205, Springer, 2016.

[148] J. An, H. Yu, R. Bai, J. Li, Y. Wang, and R. Cao, "Detection and segmentation of breast masses based on multi-layer feature fusion," *Methods*, vol. 202, pp. 54–61, 2022.

[149] F. Abdolali, J. Kapur, J. L. Jaremko, M. Noga, A. R. Hareendranathan, and K. Punithakumar, "Automated thyroid nodule detection from ultrasound imaging using deep convolutional neural networks," *Computers in Biology and Medicine*, vol. 122, p. 103871, 2020.

[150] S. Xie, J. Yu, T. Liu, Q. Chang, L. Niu, and W. Sun, "Thyroid nodule detection in ultrasound images with convolutional neural networks," in *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1442–1446, IEEE, 2019.

[151] Y. Wu and Z. Yi, "Automated detection of kidney abnormalities using multi-feature fusion convolutional neural networks," *Knowledge-Based Systems*, vol. 200, p. 105873, 2020.

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

[153] A. Bose, T. Nguyen, H. Du, and A. AlZoubi, "Faster rcnn hyperparameter selection for breast lesion detection in 2d ultrasound images," in *Advances in Computational Intelligence Systems: Contributions Presented at the 20th UK Workshop on Computational Intelligence, September 8-10, 2021, Aberystwyth, Wales, UK 20*, pp. 179–190, Springer, 2022.

[154] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," tech. rep., Stanford, 2006.

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

[156] F. D. F. A. Hancock, B. and R. Yozzo, "Notes on bayesian information criterion for x-means clustering," 2022. Available at `https://github.com/bobhancock/goxmeans/blob/master/doc/BIC_notes.pdf`, Accessed on 25.05.2021.

[157] Z. Zhao, R. Anand, and M. Wang, "Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform," in *2019 IEEE international conference on data science and advanced analytics (DSAA)*, pp. 442–452, IEEE, 2019.

[158] M. Ahmed, H. Du, and A. AlZoubi, "An enas based approach for constructing deep learning models for breast cancer recognition from ultrasound images," *arXiv preprint arXiv:2005.13695*, 2020.

[159] N. Gessert and A. Schlaefer, "Efficient neural architecture search on low-dimensional data for oct image segmentation," *arXiv preprint arXiv:1905.02590*, 2019.

[160] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7036–7045, IEEE, 2019.

Appendices

Appendix A

FRCNN Modelling Hyperparameter Investigation

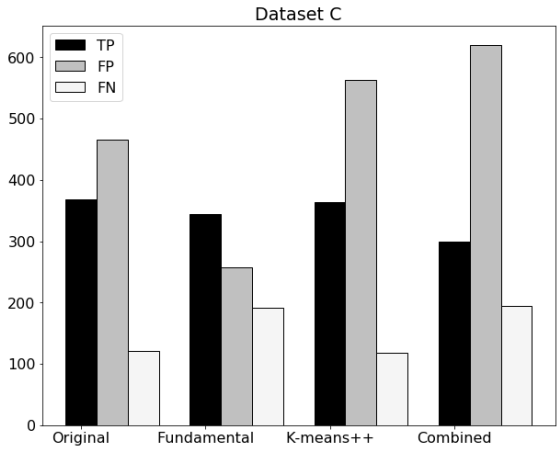


Figure A.1: Number of TP, FP and FN detection in FRCNN models trained with all anchor boxes in dataset C.

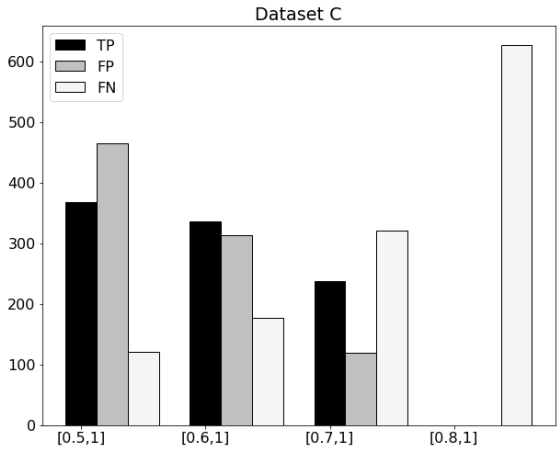


Figure A.2: Change in number of TP, FP and FN with variation in base network's positive training samples in dataset C.

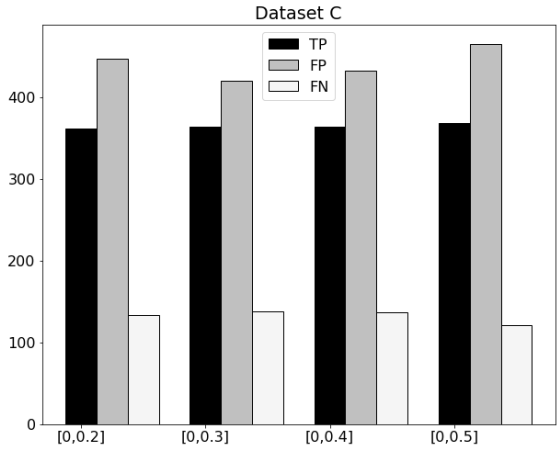


Figure A.3: Change in number of TP, FP and FN with variation in base network's negative training samples in dataset C.

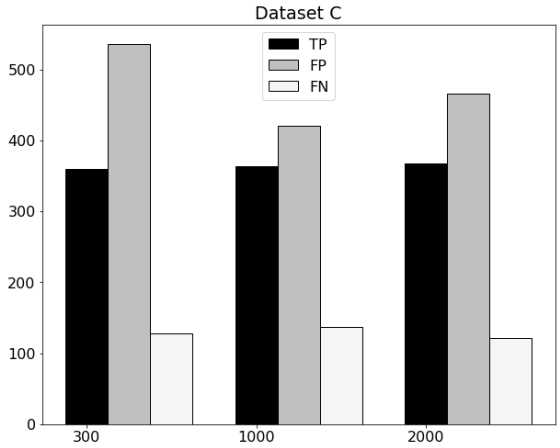


Figure A.4: Number of number of TP, FP and FN with variations in number of training proposals in dataset C.

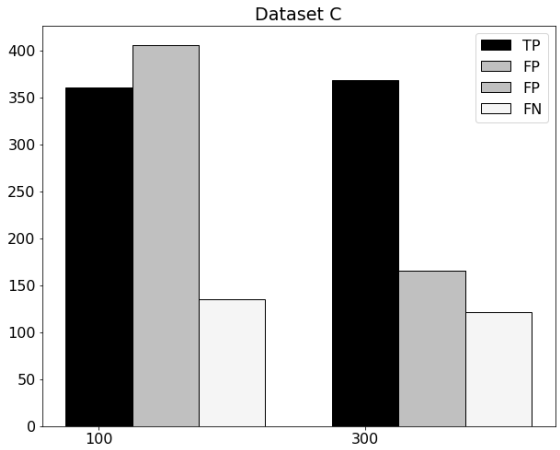


Figure A.5: FRCNN performance with variations in number of test proposals in dataset C.

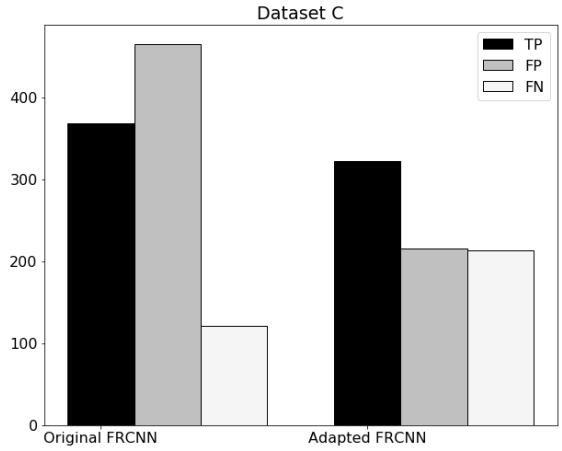


Figure A.6: Original anchor boxes with default and optimal modelling hyperparameters: Dataset C.

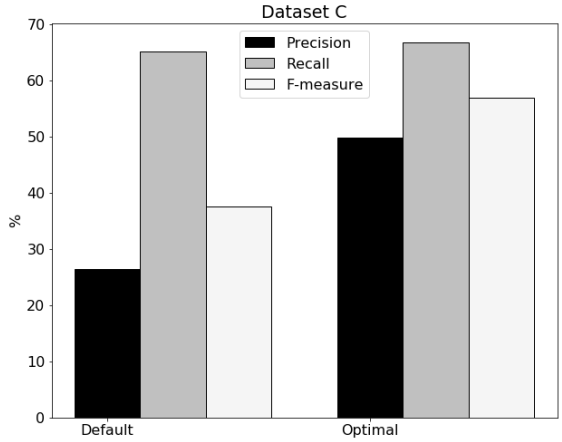


Figure A.7: K-means++ anchor boxes with default and optimal modelling hyperparameters: Dataset C

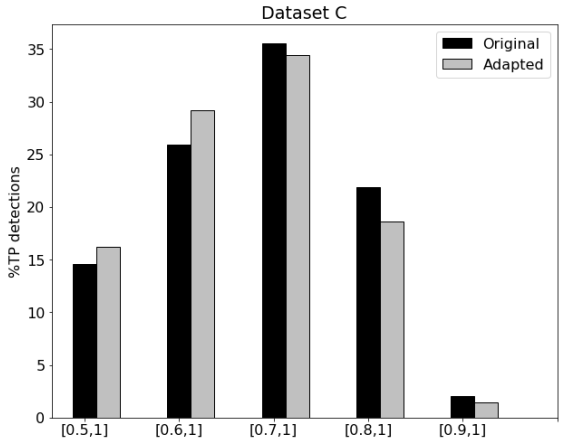


Figure A.8: IOU distribution of TPs in original and adapted FRCNN models: Dataset C.

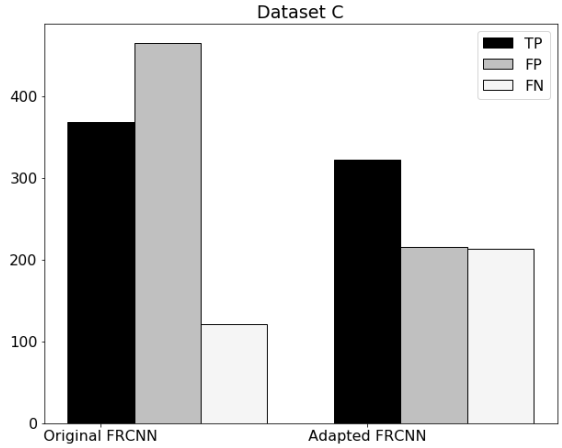


Figure A.9: Number of TP, FP and FN detections in original and adapted FRCNN models in dataset C.

Appendix B

U-Detect models

B.1 U-Detect-RPN

Table B.1 shows single-fold performance of adapted FRCNN and U-Detect-RPN. Compared to adapted FRCNN, the U-Detect-RPN model has 3.79% to 13.54% higher precision with only a small drop of 0.3% in recall in modelling dataset and 0.72% higher recall in overall unseen test sets. Thus, the overall F-measure is 2.02% to 8.94% higher. The improvement in precision and recall is due to improved filtering of proposals which restricted overshadowing of TP proposals as well as improved filtering of FP proposals. Therefore, the proposed U-Detect-RPN method successfully overcomes drawbacks of NMS.

Dataset	**Model**	**Precision**	**Recall**	**F-measure**
A	Adapted	85.23	**99.34**	91.74
	U-Detect-RPN	**89.02**	99.04	**93.76**
Overall External Test Sets	Adapted	62.88	90.17	74.09
	U-Detect-RPN	**76.42**	**90.89**	**83.03**

Table B.1: Single fold performance of U-Detect-RPN and adapted FRCNN model

B.2 U-Detect-Base

Table B.2 shows single fold performance of U-Detect-Base model in comparison to adapted FRCNN. U-Detect-Base model outperforms adapted FRCNN with 0.94% to 12.31% higher precision and only a small drop of 0.33% to 0.51% in recall. Therefore, using the clustering network to filter the final

detections of the base network is effective in addressing drawbacks of NMS while causing minimal negative impact on the number of correct detections.

Dataset	**Model**	**Precision**	**Recall**	**F-measure**
A	Adapted	85.23	**99.34**	91.74
	U-Detect-Base	**86.17**	99.01	**92.14**
Overall External	Adapted	62.88	**90.17**	74.09
	U-Detect-Base	**75.19**	89.66	**81.79**

Table B.2: Single fold performance of U-Detect-Base and adapted FRCNN

Appendix C

U-DetectH-models

C.1 SVM Training Sets Evaluation

Dataset	Accuracy	Precision	Recall	F-measure	Specificity
A	90.48	3.24	78.03	6.23	90.53
B (benign)	93.21	2.91	82.86	5.63	93.24
B (malignant)	85.52	2.29	82.76	4.46	85.53
C	47.46	0.55	90.47	1.09	47.32
D	87.16	2.68	85.17	5.21	87.71
E	92.22	4.31	74.77	8.16	92.31

Table C.1: Classification accuracy of GLCM-SVM model trained training set described in Section 7.2.1 in Chapter 7.

Dataset	Accuracy	Precision	Recall	F-measure	Specificity
A	95.13	6.58	83.59	12.20	95.17
B (benign)	95.68	5.07	90.00	9.60	95.87
B (malignant)	95.27	5.11	74.14	9.56	94.35
C	87.67	1.88	73.09	3.66	87.71
D	95.27	6.30	75.12	11.63	95.36
E	95.28	7.05	75.69	12.90	95.37

Table C.2: Classification accuracy of HOG-SVM model trained training set described in Section 7.2.1 in Chapter 7.

Dataset	Accuracy	Precision	Recall	F-measure	Specificity
A	90.90	3.06	69.95	5.86	90.98
B (benign)	93.33	1.98	54.29	3.82	93.42
B (malignant)	86.13	2.06	70.69	3.99	86.19
C	63.03	0.42	84.75	0.84	35.88
D	90.34	2.06	70.69	3.99	86.19
E	92.00	3.06	53.21	5.79	92.18

Table C.3: Classification accuracy of ULBP-SVM model trained training set described in Section 7.2.1 in Chapter 7.

Dataset	Accuracy	Precision	Recall	F-measure	Specificity
A	43.18	0.53	73.99	1.04	43.05
B (benign)	41.75	0.31	72.86	0.61	41.68
B (malignant)	40.99	0.61	87.93	1.20	40.80
C	60.01	0.66	82.63	1.31	59.93
D	44.48	0.67	90.43	1.33	40.80
E	44.19	0.75	91.74	1.50	43.97

Table C.4: Classification accuracy of combined-PCA-SVM model trained training set described in Section 7.2.1 in Chapter 7.

Dataset	Accuracy	Precision	Recall	F-measure	Specificity
A	77.28	1.29	72.73	2.53	77.30
B (benign)	80.97	0.70	54.29	1.38	81.03
B (malignant)	72.21	1.16	79.31	2.28	72.18
C	62.88	0.57	66.74	1.14	62.87
D	69.54	1.11	82.14	2.18	69.49
E	71.73	1.37	82.14	2.18	69.49

Table C.5: Classification accuracy of ULBP-M-SVM model trained training set described in Section 7.2.1 in Chapter 7.

Dataset	Accuracy	Precision	Recall	F-measure	Specificity
A	95.26	6.72	83.08	12.43	95.31
B (benign)	95.48	4.66	90.00	8.87	95.31
B (malignant)	93.47	4.69	77.59	8.84	93.53
C	88.38	1.96	72.03	3.82	88.43
D	95.70	7.19	78.79	13.18	95.77
E	94.87	6.68	77.98	12.31	94.95

Table C.6: Classification accuracy of (HOG+GLCM)-SVM model trained training set described in Section 7.2.1 in Chapter 7.

Dataset	Accuracy	Precision	Recall	F-measure	Specificity
A	97.41	8.63	56.31	14.97	97.58
B (benign)	97.09	6.29	78.57	11.65	97.13
B (malignant)	97.61	4.81	25.86	8.11	97.90
C	85.72	1.08	48.09	2.11	85.84
D	97.89	7.94	38.60	13.17	98.14
E	98.47	10.50	30.73	15.65	98.79

Table C.7: Classification accuracy of combined-SVM model trained training set TS1 described in Section 8.3 in Chapter 8.

Dataset	Accuracy	Precision	Recall	F-measure	Specificity
A	98.96	13.10	27.78	17.80	99.25
B (benign)	99.20	15.52	51.43	23.84	99.31
B (malignant)	98.76	1.64	3.45	2.22	99.15
C	98.41	3.54	15.04	5.73	98.68
D	97.35	1.99	11.16	3.37	97.71
E	99.21	3.01	2.29	2.60	99.66

Table C.8: Classification accuracy of combined-SVM model trained training set TS2 described in Section 8.3 in Chapter 8.

Dataset	Accuracy	Precision	Recall	F-measure	Specificity
A	92.81	4.07	74.24	7.71	92.88
B (benign)	94.62	3.36	75.71	6.43	94.66
B (malignant)	89.84	2.99	75.86	5.74	89.89
C	57.01	0.6	81.14	1.19	56.93
D	90.68	3.46	80.06	6.64	90.73
E	95.01	5.68	62.84	10.42	95.16

Table C.9: Classification accuracy of GLCM-new-SVM model trained training set TS3 described in Section 8.3 in Chapter 8.

Dataset	Accuracy	Precision	Recall	F-measure	Specificity
A	17.93	0.49	100.00	0.98	17.60
B (benign)	94.14	1.09	25.71	2.10	94.31
B (malignant)	72.40	0.54	36.21	1.06	72.55
C	2.07	0.27	82.63	0.54	1.81
D	50.26	0.79	95.22	1.56	50.08
E	5.30	0.48	99.54	0.96	4.86

Table C.10: Classification accuracy of combined-MRMR-SVM model trained training set described in Section 7.2.1 in Chapter 7.

C.2 U-DetectH-Base Performance

Table C.11 shows the performance of the single-feature based models on a single fold of modelling dataset. These models perform in a similar fashion in this fold. As with the 5-fold performance, combined-SVM outperforms all these models.

Dataset	U-DetectH-Base Model	Precision	Recall	F-measure
A	GLCM-SVM	**88.73**	**99.03**	**93.60**
	ULBP-SVM	86.71	99.01	92.45
	HOG-SVM	85.55	99.00	91.78
Overall External Test Sets	GLCM-SVM	76.01	**89.76**	82.31
	ULBP-SVM	73.34	89.14	80.47
	HOG-SVM	**76.18**	89.68	**82.38**

Table C.11: Performance of U-DetectH-Base using single feature-based SVM models: Single Fold.

Table C.12 shows single-fold performance of the combined-SVM model in U-DetectH-Base. Here, too, the combined-SVM model has the highest F-measure overall. The combined-SVM model has slightly lower F-measure than the U-Detect-Base model in the modelling dataset. But, in overall external test sets, combined-SVM model has 2.41% higher precision along with 0.14% higher recall than the U-Detect-Base model. This improvement in both precision and recall is due to an increase in the number of correct detections and reduction in FPs.

Dataset	Model	Precision	Recall	F-measure
A	Adapted FRCNN	85.23	**99.34**	91.74
	U-Detect-Base	**86.17**	99.01	**92.14**
	Combined-SVM	84.97	98.99	91.45
Overall External Test Sets	Adapted FRCNN	62.88	**90.17**	74.09
	U-Detect-Base	75.19	89.66	81.79
	Combined-SVM	**77.60**	89.80	**83.25**

Table C.12: Performance of U-DetectH-Base using combined-SVM model: Single Fold.