Applying Deep Learning and eXplainable AI (XAI) Techniques to Medical Imaging

By

Tan Keng Ze

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONOURS)

Faculty of Information and Communication Technology

(Kampar Campus)

FEBRUARY 2026

COPYRIGHT STATEMENT

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my supervisors, Dr Lee Wai Kong who has given me this bright opportunity to engage in eXplainable Artificial Intelligence (XAI) field. A million thanks to you.

Besides, I must say thanks to my parents and my family for their love, emotionally and physically support, and continuous encouragement throughout the course.

ABSTRACT

Incorporation of AI in medical diagnosis continues to evoke apprehensions due to the black-box nature of deep learning models. Clinicians get a prediction result without comprehending which regions of the image influenced the model's decision. The lack of model prediction interpretability constrains trust and practical implementation in clinical environments. This project aims to create a breast ultrasound lesion detection system based on Faster R-CNN with ResNet50-FPN utilizing the BUSI dataset. Contrary to pure image classification, the system concentrates on object detection by predicting the locations of bounding-boxes and class labels for benign, malignant, and normal cases. Segmentation masks from the dataset are transformed into bounding boxes to facilitate detection training. Grad-CAM visualisation is employed to assess the alignment of model attention with pertinent lesion areas. Furthermore, various masked filter pruning strategies are adapted into Faster R-CNN to evaluate the filter-importance, including random pruning, activation-only pruning, Taylor-only pruning, hybrid activation-Taylor pruning, and loss-guided Grad-CAM-style relevance pruning. Experimental results show that the baseline Faster R-CNN model attained satisfactory detection performance, with mAP@0.50 ranging from 0.70 to 0.74 and mAP@0.75 between 0.53 and 0.64 across multiple runs. Taylor-only and hybrid activation-Taylor pruning demonstrated competitive performance in specific contexts. However, the pruning outcomes should be regarded as a filter-importance ablation rather than significant model compression, as the pruning was masked and implemented solely on one designated layer. The parameter reduction was minimal, ranging from 0.25% to 1.78%, and the enhancement in inference speed was constrained. Grad-CAM visualisation offered qualitative explanatory support, although the heatmaps were not consistently well-aligned with lesion areas. A cloud-based deployment prototype was developed to showcase model accessibility via a web platform. This project introduces a development-oriented breast ultrasound detection pipeline that integrates localisation, explainability, pruning analysis, and deployment viability.

Area of Study (Minimum 1 and Maximum 2): Deep Learning, eXplainable AI (XAI)
Keywords (Minimum 5 and Maximum 10): Faster R-CNN, Visualisation, Web Application, Grad-CAM, ResNet50, Activation-based pruning, Region Proposal Network (FPN)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AFPN	Asymptotic Feature Pyramid Network
AI	Artificial Intelligence
AP	Average Precision
AUCROC	Area Under the Receiver Operating Characteristic Curve
BUSI	Breast Ultrasound Images
CAM	Class Activation Map
CNN	Convolutional Neural Network
COCO	Common Objects in Context
CPU	Central Processing Unit
CT	Computed Tomography
DL	Deep Learning
DR	Diabetic Retinopathy
F1	F1 Score
FPN	Feature Pyramid Network
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
IG	Integrated Gradients
IoU	Intersection over Union
LIME	Local Interpretable Model-agnostic Explanations
LRP	Layer-wise Relevance Propagation
mAP	mean Average Precision
MDA	Medical Device Authority
MNIST	Modified National Institute of Standards and Technology
MRI	Magnetic Resonance Imaging
PIL	Python Imaging Library
PVT	Pyramid Vision Transformer
R-CNN	Region-based Convolutional Neural Network
ROAD	Remove and Debias
ROI / RoI	Region of Interest
RPN	Region Proposal Network
SMOTE	Synthetic Minority Over-sampling Technique

SOTA	State of the Art
ViT	Vision Transformer
XAI	Explainable Artificial Intelligence

Chapter 1
Introduction

Problem statement of this project, project scope, project objectives, project significance and background information will be written in this chapter.

1.1 Problem Statement and Motivation

In medical image analysis tasks, such as disease diagnosis and anomaly classification, deep learning models like CNN and ViT have proven to be highly accurate. Nevertheless, its black-box character continues to be a significant obstacle to clinical use despite these achievements. Clinicians often cannot understand or verify how these models generate predictions, since most outputs are presented as result without interpretable reasoning. Although XAI methods have been proposed to address this issue, typically by generating heat maps that highlight influential image regions, many of these techniques remain underexplored in practical healthcare settings. Aside from the trust gap, modern deep learning models also suffer from high computational costs, even though lightweight models can often achieve comparable prediction accuracy. Large-weight models may become impractical for deployment because they consume measurable computational resources, despite offering high accuracy, especially when smaller models can deliver similar performance.

In addition to technical limitations, there is also a practical adoption gap. Many clinics underutilize online platforms that could streamline clinical workflows and provide easier access to AI-enabled tools. Beyond regulatory concerns, adoption is further hindered by the high cost of data infrastructure, the need to hire AI specialists to develop and maintain custom models, and the lack of expertise among clinicians to interpret predictions. These challenges reduce the accessibility and trustworthiness of AI-based diagnostic systems, limiting their potential impact in real-world medical practices.

This project is driven by the need to make AI in healthcare reliable, understandable, and accessible. In real world settings, the healthcare sector, especially for life-threatening conditions such as cancer or sepsis, the consequences of misdiagnosis can be severe. Therefore, it is crucial that AI diagnostic models are not only accurate but also reliable and trustworthy to support safe clinical decision-making. Beyond the need for technical transparency, computationally expensive deep learning models present major challenges for deployment. Their high computational power requirements make them costly to operate and maintain. As a result, lightweight models that offer comparable prediction accuracy provide a more practical and cost-effective alternative for real-world deployment. Other than that, the project is motivated by the need to reduce barriers to adoption in real-world clinical environments. Many clinics lack the financial and technical resources to build and maintain in-house AI infrastructure or employ specialised engineers to build an accurate prediction model and clinically relevant explanations model. Therefore, a prediction model with clinically relevant explanations model will be created and deployed to the cloud service.

1.2 Objectives

The main objective of the project is to create, assess, and implement a pruning method on CNN model on an online platform in order to solve both the black-box element of deep learning models and the underutilization of online platforms in clinical workflows. To following sub-objectives are established in order to accomplish it:

1. To develop a Faster R-CNN based breast ultrasound lesion detection model.
2. To address the CNN black-box nature through XAI visualisation technique.
3. To implement and compare various filter-importance-based pruning strategies effectiveness based on existing and proposed pruning technique.
4. To deploy the model on a cloud-based web platform to demonstrate the accessibility and user-friendly usage for clinicians.

To ensure feasibility, this project is not intended for clinical deployment or direct hospital integration. Real-world clinical use would require approval from the Medical Device Authority (MDA), as well as adherence to strict regulations concerning patient data privacy, model trustworthiness, liability for misdiagnosis, and standard operating procedures (SOPs). These regulatory and legal considerations fall beyond the scope of this experimental project. Instead, this work will focus on developing and demonstrating a research prototype.

1.3 Project Scope and Direction

This project aims to deliver a new CNN architecture that enhances both classification accuracy and clinically relevant visualisation explanations. The scope of this project includes modifying the existing model's architecture to reduce its parameter counts and achieve the same accuracy. In addition, at the end of the project, the model will be uploaded to a cloud service with a simple interface to demonstrate the feasibility of streamlining clinical workflows and providing easier access to AI-enabled tools.

One prediction model that predicts various diseases could be extremely large and computationally expensive compared to a model that focuses on specific diseases. Therefore, this project's model is narrowed down to Breast Ultrasound dataset to determine whether the breast picture is normal, benign and malignant. In addition, the last submitted proposal report writes that the project aims to adapt the existing SOTA models to get a relatively lightweight model. However, due to the SOTA models, performance is relatively great. Removing or substituting the existing layer would cause significant accuracy loss; thus, this report changed the direction to use the existing technique to prune the existing SOTA model to achieve lightweight rather than replacing part of the model architecture.

1.4 Contributions

The contributions of this project are as follows. First, the project proposes an XAI-guided pruning method on the Faster R-CNN architecture, a type of deep learning neural network that aims to have both similar detection and classification accuracy to the model and interpretability through the generation of the heat map. The heat map is the representation of how the model predicts to get such an outcome and therefore addresses the deep learning neural network's black-box nature. As a result, the clinician would have better trust in the model's outcome. In addition, by deploying the machine learning prediction pipeline through the cloud, clinicians would gain easier access to deep learning-assisted tools, such as models that provide predictions on medical images. This support can aid clinical decision-making while offering a more cost-effective option, as clinicians can use these tools without needing to develop their own deep learning models.

1.5 Report Organization

This report is organised into seven chapters: Chapter 1 Introduction, Chapter 2 Literature Review, Chapter 3 System Design, Chapter 4 System Implementation and Testing, Chapter 5 System Outcome and Discussion, Chapter 6 Conclusion. The first chapter introduces problem statements and motivation, project objectives, project scope, project contribution and report organization. The chapter 2 reviews related works on Faster R-CNN, XAI, and pruning methods. The chapter 3 shows the methodology used to develop the model, including dataset preparation, model training, Grad-CAM visualisation and pruning strategy design. The chapter 4 describes the system design of this project. It includes dataset processing, detection, explainability, pruning and evaluation modules. The chapter 5 explains the implementation details of the project. The involved components are development environment, model configuration, training process, Grad-CAM implementation and pruning implementation. The chapter 6 shows the experimental results and thorough result discussion, focusing on detection performance, pruning trade-offs, inference efficiency and comparison with related work. Lastly, the chapter 7 concludes the project works and provides recommendations for future work improvement.

Chapter 2

Literature Review

This section reviews existing works on explainable artificial intelligence (XAI) in medical imaging, with particular attention to deep learning framework. The review is organised around techniques, application domains, and explainability approaches.

2.1 XAI Visualisation Techniques Survey Paper [1]

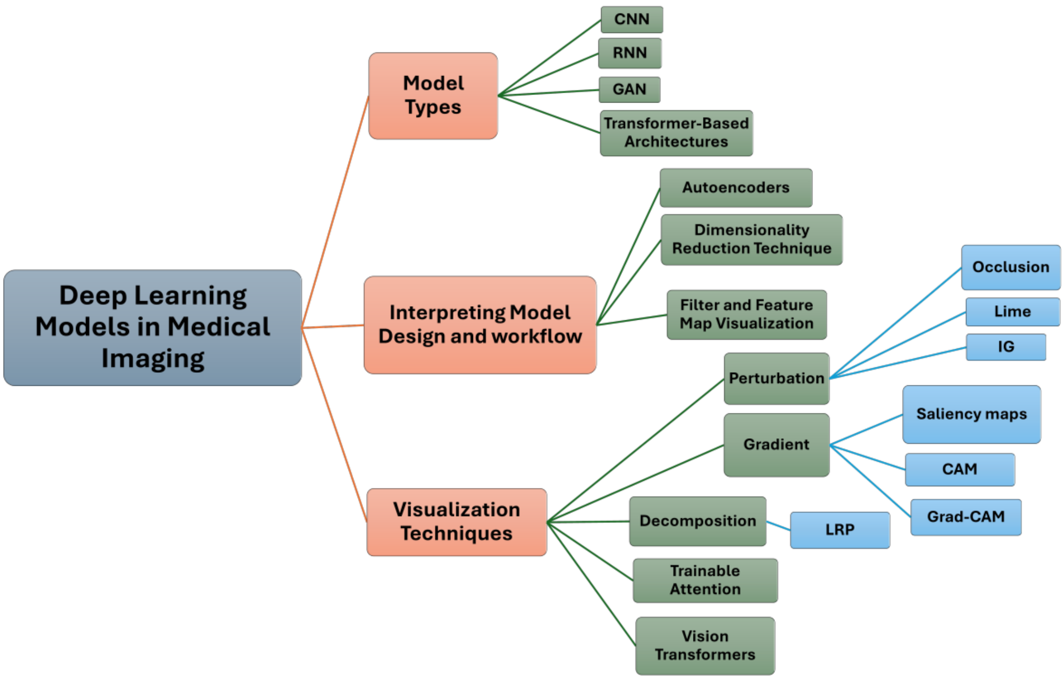


Figure 2.1: Deep Learning Models in Medical Imaging.

A thorough analysis of XAI methods applied to deep learning (DL) models in medical imaging can be found in the survey paper "A Survey on Explainable Artificial Intelligence (XAI) Techniques for Visualizing Deep Learning Models in Medical Imaging" by Bhati, Neha, and Amiruzzaman. They reviewed 240 studies of cutting-edge methods for analyzing and displaying deep learning (DL) models, providing a consolidated framework for DL in medical imaging. It highlights model types such as Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based architectures. It also discusses interpreting model designs such as autoencoders (AEs), dimensionality reduction techniques, and filter/feature map visualisation, alongside post-hoc XAI methods such as perturbation, gradient, decomposition, attention maps from Vision Transformers (ViTs), and trainable attention. Figure 2.1 illustrates an overview of common practices adopted by researchers. The authors highlight the persistent challenge of interpretability and transparency in DL models, particularly in critical healthcare contexts where reliability, ethical accountability, and clinical adoption are essential. Through a systematic literature review of 240 papers, the study categorises and evaluates a wide range of XAI methods across modalities such as MRI, CT, X-ray, ultrasound, fundus imaging and histopathology.

The review shows that methods such as Class Activation Maps (CAM) and Grad-CAM are highly effective for image classification and localisation, while Layer-wise Relevance Propagation (LRP) excels in segmentation tasks, providing clearer distinctions than gradient-based methods in conditions like Alzheimer's disease and multiple sclerosis. Integrated Gradients (IG) has been widely adopted for classification tasks with competitive AUCROC scores, whereas attention-based methods enhance both accuracy and interpretability. Vision Transformer (ViTs) demonstrate strong performance in segmentation, classification, and detection. While model-agnostic approaches like LIME are highlighted for their lower computational cost and suitability for real-time use.

The survey also tracks development trends, noting a surge in XAI research between 2017 and 2020, with gradient-based methods being the most popular. Computational comparisons reveal that perturbation-based methods are slower but model-agonistic. While gradient-based, decomposition and attention-based methods vary in efficiency and often require access to model parameters. Beyond technical findings, the paper identifies ongoing challenges such as scalability for high-resolution images, integration into clinical workflows, robustness across patient populations, lack of standardized benchmarks, and unresolved ethical concerns regarding fairness and accountability.

Finally, the authors propose number of prospective lines of research, including extending XAI to microscopic and molecular-level imaging, developing lightweight real-time methods, designing clinically intuitive interpretability tools, improving robustness across modalities and demographics, and establishing common standards for evaluation. In conclusion, the survey outlines that XAI offers significant promise in medical imaging, overcoming challenges through collaborative, standardized and ethically grounded research is crucial for successful integration into healthcare and for improving patient outcomes.

While the survey paper provides a comprehensive technical review of XAI methods, it places less emphasis on how clinicians actually perceive or interpret the generated heatmap explanation in practice. Moreover, although the paper synthesizes findings from a wide range of studies, it does not propose any standardized evaluation metrics for assessing either model performance or the quality of visual explanations. This limits its utility for guiding future research, as clear benchmarks are necessary to ensure comparability, reproducibility and clinical relevance.

2.2 SMARTSIGHT [2]

This paper proposes a framework that employs ResNet50, a CNN, together with Gradient-weighted Grad-CAM for visualisation, to detect diabetic retinopathy (DR). The authors aim to overcome the drawbacks of conventional diagnostic techniques, which depend on ophthalmologists manually examining retinal pictures and it is labour-intensive and time-consuming. The creation of an Android application that enables users to upload retinal images captured with a smartphone-based fundus camera for automatic DR detection and makes this work novel. The system is supported by Gradio, an open-source Python library, which enables real-time integration between the application and the underlying model. This ensures that any updates to the model are directly reflected in the user application without additional overhead. The framework successfully generates heatmaps overlaid on the input image to highlight regions influencing the model's decision, and the model reports a relatively high accuracy of 96%. However, the system presents notable drawbacks. Firstly, users may lack the expertise to interpret heatmaps or to determine whether the model's prediction is biases or clinically reliable. Secondly, the form of explainability offered is limited, as the heatmap merely indicates areas of focus without providing deeper insight into disease severity or diagnostic reasoning. The paper could be strengthened by incorporating a communication pipeline between end-users and ophthalmologists, ensuring that citizens receive clinically validated interpretations and thereby addressing the issue of limited access to specialized ophthalmic care.

2.3 Overview of DL for Medical Image Computing [3]

The authors of this paper proposed a comprehensive overview and implementation of popular XAI techniques within deep learning (DL) for medical image classification. They applied four CNN models to three publicly available medical datasets, skin cancer, brain tumors, and chest X-rays. and applied five commonly used XAI methods under the umbrella of Class Activation Map (CAM). A key novelty of their work was the quantitative evaluation of these XAI techniques using the “confidence increase” metric derived from the Remove and Debias (ROAD) technique. This framework systematically assessed the reliability of explanations by eliminating the features and tracks the result changes in model performance. The impact of eliminated features on decision-making is reflected in this confidence variation. This provides a precise quantitative benchmark for evaluating the effectiveness and dependability of XAI techniques.

The study primarily aimed to address two pressing issues, the urgent need for openness and interpretability in DL models for therapeutic applications, and the lack of commonly used quantitative metrics for evaluating XAI explanations, as most prior work has relied heavily on qualitative human analysis. Their experiments demonstrated that ResNet50 acquired accuracy and F1 scores across all data, including 86.31% accuracy for skin cancer classification. Quantitatively, XGradCAM showed the highest confidence increase compared to GradCAM++, and LayerCAM, excelling at visualising key features and correctly highlighting abnormal regions. From a computational perspective, LayeredCAM, EigenGradCAM and XGradCAM offered reasonable time costs suitable for resource-constrained clinical environments, whereas AblationCAM and GradCAM++ imposed significant computational burdens.

Despite these contributions, the study had several limitations. The performance of XAI methods in complex clinical cases remained weak, with some techniques, such as EigenGradCAM, proving less effective for certain skin cancer classes and failing to capture abnormal regions in melanoma. The authors also noted the need for further research into hybrid XAI techniques, optimisation of high-cost methods for scalability, and the development of higher-quality visualisation tools to improve consistency.

2.4 Skin Cancer Classification using ViT and CAM [4]

The authors proposed an explainable skin cancer classification framework that integrates DL models with XAI methods, especially utilising two ViT, three Swin Transformer, five pretrained CNN and three visual-based XAI visualisation techniques under CAM family. A key novelty of their approach was the comprehensive application and assessment of the SOTA XAI techniques to gain a deeper insight into model prediction rationale, thereby enhancing interpretability and transparency. To further strengthen classification outcomes, they employed the SMOTE to overcome class imbalance problem within the dataset.

The study sought to address the critical need for transparency and interpretability in clinical applications of deep learning for skin cancer diagnosis, given that the black-box problem of such DL model frequently prevents them from being used in medical practice.. In addition, the authors aimed to improve both classification accuracy and explainability, tackling challenges associated with subtle variations in skin lesions, class imbalance, dataset quality, and the limited interpretability of existing machine learning approaches.

Experimental results demonstrated competitive performance across the evaluated models. While ResNet50 surpassed all evaluated models with an accuracy of 88.8%, a precision of 86.9%, a sensitivity of 88.6%, an F1-score of 87.8%, and a specificity of 88.9%, ViT-large and ViT-base attained accuracies of 88.6%. The incorporation of XAI method with ResNet50 improved the reliability of automated diagnostic support by confirming that the model concentrated on clinically significant areas of the pictures when classifying lesions as benign or malignant. In comparison to bigger transformer-based models, ResNet50 was found to be a feasible choice that combines robust prediction performance with controllable computing complexity.

Despite these contributions, several limitations were highlighted. While XAI models provided useful visual explanations, some heatmaps were generated from non-essential or irrelevant image regions in complex cases, indicating limitations in the models' feature learning. Moreover, the study lacked quantitative evaluation of the XAI methods, relying solely on visual assessment. The authors acknowledged the need for future collaboration with dermatologists to verify the medical validity of the heatmaps. In addition, the reliance on a limited number of public datasets constrained the generalisability of findings, as small-scale datasets risk overfitting and poor performance in diverse real-world settings. Future research directions proposed include exploring additional XAI methods, hyperparameter optimisation, and the development of multimodal XAI frameworks to further improve both model accuracy and interpretability.

2.5 Activation-based pruning of neural network [5]

The neural networks are computationally costly to deploy on edge devices. Therefore, the author proposed a novel network compression technique called "activation-based pruning" that aims to reduce the deep learning network parameters in fully connected neural networks. The proposed method eliminates the network through the low-weight connections, which are determined by checking specific neuron weights. In more specific words, the number of times the non-linear output is non-zero during the forward pass of the training phase. This paper addresses a critical problem in traditional magnitude-based pruning.

The author highlights that the proposed method is able to reduce the deep learning model up to 9 times smaller than the original size while still remaining the same prediction accuracy. Besides, traditional method pruning would require retraining after the pruning process and might cause the eliminated neurons to connect again. However, this method could keep up to 9x smaller neuron connections completely disconnected and maintain the simplified structure. Furthermore, the importance score feature of this pruning method is the crucial part. It utilizes the neuron and weights score, computes it and calculates for shrinkage if it is below the threshold value. Lastly, it is computationally efficient through the local pruning at the same time. In the traditional method, pruning the entire network at once could cause memory overhead and therefore requires a huge amount of memory to store the network. However, this activation-based pruning takes advantage by using a local pruning method; it prunes the network layer by layer and thus speeds up the pruning process while maintaining the same pruning result as global pruning.

The limitation of this paper is that the experimental results are limited to simple image classification datasets, which are MNIST and Fashion MNIST and the results come from a 4-layer feedforward network, which is a very small neural network compared to the SOTA CNN models.

2.6 Pruning Convolutional Neural Networks for Resource Efficient Inference [6]

The authors introduce a new method to prune deep learning models, especially in the CNN field that processes pictures, which is called Taylor expansion to estimate how much the model accuracy drops if the particular filter is removed. It addresses the CNN models that require an abundant number of resources, such as memory and power, that are unable to be deployed on edge devices through permanently removing part of the chunk of the network. By removing a particular unnecessary chunk through pruning the neural network's convolutional kernels, it becomes less resource-demanding and improves the model inference speed. The Taylor expansion pruning method shows an advancement in model optimization by identifying and pruning the redundant neural network components compared to traditional techniques that solely rely on parameter size and activation frequency. As mentioned by the author, pruning would cause a few model accuracies drops as the network lost its knowledge due to the removal. Therefore, it would take a few epochs of training to recover the missing knowledge after pruning.

The limitations of this paper are that shrinking the deep learning model could be time-consuming, as it is an iterative process to delete a particular part, retrain the model and evaluate if the pruning process is effective. Besides, the proposed Taylor expansion tries to take advantage of the computational resource by only looking at the first order of the expansion to view the accuracy loss; therefore, it is just an estimation rather than accurately identifying the accuracy loss when removing the filter.

2.7 Breast Ultrasound Image Detection Based on Dual-Branch Faster R-CNN [7]

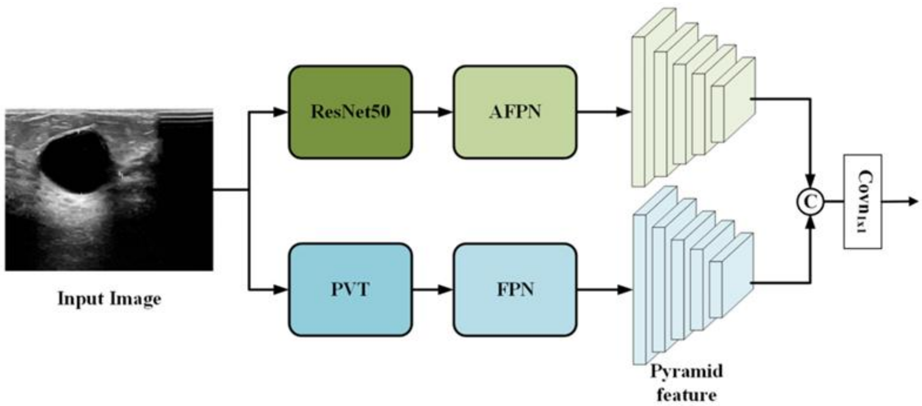


Figure 2.2: Dual-Branch Faster R-CNN Architecture. Adapted from [7].

The paper introduces a novel dual-branch deep learning model called D-faster R-CNN that aims to automate detect the tumor in BUSI dataset. It has a parallel dual-branch system that fuses the local feature extraction capabilities of ResNet50 with Pyramid vision Transformer, a type of vision transformer that capture the global contextual information of the feature maps. To process the image, the proposed architecture inputs the image into the branch simultaneously. The ResNet50 branch uses an asymptotic feature pyramid network (AFPN) to retrieve low-level and high-level features, meanwhile PVT branch utilizes a standard feature pyramid network (FPN) to get the semantic representation of the global features. The multi-scale features generated from both branches and combined together and passes through the convolutional 1x1 layer, allowing the model to perfectly merge both the data to generate candidate regions and create final bounding-box prediction.

Experimental results of this dual-branch deep learning model on the BUSI dataset achieved an AP of 38.4% and image-based sensitivity of 77.8%. Showing a better result than the standard single-branch model and modern transformer-based detectors. The model proved that it is adept at correctly locating both small benign tumors and large, irregularly shaped malignant tumors. Despite the successful result, the study has certain limitations, such as the ultrasound image flaws that all deep learning models barely could achieve high accuracy of the ultrasound image and image artifacts noise.

2.8 Summary of previous studies and Improvement could be done

Previous studies use ViT, a computationally expensive deep learning neural network, for their work. However, for a simple task that involves relatively easy data to be analyzed, such as a breast ultrasound dataset, a CNN would be sufficient to address the existing problems. Besides, CAM variations of visualisation techniques are widely used across many XAI-related papers. However, some of the visualisations are not clinically relevant to what researchers expect. Therefore, the project would try to explore various techniques to maximize the visualisation relevance. Additionally, this project would utilize the concept of activation-based pruning and Taylor expansion to integrate into the model in order to evaluate its effectiveness and compare it with the proposed pruning method.

Chapter 3

System Methodology/Approach OR System Model

3.1 Overall Project Workflow

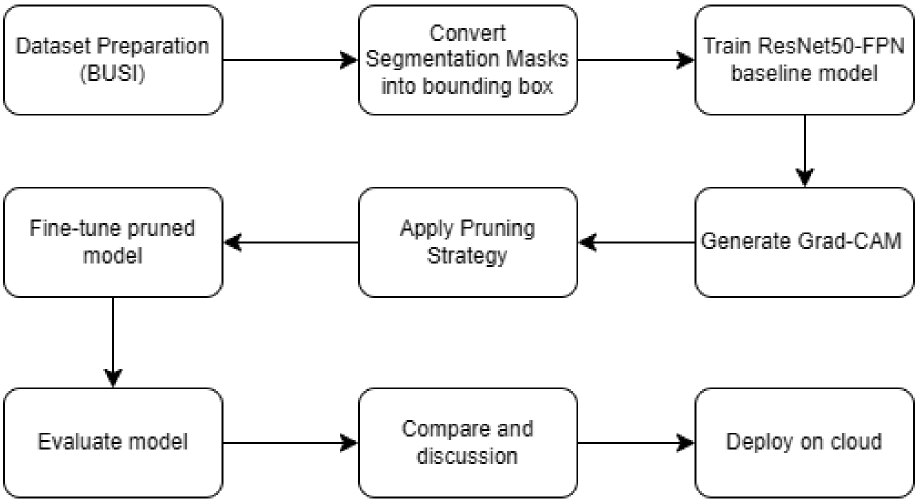


Figure 3.1: Overall Project Workflow.

The project structure follows a development-based deep learning workflow for breast ultrasound lesion detection (BUSI) dataset. Figure 3.1 shows the overall project workflow in structured flow directed by arrows. The starting of the project workflow is the dataset preparation where organize the images and segmentation masks into training, validation and test set respectively. The segmentation masks are then converted into bounding boxes for the Faster R-CNN model to use for the object detection training. Although it is an object detection training, this project would report the both classification metrics and object detection metrics to show the effectiveness of the project development.

Next, a Faster R-CNN model with ResNet50-FPN as a backbone model is developed and trained. The trained baseline model is evaluated using both detection and classification metrics. After the baseline training, Grad-CAM visualisation is applied to the model to generate heatmap to show the regions contributing to model predictions and also to address the black-box nature of the deep learning model. After that, the pruning strategies are applied to the ResNet50-FPN selected backbone layer. The pruning strategies include random pruning, activation-only pruning, Taylor-only pruning, hybrid activation-Taylor pruning, and loss-guided Grad-CAM pruning. Each pruned model is fine-tuned and evaluated to compare the detection performance, parameter changes, inference speed, and explanation quality. Lastly, the best model with the best pruning strategy would be deploy on the cloud to show the feasibility of model prediction cloud deployment idea model for the clinicians to use it throughout the world.

3.2 Dataset Preparation [8]

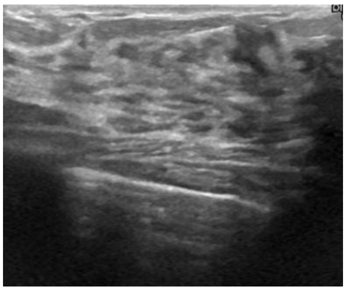


Figure 3.2: Normal Ultrasound Picture.

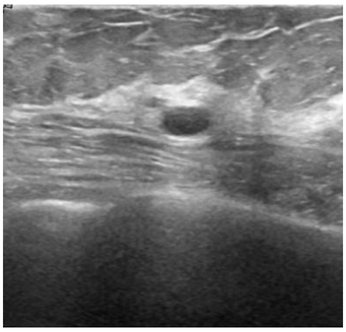


Figure 3.3: Benign Ultrasound Picture.

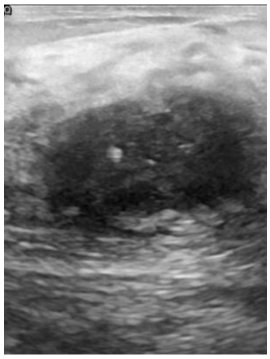


Figure 3.4: Malignant Ultrasound Picture.

The breast ultrasound image dataset is obtained from the data article "Dataset of Breast Ultrasound Images" [8]. The dataset contains 780 images, separated into three categories: normal, benign, and malignant. The normal category contains 133 images, the benign category contains 437 images, and the malignant category contains 210 images. Figure 3.2, Figure 3.3, and Figure 3.4 show the normal, benign, and malignant ultrasound images, respectively.

For data preprocessing, the images are split into 70% for the training set, 15% for the validation set, and the remaining 15% for the test set. The dataset is split to allow model building, tuning, and evaluation, thereby preventing overfitting and ensuring that the model can correctly predict inputs in real-world contexts. In addition, the images are preprocessed and loaded into memory at a fixed resolution of 224 × 224 pixels, as most CNN architectures expect a uniform input size. The images are transformed and normalized using standard. During training, data augmentation is applied to increase model robustness and reduce overfitting, especially since the dataset is relatively small compared to other standard datasets. In this project, only horizontal flipping is used as the augmentation method, as more aggressive augmentation could negatively impact training performance. The corresponding bounding boxes are also adjusted accordingly after flipping.

3.3 Segmentation Mask to Bounding Box Conversion

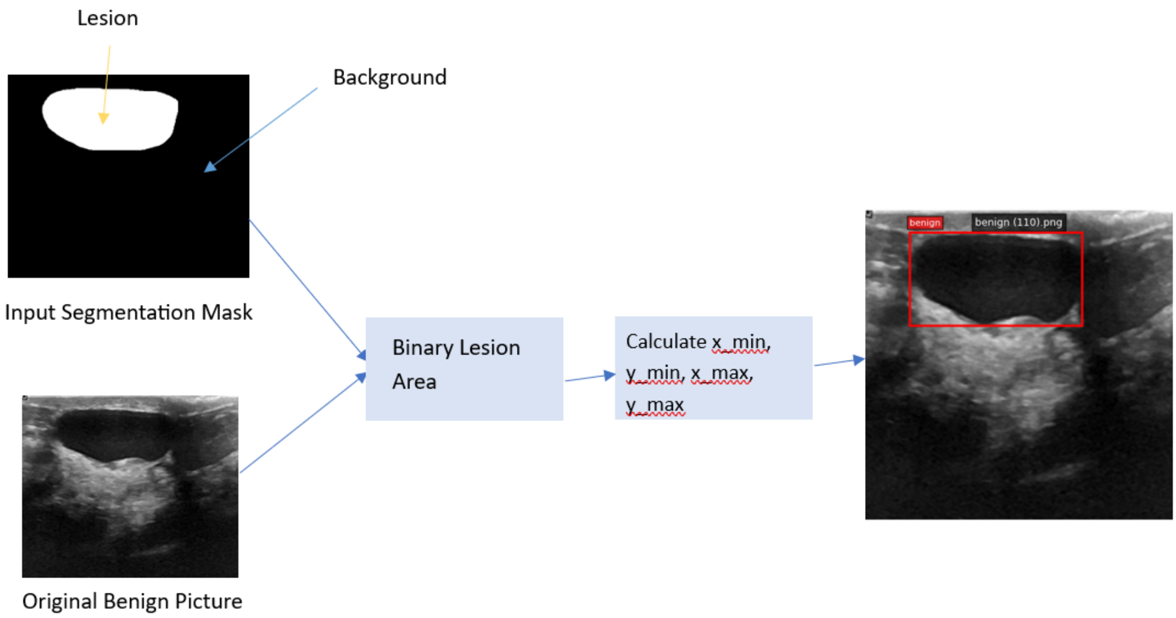


Figure 3.5: Bounding Box Conversion Process.

Since Faster R-CNN are specialized in object detection and requires bounding box annotations, the segmentation masks provided in the BUSI dataset are converted into rectangular bounding boxes through several processes. Malignant or benign dataset would contain at least 1 or more masks. Therefore, the particular picture with multiple masks would undergoes 1 more bounding box conversion process to get both bounding box coordinate and create the bounding box for detection process. The x_min, y_min, x_max and y_max are coordinate value of the lesion pixels to form the bounding box.

The specific bounding box conversion process would be loading the segmentation mask as a grayscale image. Lesion represent the white pixel whereas black pixel represent background. Next, it is converted into the mask value into a binary image using a pixel threshold and identify all foreground lesion pixels. By getting the binary image value, get the maximum and minimum x and y coordinates and stores it. Based on the x and y values, create a bounding box in the format of [x_min, y_min, x_max, y_max]. Lastly, the bounding box coordinate is resized based on the resized image dimensions which is 224x224. The number of output classes is based on the number of classes and with additional background class as Faster R-CNN requires a background category for object detection.

3.4 Faster R-CNN Detection Model Development

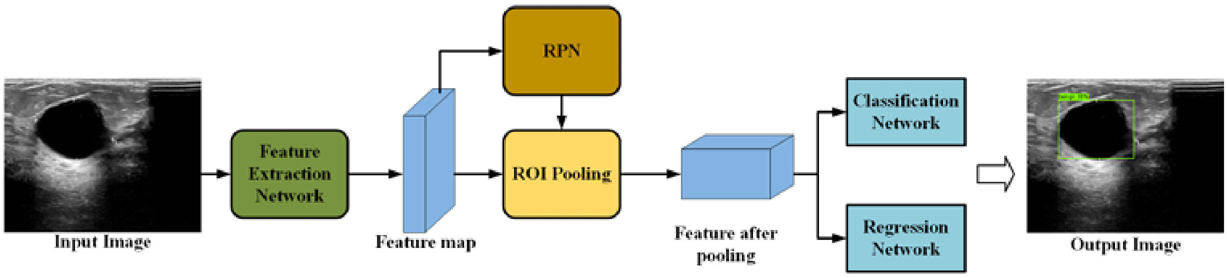


Figure 3.6: Structure of Faster R-CNN framework. Adapted from [7]

Faster R-CNN evolved from Fast R-CNN where it takes advantage of the Fast R-CNN architecture on object detection and solves the localisation speed through the Region Proposal Network (RPN) that is integrated as part of the model that runs in the GPU. Whereas Fast R-CNN uses a selective search algorithm, which is significantly slower than Faster R-CNN [9].

The main model developed in this project is using Faster R-CNN with ResNet50-FPN as a backbone for the various pruning techniques. The reason Faster R-CNN is used in the project is that Faster R-CNN is a great two-stage object detection model with strong localisation features, whereas ResNet50-FPN is used as it is a strong benchmark for this framework.

The model contains three main components. Firstly, the ResNet50-FPN backbone extracts feature maps from the BUSI image, RPN generates candidate object regions and the detection head classifies the proposed regions and refines the bounding-box coordinates. Besides that, the ResNet50-FPN baseline model is trained for 30 epochs so that the model can learn to extract the features, get the correct candidate object regions and classify it correctly through the classification loss and bounding-box regression loss. Lastly, the validation loss during the training would be monitored and the training epochs with the lowest loss would be the best-performing model and the model weights are saved for evaluation and tested for various pruning techniques.

3.5 Grad-CAM Explainability Integration

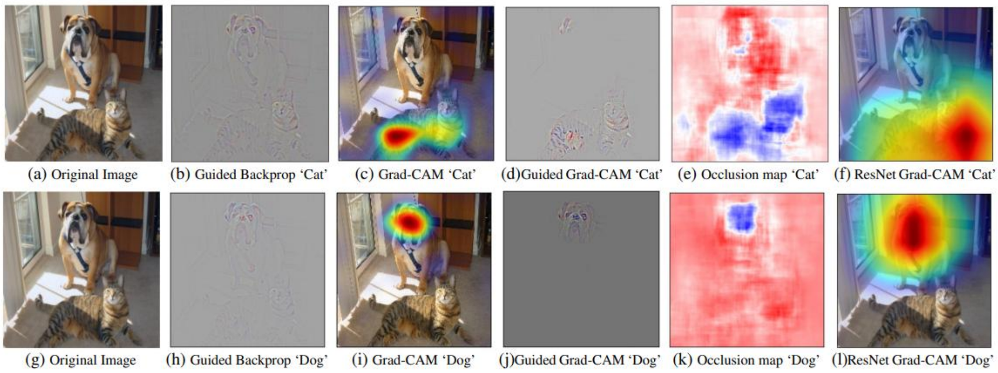


Figure 3.7: Grad-CAM Heatmap Visualisation. Adapted from [10].

Grad-CAM is a technique that aims to produce a visual explanation of the model decisions. The purpose of it is to address the black-box nature of the deep learning model to make it become more transparent and explainable. It applies the technique on the last layer of the convolutional layer of CNN as the last layer of CNN consists of high-level features that have meaningful context and it would be more visually explainable to the human to interpret it. The way it understands how the model makes such a prediction is through the pixel's gradient. Whenever a particular pixel has a high gradient score, it is saying that the neuron is important in conveying such information that the model predicts it [10].

To address the black-box nature of the CNN model, Grad-CAM is integrated into the project to provide a visual explanation of the model predictions through overlaying the heatmap on the picture. Furthermore, not only visual explanation, but also using grad-cam to assign an importance score to the filters would reduce the unnecessary filters that are involved in the backbone layer. Theoretically, pruning using Grad-CAM would reduce the heatmap noises that do not provide any information during model inference. Besides, after producing the visualisation heatmap using Grad-CAM, the generated heatmap would be resized to match the input image size and overlaid on the ultrasound image. The visualisation is used to evaluate if the model focuses on lesion regions or irrelevant background areas.

3.6 Pruning Strategy Design

The pruning strategy is implemented to compare different filter-importance in contributing the model prediction. The selected backbone layer for pruning is backbone.body.layer4[-1].conv3. This particular layer is selected is because it contains the high-level features that is safe to be pruned as pruning the front convolutional layers would affect the entire SOTA model performance. To be safe, this project would try to prune the high-level layer. There are 5 pruning strategies implemented which are random pruning, activation-only pruning, Taylor-only pruning, hybrid activation-taylor pruning, and lastly loss-guided Grad-CAM pruning. Random pruning would randomly select filters to be pruned. This is act as a baseline to compared with other 4 methods effectiveness rather than random pruning. Next is activation-only pruning. This activation-only pruning calculate the filter importance based on the average activation value of each filter. Filters with the low assigned activation value during training would be look as less activate and would be selected for pruning. Besides, Taylor-only pruning would estimate the filter importance using the product of activation and gradient value. This calculates the sensitivity of the loss to each filter and the lowest value will be eliminated. The hybrid activation-Taylor pruning combines each advantage and address the disadvantage of both pruning techniques. Activation-based pruning method only looks at the filter importance based on the filter activation score where low filter activation score's filter might be important on specific detection, whereas taylor-only pruning could support greatly to predict if removing low filter activation score would affect the model accuracy. If removing a low activated filter impacts the model sensitivity, then that particular filter is retained. As this project mainly will be doing ablation study on lightweight model through pruning, therefore ablation study of 2 hybrid technique weightage is not covered in this project. Lastly, the loss-guided Grad-CAM pruning is based on the feature activation and detector loss gradient. The filters with high loss would be selected for pruning.

As mentioned in [5,6], each pruned model must fine-tune to let the model learns back the missing knowledge to achieve back the accuracy. Therefore, this project would be undergoing 10 epochs after pruning to recover back the model performance.

3.8 Evaluation Methodology

The model performance would be evaluated under 2 different fields, which are object detection accuracy and classification accuracy.

For object detection evaluation metrics, the main evaluation metrics are mAP@0.5 and mAP@0.75. mAP@0.5 is a metrics that evaluate if the predicted bounding box overlaps with the ground-truth bounding box by at least 50% Intersection over Union (IoU). Meanwhile mAP@0.75 has a stricter localisation requirement and it is sensitive to bounding-box accuracy [12]. Furthermore, the efficiency of the model is evaluated using parameter count and inference time per image. However, the pruning is only selected to the last layer of the convolutional layer and thus the room of improvement for aggressive pruning would be lower.

For classification accuracy metrics, the main evaluation metrics are F1 score only. These metrics would be calculated if the predicted data is matched with the ground-truth required IoU threshold. Therefore, this is not considered as pure classification accuracy.

The evaluation for the Grad-CAM could be evaluated qualitatively through human inspection as there are no standard evaluation metrics for the XAI visualisation heatmap.

3.9 Cloud Deployment

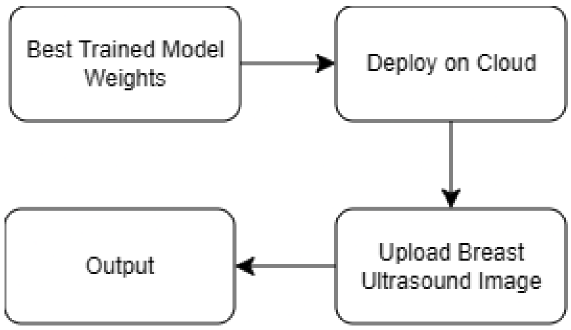


Figure 3.8: Cloud Deployment Pipeline.

The cloud deployment workflow consists of 4 stages. Firstly, the trained PyTorch model weights are exported as .pth file. Second, deploy the model on cloud and create an inference model wrapper that written on the cloud would reconstruct the Faster R-CNN architecture and loads the exported weights. Third, an uploaded ultrasound image is pre-processed and passed into the model for prediction. Finally, the system would return the predicted class label, confidence scores, bounding-box and Grad-CAM visualisation on the image.

As mentioned previously, the deployment of the model is to show the feasibility of the cloud deployment and it is not for clinical diagnosis or production medical use.

Chapter 4

System Design

The proposed system aims to create a deep learning model, specifically Faster R-CNN model for breast ultrasound lesion detection pipeline. The system receives the BUSI dataset as input and produces the defined output which are predicted class label, bounding boxes, confidence score and Grad-CAM visualisation heatmap.

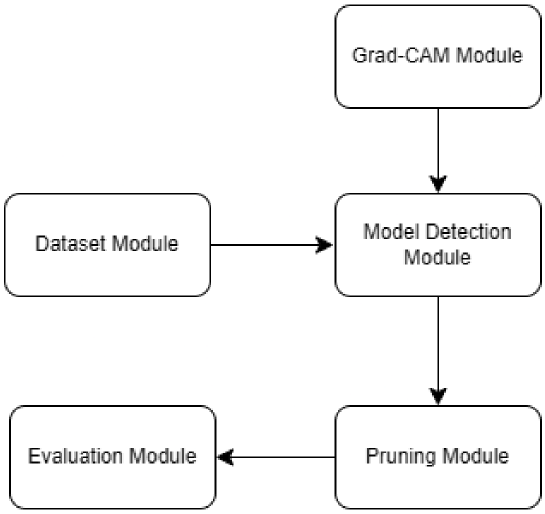


Figure 4.1: High-level System Architecture pipeline.

Figure 4.1 shows the high-level system architecture pipeline where contains various modules. Dataset module, model detection module, Grad-CAM module, pruning module and evaluation module. The dataset module converts the BUSI segmentation masks into bounding-box annotations. The model detection module trains the ResNet50-FPN backbone. The pruning module applies five different filter-importance strategies to the selected target layer. Finally, the evaluation modules that evaluate for the model and pruning effectiveness and pass all the results into output module for structured output.

4.2 System Components Specifications

4.2.1 Hardware and Platform Specification

Component	Specification
Development Platform	Google Colaboratory

GPU	Google Colab L4 GPU
Storage	Google Drive for dataset, logging and model's checkpoints
Deployment Environment	CPU-compatible Python environment
Programming Language	Python

Table 4.1: Hardware and Platform Specification Table.

4.2.2 Software Specification

Software/Library	**Purpose**
Python	Main programming language with various deep learning libraries supported
PyTorch	Model training and inference
Torchvision	Faster R-CNN and computer vision utilities
NumPy	Numerical processing
PIL	Image loading and preprocessing
Matplotlib	Plots and visualisation
Seaborn	Confusion matrix visualisation
Scikit-learn	Classification report and confusion matrix

Table 4.2: Software/Library Specification Table.

4.2.3 Dataset Specification

Item	**Description**
Dataset	Breast Ultrasound Image (BUSI)
Classes	Benign, Malignant, Normal
Original Annotation	Segmentation Masks
File Type	png
Input Image Size	224 x 224
Output	Bounding Boxes, labels, Image ID, Area, iscrowd

Table 4.3: Dataset Specification Table.

4.2.4 Model Specification

Item	**Description**
Detection Model	Faster R-CNN
Backbone	ResNet50-FPN
Classes for detection	Background, Benign, Malignant, Normal

Pretrained Weight	COCO-pretrained Faster R-CNN weight
Output	Class Label, Confidence Score, Bounding Box

Table 4.4: Faster R-CNN Model Specification Table.

4.2.5 Pruning Specification

Item	**Description**
Target Layer	Backbone.body.layer4[-1].conv3
Pruning Type	Masked filter pruning
Pruning Ratio	10%, 30%, 50%, 70%
Strategies	Random, activation-only, Taylor-only, hybrid, Grad-CAM guided
Fine-tuning	10 epochs after pruning

Table 4.5: Pruning Specification Table.

4.3 Software and Model Design

4.3.1 Dataset Module Design

The dataset module design is a crucial system component in this system architecture. It processes the raw BUSI dataset into a compatible format for the Faster R-CNN detection training. Since the original dataset of BUSI is for segmentation and has pixel-level mask annotations, this module aims to transform the mask into bounding box annotations for the detection module to draft out bounding boxes. The main component of this module is the bounding-box generation algorithms. For each ultrasound image loaded into the pipeline, it would undergo the following stages: mask binarization, foreground extraction, coordinate calculation and bounding box formulation.

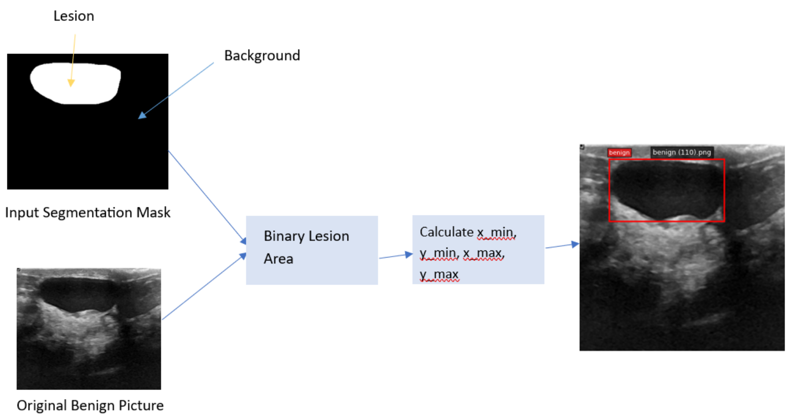


Figure 4.2: Segmentation Mask Transformation Pipeline.

Figure 4.2 shows the transformation pipeline for the segmentation mask. For the mask binarization, the segmentation mask is loaded as a grayscale image and converted into binary format through the threshold. Next, for foreground extraction, it extracts the foreground and identifies all foreground pixels that represent the lesion area. After that, for the coordinate calculation, x minimum, y minimum, x maximum and y maximum would be calculated based on the lesion's leftmost coordinate, lowest coordinate, rightmost coordinate and highest coordinate by using np.where(mask > 0) method provided by numpy. Lastly, for the bounding box formulation, coordinates of x and y are used to construct a rectangular bounding box and it is formatted as an array in the arrangement [x_min, y_min, x_max, y_max]. In the case where an image contains multiple different lesion masks, the module iteratively extracts the masks and generates the bounding boxes for that particular image. For the case of the normal breast ultrasound image, which does not have a lesion, the module will assign a full-image pseudo-bounding box for the Faster R-CNN detection pipeline, which would detect for the normal class too.

To ensure the input dimension consistency for the ResNet50-FPN backbone, the module resizes the ultrasound images and the corresponding bounding boxes into a fixed resolution, which is 224x224 pixels. Once they are resized, the images are normalized and transformed into PyTorch tensors. Besides that, after the images are converted into PyTorch Tensor data, the training set would undergo data augmentation for better model prediction, as the dataset is small. Horizontal flipping is applied on the training set and the annotations for the training set will also be calculated as well to ensure the ground-truth data consistency.

The last dataset module design is to transform the dataset value into standardized value required by the PyTorch R-CNN. The required value is:

Parameter	**Description**
boxes	Bounding box coordinate
labels	Class label in integer (1 – benign, 2 - malignant, 3 - normal)
image_id	A unique identifier for image batch.
area	Calculated area for bounding boxes for evaluation metrics
iscrowd	A boolean flag that set to 0, indicating all instances are individual distinct objects.

Table 4.6: Dataset Output Variables.

By going through the entire dataset module pipeline, the dataset is able to transform from raw picture format data into Faster R-CNN detection model required data that used for entire process.

4.3.2 Detection Model Module

The detection model module is a core component of the prediction pipeline that responsible for learning the features of the BUSI lesions to learn to predict the lesion spatial locations. This project utilizes the Faster R-CNN detection model to predict three specific foreground classes corresponding to BUSI dataset, along with background class automatically handled by the detection head.

ResNet50-FPN Backbone

The first part of the detection model is the feature extraction backbone. This project uses ResNet50 combined with a Feature Pyramid Network (FPN) which is specialized for Faster R-CNN. The purpose of the FPN integration is to transform the BUSI image into meaningful feature maps that can be used for region proposal and classification. ResNet50 extracts hierarchical image features from different network depths or in another word convolutional layer depths. First few convolutional layer aims to capture low-level visual information such as texture, edges and tissues boundaries. Whereas deeper layers capture more complex features [12] including lesion shape, boundary irregularity and semantic patterns related to benign or malignant structures. The FPN improves the model's ability as it detects objects at different scales as BUSI images' lesion would appear in different sizes and shapes. By combining feature maps from different levels, FPN could able to learn the spatial detail and high-level semantic information during detection.

Region Proposal and Region of Interest (ROI)

The second stage right after the feature extraction backbone is the detection stage where it separates into 2. The first stage is the RPN. It scams the extracted feature maps and generates the candidate regions that might contains a lesion or relevant target area. It anchors that particular coordinate and generates different sizes of candidate boxes and search for the possible object locations. For each candidate region, RPN predicts an object score and a preliminary bounding box.

The second part is the Region of Interest (RoI) head. The proposed regions from the RPN are extracted from the feature map using RoI align. The region features are than passed into the detection head. The task is classification, where the model predicts whether the region belongs to benign, malignant, normal or background. The second task of RoI is bounding-box regression, where the model refines the predicted box coordinates to better match target region.

Training Phase

During training, the model receives the dataset tensors value together with the ground-truth value that passed by the dataset module. The ground-truth value consists of bounding box value and the class label. The model compares the predictions with the both ground-truth value and calculates a combination of detection losses. The classification loss measures how accurately the model predicts the correct class label for each proposed region. The bounding-box regression loss measures how closely the predicted box matches the ground-truth box. The losses are combined into the total training loss and backpropagated through the network. The model weights are updated iteratively using an optimizer. The validation loss is monitored during training and the best-performing model weights are saved for evaluation. The best checkpoint saved could help the model from using the overfitting epoch and only uses the model checkpoint that performs well during the training process.

Inference and Confidence Thresholding

During inference process, the trained model is tested using the unseen ultrasound images to check for the model inference capability. For each input image, the model outputs the predicted bounding boxes, class labels, and confidence score. The purpose of those 3 outputs is to show the user where is the lesion located, model's guesses and how confidence does the model classify it.

A confidence threshold is applied for removing the low-confidence detections. In this project, the predictions that below the threshold value, 0.5 will be discarded and vice versa. The prediction with high confidence value will be used for evaluation, visualisation, and deployment demonstration. The reason of setting the confidence threshold is to reduce the false positive predictions.

4.4 Grad-CAM Visualisation Module

In deep learning models, particularly those used for image analysis, individuals often do not understand how the model makes predictions or derives its rationale. Therefore, a qualitative approach to addressing this issue is to use XAI visualisation. To improve its interpretability, the Grad-CAM visualisation module is integrated into the system pipeline. This approach generates a heatmap that highlights the regions of the image where the model focuses to make its predictions.

In this project, Grad-CAM is used to support the Faster R-CNN model on BUSI dataset explainability. The heatmap is overlaid on the ultrasound image together with the predicted bounding box. This could enable people to inspect the model prediction rationale and also determine if the model is focusing on the lesion region or not.

Faster R-CNN architecture is complex, as it involves FPN and RoI components. Therefore, it takes significant effort to prune the network without error. To save time and resources, this project would not prune the entire network and could only selectively pick a suitable layer that preserves low-level spatial information while maintaining high-level semantic features.

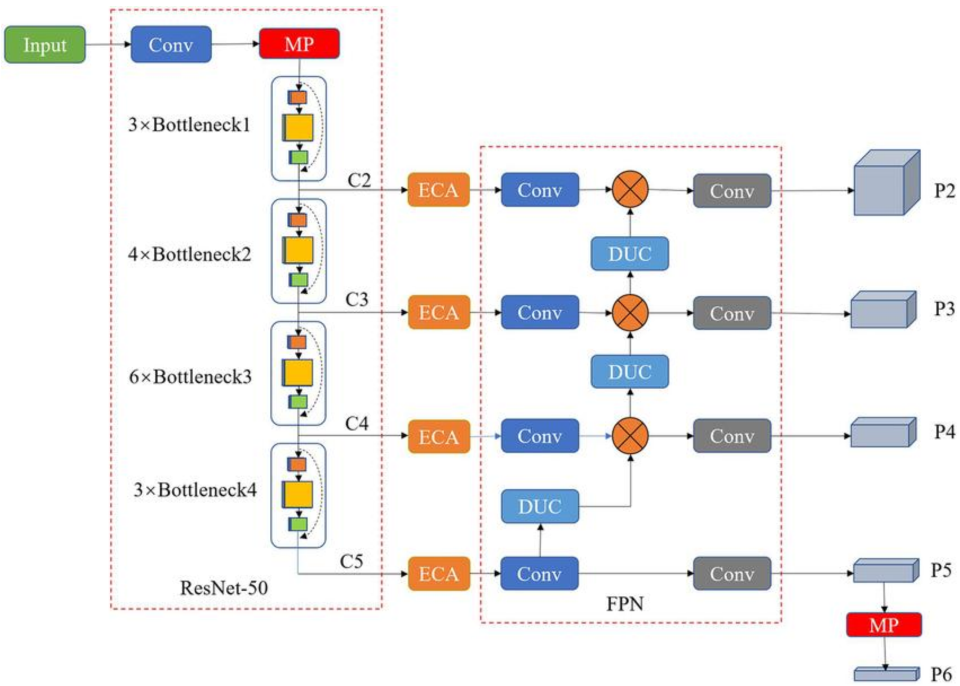


Figure 4.3: Resnet50-FPN Backbone Network. Adapted from [13].

In this project, the target layer used for Grad-CAM is:

backbone.body.layer4[-1].conv3

By referring to figure 4.3, layer 4 is located near the final stage of the backbone. It was selected because it contains high-level visual features that are useful for interpreting the model's detection decision.

Heatmap Generation and Image Overlay

To create a Grad-CAM heatmap, two variable capture data from a specific layer, a forward hook for visual patterns (activation map) and a backward hook for importance levels. Where channel weight could be obtained by summing up the entire channel importance level score. After the activation maps and channel weights are obtained, the Grad-CAM heatmap is generated through several steps. The Grad-CAM equation is the activation map score times the channel weight. The weighted activation maps are summed up to provide an attention map. After that, to know if that particular pixel is contributing on the model prediction, ReLU operation is performed to get the positive contribution. Since the model processes images at a deep and smaller neural network, the heatmap is resized back to the standard self-defined input size of 224x224 pixels. Then values are normalized between 0 and 1 to ensure the relevance score is distributed across the attention map. The red color indicating the value is relevant to model prediction and blue indicating least importance. By adding predicted bounding box on the lesion area, user can check the model's attention to detect the lesion.

Currently, Grad-CAM output is just a qualitative explanation tool, it does not has standard evaluation metrics to evaluate the effectiveness of this visual explanation and only used as supporting reference to evaluate together with other standard well-known metrics. Besides, to evaluate heatmap, if the heatmap appears around the lesion regions and appears at the bounding box area, suggesting that the heatmap visualisation could be reliable and vice versa. However, Grad-CAM should not be treated as proof of clinical correctness. It provides visual support for interpretation, but it does not replace quantitative evaluation metrics such as mAP@0.50 and mAP@0.75. In this project, Grad-CAM is therefore used together with detection performance metrics and bounding-box outputs to evaluate the model more comprehensively.

4.5 Pruning Module

The pruning module is designed for ablation study on the trained Faster R-CNN model. This module acts as an analytical tool to evaluate how different filter-importance criteria affect detection performance. To ensure a fair and consistent comparison across all experiments, pruning is strictly isolated to a single, high-level target layer: backbone.body.layer4[-1].conv3. This layer resides at the deep end of the ResNet50 backbone, making it responsible for extracting complex semantic features critical for identifying breast lesions.

Unstructured Pruning Architecture [14]

Since the Faster R-CNN architecture is complex and involves FPN and RoI components. To safely prune the architecture without causing dimensional mismatches in the downstream components, the model will only be using unstructured pruning rather than structural pruning. Unstructured pruning is a technique that only eliminates insignificant weight that does not contribute to the model prediction, whereas structural pruning is a technique that involves reconstructing the entire architecture and removing the useless component from the existing architecture to achieve effective parameter reduction and speed up inference speed [11]. This method computes importance scores and produces a binary mask for the designated layer. Filters that are labeled as unimportant by the respective pruning technique are mathematically masked and vanished inside the network. Because the filters are masked rather than physically deleted from the tensor architecture, the global parameter count and inference speed experience only minor reductions. Therefore, the primary focus of this module is evaluating accuracy trade-offs and filter redundancy instead of traditional parameter count and inference speed's boost.

Filter Importance Criteria

The main functionality of the pruning module relies on evaluating filters based on five distinct prune ratio to evaluate the pruning effectiveness. The module ranks all filters in the target layer and removes a percentage based on the user-defined pruning ratio (10%, 30%, 50%, or 70%). The five strategies implemented are:

Random Pruning (Baseline)

The random pruning acts as a baseline to compare with other pruning strategy. Filters are randomly selected without analyzing the data. This serves as a control baseline to prove whether mathematical criteria actually outperform blind luck.

Activation-Only Pruning

This activation-only pruning uses the average activation magnitude of each filter across a batch of ultrasound images. It ranks the filter and calculates the activation score. The filters that barely activate would be treated as redundant and removed.

Taylor-Only Pruning

Taylor-only pruning estimates the network's loss sensitivity. It checks if removing the filter would cause the model’s loss function instability. The equation behind it is multiplies the feature activations by the gradients of the loss function. Filters with low Taylor scores are pruned and theoretically it would cause least disruption to the total loss.

Hybrid Activation-Taylor Pruning

The pruning strategy takes advantage of both pruning strategies, whereas Taylor-only pruning takes control of the activation-only pruning technique. It normalizes and equally weights both the activation magnitudes and the Taylor sensitivity scores. This creates a balanced strategy that favors filters that are both highly active and critically linked to the detection loss. The value is derived from the equation, which is 0.5 times normalized activation pruning and 0.5 times normalized Taylor score.

Loss-Guided XAI Relevance Pruning

Inspired by Grad-CAM methodology, this strategy uses the gradients from the detection head to compute a custom relevance score. Filters that do not actively contribute to the spatial localisation of the breast lesion are targeted for pruning.

Fine-Tuning and Recovery

Pruning a model would affect the neural network performance. The immediate effect is the detection and classification accuracy degeneration. To address this, the pruning module includes a fine-tuning phase right after it is pruned. This allows the network to recover from the lost knowledge and maintain the original detection capabilities of the pruned model.

4.6 Evaluation Module

The evaluation module serves as a final component of the system. This part is used to check for the baseline Faster R-CNN model applied by various pruned techniques. Since the object detection in medical imaging has 2 main research for both detection and classification, therefore both metrics are generated for evaluation.

Localisation Metrics (mAP)

For object detection metrics, Mean Average Precision (mAP) are used as it is a common COCO (Common Objects in Context) evaluation protocol. mAP evaluates the model prediction bounding box by using the Intersection over Union (IoU) formula. The module calculates two specific values:

mAP@0.50: A general evaluation metric that requires prediction bounding box to overlap the ground truth by only 50% and above.

mAP@0.75: A stricter metric that requires the predicted bounding box to overlap the ground truth by at least 75%.

Secondary Matched-Detection Metrics

To assess how well the model distinguishes between benign and malignant lesions once it has found them, the module calculates Precision, Recall, and the F1-Score. Crucially, these are matched-detection metrics. The module only calculates these scores for predictions that successfully matched a ground-truth box at the required IoU threshold. Therefore, these metrics reflect the model's classification accuracy only among successfully localized lesions, avoiding the misleading inflation often seen in pure image-level classification.

Efficiency and Profiling Analysis

To evaluate the model pruning efficiency towards the hardware impact, the module profiles the system's computational efficiency. It takes the effective parameter count and subtract the masked filters from the total network parameters. It measures the inference time (ms/time) by averaging the execution speed across the testing dataset. Since the unstructured pruning leaves the physical architecture intact, the module is designed to interpret these efficiency metrics as secondary importance metrics rather than major model compression.

Finally, for every processed breast ultrasound image, the system would generate multiple outputs. These outputs are designed to support both detection and visual interpretation. The first output is the predicted bounding box. The bounding box contains the spatial coordinates of the detected region in the ultrasound image. These coordinates would tell the model the lesion or the location of the target region the model should focus on. The second output is the predicted class label and confidence score. The class label represents the model's predicted category, which may be benign, malignant, or normal. The confidence score indicates the model's certainty for the prediction. Any predictions below the selected confidence threshold are removed before final output to ensure the model prediction consistency instead of hallucinating. The third output is the Grad-CAM visualisation. This output is a color-coded heatmap overlaid on the original ultrasound image. Grad-CAM would highlight the image regions that contributes to the model prediction. The final output of Grad-CAM would help users inspect if the model is focusing on the relevant lesion region or unrelated background areas that does not provide any information for interpretation.

CHAPTER 5
SYSTEM IMPLEMENTATION

5.1 Development Environment

The system was implemented using Python and the PyTorch deep learning framework. The experiments were conducted in a GPU-enabled environment to accelerate Faster R-CNN training and evaluation. Google Colaboratory was used as the main experimental platform, and Google Drive was used to store the dataset, model checkpoints, logs, plots, Grad-CAM outputs, and evaluation results.

The main software libraries used in this project include PyTorch, Torchvision, NumPy, Matplotlib, Seaborn, scikit-learn, and PIL. PyTorch and Torchvision were used for model development, training, and object detection. NumPy was used for numerical processing, while Matplotlib and Seaborn were used for plotting training curves, confusion matrices, filter importance graphs, and Grad-CAM visualisations. Scikit-learn was used to generate classification reports and confusion matrices.

The main model implemented in the system was Faster R-CNN with a ResNet50-FPN backbone. The implementation was configured for breast ultrasound detection using four model classes: background, benign, malignant, and normal.

5.2 Dataset Loading and Preprocessing

To ensure fair experimental run comparison, every experimental run dataset loading should be consistent. Therefore, the BUSI dataset was organised and split into training, validation and testing folder once. Each split contained three class folders, benign, malignant, and normal. The dataset loader read each ultrasound image together with its corresponding segmentation mask files.

For benign and malignant images, the segmentation masks were used to generate lesion bounding boxes by undergo preprocessing step. Each mask was converted into a binary image, and the foreground lesion pixels were identified. The minimum and maximum coordinates of the foreground pixels were then used to create a rectangular bounding box.

All images were resized to a fixed input size of 224 × 224 pixels. The generated bounding boxes were scaled according to the resized image dimensions. The images were then converted into tensors before being passed into the model.

For training augmentation, horizontal flipping was applied randomly. When an image was flipped, its bounding-box coordinates were also updated so that the annotation remained consistent with the image.

Normal images do not contain lesion masks. In this implementation, normal images were assigned full-image pseudo-bounding boxes. This allowed the detector to include normal as a foreground class during training. However, this design was treated as a limitation because normal cases do not represent real lesion localisation.

5.3 Baseline Faster R-CNN Implementation

The baseline Faster R-CNN model was implemented by using ResNet50-FPN as backbone. The model was constructed and initialized using COCO detection training pretrained weights as the BUSI dataset is considered small. After that, the pretrained model was adapted for the BUSI dataset and replaced the prediction and classification head to match the number of the BUSI dataset classes.

The number of classes was set to four:

0: background

1: benign

2: malignant

3: normal

The baseline model was trained for 30 epochs. During training, the model received batches of ultrasound images and necessary values containing bounding boxes coordinate and class labels. The model returned a set of detection losses, and the sum of these losses was used for backpropagation. To ensure the model converges correctly based, the optimizer updated the model parameters using the configured learning rate and weight decay. Validation loss was monitored after each epoch. The model weights with the best validation loss were saved and later used as the baseline model for pruning and evaluation.

5.4 Grad-CAM Implementation

Grad-CAM was implemented to visualize the attention regions of the Faster R-CNN model. A target convolutional layer in the backbone was selected for capturing activations and gradients. The selected layer was:

backbone.body.layer4[-1].conv3

During the forward pass, the feature activations of this layer were stored using a hook. During backpropagation, the gradients with respect to this layer were also captured. The gradients were spatially averaged to produce channel weights. These weights were multiplied with the activation maps and summed to form the Grad-CAM heatmap. The heatmap was passed through a ReLU operation, resized to the input image size, and normalized. It was then overlaid on the ultrasound image using a color map. The visualisation displayed the original image, ground-truth bounding box, predicted bounding box, and Grad-CAM heatmap. Finally, the Grad-CAM results were saved as image files and used for qualitative discussion in the evaluation chapter.

5.5 Pruning Implementation

The pruning implementation focused on the selected target layer:

backbone.body.layer4[-1].conv3

For each pruning strategy, an importance score was calculated for every filter in the target layer. The filters were ranked according to their importance scores, and the lowest-scoring filters were selected for pruning.

The pruning strategies implemented were:

1. Random pruning
2. Activation-only pruning
3. Taylor-only pruning
4. Hybrid activation-Taylor pruning
5. Loss-guided Grad-CAM-style relevance pruning

Random pruning generated random filter scores. Activation-only pruning used the average activation magnitude of each filter. Taylor-only pruning used the product of activation and gradient as a loss sensitivity score. Hybrid activation-Taylor pruning combined normalized activation and Taylor scores using equal weights. Loss-guided Grad-CAM-style relevance pruning used activation and loss-gradient information to estimate filter relevance.

Pruning was implemented using a masking approach. The selected filters were set to zero, and the masks were reapplied during training so that the pruned filters remained inactive. Since the model structure was not physically changed, this method did not produce large inference speed improvements.

After pruning, each pruned model was fine-tuned for 10 additional epochs. Fine-tuning was performed to allow the model to recover from the disruption caused by pruning.

5.6 Evaluation Implementation

The evaluation implementation measured detection performance, classification performance, parameter count, inference time and Grad-CAM visualisation quality.

Detection performance was evaluated using:

mAP@0.50

mAP@0.75

mAP@0.50 measures detection performance when predicted bounding boxes overlap the ground-truth bounding boxes by at least 50% IoU. mAP@0.75 is stricter and requires stronger localisation accuracy which overlap the ground-truth bounding box by at least 75% IoU. These two metrics were treated as the primary evaluation metrics.

Precision, recall, F1-score, classification report, and confusion matrix were also generated. These were calculated using matched detections, meaning that a prediction was included only if it matched a ground-truth box with sufficient IoU. Therefore, these metrics were interpreted as secondary classification metrics for localized detections, not pure image-level classification accuracy.

Inference speed was measured by running repeated single-image inference and calculating the meantime per image. Parameter count was calculated for the baseline and pruned models. Since pruning was masked and applied only to one selected layer, the reported parameter reduction was small. Grad-CAM visual outputs were generated for selected validation images. These visualisations were used to compare whether different models focused on lesion regions after pruning.

5.7 Cloud Deployment Prototype Implementation

A lightweight cloud-deployable prototype was implemented to demonstrate the feasibility to deploy the model through the internet accessible place. Since the main focus on this project is model development and pruning evaluation, Kaggle deployment is treated as a prototype development for cloud demonstration rather than production clinical system.

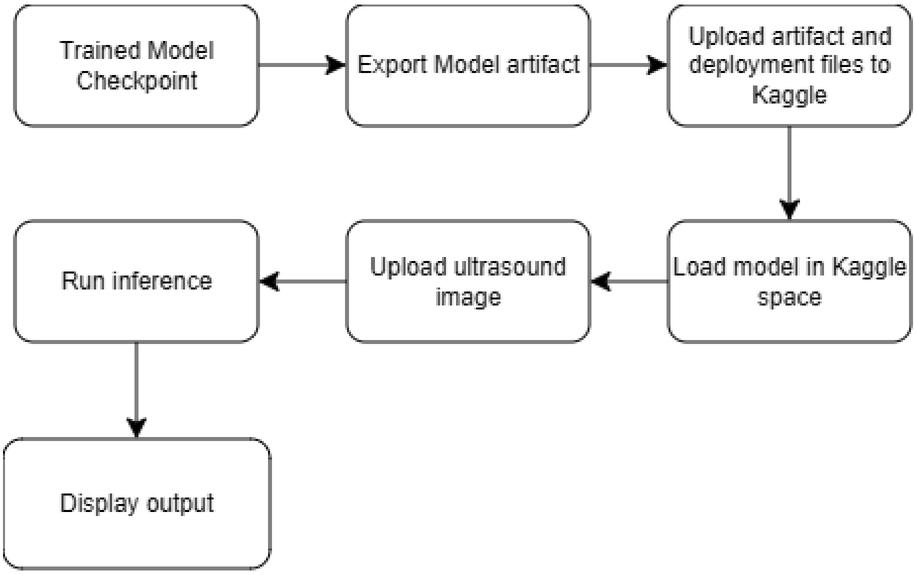


Figure 5.1: Cloud Deployment Prototype Implementation Pipeline.

The trained model checkpoint will be retrieved from the Google Drive. After that, the model artifact would be exported into model_artifact.pth to ensure the name consistency. The artifact stores the model state_dict and metadata such as project name, backbone type, pruning strategy, class names, image size, score threshold, and source weight path. This makes the model easier to load during deployment because the required inference configuration is stored together with the weights. Hugging face provides a platform called Hugging Face Space that allows researchers to upload models for online inference. Therefore, the next component is the model inference wrapper. This wrapper rebuilds the Faster R-CNN architecture, loads the saved model weights, preprocesses uploaded ultrasound images, performs inference, applies the score threshold, and returns predictions. Each prediction contains the detected class label, class ID, confidence score, and bounding-box coordinates. Lastly, when a user uploads a picture to the deployed model, the model runs the inference and displays the output in terms of the prediction label, bounding box, and confidence score.

The deployment prototype is intended only for FYP demonstration purposes. It is not a clinical diagnostic system and lacks production features such as user authentication, database storage, audit logging, model monitoring, or medical regulatory validation.

5.7 Implementation Issues and Challenges

Several implementation issues were encountered during the project. The first challenge was adapting the BUSI dataset for object detection. Since the dataset provides segmentation masks rather than bounding boxes, a mask-to-bounding-box conversion process had to be implemented.

The second challenge was integrating Grad-CAM with Faster R-CNN. Unlike simple classification networks, Faster R-CNN produces multiple predictions and includes several components such as the backbone, region proposal network, and detection head. Therefore, the Grad-CAM implementation had to capture gradients and activations from a suitable backbone layer. Therefore, the XAI visualisation quality would be weak as multiple attention required.

The third challenge was pruning Faster R-CNN. Structural pruning is a significant challenge on successful pruning as downstream components depend on feature dimensions. To avoid breaking the model architecture, unstructured pruning was used. Theoretically, this would not able to have an efficient parameter reduction and improved inference speed.

5.8 Concluding Remark

This chapter described the implementation of the proposed Faster R-CNN-based breast ultrasound lesion detection system. The implementation included dataset preparation, baseline detector training, Grad-CAM visualisation, pruning strategy implementation and model evaluation.

CHAPTER 6
RESULTS AND DISCUSSION

6.1 Experimental Setup

The experiment was conducted using Faster R-CNN with a ResNet50-FPN backbone. The BUSI dataset was divided into training, validation, and testing sets. The model was trained using breast ultrasound images with bounding-box annotations generated from the original segmentation masks.

The baseline Faster R-CNN model was trained for 30 epochs. After baseline training, pruning was applied to the selected target layer:

backbone.body.layer4[-1].conv3

Five pruning strategies were evaluated:

1. Random pruning
2. Activation-only pruning
3. Taylor-only pruning
4. Hybrid activation-Taylor pruning
5. Loss-guided XAI Grad-CAM-style relevance pruning

After pruning, each pruned model was fine-tuned for 10 additional epochs. The pruning ratios evaluated were 10%, 30%, 50%, and 70%. The primary evaluation metrics were mAP@0.50 and mAP@0.75. Precision, recall, F1-score, confusion matrix, parameter count, inference time, and Grad-CAM visualisations were also reported as secondary evaluation results. Besides, the pruning on the model was implemented as unstructured pruning on one selected layer. Therefore, the pruning experiment was designed mainly as a filter-importance ablation study, not as a major model compression experiment.

6.2 Baseline Detection Performance

The unpruned Faster R-CNN ResNet50-FPN model achieved reasonable detection performance on the BUSI test set. Across different runs, the baseline model obtained mAP@0.50 values around 0.70 to 0.74 and mAP@0.75 values around 0.53 to 0.64.

The result indicating the model was able to learn localisation features from the BUSI dataset. The mAP@0.5 score shows the detector able to locate abnormal regions with moderate overlap, meanwhile mAP@0.75 score is lower because it used a stricter localisation threshold, there the overlapping box must be over 75%. Besides, the baseline model achieved a high matched-detection classification score. Despite the scores should not directly compare with the pure classification models, the f1-score and accuracy is considered as high that it able to classify it correctly.

6.3 Grad-CAM Explainability Results

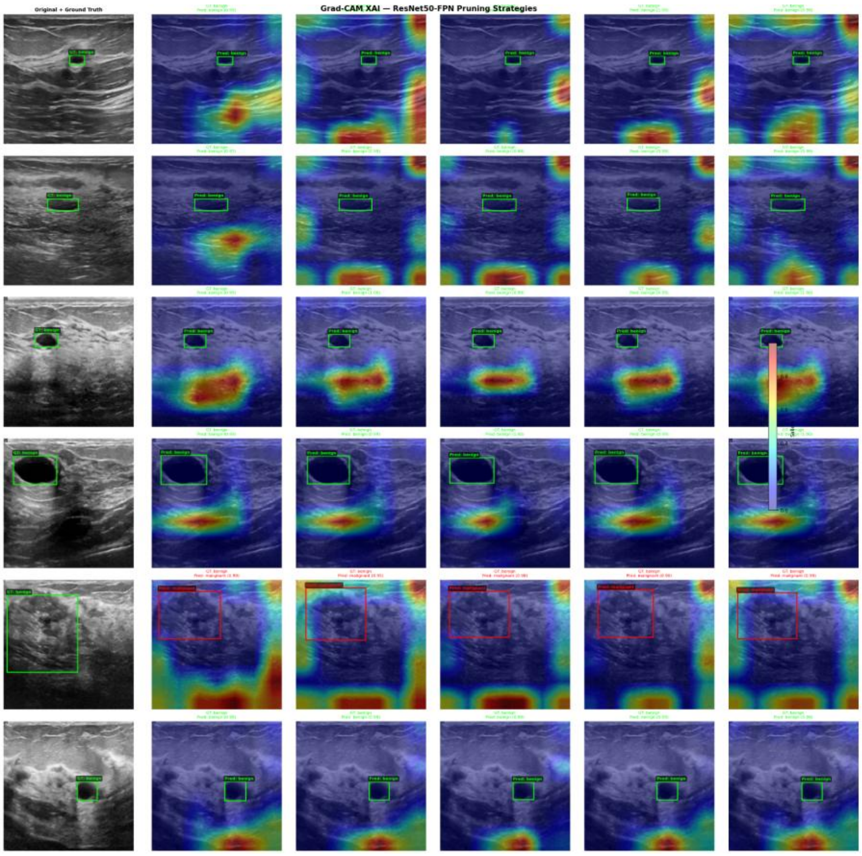


Figure 6.1: From left to right, the arrangement is original picture, baseline model, no pruning, activation-only, Taylor-only, Grad-CAM guided pruning and hybrid technique respectively.

Based on figure 6.1, the Grad-CAM heatmap visualisation did not provide strong explanatory value for this project. In the given figures, the highlighted attention regions are not within the predicted bounding box or the lesion area. Instead, the heatmap is focused on surrounding tissues regions or other irrelevant areas in the ultrasound image. As a result, the visualisation did not provide meaningful insight that the model was focusing to the actual lesion when making its prediction.

The limitation of the model that causing this issue might be due to the complexity of applying Grad-CAM to Faster R-CNN. Unlike standard classification CNN, Faster R-CNN is a two-stage object detector that produces multiple region proposals, class scores, and bounding-box predictions. Therefore, the importance score used by the Grad-CAM may not directly represent the actual final object localisation.

Therefore, the Grad-CAM should be only viewed as a supplementary visualisation support while evaluating the model prediction rationale as it is not focusing on the lesions area. The poor alignment between heatmaps and lesion bounding boxes suggests that standard Grad-CAM may not be fully suitable for explaining Faster R-CNN lesion detection outputs without further adaptation, such as proposal-specific Grad-CAM or detection-targeted saliency methods.

6.4 Pruning Strategy Ablation Results

The pruning strategy ablation compared the effect of different filter-importance criteria at the same target layer. Since all pruning strategies were applied to the same layer, the comparison mainly reflects how different scoring methods select filters for masking.

10% Pruning Result

At 10% pruning, the parameter reduction was approximately 0.25%. The results are shown below.

Strategy	**mAP@0.50**	**mAP@0.75**	**F1-score**	**Params**	**ms per image**
No pruning	0.7390	0.5991	0.9430	41,309,411	24.52
Random pruning	**0.7595**	0.5938	0.9503	41,204,555	23.71
Activation-only pruning	0.7209	0.5842	0.9302	41,204,555	23.58
Taylor-only pruning	0.7300	**0.6254**	**0.9604**	41,204,555	23.46
XAI Grad-CAM relevance pruning	0.6901	0.5398	0.9494	41,204,555	23.53
Hybrid activation-Taylor pruning	0.7450	0.5833	0.9602	41,204,555	24.09

Table 6.1: 10% Pruning Result with Both Detection and Classification Metrics.

From such outcome, Taylor-only pruning produced the strongest overall performance in this result. Compared with the no-pruning baseline, mAP@0.50 decreased slightly from 0.7390 to 0.7300. However, it improved the stricter localisation metric, where mAP@0.75 increased from 0.5991 to 0.6254. It also achieved the highest F1-score of 0.9604, compared with 0.9430 for the baseline. This indicates that Taylor-based pruning may be more effective in preserving filters that are important for accurate localisation and classification.

Hybrid activation-Taylor pruning also performed competitively. It achieved an mAP@0.50 of 0.7450, which is higher than the no-pruning baseline, and obtained a strong F1-score of 0.9602, almost similar to Taylor-only pruning. However, its mAP@0.75 decreased to 0.5833. This suggests that hybrid pruning can maintain general detection performance well, but it may be less effective than Taylor-only pruning when stricter bounding-box localisation is required.

Random pruning showed unexpectedly strong performance, reaching the highest mAP@0.50 at 0.7595 and improving the F1-score to 0.9503. However, its mAP@0.75 was slightly lower than the baseline at 0.5938. This means that random pruning detected lesions well under a looser IoU threshold, but it did not improve precise localisation.

Activation-only pruning showed a moderate decrease in performance. Its mAP@0.50 dropped to 0.7209, mAP@0.75 dropped to 0.5842, and F1-score decreased to 0.9302. This suggests that activation magnitude alone may not be a reliable indicator for selecting filters to prune, because a filter with low activation is not necessarily unimportant to the final detection result.

XAI Grad-CAM relevance pruning produced the weakest detection performance in terms of mAP, with mAP@0.50 decreasing to 0.6901 and mAP@0.75 decreasing to 0.5398. Although its F1-score remained relatively high at 0.9494, the lower mAP values suggest that Grad-CAM-based relevance may preserve some class-related information but does not sufficiently preserve localisation-sensitive features.

The parameter counts only decreased from 41,309,411 to 41,204,555, which is approximately a 0.25% reduction. Inference time also improved only slightly, with the fastest result being 23.46 ms per image for Taylor-only pruning compared with 24.52 ms per image for the baseline. This shows that the pruning did not remove a large portion of the full network. Instead, it mainly affected the selected pruning layer, so the improvement in model size and speed remained limited.

Overall, Taylor-only pruning was the best-performing strategy in this result because it achieved the highest mAP@0.75, the highest F1-score, and the fastest inference time. Hybrid activation-Taylor pruning was also a strong alternative, especially when considering mAP@0.50 and F1-score. However, the overall improvement should be interpreted carefully because the parameter reduction was small and the pruning was not a full structural compression of the entire model.

30% Pruning Result

At 30% pruning, 614 filters were masked from the target layer, resulting in approximately 0.76% effective parameter reduction.

Strategy	**mAP@0.50**	**mAP@0.75**	**F1-score**	**Params**	**ms per image**
No pruning	**0.7390**	0.5991	0.9430	41,309,411	24.52
Random pruning	0.7247	0.5731	0.9420	40,993,815	24.52
Activation-only pruning	0.7042	0.5673	0.9135	40,993,815	24.11
Taylor-only pruning	0.6854	0.5013	0.9364	40,993,815	24.10
XAI Grad-CAM relevance pruning	0.7115	0.5505	0.9474	40,993,815	24.31
Hybrid activation-Taylor pruning	0.7143	**0.6044**	**0.9490**	40,993,815	24.15

Table 6.2: 30% Pruning Result with Both Detection and Classification Metrics.

The hybrid activation-Taylor pruning strategy provides the best overall balance in this result. Although its mAP@0.50 decreases from 0.7390 to 0.7143, it slightly improves both mAP@0.75 and F1-score compared with the unpruned model. The mAP@0.75 increases from 0.5991 to 0.6044, while the F1-score improves from 0.9430 to 0.9490. This suggests that the hybrid method is able to remove less important filters while still preserving stricter localisation quality and overall classification balance.

Random pruning also performs reasonably well, especially for mAP@0.50. It achieves 0.7247, which is the highest mAP@0.50 among the pruned models. However, its mAP@0.75 decreases to 0.5731, showing that its localisation accuracy becomes weaker when a stricter IoU threshold is applied. This indicates that random pruning may still retain general detection ability, but it is less dependable for precise bounding-box prediction.

Activation-only pruning shows a more noticeable reduction in performance. Its mAP@0.50 decreases to 0.7042, while its F1-score drops to 0.9135, which is the lowest among all strategies. This suggests that activation magnitude alone may not be sufficient for deciding which filters should be removed, since filters with lower activation may still contribute to detection confidence or class prediction.

Taylor-only pruning gives the weakest detection performance in this result, especially for stricter localisation. Its mAP@0.75 decreases sharply to 0.5013. This indicates that Taylor-based sensitivity alone may be unstable in this pruning setting, particularly when the model needs to maintain accurate bounding-box localisation.

XAI Grad-CAM relevance pruning performs better than activation-only and Taylor-only pruning in terms of F1-score. It achieves an F1-score of 0.9474, which is slightly higher than the unpruned baseline. However, its mAP@0.75 decreases to 0.5505. This suggests that Grad-CAM-based relevance may help preserve class-discriminative features, but it does not fully preserve localisation-sensitive features required for accurate bounding-box prediction.

An important observation is that the parameter count only decreases from 41.31 million to 40.99 million, which is approximately a 0.76% reduction. This shows that the 30% pruning ratio does not represent a 30% reduction of the entire model. Instead, pruning is applied only to the selected target layer. The inference time also improves only slightly, from 24.52 ms per image to around 24.10-24.31 ms per image. This confirms that unstructured pruning provides limited practical acceleration unless the model architecture is physically compressed and optimized for deployment.

Overall, hybrid activation-Taylor pruning is the most effective 30% pruning strategy in this result. It achieves the highest mAP@0.75 and F1-score while maintaining competitive mAP@0.50 and slightly improving inference time. However, the improvement should be interpreted carefully because the global parameter reduction is small and the pruning method does not perform full structural compression of the network.

50% Pruning Result

At 50% pruning, the effective parameter reduction was approximately 1.27%.

Strategy	mAP@0.50	mAP@0.75	F1-score	Params	ms per image
No pruning	0.7390	**0.5991**	0.9430	41,309,411	24.52
Random pruning	0.6895	0.5422	0.9267	40,783,075	24.44
Activation-only pruning	0.7053	0.5379	0.9396	40,783,075	23.72
Taylor-only pruning	**0.7413**	0.5942	**0.9491**	40,783,075	24.29
XAI Grad-CAM relevance pruning	0.6874	0.5047	0.9486	40,783,075	24.30
Hybrid activation-Taylor pruning	0.6614	0.5346	0.9308	40,783,075	24.30

Table 6.3: 50% Pruning Result with Both Detection and Classification Metrics.

Compared with the unpruned baseline, Taylor-only pruning gives the strongest performance in this result. Its mAP@0.50 slightly increases from 0.7390 to 0.7413, while its F1-score also improves from 0.9430 to 0.9491. Although its mAP@0.75 decreases slightly from 0.5991 to 0.5942, the drop is small compared with the other pruning strategies. This suggests that Taylor-based importance estimation is the most reliable pruning criterion in this setting because it preserves detection performance and localisation accuracy more effectively.

Activation-only pruning provides moderate performance. Its mAP@0.50 decreases to 0.7053, and its mAP@0.75 decreases to 0.5379. However, its F1-score remains close to the baseline at 0.9396. This indicates that activation-based pruning can still preserve matched-detection classification balance reasonably well, but it weakens both detection accuracy and localisation precision.

Random pruning performs weaker than activation-only pruning in this result. Its mAP@0.50 decreases to 0.6895, while its mAP@0.75 decreases to 0.5422. Its F1-score also drops to 0.9267. This shows that, at this pruning level, random filter removal is less stable and may discard useful model capacity without considering feature importance.

XAI Grad-CAM relevance pruning maintains a relatively high F1-score of 0.9486. However, both mAP scores decrease, especially mAP@0.75, which falls to 0.5047. This suggests that Grad-CAM-based relevance may help preserve some class-discriminative information, but it does not sufficiently preserve localisation-sensitive features required for accurate bounding-box prediction.

Hybrid activation-Taylor pruning performs the weakest overall in this result. It obtains the lowest mAP@0.50 at 0.6614, and its mAP@0.75 also decreases to 0.5346. Although its F1-score remains acceptable at 0.9308, the larger reduction in mAP suggests that combining activation and Taylor scores did not improve filter selection in this specific pruning setting.

The parameter counts decreases from 41,309,411 to 40,783,075, which is approximately a 1.27% reduction in total parameters. The inference speed improvement is also limited. The fastest result is achieved by activation-only pruning at 23.72 ms per image, compared with 24.52 ms per image for the unpruned model. This shows that the pruning does not provide major acceleration because it is applied only to a selected layer and does not structurally compress the whole model.

Overall, Taylor-only pruning is the best-performing strategy for this pruning result. It preserves detection accuracy most effectively, slightly improves mAP@0.50 and F1-score, and only produces a small decrease in mAP@0.75. The other pruning strategies show larger performance losses, especially under the stricter localisation metric.

70% Pruning Result

At 70% pruning, the effective parameter reduction was approximately 1.78%.

Strategy	mAP@0.50	mAP@0.75	F1-score	Params	ms per image
No pruning	**0.7390**	0.5991	0.9430	41,309,411	24.52
Random pruning	0.7076	0.5891	0.9378	40,572,849	24.71
Activation-only pruning	0.7079	**0.6083**	0.9170	40,572,849	25.20
Taylor-only pruning	0.7311	0.5708	0.9235	40,572,849	24.85
XAI Grad-CAM relevance pruning	0.6955	0.5668	0.9479	40,572,849	24.46
Hybrid activation-Taylor pruning	0.7164	0.5634	**0.9593**	40,572,849	25.10

Table 6.4: 70% Pruning Result with Both Detection and Classification Metrics.

Compared with the unpruned baseline, all pruned models reduce the parameter count from 41,309,411 to 40,572,849, which is approximately a 1.78% reduction in total parameters. However, the inference time does not improve consistently. Only XAI Grad-CAM relevance pruning is slightly faster, decreasing from 24.52 ms per image to 24.46 ms per image. The other pruned models are slower than the baseline, with activation-only pruning showing the slowest inference time at 25.20 ms per image. This confirms that unstructured pruning does not necessarily produce practical acceleration.

In terms of detection performance, activation-only pruning achieves the highest mAP@0.75 at 0.6083, slightly higher than the unpruned baseline of 0.5991. This suggests that activation-based pruning can preserve stricter localisation accuracy in this result. However, its F1-score decreases to 0.9170, indicating weaker balance between precision and recall.

Hybrid activation-Taylor pruning achieves the highest F1-score at 0.9593, which is higher than the baseline F1-score of 0.9430. However, its mAP@0.50 and mAP@0.75 decrease to 0.7164 and 0.5634. This indicates that the hybrid method may preserve matched-detection classification performance, but it does not preserve overall detection and localisation performance as effectively.

Taylor-only pruning preserves mAP@0.50 best among the pruned methods, reaching 0.7311, which is only slightly below the baseline value of 0.7390. However, its mAP@0.75 decreases to 0.5708, and its F1-score drops to 0.9235. Therefore, Taylor-only pruning remains relatively stable for general detection, but it is less effective for stricter localisation in this result.

Random pruning gives moderate performance, with mAP@0.50 of 0.7076, mAP@0.75 of 0.5891, and F1-score of 0.9378. Its mAP@0.75 remains close to the baseline, but its mAP@0.50 still decreases, showing that random pruning reduces detection performance overall.

XAI Grad-CAM relevance pruning achieves the fastest inference speed and maintains a great F1-score of 0.9479. However, its mAP values are lower, especially mAP@0.50 at 0.6955. This suggests that Grad-CAM-based pruning may retain some class-relevant features, but it weakens detection coverage and localisation quality.

Overall, there is no single dominant pruning method in this result. If the priority is stricter localisation, activation-only pruning performs best because it achieves the highest mAP@0.75. If the priority is matched-detection F1-score, hybrid activation-Taylor pruning performs best. If the priority is general detection stability, Taylor-only pruning is the most stable among the pruned methods because it retains the highest mAP@0.50. However, none of the pruning strategies provides meaningful speed improvement, so this result should be interpreted as pruning behavior analysis rather than practical model acceleration.

6.5 Pruning ratio Trade-Off Analysis

Based on the actual number of parameters reduced, the pruning effect is relatively small compared with the original model size of 41,309,411 parameters.

Pruned Params	Param Reduced	Actual Reduction	Best Overall Strategy	Main Observation
41,204,555	104,856	~0.25%	Taylor-only pruning	Best accuracy-speed balance. Highest mAP@0.75, F1, and fastest inference.
40,993,815	315,596	~0.76%	Hybrid activation-Taylor pruning	Best F1 and mAP@0.75; good balance despite lower mAP@0.50.
40,783,075	526,336	~1.27%	Taylor-only pruning	Most stable. It preserves baseline accuracy with slight F1 improvement.

40,572,749	736,562	~1.78%	Mixed result	Higher pruning causes unstable results. No clear dominant method.

Table 6.5: Pruning Result Ratio Trade-off Analysis Table.

Overall, the results indicate that increasing parameter reduction does not produce a consistent improvement in inference speed. Although the number of parameters decreases from 41.31M to 40.57M, the speed improvement is small and sometimes negative. For example, at the largest reduction level, some pruned models are slower than the unpruned baseline.

In terms of accuracy, moderate pruning gives the optimal balance. The 0.25% reduction level performs especially well, where Taylor-only pruning improves mAP@0.75 from 0.5991 to 0.6254, improves F1-score from 0.9430 to 0.9604, and reduces inference time from 24.52 ms per image to 23.46 ms per image. This suggests that a small amount of pruning may remove redundant parameters and improve generalization.

At elevated reductions, the performance becomes less stable. The 1.27% reduction level still works well with Taylor-only pruning, but the 1.78% reduction level shows clearer trade-offs: some methods improve F1-score or mAP@0.75, but most lose mAP@0.50 and do not improve speed.

Consequently, through empirical parameter reduction, the best pruning trade-off is achieved at the lowest-to-moderate reduction range, especially using Taylor-only pruning. Larger parameter reduction does not inherently result in better efficiency and may cause accuracy instability.

6.6 Efficiency and Parameter Analysis

By referring to Table 6.5. The parameter reduction from pruning was small because only one layer was pruned. The inference speed from Table 6.1 to Table 6.5 improvement was also minimal and inconsistent. In several runs, pruned models had similar or slightly slower inference times than the baseline. This is expected because unstructured pruning sets filters to zero but does not physically remove channels from the model architecture. Therefore, the pruning component should not be presented as a major compression result. A more accurate interpretation is that pruning provides a controlled experiment to compare filter-importance criteria and their effect on detection performance.

6.7 Comparison with Related Work

Many breast ultrasound deep learning studies report high classification accuracy. However, these studies are usually image-level classification or region-of-interest classification studies. Their results are not directly comparable with this project because this project evaluates object detection and localisation.

6.8 Deployment of Model on Hugging Face

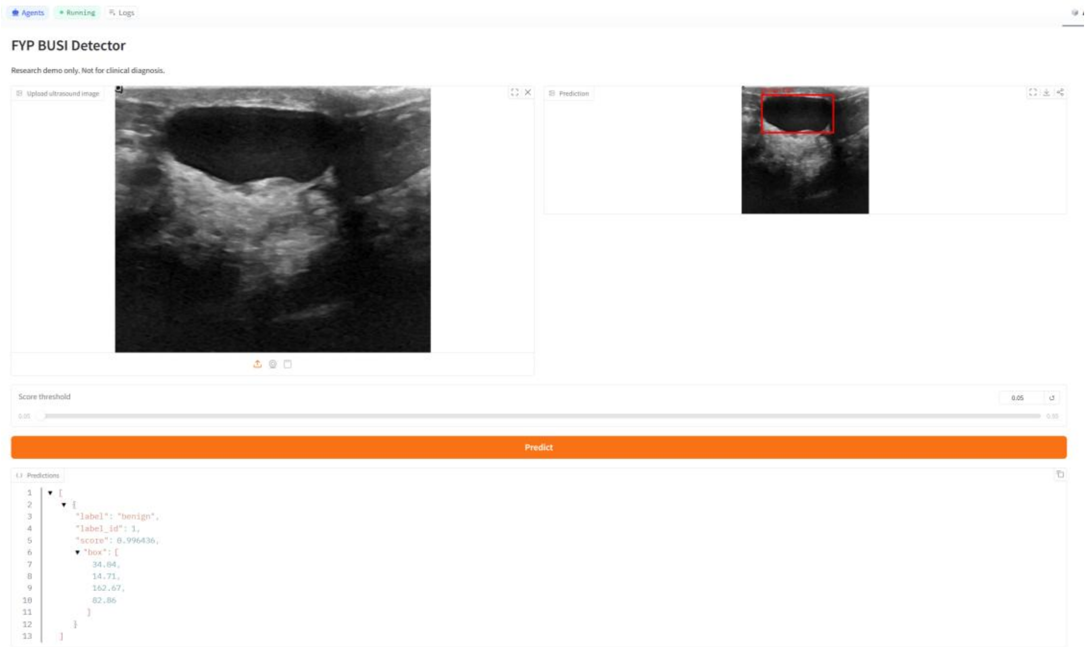


Figure 6.2: Deployment of the Model on Hugging Face Space.

The Figure 6.2 shows the model deployment on the Hugging Face Space. The deployment on the Hugging Face Space shows the feasibility of the model could be publicly deployed. A random ultrasound picture is uploaded to the interface as input, clicked predict button and the label, confidence score and the bounding box annotation are shown to the user. As this is not for clinically deployment, the deployment on Hugging Face confirms that the model usability for demonstration purpose. This supports the practical feasibility of presenting the proposed Faster R-CNN detection system as cloud accessible prototype.

6.9 Limitations

Several limitations were identified in the experiment. Firstly, the Grad-CAM visualisation was not consistently producing clinically convincing heatmaps. Although Grad-CAM was integrated to provide explainability, the heatmaps appeared to be noisy and highlighted the surrounding tissue rather than the actual lesion. This limitation may be caused by the complexity of applying Grad-CAM to Faster R-CNN, where the model produces multiple region proposals and predictions rather than a single classification score that affects the Grad-CAM importance score. Second, the pruning method was implemented as unstructured pruning on only one selected layer. Although this approach allowed controlled comparison between different filter-importance criteria, it did not physically remove channels from the network architecture. Therefore, the global parameter reduction was small, and inference speed improvements were minimal or inconsistent. Because of these limitations, the pruning results should not be presented as strong model compression results. Instead, it should be interpreted as a filter-importance ablation study.

6.10 Objectives Evaluation

The first objective was to develop a Faster R-CNN-based breast ultrasound lesion detection model. This objective was partially achieved. A Faster R-CNN ResNet50-FPN detection model was developed using the BUSI dataset, with segmentation masks converted into bounding boxes for lesion localisation. The model achieved reasonable mAP@0.50 and mAP@0.75 performance. However, the model should only be considered lightweight to a limited extent because pruning was masked and applied to one layer, giving small parameter reduction and minimal speed improvement.

The second objective is to address the CNN black-box nature through XAI visualisation techniques. This objective was achieved as a qualitative explainability component. Grad-CAM was integrated to generate heatmaps showing regions that influenced the model prediction. However, some heatmaps were broad, noisy, or not well aligned with lesion regions, so the visualisation should be treated as supporting explanation rather than clinical proof.

The third objective was to implement and compare various pruning strategies' effectiveness based on existing and novel pruning techniques. This objective was achieved. Random, activation-only, Taylor-only, hybrid activation-Taylor, and loss-guided XAI Grad-CAM-style pruning were implemented and compared. The results showed that some methods preserved performance under moderate pruning, but none consistently outperformed the unpruned baseline. The pruning study is best interpreted as a filter-importance ablation, not major model compression.

The fourth objective was to deploy the model on a cloud-based web platform to demonstrate accessibility and user-friendly usage for clinicians. This objective was partially achieved. A cloud-deployable prototype was prepared using model artifact export, inference wrapper, and simple deployment support through Kaggle. The prototype demonstrates inference feasibility but is not a production clinical platform because it lacks security, database storage, permanent hosting, scalability testing, and clinical validation.

Overall, the project successfully developed a detection, XAI, pruning, and deployment demonstration pipeline, with limitations in model compression strength, Grad-CAM quality, and production-level deployment.

6.11 Concluding Remark

This chapter presented the experimental results and discussion of the proposed Faster R-CNN breast ultrasound detection system. The baseline model achieved reasonable lesion detection performance, while Grad-CAM provided a visual explanation of model attention.

The pruning experiments showed that moderate pruning can preserve detection performance in some cases, especially for hybrid activation-Taylor pruning at 30% and Taylor-only pruning at 50%. However, pruning did not consistently outperform the unpruned baseline across all metrics. The parameter and speed improvements were small because pruning was masked and applied only to one selected layer.

Therefore, the pruning component is best interpreted as a controlled ablation of filter-importance criteria rather than a major model compression method.

CHAPTER 7
CONCLUSION AND RECOMMENDATION

7.1 Conclusion

This project developed a Faster R-CNN-based breast ultrasound lesion detection system using the BUSI dataset. Unlike image-level classification approaches, the proposed system focused on object detection and localisation by predicting both class labels and bounding boxes. Segmentation masks from the BUSI dataset were converted into bounding boxes so that the dataset could be used for Faster R-CNN training.

The main detection model used in this project was Faster R-CNN with a ResNet50-FPN backbone. The baseline model achieved reasonable detection performance, with mAP@0.50 generally around 0.70 to 0.74 and mAP@0.75 around 0.53 to 0.64 across experimental runs. These results show that the model was able to learn useful localisation patterns from breast ultrasound images.

Grad-CAM visualisation was integrated to improve interpretability. The generated heatmaps allowed visual inspection of the regions that influenced model predictions. This helped support qualitative analysis of whether the model focused on lesion regions or surrounding image areas.

The project also implemented and compared several pruning strategies, including random pruning, activation-only pruning, Taylor-only pruning, hybrid activation-Taylor pruning, and loss-guided XAI Grad-CAM-style relevance pruning. The pruning experiment was conducted on the selected layer backbone.body.layer4[-1].conv3. Since pruning was masked and applied only to one layer, the global parameter reduction was small, ranging from approximately 0.25% to 1.78%, and inference speed improvement was minimal.

The pruning results showed that moderate pruning could preserve detection performance in selected cases. At 30% pruning, hybrid activation-Taylor pruning achieved competitive performance and slightly improved mAP@0.75 compared with the unpruned baseline in that run. At 50% pruning, Taylor-only pruning preserved mAP@0.50 well among the pruned methods. However, the unpruned baseline remained strongest for stricter localisation or matched-detection classification metrics in several cases. Aggressive pruning at 70% produced fewer stable results and often reduced detection performance.

Overall, the project achieved its aim of developing a breast ultrasound detection system with explainability and pruning analysis. The pruning component should be interpreted as a controlled ablation study of filter-importance criteria rather than a major model compression method.

7.2 Project Limitations

There are several weaknesses that are identified in this project. Firstly, the pruning techniques implemented in this project are unstructured pruning rather than structural pruning. unstructured pruning is a type of pruning that only lets the selected filters zero out but does not physically remove the channel from the model architecture. In result the parameter reduction and improvement of inference speed were very limited. The pruning was applied only to one selected layer. Therefore, the project cannot claim large-scale model compression or deployment-level acceleration, as Faster R-CNN is architecture is difficult to prune. Lastly, the experiment was only run once on a small dataset. The model that trained using a small dataset is prone to the noise. A slight change in the experimental environment could change the entire model results.

7.3 Recommendations and Future Work

The enhancement of this project could be explored from various perspectives. Structural pruning on Faster R-CNN could be explored together with the proposed pruning technique. Performing pruning on the standard Faster R-CNN is difficult, as the downstream components, such as FPN and RoI, require the correct number of dimensions to pass forward for detection. Unlike unstructured pruning, structural pruning would physically remove filters or channels from the network and thus achieve a better parameter reduction and significant inference speed improvement.

Additionally, the experiment could be repeated using multiple random seeds. This would improve result reliability and help determine whether observed improvements are consistent or due to randomness. Small dataset statistics are not reliable enough and require multiple runs as other people work on the BUSI Dataset for detection accuracy. Multiple fixed-seed runs would significantly enhance the reliability of the work.

Last but not least, quantitative explainability evaluation could be added. For example, Grad-CAM heatmaps could be compared with lesion masks using overlap-based metrics. This would make the explainability analysis stronger than qualitative visual inspection alone.

REFERENCES

[1] D. Bhati, F. Neha, M. Amiruzzaman, "A Survey on Explainable Artificial Intelligence (XAI) Techniques for Visualizing Deep Learning Models in Medical Imaging," vol. 10, no. 10, p. 239, 2024.

[2] M. Bagewadi, B. M. Reshmi, N. Kalligud, K. Badiger, S. Sonagaj and S. Giraddi, "Smartsight: Android-Based Retinal Disease Detection using Retinal Fundus Images with XAI," in 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2024.

[3] A. Chaddad, Y. Hu, Y. Wu, B. Wen, R. Kateb, "Generalizable and explainable deep learning for medical image computing: An overview," Current Opinion in Biomedical Engineering, vol. 33, no. 100567, pp. 1-12, 2025.

[4] G. H. Dagnaw, M. E. Mouhtadi, M. Mustapha, "Skin cancer classification using vision transformers and explainable artificial intelligence," Journal of Medical Artificial Intelligence, vol. 7, 2024.

[5] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both Weights and Connections for Efficient Neural Networks," Oct. 2015. Available: https://arxiv.org/pdf/1506.02626

[6] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning Convolutional Neural Networks for Resource Efficient Inference," *ICLR 2017*, Jun. 2017, Accessed: May 04, 2026. [Online]. Available: https://arxiv.org/abs/1611.06440

[7] C. Liu, B. Yang, L. Zhang, C. Liu, and W. Zheng, "Breast Ultrasound Image Detection Based on Dual-Branch Faster R-CNN," Electronics, vol. 14, no. 21, p. 4247, Oct. 2025, doi: https://doi.org/10.3390/electronics14214247.

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

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localisation," International

Journal of Computer Vision, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: https://doi.org/10.1007/s11263-019-01228-7.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: https://doi.org/10.1109/tpami.2016.2577031.

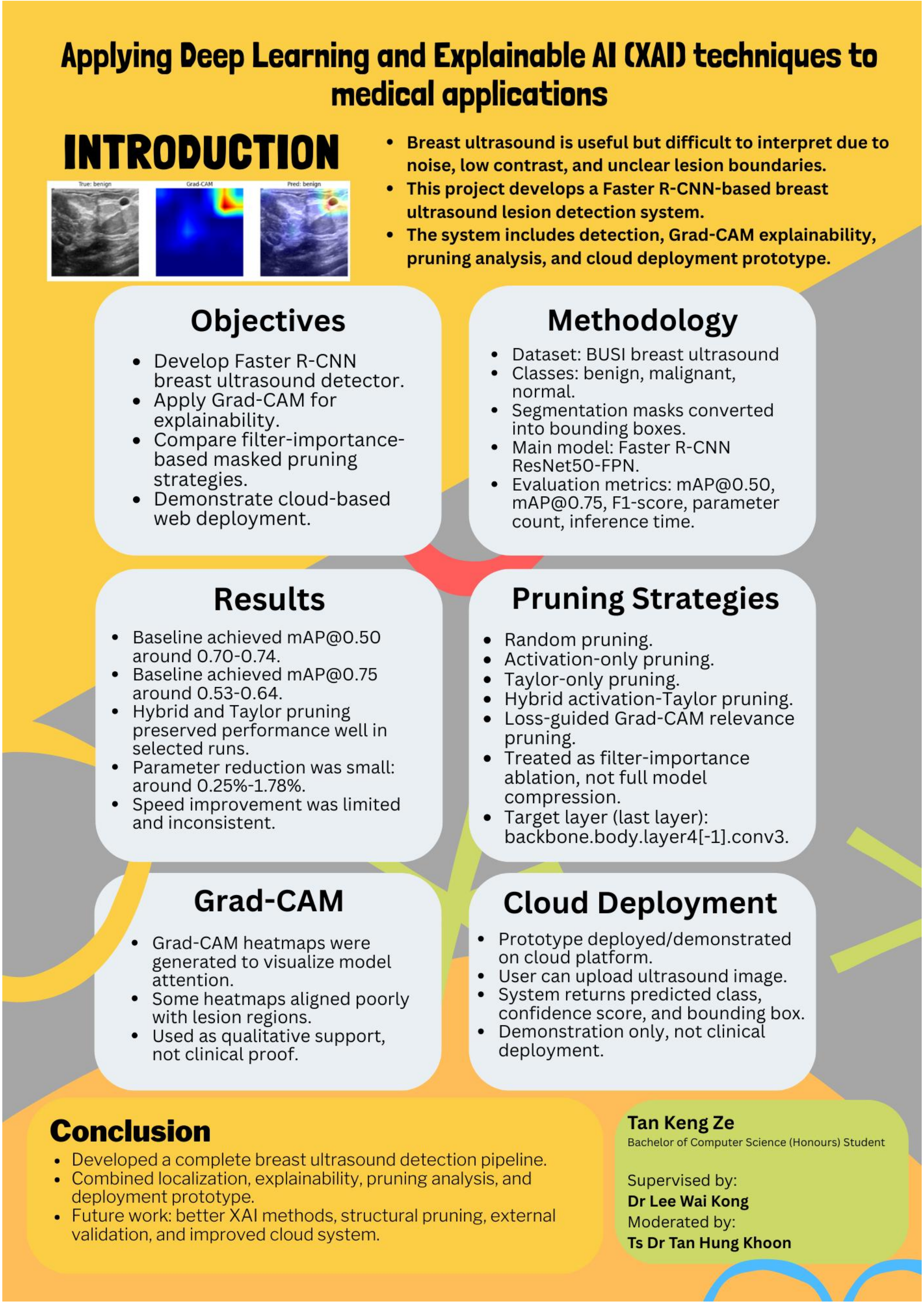
[11] E. Miller, "Layer-Aware Models: Enhancing Classification Accuracy with Neural Pruning - Magnimind Academy," *Magnimind Academy - Launch a new career with our programs*, Jan. 31, 2025. https://magnimindacademy.com/blog/layer-aware-models-enhancing-classification-accuracy-with-neural-pruning/#How_Layer-Aware_Models_Enhance_Classification_Accuracy (accessed May 04, 2026).

[12] A. Sharma, "Mean Average Precision (mAP) Using the COCO Evaluator," *PyImageSearch*, May 02, 2022. https://pyimagesearch.com/2022/05/02/mean-average-precision-map-using-the-coco-evaluator/ (accessed May 04, 2026).

[13] L. Shen et al., "Fusing attention mechanism with Mask R-CNN for instance segmentation of grape cluster in the field," Frontiers in Plant Science, vol. 13, Jul. 2022, doi: https://doi.org/10.3389/fpls.2022.934450.

[14] M. H. Siddiqui, "Difference Between Structured and Unstructured Pruning in Neural," Medium, Mar. 22, 2024. https://medium.com/@mhammadsiddiqui/difference-between-structured-and-unstructured-pruning-in-neural-cca5603581fb (accessed May 04, 2026).

POSTER


Applying Deep Learning and Explainable AI (XAI) techniques to medical applications
INTRODUCTION
True: benign
Grad-CAM
Pred: benign
• Breast ultrasound is useful but difficult to interpret due to noise, low contrast, and unclear lesion boundaries.
• This project develops a Faster R-CNN-based breast ultrasound lesion detection system.
• The system includes detection, Grad-CAM explainability, pruning analysis, and cloud deployment prototype.
Objectives
• Develop Faster R-CNN breast ultrasound detector.
• Apply Grad-CAM for explainability.
• Compare filter-importance-based masked pruning strategies.
• Demonstrate cloud-based web deployment.
Methodology
• Dataset: BUSI breast ultrasound
• Classes: benign, malignant, normal.
• Segmentation masks converted into bounding boxes.
• Main model: Faster R-CNN ResNet50-FPN.
• Evaluation metrics: mAP@0.50, mAP@0.75, F1-score, parameter count, inference time.
Results
• Baseline achieved mAP@0.50 around 0.70-0.74.
• Baseline achieved mAP@0.75 around 0.53-0.64.
• Hybrid and Taylor pruning preserved performance well in selected runs.
• Parameter reduction was small: around 0.25%-1.78%.
• Speed improvement was limited and inconsistent.
Pruning Strategies
• Random pruning.
• Activation-only pruning.
• Taylor-only pruning.
• Hybrid activation-Taylor pruning.
• Loss-guided Grad-CAM relevance pruning.
• Treated as filter-importance ablation, not full model compression.
• Target layer (last layer): backbone.body.layer4[-1].conv3.
Grad-CAM
• Grad-CAM heatmaps were generated to visualize model attention.
• Some heatmaps aligned poorly with lesion regions.
• Used as qualitative support, not clinical proof.
Cloud Deployment
• Prototype deployed/demonstrated on cloud platform.
• User can upload ultrasound image.
• System returns predicted class, confidence score, and bounding box.
• Demonstration only, not clinical deployment.
Conclusion
• Developed a complete breast ultrasound detection pipeline.
• Combined localization, explainability, pruning analysis, and deployment prototype.
• Future work: better XAI methods, structural pruning, external validation, and improved cloud system.
Tan Keng Ze
Bachelor of Computer Science (Honours) Student
Supervised by:
Dr Lee Wai Kong
Moderated by:
Ts Dr Tan Hung Khoon