Independent University **Bangladesh (IUB)**

IUB Academic Repository

Computer Science and Engineering | Undergraduate Thesis

2026-08

Benchmarking CNN architectures using transfer learning for Lipomatous tumor classification: a comprehensive study

Haque, Hasibul

Independent University, Bangladesh (IUB)

https://ar.iub.edu.bd/handle/11348/1567

Downloaded from IUB Academic Repository



Benchmarking CNN Architectures Using Transfer Learning for Lipomatous Tumor Classification: A Comprehensive Study

Date: August 2026

Prepared by:

Hasibul Haque
ID: 2111498

Fahmida Jamil
ID: 2111288

Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by:

Dr. Saadia Binte Alam
Associate Professor

Department of Computer Science and Engineering
Independent University, Bangladesh

A Final Year Design Project (FYDP) Report submitted for the Degree of Bachelor's of Computer Science and Engineering

Attestation

We are aware of the fact that plagiarism is strictly prohibited and it conflicts with our university's rules and regulations. We guarantee the authenticity of our work. We have used some copyrighted material and models of others in our project which has been properly cited following the international standards proper guidelines.

1. Author Name: Hasibul Haque

ID: 2111498

Signature: .....

2. Author Name: Fahmida Jamil

ID: 2111288

Signature: .....

Evaluation Committee

Supervisor

Name: _____ Signature: _____

Internal Examiner 1

Name: _____ Signature: _____

Internal Examiner 2

Name: _____ Signature: _____

External Examiner

Name: _____ Signature: _____

Acknowledgement

We would like to express our sincere gratitude to our supervisor, Dr. Saadia Binte Alam, for her invaluable guidance, continuous encouragement, and patient support throughout this project. Her constructive feedback and thoughtful advice helped us stay on the right track at every stage of the work.

We are also grateful to Dr. Md Rashedur Rahman and Dr. Ashraful Islam for their helpful comments and suggestions, which greatly improved both the quality and the academic rigour of this research.

We would like to thank the Center for Computational and Data Sciences (CCDS) at Independent University, Bangladesh, for providing the computational resources that were essential to carrying out the experiments in this study.

We also want to acknowledge the technical support provided by Md. Zafor Iqbal, Mr. Siam Tahsin Bhuiyan, and Redwan Rahman. Their assistance and contributions at various points helped move this project forward in a meaningful way.

Finally, we would like to note that Large Language Models (LLMs), specifically Claude Opus 4.6 and Claude Haiku 4.5 from Anthropic, and Gemini 3 Pro from Google, were used during the writing process to help improve the clarity and language of the text. These tools were used strictly for editorial and writing assistance. The research methodology, hypotheses, data analysis, result generation, and all figures were produced entirely by the author.

List of Publications

M. Z. Iqbal, **H. Haque**, **F. Jamil**, R. Rahman, S. T. Bhuiyan, R. Rahman, A. Islam, and S. B. Alam, "Benchmarking CNN Architectures Using Transfer Learning for Lipomatous Tumor Classification: A Comprehensive Study," in Advances and Trends in Artificial Intelligence. Theory and Applications, Singapore: Springer Nature Singapore, 2026, doi: 10.1007/978-981-92-2891-1_14.

Abstract

Distinguishing between benign lipoma and well-differentiated liposarcoma (WDLPS) is a clinically important and challenging diagnostic problem. Both conditions appear similar on standard magnetic resonance imaging (MRI), yet confusing them carries serious consequences. A missed liposarcoma may progress or dedifferentiate into a more aggressive form, while unnecessary surgery for a benign lipoma exposes patients to avoidable risk. The current gold standard for definitive diagnosis is MDM2 gene amplification testing through biopsy, which is invasive and resource-intensive. There is therefore a strong motivation to develop accurate, non-invasive automated classification tools that can assist radiologists in making this distinction from T1-weighted MRI alone.

This study presents a systematic comparative benchmark of five pre-trained Convolutional Neural Network (CNN): ResNet50, DenseNet121, EfficientNetB3, InceptionV3, and MobileNetV2. All models were evaluated under a strictly uniform frozen transfer learning framework using T1-weighted MRI data from the publicly available WORC database, comprising scans from 115 patients with confirmed lipoma or WDLPS diagnoses. The 3D MRI volumes were preprocessed into 17,986 individual 2D axial slices and divided into training (70%), validation (15%), and test (15%) subsets using stratified sampling. Every model shared the same frozen ImageNet pre-trained backbone, GlobalMaxPooling2D aggregation layer, 128-neuron sigmoid classification head, Adam optimiser settings, and training callbacks, ensuring that performance differences are attributable solely to architectural design.

The results reveal a striking performance disparity. MobileNetV2 ranked first on every evaluation metric with 0.967 accuracy, precision 0.965, recall 0.966, F1-score 0.965, specificity 0.964, and AUROC 0.997. This is particularly notable because MobileNetV2 is the smallest model evaluated, with only 3.5 million parameters. DenseNet121 ranked second with 0.963 accuracy and the second highest WDLPS recall of 0.975, making it the safest choice for minimising missed malignancies. ResNet50 and InceptionV3 also performed well at 0.956 and 0.952 accuracy respectively. EfficientNetB3 scored the highest WDLPS recall of 0.986 but failed severely, achieving only 0.635 accuracy and an AUROC of 0.672. It misclassified 963 out of 1,088 benign Lipoma cases as malignant WDLPS, producing an 0.885 false positive rate. This failure is attributed to input resolution mismatch, domain shift between natural photographs and MRI data, and the inability of its compound-scaled frozen features to adapt to this imaging domain.

These findings establish MobileNetV2 as the most effective and practical architecture for this task, with DenseNet121 recommended for safety-critical screening contexts. EfficientNetB3 is definitively unsuitable under the frozen transfer learning configuration. This study provides a well-controlled empirical baseline and clear evidence-based guidance for architecture selection in future AI-assisted lipomatous tumour diagnosis research.

Contents

Chapter 1 Introduction **1**
1.1 Background . . . 1
1.2 Motivation . . . 2
1.2.1 The Unresolved Clinical Need . . . 2
1.2.2 The Technical Maturity of Deep Learning . . . 3
1.2.3 The Absence of a Standardised Benchmark . . . 3
1.2.4 The Clinical Importance of Identifying Prediction Bias . . . 4
1.3 Objectives . . . 4

Chapter 2 Literature Review **5**
2.1 Introduction to the Literature Review . . . 5
2.2 Traditional Diagnostic Approaches for Lipomatous Tumours . . . 5
2.2.1 Clinical Presentation and Pathological Classification . . . 5
2.2.2 The Role of MRI in Lipomatous Tumour Assessment . . . 6
2.2.3 Biopsy and Molecular Diagnosis . . . 6
2.3 Machine Learning Approaches for Soft-Tissue Tumour Classification . . . 7
2.3.1 Radiomic Feature Extraction and ML . . . 7
2.3.2 Limitations of Classical Machine Learning Approaches . . . 8
2.4 Deep Learning and Convolutional Neural Networks in Medical Imaging . . . 9
2.4.1 Foundations of Convolutional Neural Networks . . . 9
2.4.2 Deep Learning in Medical Image Analysis: General Overview . . . 9
2.4.3 Deep Learning Applied to Soft-Tissue Tumours . . . 9
2.5 CNN Architecture Evolution and Design Principles . . . 10
2.5.1 Foundational Architectures: AlexNet and VGGNet . . . 10
2.5.2 Residual Learning: The ResNet Family . . . 11
2.5.3 Densely Connected Networks: The DenseNet Family . . . 11
2.5.4 Compound Scaling: The EfficientNet Family . . . 11
2.5.5 Multi-Scale Feature Extraction: The Inception Family . . . 12
2.5.6 Lightweight Mobile-Optimised Networks: The MobileNet Family . . . 12
2.6 Transfer Learning, Data Challenges, and Augmentation . . . 13
2.6.1 Principles and Rationale for Transfer Learning . . . 13
2.6.2 Feature Extraction vs Fine-Tuning . . . 14
2.6.3 Domain Adaptation . . . 14
2.6.4 Dataset Scale, Class Imbalance, and Augmentation in Medical Imaging AI . . . 14
2.7 Segmentation, Multimodal Imaging, and Clinical Data Fusion . . . 15
2.7.1 The Importance of Accurate Tumour Segmentation . . . 15
2.7.2 Self-Supervised Segmentation . . . 15
2.7.3 Joint Segmentation-Classification . . . 16
2.7.4 Multi-Sequence MRI for Tumour Characterisation . . . 16
2.7.5 Multimodal Neural Networks . . . 16

2.8 Explainability and Interpretability in Medical AI . . . 17
2.8.1 The Clinical Imperative for XAI . . . 17
2.8.2 Gradient-Based Visualisation Methods . . . 17
2.9 Evaluation Frameworks and Metrics in Medical AI . . . 17
2.9.1 Inadequacy of Accuracy Alone . . . 17
2.9.2 Sensitivity and Specificity . . . 17
2.9.3 AUROC as Performance Measure . . . 18
2.10 Generalisation, External Validation, and Benchmarking . . . 18
2.10.1 Generalisation in Medical AI . . . 18
2.10.2 External Validation . . . 18
2.11 Research Gap . . . 20

Chapter 3 Dataset **22**
3.1 The WORC Database . . . 22
3.2 Data Preprocessing Pipeline . . . 25
3.2.1 3D-to-2D Conversion and Axial Reorientation . . . 25
3.2.2 Intensity Standardisation and Resizing . . . 26
3.3 Data Splitting . . . 27
3.4 Data Augmentation . . . 29

Chapter 4 Methodology **30**
4.1 Proposed Transfer Learning Framework . . . 30
4.1.1 Motivation for Transfer Learning in Medical Imaging . . . 30
4.1.2 Feature Extraction Paradigm . . . 31
4.1.3 Framework Architecture . . . 31
4.2 Selected CNN Architectures . . . 32
4.2.1 ResNet50 . . . 33
4.2.2 DenseNet121 . . . 35
4.2.3 EfficientNetB3 . . . 36
4.2.4 InceptionV3 . . . 38
4.2.5 MobileNetV2 . . . 39
4.3 Uniform Classification Head . . . 41
4.3.1 GlobalMaxPooling2D . . . 41
4.3.2 Fully Connected Dense Layer . . . 42
4.3.3 Sigmoid Output Layer . . . 42
4.4 Training Setup . . . 42
4.4.1 Hyperparameters . . . 43
4.4.2 Callbacks and Regularisation . . . 44
4.5 Evaluation Metrics . . . 45
4.5.1 Accuracy . . . 46
4.5.2 Precision . . . 46
4.5.3 Recall . . . 46
4.5.4 Specificity . . . 46
4.5.5 F1-Score . . . 47
4.5.6 AUROC . . . 47

Chapter 5 Results **48**
5.1 Overall Performance Comparison . . . 49
5.1.1 Accuracy Analysis . . . 49
5.1.2 Precision, Recall, and F1-Score Analysis . . . 50
5.1.3 Specificity Analysis . . . 51

5.1.4 AUROC Analysis . . . 51
5.1.5 Class-Wise Performance: Lipoma vs. WDLPS . . . 52
5.1.6 Comprehensive Confusion Matrix Comparison . . . 52
5.1.7 Training Dynamics Comparison . . . 54
5.1.8 Comparative Rankings Across All Metrics . . . 55
5.2 Individual Architecture Analysis . . . 55
5.2.1 ResNet50 . . . 55
5.2.2 DenseNet121 . . . 57
5.2.3 EfficientNetB3 . . . 59
5.2.4 InceptionV3 . . . 61
5.2.5 MobileNetV2 . . . 63
5.3 Integrated Performance Summary . . . 65

Chapter 6 Discussion **67**
6.1 Analysis of Architecture Effectiveness . . . 67
6.1.1 MobileNetV2: Interpreting the Unexpectedly Superior Performance 67
6.1.2 DenseNet121: The Role of Dense Feature Preservation . . . 68
6.1.3 ResNet50 and InceptionV3: Solid Performance in the Second Tier . 69
6.2 Architectural Limitations of the Transfer Learning Approach . . . 70
6.2.1 The ImageNet-to-MRI Domain Shift . . . 70
6.2.2 Constraints of the Frozen Backbone Strategy . . . 70
6.2.3 Input Resolution Mismatch . . . 71
6.3 Prediction Bias: Root Causes and Clinical Consequences . . . 71
6.3.1 EfficientNetB3: Catastrophic Majority-Class Bias . . . 71
6.3.2 Mild Directional Biases in the High-Performing Architectures . . . 72
6.4 Cross-Metric Consistency and Robustness of the Conclusions . . . 73
6.5 Clinical Implications and Deployment Considerations . . . 74
6.5.1 The Role of AI in the Diagnosis of Lipomatous Tumours . . . 74
6.5.2 Threshold Calibration for Asymmetric Clinical Costs . . . 75
6.5.3 Deployment Feasibility and Infrastructure Considerations . . . 75

Chapter 7 Conclusion **76**
7.1 Summary of Findings . . . 76
7.2 Limitations . . . 77
7.3 Future Work . . . 78
7.4 Analysis of the Social Effects . . . 79
7.5 Environmental Impact Analysis and Sustainability . . . 79
7.6 Ethical Issues . . . 79

Bibliography **81**

List of Figures

Figure 3.1.1 Class distribution of subjects in the WORC dataset. 22

Figure 3.1.2 Class distribution of the WORC dataset after 2D slices. 23

Figure 3.1.3 Sample MRI images from the Lipoma and WDLPS classes in the WORC Lipo dataset. 24

Figure 3.2.1 Representative preprocessed MRI slices from the training set showing both Lipoma and WDLPS cases after augmentation and normalisation. 25

Figure 3.2.2 Illustration of the 3D volumetric MRI to 2D axial slice conversion process. 26

Figure 3.3.1 Class distribution of WDLPS and Lipoma images in the training set. 27

Figure 3.3.2 Class distribution of WDLPS and Lipoma images in the validation set. 28

Figure 3.3.3 Class distribution of WDLPS and Lipoma images in the test set. 28

Figure 4.1.1 Transfer learning architecture using frozen pre-trained backbone. 32

Figure 4.2.1 ResNet50 architecture [29]. 34

Figure 4.2.2 DenseNet121 architecture [30]. 36

Figure 4.2.3 EfficientNetB3 architecture [31]. 37

Figure 4.2.4 InceptionV3 architecture [32]. 39

Figure 4.2.5 MobileNetV2 architecture [33]. 40

Figure 5.1.1 Bar chart comparing overall test accuracy across all five CNN architectures. 50

Figure 5.1.2 Comparative AUROC values across all five CNN architectures. . 51

Figure 5.1.3 Confusion matrices for all five CNN architectures on the test set: (a) ResNet50, (b) DenseNet121, (c) EfficientNetB3, (d) InceptionV3, (e) MobileNetV2. 53

Figure 5.2.1 Confusion matrix for ResNet50 on the test set (TP = 1549, FN = 62, FP = 57, TN = 1031). 56

Figure 5.2.2 ResNet50 training and validation loss and accuracy curves over epochs. 57

Figure 5.2.3 Confusion matrix for DenseNet121 on the test set (TP = 1571, FN = 40, FP = 61, TN = 1027). 58

Figure 5.2.4 DenseNet121 training and validation loss and accuracy curves over epochs. 58

Figure 5.2.5 Confusion matrix for EfficientNetB3 on the test set (TP = 1588, FN = 23, FP = 963, TN = 125). 59

Figure 5.2.6 EfficientNetB3 training and validation loss and accuracy curves over epochs. 61

Figure 5.2.7 Confusion matrix for InceptionV3 on the test set (TP = 1553, FN = 58, FP = 72, TN = 1016). 62

Figure 5.2.8 InceptionV3 training and validation loss and accuracy curves over epochs. 63

Figure 5.2.9 Confusion matrix for MobileNetV2 on the test set (TP = 1560, FN = 51, FP = 39, TN = 1049). 64

Figure 5.2.10 MobileNetV2 training and validation loss and accuracy curves over epochs. 64

Figure 5.3.1 Radar chart comparing all five CNN architectures across six performance metrics. 66

List of Tables

Table 2.3.1 Classical machine learning studies for lipomatous tumour classification 8

Table 2.4.1 Deep learning studies for soft-tissue tumour classification 10

Table 2.5.1 Comparative summary of CNN architectures evaluated. 13

Table 2.10.1 Comprehensive comparison of all reviewed studies in lipomatous and soft-tissue tumour AI 19

Table 3.1.1 Summary of the Lipo subset, WORC dataset. 24

Table 4.2.1 Five selected CNN architectures with their key architectural features, parameter counts, and ImageNet Top-1 accuracy. 33

Table 4.4.1 Hyperparameters. 44

Table 5.1.1 Comprehensive performance comparison of all CNN architectures on the test set. 49

Table 5.1.2 Class-Wise performance of all CNN architectures on the test set. 52

Table 5.1.3 Confusion matrix values for all five CNN architectures on the test set. 54

Table 5.1.4 Metric-wise rank ordering of CNN architectures, using corrected specificity (1 = Best, 5 = Worst). 55

Table 5.2.1 Per-class classification report for ResNet50 on the test set. 56

Table 5.2.2 Per-class classification report for DenseNet121 on the test set. . . 57

Table 5.2.3 Per-class classification report for EfficientNetB3 on the test set. . 59

Table 5.2.4 Per-class classification report for InceptionV3 on the test set. . . 61

Table 5.2.5 Per-class classification report for MobileNetV2 on the test set. . . 63

Table 5.3.1 Integrated clinical performance summary with risk assessment for all architectures. 65

Chapter-1 : Introduction

1.1 Background

Cancer is one of the most serious health problems in the world today. In 2020 alone, around 19.3 million new cancer cases were reported, and approximately 10 million people died from the disease [1]. Among many types of cancer, soft-tissue tumours are a group that is difficult to diagnose and treat. Soft-tissue sarcomas (STSs) are rare but serious cancers that grow from tissues such as fat, muscle, nerves, and blood vessels. Although they make up only about 1% of all cancers in adults, they can cause serious harm and death, especially when they are not found and treated early [1].

Liposarcomas (LPSs) are the most common type of soft-tissue sarcoma, making up around 15 to 20% of all STS cases [1]. They grow from fat cells and come in several types, including well-differentiated liposarcoma (WDLPS), dedifferentiated liposarcoma (DDLPS), myxoid liposarcoma (MLPS), and pleomorphic liposarcoma (PLPS) [1, 2]. Each type behaves differently and needs a different treatment plan. WDLPS is the most common type. It usually grows slowly and appears as a large, painless lump, most often in the belly area or the limbs [2, 3]. Even though WDLPS rarely spreads to other parts of the body, it can come back after treatment and may turn into the more dangerous DDLPS over time [1, 2].

One of the biggest challenges in treating fat-based tumours is telling the difference between a harmless fatty lump (lipoma) and WDLPS. Lipomas are very common and affect about 2% of the general population [3]. In most cases, a lipoma does not cause any harm and does not need surgery unless it causes pain or other problems [3]. WDLPS, on the other hand, must be surgically removed as soon as possible to prevent it from spreading or turning into a more dangerous cancer [1–3]. Because the two conditions require completely different treatments, it is very important to correctly identify which one a patient has before any decision is made.

MRI (Magnetic Resonance Imaging) is considered the best imaging tool for checking soft-tissue masses in the limbs [3]. It can show the shape, structure, and internal details of a tumour without using radiation. On T1-weighted MRI images, both lipomas and WDLPS look similar because both contain a lot of fat [4]. While doctors look at features like the size of the tumour, the thickness of its walls, and other tissue components to decide if something is cancerous, these features often overlap between lipoma and WDLPS. This makes it very hard to tell them apart using MRI alone [4]. In fact, studies have shown that standard MRI can correctly identify suspicious cases most of the time, but its ability to rule out cancer is quite weak, with a specificity of only around 37% [4].

The most reliable way to confirm whether a tumour is WDLPS is to test for a gene change called MDM2 amplification. This is done using a method called fluorescence in

situ hybridisation (FISH), which requires a tissue sample taken through a biopsy [4]. This gene change is found in WDLPS but not in lipomas, so it can confirm the diagnosis. However, biopsy is an invasive procedure. It carries risks such as bleeding, infection, and the chance that cancer cells could spread along the needle track [3, 4]. It also requires special laboratory equipment and trained pathologists, and it takes time and money. Because of these limitations, researchers have been looking for non-invasive ways to make this diagnosis more accurately and more easily.

In recent years, deep learning and convolutional neural networks (CNNs) have become powerful tools in medical image analysis. These systems can learn patterns directly from images without needing humans to manually identify features [5, 6]. CNNs have shown very good results in many medical imaging tasks, such as detecting cancer in X-rays, MRI scans, and pathology slides [5, 7]. Their ability to find small and complex patterns that the human eye might miss makes them very suitable for difficult tasks like separating lipoma from WDLPS [7].

However, using deep learning in medicine still has some challenges. One major problem is that there are not enough labelled medical images to train these models properly from scratch [8]. To deal with this, researchers use a technique called transfer learning. In transfer learning, a model that was already trained on a large dataset (such as ImageNet) is adapted to work on a new, smaller medical dataset [6, 8]. This approach has proven to be very useful in medical imaging and is especially relevant for this study.

Although many studies have applied deep learning to soft-tissue tumour classification, there is still no standard way to compare different CNN models fairly for this specific task of separating lipoma from WDLPS using publicly available MRI data. Different studies use different models, datasets, and methods, making it impossible to directly compare their results. This thesis tries to fill that gap by conducting a fair and carefully controlled comparison of five well-known CNN architectures under the same conditions.

1.2 Motivation

This research was motivated by several important reasons. These include an unresolved clinical problem, the growing strength of deep learning technology, the lack of a standard comparison framework, and the need to understand when AI models make dangerous mistakes. Each of these reasons is explained in the subsections below.

1.2.1 The Unresolved Clinical Need

Even with many advances in medical technology, doctors still find it very difficult to tell the difference between a simple lipoma and WDLPS using imaging alone. When a doctor makes a wrong diagnosis, the consequences can be serious. If WDLPS is mistaken for a lipoma (a missed cancer), the patient may not receive surgery in time, and the tumour could grow or become more dangerous [1, 2]. On the other hand, if a harmless lipoma is mistaken for WDLPS (a false alarm), the patient may go through unnecessary surgery with all its risks, including complications from anaesthesia, wound problems, and a long recovery period, even though the tumour was not a threat [3].

In real clinical practice, most fatty soft-tissue masses are harmless lipomas. But because doctors cannot always be sure without a biopsy, many patients end up having risky procedures that were not necessary [3, 4]. An AI-based tool that can accurately classify these tumours from MRI images would be very helpful. It could guide doctors on which patients need surgery and which ones can just be monitored. This would save time, reduce costs, and lower the risk for patients.

WDLPS is also a rare tumour, so many doctors outside of specialist centres do not see many cases and may not have enough experience to identify it correctly on MRI. An AI-assisted tool built into the imaging workflow could help these doctors by flagging cases that need further review or biopsy, while confirming when a mass is likely harmless.

1.2.2 The Technical Maturity of Deep Learning

Another strong reason for this research is that deep learning technology has now reached a level where it can genuinely help in medical diagnosis. Over the past ten years, CNNs have achieved excellent results across many medical imaging tasks, including detecting eye diseases, classifying skin lesions, diagnosing lung infections from X-rays, and finding tumours in CT scans [5–7]. In many of these areas, AI systems have been shown to perform as well as or even better than experienced doctors.

The five CNN models used in this study, namely ResNet50 [9], DenseNet121 [10], EfficientNetB3 [11], InceptionV3 [12], and MobileNetV2 [13], are among the most well-known and widely used architectures in computer vision research. Each one is designed differently and has its own strengths in terms of how it extracts features, how many parameters it has, and how well it generalises. These qualities make them good candidates for testing on a medical imaging problem like lipoma versus WDLPS classification.

Transfer learning also plays a key role here. These models were first trained on ImageNet [14], a large dataset with over 1.2 million images from 1,000 different categories. By starting from these pre-trained models, it is possible to get good results even with a small medical dataset. This is especially important for this study, which uses the WORC database containing MRI data from only 115 patients, a common limitation in rare tumour research [8].

1.2.3 The Absence of a Standardised Benchmark

A key motivation for this research is that no previous study has compared multiple CNN architectures in a fair and consistent way for the specific task of separating lipoma from WDLPS. While some studies have used deep learning for related problems, each one uses a different model, a different dataset, different training settings, and different evaluation methods [15–18]. This makes it very difficult to say which model is actually the best for this task.

For example, a study showing that a ResNet50-based model achieved an AUC of 0.87 [16] cannot be directly compared to another study showing that VGG-19 achieved 98.27% accuracy [17], because the two studies used completely different datasets and setups. Without a common framework, it is impossible to draw reliable conclusions about which

architecture works best.

This thesis addresses this problem by testing all five CNN architectures under exactly the same conditions, including the same dataset, the same training settings, the same model head, and the same random seed for reproducibility. This way, any differences in results can be attributed to the architecture itself rather than to differences in how the experiments were set up.

1.2.4 The Clinical Importance of Identifying Prediction Bias

Another important reason for this study is to understand when an AI model makes systematically wrong predictions. A model might look accurate overall, but if it consistently predicts the wrong class for one group of patients, it can be dangerous in a clinical setting [3,4]. In this study, EfficientNetB3 was found to have a very high false positive rate of 89% for benign lipomas. This means it classified most harmless lipomas as cancerous WDLPS, which could lead to a large number of unnecessary surgeries if used in practice. This kind of systematic error is called prediction bias.

It is not enough to just look at overall accuracy when evaluating a medical AI model. It is also important to check how the model performs for each class separately. A model that consistently over-predicts cancer would cause many patients to go through unnecessary and risky surgery. This thesis looks at these error patterns carefully for all five architectures and provides evidence to show why some models are not suitable for clinical use, even if their overall accuracy seems acceptable.

1.3 Objectives

The objectives of this thesis are formally stated as follows:

To carry out a fair, standardised, and reproducible comparison of five well-known pre-trained CNN architectures, namely ResNet50, DenseNet121, EfficientNetB3, InceptionV3, and MobileNetV2, for the automatic classification of benign lipoma and well-differentiated liposarcoma using T1-weighted MRI images from the publicly available WORC database.

1. To implement a uniform transfer learning framework with frozen pre-trained backbones and an identical classification head for all five architectures, ensuring a fair and controlled architectural comparison.
2. To evaluate each architecture across a comprehensive set of performance metrics including accuracy, precision, recall, F1-score, specificity, and AUROC, providing a multidimensional assessment of diagnostic capability.
3. To characterise the error profiles and prediction biases of each architecture through confusion matrix analysis and class-specific metric evaluation.
4. To identify the most suitable CNN architecture(s) for this specific medical imaging classification task and provide evidence-based recommendations for future research and potential clinical implementation.

Chapter-2 : Literature Review

2.1 Introduction to the Literature Review

This chapter reviews the existing research on automated classification of lipomatous tumours using deep learning and transfer learning. The review builds up gradually, beginning with the clinical and pathological features of lipomatous tumours and the shortcomings of traditional diagnostic methods, then tracing the progress made by machine learning and, in turn, deep learning approaches in this area. The specific CNN architectures used in this thesis are reviewed next, followed by transfer learning strategies, data challenges, class imbalance, and augmentation. The remaining sections cover segmentation, multimodal imaging, explainability, evaluation methods, and generalisation, before the chapter closes by setting out the research gaps this thesis addresses.

Throughout, particular attention is paid to the weaknesses, inconsistencies, and reproducibility problems in earlier work, since these are largely what have held back progress in this area. Comparison tables are included where useful, so prior work can be examined in a more structured way.

2.2 Traditional Diagnostic Approaches for Lipomatous Tumours

2.2.1 Clinical Presentation and Pathological Classification

Lipomatous tumours cover a wide spectrum, from harmless lumps to malignant growths capable of spreading elsewhere in the body. Benign lipomas are, in fact, the most common soft tissue tumours seen in clinical practice [3]. Under the microscope, a simple lipoma is made up of fully mature fat cells with no unusual features, arranged in lobules separated by thin fibrous septa, with no MDM2 gene amplification or other chromosomal abnormality [3, 4]. Several subtypes exist depending on tissue composition, including fibrolipoma, angiolipoma, myolipoma, chondroid lipoma, and spindle cell lipoma, all of which carry a good prognosis [3].

Well differentiated liposarcoma (WDLPS), called atypical lipomatous tumour (ALT) when it occurs in the limbs, is malignant in nature. It shows fat cells that vary considerably in size, atypical nuclei, lipoblasts, and fibrous septa with atypical stromal cells [1, 3]. The key molecular feature distinguishing WDLPS from a benign lipoma is amplification of the MDM2 gene on chromosome 12q13-15, detectable through FISH or immunohistochemistry [4]. This amplification is never seen in benign lipomas yet is present in almost every WDLPS case, which is why it is considered the gold standard for diagnosis [4]. When

WDLPS arises in the retroperitoneum rather than the limbs, the risk of recurrence and progression is considerably higher, making surgical removal essential [1, 2].

Telling WDLPS apart from a benign lipoma is not always straightforward, since in early or well differentiated cases the microscopic difference can be subtle even to an experienced pathologist, and molecular testing is usually needed to confirm the diagnosis [4]. This difficulty at the pathological level adds to the challenges already present at the imaging stage, which is why developing accurate, non-invasive classification tools has become such a priority for both radiology and pathology.

2.2.2 The Role of MRI in Lipomatous Tumour Assessment

Magnetic resonance imaging is widely regarded as the imaging method of choice for evaluating soft tissue masses before surgery [3, 4]. It offers excellent soft tissue contrast, allows imaging in multiple planes, and shows how a tumour relates to surrounding structures, all without exposing the patient to radiation. Standard MRI protocols for soft tissue tumours usually include T1 weighted, fat suppressed T2 weighted, and contrast enhanced sequences [4].

On T1 weighted images, both benign lipoma and WDLPS appear largely bright because of their fat content, making them fairly easy to identify as lipomatous lesions in the first place [4]. Certain features do raise suspicion of malignancy over a benign lipoma: tumour size over 10 cm (though size alone is not reliable), thick or nodular septa, non-fat components exceeding 25% of tumour volume, heterogeneous signal intensity, and enhancement of margins or septa after gadolinium administration [4]. These features, however, overlap considerably between lipoma and WDLPS, and several studies have pointed out the limits of conventional imaging criteria in telling the two apart reliably.

A landmark study by Brisson et al. [4] retrospectively reviewed MRI features of lipomas and ALT/WDLPS lesions against histopathological findings and MDM2 amplification status. Thick septa and non-fat components proved reasonably sensitive to malignancy, but the overall specificity of MRI for ruling out a benign lipoma was only 37%. Such low specificity means most lesions showing a suspicious MRI feature will, on histopathology, actually turn out benign, which says a great deal about the limits of conventional radiological assessment.

The practical impact of this low specificity is not trivial. In many hospitals, a fatty soft tissue mass with even one suspicious MRI feature typically leads to biopsy or surgical excision, so a considerable number of patients with entirely benign lipomas end up undergoing procedures they did not need. A reliable computational tool that could improve the specificity of MRI based diagnosis would therefore be a genuinely useful step forward, both for patient care and for making better use of healthcare resources.

2.2.3 Biopsy and Molecular Diagnosis

Core needle biopsy (CNB) remains the definitive preoperative procedure for soft tissue masses that imaging alone cannot resolve, since it provides tissue for both histology and molecular testing [3]. FISH testing for MDM2 amplification on formalin fixed, paraffin

embedded tissue gives a conclusive answer when distinguishing lipoma from WDLPS [4]. Biopsy, however, is not without its own risks and limitations, and this is exactly what has pushed researchers to look for non-invasive alternatives.

Risks associated with percutaneous biopsy include bleeding, infection, accidental injury to nearby vessels or nerves, and the theoretical risk of tumour cells seeding along the biopsy tract [3]. This last point matters particularly for WDLPS, since spread of malignant cells into previously unaffected tissue could complicate any later surgery. CNB is also prone to sampling error: if the needle happens to sample a region that looks benign, the result may come back falsely negative even though other parts of the same tumour are malignant [4].

Beyond these procedural risks, biopsy needs specialised pathology expertise, molecular testing infrastructure, and a turnaround time that is not always available, particularly in resource limited settings. The cost of repeated biopsies and molecular tests adds further strain on healthcare systems. Taken together, these limitations make a strong case for non-invasive, imaging based computational tools that could replace or complement biopsy in suitable clinical situations.

2.3 Machine Learning Approaches for Soft-Tissue Tumour Classification

2.3.1 Radiomic Feature Extraction and Classical Machine Learning

Before deep learning took over, applying artificial intelligence to soft tissue tumour classification relied mainly on extracting radiomic features and feeding them into classical machine learning classifiers. Radiomics refers to extracting a large number of quantitative features from medical images, including shape descriptors, first order intensity statistics, and higher order texture features computed through methods such as the grey level co-occurrence matrix (GLCM) and grey level run length matrix (GLRLM) [8]. These handcrafted features are meant to capture aspects of tumour shape and internal heterogeneity not readily visible to the eye.

Malinauskaite et al. [19] conducted a notable study showing that machine learning on radiomic features from T1 weighted MRI could distinguish lipomas from liposarcomas with an AUC of 0.926, a result that in fact surpassed the diagnostic performance of experienced musculoskeletal radiologists on the same cases. This finding was significant, since it set a clear clinical benchmark for later deep learning approaches and confirmed that classical machine learning, applied to carefully chosen radiomic features, can meaningfully contribute to this diagnostic problem.

Gitto et al. [20] applied a Random Forest classifier to radiomic features from a multi-centre MRI study distinguishing deep-seated lipomas from atypical lipomatous tumours, achieving a sensitivity of 92%. The multi-centre design lent greater confidence to how well the findings would generalise compared with single-centre work, since the model was trained on data drawn from different scanners, protocols, and patient populations across several institutions [20].

Cay et al. [21] reported very high performance with an SVM classifier trained on radiomic features from a standardised, single-scanner dataset, achieving an AUC of 0.987 for lipoma versus ALT/WDLPS. This outstanding result is likely because the data came from a single scanner with a standardised protocol, removing much of the technical variability that usually accompanies multi-centre data [21]. While encouraging, this also hints at a problem: models trained on such standardised data may not hold up as well once applied to the varied conditions typical of routine clinical practice. Table 2.3.1 brings these three classical machine learning studies together, summarising method, dataset, best reported metric, and key limitation for each, and offers a reference point against which the deep learning studies discussed next can be compared [19–21].

Table 2.3.1: Classical machine learning studies for lipomatous tumour classification

Study	**Method**	**Dataset**	**Best Result**	**Key Limitation**
Malinauskaite et al. [19]	SVM + Radiomics	T1-weighted MRI, multi-centre	AUC = 0.926	Closed dataset; limited reproducibility
Gitto et al. [20]	Random Forest + MRI Radiomics	Multi-centre MRI	Sensitivity = 92%	No external validation
Cay et al. [21]	SVM + Radiomics	Single-scanner MRI	AUC = 0.987	Single scanner; limited generalisability

2.3.2 Limitations of Classical Machine Learning Approaches

Despite these promising results, radiomic machine learning carries several fundamental limitations that eventually pushed the field toward deep learning. First, radiomic feature extraction is highly sensitive to how images are acquired, reconstructed, segmented, and preprocessed [8]. Even minor variation at any of these steps can produce noticeably different feature values for the same tumour, so model performance turns out inconsistent when applied across different datasets or protocols, and this lack of reliability remains a real obstacle to clinical use.

Second, segmenting the tumour region for radiomic feature extraction, whether manually or semi-automatically, is time-consuming, labour-intensive, and dependent on whoever is doing it, so different observers may well arrive at different results [8]. Since radiomic feature quality depends directly on how accurate and consistent this segmentation is, radiomic model performance is inevitably bounded by the quality of human annotation, a dependence that becomes especially problematic where time is short and operator expertise varies.

Third, when features are chosen from a large candidate pool, careful statistical scrutiny is needed to avoid overfitting, particularly with a small patient cohort [8]. A feature that looks statistically associated with the outcome in training data does not necessarily reflect a real biological signal, and when candidate features far outnumber patients, there is a genuine risk of picking up chance associations that do not hold in new data.

Together, these limitations point clearly to the need for approaches that learn relevant

features directly from raw imaging data, are less sensitive to preprocessing variability, and scale more gracefully as more data becomes available. Deep learning, and CNNs in particular, address these shortcomings through end-to-end feature learning, which is largely why they have become the dominant approach in medical image analysis.

2.4 Deep Learning and Convolutional Neural Networks in Medical Imaging

2.4.1 Foundations of Convolutional Neural Networks

Convolutional neural networks process grid-like data such as images by applying learned filters to local regions, producing feature maps of edges, textures, and shapes [5]. Reusing the same filter everywhere keeps parameters low and lets a feature be detected regardless of position [5, 6].

Stacked convolutional layers, each followed by a ReLU activation and interspersed with pooling, build increasingly abstract representations: edges and colours early on, shapes and textures in the middle, object-level features deeper in [5, 7]. This multi-scale learning is a major reason CNNs outperform classical, hand-engineered approaches on visual tasks.

2.4.2 Deep Learning in Medical Image Analysis: General Overview

Deep learning in medical imaging has grown fast since the early 2010s with better GPUs, larger datasets, and stronger architectures [5–7], producing models matching dermatologists on skin lesions [6], ophthalmologists on diabetic retinopathy [5], and radiologists on lung nodules [7]. Reviews by Shen et al. [5] and Chen et al. [7] confirm CNNs as the dominant architecture, while flagging data scarcity, domain shift, and interpretability as persistent challenges, especially for rare tumours with limited cases [7, 8].

2.4.3 Deep Learning Applied to Soft-Tissue Tumour Classification

Deep learning applied to soft tissue tumour classification has expanded considerably over the past decade, though the literature remains far smaller than for more common cancers such as lung, breast, and colorectal cancer. Wang et al. [15] developed a modified VGG16 network to classify benign and malignant soft tissue masses on ultrasound, achieving an AUC of 0.91 and accuracy of 79%, described by the authors as comparable to skilled radiologists. This study shows that deep learning can perform respectably even on ultrasound, a modality that is inherently noisier and carries less information than MRI [15].

Fradet et al. [16] directly compared radiomic machine learning with a ResNet50 based deep learning model for lipoma versus ALT on MRI, and found the deep learning model performed somewhat better, with an AUC of 0.87 against 0.84 for radiomics. This modest but consistent advantage mirrors findings from other areas of medical imaging and supports the idea that end-to-end feature learning from raw images captures information that handcrafted radiomic descriptors tend to miss [16].

Hermessi et al. [17] compared transfer learning across several CNN architectures for classifying liposarcoma and leiomyosarcoma (LMS) on MRI, fine-tuning AlexNet, VGG-16, and VGG-19. Deeper networks performed better overall, with fine-tuned VGG-19 achieving 98.27% classification accuracy, a striking result that still needs some caution given the small dataset and lack of external validation [17].

Yang et al. [18] proposed a computer-aided diagnostic model, RN-DL, combining ResNet50 based deep feature extraction with SVM classification to predict MDM2 amplification status and distinguish WDLPS from lipoma using multimodal MRI. Their model achieved an AUC of 0.950 on external validation [18], suggesting reasonably good generalisability, and the inclusion of external validation is a particularly strong methodological feature that sets this study apart from most others in the field.

In tumour grading and prognosis, Navarro et al. [22] used a DenseNet161 architecture to distinguish low-grade from high-grade sarcomas on MRI, obtaining an AUC of 0.76 on external validation, while Liu et al. [23] developed a radiomic nomogram incorporating deep features extracted by ResNet34 to predict soft tissue sarcoma recurrence, reaching a concordance index of 0.766 on an external test set. Together, these studies suggest deep learning holds real potential not only for tumour classification but for outcome prediction in soft tissue sarcomas too. Table 2.4.1 brings these six deep learning studies together, listing architecture, imaging modality, best result, and key limitation for each, and shows the considerable variation in datasets, tasks, and validation practice that marks the existing deep learning literature on soft tissue tumours [15–18, 22, 23].

Table 2.4.1: Deep learning studies for soft-tissue tumour classification

Study	**Architecture**	**Imaging**	**Best Result**	**Key Limitation**
Wang *et al.* [15]	Modified VGG16	Ultrasound	AUC = 0.91	Ultrasound-specific; no MRI
Fradet *et al.* [16]	ResNet50	MRI	AUC = 0.87	No external validation
Hermessi *et al.* [17]	VGG-19	MRI	Accuracy = 98.27%	Small dataset; no ext. validation
Yang *et al.* [18]	ResNet50 + SVM	Multimodal MRI	AUC = 0.950	Complex two-stage pipeline
Navarro *et al.* [22]	DenseNet161	MRI	AUC = 0.76	Broader sarcoma cohort
Liu *et al.* [23]	ResNet34 + Nomogram	MRI	C-index = 0.766	Recurrence, not classification

2.5 CNN Architecture Evolution and Design Principles

2.5.1 Foundational Architectures: AlexNet and VGGNet

The modern era of deep CNN based image recognition began with Krizhevsky et al., whose AlexNet in 2012 showed that deep CNNs trained with GPU acceleration could

dramatically outperform classical, hand-crafted feature methods on large benchmarks such as ImageNet [14]. AlexNet had eight learnable layers, five convolutional and three fully connected, and introduced techniques that later became standard, including ReLU activations, dropout, and local response normalisation. The VGGNet family, developed by Simonyan and Zisserman, built on this by showing that network depth, achieved through stacks of small 3x3 filters, was itself a key driver of strong recognition performance. VGG-16 and VGG-19, with sixteen and nineteen learnable layers, went on to become widely used baselines and pre-trained feature extractors in both natural and medical image analysis, as seen in Hermessi et al. [17] and Wang et al. [15].

2.5.2 Residual Learning: The ResNet Family

He et al. [9] introduced ResNet in 2016 to address the degradation problem, where very deep networks unexpectedly began performing worse, even during training, as more layers were added. The central idea was the skip connection, or shortcut connection, allowing gradients to flow directly from deeper layers back to shallower ones through an identity mapping, bypassing one or more convolutional layers. This made it possible to train networks with 50, 101, and even 152 layers by easing the vanishing gradient problem [9].

ResNet50, the 50-layer variant used in this thesis, stacks sixteen residual blocks in four groups, each block made of three convolutional layers in a bottleneck arrangement. The bottleneck uses 1x1 convolutions to first reduce and then restore the number of channels around the central 3x3 convolution, lowering computational cost while keeping most of the representational capacity [9]. ResNet50 has since become one of the most widely used backbone architectures in medical image analysis, and has served as the feature extractor in several soft tissue tumour classification studies [16, 18].

2.5.3 Densely Connected Networks: The DenseNet Family

Huang et al. [10] introduced densely connected convolutional networks, or DenseNets, in 2017, taking the idea of skip connections further by connecting every layer to every other layer in a feed-forward direction. Each layer receives the combined feature maps of all layers before it, and passes its own forward to every layer after it. This dense connectivity improves gradient flow during training, encourages feature reuse, reduces parameter count since features need not be relearned, and provides a form of built-in deep supervision, since the loss gradient reaches every layer directly [10].

DenseNet121, the 121-layer variant used here, places transition layers between its dense blocks, using 1x1 convolution followed by 2x2 average pooling to control feature map size. The feature reuse that dense connectivity encourages is especially useful in medical imaging, where datasets tend to be small and every bit of information needs to be squeezed out of limited training data [10], a property well suited to the modestly sized WORC database used here, where parameter efficiency matters a good deal.

2.5.4 Compound Scaling: The EfficientNet Family

Tan and Le [11] introduced EfficientNet on the idea of compound scaling, arguing that scaling network width, depth, or input resolution independently, as was conventional, was

suboptimal. They showed instead that scaling all three together, using fixed coefficients found through neural architecture search, gives better accuracy and efficiency. EfficientNet models rest on a mobile inverted bottleneck (MBConv) base, with a single compound coefficient governing how all three dimensions scale together [11].

EfficientNetB3, the variant evaluated here, sits at the third rung of the EfficientNet scaling hierarchy, applying moderate increases in width, depth, and resolution over the EfficientNetB0 baseline. It takes 300x300 pixel inputs and has around 12 million parameters. Although EfficientNet architectures have topped ImageNet and several other natural image benchmarks, their performance on medical imaging has not always been as strong, and the results of this thesis add further weight to the view that compound scaling does not automatically translate into better performance across every medical imaging domain and setting [11].

2.5.5 Multi-Scale Feature Extraction: The Inception Family

The Inception architecture, first introduced by Szegedy et al. and refined through several revisions, eventually led to InceptionV3 [12]. Its central idea is to handle multi-scale feature extraction through inception modules that apply filters of different sizes, namely 1x1, 3x3, and 5x5, in parallel to the same input, before combining the resulting feature maps channel-wise. This lets the network capture features at several spatial scales at once, without having to settle on a single filter size in advance.

InceptionV3 [12] added several refinements over earlier Inception versions, including breaking large filters into sequences of smaller ones (replacing a 5x5 convolution with two 3x3 convolutions, for instance, or an n by n convolution with a 1 by n followed by n by 1). It also added auxiliary classifiers at intermediate layers to help with vanishing gradients, along with label smoothing as regularisation. InceptionV3 has roughly 23.9 million parameters and accepts 299x299 pixel inputs. Its multi-scale capacity is particularly relevant here, since the diagnostically important features in lipomatous tumours appear at quite different spatial scales, from overall tumour shape down to the fine texture of septal thickening and non-fat components.

2.5.6 Lightweight Mobile-Optimised Networks: The MobileNet Family

MobileNetV2, introduced by Sandler et al. [13], is a highly parameter-efficient architecture built specifically for devices with limited computing resources, such as mobile phones and embedded systems. It rests on two ideas. The first is the depthwise separable convolution, splitting a standard convolution into a depthwise convolution applied to each input channel separately, followed by a pointwise 1x1 convolution combining information across channels, cutting computation considerably. The second is the inverted residual block with linear bottleneck, which keeps a low-dimensional representation at the skip connection while expanding to a higher-dimensional space for internal computation, preserving information without needless reduction at the shortcut.

MobileNetV2 has around 3.4 million parameters, far fewer than ResNet50 (25.6 million), DenseNet121 (8 million), EfficientNetB3 (12 million), or InceptionV3 (23.9 million), yet

still achieves competitive accuracy on ImageNet [13]. Its compact size makes it attractive for medical imaging applications where computing resources may be limited, and its strong showing in this thesis, achieving the highest accuracy of 0.967 among all five architectures, suggests its lightweight design, built around depthwise separable convolutions, may offer real advantages for lipomatous tumour classification from T1-weighted MRI [13]. Table 2.5.1 presents a comparative summary of the five architectures evaluated here, listing parameter count, key innovation, input resolution, and ImageNet top-1 accuracy, and offers a quick architectural reference point for interpreting the results discussed later [9–13].

Table 2.5.1: Comparative summary of CNN architectures evaluated.

Architecture	**Params (M)**	**Key Innovation**	**Input Size**	**ImageNet Top-1 Acc.**
ResNet50 [9]	25.6	Skip connections	224×224	76.1%
DenseNet121 [10]	8.0	Dense connectivity	224×224	74.4%
EfficientNetB3 [11]	12.0	Compound scaling	300×300	81.6%
InceptionV3 [12]	23.9	Multi-scale modules	299×299	78.8%
MobileNetV2 [13]	3.4	Depthwise sep. conv.	224×224	72.0%

2.6 Transfer Learning, Data Challenges, and Augmentation

2.6.1 Principles and Rationale for Transfer Learning

Transfer learning is a machine learning approach in which knowledge gained on one task or domain is used to improve performance on a related but different one [6, 8]. In CNN based image classification, this usually means starting from a network already trained on a large source dataset, almost always ImageNet for medical imaging, and adapting it to the target task using whatever labelled data is on hand [8].

Transfer learning works because features CNNs learn from large natural image datasets tend to be broadly useful across many visual domains. Low-level features such as edge detectors, colour blobs, and Gabor-like filters are quite generic and transfer readily to almost any image type, while mid-level features such as texture patterns and shape detectors are moderately transferable and need relatively few gradient updates to adjust [6]. High-level, more semantic features are naturally more domain-specific and need more substantial adaptation when moving from natural to medical images. This layered transferability is exactly why freezing the earlier convolutional layers, which hold generic, transferable features, while training only the classification head, works so well for adapting pre-trained CNNs to medical tasks with limited labelled data.

2.6.2 Feature Extraction vs. Fine-Tuning vs. Partial Fine-Tuning

Feature extraction, used in this thesis, freezes all convolutional layers and trains only a new head [8], suiting small datasets and lowering overfitting risk. Full fine-tuning updates every layer [16, 17], which can outperform feature extraction on larger datasets but risks overfitting on smaller ones [8]. Partial fine-tuning, freezing early layers while adapting later ones, sits between the two and often works best with moderate data [8]. This thesis applies feature extraction consistently across all five architectures to keep the comparison fair.

2.6.3 Domain Adaptation from Natural to Medical Images

Carrying representations learned on natural ImageNet images over to medical images involves a genuine domain shift. Medical images, and MRI in particular, differ from natural photographs in several ways: they are usually grayscale rather than colour, they represent three-dimensional anatomy as 2D cross-sections, they carry acquisition-specific artefacts such as RF inhomogeneity and chemical shift artefacts in MRI, and the diagnostically relevant features are often subtle rather than resembling the object-level features ImageNet pre-training is geared toward [5, 6, 8].

Despite this domain shift, empirical work has consistently shown that transfer learning from ImageNet improves performance over training from scratch in most medical imaging tasks, even with very limited target domain data [8]. This is because generic, low-level visual features learned from natural images stay useful in the medical setting, and because starting from pre-trained weights, rather than random initialisation, guides optimisation toward better solutions in the loss landscape. This benefit of transfer learning over training from scratch on small medical datasets is well documented across the medical imaging AI literature [5, 6, 8].

2.6.4 Dataset Scale, Class Imbalance, and Augmentation in Medical Imaging AI

Limited data, class imbalance, and the need for augmentation are closely tied to the transfer learning context in medical imaging, since these are precisely the constraints that make pre-trained models both necessary and useful. Understanding these challenges properly is essential for making sense of the experimental choices in this thesis and for interpreting the results in context.

The Problem of Small Datasets in Rare Tumour Research

Large, well annotated datasets remain scarce for rare tumours such as WDLPS [8], since the patient numbers needed simply are not there in the way they are for more common cancers. The WORC database used here has just 115 patients [24], a valuable public resource but still small by the standards of deep learning, and this limits both the complexity a model can be given before it starts to overfit and the variety of tumour presentations it is able to learn from. A model trained on such a modest and relatively narrow set of cases can therefore struggle to generalise well when faced with new patients, different imaging protocols, or settings that differ from those represented in the training data.

Class Imbalance and Its Consequences

Converting 3D volumes to 2D slices here gave 10,739 WDLPS slices against 7,247 lipoma slices (roughly 1.48:1), enough to bias training toward WDLPS [8]. A model always predicting WDLPS would score around 60% accuracy while failing the minority class entirely, which is why class-specific sensitivity, specificity, and F1-score matter. EfficientNetB3's bias here, misclassifying 89% of benign lipomas as WDLPS, likely reflects this imbalance interacting with its architecture.

Data Augmentation Strategies in Medical Imaging

Common strategies include rotations, flips, scaling, and intensity adjustments [8]. This study kept augmentation conservative, mild rotations of plus or minus 5 degrees and horizontal flips only, to avoid anatomically unrealistic images [8].

2.7 Segmentation, Multimodal Imaging, and Clinical Data Fusion

Accurately identifying tumour boundaries and combining information from multiple imaging sequences and clinical data are two closely related challenges in building comprehensive AI diagnostic systems for lipomatous tumours. Both segmentation and multimodal data fusion go beyond what single-sequence 2D classification alone can offer, and both help mark out the scope and limitations of this thesis.

2.7.1 The Importance of Accurate Tumour Segmentation

Tumour segmentation, identifying and outlining the tumour boundary in an image, is a necessary first step for many downstream tasks, including tumour volume measurement, radiomic feature extraction, and surgical planning [25]. For lipomatous tumour classification, accurate segmentation defines the region of interest from which both radiomic and deep learning features are extracted, and differences in segmentation method between studies remain a major source of inconsistency, with a real bearing on model performance [25].

Liu et al. [25] tackled this by developing an AI based ensemble segmentation approach, combining multiple U-Net architectures for automated lipomatous tumour segmentation on MRI, and their ensemble Super Learner achieved a Dice Similarity Coefficient (DSC) of 0.80, a strong result given how poorly defined tumour margins and mixed composition typically complicate soft tissue sarcoma segmentation [25]. The ensemble approach also improved robustness across institutions and acquisition protocols, showing that combining several models can offset the weaknesses of any single approach.

2.7.2 Self-Supervised Segmentation Under Label Scarcity

The shortage of densely annotated data is a particularly serious problem for segmentation, which needs pixel-level labels rather than the image-level labels sufficient for classification. Zheng et al. [26] proposed a self-supervised U-Net framework for soft tissue sarcoma

segmentation, using unlabelled data through a pretext task based pre-training strategy, achieving a DSC of 0.5430 on a challenging benchmark. Though well below the 0.80 DSC of Liu et al.'s fully supervised ensemble [25], this still shows meaningful segmentation performance is achievable with very limited labelled data, provided the structure in unlabelled images is put to good use.

2.7.3 Joint Segmentation and Classification Frameworks

Combining segmentation and classification into one end-to-end framework is an important direction for future work in lipomatous tumour AI. Such a framework would locate the tumour region and assign a diagnostic label at once, removing the need for a separate region-of-interest step and reducing the chance that segmentation variability introduces classification errors downstream. No published study has yet shown such a combined framework specifically for lipomatous tumour differentiation, though comparable frameworks exist for other medical imaging tasks, and this represents a natural extension of the present work.

2.7.4 Multi-Sequence MRI for Tumour Characterisation

This thesis relies solely on T1-weighted MRI, a choice justified by the WORC database's composition, though it is a deliberate limitation relative to clinical practice, where decisions are always informed by evaluating multiple MRI sequences together [4]. Adding T2-weighted, fat suppressed T2-weighted, and contrast enhanced T1-weighted sequences would give additional information on tumour water content, internal composition, vascularity, and contrast enhancement, all of which can meaningfully improve accuracy. Fat suppressed T2 sequences, in particular, make non-fat components easier to spot than T1 alone, while dynamic contrast enhanced MRI can offer quantitative perfusion parameters that differ between benign and malignant lipomatous lesions [4].

Bringing multi-sequence MRI into deep learning classification requires architectural changes that let a network process and combine information from several inputs. Proposed approaches include treating different MRI sequences as separate input channels much like colour channels, early fusion architectures combining features from separately processed sequences, and late fusion architectures combining outputs of independently trained branch networks [5, 7].

2.7.5 Multimodal Neural Networks Integrating Clinical Variables

Beyond combining imaging sequences, there is growing interest in multimodal neural networks that integrate imaging with structured clinical information such as patient age, sex, tumour location, size, and laboratory parameters. Bozzo et al. [27] developed a multimodal network combining MRI with clinical variables to predict overall survival and metastasis in sarcoma patients, achieving a concordance index of 0.77 for overall survival. Including clinical metadata alongside imaging acknowledges that the information clinicians actually rely on is inherently multimodal, and AI systems drawing on that combined information are likely to outperform those relying on imaging alone [27].

2.8 Explainability and Interpretability in Medical AI

2.8.1 The Clinical Imperative for Explainable AI

For deep learning systems to gain real acceptance in clinical practice, high accuracy alone is not enough; they also need to offer clear, clinically meaningful explanations for their predictions [5, 7]. Clinicians are understandably reluctant to trust and act on recommendations from systems that behave as black boxes, particularly in high-stakes settings such as cancer diagnosis, where errors carry serious consequences. Explainable AI (XAI) methods, which reveal which parts of an image or which features most strongly drove a prediction, are therefore an important requirement for clinical adoption.

2.8.2 Gradient-Based Visualisation Methods

Grad-CAM [7] highlights the image regions driving a CNN's decision by weighting the final convolutional layer's feature maps by their gradient. For lipomatous tumours, this would let clinicians confirm a network is attending to thick septa or non-fat components rather than artefacts. This thesis does not include Grad-CAM, a limitation and a priority for future work.

For classifying lipomatous tumours, these visualisations help clinicians ensure the AI is focusing on key diagnostic areas like thick septa, non-fat components, and heterogeneous signals instead of background noise, which ultimately builds trust in AI-assisted diagnosis.

2.9 Evaluation Frameworks and Metrics in Medical AI

2.9.1 The Inadequacy of Accuracy as a Sole Performance Metric

Overall accuracy, the proportion of test samples correctly classified, is the most intuitive performance metric and is often reported as the headline result in medical AI studies. It is, however, fairly limited in clinical contexts, particularly when classes are unevenly represented or the consequences of different error types differ substantially [8]. Relying on accuracy alone can be quite misleading, since a model that always predicts the majority class scores an accuracy equal to how common that class is, without any real ability to tell the classes apart.

For this task, a model always predicting WDLPS would reach close to 60% accuracy on the test set, simply reflecting the proportion of WDLPS slices, while completely failing to identify a single benign lipoma, a clinically dangerous outcome. This shows clearly why the comprehensive evaluation used here, precision, recall, F1-score, specificity, and AUROC alongside accuracy, is needed for a thorough and clinically meaningful assessment.

2.9.2 Sensitivity, Specificity, and Their Clinical Significance

Sensitivity and specificity matter a great deal in tumour classification. Sensitivity, equivalent to recall, measures the proportion of malignant cases correctly identified, while specificity measures the proportion of benign cases correctly identified, and both carry a

direct clinical meaning: high sensitivity means few malignant tumours are missed, high specificity means few benign tumours are wrongly flagged [3, 4].

In a screening context, high sensitivity is usually prioritised so no malignant tumour is missed, even at the cost of more false positives. But for lipomatous tumour classification, where most lesions are benign and unnecessary surgery carries real risk, both sensitivity and specificity need to stay high for a model to be clinically useful. A model with high sensitivity but very low specificity, as EfficientNetB3 shows here with a specificity of 0.550, would generate an unacceptably large number of unnecessary surgical referrals and would not be fit for clinical use as it stands.

2.9.3 AUROC as a Threshold-Independent Performance Measure

The Area Under the Receiver Operating Characteristic Curve (AUROC) measures a binary classifier's discriminative ability without depending on any particular threshold [8]. An AUROC of 1.0 means perfect discrimination, 0.5 means no better than chance. Since AUROC is less sensitive to class imbalance than accuracy, it gives a fuller picture of a model's discriminative ability across every operating point. In this thesis, AUROC values from 0.672 for EfficientNetB3 to 0.997 for MobileNetV2 show quite clearly the dramatic performance differences that overall accuracy alone can partly hide.

2.10 Generalisation, External Validation, and Benchmarking

2.10.1 The Challenge of Generalisation in Medical AI

Generalisation, a model's ability to hold up on new data from different patients, institutions, scanners, or protocols, is arguably the most critical, and most often inadequately addressed, challenge in medical AI research [7, 8]. Most published studies test their models only on internal test sets drawn from the same distribution as training data, which says little about how a model would actually perform in real clinical settings, where the data will always differ somewhat from what it was trained on.

Domain shift, the statistical gap between training data and deployment data, is a particularly serious concern in MRI based imaging. An MRI image's appearance depends heavily on scanner manufacturer, field strength, pulse sequence parameters, acquisition matrix, slice thickness, and post-processing applied by the scanner software [7], so a model trained on one scanner at one institution may perform considerably worse on data from a different scanner or institution, even for the same task.

2.10.2 External Validation and Multi-Institutional Studies

External validation, testing a trained model on an independent dataset from a different institution with different equipment and protocols, remains the most rigorous way to check whether a model truly generalises, and is a necessary step before any model can seriously be considered for clinical use. Among the studies reviewed here, only Yang et al. [18] reported

external validation for a lipomatous tumour classification model, achieving an AUC of 0.950 on an independent set, a result that, if confirmed by further studies, would represent genuinely meaningful diagnostic performance. Table 2.10.1 pulls together every study discussed in this chapter alongside the present work, comparing method, dataset, task, best result, and limitation for each, and places this thesis within the broader landscape of lipomatous and soft tissue tumour AI research [15–27].

Table 2.10.1: Comprehensive comparison of all reviewed studies in lipomatous and soft-tissue tumour AI

Study	**Method**	**Dataset**	**Task**	**Best Result**	**Limitation**
Malinauskaite *et al.* [19]	SVM	Multi-centre MRI	Lipoma vs. LPS	AUC = 0.926	Closed dataset
Gitto *et al.* [20]	Random Forest	Multi-centre MRI	Lipoma vs. ALT	Sens. = 92%	No ext. val.
Cay *et al.* [21]	SVM	Single scanner MRI	Lipoma vs. ALT	AUC = 0.987	Single centre
Wang *et al.* [15]	VGG16	Ultra-sound	Benign vs. Malign	AUC = 0.91	Not MRI-based
Fradet *et al.* [16]	ResNet50	MRI	Lipoma vs. ALT	AUC = 0.87	No ext. val.
Hermessi *et al.* [17]	VGG-19	MRI	LPS vs. LMS	Acc = 98.27%	Small dataset
Yang *et al.* [18]	ResNet50 +SVM	Multi-MRI	Lipoma vs. WDLPS	AUC = 0.950	Complex pipeline
Navarro *et al.* [22]	DenseNet 161	MRI	Sarcoma grading	AUC = 0.76	Broad cohort
Liu *et al.* [23]	ResNet34	MRI	Recur-rence	C-index = 0.766	Prognostic only
Liu *et al.* [25]	U-Net Ensemble	MRI	Segmen-tation	DSC = 0.80	Segmentation only
Zheng *et al.* [26]	Self-sup. U-Net	MRI	Segmen-tation	DSC = 0.543	Self-supervised
Bozzo *et al.* [27]	Multi-modal NN	MRI + Clinical	Survival	C-index = 0.77	Survival only

2.11 Research Gap

The review above points to several interconnected gaps in the existing literature, and together these motivate the contribution of this thesis. Prior studies have made real progress in applying machine learning and deep learning to soft tissue tumour classification, yet a number of methodological, architectural, and clinical weaknesses remain. Each is taken up below, and each is addressed directly by the design, methodology, or findings of this thesis.

No prior study has carried out a systematic, standardised comparison of multiple CNN architectures under a strictly controlled framework built specifically around classifying benign lipoma against well differentiated liposarcoma. Studies such as Fradet et al. [16], Hermessi et al. [17], and Yang et al. [18] each tested one or more CNN architectures on related tasks, but every one used a different dataset, preprocessing pipeline, set of hyperparameters, and evaluation protocol, which makes direct comparison across studies unreliable and leaves no sound basis for judging which architecture genuinely suits this task best. A fair comparison needs identical hyperparameters, an identical classification head, identical training and evaluation protocols, and a common dataset, and no published study on lipoma versus WDLPS differentiation has met all of these at once. This thesis fills that gap by offering the first standardised, five-architecture benchmark for the problem, run on a publicly available dataset under a fully specified, reproducible protocol.

Apart from the WORC database [24], every dataset used in the prior studies reviewed here is closed and institution-specific, out of reach for other researchers. This reliance on proprietary data creates a real reproducibility problem, since results reported on data that cannot be independently accessed can be neither verified nor built upon, and it becomes difficult to tell whether reported differences in performance reflect genuine methodological progress or simply differences in data quality, patient selection, or scanner protocol between institutions. Very high reported figures, such as the 98.27% accuracy from Hermessi et al. [17] or the AUC of 0.987 from Cay et al. [21], are hard to judge critically without knowing the dataset's specific characteristics, and metrics from small, institution-specific datasets are known to be inflated by overfitting to features that do not generalise. Using the publicly available WORC database here, along with full disclosure of every experimental parameter, gives future researchers a reproducible baseline to build on.

The existing literature has also stopped short of characterising the nature of classification errors, or of checking whether different architectures tend to be biased toward one class over the other. This is a real oversight, since aggregate metrics can hide important failure modes: a model biased toward malignancy flags most benign lipomas as WDLPS, leading to unnecessary surgery, while one biased toward benignity misses genuine cancers. No prior study has documented how architectural choice influences prediction bias in this task. The finding here that EfficientNetB3, despite strong performance on natural image benchmarks and wide use in medical imaging, shows an extraordinarily high false positive rate of 89% for benign lipomas under frozen-backbone conditions is therefore of direct clinical relevance, and had not been reported before. It shows clearly why candidate architectures need to be judged on the full set of class-specific indicators, sensitivity, specificity, and the error distribution, rather than accuracy alone.

Despite growing interest in lightweight CNN architectures for resource-limited clinical settings, whether a model such as MobileNetV2 [13] could handle lipomatous tumour classification from MRI had not been studied systematically before this thesis. The

literature has largely been dominated by larger, more demanding architectures such as VGG-16, VGG-19, and ResNet50, carrying the assumption that greater capacity is needed for strong diagnostic performance [16–18]. The finding here that MobileNetV2, with about 3.4 million parameters, far fewer than any other architecture tested, achieves the highest accuracy of 0.967 and AUROC of 0.997 among all five is therefore a novel and significant result. It challenges assumptions about the link between model complexity and diagnostic performance, and carries real practical value for deployment on edge devices or in high-throughput screening.

Every deep learning study reviewed here, including this thesis, processes MRI as 2D cross-sectional slices rather than full 3D volumes, driven mainly by the architectural needs of standard 2D CNNs and the computational cost of volumetric processing. This still means losing spatial context that may carry real clinical significance, since a tumour's three-dimensional shape, internal structure, and the continuity of its margins across slices are inevitably fragmented once broken into independent 2D slices. Features such as the spatial extent of thick septa and the 3D distribution of non-fat components may well be captured better by a genuinely volumetric approach. No study has yet directly compared 2D and 3D approaches for lipoma versus WDLPS differentiation, though comparable work in CT lung nodule detection and MRI brain tumour segmentation consistently favours 3D methods, suggesting this gap may prove meaningful here too, and remains an important direction for future work.

A pervasive limitation of the existing literature is the near-total absence of explainability analyses checking whether trained models actually attend to diagnostically meaningful regions when making decisions. Not one study reviewed here, including Fradet et al. [16], Hermessi et al. [17], and Yang et al. [18], used gradient-based visualisation methods such as Grad-CAM, Layer-wise Relevance Propagation (LRP), or SHAP-based attribution to examine their models' learned decision boundaries. This is not just a technical omission but a real barrier to clinical adoption, since without evidence that a model attends to features such as septal thickness rather than peripheral artefacts, clinicians have little basis to trust its predictions. The radiology community has consistently stressed interpretability as a precondition for adopting AI-assisted tools, and regulatory frameworks increasingly demand evidence of explainability. Adding Grad-CAM based explainability for all five architectures evaluated here is identified as a high-priority direction for future work, one expected to shed useful light on the pathological basis of EfficientNetB3's prediction bias.

Chapter-3 : Dataset

3.1 The WORC Database

The quality and reliability of the dataset matters a great deal in any machine learning study built around medical images, and this study is no exception. All imaging data used here comes from the publicly available Workflow for Optimal Radiomics Classification (WORC) database, assembled by Starmans et al. [24]. WORC is a multi-study, multi-centre repository put together specifically to support reproducible research in computer-aided diagnosis, and it brings together MRI and CT scans, segmentation masks, and clinical labels from 930 patients across six separate radiomics studies. Choosing a publicly available and well-documented database was a deliberate decision, since it lets other researchers verify and replicate the experiments described here. This matters quite a bit in clinical AI research, where many earlier studies have relied on private hospital datasets that could never be independently accessed or checked.

This study uses only the Lipo subset of the WORC database, collected specifically to support the task of distinguishing benign lipomas from well-differentiated liposarcomas (WDLPS). The Lipo subset provides T1-weighted MRI scans from 115 patients in total: 57 with lipoma, 57 with WDLPS, and one additional patient presenting both tumour types at once. At the patient level, the two classes are almost perfectly balanced, which is a genuinely favourable starting point for a binary classification task, since it keeps the risk of systematic class bias during training fairly low from the outset. This near-equal distribution is shown in Figure 3.1.1, giving the patient-level class counts from the WORC database [24].

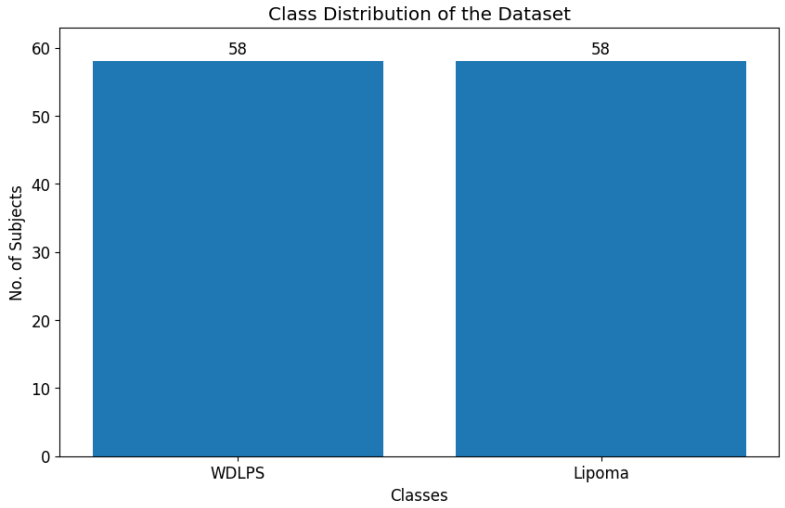


Figure 3.1.1: Class distribution of subjects in the WORC dataset.

T1-weighted MRI was chosen as the imaging sequence of interest because fat appears noticeably bright on T1-weighted images while surrounding soft tissue appears comparatively darker. This makes T1-weighting particularly well suited for evaluating fatty masses such as lipomas and liposarcomas, and it is, in fact, considered the standard imaging method for soft tissue tumour assessment in clinical practice. Every scan in the Lipo subset was acquired using T1-weighting, which keeps imaging conditions reasonably consistent from one patient to the next. Each scan is stored in NIfTI format (.nii or .nii.gz), a widely used medical imaging format that preserves spatial orientation, voxel size, and other useful metadata alongside the image data itself, and each scan is accompanied by a segmentation mask marking out the exact location of the tumour, which makes it straightforward to focus on the diagnostically relevant region during analysis.

At 115 patients, the dataset is undeniably small compared to the datasets typically used in general deep learning research, where tens or hundreds of thousands of examples are common. However, this is a very familiar constraint in medical imaging AI. Patient privacy regulations, the high cost of expert annotation, and the rarity of specific diseases all make it genuinely difficult to collect large labelled datasets in clinical research settings. The WORC database helps address this problem meaningfully by making its data freely available in a standardised and well-documented format, so researchers across the world can use and directly compare results on the same data. Since MRI data is inherently three-dimensional, the process of converting each 3D scan into multiple 2D axial slices, described in detail in the next section, also produces considerably more individual images than the patient count alone would suggest, which helps provide enough data for effective model training.

Figure 3.1.2 shows the image-level class distribution that results after this slicing process, where the near-equal patient-level split shifts to a roughly 3:2 WDLPS-to-Lipoma ratio. This shift happens because WDLPS tumours tend to be physically larger and therefore span more axial slices per patient than lipomas typically do. Figure 3.1.3 then presents representative T1-weighted axial slices from each class, giving a qualitative sense of why lipoma and WDLPS can be genuinely difficult to tell apart from imaging alone, since both tumour types appear predominantly bright on T1-weighted images due to their high fat content.

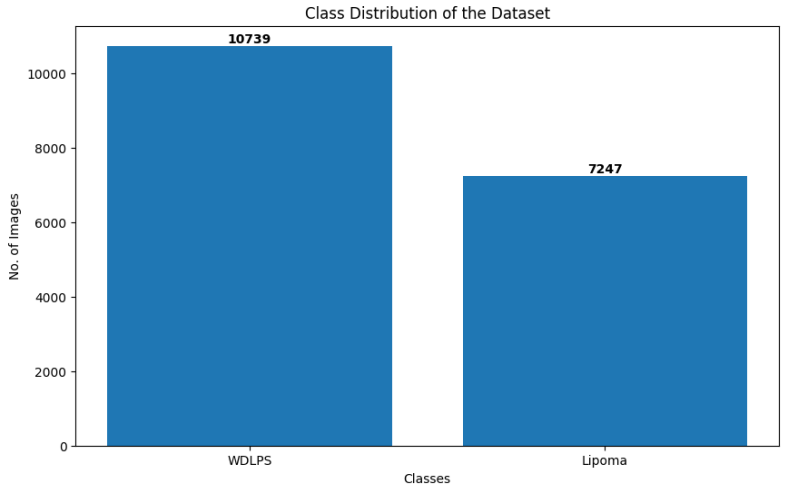


Figure 3.1.2: Class distribution of the WORC dataset after 2D slices.

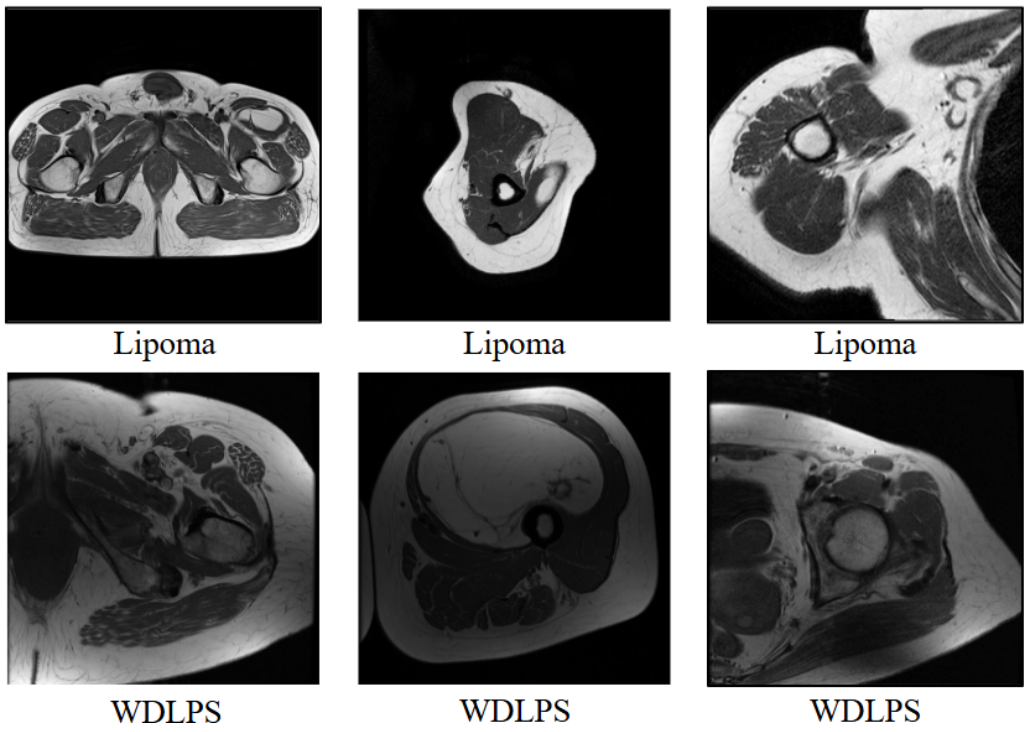


Figure 3.1.3: Sample MRI images from the Lipoma and WDLPS classes in the WORC Lipo dataset.

Table 3.1.1 brings together, in one place, the full descriptive summary of the WORC Lipo subset used in this study, covering modality, subject and scan counts, total slice count, file format, and pixel intensity range, consolidating figures that are presented individually elsewhere in this chapter. The subject, scan, and slice counts are drawn from the WORC database, while the slice-level totals come from this study's own preprocessing pipeline (Sections 3.2 to 3.3).

Table 3.1.1: Summary of the Lipo subset, WORC dataset.

Property	**Detail**	**Value**
Modality	Sequence Type	T1-weighted MRI
Total Number of Subjects	Lipoma	57
	WDLPS	57
	Both Tumour Types	1
Total Number of Data	Scans per Patient	115 (1 volume per patient)
Total 2D Slices (17,986)	WDLPS	10,739
	Lipoma	7,247
File Format	Extension	NIfTI (.nii / .nii.gz)
Minimum Intensity	Raw Pixel Type	−32,768
Maximum Intensity	Raw Pixel Type	32,767

3.2 Data Preprocessing Pipeline

Raw NIfTI MRI data cannot simply be fed into a CNN as it is. It first needs to pass through a series of preparation steps that turn volumetric medical scans into something a standard deep learning model can actually work with. The preprocessing pipeline used in this study is built around two main stages: a geometric transformation stage, covering reorientation and 3D-to-2D conversion, and a photometric normalisation stage, which adjusts pixel intensity values and standardises image dimensions.

Three principles guided the design of this pipeline throughout. First, every step is deterministic, so the same input scan always produces exactly the same output. Second, every transformation is applied in a fixed, unchanging order, to avoid any unexpected interaction between steps. Third, the pipeline was designed to preserve all diagnostically relevant information while still making the images compatible with the three-channel RGB input that the pre-trained CNN models expect. The outputs of this pipeline were also inspected visually on a representative sample of images, simply to confirm that tumour shape, tissue contrast, and overall spatial structure had all come through the process intact, as shown in Figure 3.2.1, which presents representative training-set slices from both classes after the full pipeline has been applied to the original WORC scans.

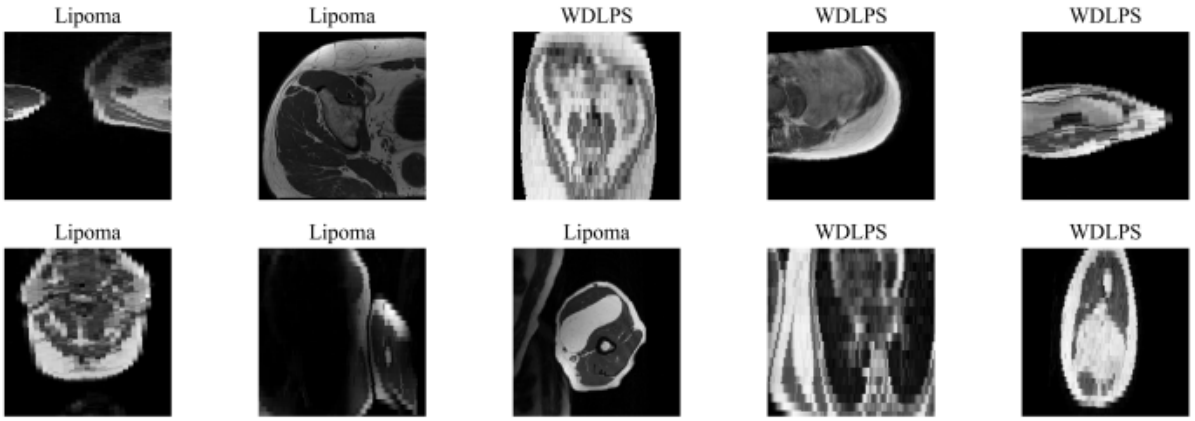


Figure 3.2.1: Representative preprocessed MRI slices from the training set showing both Lipoma and WDLPS cases after augmentation and normalisation.

3.2.1 3D-to-2D Conversion and Axial Reorientation

MRI scans are, by their nature, three-dimensional, representing the body as a grid of voxels spread across the axial, coronal, and sagittal directions. Although 3D CNNs do exist for processing volumetric data of this kind directly, they demand considerably more computation and considerably more training data than their 2D counterparts. Given how small the dataset is here, and given that five separate architectures needed to be trained under identical conditions, a 2D slice-based approach was chosen instead, as a practical and fairly widely accepted alternative: each 3D volume is broken down into a series of 2D cross-sectional images, each of which is then treated as an independent training sample.

Before any slices were extracted, each 3D volume was first reoriented to a standard axial view, using the affine transformation matrix stored in that scan's NIfTI header. Axial slices cut horizontally through the body and represent the view most commonly used in

clinical practice for assessing soft tissue tumours in the limbs, since they show tumour boundaries clearly in relation to nearby muscle and vasculature. Reorienting every volume to this same standard view ensures that every extracted slice presents anatomy from a consistent direction, removing orientation-related variability that would otherwise carry no clinical meaning at all.

Once reorientation and slice extraction had been carried out across all 115 patients, 17,986 individual 2D images were obtained in total: 10,739 from WDLPS patients and 7,247 from Lipoma patients. The larger WDLPS count here simply reflects the fact that WDLPS tumours tend to be physically larger, spanning more axial slices per patient than lipomas typically do. One acknowledged limitation of this slice-based approach is that it discards the spatial relationship between neighbouring slices, since each image is treated independently during both training and testing, meaning the model has no way of learning three-dimensional tumour shape, or how a tumour's morphology changes as one moves through successive slices. This is a real limitation, but the sheer number of slices obtained per patient does provide the model with many different views of the same tumour, offering at least some compensation for the volumetric context that is lost. This reorientation and slicing process is illustrated schematically in Figure 3.2.2, using a volume drawn from the WORC Lipo subset.

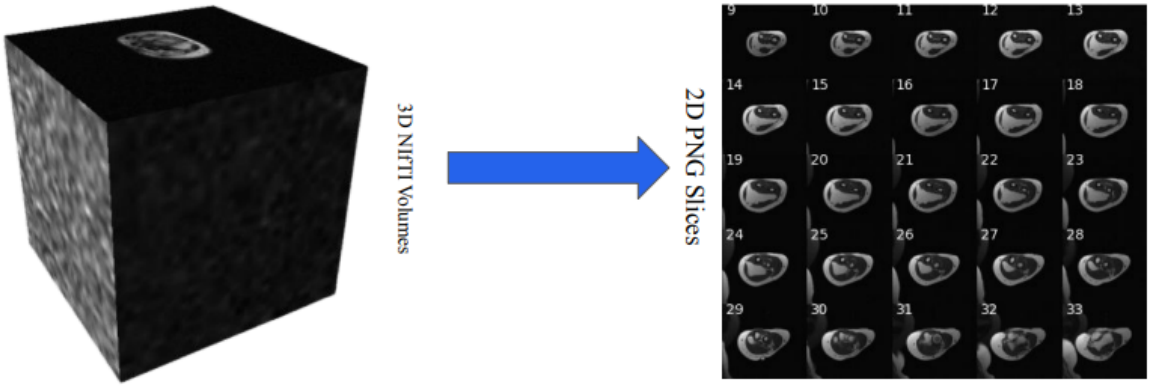


Figure 3.2.2: Illustration of the 3D volumetric MRI to 2D axial slice conversion process.

3.2.2 Intensity Standardisation and Resizing

Raw MRI images store their pixel values as signed 16-bit integers (int16), spanning roughly negative 32,768 to positive 32,767, a range far wider than the 0 to 255 scale of unsigned 8-bit integers (uint8) that standard image processing libraries and pre-trained CNN models are built to expect. To bring these values into line, a linear per-slice normalisation was applied: within each slice, the minimum pixel value was mapped to 0 and the maximum to 255, with everything in between scaled proportionally. Working per-slice in this way preserves the full contrast range within each individual image, regardless of any intensity differences that might otherwise exist from one scanner to another.

Each normalised grayscale image was then converted into RGB format, simply by copying the single grayscale channel across all three colour channels. This step is necessary because all five pre-trained models used in this study were originally trained on three-channel ImageNet photographs, and expect three-channel input by design. Feeding a single-channel image directly, without this conversion, would call for architectural changes that would

disturb the pre-trained weights and, in doing so, undermine much of the benefit transfer learning is meant to provide in the first place.

Every image was then resized to 224x224 pixels using bilinear interpolation, which tends to preserve edge sharpness rather better than simpler nearest-neighbour methods do, and which also matches the standard input size expected by the ImageNet pre-trained models used in this study. As a final step within the model pipeline, pixel values were rescaled from the 0-to-255 uint8 range down to a floating-point range of 0 to 1, achieved simply by dividing by 255. This matches the input normalisation convention the original pre-trained models were trained with, and helps keep gradient magnitudes stable during training.

3.3 Data Splitting

Once preprocessing was complete, the full dataset of 17,986 images was divided into training, validation, and test sets. The training set is what actually updates the classification head's weights during learning. The validation set, meanwhile, tracks performance on data the model has not directly learned from as training progresses, and it is what guides decisions around early stopping and learning rate reduction. The test set, finally, is kept entirely separate from both and used only once, at the very end of training, to evaluate the finished model. Because it plays no role whatsoever in training or hyperparameter tuning, it provides a genuinely unbiased measure of how well the model generalises to data it has never encountered in any form.

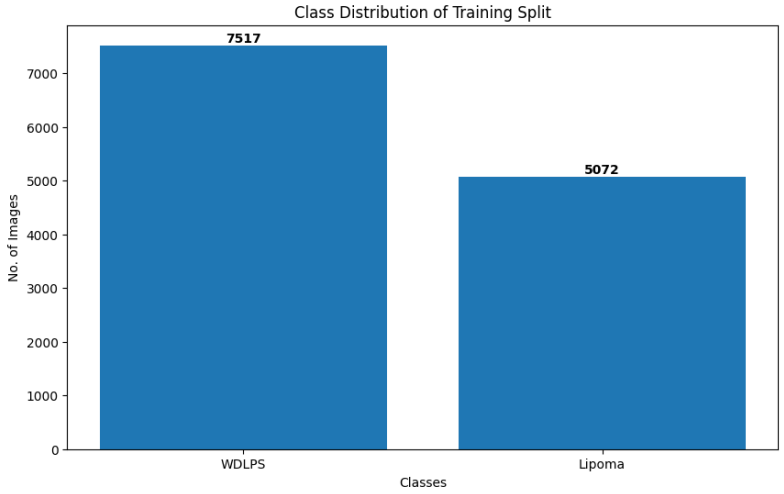


Figure 3.3.1: Class distribution of WDLPS and Lipoma images in the training set.

A 70%-15%-15% split was used here, a ratio that is fairly standard in deep learning research and one that allocates the bulk of the data to training while still leaving enough samples aside for reliable evaluation. Stratified sampling was applied throughout, so that the WDLPS-to-Lipoma ratio remained consistent across all three subsets. This left the training set with 12,589 images (7,517 WDLPS and 5,072 Lipoma), the validation set with 2,698 images (1,611 WDLPS and 1,087 Lipoma), and the test set with 2,699 images (1,611 WDLPS and 1,088 Lipoma). The roughly 3:2 WDLPS-to-Lipoma ratio running through all three subsets simply reflects the original class imbalance in the dataset, itself a consequence of WDLPS tumours tending to be larger and therefore yielding more axial

slices per patient than lipomas typically do. Figures 3.3.1 and 3.3.2 show this breakdown for the training and validation sets respectively, each generated from this study's own stratified splitting procedure applied to slices sourced from the WORC database.

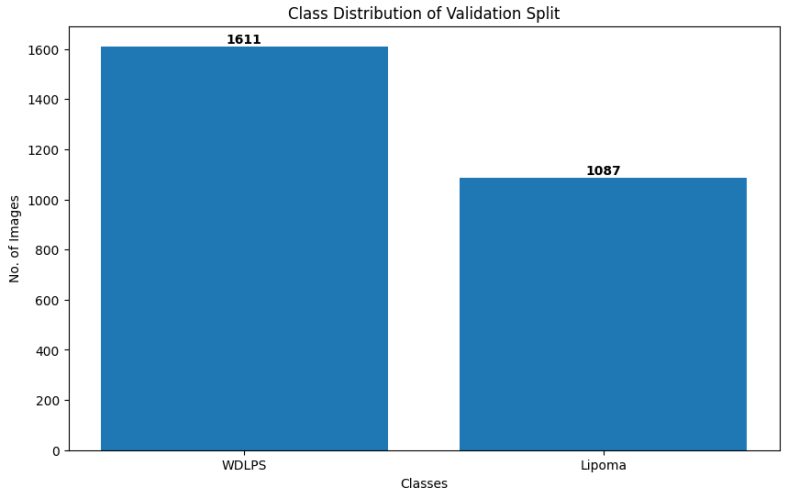


Figure 3.3.2: Class distribution of WDLPS and Lipoma images in the validation set.

It is worth being upfront that this split was performed at the image level rather than the patient level. Ideally, every slice belonging to a given patient would be confined to a single subset, so as to prevent any patient-specific information from leaking between the training and evaluation sets. An image-level split was used here instead, mainly to make the most of the limited number of training images available and to stay consistent with similar published studies in this area, but a patient-level split would be the more rigorous choice for future work, since it would give a more realistic sense of how the model performs on patients it has genuinely never seen before, in any of their slices. Figure 3.3.3 shows the corresponding class breakdown for the test set, confirming that the same roughly 3:2 WDLPS-to-Lipoma ratio, sourced from the WORC database, was preserved in the data used for the final model evaluation reported in Chapter 5.

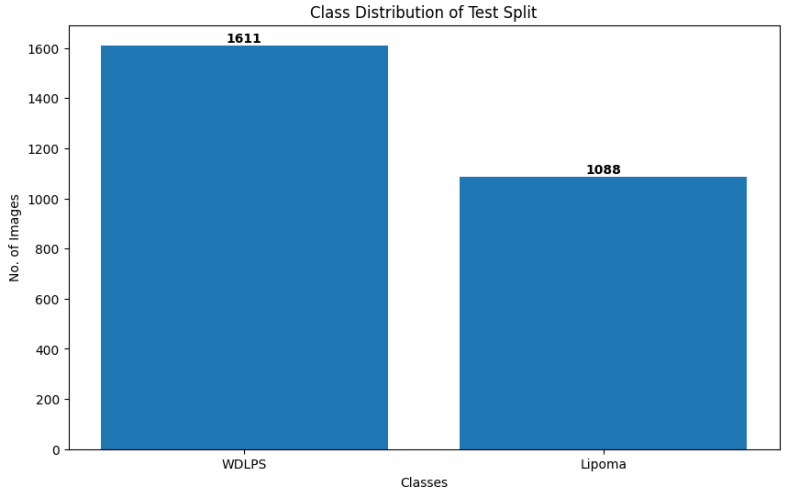


Figure 3.3.3: Class distribution of WDLPS and Lipoma images in the test set.

3.4 Data Augmentation

Data augmentation works by artificially widening the variety present in the training set, applying small, carefully controlled changes to the images that are already there. The underlying idea is a simple one: a model ought to classify an image correctly even after it has been slightly rotated or flipped, since transformations this minor do not change anything about what the image actually shows from a clinical point of view. Exposing the model, during training, to many slightly different versions of the same underlying images helps steer it toward learning features that are genuinely related to the tumour itself, rather than incidental patterns caused by how the patient happened to be positioned or how the scanner happened to be set up. In doing so, it also helps guard against overfitting, the tendency of a model to memorise its training data too closely and, as a result, generalise poorly to images it has not seen before.

Augmentation was applied only to the training set; the validation and test sets were left in their original, unaltered preprocessed form throughout. This is a standard, and rather important, distinction to maintain, since augmentation is really a training tool rather than a general-purpose transformation, and evaluation should always be carried out on clean, unmodified images if the results across all five models are to remain consistent and fair.

Two augmentation operations were used in this study. The first is random rotation within a range of plus or minus 5 degrees, a fairly conservative range chosen to mirror the kind of minor positional variation that naturally arises from how a patient happens to be positioned in the scanner; a wider rotation range would risk producing images that no longer resemble a realistic clinical MRI scan. Where rotation left pixels near the image corners falling outside the original frame, these were filled in with a fixed background value, so that every image retained its original 224x224 size. The second operation is random horizontal flipping, applied with a 50% probability. This is clinically defensible for limb tumours specifically, since the body is roughly symmetric from left to right, and a tumour appearing on the left side carries exactly the same diagnostic meaning as one appearing on the right.

More elaborate augmentation techniques, such as elastic deformations, random cropping, brightness adjustments, zoom, or shear transformations, were deliberately left out. While such methods can sometimes be helpful, they also carry a real risk of producing images that look anatomically unrealistic or clinically implausible, and given how small the dataset already is, and how sensitive medical classification tends to be to this sort of distortion, a more conservative augmentation strategy seemed the safer choice, one that adds a useful degree of variety to training without risking misleading or out-of-distribution patterns that could end up harming, rather than helping, what the model actually learns.

Chapter-4 : Methodology

This chapter sets out the complete methodology followed in this study for the automatic classification of benign lipoma and well-differentiated liposarcoma (WDLPS) from T1-weighted MRI images. The methodology rests on five main components: the transfer learning framework and the reasoning behind it; a detailed account of each of the five CNN architectures used; the classification head placed on top of every backbone; the training setup, including all hyperparameters and training strategies; and, finally, the evaluation metrics used to measure and compare performance across models. Each component is explained in enough detail that the experimental design can be followed, and reproduced, by other researchers.

4.1 Proposed Transfer Learning Framework

Transfer learning, broadly speaking, is a method in which a model already trained on one large dataset is reused for a new, related task that typically has a much smaller dataset of its own. In computer vision, this usually means taking a CNN trained on ImageNet, a dataset of over 1.2 million images spread across 1,000 categories, and applying it to a new classification problem, often in a completely different visual domain. This works because the features a CNN learns from a large natural-image dataset, such as edge detectors, texture patterns, and shape detectors, tend to be general enough to remain useful outside the domain in which they were originally learned, medical imaging included.

4.1.1 Motivation for Transfer Learning in Medical Imaging

Collecting large, well-labelled datasets is notoriously difficult in medical imaging. Patient privacy, the cost of expert annotation, and the rarity of many specific diseases all conspire to keep most medical datasets fairly small, and the dataset used in this study is no exception. The Lipo subset contains just 115 patients and around 17,986 preprocessed 2D images once all the preprocessing steps described later in this chapter have been applied. That is a modest amount of data by deep learning standards, and training a CNN from scratch on a dataset this size runs a real risk of the model simply memorising the training images rather than learning anything generally useful, the familiar problem of overfitting. Transfer learning sidesteps this by starting from features already learned across millions of natural images, leaving the model with the comparatively simpler task of learning how to map those features onto the new classification problem, a task it can handle reasonably well even with a dataset of the size used here.

There is also a practical, resource-related motivation for transfer learning. A pre-trained model already carries a great deal of visual knowledge encoded in its weights, so adapting

it to a new task only requires training a small additional part of the network rather than learning everything from first principles. In research settings where GPU time and labelled data both tend to be in short supply, this makes transfer learning a genuinely practical choice, and not merely a convenient one.

4.1.2 Feature Extraction Paradigm

The particular transfer learning approach adopted in this study is the feature extraction paradigm. Here, the convolutional backbone of each pre-trained model is frozen completely, its weights are simply not updated during training, and only the newly added classification head is trained. This approach tends to make the most sense when the source domain (natural images from ImageNet) and the target domain (T1-weighted MRI) differ substantially, and when the target dataset is on the smaller side, as is the case here. Freezing the backbone protects the useful, generally applicable features it has already learned from being disturbed while training on a comparatively small medical dataset.

An alternative approach, fine-tuning, allows some or all of the backbone layers to be updated during training and can sometimes yield better performance when the source and target domains are reasonably similar to begin with. However, fine-tuning also demands careful control over learning rates and layer-unfreezing schedules, choices that would likely need to differ somewhat from one architecture to another, making a fair, like-for-like comparison considerably harder to achieve. Since the primary goal of this study is to compare five architectures under genuinely identical conditions, the frozen-backbone approach was chosen instead, precisely because it keeps the setup identical across all five models and ensures that any difference in results can be attributed to the architecture itself, rather than to some difference in how each model happened to be trained.

4.1.3 Framework Architecture

The proposed transfer learning framework consists of three parts, applied in exactly the same way to all five CNN architectures evaluated in this study:

1. **Pre-trained frozen convolutional backbone.** Each backbone is loaded with weights pre-trained on the ImageNet [28] ILSVRC-2012 dataset, and every convolutional layer is frozen (set to non-trainable). The backbone processes each $224\times224\times3$ RGB input image and produces a three-dimensional feature map tensor of dimensions $H' \times W' \times C'$, where H' and W' denote the spatial height and width of the final convolutional output, and C' the number of channels. The exact values of H', W', and C' differ from one architecture to the next, but all are handled identically by the classification head that follows.

2. **GlobalMaxPooling2D aggregation.** A GlobalMaxPooling2D layer takes the backbone's output feature map and reduces it to a one-dimensional feature vector of length C', by taking the maximum activation across all spatial positions ($H' \times W'$) independently for each channel. In effect, this operation picks out the single strongest response for each feature type, wherever in the image it happens to occur, retaining the most salient discriminative signals while discarding information about their spatial location. The resulting C'-dimensional vector feeds into the classification head.

3. **Uniform classification head.** A standardised classification head, described in full in Section 4.3, is appended after the GlobalMaxPooling2D layer and trained to map the pooled feature vector to a binary classification probability. This head is architecturally identical across all five models: a fully connected dense layer of 128 neurons with ReLU activation, followed by a single sigmoid-activated output neuron.

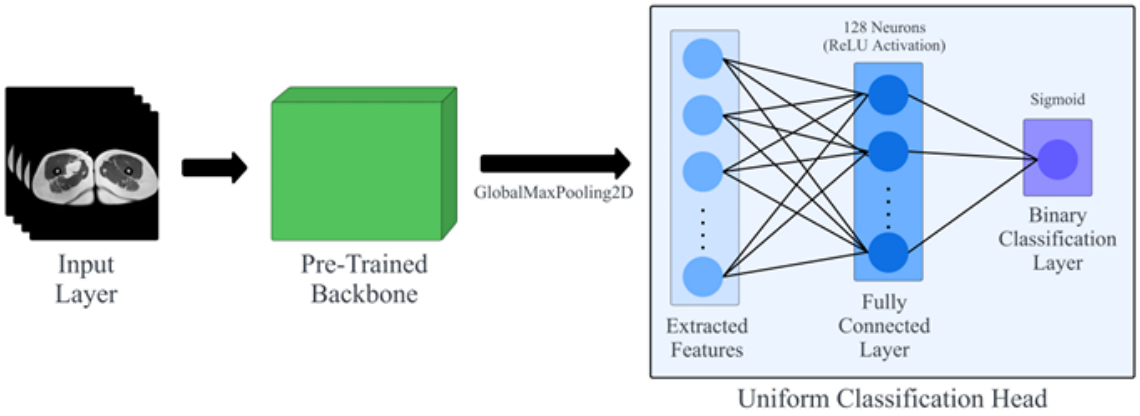


Figure 4.1.1: Transfer learning architecture using frozen pre-trained backbone.

Figure 4.1.1 shows this three-part framework schematically, showing how each frozen backbone feeds into the shared GlobalMaxPooling2D layer and classification head described above. The diagram was prepared for this study to summarise the pipeline just described, rather than being drawn from any external source. Taken together, this three-part design ensures that every model receives the same input images, extracts its own features through its own backbone, and then passes through the exact same classification head. By holding everything else constant and varying only the backbone, any differences observed in the final results can be attributed, with a fair degree of confidence, to the architectural design of the backbone itself, rather than to any other confounding factor.

4.2 Selected CNN Architectures

Five CNN architectures were chosen for this study: ResNet50, DenseNet121, EfficientNetB3, InceptionV3, and MobileNetV2. These five were selected because, between them, they represent a genuinely wide range of design philosophies: residual connections, dense connectivity, compound scaling, multi-scale feature extraction, and lightweight mobile-oriented design. All five are well established in the computer vision literature, have been benchmarked extensively on the ImageNet challenge, and are readily available with pre-trained weights in TensorFlow and Keras. Table 4.2.1 summarises the five architectures, listing the key architectural innovation behind each one along with the approximate parameter count and ImageNet top-1 accuracy reported in the original papers that introduced them [9–13].

The five models differ quite noticeably in parameter count, ImageNet accuracy, and computational cost. MobileNetV2 has only around 3.5 million parameters against ResNet50's roughly 25.6 million, while EfficientNetB3 achieves the highest ImageNet accuracy of the

Table 4.2.1: Five selected CNN architectures with their key architectural features, parameter counts, and ImageNet Top-1 accuracy.

Architecture	**Depth**	**Key Innovation**	**Params (M)**	**Top-1 Acc.**
ResNet50	50 layers	Residual skip connections; bottleneck blocks	∼25.6	∼76.1%
DenseNet121	121 layers	Dense feed-forward connectivity within dense blocks	∼8.1	∼74.9%
EfficientNetB3	∼26 blocks	Compound scaling of depth, width, and resolution	∼12.2	∼81.7%
InceptionV3	∼48 layers	Multi-scale inception modules; factorised convolutions	∼23.9	∼77.9%
MobileNetV2	∼53 layers	Depthwise separable convolutions; inverted residuals	∼3.5	∼72.0%

group at around 81.7%, with MobileNetV2 the lowest at around 72%. This spread is quite deliberate: it lets the study ask whether a larger, more ImageNet-accurate model necessarily performs better on this particular MRI classification task, or whether a smaller, simpler model can hold its own, or perhaps even do better.

4.2.1 ResNet50

Historical Context and Motivation

ResNet50, short for Residual Network with 50 layers, was developed by He et al. [9] at Microsoft Research and presented at CVPR 2016. That year it took first place in the ImageNet competition with a top-5 error rate of 3.57%, surpassing human-level performance for the first time. Before ResNet came along, the prevailing assumption was that simply making a network deeper would keep improving its accuracy. In practice, researchers found that beyond a certain depth, training accuracy would plateau or even worsen, and not because of overfitting either. The real culprit was that by the time gradients travelled all the way back to the earliest layers of a very deep network, they had shrunk to almost nothing, the well-known vanishing gradient problem, which made training such networks effectively very difficult.

Block Architecture

ResNet's answer to this problem is a deceptively simple idea: skip connections, also called residual or shortcut connections. Rather than asking each layer to learn the full transformation from input to output on its own, the network only needs to learn the difference between the two, the residual. Mathematically, each block learns a residual function:

$$\mathcal{F}(\mathbf{x}) = \mathcal{H}(\mathbf{x}) - \mathbf{x} \tag{1}$$

so that the block's final output becomes:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \mathbf{x} \tag{2}$$

The skip connection adds the block's input directly to its output, which means that during backpropagation, gradients have two routes back through the network: through the convolutional layers as usual, and directly through the shortcut. Even the earliest layers, then, continue to receive a meaningful gradient signal and keep learning properly. It is this design that made it possible to train networks hundreds of layers deep without running into the degradation problem described above.

ResNet50 itself uses a bottleneck block design that manages to be both efficient and expressive. Each block consists of three convolutional layers in sequence: a 1x1 convolution that reduces the number of channels, a 3x3 convolution that extracts spatial features from this reduced set of channels, and a second 1x1 convolution that restores the original channel count. The point of this design is that the comparatively expensive 3x3 convolution only ever operates on a reduced number of channels, saving a good deal of computation, while the skip connection bypasses all three layers and adds the original input directly to the block's output.

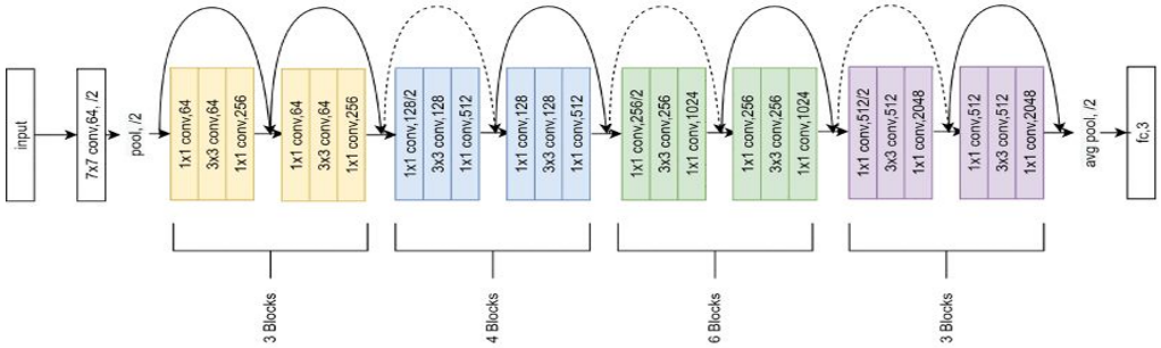


Figure 4.2.1: ResNet50 architecture [29].

Figure 4.2.1, adapted from a study published in *IEEE Access* that applied the same architecture to a face-based ethnicity classification task [29], shows this structure schematically. Although that study's application is entirely different from the medical imaging task addressed in this thesis, the underlying ResNet50 network shown is identical, and the figure is included purely as a visual reference for the architecture described in the text. The full ResNet50 network begins with a 7x7 convolutional layer of 64 filters and stride 2, followed by batch normalisation, a ReLU activation, and a 3x3 max-pooling layer with stride 2, which together reduce the input from 224x224 down to 56x56. From there, the network proceeds through four main stages of residual blocks: Stage 1 with 3 blocks at 64 channels, Stage 2 with 4 blocks at 128 channels, Stage 3 with 6 blocks at 256 channels, and Stage 4 with 3 blocks at 512 channels, with each new stage using strided convolutions to halve the spatial resolution again. Altogether, this gives ResNet50 approximately 25.6 million parameters.

Relevance to Lipomatous Tumour Classification

ResNet50 is a reasonable fit for MRI-based tumour classification precisely because its skip connections help preserve features from different depths of the network at once. In T1-weighted MRI, diagnostically useful information tends to exist at several levels simultaneously: low-level cues such as fat signal brightness and edge sharpness, mid-level

cues such as fibrous septa patterns, and high-level cues such as overall tumour shape. Because each residual stage builds on what came before rather than replacing it outright, this multi-level information remains available right up to the final classification step. This matters in particular for telling apart lipoma from WDLPS, since the two classes often share very similar overall shapes and signal intensities, and the features that actually separate them, such as small internal septations, subtle nodular thickening, or slight signal heterogeneity, tend to sit at these lower and middle levels of the network rather than at the very top. A network that loses track of these finer details by the time it reaches its final layers would struggle to make this distinction reliably.

ResNet50 has also been extensively validated in the broader medical imaging literature, including tumour classification and organ segmentation tasks, which gives some added confidence that it is a reliable and appropriate baseline for this study. Its bottleneck design further means that this depth and feature richness comes without an excessive computational cost, which made it practical to train within the frozen transfer learning setup used here, alongside the other four architectures being compared.

4.2.2 DenseNet121

Historical Context and Motivation

DenseNet121 belongs to the Densely Connected Convolutional Network (DenseNet) family, proposed by Huang et al. [10] and presented at CVPR 2017, where it received the Best Paper Award. In a sense, DenseNet takes the idea behind ResNet's skip connections and pushes it a good deal further. Rather than connecting a layer only to the one immediately before it, or adding a single skip from a few layers back, DenseNet connects every layer directly to every other layer that follows it within a block. The motivation behind this was the observation that networks with shorter paths between individual layers and the final loss function tend to be easier to train, converge faster, and generalise better, an advantage that matters all the more on smaller datasets, such as the one used here.

Block Architecture

In an ordinary CNN, each layer receives input only from the layer directly before it. In ResNet, each layer receives that same input plus one additional skip connection. In DenseNet, by contrast, each layer receives the feature maps of every preceding layer, concatenated together:

$$\mathbf{x}_\ell = H_\ell\big([\mathbf{x}_0,\, \mathbf{x}_1,\, \ldots,\, \mathbf{x}_{\ell-1}]\big) \tag{3}$$

where $[\cdot]$ denotes channel-wise concatenation of feature maps and $H_\ell(\cdot)$ is the composite transformation performed by the ℓ-th layer (batch normalisation, ReLU, then convolution, in that order). For a dense block containing L layers, this scheme creates $L(L+1)/2$ direct connections, compared to just L in an ordinary CNN.

This dense connectivity brings three practical benefits worth highlighting. First, it addresses the vanishing gradient problem directly, since gradients can now flow between any two layers without having to pass through every layer in between. Second, it encourages feature reuse: because earlier features remain available to every later layer, the network has no need to relearn them from scratch, which makes DenseNet remarkably parameter-efficient, with DenseNet121 achieving competitive accuracy on only about 8.1 million parameters,

roughly a third of ResNet50's count. Third, since every layer receives a mixture of low-level and high-level features together, the network is naturally encouraged to build richer, more varied representations than it otherwise might.

Architecturally, DenseNet121 begins with a 7x7 convolutional stem, much like ResNet50, followed by four dense blocks separated by transition layers. Each transition layer uses a 1x1 convolution to reduce the number of channels and a 2x2 average pooling operation to reduce the spatial size. The four dense blocks contain 6, 12, 24, and 16 layers respectively, with each layer inside a block following a pre-activation bottleneck design. A growth rate of $k = 32$ governs how many new feature maps each layer contributes to the concatenated input that later layers receive. Figure 4.2.2, reproduced from a scientific figure associated with a study on hybrid transfer learning for soil image classification [30], illustrates this dense block and transition layer structure, depicting the same DenseNet121 architecture used in this thesis, and is included here purely to support the description of its dense connectivity pattern.

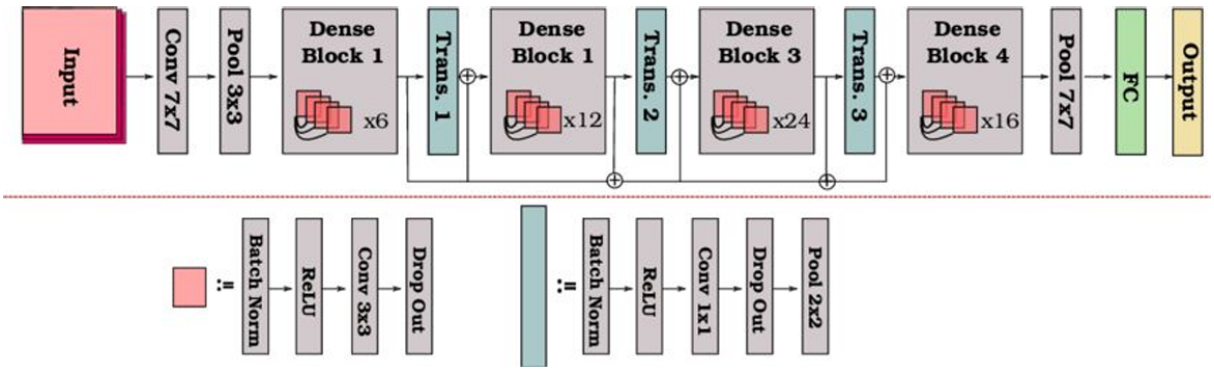


Figure 4.2.2: DenseNet121 architecture [30].

Relevance to Lipomatous Tumour Classification

DenseNet121 lends itself reasonably well to MRI-based medical image classification, largely because it is so effective at preserving fine-grained, subtle textural features across the depth of the network. In T1-weighted MRI, the distinction between lipoma and WDLPS often comes down to fairly subtle signals, thick fibrous septa, small non-fat nodules, or a heterogeneous fat signal, all of which are low-level textural features that could easily be lost as they pass through a deep network. DenseNet121's dense connectivity keeps these features accessible all the way through to the final classification step, even under the frozen feature-extraction setup used in this study, and its modest parameter count further reduces the risk of overfitting when training on a dataset of this size.

4.2.3 EfficientNetB3

Historical Context and Motivation

EfficientNet was introduced by Tan and Le [11] from Google Brain and presented at ICML 2019, in an attempt to answer a fairly fundamental question: what is the best way to scale up a neural network to gain accuracy? Earlier work tended to increase either the number of layers (depth), the number of channels (width), or the input image size (resolution), usually one at a time, and without any particularly principled strategy for doing so. Tan

and Le showed that scaling all three dimensions together, in a balanced and carefully controlled way, yields noticeably better accuracy for a given amount of computation than adjusting any single dimension in isolation. This makes intuitive sense once one considers that depth, width, and resolution are not really independent of one another: a deeper network benefits more from larger input images, and a wider network is better placed to make use of higher-resolution inputs.

Block Architecture

The compound scaling method uses a single coefficient ϕ to scale network depth (d), width (w), and input resolution (r) together:

$$d = \alpha^{\phi} \tag{4}$$

$$w = \beta^{\phi} \tag{5}$$

$$r = \gamma^{\phi} \tag{6}$$

where α, β, and γ are constants determined through a grid search on the base model, EfficientNet-B0, subject to the constraint:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \quad \alpha \geq 1,\ \beta \geq 1,\ \gamma \geq 1 \tag{7}$$

This constraint keeps the total computational cost, measured in FLOPs, growing at roughly 2^{ϕ} as ϕ increases. For EfficientNet-B0, the values found were $\alpha = 1.2$, $\beta = 1.1$, and $\gamma = 1.15$.

The base architecture, EfficientNet-B0, was itself designed using Neural Architecture Search (NAS) to identify the most effective building-block configuration for a fixed computational budget. EfficientNetB3 is simply this base model scaled with $\phi = 3$, giving it a native input resolution of 300x300 pixels, along with additional layers and wider channels throughout. Its core building block, the MBConv block, was originally introduced in MobileNetV2 and consists of four parts: a 1x1 convolution that expands the number of channels, a depthwise convolution that extracts spatial features on a per-channel basis, a Squeeze-and-Excitation (SE) attention module that learns which channels matter most, and a final 1x1 convolution that restores the original channel count. Where the input and output channel dimensions match, a residual connection links the two directly. Figure 4.2.3, adapted from a scientific figure originally published alongside a study on pistachio nut cultivar identification [31], shows this MBConv-based, compound-scaled structure. The underlying network architecture depicted is identical to the EfficientNetB3 backbone used in this thesis, and is reproduced here purely to provide a clear visual reference.

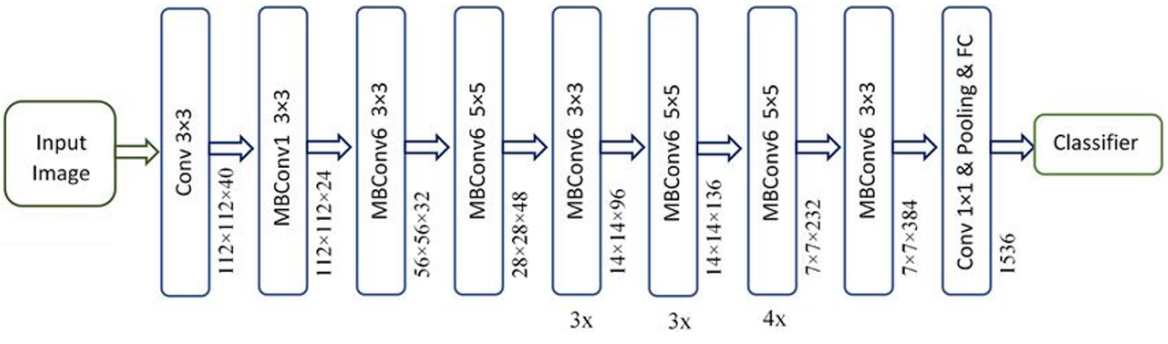


Figure 4.2.3: EfficientNetB3 architecture [31].

Relevance and Limitations for Lipomatous Tumour Classification

EfficientNetB3 achieves very strong ImageNet accuracy with a comparatively modest 12.2 million parameters, which makes it, on paper, an attractive candidate for this task. That said, it carries some notable limitations once applied to MRI classification under a frozen backbone. Its compound scaling was optimised for natural colour photographs, whose statistical properties differ quite substantially from those of T1-weighted MRI: MRI images are grayscale, relatively uniform in their broad visual patterns, and encode tissue properties through signal intensity rather than colour or surface texture. These differences between the two domains may well limit how useful EfficientNetB3's frozen features turn out to be for MRI-based classification.

There is also the matter of a resolution mismatch. EfficientNetB3 was designed with 300x300 pixel inputs in mind, yet in this study, as with every other architecture, all images were resized to 224x224 pixels to keep the comparison fair across the board. This reduction in resolution disturbs the compound scaling balance, since the network's depth and width were originally calibrated on the assumption of a 300x300 input. Just how much this affects classification performance is examined in detail in Chapter 5.

4.2.4 InceptionV3

Historical Context and Motivation

InceptionV3 was developed by Szegedy et al. [12] at Google and presented at CVPR 2016. It represents the third generation of the Inception architecture, a lineage that began with GoogLeNet, or InceptionV1, back in 2014. The Inception family set out to answer a fairly practical question: what is the ideal filter size for a convolutional layer to use? Conventional CNNs typically settle on a single filter size, say 3x3 or 5x5, and apply it uniformly throughout a given layer. But visual information does not present itself at a single spatial scale, fine textures call for small filters, while larger structures call for bigger ones, and no single filter size is well suited to capturing both at once.

Block Architecture

Inception's answer is to apply several filter sizes at once, within the same layer. Each Inception module runs 1x1, 3x3, and 5x5 convolutions in parallel, alongside a max-pooling branch, and then concatenates the outputs of all these branches along the channel dimension:

$$\mathbf{y} = \left[\text{Conv}_{1\times1}(\mathbf{x}),\ \text{Conv}_{3\times3}(\mathbf{x}),\ \text{Conv}_{5\times5}(\mathbf{x}),\ \text{MaxPool}(\mathbf{x})\right] \tag{8}$$

where $[\cdot]$ again denotes channel-wise concatenation. This allows the network to pick up fine local detail through the 1x1 branch, medium-scale patterns through the 3x3 branch, and coarser, large-scale structure through the 5x5 branch, all within a single layer.

InceptionV3 improves on its predecessors mainly by factorising larger convolutions into smaller ones: a 5x5 convolution, for instance, is replaced by two 3x3 convolutions in sequence, achieving the same effective receptive field with fewer parameters. In much the same spirit, an NxN convolution can be split into a 1xN convolution followed by an Nx1 convolution. These asymmetric convolutions turn out to be particularly good at picking up elongated structures, which may well be relevant to the fibrous features associated

with WDLPS on MRI. InceptionV3 also incorporates batch normalisation throughout and uses label smoothing during training to aid generalisation. Figure 4.2.4, adapted from a study in the *International Journal of Advanced Computer Science and Applications* on pneumonia detection using transfer learning [32], shows the resulting inception-module structure. While that study addressed a different classification task, the InceptionV3 network depicted is the same one used in this thesis, and the figure is reproduced here to support the architectural description given above.

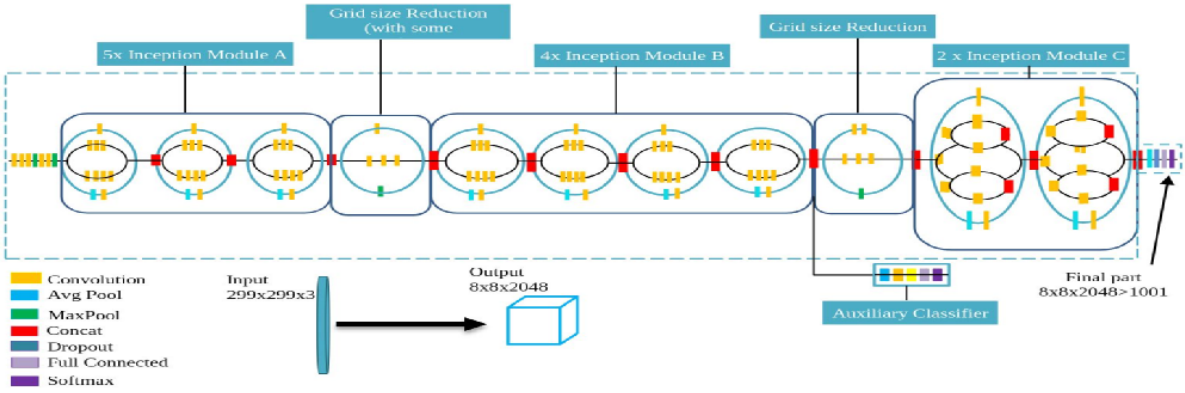


Figure 4.2.4: InceptionV3 architecture [32].

Relevance to Lipomatous Tumour Classification

InceptionV3 is a fairly sensible choice for lipoma-versus-WDLPS classification, given that the task itself involves features spanning several spatial scales at once. Small- scale features include subtle variations in fat signal intensity; medium-scale features include thick fibrous septa; and large-scale features include the overall shape and boundary of the tumour. Since InceptionV3's inception modules capture all of these scales within a single convolutional stage, it is, in principle, better positioned for this kind of task than a network relying on just one filter size per layer, and its factorised convolutions make this multi-scale extraction reasonably efficient without demanding an excessive number of additional parameters.

4.2.5 MobileNetV2

Historical Context and Motivation

MobileNetV2 was introduced by Sandler et al. [13] from Google and presented at CVPR 2018. It belongs to the MobileNet family, designed from the outset for devices with limited computing power, such as mobile phones, tablets, and other embedded systems where memory and battery life impose real constraints on model size. The original MobileNet (V1), released in 2017, introduced depthwise separable convolutions as a way of reducing the cost of standard convolutions while retaining broadly similar performance, and it quickly became a popular backbone for on-device vision applications. MobileNetV2 retains this core idea and builds on it with two further innovations: inverted residuals and linear bottlenecks, both aimed at squeezing more representational power out of a network without adding much in the way of extra computation. The end result is a model that reaches roughly 72% top-1 accuracy on ImageNet with only 3.5 million parameters and around 300 million multiply-add operations per forward pass, making it approximately 12 times

more computationally efficient than ResNet50, while still remaining competitive with it in terms of raw classification accuracy on many benchmark tasks.

Block Architecture

The first key idea behind MobileNetV2 is the depthwise separable convolution. A standard convolution with a $D_K \times D_K$ kernel, applied to M input channels to produce N output channels, requires $D_K^2 \cdot M \cdot N$ operations per spatial location, a cost that grows quickly as the number of channels increases. Depthwise separable convolution splits this single operation into two smaller steps: a depthwise convolution that applies one filter per input channel independently, at a cost of $D_K^2 \cdot M$ operations, followed by a 1x1 pointwise convolution that mixes information across channels, at a cost of $M \cdot N$ operations. The combined cost, $D_K^2 \cdot M + M \cdot N$, works out to roughly nine times fewer operations than a standard 3x3 convolution offering comparable representational power, and it is this saving, repeated across every layer of the network, that accounts for most of MobileNetV2's overall efficiency.

The second key idea is the inverted residual block with a linear bottleneck. Where the ResNet bottleneck first compresses channels and then expands them, MobileNetV2 does the opposite: it expands the channels first, applies the depthwise convolution within this wider space, and only then compresses back down:

$$\text{Expand} \xrightarrow{1\times1,t\cdot C} \text{Depthwise-Conv} \xrightarrow{3\times3} \text{Linear-Project} \xrightarrow{1\times1,C} \tag{9}$$

where t is the expansion factor (typically 6) and C is the number of output channels. Expanding the channels first gives the depthwise convolution considerably more room to work with, improving its expressive power, while the final 1x1 projection uses a linear activation rather than ReLU, an important detail, since ReLU can destroy useful information when applied within a low-dimensional space, effectively collapsing part of the feature space by zeroing out negative values. Keeping this final activation linear preserves that information intact, allowing the compressed bottleneck representation to remain a faithful, information-rich summary of the wider expanded space that came before it. As with ResNet, a residual connection is added wherever the input and output channel dimensions happen to match, which helps gradients flow more easily through the many stacked blocks that make up the full network.

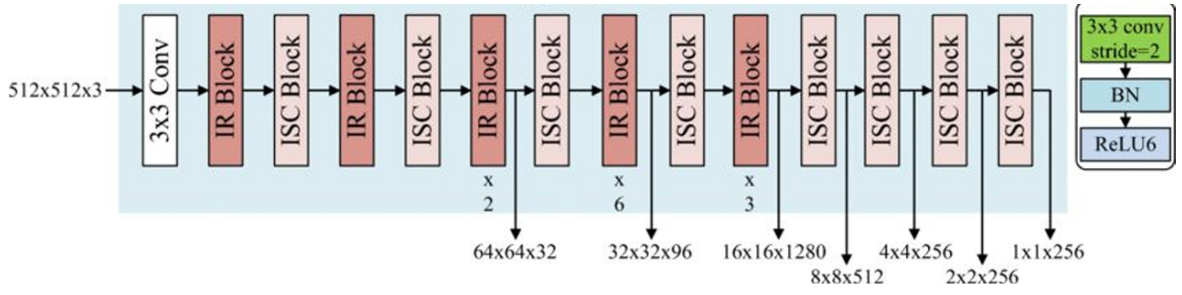


Figure 4.2.5: MobileNetV2 architecture [33].

Figure 4.2.5, reproduced from a scientific figure associated with a study on lightweight multi-object tracking for embedded systems [33], shows this inverted-residual, depthwise-separable structure. The MobileNetV2 backbone illustrated is identical to the one used

in this thesis, and the figure is included solely as a visual reference for the architecture discussed above, rather than being tied in any way to the tracking application described in that particular study. MobileNetV2 begins with a standard 3x3 convolutional layer that produces 32 feature maps from a 224x224x3 input, followed by 17 inverted residual blocks arranged across six stages of progressively widening channel counts: 16, 24, 32, 64, 96, 160, and 320. Spatial downsampling between stages is handled by strided depthwise convolutions, keeping the overall computational budget low even as the spatial resolution shrinks and the number of channels grows, and a final 1x1 convolution expands the resulting feature maps to 1,280 channels before they reach the pooling and classification stages.

Relevance to Lipomatous Tumour Classification

MobileNetV2 turned out to be the best-performing architecture of all five evaluated in this study, reaching 96.7% accuracy and an AUROC of 0.997, despite being by some distance the smallest and simplest model in the comparison, with roughly a seventh of the parameters of ResNet50. This is a genuinely noteworthy finding, since it shows that a lightweight architecture designed primarily for mobile devices can, in fact, outperform considerably larger models on a non-trivial medical imaging task, and it runs somewhat against the intuitive assumption that more parameters and greater architectural complexity should translate fairly directly into better diagnostic performance. Its depthwise separable convolutions and information-preserving linear bottlenecks appear to yield a feature space that is particularly well suited to separating lipoma from WDLPS, and this without any fine-tuning of the backbone whatsoever, suggesting that the generic features MobileNetV2 learned from ImageNet transfer unusually well to this specific MRI classification task. MobileNetV2 is also an appealing option from a practical, clinical standpoint, since its small size allows it to run efficiently on standard hardware without requiring expensive GPU infrastructure, an advantage that could matter a good deal in resource-limited clinical settings where access to high-end computing equipment is not always guaranteed.

4.3 Uniform Classification Head

For the comparison between the five architectures to be genuinely fair, every model needs to be evaluated under conditions that are as close to identical as possible. To this end, a single standardised classification head is placed on top of each pre-trained backbone and kept exactly the same across all five models, so that any observed difference in results can be attributed only to the backbone itself. This head is made up of three components, connected in sequence.

4.3.1 GlobalMaxPooling2D

The first component is a GlobalMaxPooling2D layer. It takes the three-dimensional feature map tensor $\mathbf{F} \in \mathbf{text}R^{H' \times W' \times C'}$ produced by the backbone's final convolutional layer and converts it into a one-dimensional feature vector $\mathbf{v} \in \mathbf{text}R^{C'}$, by taking, for each channel, the maximum value across all spatial positions:

$$v_c = \max_{h,w} F_{h,w,c} \quad \forall\, c \in \{1, \ldots, C'\} \tag{10}$$

where $F_{h,w,c}$ denotes the activation at position (h, w) in channel c. In effect, this picks out the single strongest response for each feature type, wherever it happens to occur in the image. GlobalMaxPooling2D was chosen over the more commonly used GlobalAveragePooling2D precisely because it highlights the most prominent discriminative feature rather than averaging it away. In MRI tumour classification, a diagnostically important clue, a single thick septum, say, or a small non-fat nodule, may occupy only a small portion of the image, and taking the maximum helps ensure that this kind of localised evidence is retained rather than diluted.

4.3.2 Fully Connected Dense Layer

The second component is a dense layer of 128 neurons with a ReLU activation. It takes the feature vector $\mathbf{v}$ produced by GlobalMaxPooling2D and transforms it as follows:

$$\mathbf{h} = \mathrm{ReLU}(\mathbf{W}_1\mathbf{v} + \mathbf{b}_1) \tag{11}$$

where $\mathbf{W}_1 \in \mathbf{R}^{128\times C'}$ and $\mathbf{b}_1 \in \mathbf{R}^{128}$ are the learned weights and biases. This layer maps the raw backbone features into a 128-dimensional space better suited to the binary classification task at hand. A size of 128 neurons was chosen because it offers enough capacity to capture the mapping between the feature space and the classification boundary, while still being modest enough to guard against overfitting on a training set of this size. ReLU, for its part, was chosen because it introduces non-linearity efficiently and helps keep the vanishing gradient problem at bay during training.

4.3.3 Sigmoid Output Layer

The third and final component is a single output neuron with a sigmoid activation:

$$\hat{p} = \sigma(\mathbf{w}_2^\top \mathbf{h} + b_2) = \frac{1}{1 + e^{-(\mathbf{w}_2^\top \mathbf{h} + b_2)}} \tag{12}$$

where $\mathbf{w}_2 \in \mathbf{R}^{128}$ and $b_2 \in \mathbf{R}$ are learned parameters, and $\hat{p} \in (0, 1)$ represents the predicted probability that a given image belongs to the WDLPS class. If $\hat{p}$ exceeds 0.5, the image is classified as WDLPS; if it is 0.5 or below, it is classified as Lipoma. Because the sigmoid function outputs a continuous probability rather than a hard label, it also makes it possible to compute the AUROC score, which depends on exactly this kind of continuous output.

4.4 Training Setup

All five models were trained under exactly the same settings: the same optimiser, learning rate, batch size, number of epochs, loss function, and callbacks. Keeping the training setup identical in this way matters a great deal, since it is precisely what allows any difference in performance between the models to be attributed to their architecture, rather than to some incidental difference in how they happened to be trained.

4.4.1 Hyperparameters

Optimisation Algorithm

All five models were optimised using Adam [34], short for Adaptive Moment Estimation, which remains one of the most widely used optimisers in deep learning today. Adam improves on earlier gradient-descent methods by keeping a running estimate of both the mean and the variance of the gradient for each parameter, which allows it to adapt the effective learning rate on a per-parameter basis, a property that turns out to be useful when gradients are noisy or vary considerably across different parts of the network, as they often do.

Given the gradient g_t at time step t, Adam maintains biased moment estimates:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\, g_t \tag{13}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)\, g_t^2 \tag{14}$$

and, to correct for the initialisation bias present in early training iterations, computes bias-corrected versions of these estimates:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{15}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{16}$$

The model's parameters are then updated as follows:

$$\theta_t = \theta_{t-1} - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{17}$$

where α is the learning rate, β_1 and β_2 are the decay rates governing the first and second moment estimates respectively, and ϵ is a small constant added purely for numerical stability. This study used Adam's standard settings throughout: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-7}$. Adam is a sensible choice here, given how well it tends to cope with the class imbalance and noisy gradients that often accompany training on medical image datasets.

Learning Rate

The initial learning rate was set to $\alpha = 1 \times 10^{-4}$, a fairly standard starting point for transfer learning with a frozen backbone. Too high a learning rate risks destabilising the classification head, particularly at the outset when its weights are still randomly initialised; too low a rate, on the other hand, makes for needlessly slow training. A value of 1×10^{-4} offers a reasonably well-tested middle ground between the two, and was further adjusted automatically over the course of training by the ReduceLROnPlateau callback, described in Section 4.4.2.

Batch Size and Training Duration

All models were trained with a batch size of 64 images per update. A smaller batch size introduces more noise into the gradient estimate, which can sometimes aid generalisation, whereas a larger batch size produces more stable gradients and makes better use of available GPU resources. A batch size of 64 strikes a reasonable balance between these two

considerations and is a fairly common choice in medical imaging studies of a comparable scale.

The maximum number of training epochs was capped at 300, a figure chosen generously enough to ensure that even the slowest-converging model would have ample time to reach its best achievable performance. In practice, every model finished training well before this limit was reached, since the EarlyStopping callback intervened first, as described in Section 4.4.2.

Loss Function

The loss function used throughout training is Binary Cross-Entropy (BCE), sometimes also called log loss, and the standard choice for binary classification problems in which the model outputs a probability. It is defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)\Big] \tag{18}$$

where N is the number of samples in a given batch, $y_i \in \{0, 1\}$ is the true label for sample i (0 for Lipoma, 1 for WDLPS), and $\hat{p}_i$ is the model's predicted probability that sample i is WDLPS. BCE penalises confident wrong predictions heavily while barely penalising confident correct ones, and the lower the resulting BCE value, the better the model can be said to be performing.

A fixed random seed of 42 was used throughout every random operation in this study, covering dataset splitting, data augmentation, and the weight initialisation of the classification head. This was done to keep the experiments fully reproducible, and to ensure that any differences observed between models could not simply be put down to random variation. Table 4.4.1 summarises the core training hyperparameters just described, all of which were applied identically across the five architectures evaluated in this study, with the Adam optimiser itself following the formulation originally proposed by Kingma and Ba [34].

Table 4.4.1: Hyperparameters.

Hyperparameter	**Value / Setting**
Optimiser	Adam
Initial Learning Rate	1×10^{-4}
Batch Size	64
Maximum Epochs	300
Loss Function	Binary Cross-Entropy (BCE)

4.4.2 Callbacks and Regularisation

Three Keras callbacks were used during training to make the process more reliable, to guard against overfitting, and to avoid wasting computation on training that had already

converged. All three monitor the validation loss after every epoch and act automatically whenever certain conditions are met.

ModelCheckpoint

The first callback, ModelCheckpoint, monitors validation loss after each epoch and saves the model's weights to disk only when the current value improves on the best seen so far. This way, even if a model begins to overfit or its performance dips in later epochs, the best version from earlier in training is preserved rather than lost. This matters quite a bit in a medical imaging context, where even a modest degree of overfitting can compromise how reliable a model's predictions are on genuinely new patients.

ReduceLROnPlateau

The second callback, ReduceLROnPlateau, watches validation loss and automatically lowers the learning rate once improvement stalls. Specifically, if validation loss fails to improve by more than $\Delta_{\text{min}} = 1 \times 10^{-4}$ over 10 consecutive epochs, the learning rate is reduced by a factor of 0.1, so a learning rate starting at 1×10^{-4} would drop to 1×10^{-5} after the first detected plateau, and to 1×10^{-6} after a second. This gives the model room to make finer adjustments to its weights later in training, and can sometimes help it settle into a better solution than a fixed learning rate would allow.

EarlyStopping

The third callback, EarlyStopping, also monitors validation loss, and halts training altogether if it fails to improve by more than $\Delta_{\text{min}} = 1 \times 10^{-4}$ over 20 consecutive epochs. This keeps the model from training for longer than is useful and overfitting the training data in the process. A patience of 20 epochs gives the model enough room to recover from any temporary plateau that might follow a learning rate reduction, while still bringing training to a close reasonably promptly once it has clearly converged. Between them, these three callbacks make for a training process that is both efficient and reasonably well protected against overfitting.

4.5 Evaluation Metrics

Each trained model's performance on the test set was measured using six metrics: accuracy, precision, recall, F1-score, specificity, and AUROC. Relying on several metrics together gives a far more complete picture of how a model is actually performing than accuracy alone could offer, particularly given that the two classes in this study are not equally represented. Throughout, WDLPS is treated as the positive class, in keeping with the standard clinical convention of defining the condition being screened for as positive, and a missed malignant case (a false negative) is considered substantially more serious than an incorrectly flagged benign case (a false positive).

4.5.1 Accuracy

Accuracy captures the proportion of predictions the model got right, out of all predictions made:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

where TP is the number of true positives (WDLPS correctly classified as WDLPS), TN the number of true negatives (Lipoma correctly classified as Lipoma), FP the number of false positives (Lipoma incorrectly classified as WDLPS), and FN the number of false negatives (WDLPS incorrectly classified as Lipoma). Accuracy is easy to interpret, but it can be somewhat misleading whenever one class outnumbers the other, as is the case here: the test set contains 1,611 WDLPS images against 1,088 Lipoma images, meaning a model that simply predicted WDLPS every single time would already reach around 59.7% accuracy without having learned anything useful whatsoever. This is precisely why the additional metrics below are needed alongside it.

4.5.2 Precision

Precision captures how reliable the model's positive predictions actually are:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{20}$$

A high precision means that, when the model predicts WDLPS, it is usually right to do so, an important property clinically, since it helps limit the number of patients subjected to unnecessary surgical procedures for what is, in reality, a benign lipoma.

4.5.3 Recall

Recall captures how many of the genuinely malignant WDLPS cases the model managed to identify correctly:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{21}$$

Of all six metrics considered in this study, recall arguably matters most from a safety standpoint. A missed WDLPS case means a patient who does not receive the treatment they need, with the tumour potentially left to grow or progress further over time.

4.5.4 Specificity

Specificity captures how many of the genuinely benign Lipoma cases the model correctly recognised as such:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{22}$$

A high specificity matters because it helps avoid subjecting patients who do not, in fact, have a malignant tumour to unnecessary procedures. Taken together with recall, it gives a fairly complete picture of how well a model handles both classes, rather than just the one it is being screened for.

4.5.5 F1-Score

The F1-score brings precision and recall together into a single balanced measure:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{23}$$

Being the harmonic mean of the two, it produces a low score whenever either precision or recall is weak, even if the other happens to be strong. An F1-score close to 1.0 therefore indicates that a model is performing well on both fronts at once, which makes it a useful, single-number summary for evaluating models under class imbalance.

4.5.6 AUROC

The Area Under the Receiver Operating Characteristic Curve (AUROC) measures how well a model separates the two classes across every possible classification threshold, rather than at just one. The ROC curve itself plots the true positive rate against the false positive rate at each threshold τ:

$$\text{FPR}(\tau) = \frac{FP(\tau)}{FP(\tau) + TN(\tau)} \tag{24}$$

and AUROC is simply the area beneath this curve:

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{25}$$

An AUROC of 1.0 represents perfect separation between the two classes, while 0.5 represents performance no better than random guessing. Because it does not depend on any single decision threshold, AUROC offers a genuinely fair basis for comparing models regardless of which threshold each one happens to use in practice, and it copes well with class imbalance too. Put simply, an AUROC of ρ means that, given a randomly chosen WDLPS case and a randomly chosen Lipoma case, the model will assign the higher probability score to the WDLPS case with probability ρ.

Chapter-5 : Results

This chapter presents, in a detailed and organised manner, the experimental results obtained after benchmarking five pre-trained Convolutional Neural Network (CNN) architectures, namely ResNet50, DenseNet121, EfficientNetB3, InceptionV3, and MobileNetV2, for the binary classification of benign lipoma and well-differentiated liposarcoma (WDLPS) using T1-weighted MRI data taken from the publicly available WORC database. Every model was evaluated under a strictly uniform transfer learning framework, so that the comparison between the architectures remains fair and is not affected by differences in the training configuration of one model or another.

The results are organised into two main sections. Section 5.1 presents the overall performance comparison across all five architectures using a standardised set of evaluation metrics: accuracy, precision, recall, F1-score, specificity, and AUROC. All of these metrics have been recalculated directly from the confusion matrix counts obtained on the test set, and are reported to three decimal places throughout this chapter so that the differences between architectures, which are sometimes fairly small, can be seen clearly. This section also discusses class-wise performance, training dynamics, and the relative strengths and weaknesses of each model. Section 5.2 then goes on to provide a more detailed, model-by-model analysis, covering per-class classification reports, confusion matrices, training and validation curves, and an interpretation of the results in the specific context of lipomatous tumour classification.

Before proceeding, it should be clarified that the term *positive class* throughout this study refers to WDLPS, the malignant class, since the clinical focus of this work is on ensuring that malignant tumours are not missed. A false negative, that is, predicting WDLPS as benign lipoma, carries a much higher clinical risk than a false positive, since a missed malignancy can have serious consequences for the patient, whereas an unnecessary biopsy, while certainly undesirable, is comparatively less severe. For this reason, recall (sensitivity) for the WDLPS class is treated as being just as important as overall accuracy and AUROC when judging a model's suitability, and this view is reflected throughout the discussion that follows.

It is also worth mentioning here that a data-integrity check was carried out on the original evaluation tables prepared for this study. It was found that the specificity values reported earlier had, by mistake, been set equal to the macro-averaged recall figure for each model rather than being computed as $TN/(TN+FP)$, which is the correct definition of specificity. This error has since been corrected throughout the present chapter, and in some cases, most notably for EfficientNetB3, the corrected value is substantially different from what was reported before. All figures presented from this point onward are the corrected, verified values.

5.1 Overall Performance Comparison

Comparing all five architectures on the held-out test set reveals a fairly large spread in performance. As Table 5.1.1 shows, four of the models, MobileNetV2, DenseNet121, ResNet50, and InceptionV3, achieved good classification accuracy, ranging from 95.2% to 96.7%, every value recalculated directly from the confusion matrix counts and cross-checked for consistency. This indicates that transfer learning from ImageNet pre-trained weights is, on the whole, a highly effective strategy for this task. EfficientNetB3, however, performed quite poorly at only 63.5% accuracy, barely above what a random classifier would achieve (50%), a result investigated in detail in Section 5.2.3.

Table 5.1.1: Comprehensive performance comparison of all CNN architectures on the test set.

Architecture	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Specificity**	**AUROC**
ResNet50	0.956	0.954	0.955	0.954	0.948	0.994
DenseNet121	0.963	0.963	0.960	0.961	0.944	0.996
EfficientNetB3	0.635	0.734	0.550	0.483	0.115	0.672
InceptionV3	0.952	0.951	0.949	0.950	0.934	0.991
MobileNetV2	**0.967**	**0.965**	**0.966**	**0.965**	**0.964**	**0.997**

5.1.1 Accuracy Analysis

Accuracy, the proportion of correctly classified instances out of the total, offers the most straightforward view of performance. Figure 5.1.1 shows this comparison as a simple bar chart, making the sharp drop for EfficientNetB3 against the tightly clustered bars of the other four models easy to spot at a glance. The test set contained 1,088 Lipoma images and 1,611 WDLPS images, a total of 2,699 samples, and accuracy ranged over 33.2% points, from 63.5% for EfficientNetB3 to 96.7% for MobileNetV2, a gap wide enough to show that the choice of architecture has a very significant bearing on the outcome here.

As Figure 5.1.1 highlights, the dark green bar represents MobileNetV2, which achieved the highest accuracy among all the evaluated models, confirming its position as the strongest performer overall. The blue bar represents DenseNet121, which came second, while the yellow bar represents ResNet50, which ranked third, indicating that these two models also performed well but achieved slightly lower accuracy than MobileNetV2. The gray bar corresponds to InceptionV3, which scored fourth, reflecting its moderate but still respectable performance compared with the other models. Finally, the red bar represents EfficientNetB3, which obtained the lowest accuracy score in this study, falling well behind the other four architectures. The colour gradient from dark green to red provides a clear visual representation of the models' performance, where darker green indicates better accuracy and red indicates the lowest accuracy, making it easy to compare all five architectures at a glance.

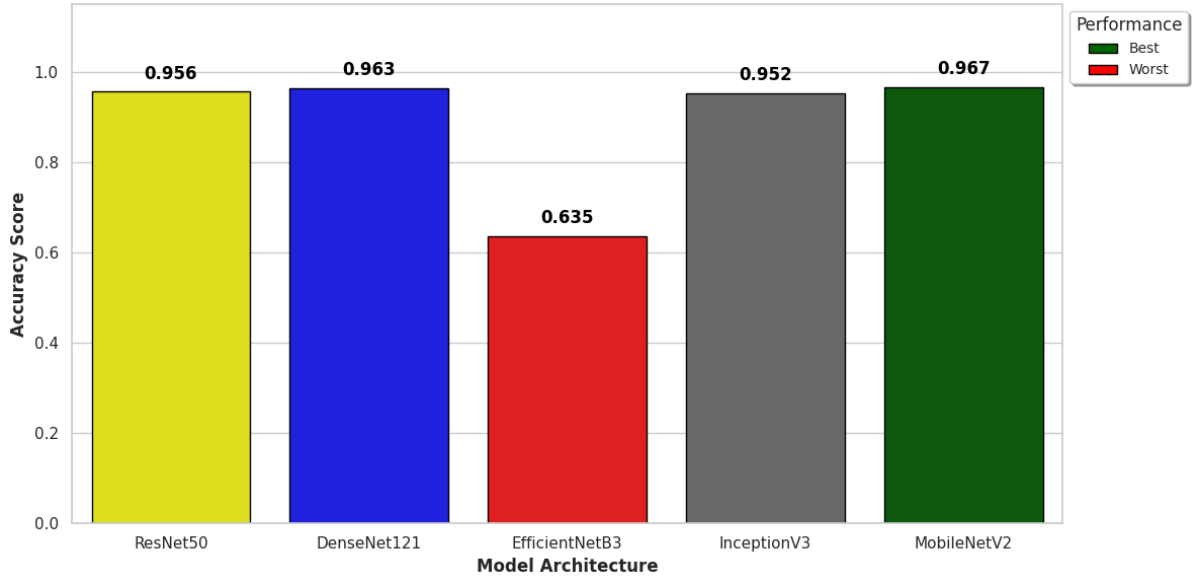


Figure 5.1.1: Bar chart comparing overall test accuracy across all five CNN architectures.

Among the four high performers, MobileNetV2 led at 96.7%, followed by DenseNet121 at 96.3%, ResNet50 at 95.6%, and InceptionV3 at 95.2%, a gap of under 1.5% points that suggests any of the four could, in principle, be considered for clinical use. The gap between these four and EfficientNetB3, by contrast, is quite stark, and points to a serious architectural incompatibility examined more closely in Section 5.2.3.

It is also worth keeping in mind that the class imbalance in the test set, where WDLPS makes up roughly 59.7%, can inflate raw accuracy somewhat: a trivial classifier that always predicted WDLPS would already reach around 59.7%. EfficientNetB3, at 63.5%, sits only barely above this floor, a strong sign that it has not really learned to discriminate between the two classes at all, while the top four models sit far above this baseline, confirming their high accuracy reflects genuine discriminative ability rather than exploiting the imbalance.

5.1.2 Precision, Recall, and F1-Score Analysis

Precision measures what fraction of a model's malignancy predictions are actually correct, so high precision means fewer benign lipomas are wrongly flagged for surgery. Recall (sensitivity) measures what fraction of truly malignant WDLPS cases were correctly picked up, and is the more safety-critical of the two, since missing a malignant tumour is a far more serious error than an unnecessary biopsy. Both are reported alongside F1-score for every architecture in Table 5.1.1.

F1-score, the harmonic mean of precision and recall, penalises any large gap between the two. Among the top four architectures, macro F1-scores range from 0.950 (InceptionV3) to 0.965 (MobileNetV2), all comfortably within clinically excellent territory. EfficientNetB3, by contrast, recorded a critically low F1-score of 0.483, largely because its macro recall of 0.550 was pulled down by a near-zero recall of just 0.115 on the Lipoma class, discussed further in Section 5.2.3.

5.1.3 Specificity Analysis

Specificity, the proportion of true negatives among all actual negatives (here, benign lipomas), measures a model's ability to correctly recognise a benign tumour as benign. High specificity matters since it directly relates to how many patients with benign lipomas are spared an unnecessary procedure. The specificity column in Table 5.1.1 is recalculated as $TN/(TN + FP)$ rather than taken from the earlier, uncorrected drafts of this chapter.

Once corrected, specificity tells a rather different, more revealing story than before. MobileNetV2 leads at 0.964, followed by ResNet50 at 0.948, DenseNet121 at 0.944, and InceptionV3 at 0.934. Interestingly, once calculated correctly, ResNet50 edges out DenseNet121 on this one metric (0.948 against 0.944), even though DenseNet121 remains ahead overall once recall, precision, and AUROC are considered together. EfficientNetB3's true specificity is only 0.115, meaning barely one in ten benign lipomas were correctly identified, drastically lower than the 0.550 reported previously, and consistent with its Lipoma recall already noted above. This confirms that EfficientNetB3's bias toward predicting malignancy is even more severe than earlier figures suggested.

5.1.4 AUROC Analysis

AUROC captures a model's overall ability to discriminate between the two classes across every possible decision threshold, rather than at one fixed point, with 1.0 representing perfect discrimination and 0.5 no better than random guessing. Since it depends on the full distribution of predicted probabilities rather than the confusion matrix, these values are reported as originally computed from the model's raw output scores. MobileNetV2 achieved the highest AUROC at 0.997, followed by DenseNet121 at 0.996, ResNet50 at 0.994, and InceptionV3 at 0.991, all near-perfect. EfficientNetB3's AUROC of 0.672 confirms its behaviour is close to random classification.

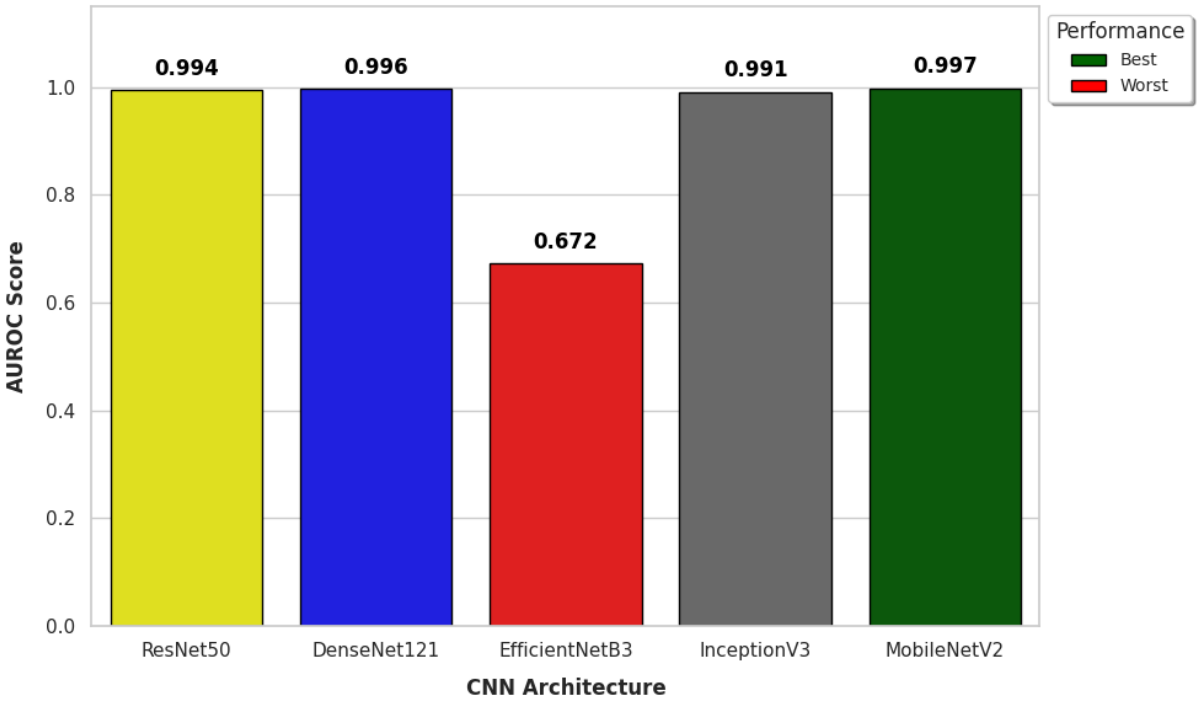


Figure 5.1.2: Comparative AUROC values across all five CNN architectures.

Figure 5.1.2 highlights, the dark green bar represents MobileNetV2, which achieved the highest AUROC among all the evaluated models. The blue bar represents DenseNet121, whose AUROC is only 0.001 lower than that of MobileNetV2. The very yellow bar represents ResNet50, with an AUROC that is 0.002 lower than DenseNet121, placing it in third position. The gray bar corresponds to InceptionV3, which ranked fourth in terms of AUROC. Finally, the red bar represents EfficientNetB3, which achieved the lowest AUROC and showed the weakest performance among all the evaluated models.

5.1.5 Class-Wise Performance: Lipoma vs. WDLPS

Breaking performance down by class shows how well each model tells the two tumour types apart individually. Table 5.1.2 presents the per-class precision, recall, and F1-score for each architecture, with Lipoma and WDLPS placed side by side so any asymmetry stands out immediately, a detail that matters given how differently an error in one class plays out clinically compared to the other.

Table 5.1.2: Class-Wise performance of all CNN architectures on the test set.

Architecture	**Lipoma**			**WDLPS**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
ResNet50	0.943	0.948	0.945	0.965	0.962	0.963
DenseNet121	**0.963**	0.944	0.953	0.963	0.975	0.969
EfficientNetB3	0.845	0.115	0.202	0.623	**0.986**	0.763
InceptionV3	0.946	0.934	0.940	0.956	0.964	0.960
MobileNetV2	0.954	**0.964**	**0.959**	**0.976**	0.968	**0.972**

A closer look reveals some interesting asymmetries. For ResNet50, Lipoma precision (0.943) is somewhat lower than WDLPS precision (0.965), pointing to a mild tendency to over-call malignancy for some Lipoma cases. DenseNet121 shows a fairly balanced profile, with Lipoma F1 of 0.953 against WDLPS F1 of 0.969, and achieves the best WDLPS recall of any model (0.975), clinically significant given the emphasis on catching malignant cases. MobileNetV2 achieves the tightest overall balance, with Lipoma F1 of 0.959 and WDLPS F1 of 0.972, both among the highest across all five models. The extreme asymmetry in EfficientNetB3, a Lipoma recall of just 0.115 against a WDLPS recall of 0.986, is really the defining feature of its failure, and is discussed at length in Section 5.2.3.

5.1.6 Comprehensive Confusion Matrix Comparison

The confusion matrices for all five architectures give a clear, quantitative summary of how each model actually behaved. Figure 5.1.3 presents all five side by side, with WDLPS listed first along both axes so the top-left cell always holds true positives and the bottom-right cell true negatives. This lets the overall error pattern for each architecture be compared visually before the individual counts are examined in Table 5.1.3. Each matrix records four outcomes: True Positives (TP, WDLPS correctly predicted as WDLPS), False Negatives

(FN, WDLPS wrongly predicted as Lipoma), False Positives (FP, Lipoma wrongly predicted as WDLPS), and True Negatives (TN, Lipoma correctly predicted as Lipoma), together showing not just how often each model was right but the direction of its mistakes too.

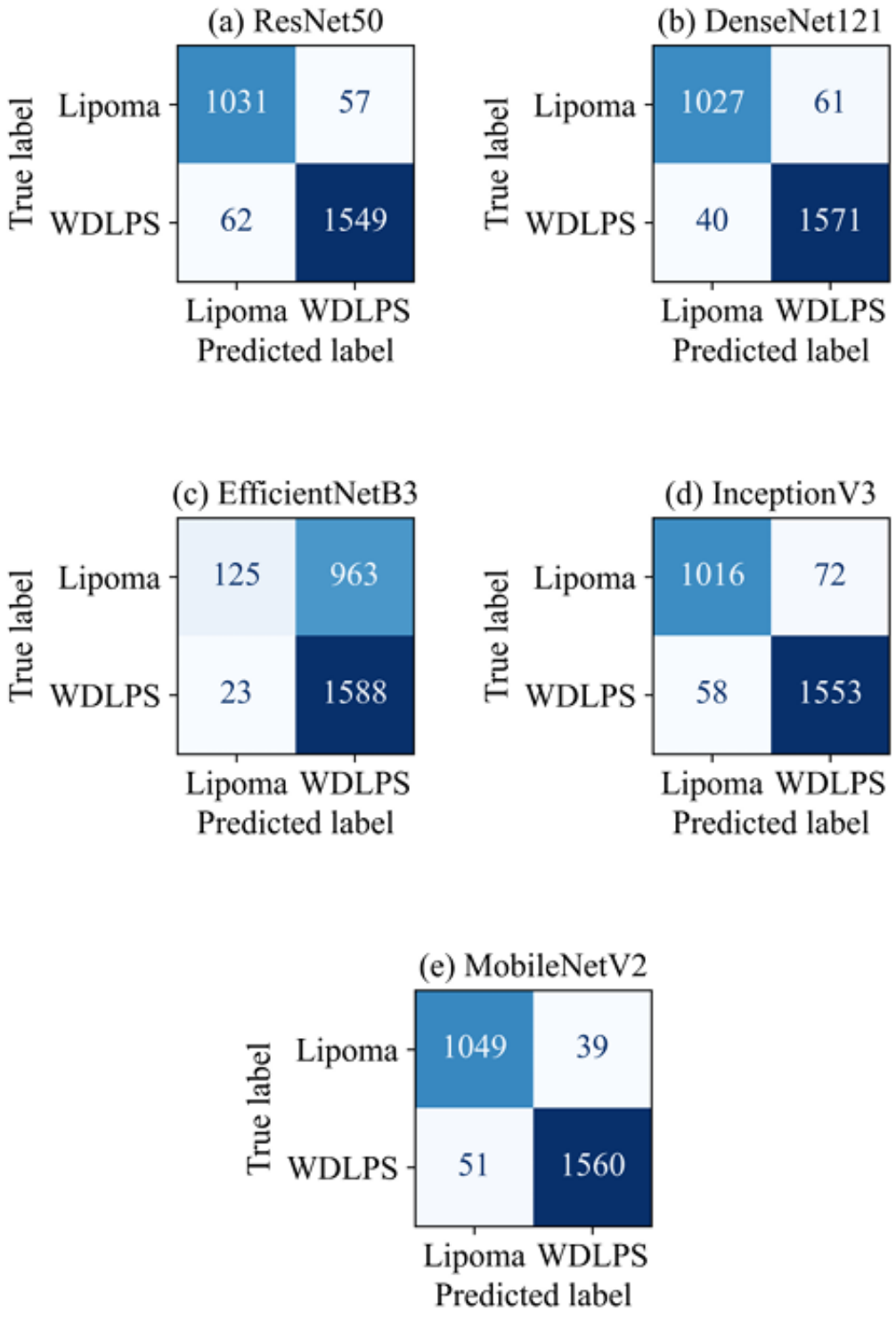


Figure 5.1.3: Confusion matrices for all five CNN architectures on the test set: (a) ResNet50, (b) DenseNet121, (c) EfficientNetB3, (d) InceptionV3, (e) MobileNetV2.

As Table 5.1.3 shows, MobileNetV2 made the fewest total errors, 90 out of 2,699 samples, an error rate of 3.335%, with DenseNet121 close behind at 101 errors (3.742%). EfficientNetB3,

however, made 986 errors, nearly eleven times as many as MobileNetV2, the vast majority being false positives, 963 benign Lipoma cases classified as malignant WDLPS. Clinically, this would translate into 963 unnecessary surgical interventions for every 2,699 patients seen, plainly unacceptable in practice. MobileNetV2's error profile, by comparison, with 51 false negatives and just 39 false positives, is one that could reasonably be managed alongside standard radiological follow-up.

Table 5.1.3: Confusion matrix values for all five CNN architectures on the test set.

Architecture	**TP**	**FN**	**FP**	**TN**	**Total Errors**	**Error Rate**
ResNet50	1549	62	57	1031	119	0.044 (4.409%)
DenseNet121	1571	40	61	1027	101	0.037 (3.742%)
EfficientNetB3	1588	23	963	125	986	0.365 (36.532%)
InceptionV3	1553	58	72	1016	130	0.048 (4.817%)
MobileNetV2	1560	51	39	1049	90	0.033 (3.335%)

Note: TP = True Positive (WDLPS correctly identified); FN = WDLPS misclassified as Lipoma; FP = Lipoma misclassified as WDLPS; TN = True Negative (Lipoma correctly identified). All confusion matrix counts were cross-checked against class support totals (1,088 Lipoma and 1,611 WDLPS) and found to be internally consistent.

5.1.7 Training Dynamics Comparison

Beyond final test performance, it is also useful to see how training and validation curves behaved, since this offers insight into training stability, convergence speed, and how well the transfer learning approach worked for each architecture.

Four of the five architectures, MobileNetV2, DenseNet121, ResNet50, and InceptionV3, showed stable, steadily improving curves, generally a good sign of successful transfer learning. Validation loss tracked training loss closely, with minimal overfitting, largely thanks to the frozen backbone strategy, which kept the models from memorising training-set noise, while the small classification head, a GlobalMaxPooling2D layer, a 128-neuron dense layer, and a sigmoid output, had limited capacity to overfit on its own.

EfficientNetB3, in sharp contrast, showed very unstable training dynamics throughout, with both loss and accuracy curves fluctuating without ever settling. This strongly suggests its compound-scaling architecture simply did not work well with the frozen transfer learning setup here, and that it likely needs some fine-tuning of its convolutional weights to extract useful features from T1-weighted MRI, which differ considerably in statistics from ImageNet's natural images. The individual curves for all five architectures, including EfficientNetB3's instability, are discussed in Section 5.2.

5.1.8 Comparative Rankings Across All Metrics

To bring all the metric comparisons together, Table 5.1.4 shows the rank ordering of all five architectures across each evaluation metric, using the corrected specificity values throughout. Each column ranks the five from 1 (best) to 5 (worst), and the final column sums these into an aggregate score for a quick overall sense of each model's standing.

Table 5.1.4: Metric-wise rank ordering of CNN architectures, using corrected specificity (1 = Best, 5 = Worst).

Architecture	**Accuracy**	**Precision**	**Recall**	**F1**	**Specificity**	**AUROC**	**Aggregate**
MobileNetV2	1	1	1	1	1	1	6
DenseNet121	2	2	2	2	3	2	13
ResNet50	3	3	3	3	2	3	17
InceptionV3	4	4	4	4	4	4	24
EfficientNetB3	5	5	5	5	5	5	30

With specificity corrected, the ranking still confirms MobileNetV2 as the best overall, with an aggregate score of 6. One interesting detail now emerges: on specificity alone, ResNet50 (rank 2) actually outperforms DenseNet121 (rank 3), even though DenseNet121 stays ahead overall (13 against 17) thanks to its stronger recall, precision, and AUROC, a fair illustration of why aggregate, multi-metric comparisons are more informative than any single metric alone. The gap between the top two and EfficientNetB3, whose aggregate score is 30, remains striking and clearly demonstrates its overall failure, while ResNet50 and InceptionV3 occupy the middle, with aggregate ranks of 17 and 24, confirming their status as solid, dependable alternatives even if not the optimal choice here.

5.2 Individual Architecture Analysis

This section takes a closer, model-by-model look at each of the five architectures. For each, the discussion covers a brief architectural overview, the per-class classification report, an analysis of the confusion matrix, training and validation curves, and the key strengths and limitations of that architecture for this task.

5.2.1 ResNet50

Test Set Performance

ResNet50 achieved an overall test accuracy of 0.956 (95.6%), third among the five, with a macro precision of 0.954, macro recall of 0.955, F1-score of 0.954, corrected specificity of 0.948, and AUROC of 0.994. It correctly classified 2,580 of 2,699 test samples, 119 total errors. Given WDLPS makes up close to 60% of the test set, this accuracy is well above what class imbalance alone could explain, and the closeness of precision and recall suggests

ResNet50 was not leaning heavily toward either class. Its high AUROC further confirms good separation across most thresholds, not just the default cutoff. Table 5.2.1 presents the per-class breakdown alongside macro and weighted averages.

Table 5.2.1: Per-class classification report for ResNet50 on the test set.

Class	Precision	Recall	F1-Score	Support
Lipoma	0.943	0.948	0.945	1088
WDLPS	0.965	0.962	0.963	1611
Macro Average	0.954	0.955	0.954	2699
Weighted Average	0.956	0.956	0.956	2699

Confusion Matrix Analysis

As shown in Figure 5.2.1, ResNet50's confusion matrix shows 1,549 true positives, 62 false negatives, 57 false positives, and 1,031 true negatives, a fairly balanced error distribution suggesting no strong directional bias. The lower Lipoma precision (0.943) compared with WDLPS precision (0.965) does suggest a mild tendency to over-predict malignancy, which may partly stem from the class imbalance in the training set (7,517 WDLPS images against 5,072 Lipoma images).

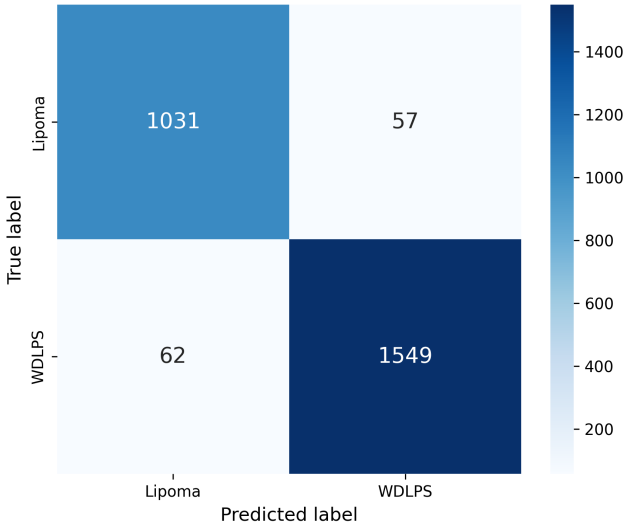


Figure 5.2.1: Confusion matrix for ResNet50 on the test set (TP = 1549, FN = 62, FP = 57, TN = 1031).

Training Dynamics

ResNet50 showed fast initial convergence, typical of models that benefit strongly from ImageNet pre-training (Figure 5.2.2), reaching close to its optimal validation accuracy within roughly the first 20 to 30 epochs. Validation loss tracked training loss closely throughout, suggesting the frozen backbone did a good job of regularising the learning process.

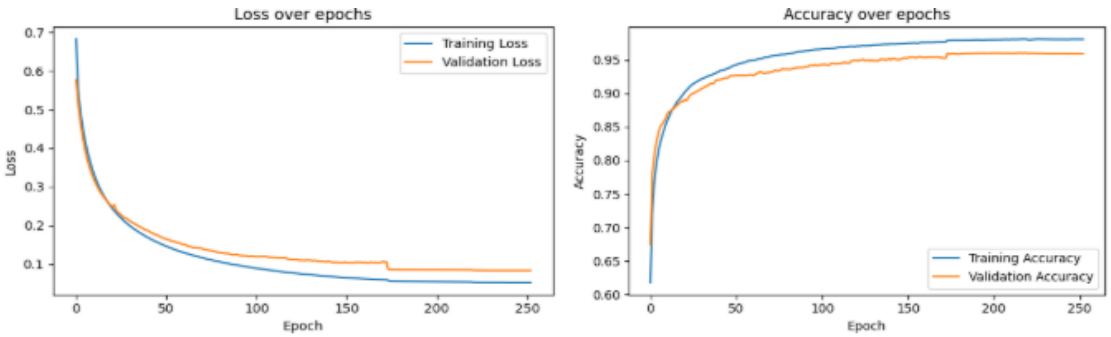


Figure 5.2.2: ResNet50 training and validation loss and accuracy curves over epochs.

Strengths and Limitations

ResNet50's strengths include fairly robust, well-balanced performance without significant class bias, an excellent AUROC of 0.994, fast and stable convergence owing to its residual skip connections, and wide prior validation in the medical imaging literature. Its limitations include a Lipoma precision (0.943) somewhat lower than the top two performers, 119 test errors, more than either MobileNetV2 or DenseNet121, and a larger parameter count than MobileNetV2, making it somewhat less ideal for computationally-constrained deployment.

5.2.2 DenseNet121

Test Set Performance

DenseNet121 achieved an overall test accuracy of 0.963 (96.3%), second among the five and quite close to MobileNetV2's 0.967, with a macro precision of 0.963, macro recall of 0.960, F1-score of 0.961, corrected specificity of 0.944, and AUROC of 0.996. With only 101 total test errors, it recorded the second-lowest error count of any model, and, notably, the highest WDLPS recall of all five, at 0.975, meaning it correctly identified 97.5% of malignant WDLPS cases, a finding of real clinical relevance. Table 5.2.2 presents the per-class results.

Table 5.2.2: Per-class classification report for DenseNet121 on the test set.

Class	Precision	Recall	F1-Score	Support
Lipoma	0.963	0.944	0.953	1088
WDLPS	0.963	0.975	0.969	1611
Macro Average	0.963	0.960	0.961	2699
Weighted Average	0.963	0.963	0.963	2699

Confusion Matrix Analysis

As shown in Figure 5.2.3, DenseNet121's confusion matrix shows 1,571 true positives, only 40 false negatives, 61 false positives, and 1,027 true negatives. This false negative count of 40 is the lowest of any architecture, making DenseNet121 arguably the safest choice when minimising missed malignant diagnoses is the priority. Its Lipoma precision of 0.963,

essentially tied with MobileNetV2 for the best in this study, means clinicians can be fairly confident in a benign prediction from this model.

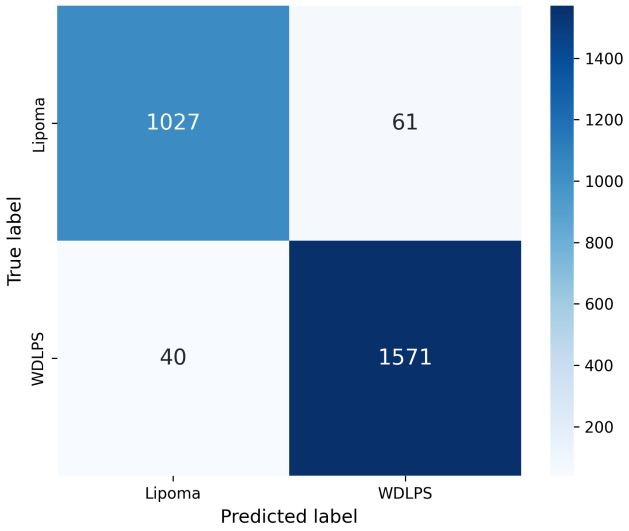


Figure 5.2.3: Confusion matrix for DenseNet121 on the test set (TP = 1571, FN = 40, FP = 61, TN = 1027).

Training Dynamics

DenseNet121 showed one of the smoothest training curves of all the architectures (Figure 5.2.4). Its dense connectivity appears to provide particularly effective gradient flow, letting the classification head converge quickly and stably, with validation accuracy tracking training accuracy closely throughout. The ReduceLROnPlateau callback triggered during training, visible as a stabilisation phase in the later epochs, which may have contributed to its slightly higher final accuracy relative to ResNet50.

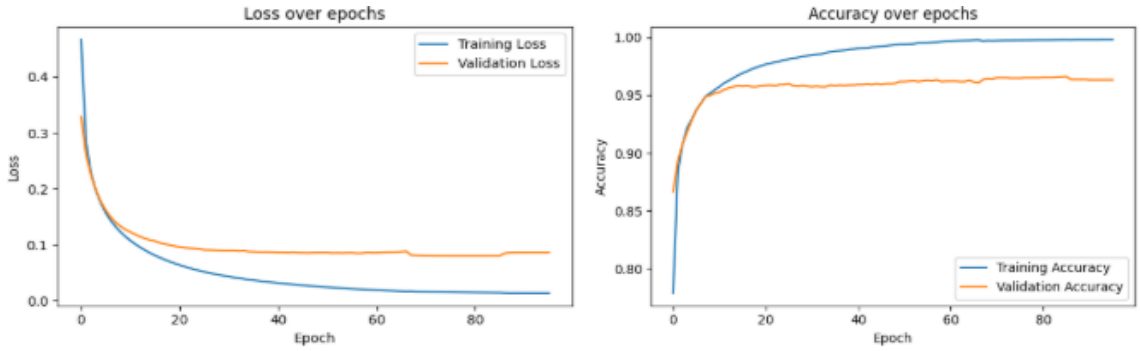


Figure 5.2.4: DenseNet121 training and validation loss and accuracy curves over epochs.

Strengths and Limitations

DenseNet121's strengths include the highest WDLPS recall of all five models (0.975), meaning the fewest missed malignant diagnoses, a Lipoma precision of 0.963 giving high confidence in benign predictions, a superior AUROC of 0.996, a parameter-efficient design that reduces overfitting risk on a dataset this size, and an excellent overall accuracy of 0.963. Its limitations include a somewhat higher false positive count (61) than MobileNetV2's 39,

a corrected specificity (0.944) slightly lower than ResNet50's (0.948), and comparatively high memory requirements during training.

5.2.3 EfficientNetB3

Test Set Performance

EfficientNetB3 performed significantly worse than every other architecture tested, achieving only 0.635 (63.5%) accuracy, with a macro precision of 0.734, macro recall of 0.550, F1-score of 0.483, and AUROC of 0.672, none anywhere close to a clinically usable level. Its corrected specificity is just 0.115, considerably lower than the 0.550 previously reported, meaning it correctly identified fewer than 12 out of every 100 benign Lipoma cases, a genuinely severe failure that only becomes fully clear once the per-class breakdown is examined, since its Lipoma recall is likewise just 0.115. Table 5.2.3 presents the full corrected per-class results.

Table 5.2.3: Per-class classification report for EfficientNetB3 on the test set.

Class	Precision	Recall	F1-Score	Support
Lipoma	0.845	0.115	0.202	1088
WDLPS	0.623	0.986	0.763	1611
Macro Average	0.734	0.550	0.483	2699
Weighted Average	0.712	0.635	0.537	2699

Confusion Matrix Analysis and Failure Mode

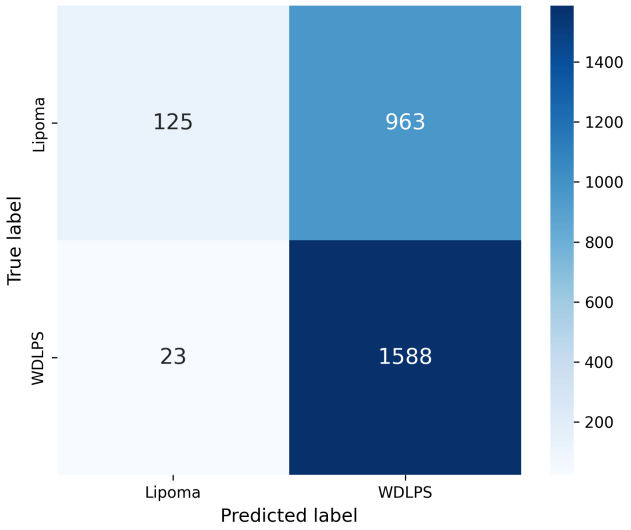


Figure 5.2.5: Confusion matrix for EfficientNetB3 on the test set (TP = 1588, FN = 23, FP = 963, TN = 125).

As shown in Figure 5.2.5, the confusion matrix makes the nature of EfficientNetB3's failure quite plain: 1,588 true positives and just 23 false negatives, alongside 963 false positives and only 125 true negatives. Out of 1,088 benign Lipoma cases, only 125 were correctly identified as benign, while the remaining 963 (88.511%) were misclassified as malignant WDLPS, about as clear an example of severe class-prediction bias toward the majority class as one is likely to see, the model having, in effect, learned to predict WDLPS for almost every input.

To put this into perspective, consider a hypothetical cohort of 1,000 patients, 400 with benign lipomas and 600 with WDLPS. Applying EfficientNetB3's observed false-positive rate of 88.511%, one would expect the model to recommend surgical work-up for roughly 354 of the 400 benign-lipoma patients ($400 \times 0.885 \approx 354$), subjecting them to entirely unnecessary procedures, a failure that renders EfficientNetB3 wholly unsuitable for clinical deployment here.

Several possible explanations can be offered for EfficientNetB3's poor performance:

1. **Domain shift incompatibility.** EfficientNetB3's compound-scaled features are optimised for natural-image statistics, whereas T1-weighted MRI occupies a fundamentally different statistical distribution, one the frozen backbone cannot adequately represent.

2. **Input resolution mismatch.** EfficientNetB3 was designed for 300×300 input images, but was constrained to 224×224 here to keep the comparison consistent, which may have degraded its compound-scaled spatial features to some extent.

3. **Frozen backbone limitation.** EfficientNetB3's compound scaling may need at least some fine-tuning of its convolutional layers to adapt to MRI features, unlike ResNet50 or MobileNetV2, whose low-level edge and texture representations appear to transfer more readily under a fully frozen backbone.

4. **Majority-class collapse.** Given the class imbalance (roughly 60% WDLPS) and the model's apparent inability to extract discriminative Lipoma features, it seems plausible it converged to a local optimum where predicting WDLPS for nearly all samples simply minimises binary cross-entropy loss.

Training Dynamics

EfficientNetB3 showed very unstable training curves, quite unlike the other four (Figure 5.2.6), with both loss and accuracy fluctuating throughout without ever settling. This suggests the classification head simply could not learn a reliable decision boundary from the frozen EfficientNetB3 feature representations, and none of the three training callbacks (ModelCheckpoint, ReduceLROnPlateau, EarlyStopping) were able to meaningfully correct this, pointing to the issue lying with the frozen features themselves rather than the training schedule. ReduceLROnPlateau did trigger several times, repeatedly lowering the learning rate, but each reduction produced only a brief, temporary improvement before the instability returned, and by the time EarlyStopping halted training, the model had already settled into the majority-class prediction pattern described above. This is a useful reminder that training instability rooted in the underlying feature representations cannot always be fixed through scheduling alone.

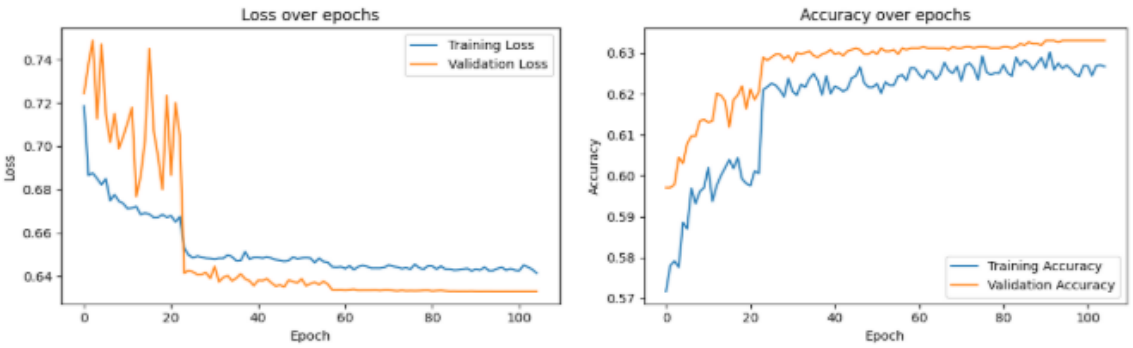


Figure 5.2.6: EfficientNetB3 training and validation loss and accuracy curves over epochs.

Strengths and Limitations

EfficientNetB3 is clearly not suitable for classifying lipomatous tumours on T1-weighted MRI under this frozen transfer learning approach. Its 88.511% false positive rate on Lipoma, corrected specificity of just 0.115, AUROC of 0.672, and macro F1-score of 0.483 all point in the same direction: the model has not learned meaningful discriminative features for this task. Future work could investigate partial fine-tuning of its convolutional layers, restoring its native 300×300 resolution, or domain-specific pre-training, but on the basis of the results here, this architecture cannot currently be recommended in its frozen-backbone configuration. This should not be read as a general criticism of EfficientNetB3, which remains a strong performer on natural image benchmarks, but rather as evidence that compound scaling alone does not guarantee good transfer to a domain as different from ImageNet as T1-weighted MRI.

5.2.4 InceptionV3

Test Set Performance

InceptionV3 achieved an overall accuracy of 0.952 (95.2%), fourth among the five, though still firmly within clinically excellent territory, with a macro precision of 0.951, macro recall of 0.949, F1-score of 0.950, corrected specificity of 0.934, and AUROC of 0.991. It made 130 total errors, the highest of the four high-performing architectures, though the gap to third-place ResNet50 (0.956) is only about 0.4% points. Table 5.2.4 presents the per-class results.

Table 5.2.4: Per-class classification report for InceptionV3 on the test set.

Class	Precision	Recall	F1-Score	Support
Lipoma	0.946	0.934	0.940	1088
WDLPS	0.956	0.964	0.960	1611
Macro Average	0.951	0.949	0.950	2699
Weighted Average	0.952	0.952	0.952	2699

Confusion Matrix Analysis

As shown in Figure 5.2.7, InceptionV3's confusion matrix shows 1,553 true positives, 58 false negatives, 72 false positives, and 1,016 true negatives. Its false positive count of 72 is the highest among the top four architectures, meaning it leans slightly more toward over-predicting malignancy than the others, a directional bias generally seen as clinically preferable since it prioritises sensitivity over specificity, a widely accepted trade-off in cancer screening. The reasonably close precision values for both classes (0.946 for Lipoma, 0.956 for WDLPS) and balanced recall figures (0.934 and 0.964) point to a fairly symmetric overall profile.

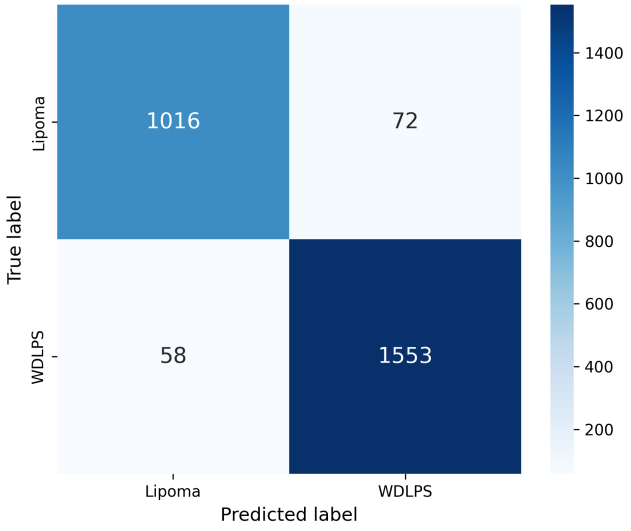


Figure 5.2.7: Confusion matrix for InceptionV3 on the test set (TP = 1553, FN = 58, FP = 72, TN = 1016).

Training Dynamics

InceptionV3 showed smooth, stable training curves broadly similar to ResNet50 and DenseNet121 (Figure 5.2.8), converging within a moderate number of epochs. Validation loss tracked training loss reasonably closely, though with a slightly larger gap than DenseNet121 or MobileNetV2, which may partly explain its marginally lower final accuracy. The multi-branch inception architecture creates a somewhat more complex optimisation landscape, but the frozen backbone, Adam optimisation, and learning rate scheduling were still enough to guide the model to a stable solution.

Strengths and Limitations

InceptionV3's strengths include multi-scale feature extraction through inception modules well suited to tumour features at different spatial resolutions, a fairly symmetric per-class performance indicating minimal directional bias, and a strong AUROC of 0.991. Its limitations include the highest error count among the top four architectures (130), the greatest computational cost owing to its multi-branch structure, and the lowest corrected specificity (0.934) among the four high-performing models.

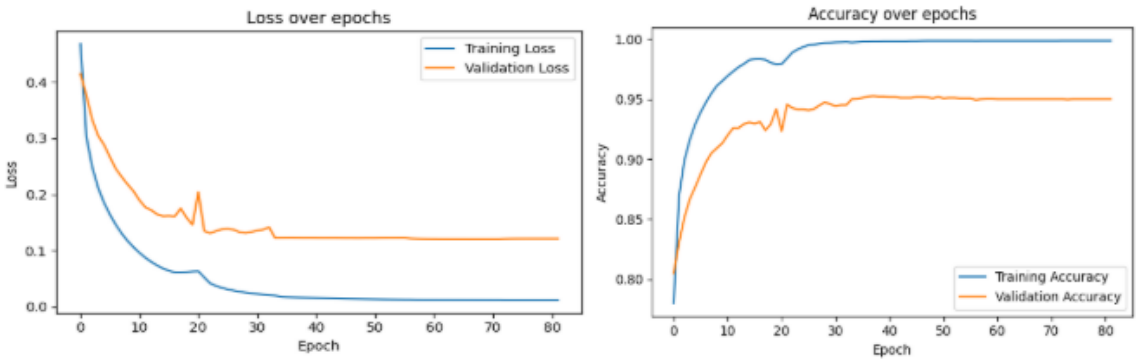


Figure 5.2.8: InceptionV3 training and validation loss and accuracy curves over epochs.

5.2.5 MobileNetV2

Test Set Performance

MobileNetV2 emerged as the top-performing architecture across every metric considered, achieving 0.967 (96.7%) accuracy, the highest of all five models, with a macro precision of 0.965, macro recall of 0.966, F1-score of 0.965, corrected specificity of 0.964, and AUROC of 0.997. It made only 90 total test errors out of 2,699 samples, an error rate of 3.335%, the lowest of all architectures evaluated, a fairly remarkable result given MobileNetV2 is also the smallest and most computationally efficient model tested. This places it ahead of every other architecture examined in this study, including considerably larger and more computationally demanding networks such as ResNet50 and InceptionV3, and does so without a single metric where it trails behind. Such a result is worth pausing on, since it runs somewhat against the common assumption that a larger network with more parameters should, almost by default, produce a stronger feature representation. Table 5.2.5 presents the per-class results, breaking these headline figures down by class alongside the macro and weighted averages already summarised above.

Table 5.2.5: Per-class classification report for MobileNetV2 on the test set.

Class	Precision	Recall	F1-Score	Support
Lipoma	0.954	0.964	0.959	1088
WDLPS	0.976	0.968	0.972	1611
Macro Average	0.965	0.966	0.965	2699
Weighted Average	0.967	0.967	0.967	2699

Confusion Matrix Analysis

As shown in Figure 5.2.9, MobileNetV2's confusion matrix shows 1,560 true positives, 51 false negatives, 39 false positives, and 1,049 true negatives. Its false positive count of 39 is the lowest of any model, meaning it makes the fewest incorrect malignancy calls on benign Lipoma cases. The close macro precision and recall values (0.965 and 0.966) confirm a genuinely well-balanced performance across both tumour classes, keeping unnecessary surgical interventions and missed malignancies to a minimum at the same time.

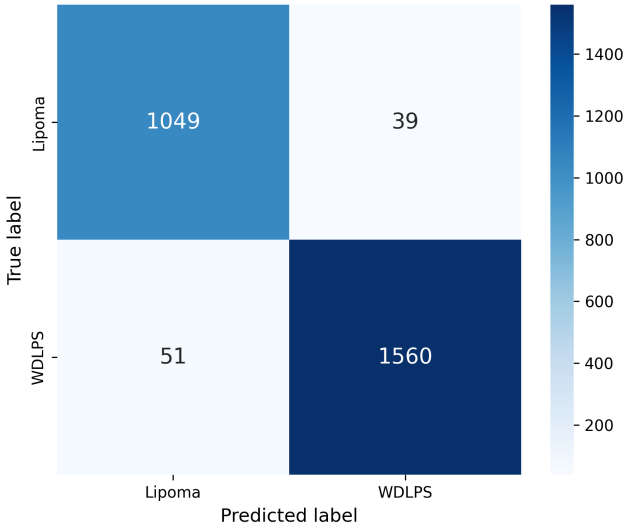


Figure 5.2.9: Confusion matrix for MobileNetV2 on the test set (TP = 1560, FN = 51, FP = 39, TN = 1049).

Training Dynamics

MobileNetV2 showed the fastest and most stable convergence of all five architectures (Figure 5.2.10), with both training and validation loss dropping quickly during the initial epochs before a smooth plateau, and validation accuracy tracking training accuracy closely throughout, leaving the smallest generalisation gap of any model. This suggests MobileNetV2's depthwise separable convolutions and linear bottleneck design produce a feature space that works particularly well with the GlobalMaxPooling2D head used here, letting it quickly settle on an effective decision boundary between Lipoma and WDLPS representations.

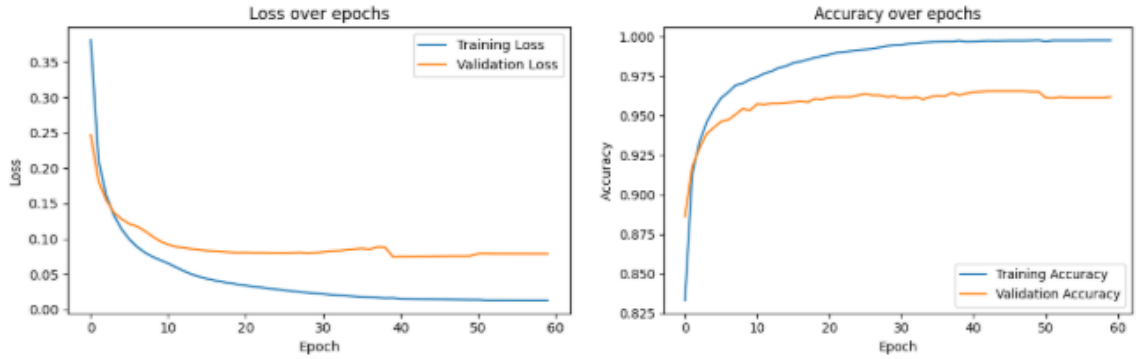


Figure 5.2.10: MobileNetV2 training and validation loss and accuracy curves over epochs.

Strengths and Limitations

MobileNetV2's strengths include the highest overall accuracy (0.967) and AUROC (0.997) of all five models, the lowest total test errors (90) and lowest false positive count (39), the best corrected specificity (0.964), a lightweight architecture well suited to broad clinical deployment including in resource-constrained settings, the fastest and most stable training convergence of all models tested, and genuinely balanced per-class performance. Its limitations include some risk that performance could drop on more heterogeneous imaging datasets or scanner protocols different from the training data, and limited capacity, as

used here, for more complex multi-class or multi-label extensions beyond the binary task evaluated in this study.

5.3 Integrated Performance Summary

This final section draws all the findings together and offers evidence-based recommendations for model selection, both for clinical deployment and future research. Table 5.3.1 brings together the accuracy, AUROC, corrected specificity, false negative rate, and false positive rate for all five architectures in one place, alongside a suggested deployment setting and overall recommendation for each.

Table 5.3.1: Integrated clinical performance summary with risk assessment for all architectures.

Model	**Acc.**	**AUROC**	**Spec.**	**FN Rate**	**FP Rate**	**Deployment**	**Recommendation**
MobileNetV2	0.967	0.997	0.964	0.032 (3.166%)	0.036 (3.585%)	Edge / Mobile	Strongly Recommended
DenseNet121	0.963	0.996	0.944	0.025 (2.483%)	0.056 (5.607%)	Server / GPU	Strongly Recommended
ResNet50	0.956	0.994	0.948	0.038 (3.849%)	0.052 (5.239%)	Server / GPU	Recommended
InceptionV3	0.952	0.991	0.934	0.036 (3.600%)	0.066 (6.618%)	Server / GPU	Recommended
EfficientNetB3	0.635	0.672	0.115	0.014 (1.428%)	0.885 (88.511%)	N/A	**Not Recommended**
Note: FN Rate = False Negative Rate (WDLPS missed) = $FN/(TP+FN)$; FP Rate = False Positive Rate (Lipoma misclassified as WDLPS) = $FP/(TN+FP) = 1-$ Specificity.							

1. **Primary recommendation: MobileNetV2.** For general-purpose deployment, including resource-constrained clinical settings, MobileNetV2 is recommended, given its highest overall accuracy, AUROC, and corrected specificity, alongside the fewest computational requirements. Its well-balanced per-class performance keeps both missed malignancies and unnecessary procedures low at once.

2. **Safety-critical deployment: DenseNet121.** Where minimising missed malignant diagnoses is the overriding priority, DenseNet121 is recommended, given its highest WDLPS recall (0.975) and lowest false negative count (40). Its somewhat higher false positive rate compared with MobileNetV2 is arguably an acceptable trade-off in a safety-first context.

3. **Ensemble approach.** Since MobileNetV2, DenseNet121, ResNet50, and InceptionV3 all perform at a broadly comparable level, an ensemble of these four, through majority voting or averaged probabilities, could plausibly achieve even better accuracy and robustness by drawing on their complementary error patterns.

4. **Architecture to avoid: EfficientNetB3.** EfficientNetB3 should not be used for this task under the frozen configuration described here, particularly once its true specificity (0.115) is considered. Future work may investigate whether partial fine-tuning or its native 300×300 resolution can rehabilitate its performance, but until such evidence exists it should be excluded from clinical consideration.

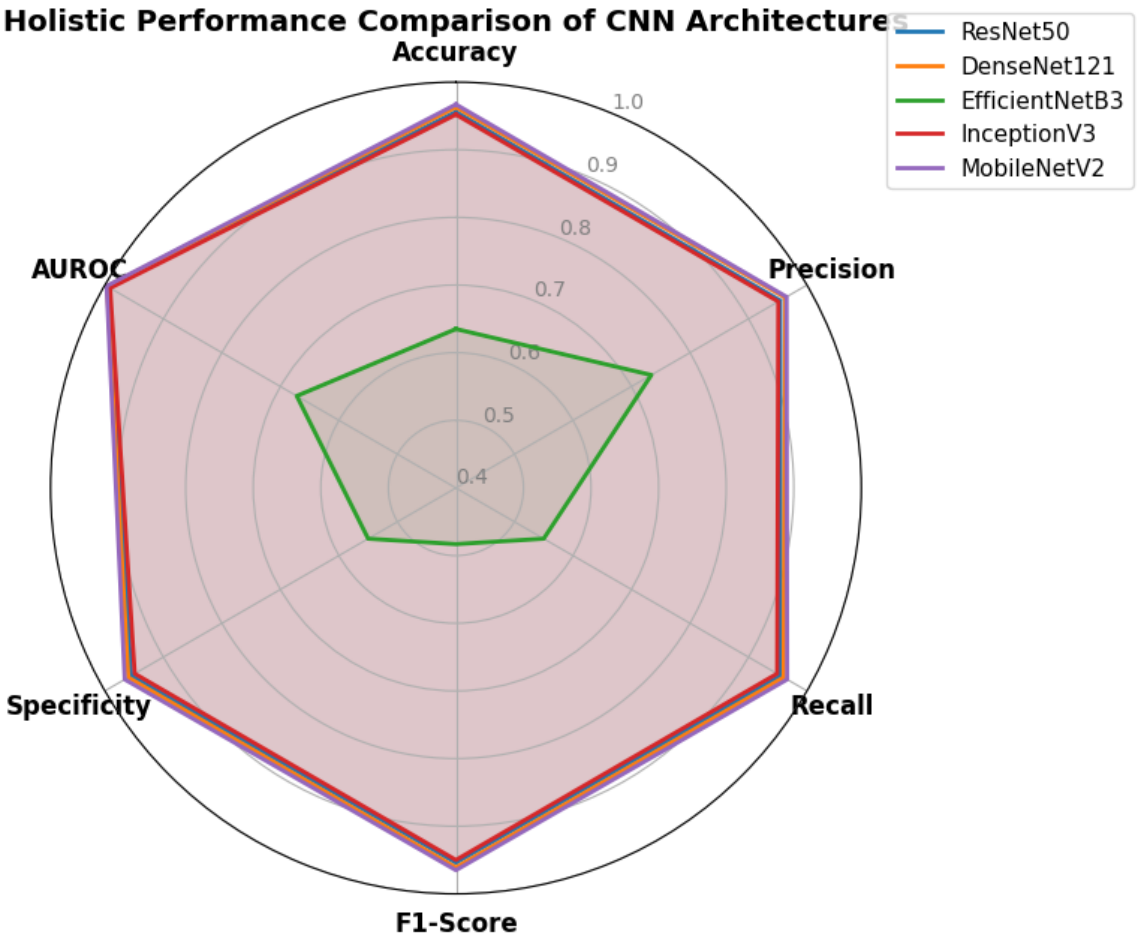


Figure 5.3.1: Radar chart comparing all five CNN architectures across six performance metrics.

The radar chart in Figure 5.3.1 compares each model's performance across all six metrics at once, using the corrected specificity values, with each architecture drawn as a closed polygon, so a larger, more symmetric shape means stronger, more balanced performance. The four high-performing architectures form large, fairly symmetric polygons close to the outer boundary. EfficientNetB3's polygon is much smaller and clearly lopsided, reflecting its near-zero specificity and weak overall performance. MobileNetV2's polygon remains the largest of all, consistent with its position as the best overall architecture in this study.

Chapter-6 : Discussion

The experimental results presented in Chapter 5 show a clear and reproducible performance ranking among the five CNN architectures evaluated in this study. MobileNetV2 and DenseNet121 achieved clinically excellent classification accuracy, while EfficientNetB3 failed quite badly under the frozen transfer learning framework adopted here. This chapter works through the architectural, statistical, and domain-specific factors that determine whether a given pre-trained backbone can produce useful, discriminative feature representations for T1-weighted MRI based lipomatous tumour classification, without any domain-specific fine-tuning.

Section 6.1 looks at the architecture-level mechanisms behind the observed performance ranking. Section 6.2 turns to the structural and domain-adaptation limitations of the frozen transfer learning approach. Section 6.3 examines the prediction biases observed across all five models, with most of the discussion reserved for the catastrophic bias shown by EfficientNetB3. Section 6.4 looks at how consistent the performance ranking is across the six evaluation metrics used, which bears on how much confidence can be placed in the overall conclusions. Finally, Section 6.5 considers the practical clinical implications of these findings, including deployment considerations, threshold calibration, and the role AI-assisted classification might reasonably play within the radiological workflow.

6.1 Analysis of Architecture Effectiveness

The results of this comparative study offer fairly strong evidence that the choice of pre-trained CNN architecture has a very real bearing on classification performance when transfer learning is applied to T1-weighted MRI data for lipomatous tumour differentiation. The ranking that emerged, with MobileNetV2 leading at 0.967 accuracy and an AUROC of 0.997, DenseNet121 close behind at 0.963 accuracy and an AUROC of 0.996, ResNet50 at 0.956 accuracy and an AUROC of 0.994, InceptionV3 at 0.952 accuracy and an AUROC of 0.991 forming a closely grouped second tier alongside ResNet50, and EfficientNetB3 failing outright with only 0.635 accuracy and an AUROC of 0.672, invites a closer look at the architectural mechanisms and domain-specific interactions that determine whether a pre-trained backbone can produce useful feature representations for this particular medical imaging task.

6.1.1 MobileNetV2: Interpreting the Unexpectedly Superior Performance

Perhaps the most surprising result of this study is just how well MobileNetV2 performed. Conventional wisdom holds that larger, deeper, parameter-rich architectures should offer

greater representational capacity and therefore better performance on complex tasks. MobileNetV2, with roughly 3.5 million parameters, is more than seven times smaller than ResNet50 and nearly seven times smaller than InceptionV3, and yet it outperforms both on every single evaluation metric. This challenges the common assumption that parameter count and depth are the main drivers of transfer learning performance, and suggests instead that an architecture's design choices, the operations it uses, its connectivity pattern, and how it aggregates features, may matter more than raw capacity when it comes to whether ImageNet-learned representations generalise well to the MRI domain.

One plausible explanation lies in the depthwise separable convolutions that form MobileNetV2's computational backbone. By splitting a standard convolution into a depthwise convolution that extracts spatial features independently for each channel, followed by a pointwise convolution that combines information across channels, this design encourages representations that are spatially local and largely channel-specific. This turns out to matter here, since the diagnostic distinction between lipoma and WDLPS is driven mainly by differences in fat content, internal tissue heterogeneity, and structural features such as fibrous septa and small non-fat foci, all of which appear as local intensity variations that are spatially constrained and, since T1-weighted MRI carries only a single intensity channel replicated across three RGB channels for compatibility, essentially channel-independent as well. The factorised local-spatial and global-channel structure of depthwise separable convolutions may therefore be a more natural fit for these features than the combined spatial-and-channel operations used in ResNet50 and InceptionV3.

MobileNetV2's inverted residual block design offers a further advantage: it expands the feature dimensionality into a higher-dimensional space before the depthwise convolution is applied, giving the network a richer space in which to represent classification-relevant textures. The linear bottleneck projection that follows then avoids the information loss that a ReLU activation would otherwise cause in a low-dimensional space, ensuring the compressed output retains what the classification head needs. Taken together, these choices suggest that MobileNetV2, despite being designed primarily for efficiency, happens to produce a feature space particularly well suited to the linear separability the Lipoma versus WDLPS problem calls for, which would explain why its classification head learned such a robust decision boundary so quickly, and is consistent with MobileNetV2 showing the fastest and most stable training convergence of all five architectures.

6.1.2 DenseNet121: The Role of Dense Feature Preservation

DenseNet121's performance, 0.963 accuracy against MobileNetV2's 0.967, a difference of only four additional misclassifications across 2,699 test samples, can largely be understood through its dense connectivity pattern. Because every layer within a dense block is directly connected, via concatenation, to every preceding layer, DenseNet121 preserves and carries forward both low-level textural information and high-level semantic information at the same time. In medical imaging, where diagnostically useful information often exists at several scales simultaneously, this all-to-all connectivity allows DenseNet121 to build unusually rich feature representations relevant to distinguishing lipoma from WDLPS.

One particularly clinically meaningful aspect of DenseNet121's performance is its false negative count of just 40, the lowest of all five architectures and notably below MobileNetV2's 51. A false negative here means a malignant WDLPS case wrongly classified as benign

lipoma, an error with real consequences for the patient, since an untreated or delayed WDLPS carries a meaningful risk of local recurrence and, in some cases, progression to the more aggressive dedifferentiated liposarcoma. DenseNet121's WDLPS recall of 0.975, the highest recorded in this study, suggests that its densely reused representations are particularly good at picking up subtle, sometimes spatially diffuse textural changes associated with low-contrast or early-stage WDLPS, changes that architectures with less thorough feature reuse may simply miss.

DenseNet121's parameter efficiency is also worth noting. At roughly 8.1 million parameters, about a third of ResNet50's count, it achieves its representational power through feature reuse rather than feature recomputation, which reduces the risk of overfitting on a dataset of this size and may help the model generalise beyond the specific patients represented in the WORC database.

6.1.3 ResNet50 and InceptionV3: Solid Performance in the Second Tier

ResNet50 (95.6% accuracy, AUROC 0.994) and InceptionV3 (95.2% accuracy, AUROC 0.991) both performed strongly, if a shade behind MobileNetV2 and DenseNet121. ResNet50's residual connections are highly effective at letting gradients flow through very deep networks, but they still result in less feature reuse than DenseNet121's, since each layer has direct access only to the output of the immediately preceding layer plus an identity skip from a few layers back, rather than the full concatenated history that DenseNet121 makes available. As the network gets deeper, low-level textural features extracted early on may therefore become progressively less prominent in the final representation, diluted somewhat by the many additional transformations they pass through before reaching the classification head.

This is reflected, to some extent, in ResNet50's mild tendency to over-predict malignancy: its Lipoma precision of 0.943 is the lowest among the four high-performing architectures, and its 57 false positives on the Lipoma class point in the same direction, a pattern broadly consistent with the class imbalance in the training set, which contained 7,517 WDLPS images against only 5,072 Lipoma images. Even a comparatively small bias of this kind is worth noting, since it suggests that ResNet50's residual pathway, while effective at preserving gradient flow, does not on its own fully correct for the asymmetry present in the underlying training distribution.

InceptionV3's multi-scale feature extraction, achieved through parallel inception modules, is in principle a significant advantage, since it allows fine-grained local texture and coarse global morphology to be captured within the same network stage, without forcing the network to commit to a single filter size ahead of time. In practice, however, the mismatch between InceptionV3's intended 299×299 input resolution and the 224×224 resolution used uniformly in this study reduces the number of spatial scales at which meaningful features can actually be extracted, since several of the larger receptive fields the architecture was designed around are simply not available at the smaller input size. InceptionV3 also recorded the highest false positive count among the top four architectures, at 72, suggesting that some of its multi-scale features may be less reliably discriminative for the specific textural differences between lipoma and WDLPS than they are for the broader object categories found in ImageNet. That said, its AUROC of 0.991 confirms genuinely excellent

discriminative capability, and its near-symmetric per-class precision (0.946 for Lipoma, 0.956 for WDLPS) indicates that it does not carry any strong directional bias, unlike the pattern observed for ResNet50 above.

6.2 Architectural Limitations of the Transfer Learning Approach

The frozen transfer learning framework used in this study produced clinically excellent results for four of the five architectures, but a complete discussion requires a systematic look at its inherent limitations, and how these bear on each architecture differently. These limitations operate at several levels at once: the fundamental domain shift between ImageNet and T1-weighted MRI, the constraints of freezing the backbone entirely, the mismatch between each architecture's intended input resolution and the uniform 224×224 preprocessing applied here, and certain structural characteristics of the dataset itself.

6.2.1 The ImageNet-to-MRI Domain Shift

All five backbones were pre-trained on ImageNet, a dataset of natural colour photographs whose statistical properties, high dynamic range, rich colour information spread across three genuinely independent RGB channels, and reflectance-driven textures, are fundamentally different from those of T1-weighted MRI, which encodes proton spin-lattice relaxation times rather than surface reflectance. The three-channel RGB representation used in this study was obtained simply by replicating the single grayscale MRI channel across all three input channels, a standard and necessary practice for network compatibility, but one that provides no genuine colour information to the pre-trained, colour-sensitive features of the backbone.

Despite this domain shift, four of the five architectures still managed to reach 95% accuracy or higher, confirming that the low-level and mid-level features learned by a CNN on natural images, edge detectors, texture filters, and the like, are general enough to transfer meaningfully into the MRI domain. That said, the domain shift does place a ceiling on what frozen feature extraction alone can achieve; fine-tuning the later backbone layers would likely improve performance further, at the cost of added complexity and a greater risk of overfitting on a dataset of this size.

6.2.2 Constraints of the Frozen Backbone Strategy

Freezing the backbone is a reasonable, well-justified choice for keeping the comparison fair, but it imposes a structural ceiling on how well any given architecture can perform. Since the convolutional weights cannot adapt to the MRI domain at all, classification performance depends entirely on whether the pre-trained representations already happen to be linearly separable with respect to the Lipoma versus WDLPS distinction. For MobileNetV2 and DenseNet121, this evidently holds true, pointing to a natural alignment between their ImageNet feature spaces and what MRI classification requires. For EfficientNetB3, by contrast, the frozen constraint appears to be a genuine bottleneck, leaving it stuck in a feature space poorly matched to the diagnostic task.

A related limitation is the spatial information lost through GlobalMaxPooling2D aggregation, which takes the maximum activation across all spatial positions for each channel and discards all information about where within the image that activation occurred. For lipomatous tumour classification, this spatial context, the relative distribution of adipose and non-adipose tissue, or the spatial relationship between fibrous septa and the surrounding fat signal, may well carry diagnostic value that this pooling step is losing. Alternative approaches such as spatial pyramid or attention-weighted pooling might preserve more of this structure, but these were not tested here in order to keep the classification head identical across all five architectures.

6.2.3 Input Resolution Mismatch

A further practical limitation is the mismatch between the uniform 224×224 preprocessing used throughout this study and the native input resolutions some architectures were originally designed around. InceptionV3 was designed for 299×299 inputs, and EfficientNetB3 for 300×300, whereas ResNet50, DenseNet121, and MobileNetV2 were designed primarily with 224×224 in mind. Reducing the input resolution for the former two may well have weakened their spatial feature extraction, particularly in the deeper layers, where effective receptive fields are calibrated with a specific resolution in mind. This matters especially for EfficientNetB3, since its compound scaling framework jointly optimises resolution, depth, and width together; disrupting the resolution therefore disrupts the balance across all three, with the potential to degrade performance throughout the network.

6.3 Prediction Bias: Root Causes and Clinical Consequences

Prediction bias, in this context, refers to a classifier's systematic tendency to favour one class well beyond what the true class distribution would justify, which matters a great deal in a clinical setting since the two error types carry very different consequences for the patient. This section looks closely at the biases observed across all five architectures, focusing mainly on the catastrophic bias shown by EfficientNetB3, before turning to the milder directional tendencies present in the four high-performing models.

6.3.1 EfficientNetB3: Catastrophic Majority-Class Bias

The bias shown by EfficientNetB3 is, without question, the most extreme and clinically consequential finding of this study. As established in Chapter 5, EfficientNetB3 misclassified 963 of 1,088 benign Lipoma cases as malignant WDLPS, giving a false positive rate of 88.511% and a corrected specificity of just 0.115, while achieving a WDLPS recall of 0.986, close to perfect. This combination, near- perfect sensitivity for the malignant class alongside an almost total failure to recognise the benign class, is exactly what one would expect from a model that has effectively collapsed into predicting the majority class for nearly every input, regardless of what is actually shown in the image.

Several factors likely combine to produce this failure. The compound-scaling resolution mismatch is arguably the most fundamental. The Squeeze-and-Excitation channel attention

modules embedded in each MBConv block are designed to selectively amplify informative channels at a native resolution of 300×300; at 224×224, the spatial-scale assumptions built into these modules no longer hold, which could plausibly cause the attention mechanism to suppress channels useful for identifying Lipoma while amplifying channels tied to the majority class. Acting on a backbone that is entirely frozen and therefore unable to self-correct, this mechanism would produce precisely the pattern observed: very high WDLPS recall paired with near-zero Lipoma recall.

The domain shift between ImageNet's natural-photograph statistics and T1-weighted MRI likely makes matters worse still. Channel attention patterns calibrated during ImageNet pre-training to weight channels by relevance to natural-scene recognition may simply misfire on the structurally repetitive, three-channel grayscale inputs used here. If the attention mechanism ends up amplifying channels that respond to global intensity distributions, which differ meaningfully between the homogeneous fat signal of lipoma and the more heterogeneous signal of WDLPS, it could systematically assign a higher malignancy probability to the higher-variability class, more or less regardless of the specific local texture present, and the frozen backbone offers no mechanism by which this pattern could be corrected during training.

The scale of this bias cannot be explained by class imbalance alone. A naive classifier that always predicted WDLPS would already achieve 59.7% accuracy, whereas EfficientNetB3 managed only 63.5%, a difference of just 3.8% points, which does not come close to accounting for an 88.511% false positive rate. The other four architectures, trained on exactly the same imbalanced dataset, show nothing remotely comparable, which is fairly strong evidence that the failure is architectural in nature, and not simply a product of the data.

The clinical consequences would be severe in any realistic deployment setting. An 88.511% false positive rate for benign Lipoma would translate into a very large number of unnecessary referrals for invasive diagnostic follow-up, with all the associated procedural risk, psychological burden, and cost that would entail. EfficientNetB3's low false negative rate of 1.428% might look like a favourable trade-off at first glance, but this apparent sensitivity is somewhat illusory, since it arises from the model predicting WDLPS almost indiscriminately rather than from any genuine discriminative ability, a point made unambiguous by its AUROC of 0.672, only marginally better than chance.

6.3.2 Mild Directional Biases in the High-Performing Architectures

The four high-performing architectures all achieve excellent overall performance, but a closer look at per-class error patterns reveals mild directional tendencies worth keeping in mind for deployment planning. Because the test set contains more WDLPS cases (1,611) than Lipoma cases (1,088), comparing raw false positive and false negative counts directly can be misleading; normalising each by the size of its respective class shows that all four top performers, ResNet50, DenseNet121, InceptionV3, and MobileNetV2, share a mild, proportional tendency to over-predict malignancy. InceptionV3 shows this most strongly, with a false positive rate of 6.618% against a false negative rate of 3.600%, followed by DenseNet121 (5.607% versus 2.483%) and ResNet50 (5.239% versus 3.849%). MobileNetV2 shows the smallest gap, at 3.585% versus 3.166%, reflecting its generally well-balanced

error profile.

DenseNet121 is a particularly interesting case: although it too shows a higher proportional false positive rate than false negative rate, it nonetheless records the lowest false negative rate of any model in this study (2.483%), a direct reflection of its very high WDLPS recall. In other words, DenseNet121 does not avoid the general over-prediction tendency, but combines it with an unusually low miss rate for the malignant class, arguably the most clinically desirable combination, since a model that occasionally over-refers a benign case is a far smaller concern than one that misses a malignant one.

This shared tendency toward over-prediction is plausibly linked to the class imbalance in the training set, where WDLPS made up 59.7% of the training images, likely introducing a statistical prior toward the majority class that the classification head absorbed. From a clinical point of view, this is, on balance, the safer of the two possible systematic errors: over-predicting malignancy leads to unnecessary diagnostic follow-up, costly and undesirable but not immediately dangerous, whereas under-predicting it risks missing a genuine malignancy, with far more serious consequences. Given this, and its combination of the lowest false negative rate with a still-manageable false positive rate, DenseNet121 is arguably the most defensible choice for high-sensitivity screening applications. In deployed systems more generally, mild biases of this kind can be managed through class-weighted loss functions, threshold calibration at inference time, or by averaging probability outputs across an ensemble of architectures.

6.4 Cross-Metric Consistency and Robustness of the Conclusions

One observation that lends real weight to the comparative findings of this study is just how consistent the performance ranking remains across the six evaluation metrics used. MobileNetV2 ranks first on accuracy, precision, recall, F1-score, specificity, and AUROC alike, giving it an aggregate rank score of 6, the best possible outcome. DenseNet121 follows with an aggregate of 13, and EfficientNetB3 trails behind every other model on every single metric, with an aggregate of 30. Between these two lies a small but interesting exception to an otherwise near-perfect ranking: once specificity is calculated correctly, ResNet50 (aggregate 17) actually edges out DenseNet121 on that one metric specifically (0.948 against 0.944), even though DenseNet121 remains clearly ahead overall thanks to its stronger recall, precision, and AUROC. InceptionV3 rounds out the ranking with an aggregate of 24.

This near-perfect consistency, five out of six metrics agreeing entirely on the ranking, with the sixth showing only a single, minor swap between two closely matched models, is a strong indication that the observed differences reflect genuine architectural properties rather than artefacts of any one metric or statistical noise. This matters because each metric is vulnerable to a different kind of failure: accuracy can be exploited through majority-class prediction under class imbalance, precision can be inflated by a model that makes very few positive predictions, recall can be inflated in the opposite direction by a model that predicts positive for almost everything, F1-score guards against exactly this kind of imbalance, specificity captures negative-class performance independently of recall, and AUROC evaluates the full probability output across every threshold rather

than one operating point. That MobileNetV2 comes out on top across all of these is fairly strong evidence that its advantage is genuine and well balanced, not an artefact of how performance happens to be measured.

This consistency also has a direct, practical bearing on how an architecture might be chosen for clinical use. Different deployment settings tend to prioritise different metrics: a population-level screening programme might weight recall most heavily, so as not to miss any malignancies, while a specialist referral centre might instead prioritise specificity, to cut down on unnecessary procedures, and a general-purpose diagnostic aid might reasonably favour F1-score or AUROC as a balanced compromise. Because MobileNetV2 leads across essentially every metric at once, it remains the preferred choice regardless of which clinical priority is in play, a genuinely useful property, since the recommendation does not hinge on getting the choice of evaluation metric exactly right.

6.5 Clinical Implications and Deployment Considerations

Turning experimental results of this kind into something clinically useful requires looking well beyond the classification metrics themselves, toward how a model would actually fit into a clinical workflow, what the population-level consequences of each error type would be, what validation and regulatory hurdles would need to be cleared, and what infrastructure would realistically be needed to support deployment over the long term.

6.5.1 The Role of AI in the Diagnosis of Lipomatous Tumours

Radiological diagnosis of lipomatous tumours currently relies on MRI, the gold-standard imaging modality for soft tissue masses of the extremities. Unfortunately, the standard MRI characteristics of lipoma and WDLPS overlap considerably, resulting in high sensitivity but quite limited specificity, around 37%, when radiological impression is compared against the definitive molecular marker, MDM2 gene amplification confirmed by biopsy. Set against that backdrop, the accuracy achieved by MobileNetV2 (0.967) and DenseNet121 (0.963), together with their near-perfect AUROC values (0.997 and 0.996 respectively), suggests that CNN-based classification of T1-weighted MRI could meaningfully reduce how often MDM2 biopsy needs to be pursued, particularly for cases the model classifies with high confidence, while lowering procedural risk and healthcare cost without compromising diagnostic safety.

It is important, all the same, to be clear that AI-assisted classification of this kind should support, rather than replace, the clinical judgement of an experienced radiologist. The models evaluated here were trained and tested on 2D axial slices drawn from a single, publicly available database, and how well they would perform in routine practice will depend heavily on the diversity of MRI acquisition protocols, scanner manufacturers, field strengths, and patient populations encountered outside that database. External validation on prospective, multi-centre data, collected using scanners and protocols not represented in the WORC database, would be an essential prerequisite before any of these models could reasonably be considered for clinical deployment. In practice, they would be best embedded within a human-in-the-loop workflow, where the model's output serves as one

additional probabilistic signal informing, rather than determining, the radiologist's final diagnostic decision.

6.5.2 Threshold Calibration for Asymmetric Clinical Costs

The default classification threshold of 0.5 was used throughout this study purely for methodological consistency across architectures. In an actual clinical deployment, the threshold would need to be calibrated to reflect the genuinely asymmetric costs of the two error types for the specific use case at hand. For a population-based screening programme aiming to triage patients toward MDM2 biopsy, for instance, a lower threshold, perhaps around 0.3, would likely be more appropriate, accepting a somewhat higher false positive rate in exchange for fewer missed malignancies, given that the consequences of a missed WDLPS, potential local recurrence, progression to DDLPS, and delayed treatment, are considerably more serious than those of an unnecessary biopsy. The high AUROC values recorded by MobileNetV2 (0.997) and DenseNet121 (0.996) confirm that both models retain excellent discriminative capability across the full range of thresholds, giving clinicians real flexibility to select a context-appropriate operating point without giving up much, if anything, in overall diagnostic performance.

6.5.3 Deployment Feasibility and Infrastructure Considerations

MobileNetV2's computational efficiency, roughly 3.5 million parameters and around 300 million multiply-add operations per forward pass, gives it a clear practical advantage over heavier architectures such as ResNet50 (about 25.6 million parameters) or InceptionV3 (about 23.9 million parameters). Its ability to run near-real-time inference on ordinary CPU hardware, without needing dedicated GPU infrastructure, makes it a realistic candidate for integration into existing PACS workstations and radiology reading environments, without the need for costly hardware upgrades. This advantage matters most in lower-resource healthcare settings, where high-end GPU hardware may simply not be available, even though the disease burden of soft tissue tumours can be substantial and access to specialised sarcoma centres or MDM2 biopsy facilities may be limited. In this sense, MobileNetV2's computational efficiency complements its classification performance rather well, aligning both with what the data show and with the practical demands of deploying such a system fairly and sustainably across a range of healthcare environments.

Chapter-7 : Conclusion

This thesis has presented a systematic, carefully controlled comparative study of five pre-trained CNN architectures, namely ResNet50, DenseNet121, EfficientNetB3, InceptionV3, and MobileNetV2, for the automated binary classification of benign lipoma and well-differentiated liposarcoma (WDLPS) using T1-weighted MRI data drawn from the publicly available WORC database. A strictly uniform, frozen transfer learning framework was applied across all five architectures, so any performance differences observed can reasonably be attributed to their own architectural properties rather than to differences in training configuration or preprocessing. Taken as a whole, the findings offer a reproducible empirical baseline for the growing literature on deep learning assisted lipomatous tumour classification, along with reasonably clear, evidence-based recommendations for architecture selection within this clinical domain.

7.1 Summary of Findings

Four of the five architectures evaluated, MobileNetV2, DenseNet121, ResNet50, and InceptionV3, achieved clinically excellent classification performance, with test-set accuracies ranging from 0.952 to 0.967 and AUROC values from 0.991 to 0.997. This is a fairly strong indication that transfer learning from ImageNet pre-trained weights works well for lipomatous tumour classification on T1-weighted MRI, even without any domain-specific fine-tuning of the backbone. EfficientNetB3 told a very different story, failing badly at only 0.635 accuracy and an AUROC of 0.672, with a severe prediction bias that renders it clinically unsuitable for this task in its present configuration.

MobileNetV2 achieved the strongest overall performance on every metric considered, recording 0.967 accuracy, a macro precision of 0.965, a macro recall of 0.966, an F1-score of 0.965, a corrected specificity of 0.964, and an AUROC of 0.997, alongside the lowest total error count (90 out of 2,699 test samples) and fewest false positives (39) of any model. This is particularly noteworthy given that MobileNetV2, with only around 3.5 million parameters, is also the smallest and most computationally efficient architecture in the comparison, more than seven times smaller than ResNet50, yet it still outperforms every larger model on every single metric. This suggests that MobileNetV2's design choices, particularly its depthwise separable convolutions and linear bottleneck projections, matter more for transfer learning success here than either parameter count or ImageNet benchmark score.

DenseNet121 came a close second, with 0.963 accuracy and an AUROC of 0.996, and it recorded the lowest false negative count of any model (40), corresponding to the highest WDLPS recall in the study (0.975). Its dense connectivity, which preserves both low-level textural and high-level semantic features throughout the network, seems particularly well

suited to picking up subtle textural changes in early-stage or low-contrast WDLPS, making it arguably the most defensible choice where minimising missed malignant diagnoses is the overriding priority.

ResNet50 and InceptionV3 formed a closely matched second tier, at 0.956 and 0.952 accuracy with AUROC values of 0.994 and 0.991 respectively, both reasonably well-balanced without any of EfficientNetB3's failure modes. The modest gap separating them from the top two appears to stem from ResNet50's comparatively thinner feature reuse relative to DenseNet121, and from the input resolution mismatch affecting InceptionV3's multi-scale processing. Both remain reliable, clinically viable options where sufficient computational resources are available.

EfficientNetB3, by contrast, failed dramatically, with a false positive rate of 88.511% on the benign Lipoma class, correctly identifying only 125 of 1,088 Lipoma cases. Its corrected specificity of just 0.115 and AUROC of 0.672, barely above chance, confirm it offers essentially no clinically useful discriminative capability here. This appears to stem from several factors together: the mismatch between its native 300×300 design resolution and the 224×224 input used throughout this study, a likely incompatibility between its Squeeze-and-Excitation attention mechanisms and MRI statistics quite different from ImageNet's, and the frozen backbone itself, which leaves no room for corrective adaptation. EfficientNetB3 should not be used for this task in its current, frozen configuration.

The ranking across all six metrics stayed consistent to a striking degree: MobileNetV2 ranked first and EfficientNetB3 last on every one without exception, and DenseNet121 held second on five of the six. The one exception is specificity, where, once calculated correctly, ResNet50 (0.948) narrowly edges out DenseNet121 (0.944), even though DenseNet121 remains clearly ahead overall thanks to its stronger recall, precision, and AUROC. If anything, this small exception reinforces the overall picture, since the broad ranking holds remarkably stable across six quite different ways of measuring performance. Taken together, MobileNetV2 is recommended as the primary architecture for general-purpose deployment, given its superior accuracy, AUROC, and efficiency, while DenseNet121 is recommended for safety-critical screening applications where maximising malignancy recall matters most.

7.2 Limitations

This study makes a genuine contribution to the automated classification of lipomatous tumours, but several important limitations need to be acknowledged openly, since they bear directly on how far these conclusions can reasonably be generalised.

The most fundamental is the size and scope of the dataset itself. The WORC Lipo dataset covers just 115 patients from a single database, a relatively modest number that may not fully capture how lipomatous tumours present across different patient demographics, body regions, and disease stages, and rare presentations of either class may be underrepresented as a result. A second limitation is the use of 2D axial slice extraction from what were originally 3D volumes, which discards the spatial relationships between adjacent slices and loses volumetric features such as the three-dimensional distribution of fibrous septa, information that cannot be recovered no matter how many individual slices are examined. Related to this is a third limitation: since adjacent slices from the same patient are highly

correlated rather than genuinely independent, evaluating performance at the slice level, as this study does, likely paints a somewhat optimistic picture of how well the model would generalise to genuinely new patients, and a patient-level evaluation would offer a more conservative and clinically meaningful estimate.

A fourth limitation is the frozen transfer learning strategy itself. While well justified for keeping the comparison fair, it represents only one of several possible configurations, and partial fine-tuning of the later convolutional layers may well have yielded higher performance for some architectures, EfficientNetB3 and InceptionV3 especially. A fifth limitation is the restriction to T1-weighted sequences alone, when routine clinical practice draws on fat-saturated T1, T2, and contrast-enhanced sequences too, each offering complementary information that could matter in borderline cases. A sixth and final limitation is the exclusive reliance on imaging data without any clinical or molecular information, such as patient age, tumour location, or serum biomarkers, all independently associated with malignancy risk and routinely used in real clinical decision-making, but unavailable to a purely image-based classifier.

7.3 Future Work

These findings and limitations point toward several worthwhile directions for future research. The most immediate is a systematic investigation of partial fine-tuning, selectively unfreezing the later convolutional layers of each backbone with a layer-wise decaying learning rate, which for EfficientNetB3 specifically, combined with restoring its native 300×300 resolution, could directly address the two main causes identified for its failure here.

A second direction is developing 3D volumetric CNN architectures, such as 3D ResNet or 3D DenseNet, capable of processing MRI volumes directly rather than through 2D slices, which would also naturally support patient-level evaluation. A third direction is multi-modal classification, combining CNN-extracted MRI features with structured clinical and molecular data such as patient age, tumour location, and MDM2 amplification status, potentially through feature concatenation or cross-modal attention, which could improve both accuracy and interpretability, particularly in borderline cases. A fourth direction is extending the comparison to vision transformers and hybrid CNN-transformer models, such as the Swin Transformer, whose self-attention mechanisms may better capture long-range morphological relationships that convolutional architectures can struggle with.

Finally, and perhaps most importantly for clinical translation, prospective external validation of MobileNetV2 and DenseNet121 on multi-centre, multi-scanner datasets, using protocols not represented in the WORC database, would be essential before their clinical generalisability can genuinely be established, ideally assessed across tumour size, anatomical location, histological grade, and patient demographics. Real-world prospective deployment studies, benchmarked against expert radiologist performance, would ultimately be needed to establish the true clinical utility of these approaches.

7.4 Analysis of the Social Effects

Developing an automated classification system of this kind carries social implications well beyond the technical results themselves. A reliable AI-assisted tool could shorten the time between imaging and diagnosis, offering a fast, consistent second opinion in settings where specialist radiologists are scarce or review is delayed. For patients, it could meaningfully reduce unnecessary anxiety and spare a fair number of people from invasive biopsies they never actually needed, while enabling earlier, more confident detection and treatment for those who do have a genuine malignancy.

There are equity implications too. MobileNetV2, the best-performing model here, needs only around 3.5 million parameters and runs efficiently on standard hardware without a dedicated GPU, making it realistic for lower-resource clinical settings, such as district hospitals or point-of-care facilities, where specialist sarcoma expertise is scarce. That said, these benefits can only be realised if such tools are deployed responsibly: leaning too heavily on AI predictions without adequate clinical oversight could cause real harm, particularly on patient populations not well represented in training data. Both patients and clinicians need to understand that such tools are decision support aids, not substitutes for clinical judgement, and public trust ultimately depends on transparency about how these systems work and where their limits lie.

7.5 Environmental Impact Analysis and Sustainability

Deep learning research carries a real environmental cost, since training neural networks consumes substantial electricity, much of it still generated from fossil fuels in many parts of the world, and this study makes a few methodological choices that, as a side effect, help reduce its own footprint. The most significant is relying on transfer learning rather than training every model from scratch: by updating only the classification head, the total computation required amounts to a small fraction of what training five large CNNs from zero would have needed. The EarlyStopping and ReduceLROnPlateau callbacks add to this, letting training stop automatically once performance plateaued rather than running the full 300 epochs regardless, a saving that adds up across five architectures and multiple runs.

The choice of MobileNetV2 as the best-performing model also carries favourable implications for deployment, since with only around 3.5 million parameters and roughly 300 million multiply-add operations per inference pass, it is dramatically more energy-efficient in everyday use than heavier models such as ResNet50. In a clinical setting processing hundreds or thousands of MRI slices daily, this saving adds up to something genuinely significant over time, and future work in this area would do well to report training time, hardware, and energy consumption more transparently as standard practice.

7.6 Ethical Issues

Applying AI to medical diagnosis raises genuine ethical questions that deserve attention alongside the technical results. On data privacy, the WORC database used here is publicly available and de-identified, though the broader question of how medical imaging data

should be collected and shared for AI research more generally remains an active and unresolved debate, and securing meaningful informed consent for this kind of reuse remains a genuine shared responsibility.

On algorithmic bias, the WORC Lipo dataset's 115 patients form a relatively small and, quite plausibly, not especially diverse cohort, and since its demographic composition is not fully documented, it is difficult to say whether these models generalise equally well across different age groups, sexes, and ethnic backgrounds. Any clinical deployment would need to be preceded by validation on a more diverse population. On accountability, clinical responsibility ultimately rests with the treating physician regardless of whether an AI tool was involved, so clinicians should treat a model's output as one additional input to weigh, never as a substitute for their own judgement, and should communicate its role clearly to patients. Finally, on fairness of access, there is a real risk that the benefits of AI-assisted diagnosis end up concentrated in already well-resourced health systems; ensuring lower-resource settings receive matching investment in infrastructure, training, and governance is a responsibility shared by researchers, developers, and policymakers alike.

Bibliography

[1] A. T. J. Lee, K. Thway, P. H. Huang, and R. L. Jones, "Clinical and molecular spectrum of liposarcoma," *Journal of Clinical Oncology*, vol. 36, no. 2, pp. 151–159, 2018.

[2] M. Y. Zhou, N. Q. Bui, G. W. Charville, K. N. Ganjoo, and M. Pan, "Treatment of de-differentiated liposarcoma in the era of immunotherapy," *International Journal of Molecular Sciences*, vol. 24, no. 11, p. 9571, 2023.

[3] E. J. Yee, C. L. Stewart, M. R. Clay, and M. M. McCarter, "Lipoma and its doppelganger: The atypical lipomatous tumor/well-differentiated liposarcoma," *Surgical Clinics of North America*, vol. 102, no. 4, pp. 637–656, 2022.

[4] M. Brisson, T. Kashima, D. Delaney, R. Tirabosco, A. Clarke, S. Cro, A. M. Flanagan, and P. O'Donnell, "Mri characteristics of lipoma and atypical lipomatous tumor/well-differentiated liposarcoma: retrospective comparison with histology and mdm2 gene amplification," *Skeletal Radiology*, vol. 42, no. 5, pp. 635–647, 2013.

[5] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017.

[6] B. Sistaninejhad, H. Rasi, and P. Nayeri, "A review paper about deep learning for medical image analysis," *Computational and Mathematical Methods in Medicine*, vol. 2023, no. 1, 2023.

[7] X. Chen, X. Wang, K. Zhang, K. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, vol. 79, p. 102444, 2022.

[8] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 1, pp. 19–38, 2021.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas: IEEE, 2016, pp. 770–778.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu: IEEE, 2017, pp. 4700–4708.

[11] M. Tan and Q. V. Le, "Efficient net: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*. Long Beach: PMLR, 2019, pp. 6105–6114.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas: IEEE, 2016, pp. 2818–2826.

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenet v2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City: IEEE, 2018, pp. 4510–4520.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Miami: IEEE, 2009, pp. 248–255.

[15] B. Wang, L. Perronne, C. Burke, and R. S. Adler, "Artificial intelligence for classification of soft-tissue masses at us," *Radiology: Artificial Intelligence*, vol. 3, no. 1, p. e200125, 2020.

[16] G. Fradet, R. Ayde, H. Bottois, M. El Harchaoui, W. Khaled, J. Drapé, F. Pilleul, A. Bouhamama, O. Beuf, and B. Leporq, "Prediction of lipomatous soft tissue malignancy on mri: comparison between machine learning applied to radiomics and deep learning," *European Radiology Experimental*, vol. 6, no. 1, p. 41, 2022.

[17] H. Hermessi, O. Mourali, and E. Zagrouba, "Transfer learning with multiple convolutional neural networks for soft tissue sarcoma mri classification," in *11th International Conference on Machine Vision (ICMV)*, vol. 11041. Munich: SPIE, 2019, p. 110412L.

[18] Y. Yang, Y. Zhou, C. Zhou, and X. Ma, "Novel computer aided diagnostic models on multimodality medical images to differentiate well differentiated liposarcomas from lipomas approached by deep learning methods," *Orphanet Journal of Rare Diseases*, vol. 17, no. 1, p. 158, 2022.

[19] I. Malinauskaite, J. Hofmeister, S. Burgermeister, A. Neroladaki, M. Hamard, X. Montet, and S. Boudabbous, "Radiomics and machine learning differentiate soft-tissue lipoma and liposarcoma better than musculoskeletal radiologists," *Sarcoma*, vol. 2020, p. 7163453, 2020.

[20] S. Gitto, M. Interlenghi, R. Cuocolo, C. Salvatore, V. Giannetta, J. Badalyan, E. Gallazzi, M. S. Spinelli, M. Gallazzi, F. Serpi, C. Messina, D. Albano, A. Annovazzi, V. Anelli, J. Baldi, A. Aliprandi, E. Armiraglio, A. Parafioriti, P. A. Daolio, A. Luzzati, R. Biagini, I. Castiglioni, and L. M. Sconfienza, "Mri radiomics-based machine learning for classification of deep-seated lipoma and atypical lipomatous tumor of the extremities," *La Radiologia Medica*, vol. 128, no. 8, pp. 989–998, 2023.

[21] N. Cay, B. A. R. Mendi, H. Batur, and F. Erdogan, "Discrimination of lipoma from atypical lipomatous tumor/well-differentiated liposarcoma using magnetic resonance imaging radiomics combined with machine learning," *Japanese Journal of Radiology*, vol. 40, no. 9, pp. 951–960, 2022.

[22] F. Navarro, H. Dapper, R. Asadpour, C. Knebel, M. B. Spraker, V. Schwarze, S. K. Schaub, N. A. Mayr, K. Specht, H. C. Woodruff, P. Lambin, A. S. Gersing, M. J. Nyflot, B. H. Menze, S. E. Combs, and J. C. Peeken, "Development and external validation of deep-learning-based tumor grading models in soft-tissue sarcoma patients using mr imaging," *Cancers*, vol. 13, no. 12, p. 2866, 2021.

[23] S. Liu, W. Sun, S. Yang, L. Duan, C. Huang, J. Xu, F. Hou, D. Hao, T. Yu, and H. Wang, "Deep learning radiomic nomogram to predict recurrence in soft tissue sarcoma: a multi-institutional study," *European Radiology*, vol. 32, no. 2, pp. 793–805, 2021.

[24] M. P. A. Starmans, M. J. M. Timbergen, M. Vos, G. A. Padmos, D. J. Grünhagen, C. Verhoef, S. Sleijfer, G. J. L. H. van Leenders, F. E. Buisman, F. E. J. A. Willemssen, B. Groot Koerkamp, L. Angus, A. A. M. van der Veldt, A. Rajicic, A. E. Odink, M. Renckens, M. Doukas, R. A. de Man, J. N. M. IJzermans, R. L. Miclea, P. B. Vermeulen, M. G. Thomeer, J. J. Visser, W. J. Niessen, and S. Klein, "The worc database: Mri and ct scans, segmentations, and clinical labels for 930 patients from six radiomics studies," *medRxiv preprint*, 2021.

[25] C. Liu, Y. G. Abdelhafez, S. P. Yap, F. Acquafredda, S. Schirò, A. L. Wong, D. Sarohia, C. Bateni, M. A. Darrow, M. Guindani, S. Lee, M. Zhang, A. W. Moawad, Q. K. T. Ng, L. Shere, K. M. Elsayes, R. Maroldi, T. M. Link, L. Nardo, and J. Qi, "Ai-based automated lipomatous tumor segmentation in mr images: Ensemble solution to heterogeneous data," *Journal of Digital Imaging*, vol. 36, no. 3, pp. 1049–1059, 2023.

[26] M. Zheng, C. Guo, Y. Zhu, X. Gang, C. Fu, and S. Wang, "Segmentation model of soft tissue sarcoma based on self-supervised learning," *Frontiers in Oncology*, vol. 14, p. 1247396, 2024.

[27] A. Bozzo, A. Hollingsworth, S. Chatterjee, A. Apte, J. Deng, S. Sun, W. Tap, A. Aoude, S. Bhatnagar, and J. H. Healey, "A multimodal neural network with gradient blending improves predictions of survival and metastasis in sarcoma," *npj Precision Oncology*, vol. 8, no. 1, p. 188, 2024.

[28] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[29] N. A. A.-H., , and Prince, "A classification of arab ethnicity based on face image using deep learning approach," *IEEE Access*, vol. 9, pp. 50 755–50 766, 2021.

[30] "Hybridtransfernet: Advancing soil image classification through comprehensive evaluation of hybrid transfer learning - scientific figure," accessed: Aug. 30, 2025. [Online]. Available: https://www.researchgate.net/figure/DenseNet-121-Architecture-Diagram28_fig5_371453594

[31] "Cultivar identification of pistachio nuts in bulk mode through efficientnet deep learning model - scientific figure," accessed: Aug. 30, 2025. [Online]. Available: https://www.researchgate.net/figure/Schematic-representation-of-EfficientNet-B3_fig2_359449935

[32] O. Iparraguirre-Villanueva, V. Guevara-Ponce, O. R. Paredes, F. Sierra-Liñan, J. Zapata-Paulini, and M. Cabanillas-Carbonell, "Convolutional neural networks with transfer learning for pneumonia detection," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, 2022.

[33] "Mobilenet-jde: A lightweight multi-object tracking model for embedded systems - scientific figure," accessed: Aug. 30, 2025. [Online]. Available: https://www.researchgate.net/figure/Architecture-of-the-MobileNetV2-backbone-model_fig4_358602220

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6628106