Independent University **Bangladesh (IUB)**

IUB Academic Repository

Computer Science and Engineering | Undergraduate Thesis

2026-06

Ultrasound-based liver disease classification using deep convolutional neural networks

Noor, Sazin Bin

Independent University, Bangladesh

https://ar.iub.edu.bd/handle/11348/1403

Downloaded from IUB Academic Repository



Ultrasound-Based Liver Disease Classification Using Deep Convolutional Neural Networks

June 2026

Prepared by:

Sazin Bin Noor

ID: 2110806

Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by

Md. Rashedur Rahman, D. Eng.
Assistant Professor
Department of Computer Science and Engineering
Independent University, Bangladesh

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.

Author Name:

.....

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 deepest gratitude to Md. Rashedur Rahman, D.Eng., Assistant Professor, Independent University, Bangladesh, for his invaluable guidance, continuous encouragement, and expert insights throughout the course of this research. His mentorship and thoughtful feedback have been instrumental in shaping the direction and quality of our work.

We also extend our sincere appreciation to Fatin Israq Tabib, Research Assistant at the Center for Computational & Data Sciences (CCDS Lab), for his dedicated support, technical assistance, and valuable suggestions during the development and experimentation phases of this project.

Our heartfelt thanks go to the CCDS Lab (Center for Computational & Data Sciences) for providing a collaborative and resourceful research environment. The lab's consistent support and access to essential resources greatly contributed to the successful completion of this study.

Finally, we are grateful to all individuals who have supported and encouraged us, both directly and indirectly, during this endeavor. Their contributions have played a meaningful role in the realization of this work.

Abstract

Liver disease is a significant health concern in the world, which has led to millions of deaths annually. Complications associated with the disease include cirrhosis, liver failure, and hepatocellular carcinoma. Early and correct diagnosis minimizes these risks, enhances treatment results and facilitates timely clinical decisions. One of the most common diagnostic methods to assess the liver is ultrasound imaging, as it is safe, non-invasive, cheap and can be used in both urban and resource-restricted healthcare facilities. Nonetheless, ultrasound interpretation is usually subjective and relies on operator expertise, level of experience and changes in image quality despite its availability. This renders automated and standardized diagnostic support as a significant field of study.

This paper explores the potential of modern convolutional neural network (CNN) architectures to aid in classifying liver ultrasound images into three clinically relevant categories, which include normal, benign, and malignant. We concentrate on three popular CNN families EfficientNet, MobileNet, and ConvNeXt each with different design philosophies in deep learning. MobileNet focuses on lightweight operations and algorithms, EfficientNet provides balanced scaling of compounds to enhance accuracy at fewer parameters, and ConvNeXt applies modernized convolutional architectures based on the transformer-like principles. Each of the models was trained in a frozen-backbone transfer learning configuration, with all experiments being fairly compared by maintaining similar preprocessing, training parameters, and classification layers.

This study uses a dataset of annotated ultrasound images, which is publicly available. Though the dataset is quite small and unbalanced, it is representative of real-world clinical situations in which malignant cases are more likely to be reported. All the models were trained on a single pipeline. To measure performance, Accuracy, Precision, Recall, F1-Score, and AUROC were used.

MobileNetV2 was the most successful of all the tested models, with an F1-Score of 75.20%, an Accuracy of 78.37%, and a recall of 78.03%. These findings indicate that not only did MobileNetV2 make accurate predictions but also it was highly effective in the detection of malignant cases. EfficientNetB2 was also quite stable, achieving 77.02% Accuracy and showing good balance in all measures. ConvNeXt variants were competitive but did not outperform MobileNetV2 or EfficientNetB2, probably because they have more capacity and are sensitive to small datasets trained with frozen backbone.

Overall, the findings show that lightweight and moderately sized CNNs can provide strong diagnostic support for liver ultrasound classification without requiring large number of computational resources. These results highlight the potential for accessible AI-driven tools that support radiologists, reduce interpretation variability, and improve early detection of liver disease.

Contents

Chapter 1: Introduction 10
1.1 Background 11
1.2 Motivation 13
1.3 Scope of Study 13
1.4 Research Questions 14
1.5 Objective 15
1.5.1 Primary Objective 15
1.5.2 Secondary Objective 15
1.6 Contribution 15
Chapter 2: Literature Review 16
Chapter 3: Dataset 19
Chapter 4: Methodology 21
4.1 CNN Models 21
4.1.1 MobileNet Family 21
4.1.2 ConvNeXt Family 28
4.1.3 EfficientNet Family 33
4.2 Training Setup 39
4.3 Evaluation Metrics 40
Chapter 5: Results and Discussion 42
5.1 Overview of experiments 42
5.2 MobileNet Results 42
5.2.1 Performance Summary 42
5.2.2 Training and Validation Curves 43
5.2.3 Confusion Matrices 45
5.2.4 ROC Curves 47
5.3 ConvNeXt Results 49
5.3.1 Performance Summary 49
5.3.2 Training and Validation Curves 50
5.3.3 Confusion Matrices 52

5.3.4 ROC Curves 54
5.4 EfficientNet Results 56
5.4.1 Performance Summary 56
5.4.2 Training and Validation Curves 57
5.4.3 Confusion Matrices 59
5.4.4 ROC Curves 61
5.5 Comparison of the best models 63
Chapter 6: Conclusion 65
6.1 Summary 65
6.2 Limitations 65
6.3 Future Work 66
References 68

Table of Figures

Figure 3.1: Distribution of ultrasound images across malignant, benign, and normal classes in the annotated liver dataset. 19
Figure 3.2: Example ultrasound images representing the three classes in the dataset: normal, benign, and malignant..... 20
Figure 4.5: ConvNeXT Architecture..... 30
Figure 4.6 EfficientNetB1 Architecture 34
Figure 4.7: EfficientNetB2 Architecture 36
Figure 4.8: EfficientNetB6 Architecture 38
Figure 4.9: Model architecture used across all experiments. 39
Figure 5.1: Training & Validation Curve of MobileNet 43
Figure 5.2: Training & Validation Curve of MobileNetV2..... 44
Figure 5.3: Training & Validation Curve of MobileNetV3Large 44
Figure 5.4: Confusion Matrix of MobileNet..... 45
Figure 5.5: Confusion Matrix of MobileNetV2 46
Figure 5.6: Confusion Matrix of MobileNetV3Large..... 46
Figure 5.7: Receiver Operating Characteristic (ROC) curves for MobileNet 47
Figure 5.8: Receiver Operating Characteristic (ROC) curves for MobileNetV2 48
Figure 5.9: Receiver Operating Characteristic (ROC) curves for MobileNetV3Large 48
Figure 5.10: Training & Validation Curve for ConvNeXt Small..... 50
Figure 5.11: Training & Validation Curve for ConvNeXt Large 50
Figure 5.12:Training & Validation Curve for ConvNeXtXLarge 51
Figure 5.13: Confusion matrices for ConvNeXt Small 52
Figure 5.14: Confusion matrices for ConvNeXt Large 53
Figure 5.15: Confusion matrices for ConvNeXtXLarge..... 53
Figure 5.16: Receiver Operating Characteristic (ROC) curves for ConvNeXt Small 54
Figure 5.17: Receiver Operating Characteristic (ROC) curves for ConvNeXtLarge 55
Figure 5.18: Receiver Operating Characteristic (ROC) curves for ConvNeXtXLarge 55
Figure 5.19: Training & Validation Curve for EfficientNetB1 57
Figure 5.20: Training & Validation Curve for EfficientNetB2 58
Figure 5.21: Training & Validation Curve for EfficientNetB6 58
Figure 5.22: Confusion matrices for EfficientNetB1..... 59
Figure 5.23: Confusion matrices for EfficientNetB2..... 60
Figure 5.24: Confusion matrices for EfficientNetB6..... 60
Figure 5.25: Receiver Operating Characteristic (ROC) curves for EfficientNetB1 61
Figure 5.26: Receiver Operating Characteristic (ROC) curves for EfficientNetB2 62
Figure 5.27: Receiver Operating Characteristic (ROC) curves for EfficientNetB6 62

Table of Tables

Table 4.1: Hyperparameter Summary 40
Table 5.1: Performance metrics for MobileNet, MobileNetV2, and MobileNetV3Large on liver ultrasound classification. 42
Table 5.2: Performance metrics for ConvNeXtSmall, ConvNeXtLarge, and ConvNeXtXLarge. 49
Table 5.3: Performance metrics for EfficientNetB1, EfficientNetB2, and EfficientNetB6. 56
Table 5.4: Performance comparison of MobileNetV2, EfficientNetB2, and ConvNeXtXLarge as the top-performing variants in each model family. 63

Chapter 1: Introduction

Liver disease includes many conditions that harm the liver and disrupt the essential work it performs every day. The liver helps the body digest food, process nutrients, remove toxins, store energy, and support the immune system. When the liver becomes damaged whether suddenly or slowly over many years these functions weaken, and serious health problems can follow. Liver disease may be acute, where the damage happens quickly, or chronic, where the liver gets worse over a long period. Chronic liver disease usually progresses through well-known stages: inflammation, fibrosis (scarring), cirrhosis (advanced scarring), and in many cases liver cancer, especially hepatocellular carcinoma (HCC), which is the most common type of primary liver cancer worldwide [1].

There are many causes of liver disease. Viral hepatitis remains one of the leading global causes. Chronic hepatitis B (HBV) and hepatitis C (HCV) infect liver cells and cause long-term inflammation that can progress to cirrhosis or liver cancer if untreated [3]. Another major cause is metabolic dysfunction–associated steatotic liver disease (MASLD), which was previously called non-alcoholic fatty liver disease (NAFLD). MASLD is strongly linked to high blood pressure and type 2 diabetes. MASLD starts with fat accumulating in the liver and can progress to cirrhosis if not managed early [4]. Alcohol-associated liver disease (ALD) also affects millions of people. Viral hepatitis and MASLD together account for most chronic liver disease cases globally.

Autoimmune hepatitis happens when the body's own immune system mistakenly attacks liver tissue. And certain medications or supplements can cause drug induced liver injury, which may progress rapidly if not identified. Genetic conditions like hemochromatosis and Wilson disease also lead to long-term liver damage if untreated [2].

Liver diseases are also classified by type and severity. Early stages include steatosis (fat buildup) or hepatitis (inflammation). More advanced conditions include fibrosis and cirrhosis. Another major group of liver diseases involves tumors. These tumors fall into two categories: benign and malignant, Benign lesions include simple cysts and focal nodular hyperplasia. These do not spread but may still require monitoring, whereas Malignant lesions, such as hepatocellular carcinoma and intrahepatic cholangiocarcinoma can grow quickly and spread to other parts of the body, making early detection essential [1].

Ultrasound is often the first imaging method doctors use to evaluate the liver because it is widely available and does not involve radiation. However, ultrasound images can be hard to interpret. The operator's skill and patient's body type play a major role in determining the reliability of ultrasound images. Classification can be made difficult even for experienced radiologists when benign and malignant lesions appear very similar. As a result, achieving early and accurate detection continues to be a significant challenge.

As the global burden of liver disease continues to increase, the need for more effective diagnostic tools becomes increasingly obvious. Recent reports show rising rates of MASLD worldwide [4], millions living with chronic viral hepatitis [3], and liver cancer continuing to be a leading cause of cancer-related deaths [1]. These trends make it crucial to explore computer-based solutions that can support radiologists and improve consistency in clinical practice.

This thesis builds on that need by investigating whether modern convolutional neural networks can classify liver ultrasound images more accurately. The goal is to compare different CNN architectures and their performance across standard evaluation metrics.

1.1 Background

Liver disease remains a significant global health challenge, affecting hundreds of millions of people worldwide. Chronic viral hepatitis continues to represent a major contributor to this burden. In 2022, an estimated 254 million individuals were living with chronic hepatitis B and approximately 50 million with chronic hepatitis C, both of which remain leading causes of cirrhosis and hepatocellular carcinoma [3]. Beyond viral hepatitis, metabolic-associated steatotic liver disease (MASLD/NAFLD) has become one of the most common chronic liver conditions globally. A large meta-analysis reported that nearly one-third of adults worldwide are affected, reflecting rising rates of obesity and metabolic dysfunction [4]. Together, these conditions significantly contribute to chronic liver disease, which accounts for more than two million deaths annually [2].

The impact of liver cancer extends beyond mortality statistics and highlights the need for more effective diagnostic strategies. Primary liver cancer ranks among the leading causes of cancer-related deaths worldwide [1]. Early detection is essential for improving prognosis, yet liver diseases often progress silently until advanced stages, making imaging-based detection a critical component of preventive care.

Ultrasound imaging plays a central role in liver disease evaluation. Its advantages non-invasiveness, affordability, real-time capability, and widespread availability make it the preferred first-line modality for screening and follow-up, particularly in low-resource settings. However, despite these strengths, ultrasound interpretation poses challenges. Image quality is affected by patient anatomy, operator skill, and device variability. Subtle early-stage liver abnormalities, such as mild steatosis, early fibrosis, or small focal lesions, often present with minimal grayscale or textural differences, increasing the likelihood of missed or inconsistent interpretations.

Growing interest has been directed toward computer-aided diagnosis (CAD) and deep learning techniques as potential solutions to these limitations. Deep learning has demonstrated strong performance across medical imaging tasks, showing value in complex pattern recognition [5]. In ultrasound, review studies indicate that CNN-based models can significantly augment traditional interpretation by improving sensitivity and reducing variability [6]. In liver steatosis

assessment, transfer learning and feature extraction showed promising results for improving diagnostic accuracy in early studies [7]. Later studies expanded on these approaches by incorporating cascaded networks and multi-view ultrasound strategies. As a result, prediction performance improved across heterogeneous datasets [8, 9, 10, 11].

Advancements have also extended to focal liver lesions. Deep learning models have achieved promising accuracy in classifying benign and malignant solid liver lesions directly from B-mode ultrasound [13]. Additional work in contrast-enhanced ultrasound demonstrates that deep learning can support malignancy diagnosis comparable to expert interpretation [14]. Lesion recognition and classification performance improved further when ultrasound video frames or multi-phase imaging were incorporated into more complex diagnostic pipelines [20]. Emerging research has explored ultrasound radiomics combined with machine learning to support diagnosis of advanced liver diseases such as steatohepatitis [12].

Despite these advancements, several challenges persist. Many studies rely on institution-specific datasets with limited size or diversity, restricting model generalizability. Ultrasound is highly operator-dependent, and variations in acquisition protocols, device manufacturers, and patient characteristics introduce domain shifts that degrade performance when models are deployed across different clinical environments. Robust clinical deployment of AI-assisted ultrasound diagnostics depends on more than model performance alone. Domain adaptation, multi-center training, and standardized annotation protocols have emerged as key requirements for achieving reliable results across clinical settings [23, 24]. Another limitation is the lack of large, annotated datasets, though efforts to develop high-quality public ultrasound datasets have increased in recent years [21, 25].

Deep learning techniques have also been explored for broader liver disease detection, including fibrosis staging, lesion localization, and segmentation. Several studies reported strong performance using UNet-based architectures for liver and lesion segmentation, improving downstream classification accuracy and enabling semi-automated interpretation workflows [26]. In addition, recent investigations emphasize the importance of model interpretability and clinician-AI collaboration to enhance trust and integration into routine workflows [27].

Although deep learning methods have shown considerable promise, questions remain about which CNN architectures are best suited for liver ultrasound classification. Recent deep learning innovations such as EfficientNet, MobileNet families, and ConvNeXt offer enhanced efficiency, scalability, and representational capability [15–19]. Comparative analysis of this architecture families can provide insights into performance trade-offs suitability for real-world implementation. Ultimately, more accessible diagnostic support could be enabled through advances in liver ultrasound classification.

1.2 Motivation

Liver disease continues to rise globally, creating a need for timely and accurate diagnosis. Ultrasound remains one of the most widely used imaging modalities for liver assessment due to its safety and accessibility, especially in rural areas. However, ultrasound interpretation is inherently challenging; image quality can vary significantly, and diagnostic accuracy often depends on the experience and judgment of the interpreting clinician. As a result, higher diagnostic consistency may be achieved through tools designed to reduce missed findings.

Few studies have systematically compared different convolutional neural network (CNN) architectures for liver ultrasound classification. Many existing works focus on a single model or evaluate architecture under inconsistent experimental settings, making it difficult to reach conclusions about their relative strengths and limitations. To enable a fair comparison, multiple CNN families are evaluated under a unified training pipeline. Experimental variability is minimized so that architectural differences can be assessed more clearly.

Questions remain about how architectural design choices influence model performance in liver ultrasound classification. Modern CNN architectures such as MobileNet, EfficientNet, and ConvNeXt have fundamentally different design philosophies, yet their behavior on medical ultrasound data remains unexplored. A systematic analysis of these design choices can reveal which features are most effective for liver ultrasound classification. Their impact under limited and imbalanced data conditions can also be assessed more clearly.

1.3 Scope of Study

Deep learning models are evaluated for their ability to classify liver ultrasound images into three categories: normal, benign, and malignant. Performance differences among CNN architectures are examined, such as, their ability to capture complex ultrasound features. The analysis covers three major CNN families: EfficientNet, MobileNet, and ConvNeXt, each with different features, like balancing accuracy, and model complexity.

Multiple variants within each model family are included, ranging from compact architectures optimized for real-time or resource limited setting to more advanced models designed for higher predictive accuracy. By training and assessing these networks the study compares how model size, scaling strategies, and dataset limitations affect classification performance.

Comparing these architectures provides a clearer understanding of their strengths and limitations in liver ultrasound classification. The results can help guide future research and model development by identifying suitable approaches for medical imaging applications where accuracy, efficiency, and reliability are essential.

1.4 Research Questions

This study is driven by several research questions that define the objectives and scope of evaluating deep learning models for liver ultrasound classification. These questions focus on understanding model performance, architectural suitability, and the challenges associated with medical imaging data.

Primary Research Question

- How effectively can different convolutional neural network (CNN) architectures classify liver ultrasound images into normal, benign, and malignant categories?

Secondary Research Questions

- Which CNN model families (EfficientNet, MobileNet, ConvNeXt) demonstrate the strongest suitability for medical ultrasound data, and what architectural characteristics contribute to their performance differences?

- To what extent do model sizes such as deeper architecture impact classification accuracy, especially when training data are limited or imbalanced?

- How stable are the models during training, given the inherent noise, variability, and irregular texture patterns present in ultrasound images?

- How do different architectures handle class imbalance, particularly when malignant cases dominate the dataset, and do certain models exhibit better resilience to skewed distributions?

- What observations from this study suggest future directions for improving performance, such as the need for enhanced augmentation strategies, alternative loss functions, or expanded datasets?

Exploratory/Theoretical Questions (Beyond the Scope of This Thesis)

- How well would the evaluated models generalize to ultrasound images acquired from different hospitals, imaging devices, or scanning protocols?

- Could emerging architectures, such as Vision Transformers and hybrid CNN-Transformer models potentially outperform conventional CNNs in liver ultrasound classification?

1.5 Objective

1.5.1 Primary Objective

The primary objective of this study is to systematically compare different convolutional neural network (CNN) architectures for liver ultrasound image classification. The study evaluates three major CNN families EfficientNet, MobileNet, and ConvNeXt each representing distinct architectural design principles. By training all models under a unified and controlled experimental setup, this work aims to identify how differences in architecture influence performance in classifying ultrasound images into normal, benign, and malignant categories.

1.5.2 Secondary Objective

Model performance is assessed using standard classification metrics, including Accuracy, Precision, Recall, F1-score, and AUROC. Together, these measures provide a comprehensive view of each model's ability to classify images correctly, identify clinically important cases, and maintain balanced performance across classes. The evaluation also reveals which architecture delivers the most reliable and effective results for liver ultrasound classification.

1.6 Contribution

A systematic comparison of convolutional neural network (CNN) architectures for liver ultrasound image classification is presented in this study. EfficientNet, MobileNet, and ConvNeXt are evaluated under a unified experimental setup in which all models are trained and tested under identical conditions. This consistency reduces the influence of experimental variability and allows architectural differences to be examined more clearly.

Model performance is assessed using a range of evaluation metrics, including Accuracy, Precision, Recall, F1-score, and AUROC. Together, these measures provide a comprehensive view of each model's strengths and limitations, particularly in identifying clinically important cases that may be overlooked by less balanced evaluation strategies. Relying on multiple metrics rather than a single measure also enables a more nuanced assessment of classification performance across categories.

The influence of network depth, scaling strategies, and overall model complexity is examined through comparative analysis of the selected architectures. Performance trends observed across the evaluated models provide practical insight into their behavior under constrained conditions, including limited and imbalanced datasets that are commonly encountered in medical imaging applications.

Beyond comparing classification performance, the analysis offers a structured perspective on the characteristics of modern CNN architectures and their design philosophies. The resulting observations may help guide future research, model selection, and optimization efforts for liver ultrasound classification as well as related medical imaging tasks.

Chapter 2: Literature Review

Deep learning has transformed medical image analysis by enabling models to learn complex patterns directly from imaging data instead of relying on handcrafted features. Early surveys on deep learning in medical imaging highlighted its rapid growth across modalities such as CT, MRI, and ultrasound, while also identifying key challenges including limited datasets, inconsistent annotation quality, and variability across imaging devices [5]. In the area of ultrasound, deep learning has gained particular attention because it handles speckle noise, operator dependence, and subjective interpretation issues that often reduce diagnostic reliability in routine clinical settings [6]. These foundational reviews create a strong motivation for using deep learning to improve the accuracy and consistency of liver ultrasound interpretation. More recent work has also emphasized the need for architectures and training strategies that explicitly address the small-sample, high-variability nature of ultrasound data, including transfer learning, data augmentation, and synthetic data generation pipelines tailored to liver imaging [10, 11, 29, 31].
Further improvements in classification stability were reported through multi-view ultrasound approaches. Combining several images of the same patient, acquired from different scanners, was shown to reduce model sensitivity to device variability while improving overall classification performance [9].

High classification accuracy for NAFLD was also achieved using a multi-scale residual CNN architecture that integrates B-mode ultrasound images with phase- and symmetry-based transforms [29]. The study highlighted the value of combining multiple feature representations when analyzing liver ultrasound data.

To mitigate data scarcity, a GAN-based framework was later developed for realistic liver ultrasound image synthesis. When synthetic and real images were used together, CNN performance improved significantly, indicating the potential of data augmentation strategies in ultrasound-based diagnosis [31].

The application of deep learning was later extended beyond steatosis to diffuse liver disease classification. Ensembles of pretrained CNNs, including ResNet and ResNeXt variants, were used to distinguish normal liver, hepatitis, and cirrhosis, revealing the advantages of hybrid multi-network classifiers for ultrasound-based liver staging [30].

While deep learning classification has dominated much of the liver ultrasound literature, parallel research explores ultrasound radiomics, where handcrafted features are extracted and fed into machine learning or combined with CNNs. For example, a recent radiomics-based study for diagnosing non-alcoholic steatohepatitis (NASH) showed promising performance using ultrasound texture descriptors and machine learning classifiers, demonstrating that radiomics can complement deep-learning-based systems or serve as an alternative when data are scarce [12]. This diversity of methods deep learning, radiomics, and hybrid models illustrates the growing sophistication of noninvasive liver disease assessment. More recently, “deep radiomics” approaches have attempted to integrate radiomics-style feature interpretation

with deep feature extraction. Liu et al. developed a 3D CNN-based radiomics model that combines long-range contrast-enhanced ultrasound (CEUS) cines with clinical variables such as hepatitis status and alpha-fetoprotein, achieving excellent performance in benign–malignant focal liver lesion (FLL) discrimination and underscoring the value of jointly modeling imaging and clinical data [37].

Beyond steatosis, deep learning has also been applied to focal liver lesion classification, including the differentiation of benign lesions such as cysts or hemangiomas from malignant tumors such as hepatocellular carcinoma (HCC). Xi et al. trained and fine-tuned a ResNet-50 model on B-mode ultrasound images and found that it achieved high diagnostic accuracy, in some cases matching or exceeding the performance of experienced radiologists [13]. In contrast-enhanced ultrasound (CEUS), dynamic temporal information offers additional diagnostic cues. Li et al. developed a deep learning system capable of analyzing CEUS video sequences end-to-end, allowing the model to learn temporal enhancement patterns associated with malignant tumors without manual feature engineering [14]. Together, these studies demonstrate the potential of CNNs to support radiologists across both standard and contrast-enhanced ultrasound modalities. Earlier work by Hassan et al. used deep architectures to classify focal liver diseases from ultrasound, showing that deep representations substantially outperform conventional texture-based features for multi-class lesion categorization [36]. Schmauch et al. proposed a supervised-attention CNN that simultaneously detects and characterizes FLLs (benign vs. malignant) from 2D ultrasound images, achieving high ROC-AUC in both detection and characterization tasks and illustrating how attention mechanisms can focus the model on lesion regions without explicit segmentation [35]. Two large multi-centre EBioMedicine studies further demonstrated that deep CNNs can improve B-mode ultrasound diagnostic performance for FLLs and assist radiologists across different experience levels, particularly in low-resource settings where CT or MRI may be less accessible [33, 34]. In addition, DL-radiomics models that integrate CEUS cines, spatial–temporal features, and clinical factors have been shown to rival or exceed expert performance for benign–malignant differentiation [37]. Recent work has also explored real-time or near real-time AI assistance using ultrasound videos to detect FLLs at the point of care, reflecting an emerging focus on workflow integration rather than purely retrospective analysis [38].

Despite progress, the field lacks large, standardized, publicly available ultrasound datasets for liver disease classification. One of the few open datasets is the Annotated Ultrasound Liver Images dataset, containing 735 images across three categories: malignant (435), benign (200), and normal (100). This dataset was originally released as part of an AI pipeline for ultrasound-based malignancy diagnosis published in Briefings in Bioinformatics [20]. The Zenodo dataset record formalized the release and provided labeled images and segmentation masks to facilitate reproducible research [21]. Though community usage exists via Kaggle and GitHub notebooks, few peer-reviewed studies have performed systematic, multi-architecture comparisons using this dataset.

Few peer-reviewed studies have systematically compared multiple CNN architecture families using this dataset under consistent experimental conditions. Benchmarking EfficientNet,

MobileNet, and ConvNeXt within a unified framework provides an opportunity to examine their relative strengths and limitations more objectively.

Related large-scale efforts include multi-centre cohorts of B-mode FLL ultrasound images used to train and evaluate deep ensembles across numerous CNN backbones [32]. Results from these studies suggest that ensemble learning strategies can improve multi-class hepatic mass classification beyond the performance achieved by individual models. Other multi-centre evaluations and long-range CEUS cohorts similarly underscore the importance of diverse imaging conditions, multiple scanners, and external validation when developing robust liver ultrasound AI systems [32, 33, 37, 38].

These architectures represent three influential CNN design philosophies. MobileNet, introduced as a lightweight model for mobile and edge devices, uses depthwise separable convolutions to dramatically reduce computation without sacrificing accuracy [16]. Its successor, MobileNetV2, introduced inverted residuals and linear bottlenecks to improve representational efficiency [17]. MobileNetV3 further refined the design through neural architecture search and squeeze-and-excitation layers. These additions improved efficiency while maintaining competitive performance across a range of tasks [18]. EfficientNet scales networks using a compound coefficient to balance depth, width, and resolution. This strategy often yields strong performance on small datasets, an important advantage for medical imaging applications where data availability is limited [15]. ConvNeXt, meanwhile, modernizes classic CNNs by adopting design choices from Vision Transformers, including large kernels and layer normalization, while retaining convolutional operations for computational efficiency [19]. Comparing these architectures under identical data preprocessing, splits, and training configurations offers insight into which design strategies best generalize to liver ultrasound. Recent liver ultrasound studies have largely focused on ResNet-, Inception-, and DenseNet-based backbones [7, 10, 13, 29, 30, 35]. As a result, efficient architectures such as MobileNet, EfficientNet, and ConvNeXt remain relatively underexplored in this application area, making their comparative evaluation particularly valuable.

Across the literature, evaluation practices vary significantly. Many papers report only accuracy, while others include AUROC, sensitivity, or specificity. For multi-class problems such as normal/benign/malignant classification, the recommended approach is to report class-wise precision, recall, and F1-scores, and to compute the multiclass AUROC using the generalized formulation proposed by Hand and Till [22]. Despite the importance of model explainability, only a minority of studies incorporate visualization techniques like Grad-CAM, which are essential for assessing model trustworthiness in clinical settings. Furthermore, only a subset of works report performance separately for clinically important subgroups (for example, lesion size, etiology, or scanner type), and relatively few include external test cohorts; this heterogeneity in evaluation design makes it difficult to compare models fairly across studies and strengthens the case for well-controlled benchmark experiments using shared datasets and standardized metrics [11, 20, 29, 32, 33].

Chapter 3: Dataset

The Annotated Ultrasound Liver Images Dataset was selected for this study and obtained from the Zenodo repository [21]. Originally released as part of an artificial intelligence pipeline for liver malignancy diagnosis, it remains one of the few publicly available liver ultrasound datasets developed specifically for machine learning research. The dataset contains 735 ultrasound images labeled as malignant, benign, or normal. The distribution includes 435 malignant cases, 200 benign cases, and 100 normal cases (Fig. 3.1). Although this class imbalance may appear challenging from a modeling perspective, it reflects realistic clinical environments where malignant cases are documented more frequently because of their urgency.

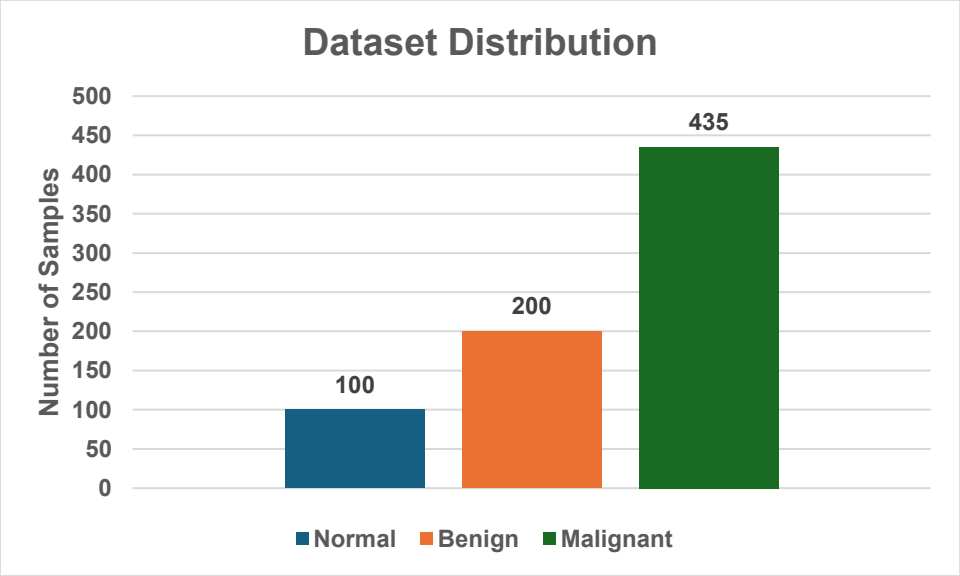


Figure 3.1: Distribution of ultrasound images across malignant, benign, and normal classes in the annotated liver dataset.

The images in the dataset come in standard formats such as JPEG and PNG, allowing for straightforward loading and processing using deep learning frameworks. The original resolution of the images is 1024 × 768 pixels, a size that retains fine-grained ultrasound features such as echotexture, lesion shape, and boundary characteristics. These structural details are essential for model training, as subtle differences in texture and contrast can be key indicators of liver disease. Ultrasound inherently contains speckle noise and operator-dependent variability that makes it more challenging to classify these images.

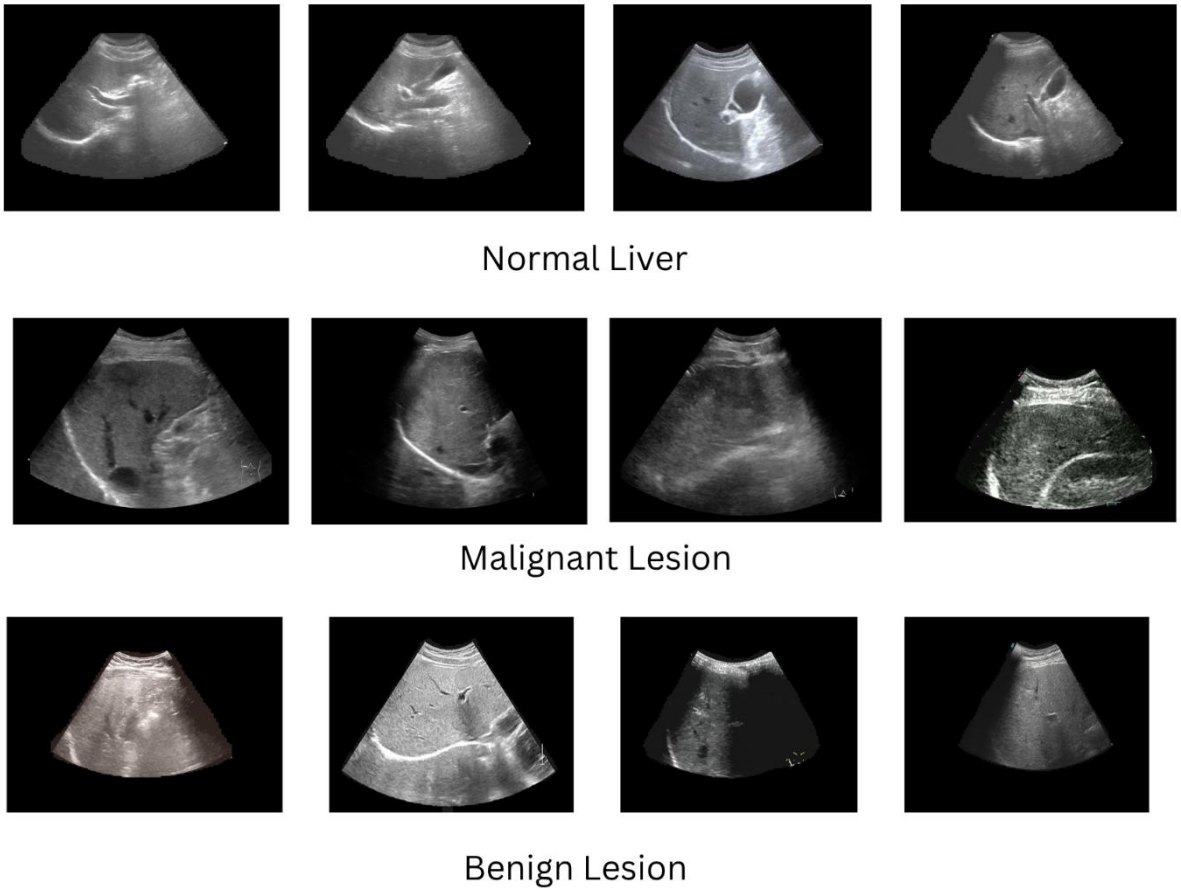


Figure 3.2: Example ultrasound images representing the three classes in the dataset: normal, benign, and malignant.

Chapter 4: Methodology

4.1 CNN Models

Deep learning has become a widely used approach in medical image analysis because it can automatically learn hierarchical patterns from data without relying on handcrafted features. Among deep learning methods, convolutional neural networks (CNNs) are the standard architecture for image classification. CNNs use convolutional filters to extract spatial and textural information, making them effective for processing ultrasound images, which often contain noise and subtle structural variations.

Three architecturally distinct CNN families MobileNet, ConvNeXt, and EfficientNet are evaluated for liver ultrasound classification into normal, benign, and malignant categories. MobileNet represents lightweight models that use depthwise-separable convolutions for efficiency. ConvNeXt modernizes traditional CNN designs through the incorporation of features inspired by transformer architectures. These design choices enhance scalability and representation quality. EfficientNet applies a compound scaling method that balances depth, width, and resolution to achieve high accuracy with optimized computation.

4.1.1 MobileNet Family

The MobileNet family was designed to solve a common problem in computer vision: how to keep accuracy high while making computation and memory small. That goal fits medical ultrasound very well. Many clinics rely on standard CPUs or modest GPUs, and they need fast, reliable results without heavy hardware. MobileNet meets this need by replacing standard convolutions with depthwise separable convolutions. Instead of filtering and mixing channels in one costly step, the model first applies a depthwise spatial filter to each input channel and then uses a 1×1 pointwise convolution to mix channels. This split reduces parameter count and multiply–accumulate operations while keeping feature quality competitive, so models train faster and run quickly at inference time [16]. MobileNet also has practical controls like width and resolution multipliers, which can dial performance up or down to match a given device or latency target without redesigning the network.

The family evolves across three versions we use in this thesis. MobileNet establishes the efficiency blueprint and proves that depthwise separable convolutions can deliver strong accuracy with a fraction of the cost of standard convolutions. MobileNetV2 rethinks the residual block to improve information flow. It uses inverted residuals and linear bottlenecks: expand channels, do lightweight spatial processing, then project back down through a linear layer. This keeps nonlinear operations in the high-dimensional space, preserves detail, and makes deeper but still efficient models easier to train [17]. MobileNetV3 then mixes neural architecture search with lightweight attention and mobile-friendly activations. Selected blocks include squeeze-and-excitation-style channel attention; the network also uses hard-swish and hard-sigmoid to speed inference while keeping accuracy high. V3 provides “Small” and “Large” options to match deployment constraints [18].

For our classification task (normal, benign, malignant) we resize all ultrasound images to 224 × 224, which is a natural fit for MobileNet-style backbones. This input size is common in transfer learning and lets us keep the compute budget predictable. The family's scaling knobs also help: if a site wants faster throughput, we can reduce width or resolution; if a site wants more accuracy, we can increase them. In our controlled setup same preprocessing (resize only), same training recipe, and the same shared classification head we compare MobileNet, MobileNetV2, and MobileNetV3Large to understand how each design choice translates to clinical images. Our experiments show that MobileNetV2 often gives the best balance of accuracy and speed on this dataset, while MobileNetV3Large offers more sophisticated blocks and attention but can be sensitive to small-data training dynamics. Overall, the MobileNet family provides a practical path to deploying AI support in ultrasound environments that cannot host large models, while still producing competitive results [16–18].

4.1.1.1 MobileNet

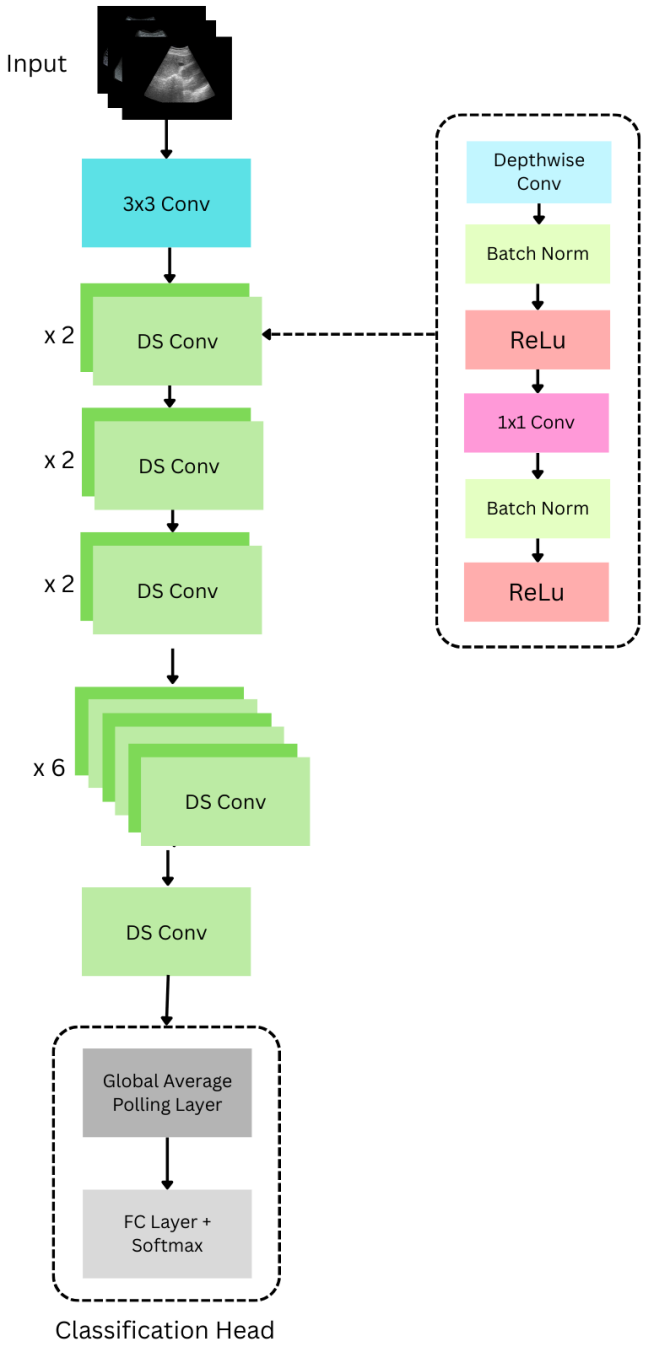


Figure 4.1: MobileNet Architecture

MobileNet introduced the key idea that made mobile-scale CNNs practical: depthwise separable convolutions (Fig 4.1). A standard 3×3 convolution filters across spatial dimensions and channel dimensions at the same time; this is accurate but heavy. MobileNet splits that work into two light steps. First, a depthwise 3×3 convolution filters each input channel separately. Second, a pointwise 1×1 convolution mixes those filtered channels to create new feature maps. This two-step pattern reduces both parameters and floating-point operations by

a large margin compared to a standard convolution with the same width. Blocks are kept simple with ReLU activations, and they are stacked to build up high-level features as the network goes deeper [16].

Another practical contribution to MobileNet is the use of width and resolution multipliers. The width multiplier (often written as α) uniformly scales channel counts in every layer. The resolution multiplier scales the input image size. These levers let us set a target latency or memory budget without redesigning the network from scratch. In our project, we use an input of 224 × 224, which MobileNet handles well, and we keep the backbone compact so we can place our shared classification head on top without exceeding memory limits. This setup makes iteration fast: we can train, evaluate, and adjust the learning schedule quickly, which matters when we work with a relatively small dataset and want to avoid overfitting.

In medical ultrasound, images are noisy and details can be subtle. MobileNet's efficiency does not automatically guarantee accuracy, but it enables more experiments and faster feedback, which often leads to better model tuning. The architecture learns useful spatial filters for echotexture, lesion edges, and coarse shape while keeping the compute budget modest. The trade-off is that MobileNet's representational capacity is limited compared with newer designs. When classes are very similar, the model can plateau. That is why later versions refine the block design to preserve more information. Still, as a baseline for liver ultrasound classification, MobileNet is valuable: it proves that a lightweight backbone can deliver meaningful accuracy, run on ordinary hardware, and support real-time or near-real-time use in clinics.

4.1.1.2 MobileNetV2

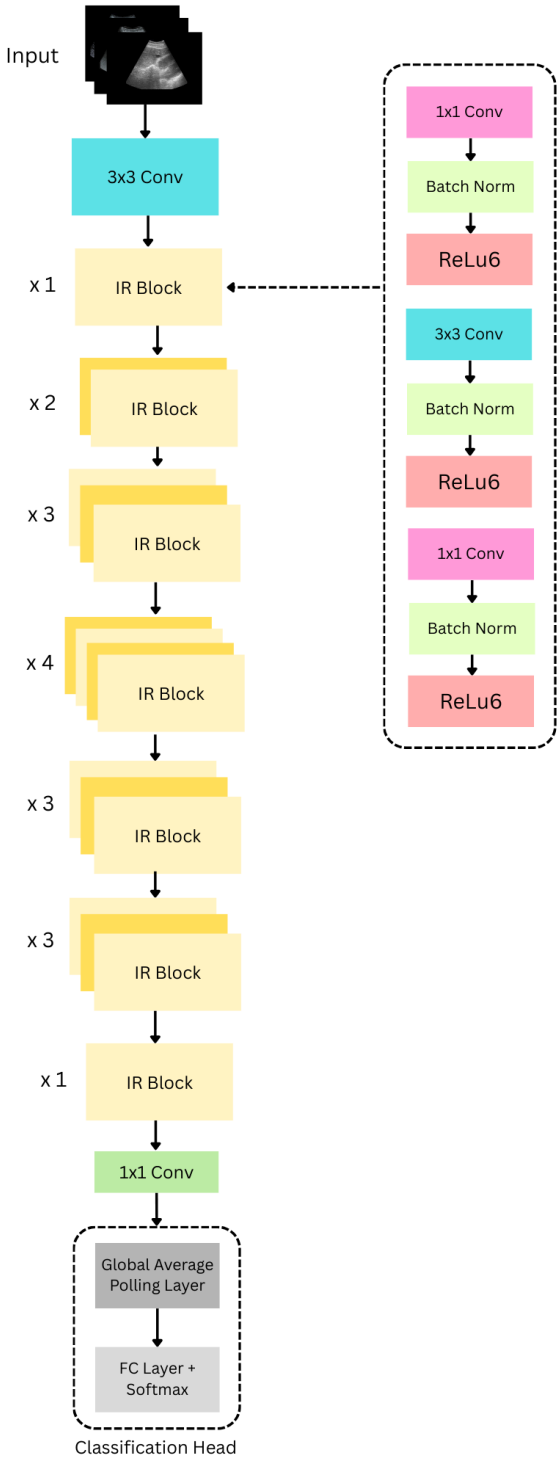


Figure 4.2: MobileNetV2 Architecture

MobileNetV2 raises the ceiling on what an efficient backbone can learn by redesigning the core block around inverted residuals and linear bottlenecks. In an inverted residual block, the network first expands channels with a 1×1 convolution, applies a depthwise 3×3 convolution to perform the spatial filtering, and then projects features back down with another 1×1 convolution to a linear (non-activated) low-dimensional space (Fig 4.2). When the stride is 1,

the input and output connect with a residual skip to help gradients flow; with stride 2, block down samples and omits the skip. The key idea is to keep most nonlinear operations in the expanded, high-dimensional space (where the model can express richer patterns) and to avoid squeezing features through a nonlinearity at the final bottleneck, which can destroy information. This simple blueprint improves both accuracy and stability at nearly the same compute as MobileNet [17].

For liver ultrasound, these choices matter. Ultrasound images include fine textures, low-contrast boundaries, and small lesions. If a model compresses too early or applies aggressive nonlinearities at narrow layers, it can lose the very signals we care about. MobileNetV2 avoids that by letting features evolve in the expanded space and by using a linear projection to preserve detail at the bottleneck. The blocks remain efficient thanks to depthwise separable convolutions, so the model stays fast at 224 × 224 inputs and fits easily into memory on standard hardware. In our study, this design translated into stable training and strong operating performance, especially on recall and F1-score for the malignant class, which is clinically important.

MobileNetV2 also scales smoothly. If we need lower latency, we can reduce the width; if we need higher accuracy, we can increase it or fine-tune for longer without changing the overall pipeline. This flexibility helps when we prepare models for different clinical settings. Compared to heavier backbones, V2 tends to overfit less on small datasets because it packs more effective feature learning into lean blocks rather than simply adding depth or width. In our controlled comparison same resize to 224 × 224, same training schedule, same classification head V2 consistently struck the best balance between speed and accuracy within the MobileNet family. It shows that thoughtful block design can outperform larger models when data are limited and deployment constraints are real [17].

4.1.1.3 MobileNetV3Large

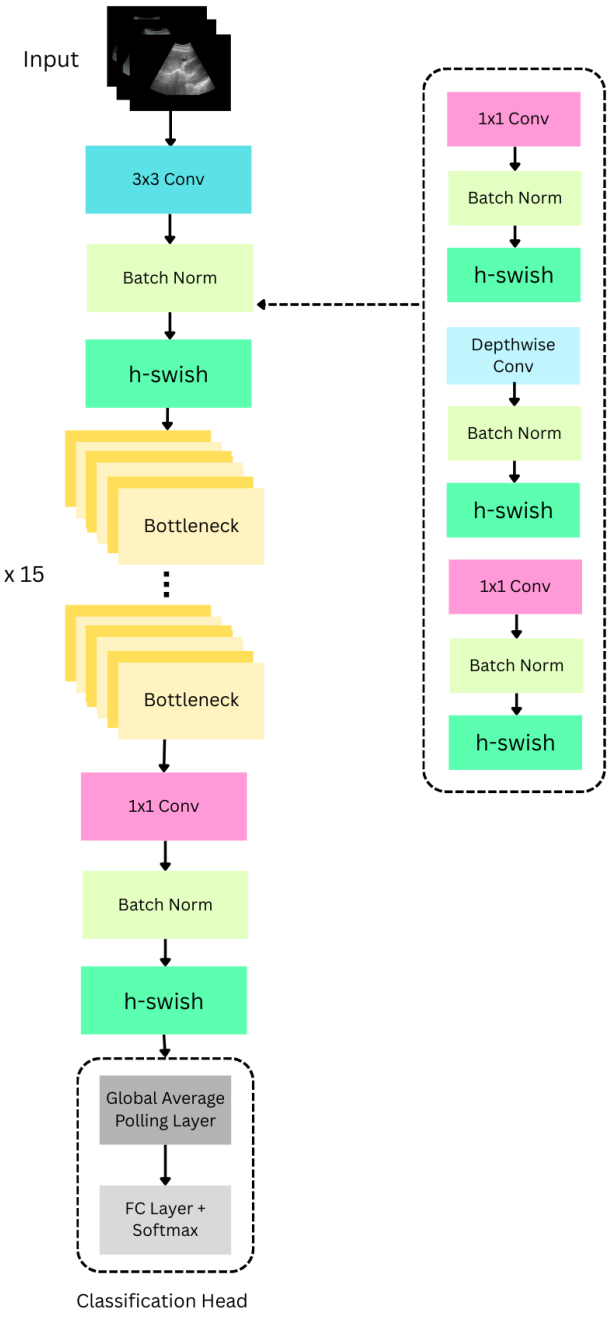


Figure 4.3: MobileNetV3 Large Architecture

MobileNetV3Large builds on the MobileNet/ MobileNetV2 foundations and uses neural architecture search (NAS) plus lightweight attention and fast activations to push the accuracy–efficiency frontier further. Selected bottleneck blocks include squeeze-and-excitation (SE) modules that compute a simple channel-wise attention signal from global context and reweight feature channels accordingly (Fig 4.3). This helps the network highlight informative channels (for example, those responding to lesion texture) and suppress less useful ones. V3Large also adopts hard-swish and hard-sigmoid activations, which approximate swish and sigmoid while

being cheaper to compute, making them well suited for CPU and mobile inference. The final architecture emerges from a search-and-refine process that considers real latency on target hardware, not just theoretical FLOPs, so it tends to deliver speed that users actually feel in practice [18].

On our ultrasound classification task, these features are attractive for two reasons. First, the attention mechanism can help when classes differ by subtle channel patterns, as often happens with benign versus malignant echotexture. Second, the hardware-aware design aims to keep latency low even as accuracy improves, which aligns with our goal of usable tools in clinics. We use a 224 × 224 input, which is native to MobileNet-style training, and we keep the same training recipe and shared classification head across models to ensure a fair comparison. In this setting, MobileNetV3Large did not always surpass V2 on our relatively small and imbalanced dataset. This likely reflects data sensitivity rather than a flaw in the architecture: V3Large's added capacity and attention may need more data, stronger regularization, or slightly different optimization to show clear gains. On small medical datasets, simpler information-preserving designs like V2 can sometimes generalize better.

Even so, V3Large remains promising for future expansion. If we grow the dataset, incorporate richer training strategies, or move toward real-time integration on ultrasound devices, the SE attention, NAS-tuned block layouts, and fast activations may pay off more clearly. The key contribution of V3Large is its deployment-first mindset: it treats latency and memory as core design targets alongside accuracy, which is essential for bedside tools. In summary, MobileNetV3Large offers a modern, attention-enhanced backbone that keeps MobileNet's efficiency mission intact, and it provides a solid alternative to V2 when conditions favor more capacity and hardware-aware optimization [18].

4.1.2 ConvNeXt Family

ConvNeXt is a "modernized CNN" that keeps the core strengths of convolution translation equivariance, efficient dense computation, and hardware friendliness while adopting design choices that helped Vision Transformers succeed. The key ideas include large depthwise convolution kernels to widen the receptive field, LayerNorm in place of BatchNorm, inverted-bottleneck style channel expansion within each block, and clean stage downsampling with dedicated transitions. Together, these choices help the network capture both local textures and broader context without leaving the convolutional paradigm [19]. For medical ultrasound, this is attractive: we want a backbone that is sensitive to fine echotexture and edges, yet still able to integrate regional cues that distinguish benign from malignant patterns.

Each ConvNeXt stage contains multiple identical ConvNeXt blocks. A block typically applies a depthwise convolution with a large kernel (commonly 7×7) to aggregate spatial context, then normalizes features with LayerNorm (channels-last format), then uses a pointwise expansion (roughly 4×) with a nonlinearity such as GELU, followed by a pointwise projection back to the base width, plus a residual connection. Between stages, the model downsamples with a strided convolution, increasing channels and reducing spatial size. Compared to classical

ResNets, ConvNeXt reduces activation count, simplifies the stem, and streamlines the block so most compute sits in a small set of well-optimized ops. The result is a CNN that trains stably, scales well, and reaches transformer-level accuracy on natural images all while staying fully convolutional and efficient on today's hardware [19].

For our 224 × 224 ultrasound images, ConvNeXt's design gives us a good balance between texture sensitivity and context capture. Because we use transfer learning, ImageNet pretraining initializes features that respond to edges, blobs, and frequency patterns, and fine-tuning adapts them to speckle-rich liver ultrasound. In this thesis we evaluate ConvNeXtSmall, ConvNeXtLarge, and ConvNeXtXLarge. These variants scale depth and width across stages to trade compute for capacity. Small emphasizes speed and efficiency; Large increases representational power; XLarge pushes capacity further and is best suited for larger datasets or stronger regularization. We attach the same classification head used for all backbones to keep the comparison fair, so any performance differences reflect the backbone rather than the head or optimizer. This family allows us to test whether "modern CNN" refinements can close the gap with heavier networks while remaining practical for clinical deployment [19].

4.1.2.1 ConvNeXtSmall

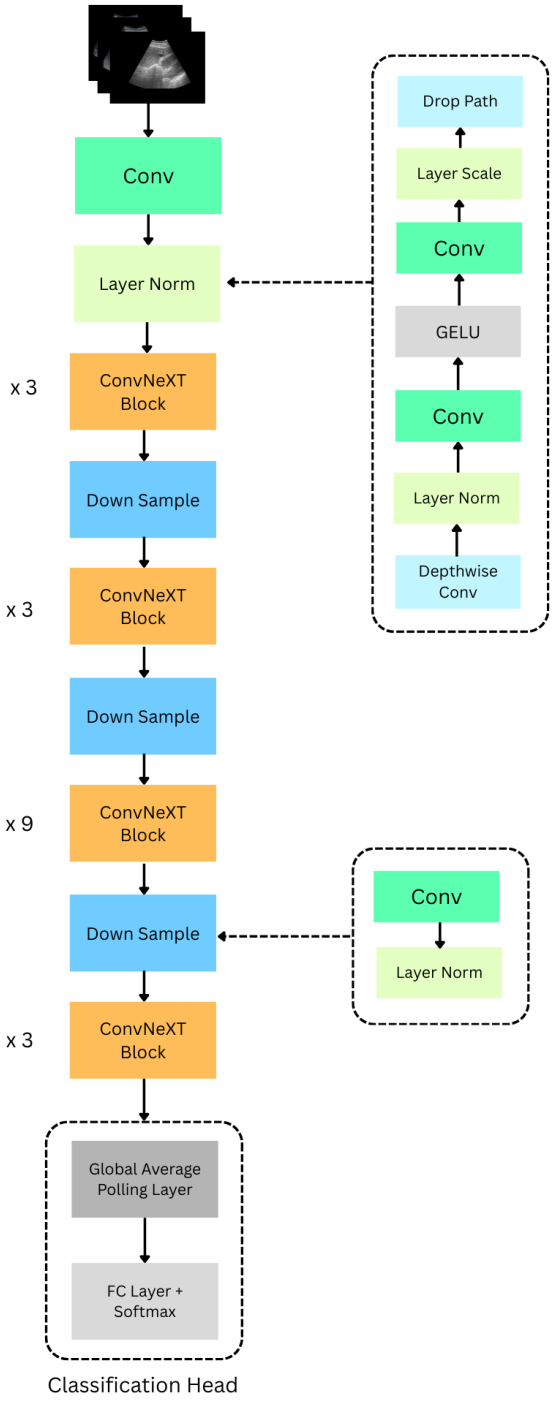


Figure 4.1: ConvNeXT Architecture

ConvNeXtSmall is the entry point into the family. It follows the same building blocks as the larger variants large-kernel depthwise convolution, LayerNorm, pointwise expansion and projection, and residual connections but it keeps the stage widths and depths modest to reduce parameters and inference time (Fig 4.5) [19]. The model begins with a streamlined stem that converts the input image into a lower-resolution feature map, then proceeds through four stages. Each stage increases channels and reduces spatial size through a strided downsampling

layer, while stacking several ConvNeXt blocks to grow semantic abstraction. Because most spatial mixing happens in the large depthwise kernels, ConvNeXtSmall still gathers substantial context at each block, despite its compact size.

For ultrasound liver classification at 224 × 224, this balance is useful. Ultrasound features are often subtle, but not every case needs heavy capacity. With transfer learning, the pretrained filters already capture generic edge and texture primitives; fine-tuning adapts them to echotexture, posterior shadowing, and lesion boundaries. ConvNeXtSmall's lower parameter count reduces overfitting risk on small datasets and allows faster iteration when refining the learning rate schedule or early-stopping rules. Its memory footprint is also friendly to mid-range GPUs and even high-end CPUs, which supports broader access in routine clinical environments.

In practice, we find ConvNeXtSmall trains stably under the same recipe we use for other backbones and integrates cleanly with our shared classification head. Where it may fall short is in highly complex cases, where malignant lesions present irregular textures or when benign lesions mimic malignant patterns. Those scenarios reward more capacity, which is why we also evaluate ConvNeXtLarge and ConvNeXtXLarge. Still, as a compact backbone, ConvNeXtSmall shows how the family's modern block large-kernel depthwise conv + LayerNorm + pointwise MLP-style expansion can deliver competitive performance without heavy compute. It is a strong baseline for real-time or near-real-time use, and a practical choice when hardware and dataset size argue for efficiency first [19].

4.1.2.2 ConvNeXtLarge

ConvNeXtLarge only scales width (channels) relative to the Small variant (Fig 4.5), increasing the model's capacity to learn richer, higher-level features from ultrasound images [19]. The underlying block remains the same 7×7 depthwise convolution, LayerNorm, pointwise expansion with a nonlinearity, pointwise projection, and a residual skip. However, by widening channels and stacking more blocks per stage, ConvNeXtLarge can represent more diverse textures, capture more complex boundaries, and integrate context across larger regions. The downsampling transitions between stages continue to raise channel count and reduce spatial size, allowing the deeper network to focus compute where it is most impactful.

On our 224 × 224 input, ConvNeXtLarge often shows advantages in cases where lesion appearance is heterogeneous or where benign and malignant patterns overlap in subtle ways. The extra capacity helps the model separate look-alike textures through a sequence of refined transformations. Transfer learning provides a strong starting point, but the deeper stack gives fine-tuning room to carve class-specific features that are sensitive to echotexture variations. In clinical terms, this can translate to better recall on challenging malignant cases while maintaining reasonable precision, provided the training is regularized to avoid overfitting.

The trade-off is compute and memory. ConvNeXtLarge requires more resources than Small, so batch sizes may need to be smaller on limited GPUs, and inference may be slower on CPU-only setups. For hospitals with capable hardware, that cost can be acceptable if it yields

higher diagnostic reliability. In our standardized pipeline same preprocessing and shared classification head ConvNeXtLarge represents the “capacity bump” within the family: it keeps the modern ConvNeXt block intact but scales it to better model complex ultrasound patterns. When datasets are modest, careful early stopping and learning-rate scheduling help the Large variant realize its benefits without memorizing noise. Overall, ConvNeXtLarge strikes a useful middle ground between practicality and representational power, making it a strong candidate when we need more than a compact model can offer [19].

4.1.2.3 ConvNeXtXLarge

ConvNeXtXLarge is the highest-capacity variant we evaluate. It extends the same block structure large-kernel depthwise conv, LayerNorm, pointwise expansion and projection, residual links, and clean downsampling to greater width, yielding the strongest representational power in the family [19]. With more channels per stage and more blocks stacked in sequence, XLarge can model highly complex textures and multi-scale cues that may appear in difficult liver ultrasound cases. The large depthwise kernels provide broad spatial context within each block, while the deeper stack supports hierarchical refinement from coarse patterns to fine lesion details.

For our 224 × 224 pipeline, ConvNeXtXLarge offers a test of the upper bound of capacity within a convolutional design. When the dataset is small and imbalanced, as in our setting, such capacity can be a double-edged sword. On one hand, it can capture rare or subtle malignant signatures that simpler models miss. On the other hand, without strong regularization or more data, it can overfit to scanner-specific noise or patient-specific features. To keep the comparison fair, we fine-tune XLarge under the same training recipe and attach the same shared classification head as for all other backbones. This allows us to attribute any performance differences to the backbone itself rather than to changes in optimization or head capacity.

In realistic clinical deployments, ConvNeXtXLarge demands more compute and memory than Small or Large, which can limit batch sizes and throughput on typical hospital hardware. For research and for high-resource centers, XLarge can serve as a reference point: it shows what the family can do when capacity is not the primary constraint. For broader deployment, we weigh its incremental accuracy against latency and cost. In our thesis results, we analyze whether the jump to XLarge yields consistent gains over Large and Small on this dataset, and we discuss where the diminishing returns begin. In short, ConvNeXtXLarge preserves the modern ConvNeXt principles but stretches them to maximum capacity, offering the richest features at the price of higher computational demand a trade-off that each clinical site must evaluate based on its hardware and workload [19].

4.1.3 EfficientNet Family

EfficientNet is a family of convolutional networks designed to reach strong accuracy with far fewer parameters and operations than older CNNs. The key idea is compound scaling: instead of scaling only depth or only width or only input resolution, EfficientNet scales all three together using a single coefficient. This balanced approach avoids the waste that comes from making a network deep without enough width, or wide without enough resolution. It gives a set of models (B0 to B7) that step up in capacity in a predictable way. Under the hood, EfficientNet builds on MBConv blocks (mobile inverted bottleneck convolutions) which include depthwise separable convolution, pointwise expansion and projection, and channel recalibration inside the block. These blocks keep the compute low while preserving representational power, so the network can learn useful filters without growing too large [15].

For medical ultrasound, EfficientNet's design is helpful for two reasons. First, ultrasound images include subtle textures and edges that need careful feature extraction, but many clinics cannot afford heavy models. EfficientNet offers a good middle path: it tends to achieve solid accuracy with a moderate number of parameters. Second, because the family scales up (B1 → B2 → B6), we can study how added capacity affects performance on a small, three-class dataset. That is important here, because very large models can overfit when data is limited. By selecting B1, B2, and B6, we cover a small, medium, and larger option that share the same building blocks but differ in depth/width/resolution scaling.

In our pipeline, we resize all images to 224 × 224 to keep training uniform across backbones. Although some EfficientNet variants are often trained at higher resolutions, keeping 224 × 224 for all models ensures a fair comparison. We then attach the same classification head (two dense layers followed by a three-way Softmax) so differences in results reflect the backbone, not the head. In short, EfficientNet lets us test whether balanced scaling and efficient blocks can deliver reliable liver ultrasound classification without the cost of much larger architectures.

4.1.3.1 EfficientNetB1

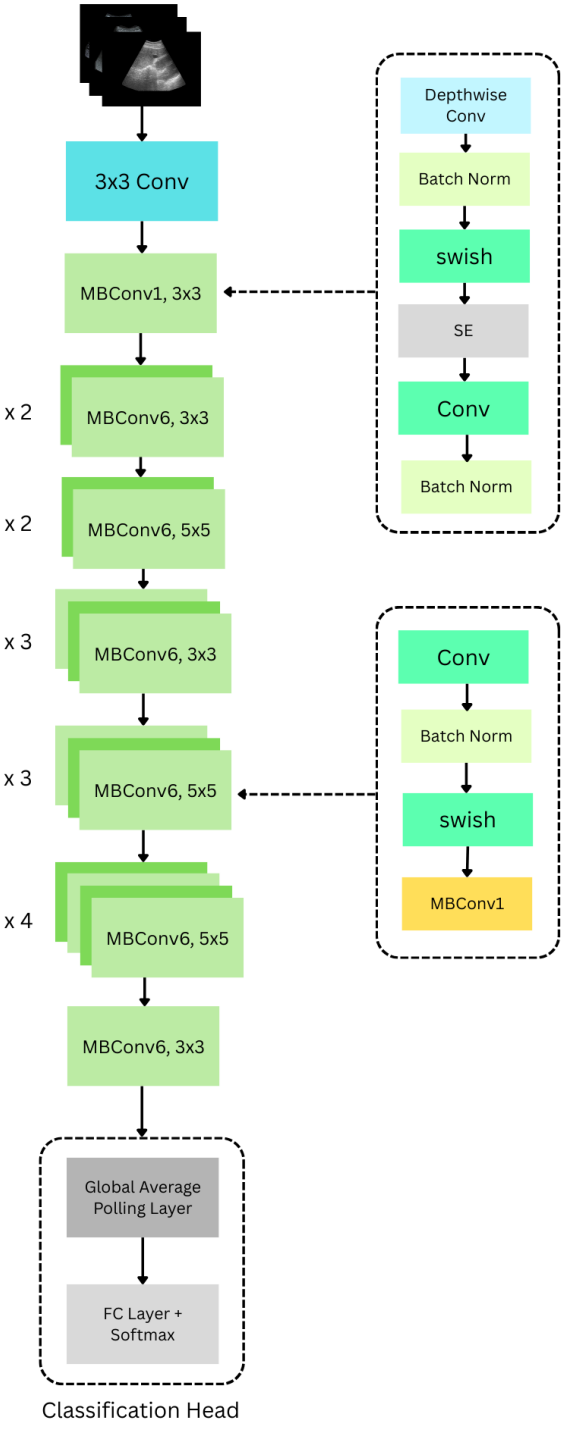


Figure 4.2 EfficientNetB1 Architecture

EfficientNetB1 represents the lightweight end of the family options we use. It keeps the same MBConv building blocks and compound scaling rules as the larger models but uses fewer channels and layers. This keeps computation and memory low while giving enough capacity to learn meaningful features from ultrasound. Each MBConv block first expands channels, applies a depthwise spatial convolution to capture local patterns, and then projects back to a

narrower representation. The block also includes channel reweighting and a residual connection in compatible cases, which improves gradient flow. Because these blocks are efficient, B1 builds a reasonably deep network without a large parameter count, making it a good starting point for transfer learning on medical images [15].

On our 224 × 224 inputs, B1 fine-tunes quickly and trains stably. It is well suited to small datasets, where simpler models can generalize better. Ultrasound classification often rewards careful texture modeling rather than brute-force capacity, and B1's compact structure keeps the model focused. In real clinical settings, a backbone like B1 also helps with latency and memory. It can run on entry-level GPUs and even some CPUs with acceptable speed, which broadens the number of sites that can deploy the model. The trade-off is that B1 may not capture the most complex, heterogeneous lesion patterns as consistently as a larger model. If benign and malignant textures overlap or if boundaries are irregular and faint, a bigger network can help. That is why we also evaluate B2 and B6: to see where the gains from capacity begin to matter.

In our standardized setup same input size and shared classification head EfficientNetB1 sets a baseline for the family. If it performs close to the larger variants, that suggests the task does not require much capacity. If it falls behind, it tells us that this dataset benefits from deeper or wider models. Either way, B1 shows the core EfficientNet idea in action: a compact network, built from efficient blocks and balanced scaling, can deliver competitive accuracy for liver ultrasound at low cost [15].

4.1.3.2 EfficientNetB2

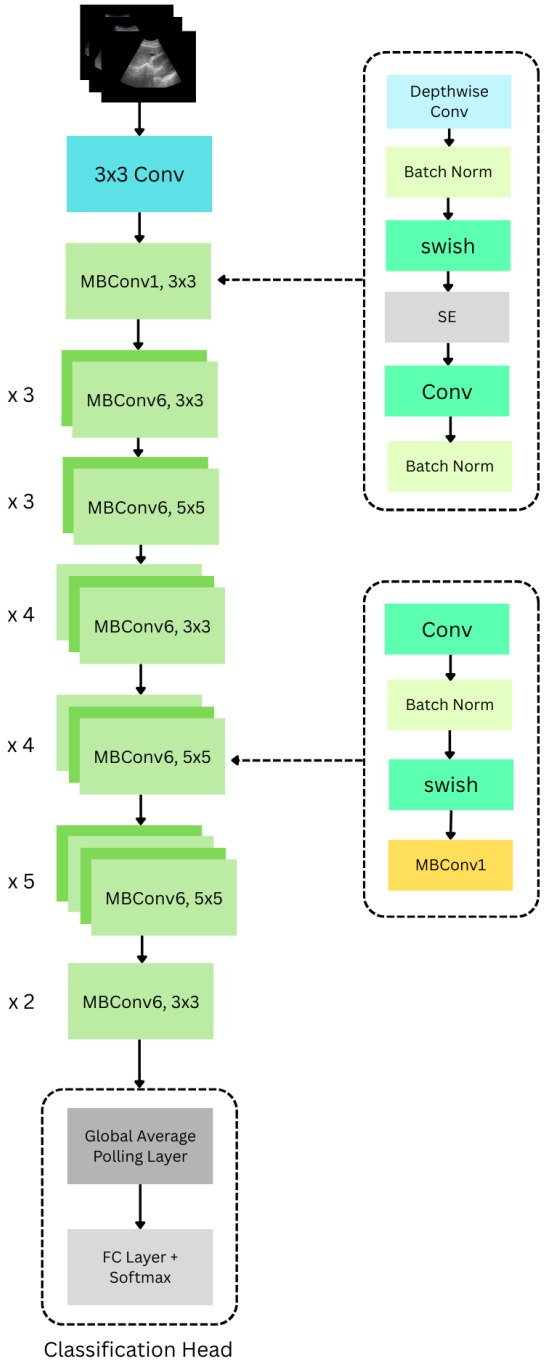


Figure 4.3: EfficientNetB2 Architecture

EfficientNetB2 moves one step up the ladder. It applies the same architecture and scaling rules as B1 but adds a little more depth and width (and in typical setups, resolution) to improve feature quality. The aim is to capture texture differences and boundary cues that a smaller network might miss, while staying close to the efficiency profile that makes EfficientNet attractive in the first place. Because the building blocks are the same MBConv with depthwise

spatial filtering, expansion, projection, and channel reweighting the training dynamics remain familiar, but B2 has more room to form class-specific patterns [15].

In our pipeline, we still use 224 × 224 inputs for fairness across backbones. Even at that resolution, B2 benefits from its extra capacity: it can model subtle echotexture, combine cues from neighboring regions more effectively, and represent irregular lesion edges with more stability. In practice, this often shows up as gains in recall and F1-score for malignant cases, where missing a lesion is costly. B2 remains manageable in compute and memory compared to much larger networks, so it is a practical choice when a clinic can allow a modest bump in hardware demand for a meaningful bump in accuracy.

From a training standpoint, B2 tends to be forgiving on small datasets: it offers a capacity increase without the fragility that sometimes appears in very deep models. Early stopping and careful learning-rate scheduling are still important, but the model usually converges smoothly. In our uniform comparison same preprocessing and shared classification head B2 lets us test whether moderate scaling is the sweet spot for liver ultrasound. If B2 consistently outperforms B1 and narrows the gap with much larger backbones, it strengthens the case for balanced, efficient scaling as a practical strategy in medical imaging settings [15].

4.1.3.3 EfficientNetB6

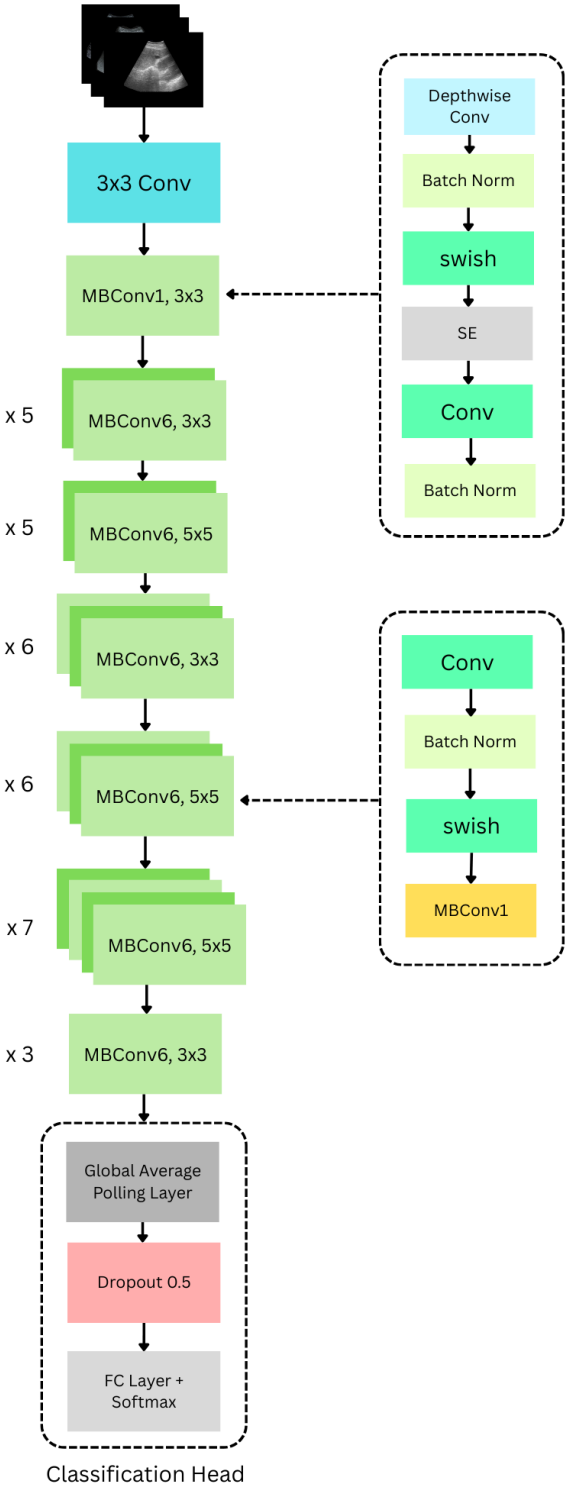


Figure 4.4: EfficientNetB6 Architecture

EfficientNetB6 represents a larger step in the family, scaling depth and width more aggressively to deliver higher representational capacity. The architecture remains the same at the block level MBConv with depthwise separable convolution, expansion, projection, and channel reweighting but the model stacks more blocks and wider stages, which allows it to learn complex, multi-scale features. On large, diverse datasets, this extra capacity can translate

to strong accuracy gains. However, when datasets are small or imbalanced, as is common in medical imaging, large models can overfit if training is not carefully controlled [15].

We keep 224 × 224 inputs and attach the same classification head to B6 as we do for all other backbones. In this controlled setting, B6 tests how much performance we can gain by scaling EfficientNet beyond the compact B1/B2 range. It can capture heterogeneous lesions and fine boundary details better, and it can integrate broader context across the field of view. The tradeoffs are increased compute, memory, and potentially a higher risk of overfitting without stronger regularization or more aggressive data strategies. In clinics with powerful GPUs, the cost may be acceptable; in low-resource settings, latency and throughput can become concerns.

In our comparison, B6 helps map the diminishing returns curve. If B6 clearly outperforms B2 and B1, the dataset benefits from higher capacity. If the margin is small, it suggests the efficient, moderate variants already capture the critical patterns in liver ultrasound. That insight matters for real deployments, where the choice of backbone affects both hardware cost and clinical workflow. Overall, EfficientNetB6 shows the upper bound of what balanced scaling can offer in this family for our task: richer features and potentially higher accuracy, offset by heavier requirements and tighter training margins on small medical datasets [15].

4.2 Training Setup

We trained all backbones under a single, consistent transfer-learning pipeline so that any performance differences come from the feature extractor rather than the training configuration. For each run, we load a pre-trained CNN backbone and freeze every backbone layer, turning it into a fixed feature extractor. On top, we attach the same classification head for all models: Two fully connected layers with 1024 and 512 units (both with ReLU), and a final 3-unit Softmax layer for the three classes (normal, benign, malignant). Keeping the head identical ensures a fair backbone comparison.

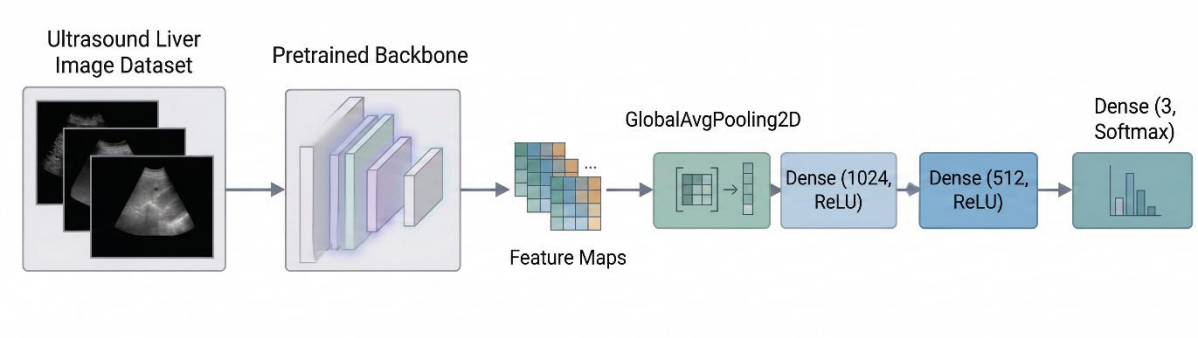


Figure 4.5: Model architecture used across all experiments.

All ultrasound images are resized to 224 × 224 and fed directly to the models. We fix the input resolution, so each model processes a similar amount of image context at each stage, making the comparison fair. We set the batch size to 16, which fits comfortably on modest hardware and keeps iteration quick enough for frequent model checkpoints and monitoring.

We train for up to 50 epochs using Adam with a learning rate of 0.0001. To improve convergence and prevent overfitting on this relatively small dataset, we rely on three callbacks that all monitor validation loss. ModelCheckpoint saves the best model (full weights) when validation loss reaches a new minimum; EarlyStopping has a patience of 10 epochs and restores the best weights when stopping is triggered; and ReduceLROnPlateau decreases the learning rate after 5 stagnant epochs to help the optimizer settle into a better minimum. A lightweight learning-rate logger prints the effective rate at the end of each epoch so we can verify that the scheduler is working as intended.

Table 4.1: Hyperparameter Summary

Hyperparameter Information	
Epoch	50
Loss Function	Cross Entropy
Learning Rate	0.0001
Optimizer	Adam
LR Scheduler	ReduceLROnPlateau
Batch Size	16

We split the dataset into 80% training, 10% validation, and 10% testing and keep this split fixed for all experiments. We use a stratified split so that the class distribution is consistent across subsets, which helps the model see representative samples during learning and provides a reliable validation signal during training. This is particularly important for imbalanced datasets, as it prevents any subset from becoming overly dominated by a single class.

4.3 Evaluation Metrics

To measure how well each model classified liver ultrasound images we used a combination of standard quantitative metrics. These metrics help us understand how accurate the predictions are overall and how the model works in more sensitive areas, such as detecting difficult or minority cases. All metrics are computed from four outcomes: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Together, these values describe how often the model is right or wrong for each class. In this context, true positives are correctly identified cases, false positives are incorrect positive predictions, false negatives are missed cases, and true negatives are correctly identified non-cases.

The core metrics we used are Accuracy, Precision, Recall, and F1-Score, because they provide a balanced view of performance. Accuracy measures overall correctness. Precision tells us how often the model's positive predictions are correct. Precision is important as false alarms can cause unnecessary follow up tests. Recall measures how well, the model predicts actual positive cases. Recall is especially important in medical imaging as missing a malignant case can be dangerous. The F1-Score combines Precision and Recall into one value. It evaluates performance when the dataset is imbalanced, as is common in medical imaging.

The formulas used for these metrics are:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

These metrics let us examine different aspects of model performance beyond just correctness. For example, a model may have high Accuracy but low Recall if it frequently misses malignant cases. In a medical setting, this would be unacceptable. Therefore, Recall and the F1-score better represent the model's reliability.

We also used the Area Under the Receiver Operating Characteristic Curve (AUROC) to see how well our model can tell apart the three liver conditions. Unlike Accuracy and F1-Score, which only looks at how well the model does at one specific threshold, AUROC gives us a better idea of how well the model can really distinguish between the different conditions. A higher AUROC indicates that the model is better at ranking images so that malignant, benign, and normal cases are separated effectively. This is useful because different hospitals or diagnostic tools may use slightly different confidence thresholds.

By combining these metrics, we gain a comprehensive understanding of how each model behaves, where it performs well, and where it might struggle. This multi-metric evaluation ensures that the comparison between MobileNet, EfficientNet, and ConvNeXt is fair, balanced, and aligned with real-world clinical needs.

Chapter 5: Results and Discussion

5.1 Overview of experiments

This chapter presents the experimental results for all models used in this thesis and discusses how well they classify liver ultrasound images into normal, benign, and malignant categories. We evaluate three convolutional neural network (CNN) families MobileNet, ConvNeXt, and EfficientNet using multiple variants from each family. All models share the same training pipeline: frozen pre-trained backbone, a common classification head, resized inputs of 224 × 224, and identical optimization settings. This controlled setup ensures that any performance differences mainly reflect the backbone architecture, not changes in preprocessing or training strategy. For each family, we report the main quantitative metrics (Precision, Recall, F1-Score, AUROC, and Accuracy) in a dedicated table and illustrate model behaviour with training curves, confusion matrices, and ROC curves.

5.2 MobileNet Results

This section focuses on the performance of the MobileNet family: MobileNet, MobileNetV2, and MobileNetV3Large. All three models use the same dataset split, the same 224 × 224 input size, and the same classification head. This makes it possible to directly compare how architectural changes inside the backbone affect classification performance.

5.2.1 Performance Summary

Table 5.1: Performance metrics for MobileNet, MobileNetV2, and MobileNetV3Large on liver ultrasound classification.

Models	Precision	Recall	F1 Score	AUROC	Accuracy
MobileNet	0.7008	0.7121	0.7064	0.8997	0.7837
MobileNetV2	0.7256	0.7803	0.7520	0.8853	0.7837
MobileNetV3Large	0.6078	0.5879	0.5977	0.8913	0.7297

Table 5.1 summarizes the performance of the MobileNet variants. Both MobileNet and MobileNetV2 achieved the highest Accuracy in the family, with values of 0.7837. However, MobileNetV2 clearly outperformed the others in terms of Recall (0.7803) and F1-Score (0.7520). MobileNet obtained a Precision of 0.7008, Recall of 0.7121, F1-Score of 0.7064, and the highest AUROC (0.8997) among the three. In contrast, MobileNetV3Large showed weaker threshold-based performance, with Precision 0.6078, Recall 0.5879, F1-Score 0.5977, Accuracy 0.7297, and AUROC 0.8913.

These numbers suggest that MobileNetV2 is the most balanced model in this family. It achieves the best combination of Recall and F1-Score while keeping Accuracy high. MobileNet is still competitive and ranks well in terms of AUROC, but it does not convert that ranking ability into as many correct positive predictions as V2 does. MobileNetV3Large, despite having a respectable AUROC, struggles more with correct classification at the chosen decision threshold.

5.2.2 Training and Validation Curves

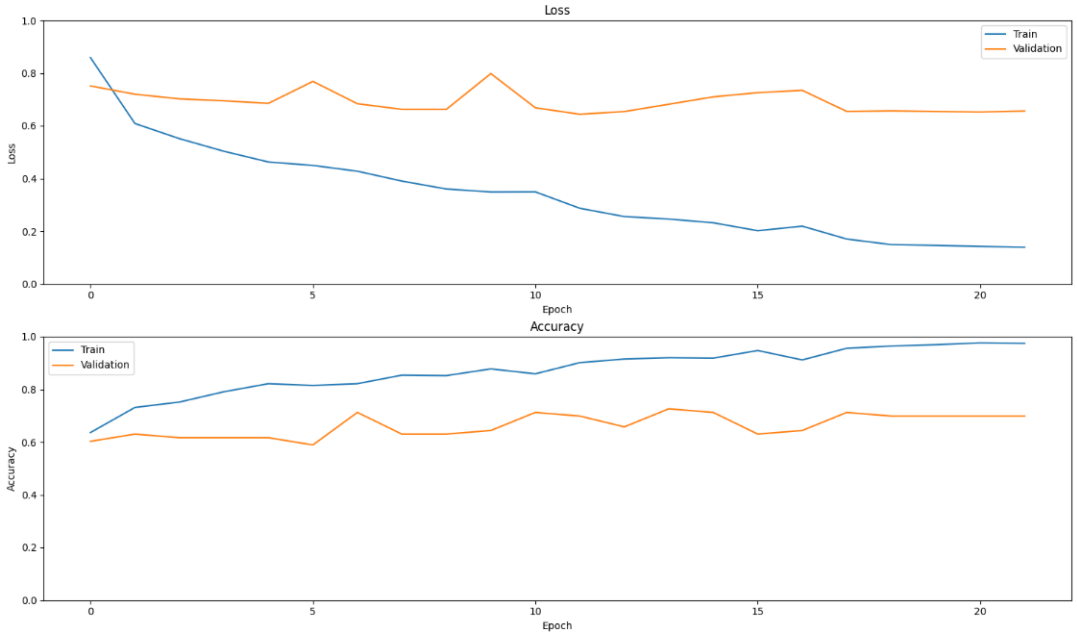


Figure 5.1: Training & Validation Curve of MobileNet

MobileNet learns the training data steadily, as shown by the continuous drop in training loss and the rise in training accuracy over epochs (Fig 5.1). However, the validation loss remains unstable and considerably higher, while validation accuracy improves only slightly and then plateaus. This pattern indicates that MobileNet fits the training set well but shows limited generalization to unseen ultrasound images.

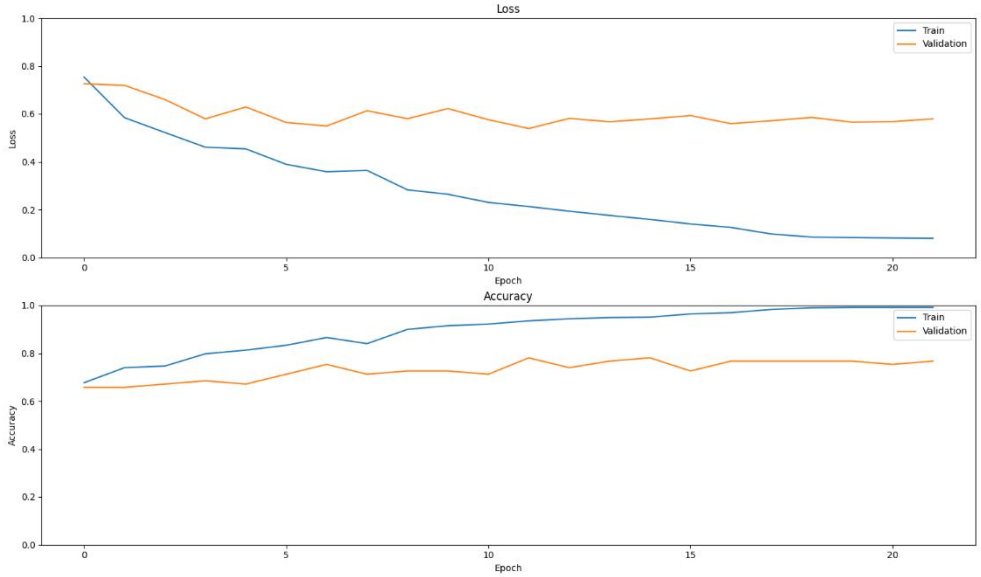


Figure 5.2: Training & Validation Curve of MobileNetV2

MobileNetV2 demonstrates smooth and stable learning, with training loss decreasing consistently and validation loss remaining relatively controlled (Fig 5.2). Validation accuracy improves steadily and stays close to training accuracy, indicating better generalization compared to MobileNet. This behavior suggests that MobileNetV2 achieves a strong balance between learning capacity and robustness on unseen data.

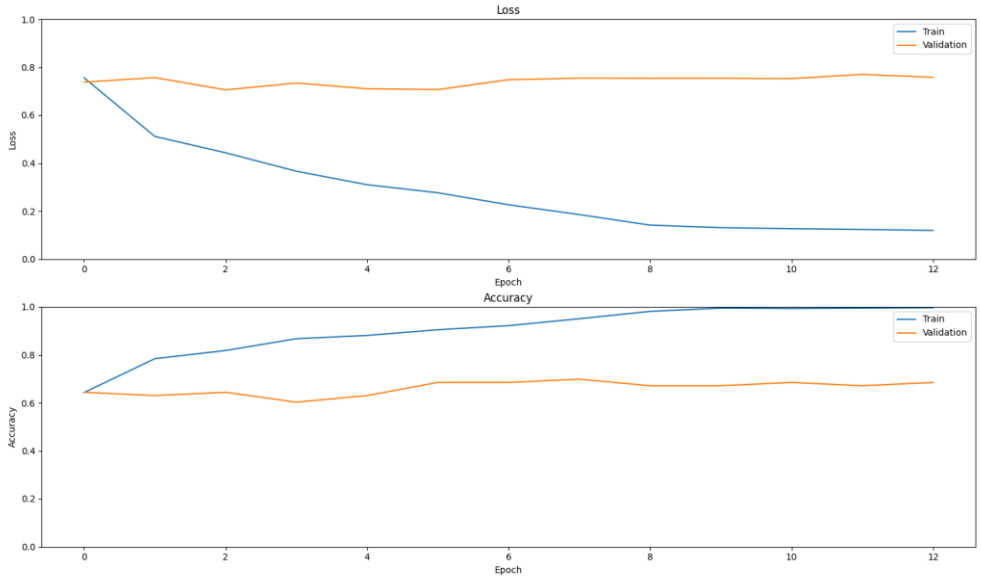


Figure 5.3: Training & Validation Curve of MobileNetV3Large

MobileNetV3Large quickly fits the training data, as shown by the sharp decrease in training loss and rapid increase in training accuracy (Fig 5.3). However, the validation loss remains almost flat and high, while validation accuracy shows minimal improvement. This behavior indicates clear overfitting, where the model fails to generalize effectively to unseen data.

Overall, the training curves support the idea that MobileNetV2 offers the best balance between learnability and generalization in this family. The model learns quickly, stabilizes well, and does not show signs of severe overfitting under the chosen training configuration.

5.2.3 Confusion Matrices

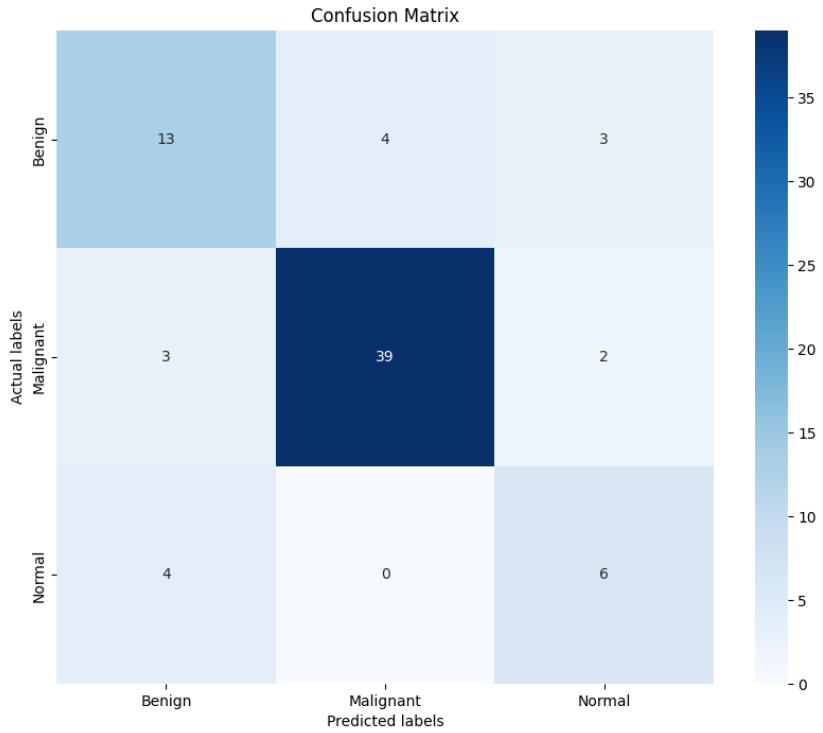


Figure 5.4: Confusion Matrix of MobileNet

MobileNet classifies malignant cases most effectively, with the highest number of correct predictions in that class (Fig 5.4). However, it shows noticeable confusion between benign and malignant samples, and it also misclassifies several normal cases as benign. This pattern suggests that while the model detects major abnormalities reasonably well, it struggles to separate finer class-specific features consistently.

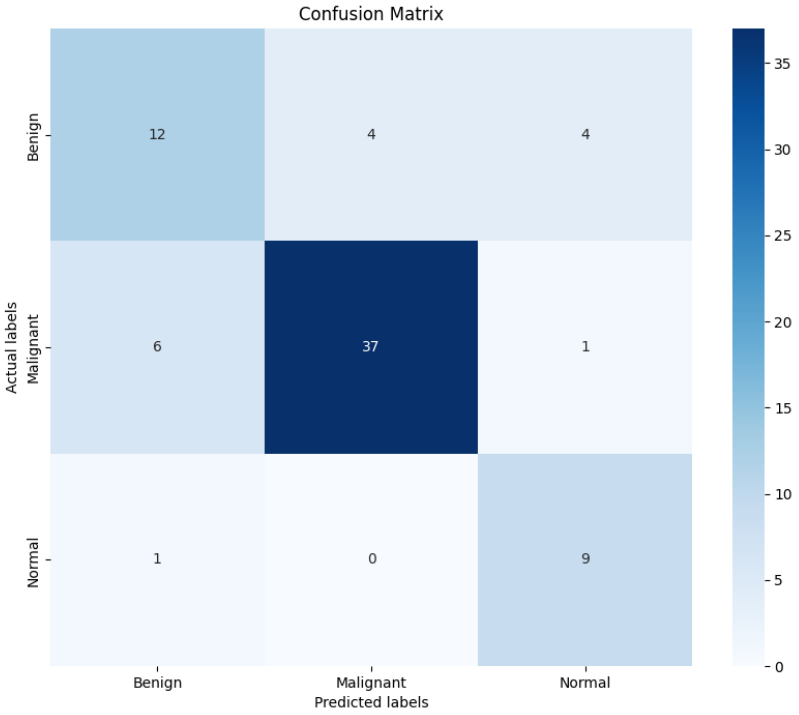


Figure 5.5: Confusion Matrix of MobileNetV2

MobileNetV2 correctly identifies most malignant and normal cases, showing improved classification balance across classes (Fig 5.5). Misclassification between benign and malignant cases is reduced, although some confusion still remains. Overall, the model demonstrates stronger class discrimination and more consistent performance compared to MobileNet.

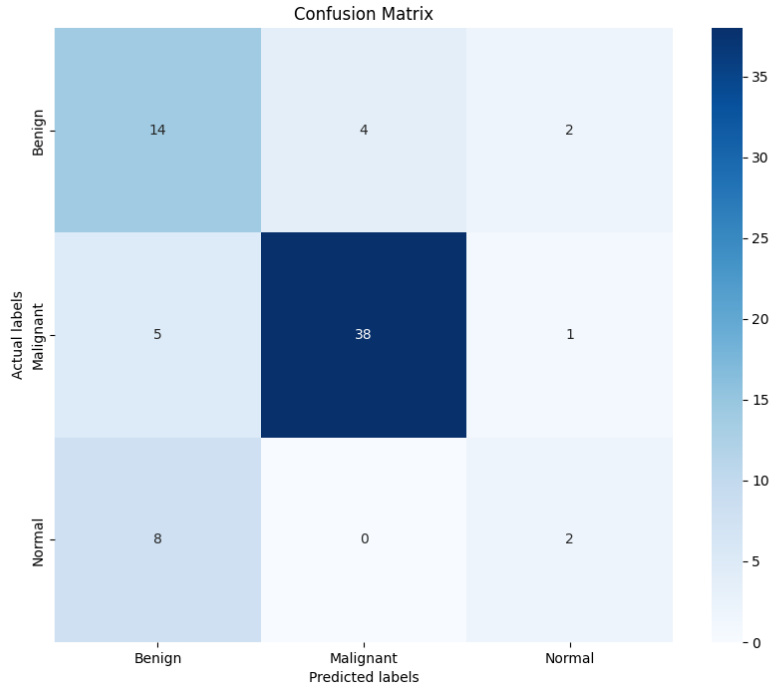


Figure 5.6: Confusion Matrix of MobileNetV3Large

MobileNetV3Large correctly detects a high number of malignant cases but shows increased misclassification in the normal class (Fig 5.6). Several normal samples are incorrectly labeled as benign, indicating weaker class separation. Overall, the model demonstrates less balanced performance, with noticeable confusion across non-malignant classes.

5.2.4 ROC Curves

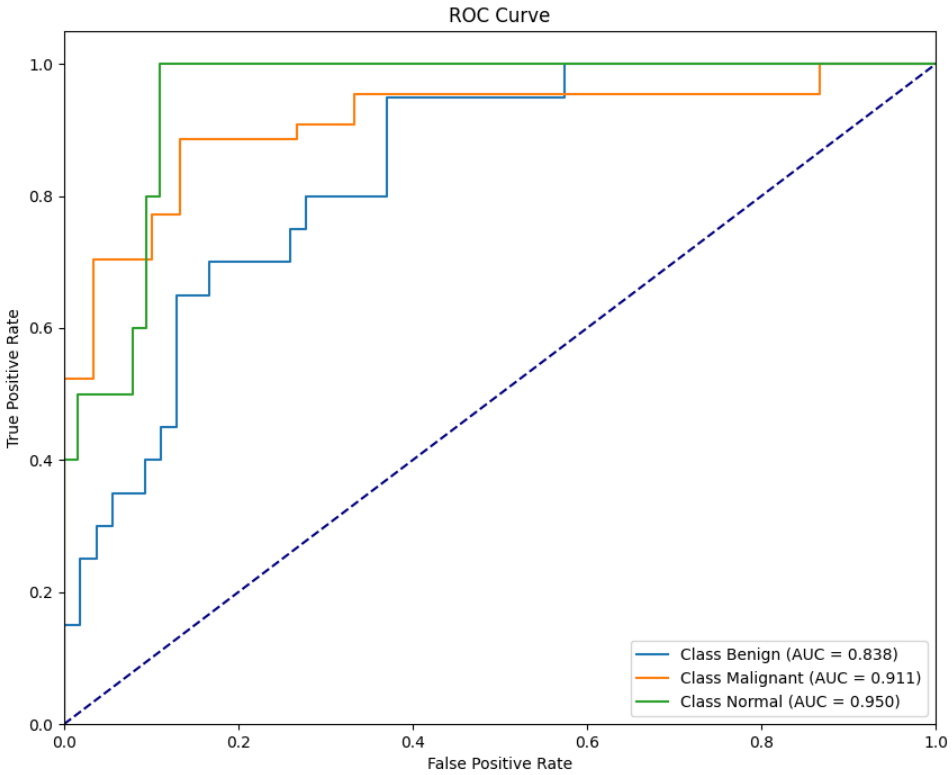


Figure 5.7: Receiver Operating Characteristic (ROC) curves for MobileNet

MobileNet demonstrates strong discriminative ability across all classes, with the highest AUC observed for the normal class (0.950) (Fig 5.7). The malignant class also shows high separability (AUC = 0.911), while the benign class performs comparatively lower. Overall, the model effectively ranks samples across thresholds, despite weaker performance in threshold-based classification.

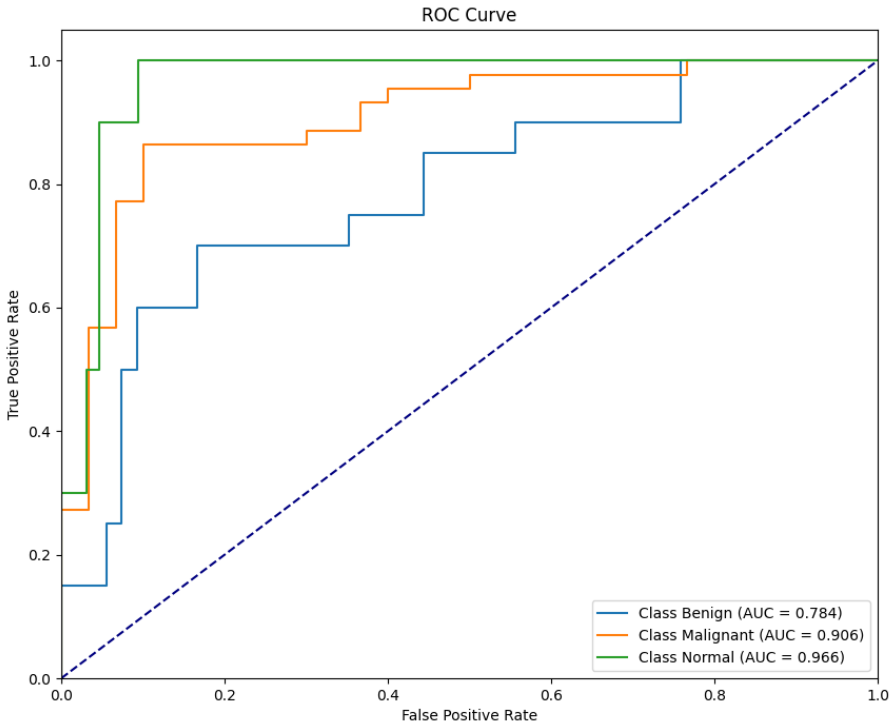


Figure 5.8: Receiver Operating Characteristic (ROC) curves for MobileNetV2

MobileNetV2 shows strong class separability, with the normal class achieving the highest AUC (0.966) and excellent discrimination (Fig 5.8). The malignant class also maintains high performance (AUC = 0.906), while the benign class shows comparatively lower separability. Overall, the model demonstrates reliable ranking ability across classes, supporting its strong classification performance.

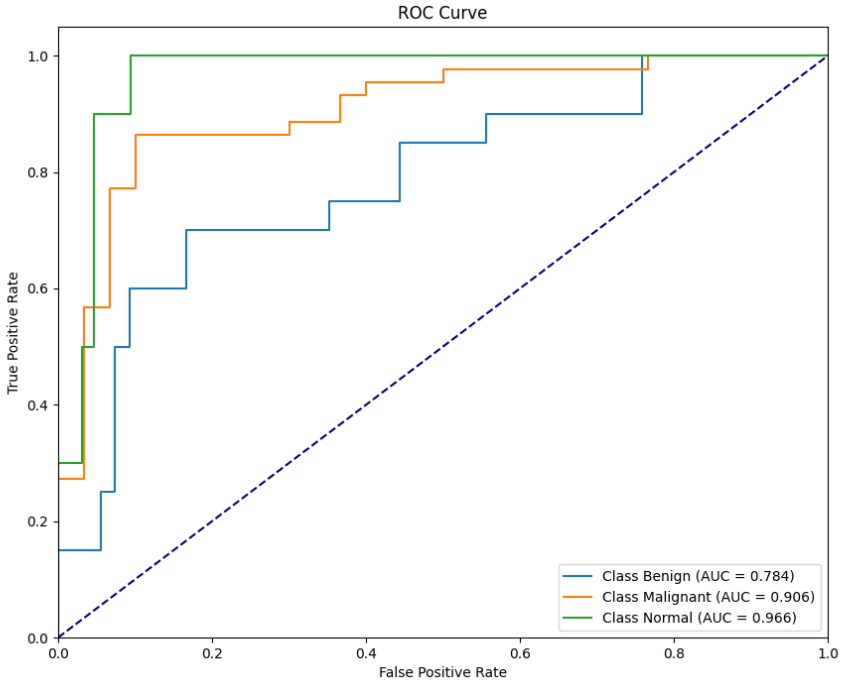


Figure 5.9: Receiver Operating Characteristic (ROC) curves for MobileNetV3Large

MobileNetV3Large shows strong discriminative performance for the normal class, achieving the highest AUC (0.966), while the malignant class also maintains high separability (AUC = 0.906) (Fig 5.9). However, the benign class continues to lag behind with lower AUC, indicating weaker distinction. Overall, the model demonstrates good ranking ability but less balanced performance across all classes

5.3 ConvNeXt Results

The ConvNeXt family represents a modern reinterpretation of convolutional networks, inspired by design elements borrowed from Vision Transformers but implemented using standard convolutional operations. In this study, we evaluated ConvNeXtSmall, ConvNeXtLarge, and ConvNeXtXLarge under the same training pipeline as the other backbone families. Although these models differ significantly in depth, width, and representational capacity, they share a common architectural foundation that emphasizes large kernel sizes, inverted bottlenecks, and improved normalization. This section presents their performance metrics, visual behaviors, and key observations regarding their strengths and limitations when classifying liver ultrasound images.

5.3.1 Performance Summary

Table 5.2: Performance metrics for ConvNeXtSmall, ConvNeXtLarge, and ConvNeXtXLarge.

Models	Precision	Recall	F1 Score	AUROC	Accuracy
ConvNeXtLarge	0.7222	0.7000	0.7109	0.8610	0.7297
ConvNeXtSmall	0.6818	0.6727	0.6773	0.8610	0.7297
ConvNeXtXLarge	0.7189	0.7061	0.7124	0.8693	0.7432

The performance metrics for the ConvNeXt family are shown in Table 5.2. ConvNeXtXLarge obtained the strongest results in this family, with Precision 0.7189, Recall 0.7061, F1-Score 0.7124, AUROC 0.8693, and Accuracy 0.7432. ConvNeXtLarge performed slightly below this with Precision 0.7222, Recall 0.7000, F1-Score 0.7109, AUROC 0.8610, and Accuracy 0.7297. ConvNeXtSmall, the lightest model in this family, delivered the weakest performance: F1-Score 0.6773, Accuracy 0.7297, and AUROC 0.8610.

Across all three, Accuracy values cluster around 0.73–0.74, and F1-Scores remain in a narrow band except for ConvNeXtSmall. This pattern suggests that increasing model size from Small → Large → XLarge yields only small gains, and those gains are not as large as one might expect given the dramatic difference in parameter count. ConvNeXtXLarge does rank highest, but its improvement is incremental rather than transformative.

When compared to MobileNetV2 (Accuracy 0.7837, F1-Score 0.7520), all ConvNeXt variants fall short. This indicates that model size and capacity alone do not guarantee superior performance, especially on smaller, medical imaging datasets where overfitting and limited texture diversity can constrain the usefulness of deeper architectures.

5.3.2 Training and Validation Curves

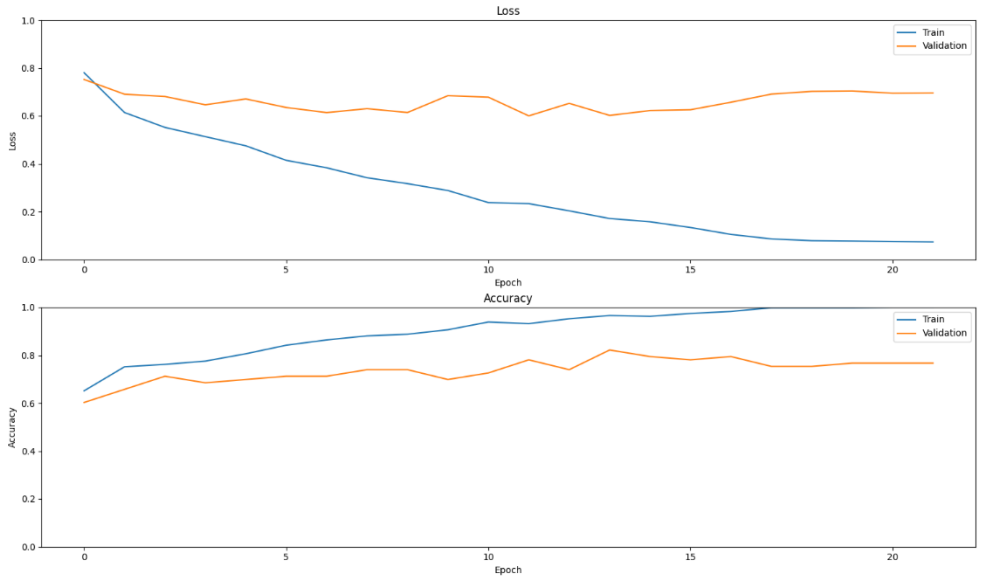


Figure 5.10: Training & Validation Curve for ConvNeXt Small

ConvNeXtSmall shows steady improvement in training performance, with decreasing loss and increasing accuracy over epochs (Fig 5.10). However, the validation loss fluctuates and trends upward slightly, while validation accuracy plateaus after early gains. This pattern suggests moderate overfitting and limited generalization to unseen data.

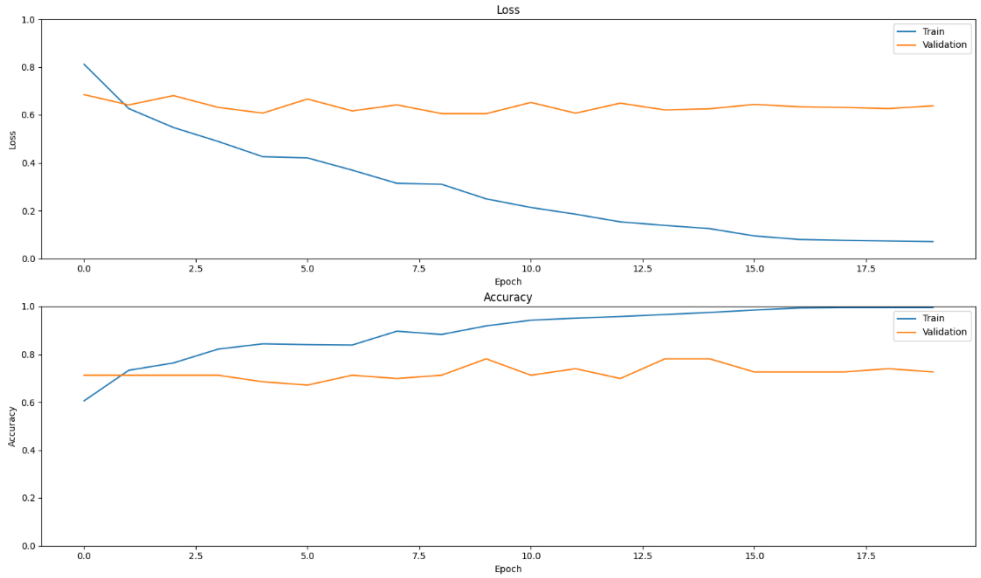


Figure 5.11: Training & Validation Curve for ConvNeXt Large

ConvNeXtLarge rapidly improves training performance, with a sharp decrease in training loss and near-perfect training accuracy (Fig 5.11). However, validation loss remains relatively flat and validation accuracy fluctuates without clear improvement. This indicates overfitting, where the model fails to generalize despite strong learning on the training data.

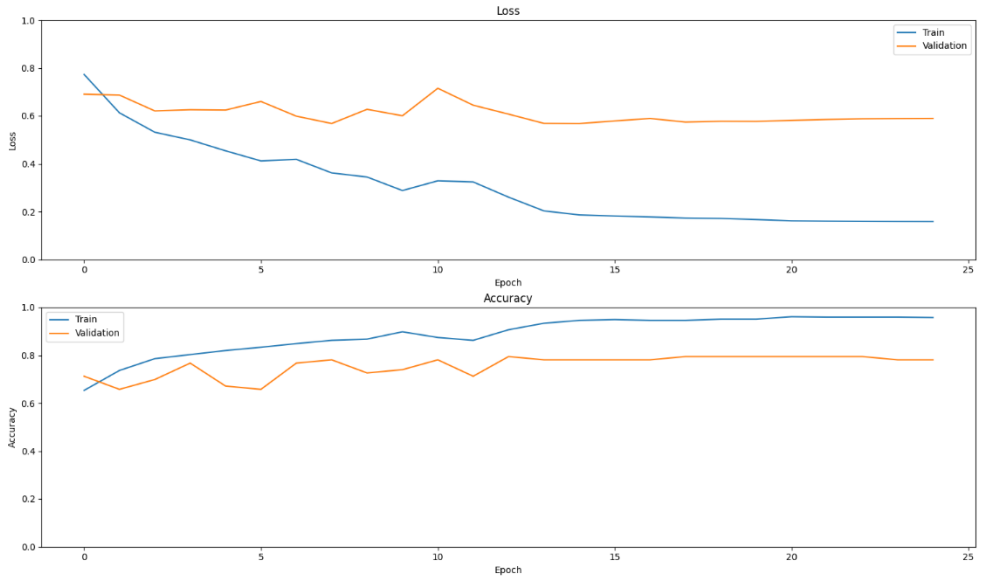


Figure 5.12:Training & Validation Curve for ConvNeXtXLarge

ConvNeXtXLarge shows strong training performance with steadily decreasing loss and high training accuracy (Fig 5.12). However, validation loss remains relatively unstable and higher, while validation accuracy plateaus after initial improvement. This indicates overfitting, with the model struggling to generalize despite its high capacity.

5.3.3 Confusion Matrices

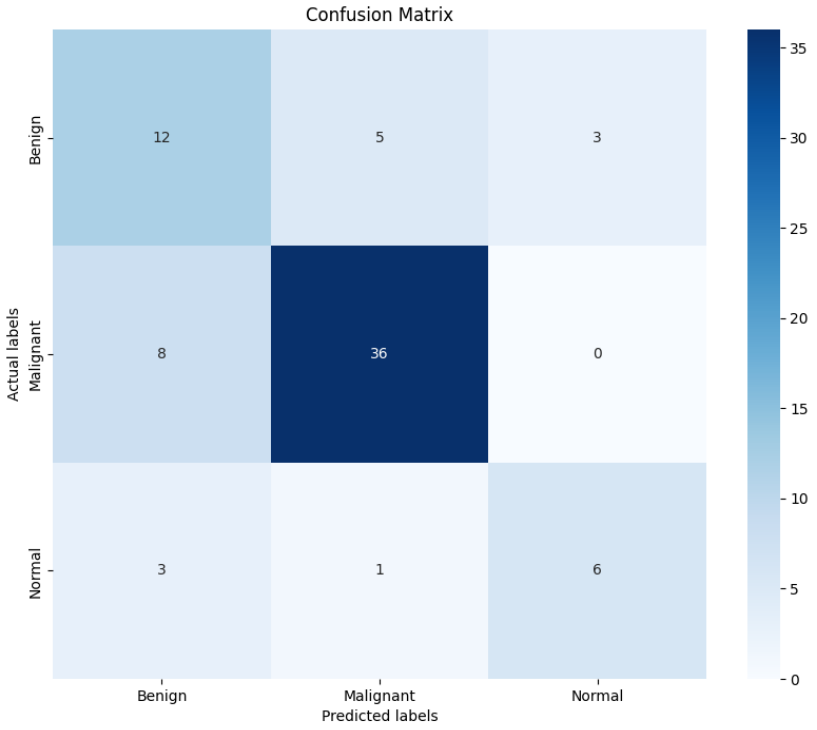


Figure 5.13: Confusion matrices for ConvNeXt Small

ConvNeXtSmall correctly identifies most malignant cases but shows noticeable confusion between benign and malignant samples (Fig 5.13). It also misclassifies some normal cases as benign, indicating weaker separation of non-malignant classes. Overall, the model demonstrates moderate performance with limited class discrimination.

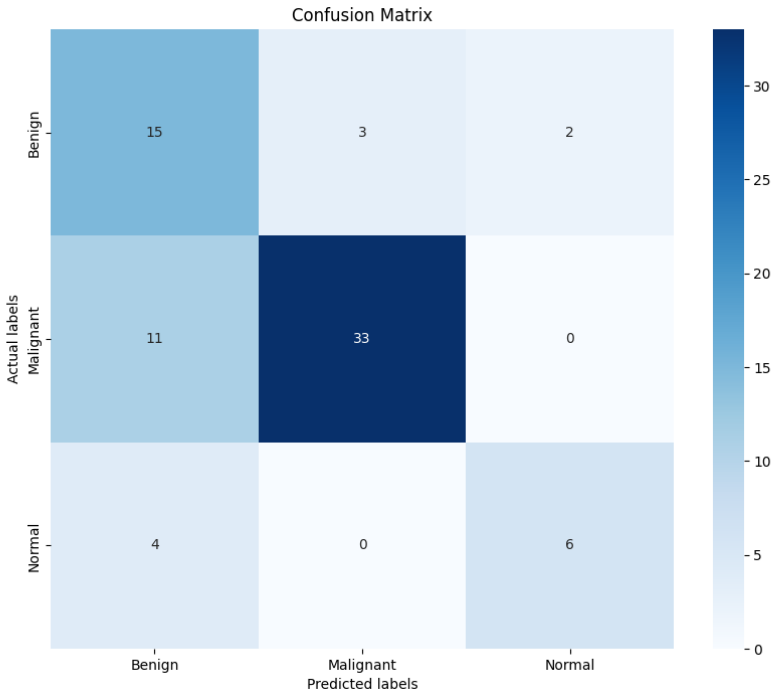


Figure 5.14: Confusion matrices for ConvNeXt Large

ConvNeXtLarge correctly identifies a substantial number of benign and malignant cases but shows increased misclassification of malignant samples as benign (Fig 5.14). Normal cases are moderately well classified, though some are still confused with benign. This indicates that the model struggles to clearly separate malignant from benign patterns despite higher capacity.

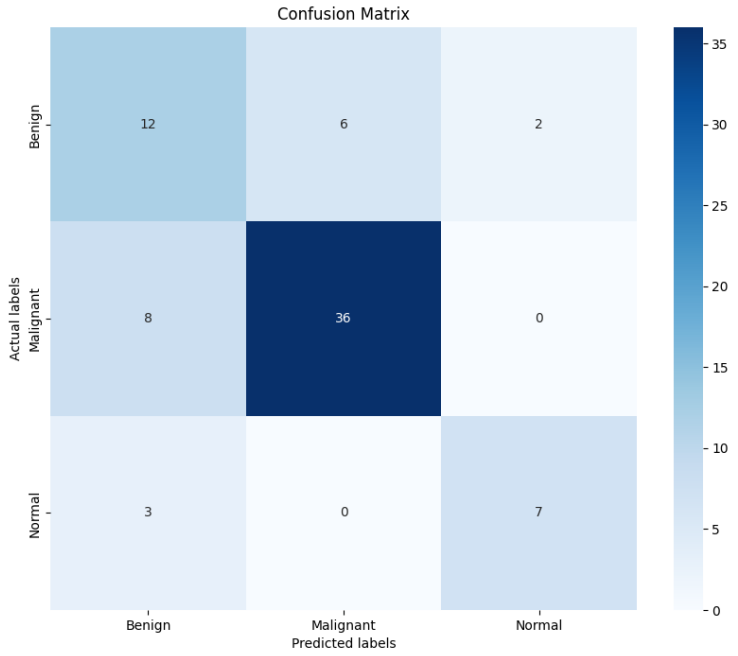


Figure 5.15: Confusion matrices for ConvNeXtXLarge

ConvNeXtXLarge correctly identifies most malignant and normal cases but continues to misclassify a notable number of benign samples (Fig 5.15). It also shows confusion between benign and malignant classes, similar to smaller variants. Overall, the model achieves slightly improved balance but still struggles with clear class separation.

A key observation from the confusion matrices is that ConvNeXt models are consistent but not dominant. They rarely collapse or misclassify severely, but they also fail to clearly outperform lightweight models like MobileNetV2. This supports the idea that ConvNeXt's deeper blocks are not fully exploited in a frozen-backbone setup where only the final layers are trainable.

5.3.4 ROC Curves

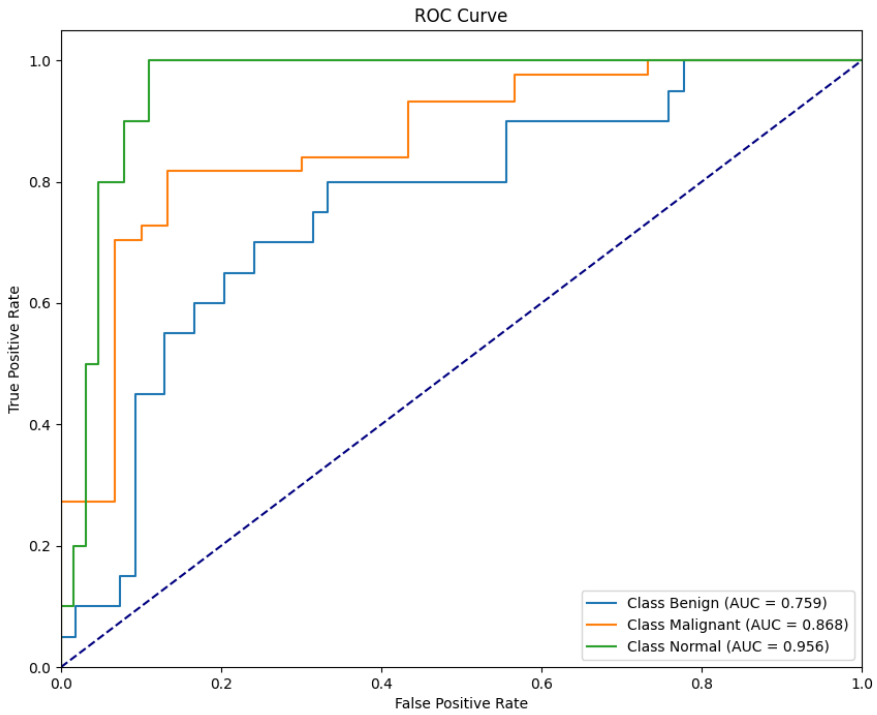


Figure 5.16: Receiver Operating Characteristic (ROC) curves for ConvNeXt Small

ConvNeXtSmall shows strong discriminative ability for the normal class (AUC = 0.956), while malignant class performance remains moderate (AUC = 0.868) (Fig 5.16). The benign class has the lowest AUC, indicating weaker separation from other classes. Overall, the model demonstrates reasonable ranking ability but lacks balanced performance across all categories.

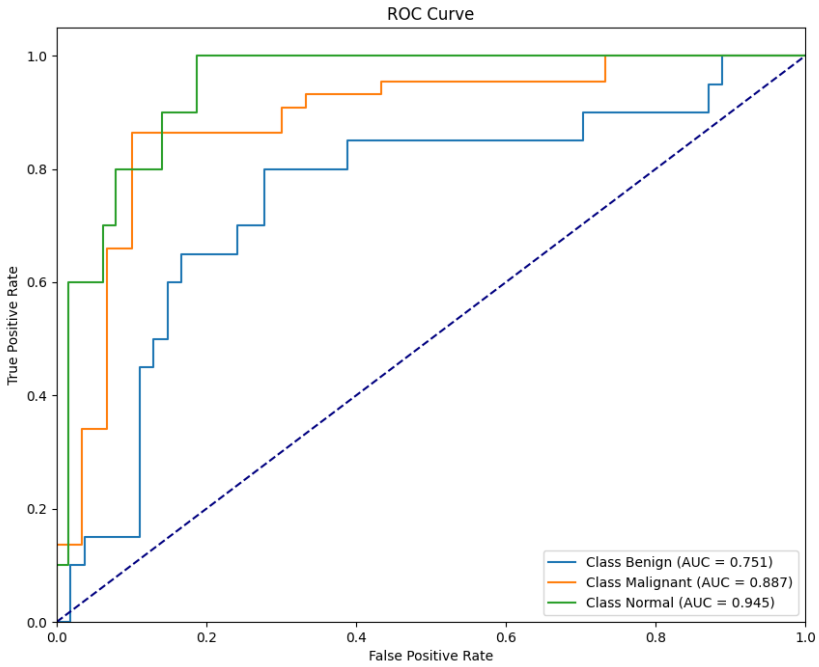


Figure 5.17: Receiver Operating Characteristic (ROC) curves for ConvNeXtLarge

ConvNeXtLarge maintains strong discriminative performance for the normal class (AUC = 0.945), with improved separation for the malignant class (AUC = 0.887) (Fig 5.17). However, the benign class continues to show weaker separability with the lowest AUC. Overall, the model demonstrates moderate improvement but still lacks balanced class-wise discrimination.

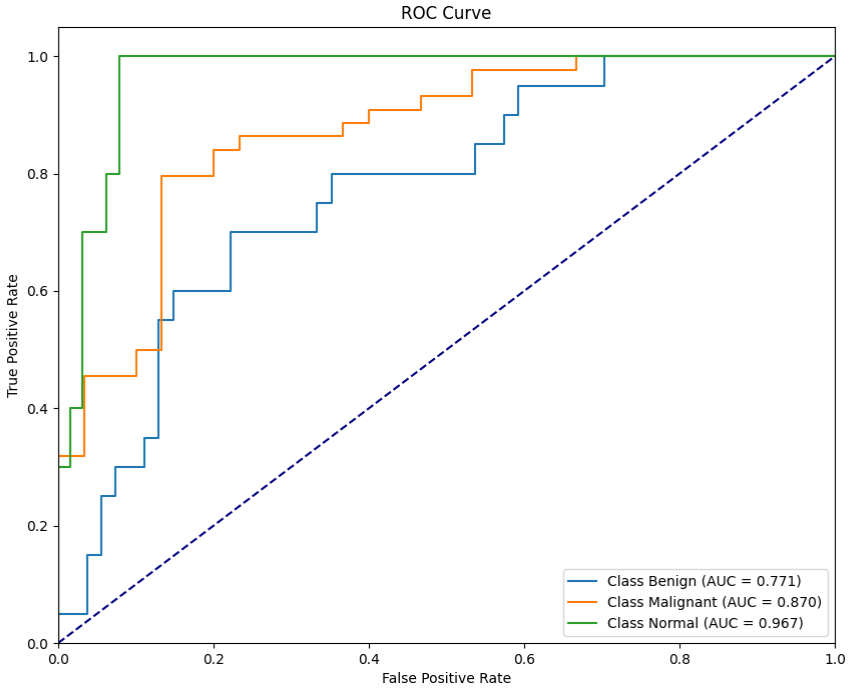


Figure 5.18: Receiver Operating Characteristic (ROC) curves for ConvNeXtXLarge

ConvNeXtXLarge achieves the highest discriminative performance for the normal class (AUC = 0.967), with consistent separation (Fig 5.18). The malignant class also shows strong performance (AUC = 0.870), while the benign class remains the weakest. Overall, the model improves slightly over smaller variants but still shows imbalance across classes.

The ROC curves collectively show that ConvNeXt models are stable rankers but not top performers.

5.4 EfficientNet Results

The EfficientNet family represents a class of models built around compound scaling, where depth, width, and resolution increase in a balanced way. Unlike architectures that scale one dimension at a time, EfficientNet expands all three according to a single scaling coefficient. This design makes the family highly parameter-efficient while still achieving strong performance on a variety of computer vision tasks. In this study, we evaluate EfficientNetB1, EfficientNetB2, and EfficientNetB6. These models differ in size and representational capacity, but all share the EfficientNet design principles of mobile inverted bottleneck blocks (MBConv), depthwise convolutions, and squeeze-and-excitation modules. As before, all variants use the same training pipeline, classification head, and evaluation metrics, making it possible to directly compare how compound scaling affects performance on liver ultrasound classification.

5.4.1 Performance Summary

Table 5.3: Performance metrics for EfficientNetB1, EfficientNetB2, and EfficientNetB6.

Models	Precision	Recall	F1 Score	AUROC	Accuracy
EfficientNetB1	0.6865	0.6985	0.6924	0.8927	0.7432
EfficientNetB2	0.7049	0.7197	0.7122	0.8867	0.7702
EfficientNetB6	0.6533	0.6545	0.6539	0.8577	0.7297

Table 5.3 presents the results of the EfficientNet variants. EfficientNetB2 achieves the strongest and most balanced performance among the three models, with Precision 0.7049, Recall 0.7197, F1-Score 0.7122, AUROC 0.8867, and Accuracy 0.7702. This makes B2 the top performer in the EfficientNet family. EfficientNetB1, one step smaller than B2, attains Precision 0.6865, Recall 0.6985, F1-Score 0.6924, AUROC 0.8927, and Accuracy 0.7432. Interestingly, B1 achieves the highest AUROC in this family (0.8927), even though its thresholder metrics are slightly lower compared to B2. EfficientNetB6, the largest model in this group, performs the weakest, with F1-Score 0.6539, Accuracy 0.7297, and AUROC 0.8577.

The performance pattern reveals a key insight: moderate scaling (B2) works best, while aggressive scaling (B6) produces diminishing returns or even degradation when applied to a relatively small and specialized dataset like liver ultrasound images. EfficientNetB1 already demonstrates solid performance with fewer parameters, but B2 adds meaningful

representational power without overfitting. B6, with its greatly increased depth and resolution, may be unnecessarily complex for this dataset and training setup.

5.4.2 Training and Validation Curves

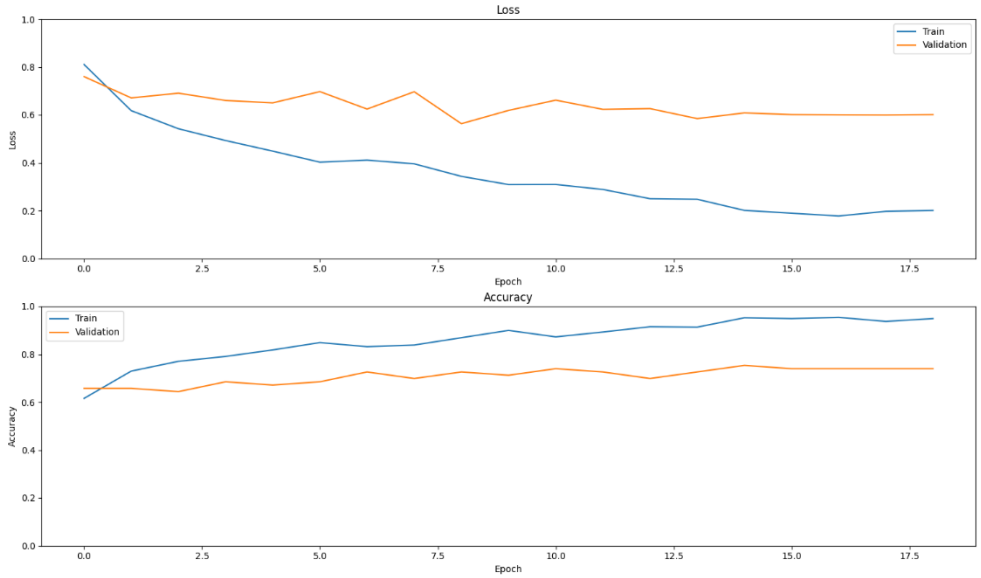


Figure 5.19: Training & Validation Curve for EfficientNetB1

EfficientNetB1 shows steady improvement in training performance, with decreasing loss and increasing accuracy over epochs (Fig 5.19). However, validation loss fluctuates and remains higher, while validation accuracy plateaus after initial gains. This indicates moderate overfitting and limited generalization to unseen data.

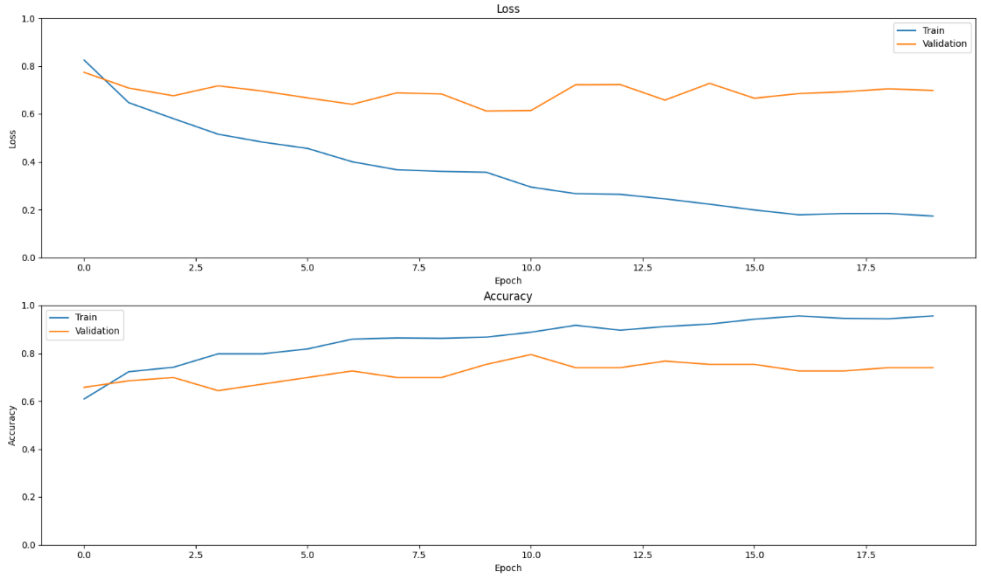


Figure 5.20: Training & Validation Curve for EfficientNetB2

EfficientNetB2 demonstrates stable and consistent learning, with training loss decreasing smoothly and validation loss remaining relatively controlled (Fig 5.20). Validation accuracy improves steadily and stays closer to training accuracy compared to B1. This indicates better generalization and a more balanced learning behavior.

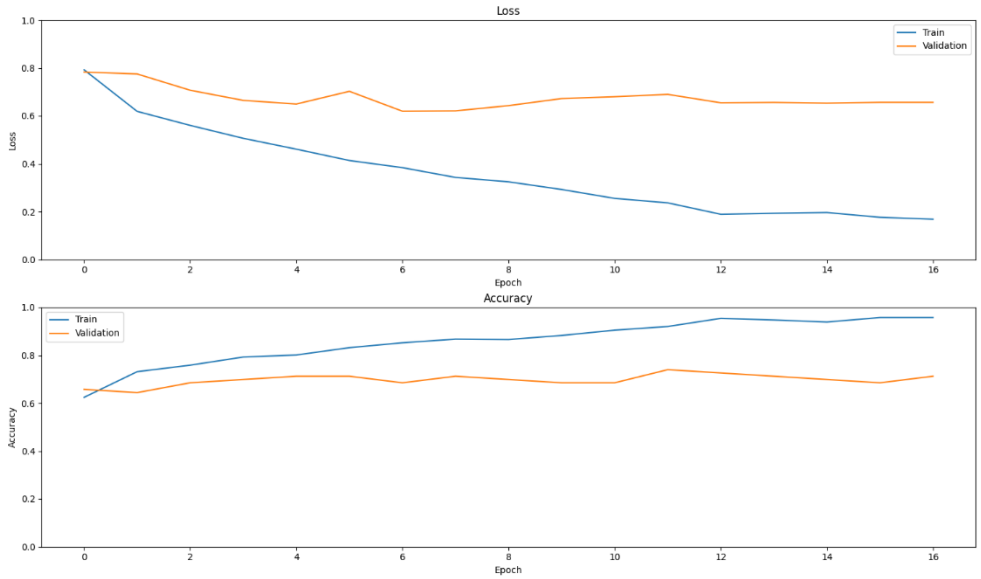


Figure 5.21: Training & Validation Curve for EfficientNetB6

EfficientNetB6 shows strong training performance with steadily decreasing loss and increasing accuracy (Fig 5.21). However, validation loss remains relatively high and unstable, while validation accuracy plateaus early. This indicates overfitting, with the model struggling to generalize despite its higher capacity.

Overall, the training curves highlight that moderate-scale EfficientNet variants are more compatible with frozen-backbone transfer learning, while very large variants may require deeper fine-tuning, more data, or increased regularization to shine.

5.4.3 Confusion Matrices

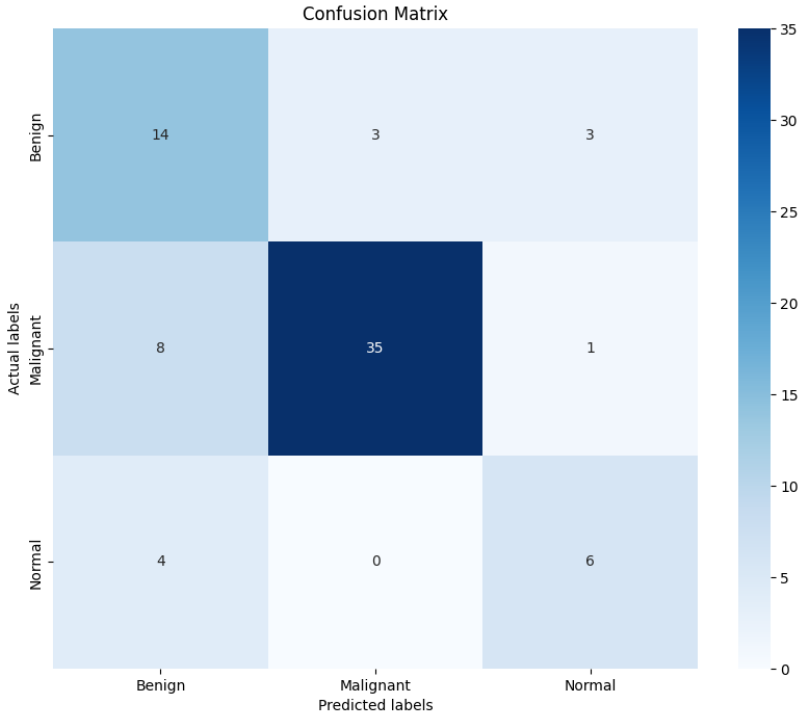


Figure 5.22: Confusion matrices for EfficientNetB1

EfficientNetB1 correctly identifies most malignant cases but shows confusion between benign and malignant samples (Fig 5.22). It also misclassifies some normal cases as benign, indicating weaker separation of non-malignant classes. Overall, the model demonstrates moderate class discrimination with some imbalance.

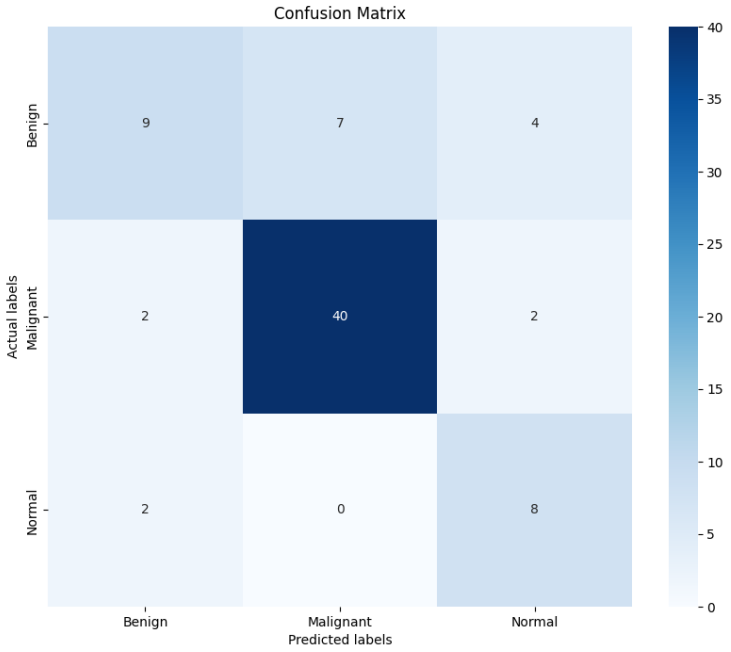


Figure 5.23: Confusion matrices for EfficientNetB2

EfficientNetB2 correctly identifies a high number of malignant cases with fewer misclassifications compared to B1 (Fig 5.23). However, it shows increased confusion in the benign class, with several samples misclassified as malignant or normal. Overall, the model demonstrates improved detection of critical cases but less balanced performance across all classes.

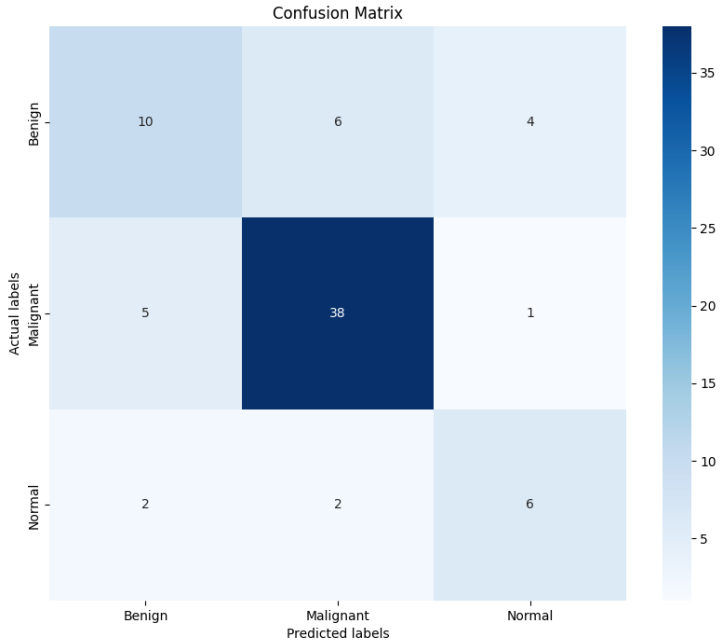


Figure 5.24: Confusion matrices for EfficientNetB6

EfficientNetB6 correctly identifies most malignant cases but still shows confusion between benign and malignant samples (Fig 5.24). It also misclassifies several benign and normal cases,

indicating weaker class separation despite higher model capacity. Overall, the model demonstrates strong detection of critical cases but lacks balanced performance across all classes.

All three models classify normal cases more accurately than diseased cases. However, since the malignant class is the most critical in terms of clinical impact, B2's relative strength in this area makes it the most reliable EfficientNet variant for this dataset.

5.4.4 ROC Curves

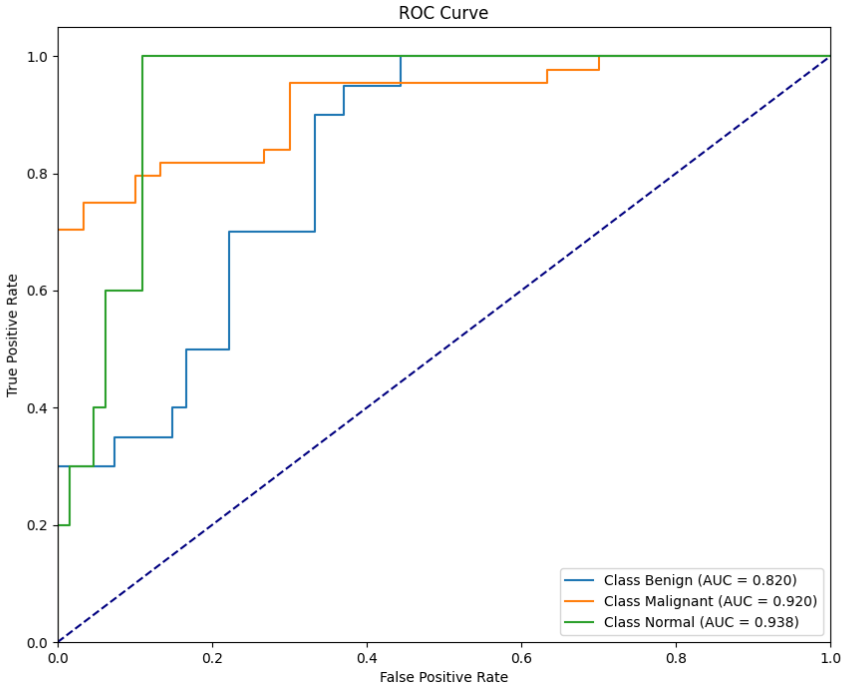


Figure 5.25: Receiver Operating Characteristic (ROC) curves for EfficientNetB1

EfficientNetB1 shows strong discriminative performance across all classes, with the highest AUC for the normal class (0.938) (Fig 5.25). The malignant class also achieves high separability (AUC = 0.920), while the benign class performs comparatively lower. Overall, the model demonstrates effective ranking ability with relatively balanced class performance.

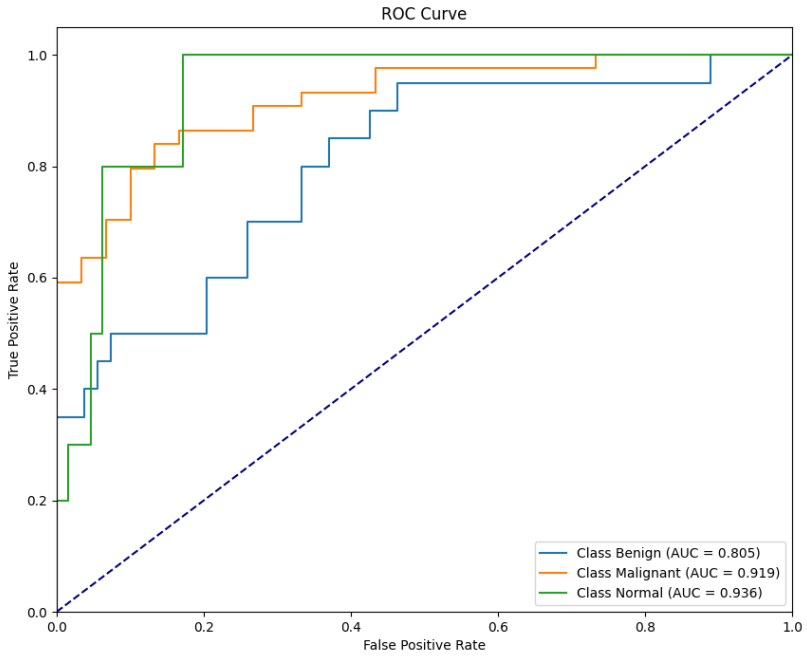


Figure 5.26: Receiver Operating Characteristic (ROC) curves for EfficientNetB2

EfficientNetB2 maintains strong discriminative performance, with the normal class achieving the highest AUC (0.936) and consistent separation (Fig 5.26). The malignant class also performs well (AUC = 0.919), while the benign class remains slightly lower. Overall, the model demonstrates balanced and reliable class-wise ranking performance.

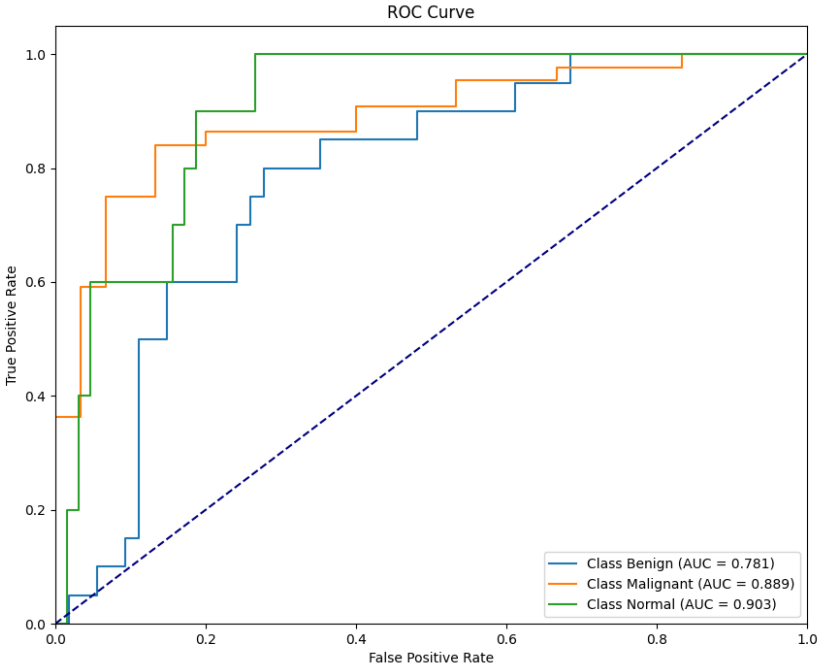


Figure 5.27: Receiver Operating Characteristic (ROC) curves for EfficientNetB6

EfficientNetB6 shows strong discriminative performance, particularly for the malignant class (AUC = 0.889) and normal class (AUC = 0.903) (Fig 5.27). However, the benign class again records the lowest AUC, indicating weaker separability. Overall, the model maintains good ranking ability but does not achieve balanced performance across all classes.

5.5 Comparison of the best models

Table 5.4: Performance comparison of MobileNetV2, EfficientNetB2, and ConvNeXtXLarge as the top-performing variants in each model family.

Models	Precision	Recall	F1 Score	AUROC	Accuracy
EfficientNetB2	0.7049	0.7197	0.7122	0.8723	0.7702
MobileNetV2	0.7256	0.7803	0.7520	0.8853	0.7837
ConvNeXtXLarge	0.7189	0.7061	0.7124	0.8693	0.7432

To understand how each architecture family performs when represented by its strongest variant, this section compares MobileNetV2, EfficientNetB2, and ConvNeXtXLarge. These models were selected because they achieved the highest F1-Score or strongest overall balance within their respective families. Table 5.4 summarizes their quantitative performance.

Across all metrics, MobileNetV2 stands out as the most reliable model for classifying liver ultrasound images under the conditions of this study. It achieved the highest F1-Score (0.7520) and highest Recall (0.7803) among all tested models, while maintaining the highest overall Accuracy (0.7837) tied with the original MobileNet. This strong Recall is particularly important in medical imaging, where the cost of missing a malignant case is high. The training curves for MobileNetV2 show steady convergence and minimal instability, indicating that the model learned efficiently and generalized well despite the limited dataset size. Its confusion matrix also shows fewer malignant-to-benign misclassifications compared to the other two best models, reinforcing its ability to detect clinically important cases.

EfficientNetB2 follows closely behind MobileNetV2, achieving an F1-Score of 0.7122 and Accuracy of 0.7702. Although its Recall is slightly lower, its performance across precision, AUROC, and accuracy remains consistently strong, making it the second-best model overall. EfficientNetB2 provides a good balance between capacity and generalization. The compound scaling strategy used in EfficientNetB2 seems to offer useful representational power without leading to overfitting, which was more noticeable in EfficientNetB6. The ROC curve for B2 shows smooth and stable separation between classes, suggesting that even if its thresholder Recall is slightly lower than MobileNetV2's, its ranking ability remains dependable.

ConvNeXtXLarge, the strongest variant in the ConvNeXt family, performs competitively but does not surpass the best models from the other two families. Its F1-Score of 0.7124 is nearly identical to EfficientNetB2's, but its Accuracy (0.7432) and Recall (0.7061) trail behind MobileNetV2's by a clear margin. The larger model size does not give ConvNeXtXLarge a decisive advantage in this setting, especially with a frozen backbone. Its confusion matrix

shows more misclassifications between benign and malignant categories, and its training curves display mild overfitting tendencies. These patterns suggest that ConvNeXtXLarge requires more data or fine-tuning to fully realize its potential.

Overall, the comparison shows that lightweight and mid-scale models outperform large-scale models on this relatively small ultrasound dataset. MobileNetV2's efficient architecture designed to preserve information flow with inverted residuals and linear bottlenecks appears well suited to the subtle textures of liver ultrasound images. EfficientNetB2 also performs strongly, supporting the idea that well-balanced compound scaling works well when data is limited. ConvNeXtXLarge, while advanced and powerful, does not provide enough benefit to justify its complexity in this context.

In practical terms, if a single model were to be deployed in a clinical decision-support tool based on the conditions of this study, MobileNetV2 would be the most appropriate choice. It offers the strongest combination of accuracy, sensitivity to malignant cases and stability during training.

Chapter 6: Conclusion

6.1 Summary

The effectiveness of three widely adopted convolutional neural network (CNN) families, namely MobileNet, EfficientNet, and ConvNeXt, was evaluated for classifying liver ultrasound images into normal, benign, and malignant categories. All experimental conditions were standardized, including the use of frozen backbones and a uniform classification head. This ensured that performance differences could be attributed primarily to architectural design rather than variations in training procedures. As a result, a fair comparison was possible across variants within the three CNN families.

Across all tested configurations, MobileNetV2 delivered the strongest and most consistent results. It achieved the best overall balance between Accuracy, Recall, and F1-Score. The results suggest that lightweight architectures can perform exceptionally well even when trained on relatively small medical imaging datasets. MobileNet's depthwise-separable convolution strategy likely contributed to efficient feature extraction without introducing unnecessary complexity. This characteristic makes the architecture particularly suitable for ultrasound images, which often contain subtle textures and noise.

EfficientNetB2 also produced competitive results, reinforcing the value of compound scaling for improving feature representation while maintaining computational efficiency. Its performance indicates that moderate scaling can provide meaningful benefits without substantially increasing the risk of overfitting under frozen backbone constraints. ConvNeXtXLarge, by comparison, achieved stable but not superior performance. One possible explanation is the limited dataset size. Another is the inability of the model to update its deeper layers during training. Larger architecture often requires extensive finetuning before they can fully adapt to domain-specific variations in medical imagery.

Taken together, these findings indicate that carefully selected lightweight CNNs can achieve strong performance in liver ultrasound classification. Such models may offer affective solutions for clinical environments where computational resources and dataset sizes are limited.

6.2 Limitations

Several limitations should be considered when interpreting the results of this study. The relatively small dataset size is among the most important. Although sufficient for comparative evaluation, it does not fully support the learning capacity of deeper models such as ConvNeXtXLarge or EfficientNetB6. Large amounts of data are typically required for these architectures to perform effectively while avoiding overfitting. Dataset diversity is also limited, as images from multiple institutions, different ultrasound devices, and varied patient

populations are not represented. Consequently, the ability of the models to generalize to real-world clinical settings may be reduced.

A frozen-backbone transfer learning strategy was adopted to ensure fair comparison across all models. While this approach maintains consistency, it prevents the networks from adapting their internal representations specifically to liver ultrasound data. Important image patterns may therefore remain underutilized. Fine-tuning could potentially improve performance, particularly for deeper architectures, but was intentionally excluded to preserve consistent experimental conditions.

The classification task was restricted to three categories: normal, benign, and malignant. This simplification makes the problem more manageable but does not capture the full spectrum of liver conditions encountered in clinical practice. As a result, the applicability of the models remains limited in broader diagnostic scenarios.

Ultrasound image quality can vary considerably across examinations. Differences in operator technique, imaging equipment, and patient-specific factors all influence image appearance. Such variability may affect model performance and presents an ongoing challenge when developing systems that must operate reliably across diverse clinical environments.

6.3 Future Work

This work establishes a basis for further research in deep learning–based liver ultrasound analysis. Expanding the dataset remains one of the most important next steps. Greater diversity in patient populations, institutions, ultrasound devices, and scanning conditions would improve model generalization and strengthen the reliability of future findings. More balanced class distributions should also be considered to ensure that each category is represented appropriately during training.

Further gains in performance may be achieved through alternative training strategies. Instead of keeping the backbone frozen, pre-trained networks could be fine-tuned to adapt their learned representations more effectively to liver ultrasound data. Additional CNN architectures may also warrant investigation, particularly if different design choices offer advantages for this classification task.

Increased data diversity can be achieved through augmentation techniques such as rotation, flipping, and contrast adjustment. These methods may improve robustness while reducing the likelihood of overfitting. Their value becomes particularly apparent when only limited medical imaging data are available.

The current framework is limited to three categories: normal, benign, and malignant. Expanding the classification scheme to include more detailed clinical classes would better reflect real diagnostic scenarios. Models developed under such conditions may be capable of distinguishing a broader spectrum of liver diseases and pathological variations.

Future developments need not be restricted to image classification alone. Clinically useful functionality could be enhanced through lesion localization and segmentation, allowing both the presence and location of abnormalities to be identified. Greater emphasis on explainability, user-friendly interfaces, and evaluation with practicing radiologists would further support the transition of AI-assisted diagnostic systems from research environments to routine clinical practice.

AI Acknowledgement

The authors acknowledge the use of AI-assisted language tools, including QuillBot and ChatGPT, exclusively for improving the clarity, grammar, and readability of the manuscript. These tools were not used to generate data, images, experimental results, or analyses. Following their use, the authors carefully reviewed and edited the manuscript and assume full responsibility for the content of this publication.

References

[1] Maucort-Boulch, D., et al. (2022). Global burden of primary liver cancer in 2020 and predictions to 2040. Journal of Hepatology, 77(6), 1598–1606. https://doi.org/10.1016/j.jhep.2022.08.021

[2] Devarbhavi, H., Asrani, S. K., Arab, J. P., Nartey, Y., Pose, E., & Kamath, P. S. (2023). Global burden of liver disease: 2023 update. Journal of Hepatology, 79(2), 516–537. https://doi.org/10.1016/j.jhep.2023.03.017

[3] World Health Organization. (2024). Global hepatitis report 2024. https://iris.who.int/bitstreams/4fc7b178-5017-4065-9970-fc5a7533cb5f/download

[4] Riazi, K., Azhari, H., Charette, J. H., Underwood, F. E., King, J. A., Afshar, E. E., Swain, M. G., Congly, S. E., Kaplan, G. G., & Shaheen, A. A. (2022). The prevalence and incidence of NAFLD worldwide: A systematic review and meta-analysis. The Lancet Gastroenterology & Hepatology, 7(9), 851–861. https://doi.org/10.1016/S2468-1253(22)00165-0

[5] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

[6] Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., … Wang, T. (2019). Deep learning in medical ultrasound analysis: A review. Engineering, 5(2), 261–275. https://doi.org/10.1016/j.eng.2018.11.020

[7] Byra, M., Styczynski, G., Szmigielski, C., Kalinowski, P., Michalowski, L., Paluszkiewicz, R., … Nowicki, A. (2018). Transfer learning with deep convolutional neural networks for liver steatosis assessment in ultrasound images. International Journal of Computer Assisted Radiology and Surgery, 13, 1895–1903. https://doi.org/10.1007/s11548-018-1843-2

[8] Acharya, U. R., Mookiah, M. R. K., Koh, J. E. W., Tan, J. H., Nayak, K., Bhandary, S. V., & Chua, C. K. (2021). Cascaded deep learning neural network for fully automated liver steatosis prediction using ultrasound images. Sensors, 21(16), 5304. https://doi.org/10.3390/s21165304

[9] Kim, T., Lee, D. H., Park, E.-K., & Choi, S. (2021). Deep learning techniques for fatty liver using multi-view ultrasound images scanned by different scanners. JMIR Medical Informatics, 9(11), e30066. https://doi.org/10.2196/30066

[10] Deep learning with ultrasound images enhances the diagnosis of nonalcoholic fatty liver disease. (2024). Ultrasound in Medicine & Biology. https://doi.org/10.1016/j.ultrasmedbio.2024.07.014

[11] Diagnostic accuracy of convolutional neural networks in classifying hepatic steatosis on ultrasound: A systematic review and meta-analysis. (2025). https://doi.org/10.1016/j.lansea.2025.100644

[12] Xia, F., Wei, W., Wang, J., Duan, Y., Wang, K., & Zhang, C. (2024). Machine learning model for non-alcoholic steatohepatitis diagnosis based on ultrasound radiomics. BMC Medical Imaging, 24, 221. https://doi.org/10.1186/s12880-024-01398-y

[13] Xi, I. L., Wu, J., Guan, J., Zhang, P. J., Horii, S. C., Soulen, M. C., … Bai, H. X. (2021). Deep learning for differentiation of benign and malignant solid liver lesions on ultrasonography. Abdominal Radiology, 46, 534–543. https://doi.org/10.1007/s00261-020-02564-w

[14] Li, Z., Wang, S., Zhang, J., … Liu, Y. (2024). Malignancy diagnosis of liver lesion in contrast-enhanced ultrasound using an end-to-end deep learning network. BMC Medical Imaging, 24, 247. https://doi.org/10.1186/s12880-024-01247-y

[15] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In ICML 2019. https://arxiv.org/abs/1905.11946

[16] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. https://arxiv.org/abs/1704.04861

[17] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In CVPR 2018. https://doi.org/10.1109/CVPR.2018.00474

[18] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., … Le, Q. V. (2019). Searching for MobileNetV3. In ICCV 2019. https://doi.org/10.1109/ICCV.2019.00140

[19] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s (ConvNeXt). https://arxiv.org/abs/2201.03545

[20] Xu, Y., Gao, S., Chen, Y., … Zhang, T. (2023). Improving artificial intelligence pipeline for liver malignancy diagnosis using ultrasound images and video frames. Briefings in Bioinformatics, 24(1), bbac569. https://doi.org/10.1093/bib/bbac569

[21] Xu, Y. (2022). Annotated Ultrasound Liver Images (Version 1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7272660

[22] Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning, 45, 171–186. https://doi.org/10.1023/A:1010920819831

[23] Zhang, Y., Jiang, J., Chen, Y., & Zhang, Y. (2021). Domain adaptation for ultrasound image classification: A systematic study. Medical Image Analysis, 70, 101990. https://doi.org/10.1109/TBME.2021.3117407

[24] Hussain, M., Anwar, S. M., Majid, M., & Khan, M. K. (2018). Quantitative analysis of liver ultrasound images using deep learning for chronic liver disease classification. Computers in Biology and Medicine, 95, 24–33. https://doi.org/10.1016/j.compbiomed.2018.02.011

[25] Smistad, E., Falch, T. L., Pedersen, A., & Lindseth, F. (2017). Deep learning-based automatic segmentation of liver and hepatic lesions in ultrasound images. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 64(10), 1753–1764. https://doi.org/10.1109/TUFFC.2017.2723044

[26] Mishra, D., Saha, P. K., & Chakraborti, T. (2020). Ultrasound liver and lesion segmentation using improved UNet architecture. Computerized Medical Imaging and Graphics, 84, 101749. https://doi.org/10.1016/j.compmedimag.2020.101749

[27] H. M., Kuijf, H. J., Kuhlmann, K. F. D., & van Ginneken, B. (2022). Explainable artificial intelligence improves radiologist confidence in liver ultrasound interpretation. European Radiology, 32, 5372–5382. https://doi.org/10.1007/s00330-022-08720-9

[28] Schalkx, H., Willemink, M. J., Tak, S., et al. (2021). Deep learning for liver disease detection in ultrasound: A multi-center evaluation. European Journal of Radiology, 141, 109813. https://doi.org/10.1016/j.ejrad.2021.109813

[29] Brown, L. G., Foran, D. J., Nosher, J. L., & Hacihaliloglu, I. (2021). Liver disease classification from ultrasound using multi-scale CNN. International Journal of Computer Assisted Radiology and Surgery, 16(9), 1537–1548. https://doi.org/10.1007/s11548-021-02414-0

[30] Pasyar, P., Mahmoudi, T., Mousavi Kouzehkanan, S.-Z., Ahmadian, A., Arabalibeik, H., Soltanian, N., & Radmard, A. R. (2021). Hybrid classification of diffuse liver diseases in ultrasound images using deep convolutional neural networks. Informatics in Medicine Unlocked, 22, 100496. https://doi.org/10.1016/j.imu.2020.100496

[31] Che, H., Ramanathan, S., Foran, D., Nosher, J. L., Patel, V. M., & Hacihaliloglu, I. (2021). Realistic ultrasound image synthesis for improved classification of liver disease. In Advances in Simplifying Medical Ultrasound (ASMUS 2021), Lecture Notes in Computer Science (Vol. 12929, pp. 187–196). Springer. https://doi.org/10.1007/978-3-030-87583-1_18

[32] Nakata, N., & Siina, T. (2023). Ensemble learning of multiple models using deep learning for multiclass classification of ultrasound images of hepatic masses. Bioengineering, 10(1), 69. https://doi.org/10.3390/bioengineering10010069

[33] Yang, Q., Wei, J., Hao, X., Kong, D., Yu, X., Jiang, T., … Wang, Y. (2020). Improving B-mode ultrasound diagnostic performance for focal liver lesions using deep learning: A multicentre study. EBioMedicine, 56, 102777. https://doi.org/10.1016/j.ebiom.2020.102777

[34] Oyama, A., Inokuchi, R., … Yamada, T. (2020). Deep learning promotes B-mode ultrasound screening for focal liver lesions. EBioMedicine, 56, 102814. https://doi.org/10.1016/j.ebiom.2020.102814

[35] Schmauch, B., Herent, P., Jehanno, P., Dehaene, O., Saillard, C., Aubé, C., … Lassau, N. (2019). Diagnosis of focal liver lesions from ultrasound using deep learning. Diagnostic and Interventional Imaging, 100(4), 227–233. https://doi.org/10.1016/j.diii.2019.02.009

[36] Hassan, T. M., Elmogy, M., & Sallam, E.-S. (2017). Diagnosis of focal liver diseases based on deep learning technique for ultrasound images. Arabian Journal for Science and Engineering, 42(8), 3127–3140. https://doi.org/10.1007/s13369-016-2387-9

[37] Liu, L., Tang, C., Li, L., Chen, P., Tan, Y., Shang, Y., … Guo, Y. (2022). Deep learning radiomics for focal liver lesions diagnosis on long-range contrast-enhanced ultrasound and clinical factors. Quantitative Imaging in Medicine and Surgery, 12(8), 3979–3993. https://doi.org/10.21037/qims-21-1004

[38] Tiyarattanachai, T., Apiparakoon, T., Marukatat, S., Sukcharoen, S., Yimsawad, S., Chaichuen, O., … Chaiteerakij, R. (2022). The feasibility to use artificial intelligence to aid detecting focal liver lesions in real-time ultrasound: A preliminary study based on videos. Scientific Reports, 12, 7749. https://doi.org/10.1038/s41598-022-11506-z