MedFoundX: A Foundation Model for Biomedical Image Classification and Segmentation

by

Md Mehedi Hasan Shawon
22166038

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
July 2025

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing my degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate references.
3. The thesis does not contain material that has been accepted or submitted for any other degree or diploma at a university or other institution.
4. I have acknowledged all main sources of help.

Student's Full Name & Signature:

Md Mehedi Hasan Shawon
22166038

Approval

The thesis titled "MedFoundX: A Foundation Model for Biomedical Image Classification and Segmentation" submitted by

1. Md Mehedi Hasan Shawon (22166038)

of Summer 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science and Engineering on **27th of July, 2025**.

Examining Committee:

Examiner:
(External)

Dr. Md. Mahbubur Rahman
Professor
Department of Computer Science and Engineering
Military Institute of Science and Technology

Examiner:
(Internal)

Dr. Muhammad Iqbal Hossain
Associate Professor
Department of Computer Science and Engineering
BRAC University

Supervisor:

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:

Dr. Md. Sadek Ferdous
Professor
Department of Computer Science and Engineering
BRAC University

Chairperson:

Dr. Sadia Hamid Kazi
Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

The demand for efficient and generalizable artificial intelligence models in medical imaging is rapidly increasing. This thesis introduces MedFoundX, a unified foundation model designed for classification and segmentation across various biomedical imaging modalities. The model's architecture features a pre-trained EfficientNet-B3 backbone, integrated with Convolutional Block Attention Modules (CBAM) and Multi-Head Attention. A sequential weight transfer training protocol is applied to eight publicly available and clinically relevant datasets, encompassing multiple imaging types, including MRI, CT, X-ray, and colonoscopy. MedFoundX achieved nearly perfect classification performance in several datasets, significantly exceeding established models such as CNN, KAN, ResNet-50, Swin-Transformer and VGG-16. In segmentation tasks, it reached mean Dice coefficients of up to 0.964 on the MedSeg-Liver dataset and 0.941 on the Kvasir-SEG dataset, considerably outperforming other models like CNN, KAN, and DeepLabV3. Furthermore, MedFoundX was tested on two unseen datasets with fine-tuning. It achieved an impressive accuracy of 98.5% on the unseen PMRAM classification dataset, with only six misclassifications out of 400 scans. In the CVC-ClinicDB segmentation dataset, MedFoundX recorded a mean Dice score of 0.83 during inference, along with high sensitivity (0.94) and specificity (0.988). Computational investigation revealed that MedFoundX has 22.21 million parameters, requires 2.8 billion FLOPs, and occupies 84.7 MB of memory, allowing real-time inference and outperforming larger models such as ResNet-50, DeepLabV3, and VGG-16. This work effectively addresses the performance-efficiency gap in clinical AI by providing a single, attention-enhanced architecture that generalizes well across various modalities and tasks, achieving state-of-the-art classification and segmentation without compromising computational feasibility. Its successful application to unseen datasets demonstrates its robust generalization capabilities.

Keywords: Biomedical imaging, Foundation model, Attention mechanism, Classification, Segmentation, Deep learning.

Dedication

To my beloved wife, Sumaiya Akter, and our precious daughter, Inaaya Hasan Yara. My wife's continuous encouragement and unwavering support have guided me through every stage of this work. My daughter's smile and boundless energy have continually renewed my motivation.

Acknowledgement

All praise and gratitude go to Almighty Allah, whose infinite mercy and guidance have made the completion of this thesis possible. I extend my heartfelt thanks to my respected advisor, Dr. Md. Golam Rabiul Alam, for his persistent support, insightful advice, and constant encouragement. His guidance was instrumental, and his mentorship has been invaluable in shaping this work. I also thank Md. Jobayer for his responsible collaboration and support in this project. His commitment to teamwork has been instrumental in the implementation of this project. Finally, I owe an immeasurable debt of gratitude to my parents, whose endless love, prayers, and support have carried us through every step of this journey. Their sacrifices and encouragement have been the foundation of my success and have guided me to the cusp of graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviations used in the document.

ϵ Epsilon

υ Upsilon

AI Artificial Intelligence

$CBAM$ Convolutional Block Attention Module

CNN Convolutional Neural Network

CT Computed Tomography

$DenseNet$ Densely Connected Convolutional Network

DL Deep Learning

$EfficientNet$ Efficient Neural Network Architecture

$fMRI$ Functional Magnetic Resonance Imaging

IoU Intersection over Union

$LIME$ Local Interpretable Model-agnostic Explanations

ML Machine Learning

MRI Magnetic Resonance Imaging

$ResNet$ Residual Network

$SHAP$ SHapley Additive exPlanations

Chapter 1

Introduction

Biomedical imaging techniques, such as computed tomography (CT), magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), ultrasound imaging, nuclear medicine imaging, etc, have become essential tools in modern healthcare. These imaging modalities provide critical information on anatomical structures, physiological functions, and pathological conditions, enabling accurate diagnosis, prognosis, and treatment planning [1], [2]. Typically, healthcare professionals use these biomedical images to diagnose diseases and develop a treatment plan. However, different personal interpretations can influence how people visually perceive these images. Errors can occur due to fatigue, outside distractions, and the inherent subjectivity of this type of analysis. In contrast, utilizing automated methods for image analysis with appropriate computer techniques promotes objective evaluation by specialists, thus enhancing both diagnostic precision and confidence in the analysis [3]. Significant advances in artificial intelligence (AI), particularly in deep learning techniques, have substantially improved the automation and improvement of biomedical image classification and segmentation tasks [4]. Deep learning architectures, particularly convolutional neural networks (CNN) and pre-trained models such as ResNet, DenseNet, and EfficientNet, have achieved satisfactory performance in various biomedical imaging domains [5], [6]. However, these models typically require task-specific designs and extensive labeled datasets and are prone to performance degradation when applied to diverse imaging modalities or clinical sites. This limits their generalization and scalability in real-world applications, which poses significant challenges for widespread clinical adoption. The advancements in the field of AI underscore the importance of foundation models, which are large-scale neural networks pre-trained on a diverse range of datasets. These models exhibit remarkable generalizability and possess strong transfer learning capabilities, making them valuable for a wide range of applications. [7], [8]. Therefore, using a single, unified foundation model for multiple biomedical imaging tasks, such as classification and segmentation, could significantly streamline workflows, reduce computational resource demands, and enhance the model's adaptability across different medical applications. This research aims to design and validate a generalizable, efficient foundation model that reliably handles both biomedical image classification and segmentation, bridging the gap between high performance and practical deployment to advance AI-driven medical imaging in clinical settings.

1.1 Problem Statement

Recent advances in biomedical imaging, primarily driven by deep learning techniques, have significantly improved diagnostic accuracy, allowing faster and more reliable patient evaluations. Despite these significant advances, several critical limitations hinder the widespread adoption and effective integration of these technologies into clinical practice. One primary concern is the limited generalizability of existing models; they often demonstrate impressive performance on benchmark datasets, but fail to replicate similar results when faced with diverse imaging modalities, such as MRI, CT, and X-ray. These discrepancies arise due to substantial variability between modalities in image characteristics, anatomical structures, imaging protocols, patient demographics, and pathological conditions. Models trained on specific datasets or modalities often struggle to generalize beyond their narrow training scenarios, thereby reducing their applicability and reliability in heterogeneous, real-world clinical environments. In addition, computational efficiency remains a significant barrier to practical deployment, particularly in resource-constrained clinical settings. Many state-of-the-art deep learning models require extensive computational resources, making them unsuitable for real-time processing or deployment on edge devices commonly used in hospitals and clinics. Moreover, computationally intensive models can introduce latency, potentially delaying critical clinical decisions and negatively impacting patient outcomes. Interpretability and transparency constitute another significant challenge. Deep learning models are often regarded as black boxes due to their complex internal mechanisms, which makes it challenging for clinicians to fully trust their decisions. In clinical practice, where interpretability and transparency are paramount, the lack of clear reasoning behind model predictions can severely undermine clinician confidence, impede regulatory approvals, and limit clinical adoption. Addressing these multifaceted challenges requires a robust, computationally efficient, and interpretable unified foundation model that can consistently perform well in multiple biomedical imaging modalities and tasks, including classification and segmentation. This unified model must leverage the diverse characteristics of various modalities to enhance generalizability and optimize computational efficiency, facilitating seamless integration into clinical workflows and improving diagnostic reliability. Although our proposed research significantly focuses on generalization and computational efficiency, further advancements in interpretability and transparency remain imperative to fully realize the potential of AI-driven biomedical imaging in clinical practice.

1.2 Research Objectives

This research aims to address the limitations found in current deep learning approaches for biomedical imaging, specifically their limited generalizability, high computational costs, and lack of scalability. The goal is to explore the potential of a unified and modality-agnostic foundation model that can address these complex challenges. The main objective of this study is to design, implement, and validate a deep learning architecture that is generalizable and computationally efficient, capable of performing robust classification and segmentation across various modalities of biomedical imaging. To achieve this, the research is guided by the

following specific objectives:

- **Development of a Unified Foundation Model:** Design a single, end-to-end deep learning framework that can simultaneously handle classification and segmentation tasks. The model architecture is designed to promote shared feature learning, reduce redundancy, and enhance scalability in various clinical applications.
- **Enhancement of Generalizability:** Improve the model's robustness across varying clinical conditions and imaging domains by training on multi-modal datasets (e.g., MRI, CT, X-ray, colonoscopy). Techniques such as cross-modal learning and data augmentation will be employed to enhance the model's adaptability and reduce overfitting.
- **Optimization for Computational Efficiency:** Develop a lightweight model architecture with a reduced parameter count and memory footprint, enabling deployment on resource-constrained edge devices commonly found in clinical environments. The model should maintain competitive diagnostic accuracy while reducing model size and resource demand.

These objectives are translated into a set of focused research questions that define the scope and trajectory of this work:

- Can a single foundation model deliver comparable or superior performance in both classification and segmentation tasks across diverse biomedical imaging modalities, relative to modality-specific models?
- Does training on heterogeneous, multi-modal biomedical datasets improve the model's ability to generalize to unseen modalities, institutions, or patient populations?
- How can architectural design reduce computational complexity without significantly compromising diagnostic accuracy, thus supporting real-time or edge-based clinical deployment?

Together, these objectives and questions form the core of this investigation, which aims to bridge the gap between performance, practicality, and clinical utility in AI-powered biomedical imaging.

1.3 Contributions

In pursuit of the research objectives and response to the critical limitations outlined in the background and problem statement, this thesis delivers the following key contributions to the fields of biomedical image analysis and deep learning:

- **A Unified Foundation Model for Multi-task Biomedical Imaging:** This research presents a novel, modality-agnostic deep learning framework that integrates both classification and segmentation tasks within a single architecture. By consolidating tasks under a shared feature backbone and task-specific heads, the model eliminates the need for separate models, enhancing scalability and simplifying clinical deployment.

- **Multi-modal Training for Model Generalization:** A comprehensive training methodology is deployed using various biomedical imaging datasets that span multiple modalities. This strategy improves model generalization to new imaging types and clinical settings with minimal fine-tuning, demonstrating adaptability across different anatomical regions, disease conditions, and patient populations.
- **Efficient Model Design for Edge Deployment:** The proposed model is intentionally designed for low-resource clinical environments, with a strong focus on computational efficiency and deployment feasibility, by leveraging a lightweight pre-trained backbone (EfficientNet-B3) and attention modules. The model achieves an optimal balance between diagnostic accuracy and efficiency. With only 22.21 million parameters and a compact model size of approximately 85MB, it is well-suited for real-time inference and deployment on edge devices in resource-constrained healthcare settings.

These contributions address the interrelated challenges of generalization, efficiency, and task integration in biomedical imaging. The findings from this research may facilitate broader clinical adoption of AI-based solutions and establish a foundation for future advancements in scalable and interpretable medical image analysis.

1.4 Thesis Organization

This thesis is organized into five chapters, each focused on a specific component of the research, from motivation and methodology to results, discussions, and future directions.

Chapter 1: Introduction lays the foundation for the thesis by discussing the background and motivation behind developing a unified foundation model for biomedical image analysis. It articulates the core problem statement, outlines the primary research objectives and questions, and highlights the key contributions of this work. The chapter concludes with an overview of the thesis structure.

Chapter 2: Literature Review provides an in-depth overview of the biomedical image analysis landscape, examining the evolution of classification and segmentation approaches. The chapter discusses recent advances in foundation models and transfer learning, with a focus on their applications in medical imaging. The chapter also explores computational efficiency strategies and concludes by identifying critical gaps in the literature that motivate the proposed work.

Chapter 3: Methodology details the datasets, and model architecture. The chapter begins by describing the biomedical imaging datasets used for classification and segmentation. It then explains data pre-processing techniques, including augmentation, normalization, and data loading. The core architecture of the proposed MedFoundX model is elaborated with a focus on its backbone (EfficientNet-B3), CBAM, Multi-Head Attention, and task-specific heads.

Chapter 4: Results and Discussion presents a comprehensive evaluation of the proposed model. The chapter discusses the implementation details, including

optimizer settings, training strategy and hardware details. Then, it defines the evaluation metrics for both classification and segmentation tasks. This is followed by detailed training performance results across various datasets, fine-tuning results on target datasets (PMRAM and CVC-ClinicDB), and an analysis of inference. Comparative analyses are provided for baseline models, such as CNN, KAN, ResNet, Swin-T, DeeplabV3, etc., in terms of diagnostic performance and computational efficiency. The chapter concludes with a detailed discussion of findings across tasks, highlighting key performance trends and practical implications.

Chapter 5: Conclusion and Future Work summarizes the major findings and contributions of the research. It critically reflects on the limitations of the current model and proposes future research directions, including improvements in generalization, interpretability, and real-world deployment. This chapter emphasizes the potential of MedFoundX as a scalable and clinically applicable foundation model for diverse biomedical imaging tasks.

Chapter 2

Literature Review

Biomedical imaging supports modern diagnostics by capturing structural, functional, and molecular information in a variety of modalities, including X-ray, CT, MRI, positron emission tomography (PET), ultrasound, and histopathology. During the past several decades, the field has moved from manual interpretation and hand-crafted feature pipelines to automated, data-driven analysis enabled by machine learning and deep learning [9], [10]. Early computer-aided detection systems relied on expert-defined descriptors (e.g. texture features, shape descriptors) and offered limited scalability and robustness. More recently, CNN and vision transformers (ViTs) have revolutionized classification, segmentation, and detection tasks by learning hierarchical features directly from images and modeling global context. These advances have spurred the development of large network foundation models pre-trained in extensive heterogeneous imaging datasets, which can be adapted to specific clinical applications with minimal supervision [11]. Despite these breakthroughs, biomedical imaging faces persistent challenges: the scarcity of annotated data, variability between scanners and institutions, the high computational demands of three-dimensional and time series data, and the need for interpretability and trustworthiness in clinical settings. Hybrid architectures and lightweight models aim to balance accuracy with efficiency, while self-supervised and multimodal foundation models promise improved generalization across modalities and tasks. The following sections delve into classification, segmentation, foundation models, and computational efficiency in more detail, discussing current methods and open challenges in each area.

2.1 Background of Classification Task

The early biomedical image classification pipelines were based on hand-made features derived from domain knowledge. Texture analysis, introduced by Haralick in the 1970s, employed gray-level co-occurrence matrices to capture the spatial arrangements of intensities. These features were simple and interpretable, but lacked a standardized workflow and were sensitive to imaging parameters [12]. On the other hand, hand-crafted descriptors, histograms of oriented gradients (HOG), and local binary patterns (LBP) dominated early medical image analysis and required expert-defined pre-processing pipelines [9]. Traditional computer-aided detection systems combined these features with rule-based heuristics (e.g. asymmetry, border irregularity, or the ABCD rule for dermoscopic images) and

machine learning classifiers, such as support vector machines or extreme gradient boosting [13]. These systems were poorly generalized because the feature descriptors and pre-processing were highly task-specific.

The advent of deep learning shifted classification to data-driven feature learning. Convolutional neural networks (CNN) automatically learn hierarchical representations from raw pixels and have become the workhorse of medical image classification. Modern CNN architectures, such as AlexNet, VGG, ResNet, and DenseNet, achieve superior performance on tasks like chest X-ray diagnosis, lesion detection, and retinal vessel segmentation [10]. CNN has inductive biases, translation equivariance, and local connectivity, which make it robust in limited datasets; however, their local receptive fields restrict global contextual awareness. Transfer-learning techniques mitigate data scarcity by using pre-trained CNNs (e.g. Inception, ResNet) as feature extractors or by fine-tuning their weights, saving time and hardware resources, and often outperform hand-crafted features [9]. Self-supervised pre-training (e.g. SimCLR or MoCo) on large volumes of unlabeled medical images further improves robustness and reduces the need for annotated data.

ViT models an image as a sequence of patches, each processed with self-attention. They capture long-range dependencies and offer intrinsic explainability via attention maps, which helps clinicians identify salient regions [11]. However, ViTs do not have built-in spatial inductive biases and thus require large training datasets and extensive computational resources. Hybrid architectures that combine CNN tokenizers with transformer encoders (e.g. TransUNet, SwinUNet) can alleviate data demands but remain expensive. Lightweight vision transformers address efficiency concerns by fusing convolution and attention modules. MobileViT treats transformers as convolutions: local feature extraction uses separable convolutions in depth and bottleneck layers, while the global context is modeled with a light self-attention block; despite having only about 6 million parameters, it outperforms MobileNetV3 and DeiT on ImageNet [14]. HI-MViT further reduces computation by combining shortcut connections and depth-wise separable convolutions, as well as embedding MobileViT blocks. It efficiently encodes local and global information, achieving superior accuracy in dermoscopy data sets while remaining deployable on mobile devices [13]. AlzheimerViT adapts MobileViT to brain MRI classification; convolutional MV2 blocks extract local features and transformer blocks capture global context, allowing accurate Alzheimer's diagnosis with modest computational demands [15].

2.2 Background of Segmentation Task

Medical image segmentation has followed a similar trajectory. The U-Net encoder-decoder architecture [16] introduced in 2015 established a standard deep learning baseline for biomedical segmentation, with a contracting path that captures context and an expansive path that features skip connections, allowing precise localization. Extensions such as UNet++ [17], Attention UNet [18], and DeepLabV3+ [19] incorporate dense skip connections, attention gates, and attractive spatial pyramid clustering to provide multiscale context and focus on salient structures. Attention modules, such as the squeeze-and-excitation or convolutional block attention

module, enhance feature weighting but increase model complexity.

Transformer-based segmentation leverages self-attention to achieve a holistic understanding of spatial relationships. Vision transformers process all pixel positions simultaneously and employ positional encoding to maintain spatial order, whereas pure transformers struggle with small datasets and high computational costs [20]. Hybrid architectures mitigate this by combining convolutional encoders with transformer decoders. A recent survey catalogs dozens of transformer-based networks (UNETR, TransBTS, CoTr, MISSFormer, etc.) that integrate modified self-attention and multi-scale connections [20]. For volumetric data, strategies include dividing volumes into patches or using convolutional layers to reduce feature maps before applying attention, balancing global context and computational efficiency.

Generative adversarial networks (GAN) have been utilized for adversarial segmentation, where a discriminator enforces anatomical realism in the predicted masks, while diffusion models offer high-fidelity segmentation by iteratively denoising predictions. Graph neural networks perform segmentation when anatomical structures form graphs (e.g., vascular trees) to capture relational patterns that CNN may miss. Morphological operations remain valuable for pre- and post-processing; for example, ground-truth lung masks for CT segmentation can be extracted via morphological operators and manual refinement before training a U-Net variant [21]. Hybrid architectures sometimes integrate morphological layers or differentiable soft morphological filters to refine segmentation boundaries.

2.3 Foundation Models

Foundation models (FM) represent a paradigm shift in medical imaging. These large, pre-trained networks learn general-purpose representations from vast and diverse datasets via self-supervised learning and can be adapted to numerous downstream tasks with minimal supervision. They typically combine large-scale pretraining, self-supervised objectives (such as contrastive, masked modeling, or generative tasks), and flexible adaptation strategies, including prompt engineering, linear probing, parameter-efficient fine-tuning, and full fine-tuning [11]. CNN backbones offer strong inductive biases and perform well with limited data, whereas vision transformers scale better to large datasets but require more resources. Hybrid backbones (e.g. Swin or convolutional vision transformers) combine local and global processing and are often employed in FM.

Self-supervised pre-training reduces reliance on annotations. Contrastive methods (SimCLR, MoCo) and masked image modeling (MAE, BEiT) train encoders to differentiate between positive and negative pairs or reconstruct masked patches. In the context of medical image segmentation, self-supervised pre-training in tens of thousands of unlabeled CT or MRI volumes produces representations that generalize across modalities [22]. Multimodal FMs, such as CLIP, extend contrastive learning to image-text pairs: positive pairs comprise images and their corresponding captions, encouraging the model to learn a unified representation space [23]. Self-prediction models mask parts of an image or text and train the model to reconstruct the missing

content. Vision Transformer–based approaches, such as BEiT and MAE, have shown superior performance after fine-tuning. Generative models (autoencoders, GAN, and diffusion models) learn to reconstruct or generate data, while generative vision–language models use instructions and produce outputs such as radiology reports by combining image encoders with language decoders [23].

Adapting FM to biomedical tasks can be achieved through several strategies. Full fine-tuning updates all parameters but is resource-intensive; parameter-efficient methods, such as low-rank adaptation (LoRA) and adapters, insert small, trainable modules into an otherwise frozen model, thereby reducing computational cost and mitigating catastrophic forgetting [22]. Adapter modules can be designed to capture 3D spatial or temporal information, enabling 2D pretrained encoders to process volumetric data. Prompt engineering adjusts the input to guide the model. Manual prompts (e.g., point prompts in the Segment Anything Model) achieve zero-shot segmentation; automated prompt-tuning methods iteratively refine prompts by assessing segmentation quality [22]. These strategies enable FM to be customized for specific organs, diseases, or modalities without requiring retraining from scratch.

2.4 Computational Efficiency

Deploying deep models in clinical environments requires careful attention to computational efficiency. High-capacity models can overwhelm edge devices and delay real-time inference. Several strategies address this challenge, like model compression, light-weight architectures, hardware-aware neural architecture search, parameter-efficient fine-tuning of foundation models, and test-time and style adaptation.

Pruning removes unimportant weights or neurons to obtain a sparse network. Early techniques, such as Optimal Brain Damage and Optimal Brain Surgeon, demonstrated that pruning could significantly reduce parameters with minimal loss of accuracy. Deep compression combines pruning with quantization and Huffman coding; grouping weights into shared quantized values reduces memory footprint and enables efficient inference [24]. Mixed-precision quantization is widely supported. Knowledge distillation trains a compact student model to mimic a large teacher model; the student can match or exceed the teacher's performance, although the process may be time-consuming and depends on the teacher's quality [25].

CNN with depth-wise separable convolutions (e.g., MobileNet) reduce multiply-accumulate operations. MobileViT extends this idea by embedding a small self-attention block to capture global context while maintaining a small footprint. HI-MViT utilizes an Improved-MV2 module with shortcut and depth-wise convolution to reduce computation, and MobileViT blocks that efficiently encode local and global information [13]. In AlzheimerViT, MobileViT layers downsample the input and employ lightweight transformer blocks, enabling efficient MRI classification [15].

Instead of manually designing efficient networks, hardware-aware NAS automatically searches for architectures that strike a balance between accuracy and latency.

MnasNet explicitly incorporates the inference latency measured on real devices into its optimization objective and uses a hierarchical search space; it achieves 75.2 % top-1 accuracy with 78 ms latency on a Pixel phone, outperforming MobileNetV2 and NASNet for the same parameter budget [26]. Such approaches can tailor models to specific hardware (e.g., mobile phones or specialized accelerators) and are promising for edge deployment. Methods like LoRA and adapters not only facilitate adaptation, but also reduce memory usage by keeping most parameters frozen [22]. Because only a small number of parameters are trained, these methods drastically lower GPU memory requirements and computation.

Data augmentation (e.g., rotations, translations) and style transfer (e.g., CycleGAN) increase data diversity and improve generalization, but may introduce artifacts. Test-time adaptation dynamically updates model parameters during inference to match the target distribution; however, it can be computationally expensive and requires careful consideration to avoid degrading performance. Collectively, these strategies enable high-performance biomedical image analysis while keeping resource consumption within the constraints of clinical workflows and edge devices.

2.5 Discussion and Research Gaps

The evolution of biomedical image analysis, from early rule-based and feature-engineered techniques to modern deep learning approaches, has led to substantial improvements in classification and segmentation accuracy. CNN has revolutionized the field by enabling automated feature extraction, while attention-based architectures and vision transformers have further improved performance by modeling global dependencies. Hybrid models, such as TransUNet and Swin-UNet, exemplify the state of the art, combining convolutional tokenization with self-attention mechanisms to improve spatial and contextual awareness. Although self-supervised and transfer learning approaches have shown promise, their success is often contingent on the alignment between pre-training and target domains. Techniques such as GAN-based augmentation, domain adaptation, and federated learning address data scarcity and heterogeneity but introduce new challenges, including potential artifact generation, instability during training, and complexity in implementation. Some of the research gaps identified from the literature review are given below:

- **Fragmented Task-Specific Solutions:** The current landscape remains largely fragmented, with separate models developed for classification, segmentation, or specific imaging modalities. This leads to redundancy in model training, increased maintenance burden, and challenges in unified deployment pipelines.
- **Limited Generalizability:** Many existing models perform well on controlled datasets but struggle to generalize across diverse imaging modalities (e.g., MRI, CT, ultrasound), scanner types, and patient demographics. This restricts their real-world applicability and reliability in heterogeneous clinical environments.

- **Computational Complexity:** Transformer-based models and high-capacity CNN often require substantial computational resources, hindering their deployment in edge devices or in time-sensitive clinical workflows. Even powerful models become impractical if they cannot meet the efficiency requirements of real-time applications or resource-constrained settings.
- **Interpretability Challenges:** As models become more complex, their decision-making processes become increasingly opaque. This lack of interpretability undermines clinician trust and hinders clinical integration, particularly in high-stakes diagnostic scenarios.

The gaps in current biomedical imaging underscore the urgent need for a generalizable, efficient, and interpretable foundation model. This model should utilize shared representations and attention mechanisms, support multitask learning, and be optimized for various clinical settings. Addressing these challenges is essential for advancing biomedical AI from experimental phases to real-world clinical applications. Through our research, we aim to introduce a unified framework that can effectively perform both classification and segmentation across multiple imaging modalities, while ensuring clinical-grade efficiency and robustness. We plan to incorporate multi-site and multi-modal training strategies, which are vital for achieving strong generalization in diverse clinical environments. Additionally, we seek to address the lack of emphasis on computational feasibility in existing models, as few architectures are designed to be lightweight and suitable for real-time or edge deployment. We will also focus on balancing the advantages of attention mechanisms with their computational costs, which is critical for practical implementation in resource-constrained environments.

Chapter 3

Methodology

This section provides a comprehensive overview of the methodology used to develop and evaluate our unified biomedical imaging model. It details the composition and pre-processing of various diverse datasets, describes the general architecture of the unified framework for handling both classification and segmentation tasks, and outlines the training strategy implemented to ensure robust performance across different modalities. Additionally, it summarizes the approach to model evaluation, including the use of standard performance metrics and computational considerations. After training the model on a collection of eight datasets, we further fine-tuned it on two additional datasets: one for a classification task and another for a segmentation task, in preparation for inference. The fine-tuned models were then evaluated on their respective datasets. We also benchmarked our model, MedfoundX, against established baseline models using both accuracy-focused evaluation metrics and computational efficiency metrics.

3.1 Datasets

To ensure comprehensive training and evaluation of our proposed model, we used ten diverse biomedical imaging datasets sourced from established and reputable public repositories. These datasets encompass various imaging modalities, including MRI, CT, and endoscopy, and address both classification and segmentation tasks. This variety enables a thorough assessment of the model's generalizability and performance across various domains. A detailed summary of the datasets used in this study is provided in Table 3.1.

3.1.1 Brain MRI ND-5 Dataset

The Brain MRI ND-5 collection [27] contains 17784 axial T1-weighted scans sourced from multiple tertiary hospitals. Board-certified neuroradiologists manually annotate each image for four mutually exclusive categories: glioma, meningioma, pituitary tumor, and no visible tumor. The breadth of the data set, encompassing tumor subtypes and patient demographics, makes it attractive for benchmarking generalizable tumor detection pipelines and transfer learning studies aimed at early diagnosis and surgical planning.

Table 3.1: Summary of the Datasets

Dataset	Modality	Primary Task(s)	Reference
Brain MRI ND-5	MRI	Multi-class classification, One-vs-rest classification	[27]
Brain Tumor MRI	MRI	Brain tumor classification	[28], [29]
Mammogram Mastery	X-ray	Binary classification, Saliency-map localization	[30]
Pancreas CT	CT	Binary classification, Pancreas segmentation	[31]
PMRAM	MRI	Multi-class classification	[32]
Brain Stroke CT	CT	Binary classification, Lesion segmentation	[33]
CVC-ClinicDB	Colonoscopy	Polyp segmentation, Detection	[34]
Kvasir-SEG	Colonoscopy	Polyp segmentation, Object detection	[35]
MedSeg COVID CT	CT	Semantic segmentation	[36]
MedSeg Liver CT	CT	Liver segmentation	[37]

3.1.2 Brain Tumor MRI Dataset

This dataset, created by combining data from Figshare, SARTAJ, and Br35H, supports supervised learning for brain tumor detection. It includes 7,023 axial T1-weighted MRI slices (PNG, 512×512 px) across four categories: Glioma Tumour, Meningioma Tumour, Pituitary Tumour, and No Tumour. The 'No Tumour' images are exclusive to the Br35H cohort, while the tumour images primarily originate from Figshare and a cleaned subset of SARTAJ, which was re-audited to correct misclassifications. All volumes were converted to 2-D slices, skull-stripped, bias-field corrected, and z-score normalized, then resized to 240×240 px. By integrating data from various institutions and scanners, the dataset introduces variability in acquisition methods, making it suitable for benchmarking domain adaptation and label shift techniques in tumor classification.

3.1.3 Mammogram Mastery Dataset

Compiled by Ahmed et al. [30], this collection includes anonymised digital breast X-rays from teaching hospitals in Sulaymaniyah, Iraq, acquired between 2020 and 2023. Each study was double-read by two radiologists and reviewed by a senior breast-imaging specialist, resulting in binary labels: *Cancer* and *No Cancer*. The dataset contains:

- **745 source images** (DICOM to 1024×1024 PNG): 378 malignant and 367 benign/normal cases.
- **9685 augmented projections** through rotations ($\pm 15°$), flips, and other techniques, totaling 10430 images for model development.

All scans underwent preprocessing, including de-identification, artefact removal, and intensity normalisation, to preserve full-field anatomical context for multi-scale analysis. With its Middle-Eastern demographic, this dataset complements existing mammography corpora, such as CBIS-DDSM and INbreast, thereby enhancing model generalisability. The curated labels and augmentation techniques support benchmarking for breast-cancer detection and related studies.

3.1.4 Pancreas CT Dataset

Collected jointly by West China Hospital and Sichuan People's Hospital, this contrast-enhanced abdominal CT set distinguishes *P-Pancreas* (acute or chronic pancreatitis) from *N-Pancreas* (radiologically normal pancreas) [31]. The images are reconstructed at 1 mm isotropic resolution with standard soft tissue kernels. Institutional review boards provided ethical clearance and a balanced design (913 cases per class) to avoid bias. The dataset underpins pancreas segmentation pre-training and binary pathology classifiers, supplying crucial diversity in imaging modalities for our multitask foundation model.

3.1.5 PMRAM: Bangladeshi Brain-Cancer MRI Dataset

The PMRAM dataset [32] aggregates 6000 T1-weighted MRIs from four Bangladeshi hospitals, providing 1500 images for each of *Glioma*, *Meningioma*, *Pituitary*, and *No Tumour*. Original 512×512 images underwent data-augmentation (rotation, translation, shear, zoom, horizontal flips) to mitigate overfitting while maintaining radiological plausibility through the fill_mode=nearest strategy. Clinical oversight by a multidisciplinary team (four neuro-oncologists and neuro-radiologists) ensures high-quality ground truth. Its geographic and scanner diversity complements ND-5, allowing robustness checks across populations in the final unified model. Together, these four datasets span two imaging modalities (MRI, CT) and multiple organ systems (brain, pancreas) for classification tasks, thus providing a comprehensive test bed for evaluating the proposed foundation model.

3.1.6 The Brain Stroke CT Dataset

The Brain Stroke CT Dataset [33], created by Türkiye's Ministry of Health for the TEKNOFEST-2021 AI in Healthcare competition, includes 877 CT and 230 MRI series from 819 patients, derived from 67000 CT and 18000 MR studies (2019–2020). It contains 6651 axial slices (2-5 mm thick): 4427 without acute stroke, 1131 with hyperacute/acute ischemia, and 1093 with intracranial hemorrhage. The blind test sets consist of 200 slices (130 normal/chronic, 35 ischemic, 35 hemorrhagic) and 100 slices (48 ischemic, 47 hemorrhagic, 5 mixed). Seven radiologists delineated every lesion using a RAD-X PACS workstation, and each slice is provided as four files: 16-bit DICOM, PNG, single-channel mask, and color overlay, all fully anonymized. The dataset supports two tasks: (i) binary stroke classification and (ii) pixel-wise lesion segmentation using a dilated and eroded intersection-over-unit metric.

3.1.7 CVC-ClinicDB Dataset

CVC-ClinicDB [34] is an open-access repository of colonoscopy images launched in 2015 by the Computer Vision Center and Hospital Clínic de Barcelona. It was created for the ISBI 2015 Endoscopic Vision Challenge to aid in the automatic detection of polyps. The dataset contains 612 RGB frames of size 384 $\times$ 288 pixels, sampled from 29 video sequences of 25 colonoscopy studies. Each frame shows one polyp and comes with a pixel-perfect binary mask. The images and masks are organized in TIFF folders without predefined splits, allowing custom train/validation/test partitions. The dataset also includes clinical metadata on polyp size, morphology, and histology, supporting performance evaluation based on relevant factors. CVC-ClinicDB focuses on small lesions that can lead to missed diagnoses, serving as a benchmark for polyp segmentation, detection, and saliency-based localization.

3.1.8 Kvasir-SEG

Kvasir-SEG [35] is an open-access benchmark for colorectal polyp segmentation that enhances the Kvasir GI dataset with precise pixel annotations and bounding-box metadata. It includes 1,000 RGB colonoscopy images in JPEG format, each paired with a binary PNG mask where polyp tissue is shown in white against a black background. A JSON file provides the four-corner bounding box for each lesion. The masks were manually delineated by engineers and physicians and verified by an experienced gastroenterologist. This dataset supports the evaluation of semantic segmentation and object detection algorithms and serves as a common baseline for advanced models like ResUNet and FCM clustering.

3.1.9 MedSeg Dataset

MedSeg, a Norwegian start-up, offers several public datasets, including the MedSeg COVID-19 CT and MedSeg Liver CT datasets. The MedSeg COVID-19 CT dataset comprises 100 axial slices from over 40 RT-PCR-confirmed COVID-19 patients, each slice annotated by a radiologist for ground-glass opacity, consolidation, and pleural effusion. The images are formatted at 512 $\times$ 512 PNG. Meanwhile, the MedSeg Liver CT dataset features 90 contrast-enhanced abdominal CT volumes from the Medical Decathlon Liver dataset, complete with 8-channel masks that outline Couinaud segments and a 5-segment mask. All annotations are conducted by a team of radiologists for high accuracy. Both datasets are organized into image and mask folders, distributed without restrictive licenses, and designed to facilitate research in medical image segmentation.

3.2 Dataset Pre-processing

The pre-processing pipeline prepares medical imaging data for classification and segmentation tasks, ensuring robust and consistent model training. Images are converted to RGB color space and resized to 64 $\times$ 64 pixels for segmentation tasks. Binary masks are resized using nearest-neighbor interpolation to preserve mask integrity.

3.2.1 Data Augmentation

Data augmentation enhances model generalization by introducing variability during training, using the Albumentations library. The pipeline applies horizontal flips, vertical flips, random rotations, and random adjustments to brightness and contrast. These transformations enhance robustness to variations in medical images, including brain MRI and segmentation datasets. Augmentation is disabled during validation and testing to ensure consistent evaluation.

3.2.2 Normalization and Data Loading

Images are normalized using ImageNet-standardized mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. Datasets are split into training and validation sets with an 80:20 ratio, and random shuffling is used for unbiased sampling. Data loaders use a batch size of 32.

3.3 Model Specification

Medical image analysis is integral to modern healthcare, enabling critical tasks such as disease diagnosis through classification and the delineation of anatomical structures through segmentation. These applications require models that can capture and interpret complex patterns across various imaging modalities, including X-rays, MRI, and CT scans. In response to this need, we introduce **MedFoundX**, a unified deep learning framework specifically designed to perform classification and segmentation within a single architecture optimized for medical imaging. MedFoundX incorporates a pre-trained CNN backbone called EfficientNet-B3 [38], augmented with Convolutional Block Attention Modules (CBAM) [39] and Multi-Head Attention mechanisms inspired by the Transformer architecture [40]. These enhancements enable the model to focus on clinically relevant features, improving its sensitivity to subtle pathological variations. Furthermore, the framework employs modular, task-specific heads, allowing it to flexibly adapt to a wide range of clinical scenarios, such as identifying pneumonia on chest radiographs or segmenting tumors in brain MRIs, while maintaining high performance across multiple datasets. This section provides a comprehensive theoretical and mathematical exposition of MedFoundX, detailing its components.

3.3.1 Overview of the MedfoundX

The MedFoundX model is a multi-task architecture that integrates feature extraction, attention mechanisms, and task-specific processing to handle classification and segmentation. Its design is motivated by the need for a unified model that can adapt to the heterogeneity of medical imaging datasets, where tasks vary in complexity and data availability is often limited. The model consists of three primary components. Backbone: A pre-trained CNN (EfficientNet-B3) extracts multiscale feature maps, providing robust initial representations. Classification Head: This module utilizes CBAM, Multi-Head Attention, and a classifier to generate class scores, making it suitable for diagnostic tasks. Segmentation Head: This module employs a UNet-style architecture with CBAM, Multi-Head Attention,

and a decoder block to generate pixel-wise segmentation masks. The illustration of the MedFoundX architecture is shown in Figure 3.1.

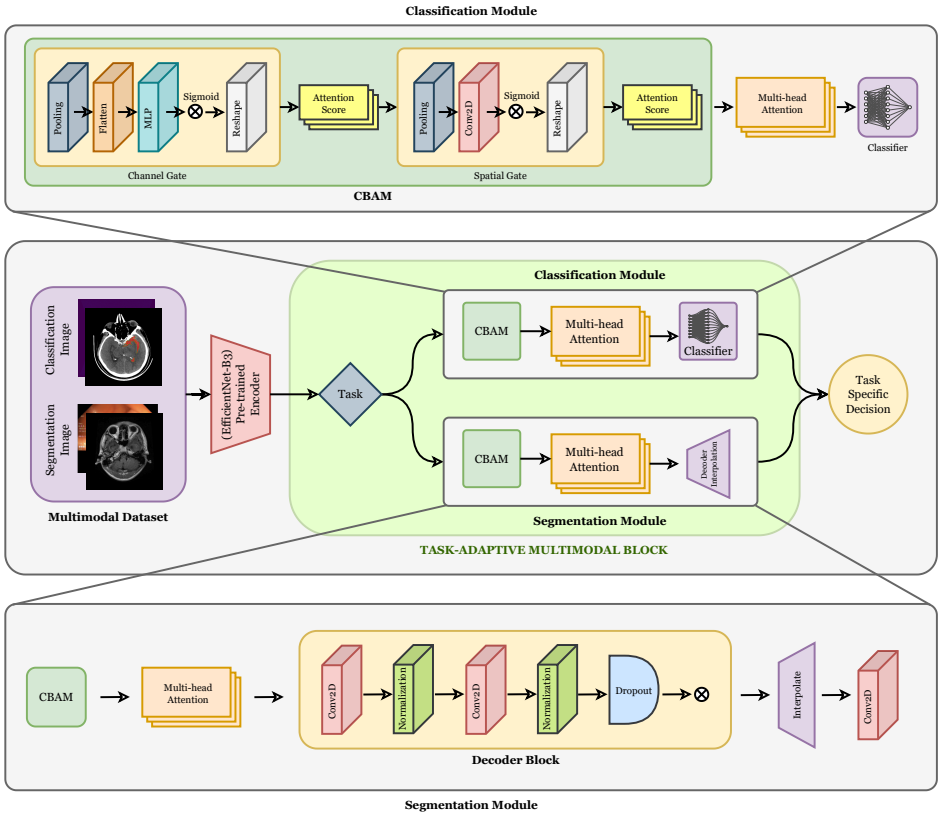

Figure 3.1: Illustration of the MedFoundX Model Architecture

The forward pass is defined by the task type (`classification` or `segmentation`) and dataset name, routing features to the appropriate head. Mathematically, the model is expressed as follows:

$$\mathcal{M}(x,t,d) = \begin{cases} \mathcal{H}_c^d\left(\mathcal{F}_b(x)_{-1}\right) & \text{if } t = \texttt{classification} \\ \mathcal{H}_s^d\left(\mathcal{F}_b(x),(H,W)\right) & \text{if } t = \texttt{segmentation} \end{cases} \tag{3.1}$$

where: $x \in \mathbb{R}^{B\times C\times H\times W}$: Input image tensor (batch size B, channels C, height H, width W), t: Task type (`classification` or `segmentation`), d: Dataset name, $\mathcal{F}_b$: Backbone feature extractor, outputting a list of feature maps, $\mathcal{H}_c^d$: Classification head for dataset d, $\mathcal{H}_s^d$: Segmentation head for dataset d.

3.3.2 Backbone: EfficientNet-B3

The backbone of the proposed foundation model is EfficientNet-B3, a CNN from the EfficientNet family, designed to optimize both accuracy and computational efficiency through a compound scaling strategy [38]. This method uniformly scales the network depth, width, and input resolution using fixed scaling coefficients, allowing a better performance-efficiency trade-off compared to traditional scaling methods.

EfficientNet-B3 is constructed using *Mobile Inverted Bottleneck Convolution* (MBConv) blocks, initially introduced in MobileNetV2 [41]. Each MBConv block consists of a point-wise expansion layer, a depth-wise separable convolution, and a point-wise projection. These are further augmented with Squeeze-and-Excitation (SE) modules [42], which adaptively recalibrate channel-wise feature responses. The architecture employs the Swish activation function to enhance non-linear representation capability. The network begins with a 3×3 stem convolution using 40 filters and progresses through multiple stages composed of MBConv blocks with kernel sizes of 3×3 or 5×5, depending on the stage. It concludes with a 1×1 convolution, global average pooling, and a fully connected output layer [43]. Although the total number of layers in EfficientNet-B3 depends on the counting method, estimates may exceed 200 when accounting for all parameterized operations. The model contains approximately 12 million parameters and is organized for modular extensibility and efficiency. This architectural design achieves a strong balance between performance and resource usage, making EfficientNet-B3 particularly well-suited as a feature extractor in medical imaging pipelines, where preserving high-resolution detail and ensuring deployability on resource-constrained hardware are critical. Table 3.2 provides a comparative overview of the variants of EfficientNet in terms of resolution, parameter count, and computational cost.

Table 3.2: Comparison of EfficientNet Variants (B0 to B7)

Model	**Input Resolution**	**Parameters (M)**	**FLOPs (B)**	**Top-1 Accuracy (%)**
EfficientNet-B0	224×224	5.3	0.39	77.1
EfficientNet-B1	240×240	7.8	0.70	79.1
EfficientNet-B2	260×260	9.2	1.0	80.1
EfficientNet-B3	300×300	**12.0**	**1.8**	**81.6**
EfficientNet-B4	380×380	19.0	4.2	82.9
EfficientNet-B5	456×456	30.0	9.9	83.6
EfficientNet-B6	528×528	43.0	19.0	84.0
EfficientNet-B7	600×600	66.0	37.0	84.3

In the context of this work, EfficientNet-B3 was selected as the backbone due to its favorable trade-off between model size and expressive power. Its ability to preserve fine-grained spatial features while maintaining a relatively modest parameter footprint makes it particularly suitable for medical image classification and segmentation tasks within a unified framework.

3.3.3 Convolutional Block Attention Module

The Convolutional Block Attention Module (CBAM) is a lightweight attention mechanism that refines feature maps by emphasizing important channels and spatial regions. It is integrated into both classification and segmentation heads to improve feature discriminability, particularly for medical images, where subtle differences (e.g., lesion boundaries) are crucial.

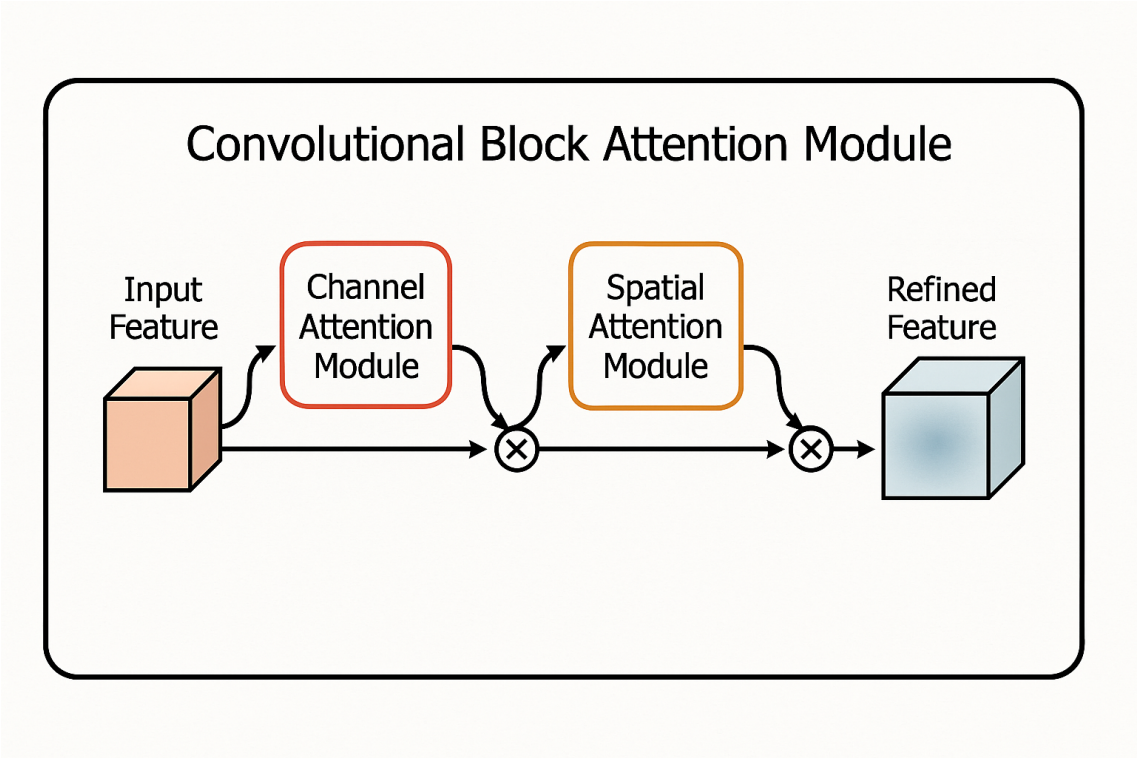

Figure 3.2: Illustration of the CBAM Model [39]

CBAM consists of two sequential modules: Channel Attention and Spatial Attention. The sequential application of channel and spatial attention by CBAM mimics human visual attention, prioritizing relevant features while suppressing noise, as shown in Figure 3.2. Its lightweight design (minimal parameters) makes it suitable for medical imaging, where computational resources may be limited. The use of multiple pooling methods enhances robustness by capturing complementary feature statistics, thus improving the overall accuracy. The general operation is as follows:

$$f_{\text{CBAM}} = \text{CBAM}(f) = f \cdot a_c \cdot a_s \tag{3.2}$$

where $f \in \mathbb{R}^{B\times C\times H\times W}$ is the input feature map, $a_c \in \mathbb{R}^{B\times C\times 1\times 1}$ is the channel attention map, and $a_s \in \mathbb{R}^{B\times 1\times H\times W}$ is the spatial attention map.

Channel Attention

Channel attention helps the network discern which feature channels contain the most critical information. It accomplishes this by assigning importance weights to each channel based on its relevance to the task. Specifically, the channel attention module applies both average pooling and max pooling across spatial dimensions to capture a comprehensive summary of feature responses:

$$f_{\text{avg}} = \text{AvgPool}_{\text{2D}}(f) \in \mathbb{R}^{B\times C\times 1\times 1}, \tag{3.3}$$

$$f_{\text{max}} = \text{MaxPool}_{\text{2D}}(f) \in \mathbb{R}^{B\times C\times 1\times 1}. \tag{3.4}$$

These pooled features are then processed through a shared multi-layer perceptron (MLP), which includes a dimensionality reduction step controlled by a reduction

ratio parameter $r = 16$. This helps the module efficiently learn channel-wise dependencies:

$$a_{\text{avg}} = W_1 \cdot \text{ReLU}(W_0 \cdot \text{Flatten}(f_{\text{avg}})), \tag{3.5}$$
$$a_{\text{max}} = W_1 \cdot \text{ReLU}(W_0 \cdot \text{Flatten}(f_{\text{max}})), \tag{3.6}$$

where $W_0 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_1 \in \mathbb{R}^{C \times \frac{C}{r}}$.
The resulting features from average and max pooling paths are summed and passed through a sigmoid activation to produce a normalized attention map:

$$a_c = \sigma(a_{\text{avg}} + a_{\text{max}}). \tag{3.7}$$

Finally, this channel-wise attention map a_c is broadcast across the spatial dimensions and multiplied element-wise with the original input feature map, effectively recalibrating the importance of each channel:

$$f' = f \cdot a_c. \tag{3.8}$$

Spatial Attention

The Spatial Attention module focuses on identifying *where* critical features appear within the spatial domain of feature maps. It achieves this by producing a spatial attention map that highlights regions containing valuable information.
Initially, the module applies two complementary pooling strategies: max pooling and average pooling in the channel dimension to distill essential spatial characteristics.

$$p_{\text{max}} = \text{MaxPool}_{\text{channel}}(f) \in \mathbb{R}^{B \times 1 \times H \times W}, \tag{3.9}$$
$$p_{\text{avg}} = \text{AvgPool}_{\text{channel}}(f) \in \mathbb{R}^{B \times 1 \times H \times W}. \tag{3.10}$$

These two feature maps, capturing distinct yet complementary aspects, are concatenated along the channel axis to form an enhanced representation:

$$p = [p_{\text{max}}, p_{\text{avg}}] \in \mathbb{R}^{B \times 2 \times H \times W}. \tag{3.11}$$

Subsequently, a convolutional operation with a kernel size of 7×7 is applied to this combined representation, effectively aggregating local contextual information to produce the final spatial attention map:

$$a_s = \sigma\left(\text{Conv}_{7\times7}(p)\right) \in \mathbb{R}^{B \times 1 \times H \times W}. \tag{3.12}$$

Finally, this spatial attention map a_s is applied multiplicatively to the input feature map f' obtained from the preceding Channel Attention step, emphasizing the spatial regions of highest relevance:

$$f'' = f' \cdot a_s. \tag{3.13}$$

The choice of a 7×7 convolutional kernel is deliberate, as it strikes an optimal balance between receptive field size and computational efficiency, effectively capturing both local and contextual spatial information.

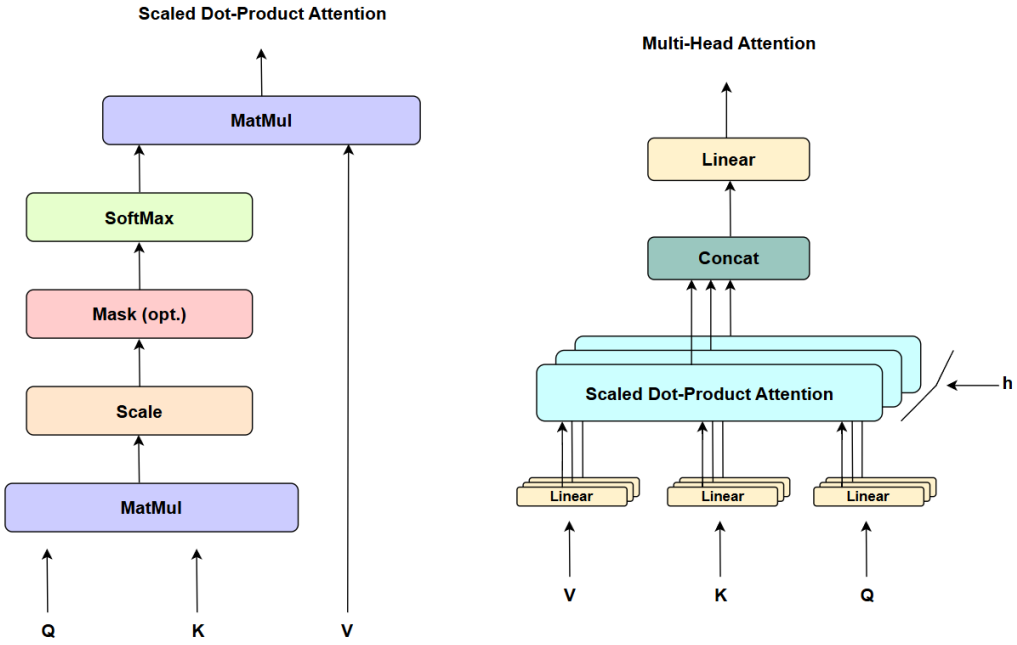

Figure 3.3: Illustration of the Multi-head Attention Mechanism [40]

3.3.4 Multi-Head Attention

The Multi-Head Attention mechanism enhances the model's ability to capture relationships across spatial regions by simultaneously focusing on multiple positions within the feature maps. This process allows the model to aggregate context-sensitive information from different spatial locations and enhances representation learning significantly. First, the spatial dimensions of the feature map f_4 are flattened into a sequence to facilitate attention computation:

$$f_{\text{flat}} = \text{Reshape}(f_4) \in \mathbb{R}^{B\times(H_4\cdot W_4)\times C_4}. \tag{3.14}$$

From this flattened representation, three linear transformations produce distinct sets of vectors: queries Q, keys K, and values V. Each captures different aspects of the input data, enabling fine-grained attention computation:

$$Q = f_{\text{flat}}W_q, \quad K = f_{\text{flat}}W_k, \quad V = f_{\text{flat}}W_v, \tag{3.15}$$

where $W_q, W_k, W_v \in \mathbb{R}^{C_4\times C_4}$ are learnable parameter matrices.
To allow the model to simultaneously attend to multiple representation subspaces, queries, keys, and values are divided into multiple attention heads. In our implementation, we use $N_h = 8$ heads, reshaping the matrices accordingly:

$$Q_h = \text{Reshape}(Q) \in \mathbb{R}^{B\times N_h\times(H_4\cdot W_4)\times \frac{C_4}{N_h}}, \tag{3.16}$$

$$K_h = \text{Reshape}(K) \in \mathbb{R}^{B\times N_h\times(H_4\cdot W_4)\times \frac{C_4}{N_h}}, \tag{3.17}$$

$$V_h = \text{Reshape}(V) \in \mathbb{R}^{B\times N_h\times(H_4\cdot W_4)\times \frac{C_4}{N_h}}. \tag{3.18}$$

Attention scores are then computed for each head independently using a scaled dot-product operation, ensuring stable gradients and effective learning:

$$A_h = \text{Softmax}\left(\frac{Q_h K_h^\top}{\sqrt{C_4/N_h}}\right). \tag{3.19}$$

Each head's output O_h is calculated as a weighted combination of the corresponding value vectors:

$$O_h = A_h V_h. \tag{3.20}$$

Subsequently, outputs from all heads are concatenated and linearly projected back into the original feature dimension, thus aggregating diverse context into a unified representation:

$$O = \text{Concat}(O_h), \quad O_{\text{proj}} = OW_o. \tag{3.21}$$

Finally, the result is added to the initial flattened input through a residual connection and reshaped back to its original spatial dimensions, ensuring seamless integration into subsequent processing steps:

$$f_{\text{attn}} = f_{\text{flat}} + O_{\text{proj}}, \tag{3.22}$$

$$f' = \text{Reshape}(f_{\text{attn}}) \in \mathbb{R}^{B \times C_4 \times H_4 \times W_4}. \tag{3.23}$$

This multi-head approach significantly improves the network's ability to model complex spatial relationships and capture global contextual dependencies within medical images.

3.3.5 Classification Head

The classification head is designed to generate diagnostic predictions from the feature representation $f_4 \in \mathbb{R}^{B \times C_4 \times H_4 \times W_4}$ obtained from earlier stages of the architecture. It leverages both the CBAM, as described in Section 3.3.3, and Multi-Head Attention, as detailed in Section 3.3.4, to refine channel-wise and spatial representations, facilitating the modeling of complex feature interactions and global contextual dependencies. Initially, the input feature map f_4 is enhanced through the CBAM module, which emphasizes informative channels and spatial regions. Subsequently, Multi-Head Attention is applied to capture long-range interactions across spatial locations, resulting in a refined feature representation f'.

To generate class predictions, the refined feature map f' undergoes several transformations. First, global spatial information is aggregated through average pooling across the spatial dimensions:

$$p = \text{AvgPool}_{\text{2D}}(f') \in \mathbb{R}^{B \times C_4 \times 1 \times 1}. \tag{3.24}$$

The resulting tensor p is then flattened into a vector to facilitate subsequent linear operations:

$$p_{\text{flat}} = \text{Flatten}(p) \in \mathbb{R}^{B \times C_4}. \tag{3.25}$$

To enhance stability and accelerate convergence during training, layer normalization is applied to this flattened representation:

$$p_{\text{norm}} = \text{LayerNorm}(p_{\text{flat}}). \tag{3.26}$$

The normalized features are then passed through a fully connected layer, preceded and followed by dropout for regularization. A rectified linear unit (ReLU) activation introduces non-linearity, allowing the network to model complex decision boundaries:

$$z = \text{Dropout}\left(\text{ReLU}(W_1 \cdot \text{Dropout}(p_{\text{norm}}))\right), \tag{3.27}$$

where $W_1 \in \mathbb{R}^{512 \times C_4}$ represents learnable parameters mapping the input features to a 512-dimensional intermediate representation. Finally, a linear projection transforms this intermediate representation into class scores:

$$y = W_2 \cdot z \in \mathbb{R}^{B \times K}, \tag{3.28}$$

where $W_2 \in \mathbb{R}^{K \times 512}$ contains the learnable parameters that produce class-specific scores, and K denotes the total number of classes. This structured approach ensures robust classification by effectively integrating channel-wise, spatial, and global contextual information into the final prediction, enhancing diagnostic accuracy and reliability.

3.3.6 Segmentation Head

The segmentation head generates high-resolution, pixel-wise segmentation masks by decoding the hierarchical feature representations extracted by the encoder. The decoding process begins by passing it to the encoder feature map using the CBAM, followed by Multi-Head Attention. This ordering ensures that both local feature recalibration and global contextual dependencies are modeled before spatial upsampling and refinement. Initially, the encoder generates hierarchical feature maps:

$$\mathcal{F}_b(x) = \big[f_0,\, f_1,\, f_2,\, f_3,\, f_4\big], \quad f_i \in \mathbb{R}^{B \times C_i \times H_i \times W_i},$$

which are reordered from deepest to shallowest to match decoder progression:

$$f' * i = f * 4 - i, \quad i = 0, 1, \ldots, 4. \tag{3.29}$$

The deepest encoder feature f'_0 is processed as follows. First, CBAM is applied:

$$\tilde{f}\_0 = \text{CBAM}(f'\_0). \tag{3.30}$$

Next, Multi-Head Attention captures long-range dependencies within the recalibrated features:

$$\hat{f}\_0 = \text{MultiHeadAttention}(\tilde{f}\_0). \tag{3.31}$$

The output is then refined using a dedicated decoding module, the DecoderBlock, defined as:

$$d\_0 = \text{DecoderBlock}(\hat{f}\_0) \in \mathbb{R}^{B \times D\_0 \times H\_4 \times W\_4}, \tag{3.32}$$

where the DecoderBlock is a stack of convolution, batch normalization, ReLU activation, and dropout layers, defined as:

$$d = \text{ReLU}\left(\text{BN}\left(\text{Conv} * \left(\text{Dropout}\left(\text{ReLU}\left(\text{BN}\left(\text{Conv} *(x)\right)\right)\right)\right)\right)\right). \tag{3.33}$$

After the initial decoding stage, the decoder features are interpolated directly back to the original input image resolution:

$$u\_\text{final} = \text{Interpolate}\big(d\_0\big). \tag{3.34}$$

The final segmentation mask is obtained via passing it to a convolution layer, mapping the refined features to the desired number of segmentation classes K:

$$s = \text{Conv} * \big(u\_\text{final}\big) \in \mathbb{R}^{B\times K\times H\times W}. \tag{3.35}$$

This decoding strategy leverages skip connections between corresponding encoder-decoder layers, ensuring detailed spatial information is preserved and effectively integrated into final predictions. The combined use of CBAM and multi-head attention significantly boosts the model's sensitivity to critical spatial structures, such as fine anatomical boundaries and subtle lesion features, making it highly suitable for precise biomedical segmentation tasks.

The proposed foundation model utilizes task-specific classification and segmentation heads, enabling efficient multi-task learning with a shared backbone. Given a set of tasks $T = \{t_{\text{cls}}, t_{\text{seg}}\}$, the overall operation can be expressed as:

$$y = \begin{cases} H^t_{\text{cls}}(F_b(x)), & t \in t_{\text{cls}} \\ H^t_{\text{seg}}(F_b(x)), & t \in t_{\text{seg}} \end{cases} \tag{3.36}$$

where H^t_{cls} and H^t_{seg} represent the task-specific heads for classification and segmentation tasks, respectively, and $F_b(x)$ denotes the shared backbone features.

Dynamic addition and removal of these heads enable quick adaptation to new datasets without retraining the entire backbone. Practically, the model benefits from multi-scale feature integration, robust pre-training, modularity, and attention-driven feature refinement, all of which are crucial for precise medical imaging tasks. Hyperparameters, such as the CBAM reduction ratio r, the number of attention heads N_h, and dropout rates, remain critical for fine-tuning performance. Moreover, grayscale inputs require careful preprocessing, as the pre-trained backbone is designed to expect RGB images. Compared to standard architectures, like ResNet [44], the proposed model offers superior adaptability and a richer representation of features.

3.4 Loss Functions and Multi-task Weights

The loss function framework supports classification and segmentation tasks, addressing class imbalances in medical imaging datasets. For classification, cross entropy loss computes the negative log-likelihood of predicted versus true class probabilities. For segmentation, a combined loss balances pixel-wise accuracy and region overlap using cross-entropy loss and dice loss, weighted equally:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Dice}}$$

In multi-task learning, the total loss is a weighted sum:

$$\mathcal{L}_{\text{total}} = \sum_{t\in\{\text{cls,seg}\}} w_t \mathcal{L}_t,$$

with task weights $w_t = 1.0$, adjustable for task prioritization, ensuring robust optimization across tasks like tumor classification, polyp segmentation, etc.

Chapter 4

Results and Discussion

4.1 Implementation Details

This section gives a quick snapshot of the implementation. It summarizes the optimization and learning-rate schedule, explains the staged training across eight datasets with logging and checkpoints, sketches the fine-tuning approach for new tasks, and notes the hardware and software setup used for all experiments.

4.1.1 Optimizer and Learning Rate Schedule

The AdamW optimizer is used, leveraging adaptive learning rates and decoupled weight decay. It is configured with a learning rate of 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 10^{-4}. Only the EfficientNet-B3 backbone and task-specific head (classification or segmentation) are optimized. A ReduceLROnPlateau scheduler reduces the learning rate by a factor of 0.5 if validation loss plateaus for 5 epochs, ensuring efficient convergence. Training runs for 500 epochs, with checkpoints saved every 25 epochs and for the best validation loss. Early stopping is disabled during initial training, allowing for complete training cycles across datasets.

4.1.2 Training Procedure

The training process sequentially processes eight datasets: four classifications (Brain MRI-ND5, Brain Tumor MRI, Mammogram Mastery, and Pancreas CT) and four segmentations (Kvasir-SEG, MedSeg-COVID, MedSeg-Liver, and The Brain Stroke CT). The `FoundationModelTrainer` initializes task-specific data loaders, loss functions, and metrics (accuracy/F1 for classification, Dice/IoU for segmentation). The model, with an EfficientNet-B3 backbone enhanced by CBAM and multi-head attention with 8 heads, optimizes the shared backbone and relevant head, freezing the other to prevent interference. Each dataset is trained for 500 epochs with a batch size of 32. Metrics are logged, and visualizations (e.g., training history, metrics comparisons) are generated. Checkpoints are saved every 25 epochs and for the best validation loss. Early stopping is disabled, ensuring thorough training. The results are saved in JSON and CSV formats, enabling performance analysis across tasks.

4.2 Fine-tuning Strategy

Finetuning adapts the pretrained foundation model to task-specific datasets, such as brain stroke segmentation. It uses a learning rate of 10^{-3}, 500 epochs, and a batch size of 32. Finetuning employs the same preprocessing pipeline (using 64×64 images and augmentations) and a combined loss (CrossEntropy + Dice) for segmentation tasks. This strategy ensures the model adapts to specialized medical imaging tasks while leveraging pre-trained features, enhancing performance on different datasets.

4.3 Hardware and Environment

All experiments were carried out on a workstation equipped with an Intel Core i7-14700 CPU, an NVIDIA RTX 4070 Ti Super GPU with 16 GB of VRAM, and 64 GB of RAM. This configuration efficiently supports mini-batch training of size 32 on images resized to 64×64, with GPU memory usage remaining below 80%. The implementation was developed using the PyTorch deep learning framework. This setup ensured stable and efficient model training in both classification and segmentation tasks, while maintaining scalability and reproducibility for deployment in various computational environments.

4.4 Evaluation Metrics

This section outlines the evaluation criteria for our model in brief. For classification tasks, we report accuracy, precision, recall, and F1 scores, each weighted to account for class imbalance. For segmentation tasks, we track pixel-level accuracy alongside two overlap measures: mean Dice and mean IoU, providing a balanced view of both boundary quality and regional agreement between the predicted and ground-truth masks.

4.4.1 Classification Metrics

For classification tasks, such as brain tumor and pancreas classification, model performance is evaluated using accuracy, precision, recall, and the F1 score, calculated per batch and aggregated across epochs. Accuracy measures the proportion of correctly classified samples:

$$\text{Accuracy (\%)} = \left(\frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \right) \times 100$$

Precision, recall, and F1 score are calculated with weighted averaging to account for class imbalances, common in medical datasets:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}},$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}},$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

4.4.2 Segmentation Metrics

For segmentation tasks, such as polyp segmentation and liver segmentation, performance is assessed using pixel accuracy, mean Dice score, and mean Intersection over Union (IoU). Pixel accuracy measures the proportion of correctly classified pixels:

$$\text{Pixel Accuracy (\%)} = \left(\frac{\text{Correctly Classified Pixels}}{\text{Total Pixels}} \right) \times 100$$

The Dice score quantifies overlap between predicted and true masks:

$$\text{Dice} = \frac{2 \cdot |\text{Pred} \cap \text{True}|}{|\text{Pred}| + |\text{True}|},$$

where $|\text{Pred} \cap \text{True}|$ is the intersection of predicted and true mask pixels. IoU measures the ratio of intersection to union:

$$\text{IoU} = \frac{|\text{Pred} \cap \text{True}|}{|\text{Pred} \cup \text{True}|}.$$

Both Dice and IoU are averaged across classes (e.g., binary segmentation with 2 classes), with per-class values reported for detailed evaluation.

4.5 Foundation Model Training

During the training session, we trained one dataset at a time and transferred the updated weights to the following dataset, allowing it to learn from the different datasets gradually. Here, the training of each dataset is explained in terms of the training sequences.

4.5.1 Classification Datasets Training Performance

For the Brain MRI ND-5 dataset, training and validation loss curves drop dramatically in the first few epochs, as shown in Figure 4.1, indicating that the network quickly learns key visual patterns in the images. The curves are nearly indistinguishable, with validation occasionally dipping below training, indicating strong regularization and no overfitting. The F1 score effectively measures the performance of the balance of classes, increasing along with the accuracy to almost 0.999 at the end of the training in Figure 4.2. The overlap in training and validation lines indicates that the model consistently balances false positives and false negatives for both known and unseen scans. Precision and recall start slightly apart but quickly converge in Figure 4.3, both exceeding 0.97 within 30 epochs and oscillating around 0.999 thereafter. Their near-identity suggests that the classifier effectively distinguishes between healthy and diseased MRIs. The training and validation trials overlap, which confirms the strong performance of the network on new data.

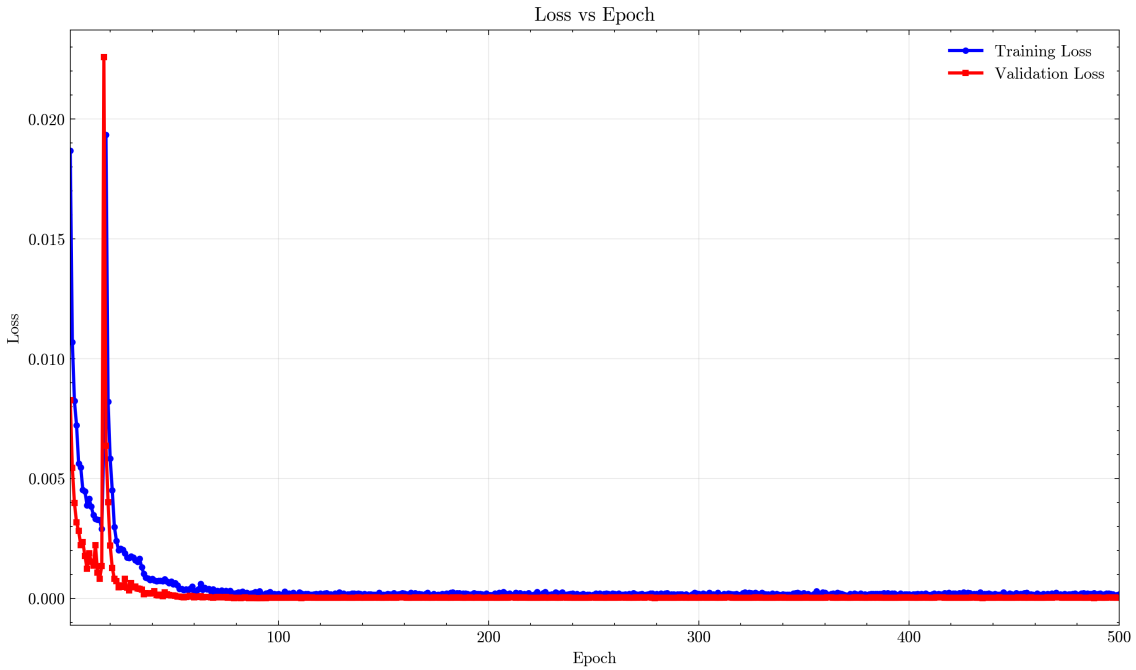

Figure 4.1: Brain MRI ND5 Dataset Training (Loss Vs Epoch)

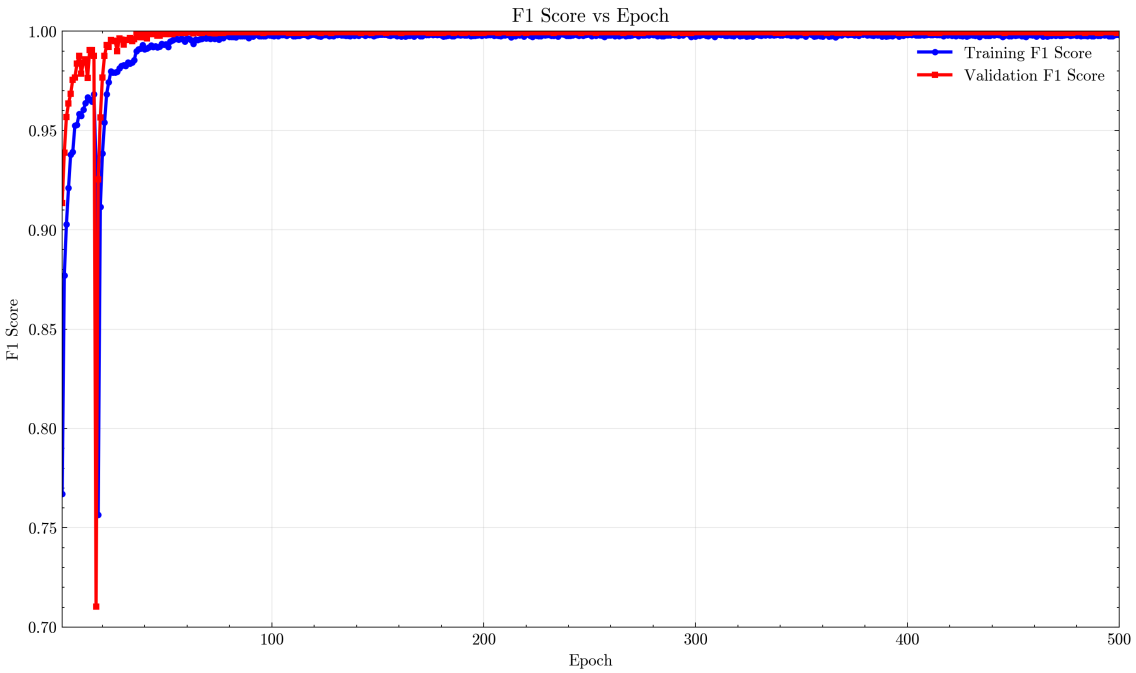

Figure 4.2: Brain MRI ND5 Dataset Training (F1 Score Vs Epoch)

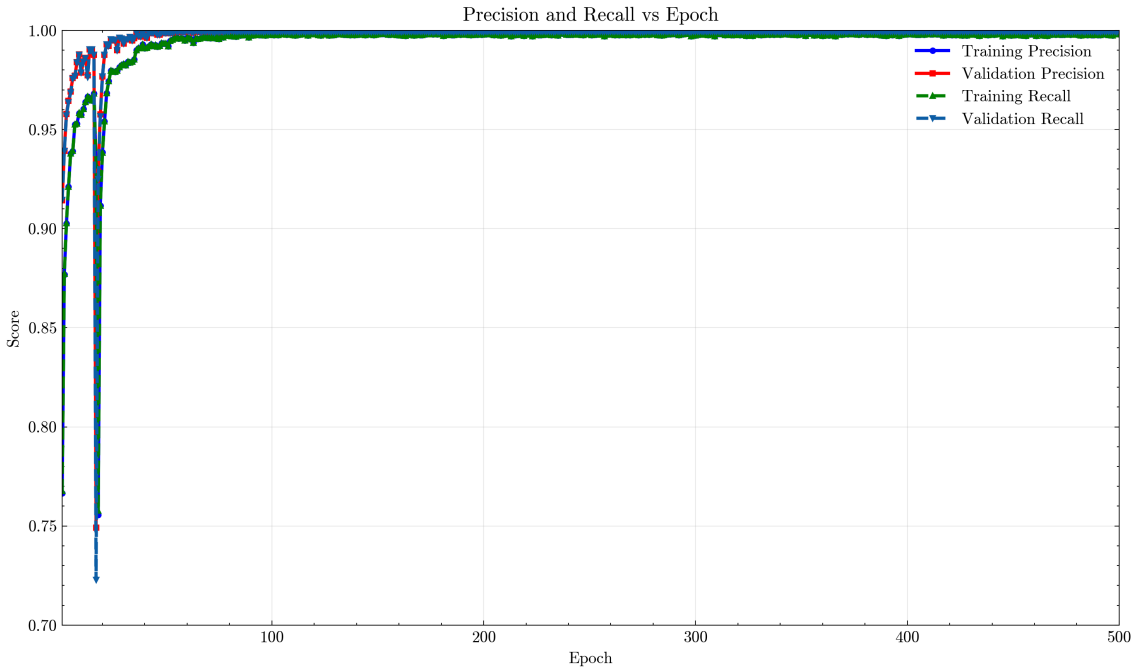

Figure 4.3: Brain MRI ND5 Dataset Training (Precision & Recall Vs Epoch)

For the Brain Tumor MRI dataset, the network cross-entropy loss is low for training and slightly higher for validation in Figure 4.4. In the first ten to fifteen epochs, both losses drop sharply as the model learns to distinguish the tumor from healthy tissue. The F1 curve follows closely with accuracy, exceeding 0.99 early and nearing 0.999 at the end of the training in Figure 4.5. The training and validation lines are nearly identical, indicating a consistent balance between false positives and negatives on new data. Precision and recall initially differ slightly, with precision lagging by a fraction, but converge in 30 epochs, stabilizing around 0.999 in Figure 4.6. This indicates that the classifier rarely mislabels healthy tissue as tumors and seldom misses real tumors. The validation metrics closely align with the training metrics, demonstrating the model's ability to generalize to unseen images without degradation.

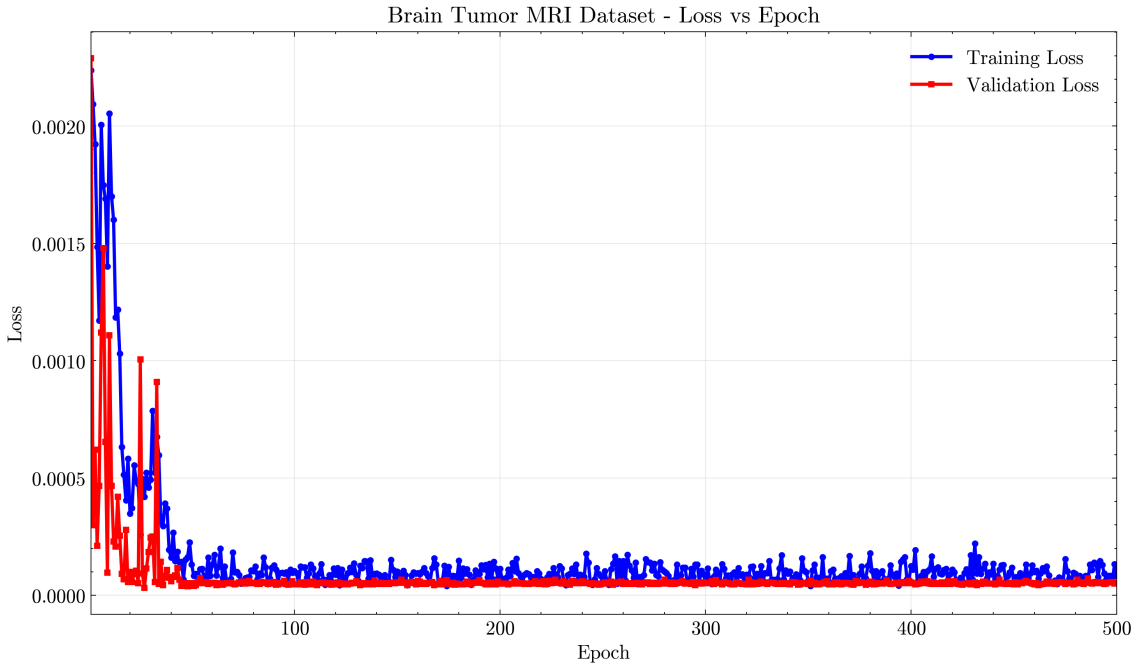

Figure 4.4: Brain Tumor MRI Dataset Training (Loss Vs Epoch)

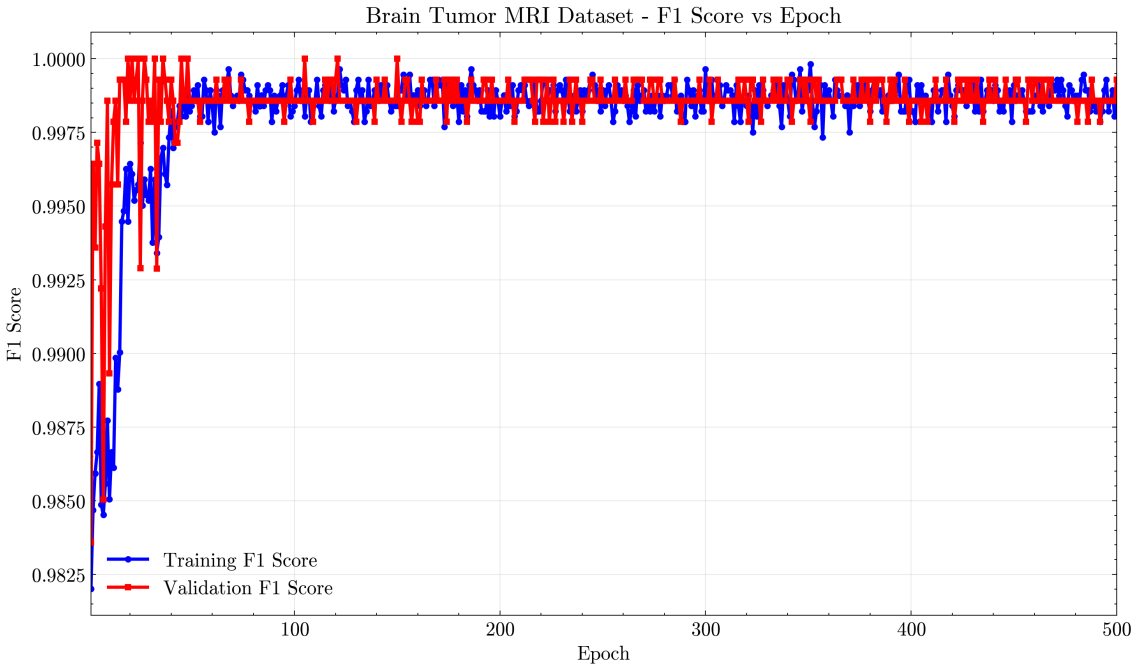

Figure 4.5: Brain Tumor MRI Dataset Training (F1 Score Vs Epoch)

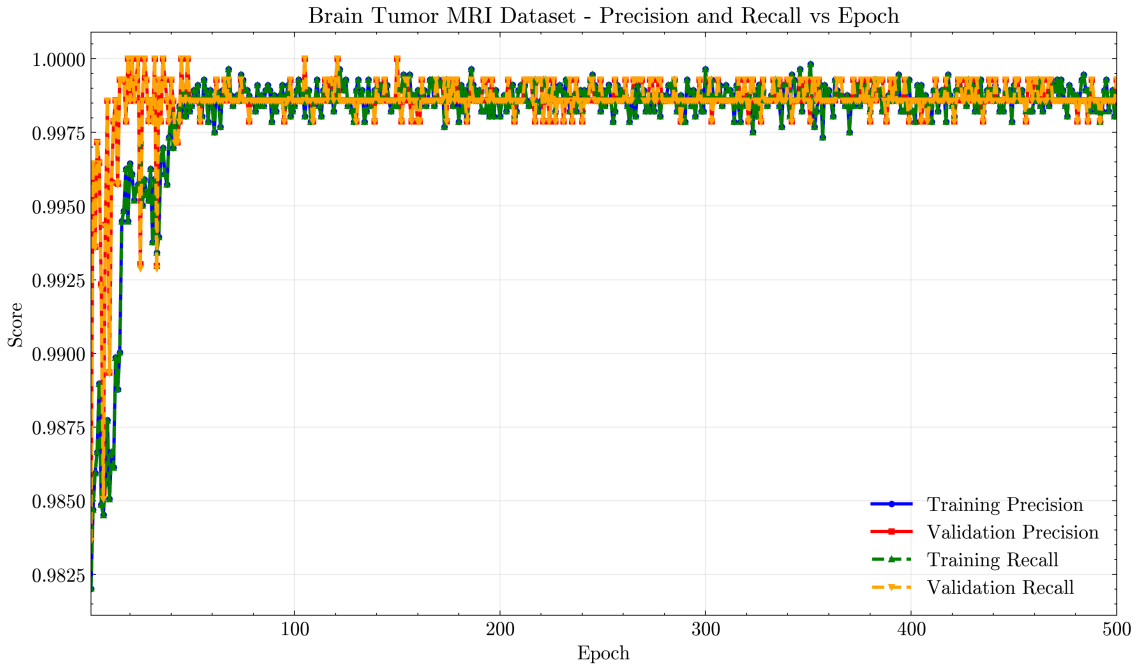

Figure 4.6: Brain Tumor MRI Dataset Training (Precision & Recall Vs Epoch)

During the first few epochs of the Pancreas CT dataset, the training loss decreases dramatically, while the validation loss decreases even more dramatically in Figure 4.7. After this initial learning spike, both losses continue to decline steadily, with the validation loss staying just below the training loss. The F1 score in Figure 4.8 shows that high accuracy does not mask class imbalance problems. The training F1 score increases from 0.70 to over 0.98 and approaches 0.999, while the validation F1 curve reaches 1.0 by epoch 10 and remains at this level, aligning with the validation accuracy. The overlap of the curves suggests that the classifier maintains a consistent balance between false positives and false negatives with unseen scans. Initially, precision (0.68) lags behind recall (0.73) in the training set, but both metrics converge to more than 0.99 within the first dozen epochs in Figure 4.9. From Epoch 10 onward, validation precision and recall reach 1.0, while training values approach 1.0 by the early 40th epoch. The convergence indicates that the network effectively balances precision and recall, rarely mislabeling healthy slices as diseased, and rarely missing actual tumors.

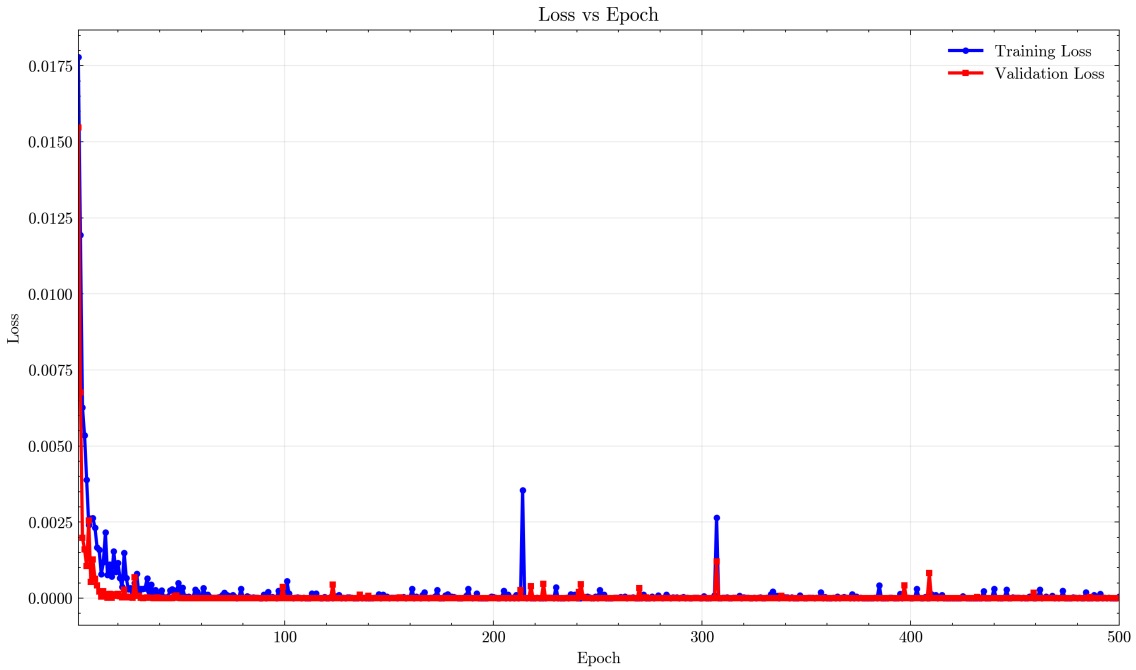

Figure 4.7: Pancreas CT Dataset Training (Loss Vs Epoch)

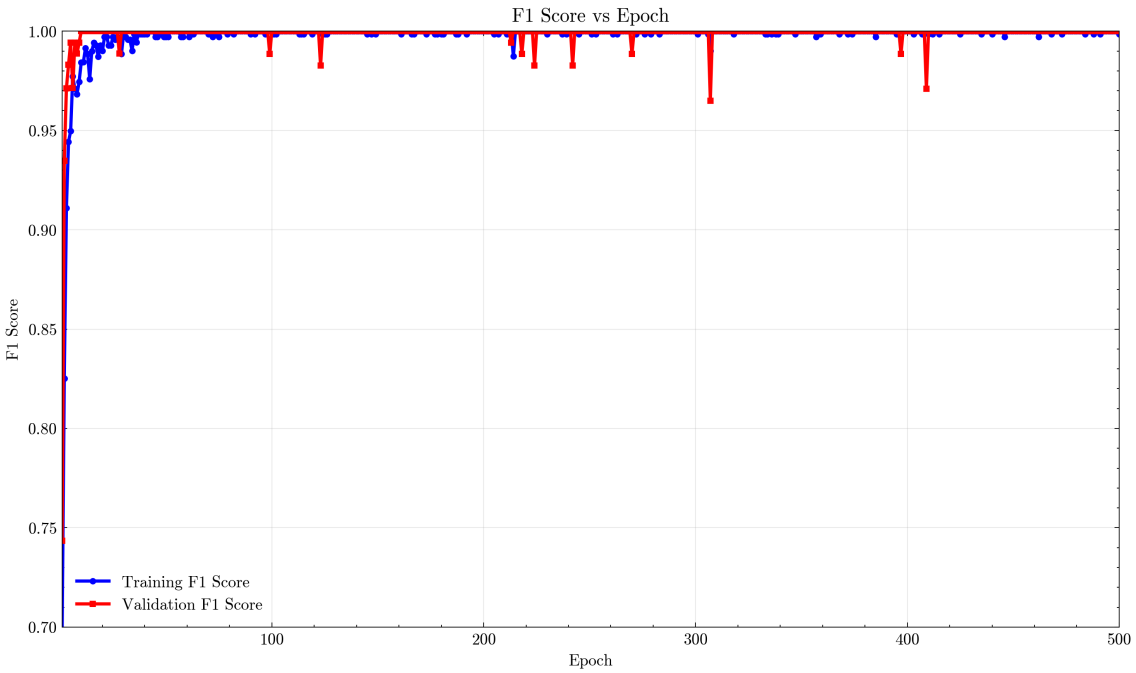

Figure 4.8: Pancreas CT Dataset Training (F1 Score Vs Epoch)

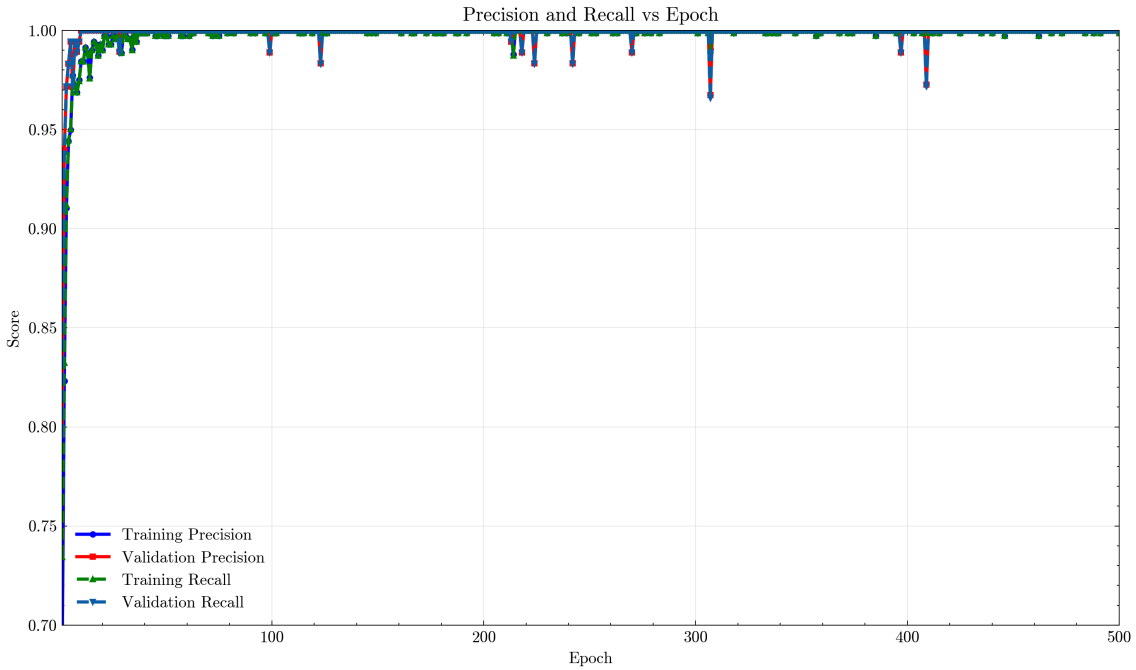

Figure 4.9: Pancreas CT Dataset Training (Precision & Recall Vs Epoch)

Regarding the Mammogram Mastery dataset, training and validation curves both decrease significantly in the first ten to fifteen epochs, indicating that the model quickly learns the discriminative characteristics of mammography in Figure 4.10. After 30 epochs, both lines stabilize, with validation loss remaining lower than training loss. F1 harmonizes precision and recall, reflecting the prediction reliability for both classes in Figure 4.11. The training F1 score starts at 0.79 and steadily climbs, catching up to the validation curve (beginning at 0.93) around epoch 35 and reaching approximately 0.993 at the end. The validation F1 score reaches 1.0 by epoch 32, indicating an optimal balance between false positives and negatives on unseen images. Initially, precision is slightly lower than recall in the training set (0.78 vs 0.81), while both metrics are high in validation, around 0.93 in Figure 4.12. By the end of thirty epochs, precision and recall, along with their validation counterparts, exceed 0.99. From Epoch 32 onward, the validation metrics reach 1.0, and the training metrics remain very close, indicating that the classifier is highly calibrated with high precision and recall. The overlap between training and validation traces confirms that this balance generalizes well to new data.

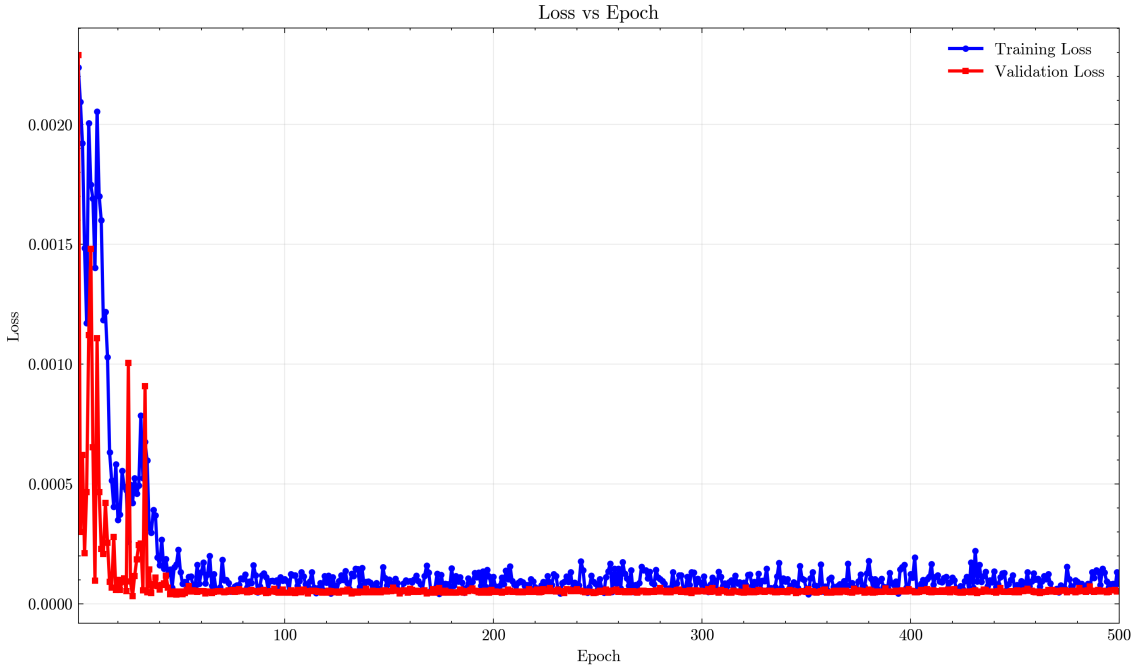

Figure 4.10: Mammogram Mastery Dataset Training (Loss Vs Epoch)

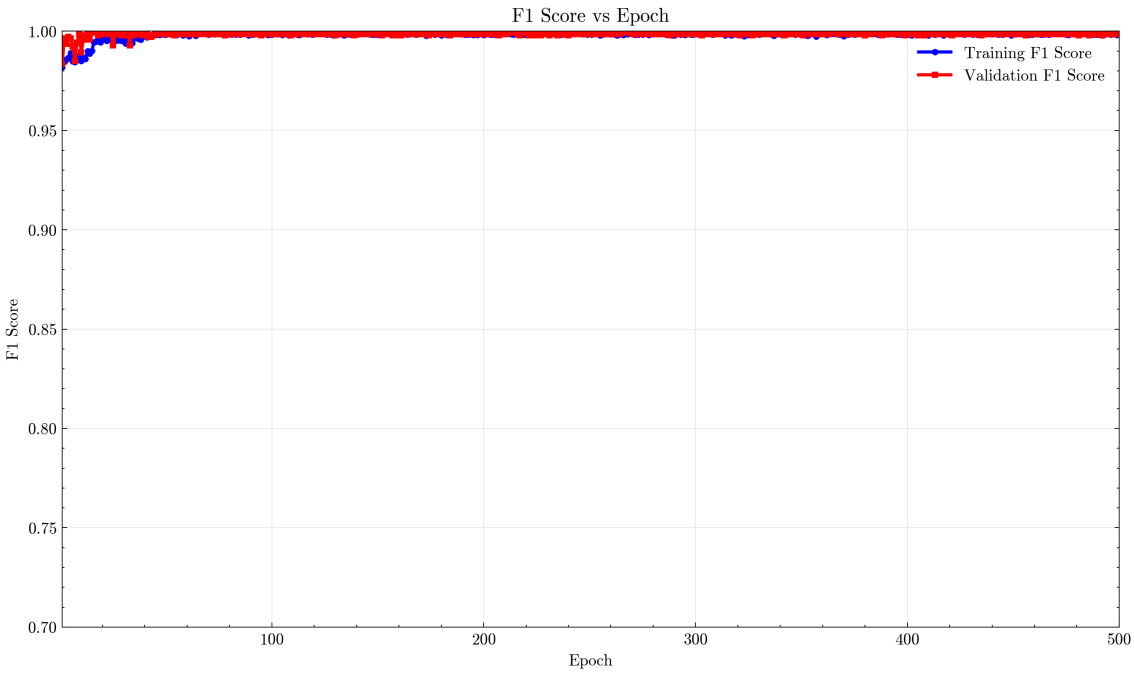

Figure 4.11: Mammogram Mastery Dataset Training (F1 Score Vs Epoch)

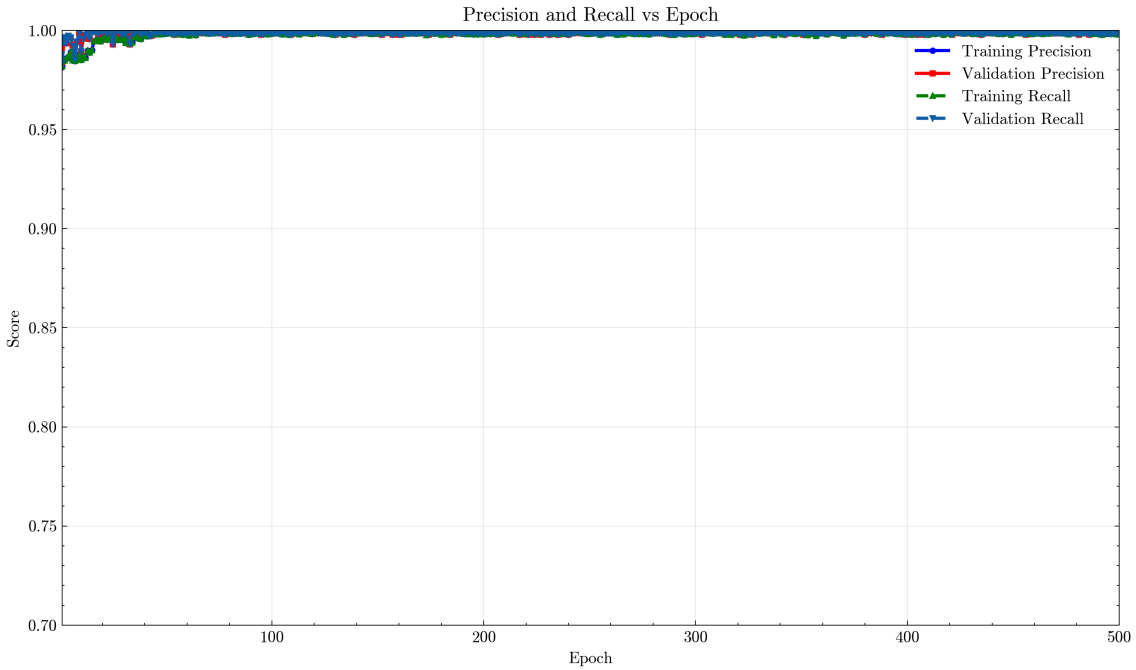

Figure 4.12: Mammogram Mastery Dataset Training (Precision & Recall Vs Epoch)

4.5.2 Segmentation Datasets Training Performance

The MedSeg-COVID collection features thin-section chest CT slices annotated for COVID-19 pneumonia, characterized by a severe class imbalance. Despite achieving a categorical accuracy of nearly 0.98, the performance of segmentation-specific models plateaus significantly earlier than in other applications. In the loss versus epoch graph, the training loss decreases from approximately 0.025 to approximately 0.016 by epoch 30. In contrast, the validation loss drops to a minimum of 0.0186 at epoch 7 before stabilizing around 0.022 in Figure 4.13. This indicates that optimization saturates after epoch 10, with the subsequent epochs reinforcing existing weights. The mean Dice score curve shows a peak validation Dice of 0.496 at epoch 17, oscillating afterward and closing at 0.495 by epoch 500, indicating no overfitting, but a limit imposed by the dataset's class imbalance. The mean IoU plot reflects similar trends, finishing at 0.491 in Figure 4.14. In general, the results demonstrate an initial rapid improvement, followed by convergence at epoch 30 in Figure 4.15. Achieving higher performance beyond this will require modifications to loss functions, enhanced augmentation, or a different architectural approach focusing on the small infected regions.

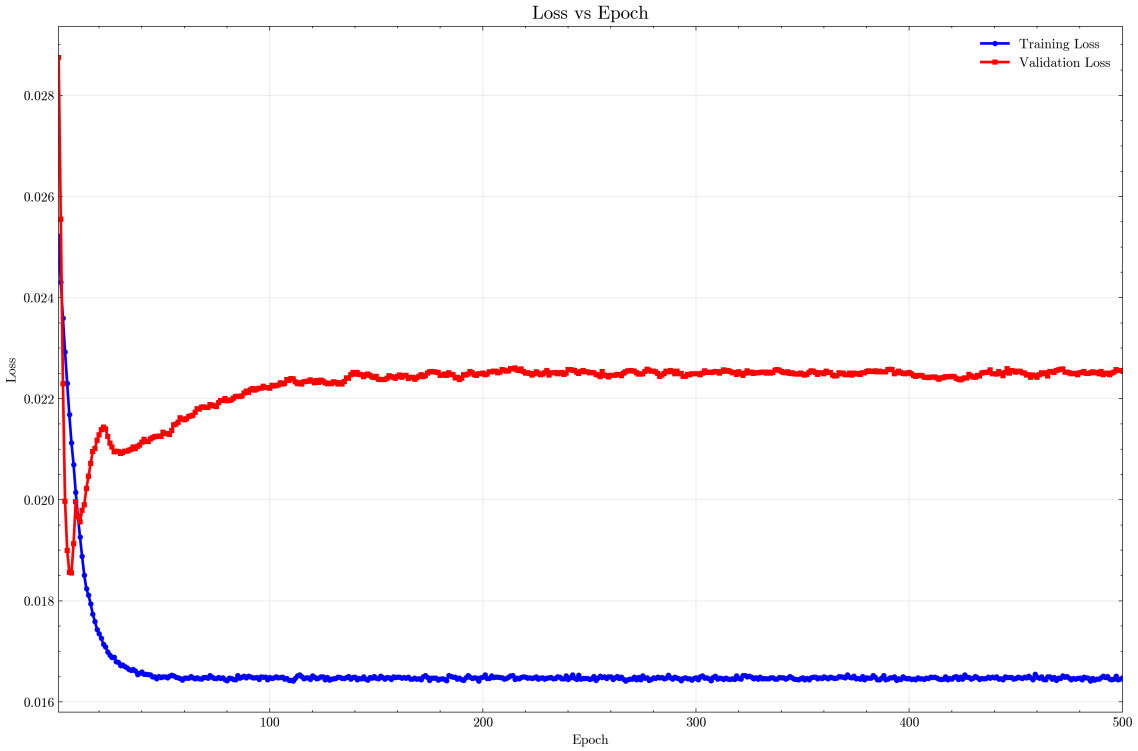

Figure 4.13: MedSeg-Covid Dataset Training (Loss Vs Epoch)

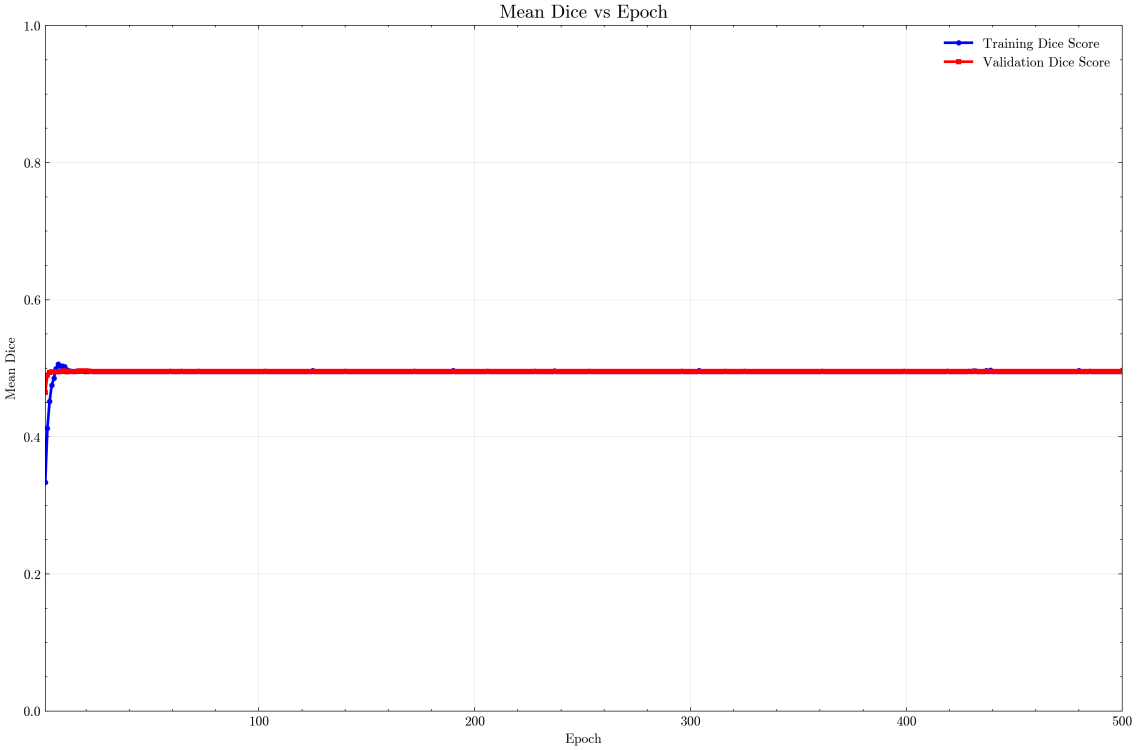

Figure 4.14: MedSeg-Covid Dataset Training (Dice Score Vs Epoch)

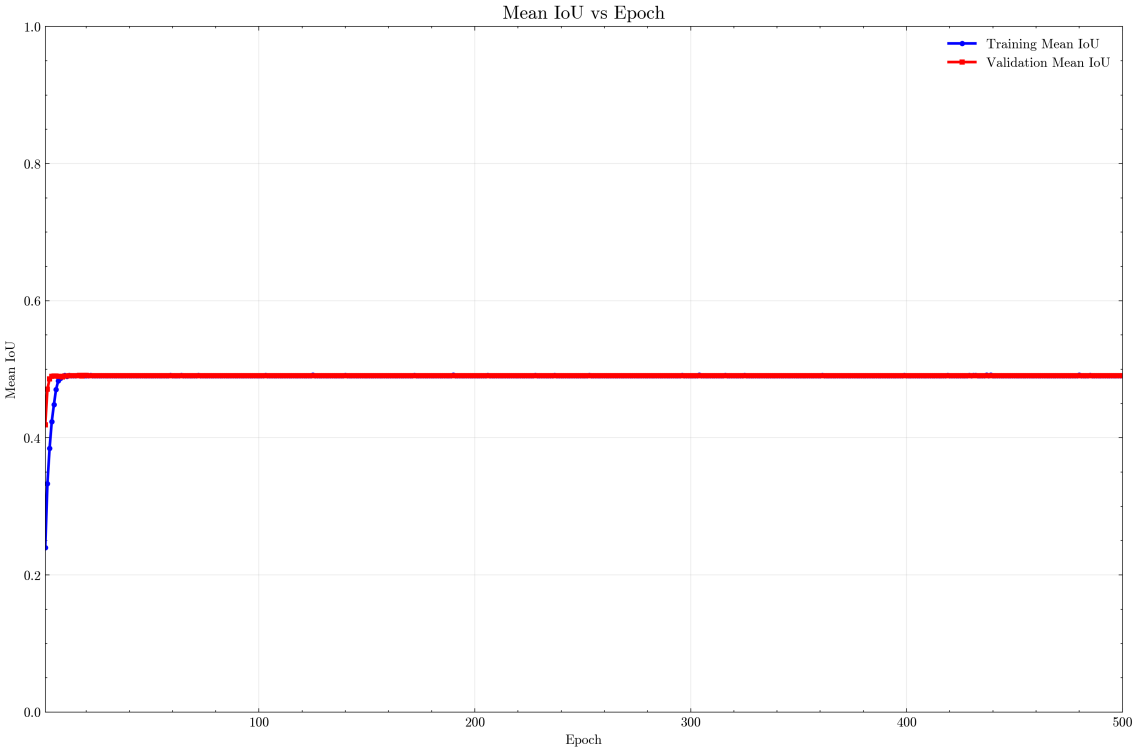

Figure 4.15: MedSeg-Covid Dataset Training (IoU Vs Epoch)

The MedSeg-Liver collection offers thin-section abdominal CT volumes with pixel-level masks that outline the liver and, in some cases, tumors. The model demonstrates strong performance despite boundary ambiguities, showing steady improvement and a prolonged period of stable refinement, as indicated by nearly overlapping training and validation curves. In the loss versus epoch trace, the training loss decreases from 0.020 to 0.0039 by epoch 190 in Figure 4.16, while the validation loss reaches a minimum of 0.006 by epoch 492. Both curves flatten past epoch 50, suggesting the optimizer converged early, indicating room for an early-stopping criterion based on validation loss. The mean Dice score curve increases rapidly, jumping from 0.49 to 0.91 in the first twenty epochs, reaching 0.968 by epoch 112, and ending at 0.964 by epoch 500 in Figure 4.14. This indicates that the overfitting is minimal and that the model aligns with expert tracings on more than 96% of the liver pixels. The mean IoU trajectory follows a similar pattern, but is about three points lower, reaching a peak of 0.94 by epoch 112 and settling at 0.932 4.18. The small gap between IoU training and validation highlights the stability of the model. Overall, the three plots demonstrate that the network effectively captures both organ topology and edge detail from the outset, with consistent performance thereafter. A checkpoint around epoch 120 would capture optimal accuracy while reducing training time, with further enhancements likely dependent on loss functions or post-processing rather than extended training.

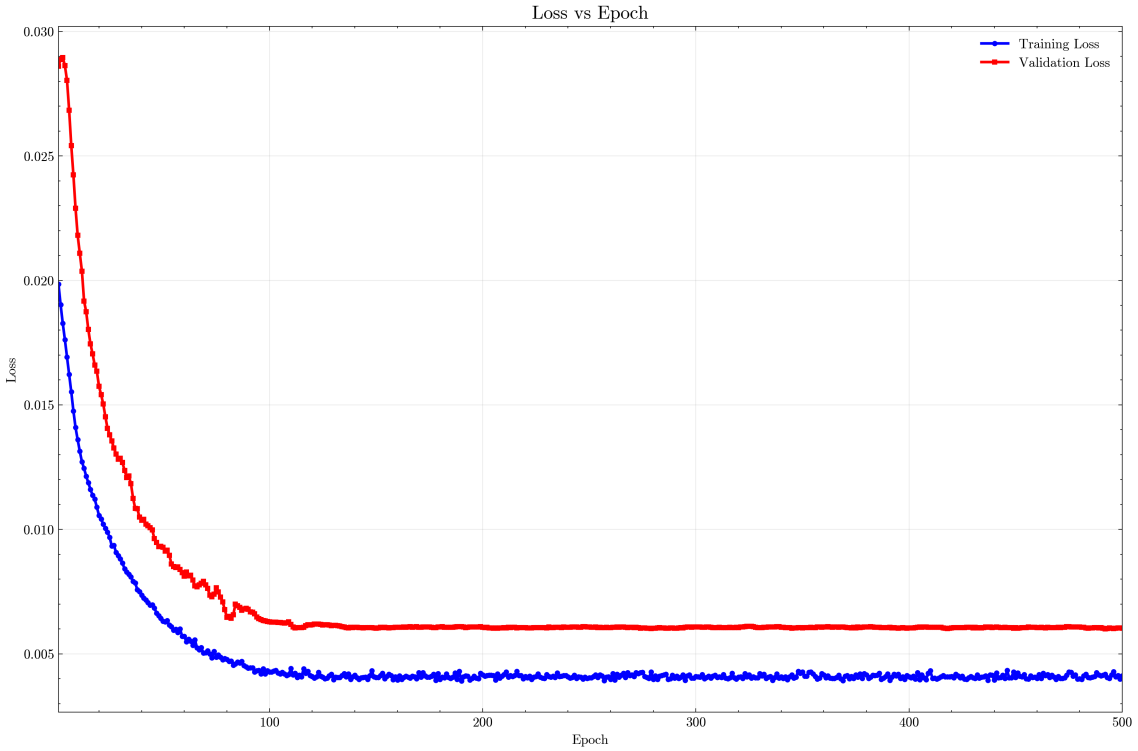

Figure 4.16: MedSeg-Liver Dataset Training (Loss Vs Epoch)

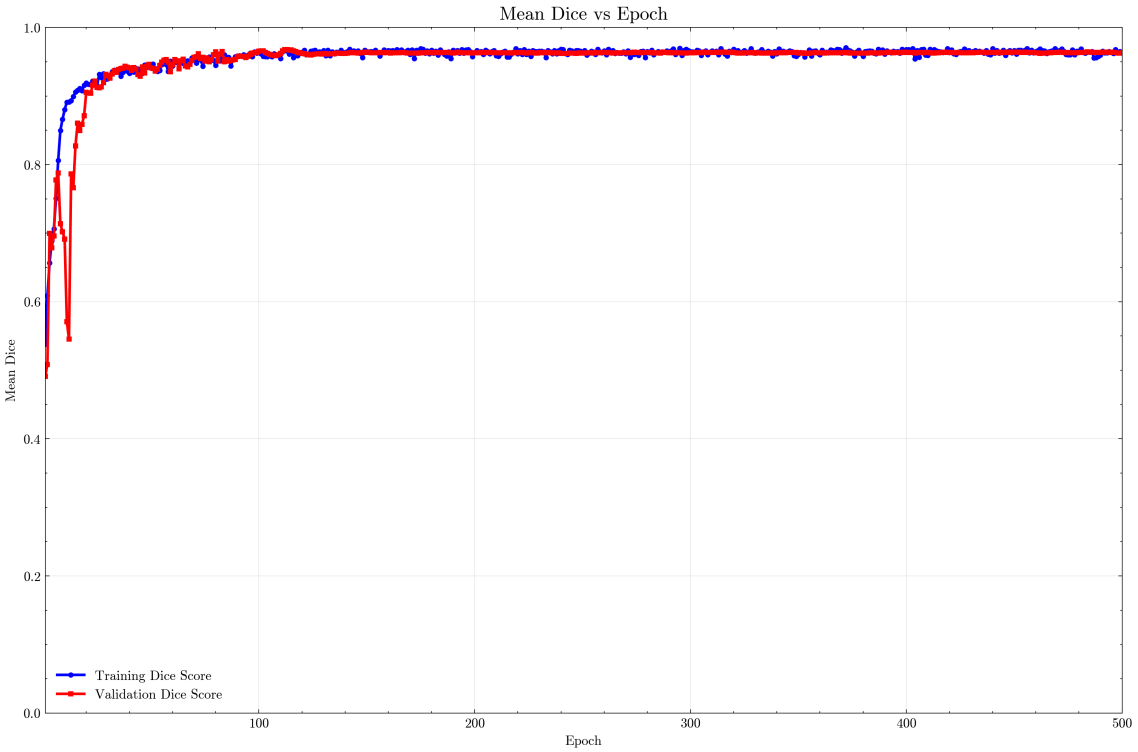

Figure 4.17: MedSeg-Liver Dataset Training (Dice Score Vs Epoch)

Figure 4.18: MedSeg-Liver Dataset Training (IoU Vs Epoch)

The Kvasir-SEG collection comprises 1000 gastrointestinal endoscopy frames paired with precise polyp masks, presenting a challenge in detecting small objects amidst clutter. The 500-epoch training shows a steady optimization process, with loss decreasing from 0.031 to approximately 0.0046, as shown in Figure 4.19. This indicates that an early stop checkpoint could potentially save computational costs while maintaining accuracy. The mean Dice score of the validation improves from 0.49 to 0.79 in the first two epochs, reaches 0.90 in epoch 40, and peaks at 0.927 in epoch 325 in Figure 4.20. The training mean Dice is slightly higher at 0.947, suggesting only mild overfitting that can be addressed with augmentation or dropout techniques. The mean IoU follows a similar trend, starting at 0.33 and reaching 0.8687 at epoch 325, settling at 0.8668 at the last epoch in Figure 4.21. This performance demonstrates strong generalization capabilities, positioning the model among the top methods for polyp segmentation. In summary, the learning curves show rapid convergence, consistent loss reduction, and only a modest training-validation gap. Checking the model between epochs 130 and 150 would preserve the top performance while significantly reducing training time, with further improvements likely to be achieved by tuning loss functions or refining augmentations.

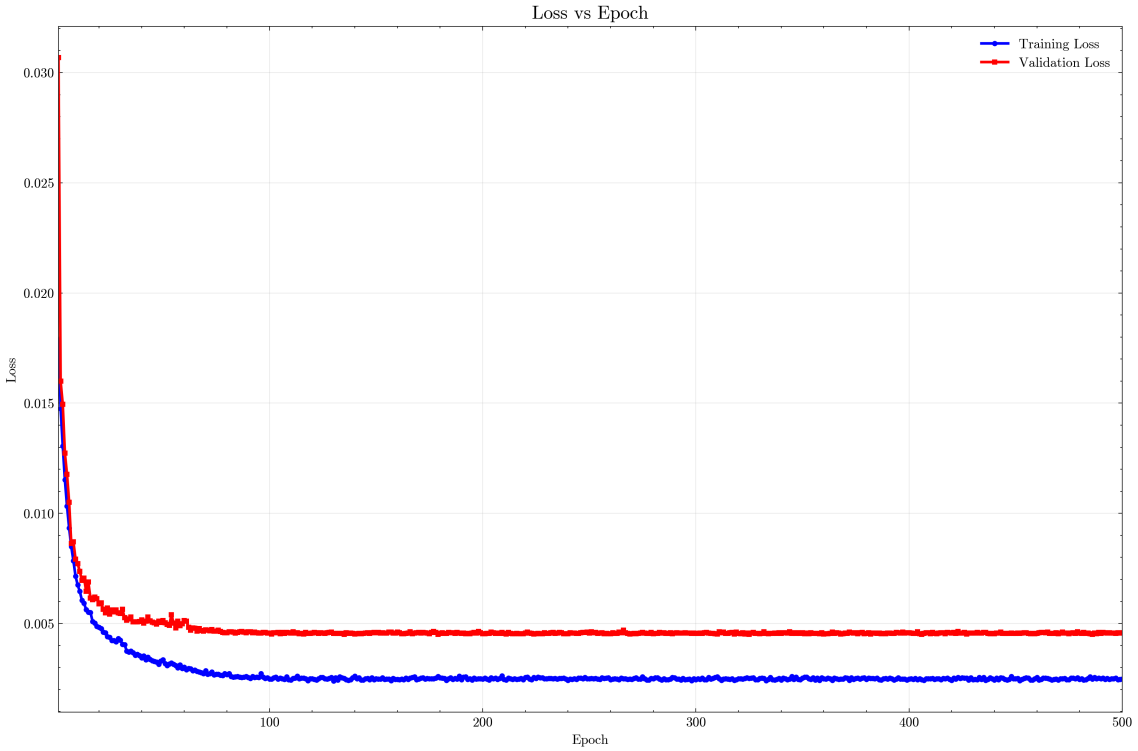

Figure 4.19: Kvasir-SEG Dataset Training (Loss Vs Epoch)

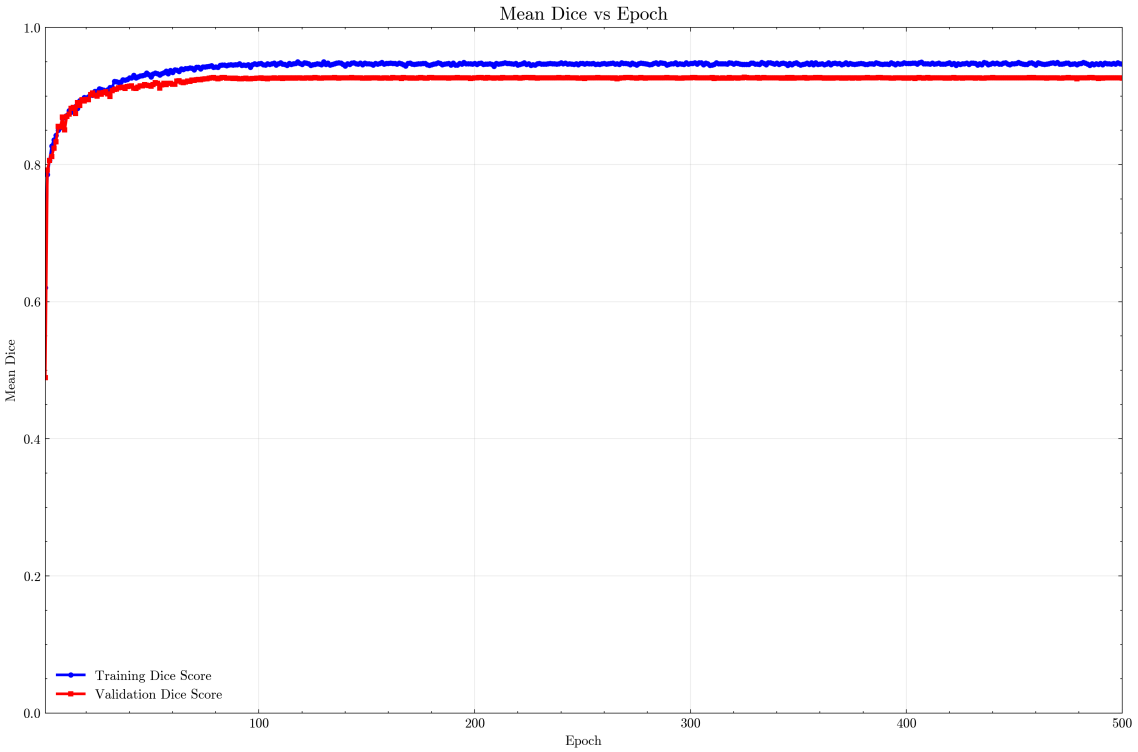

Figure 4.20: Kvasir-SEG Dataset Training (Dice Score Vs Epoch)

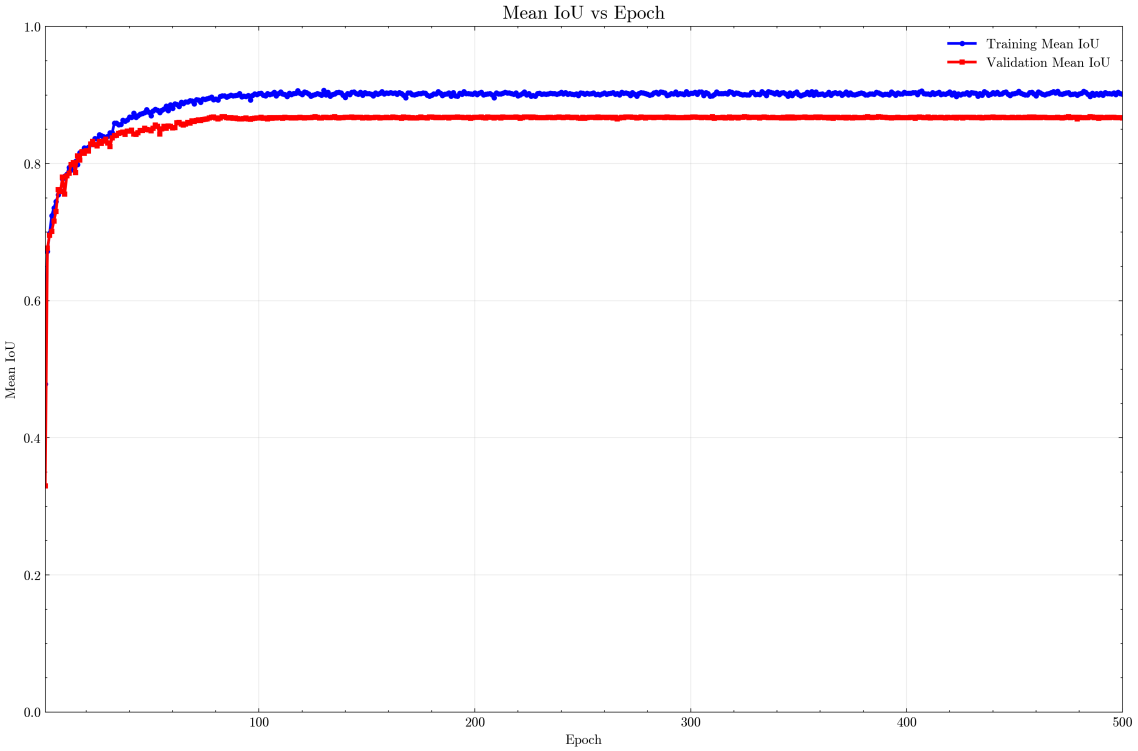

Figure 4.21: Kvasir-SEG Dataset Training (IoU Vs Epoch)

The Brain Stroke CT Dataset comprises non-contrast head CT slices with pixel-level masks for ischemic and hemorrhagic lesions, presenting challenges such as class imbalance and visual variability. Training over 500 epochs demonstrated effective optimization, with training and validation metrics closely tracked, and validation metrics eventually surpassing those of training. The loss curve in Figure 4.22 begins around 0.012, dropping to approximately 0.007 within 30 epochs as the model learns the structure of the brain background. It continues to decrease to about 0.002 by epoch 100. At the same time, the validation loss remains slightly lower than the training loss, indicating that early stopping between epochs 100 and 120 could preserve accuracy and reduce computational costs. The mean Dice score increases from 0.50 to 0.824 in Epoch 87, ending at 0.815 in Epoch 500 in Figure 4.23, with the mean Dice of validation consistently exceeding the Dice of training by approximately 0.026. This indicates that data augmentation improves performance, with a final mean Dice score of 0.82 showing agreement with the expert annotations 80%. The mean IoU score follows a similar trend, climbing to 0.739 at epoch 87 and ending at 0.729 in Figure 4.24. In general, optimization demonstrates rapid learning, stability, and strong Dice and IoU scores, suggesting effective capture of the lesion details without overfitting. A checkpoint around epochs 90-120 may be beneficial for further deployment or experimentation.

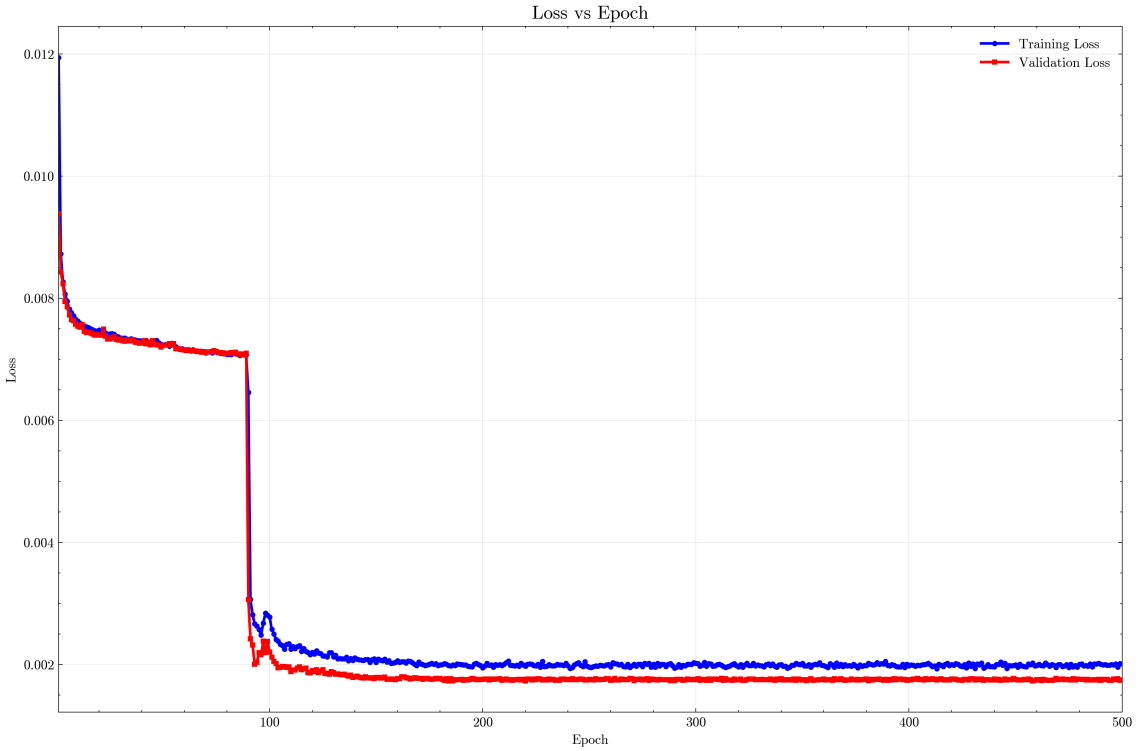

Figure 4.22: The Brain Stroke CT Dataset Training (Loss Vs Epoch)

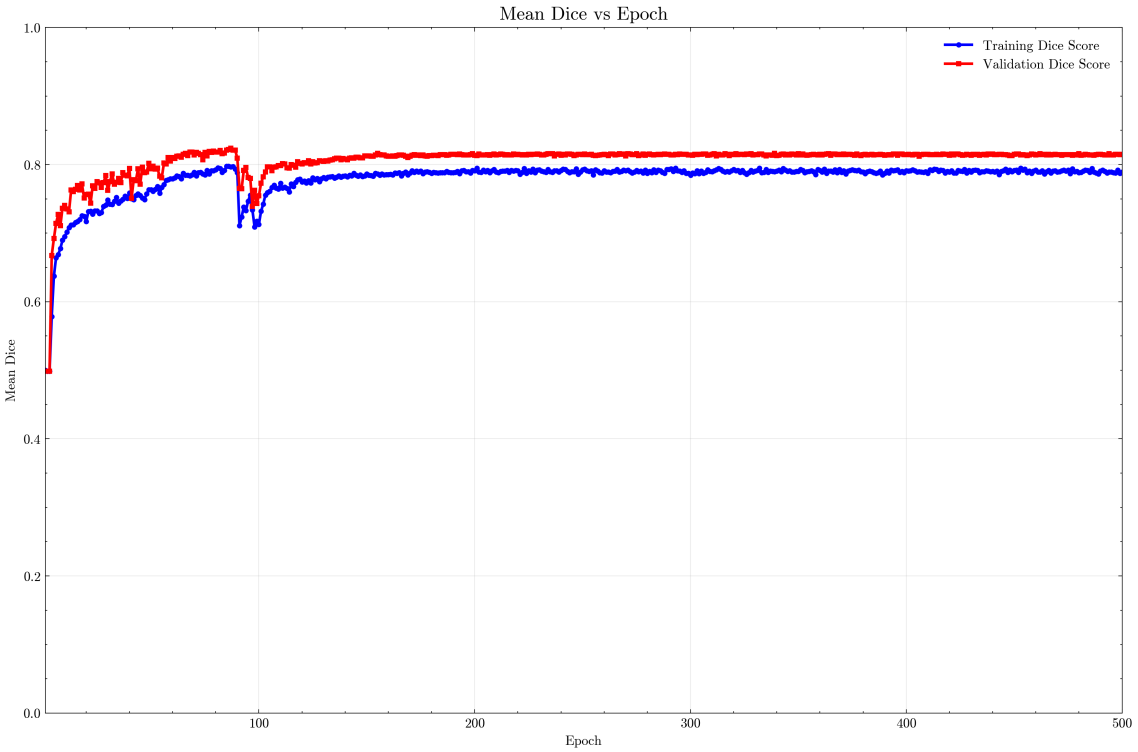

Figure 4.23: The Brain Stroke CT Dataset Training (Dice Score Vs Epoch)

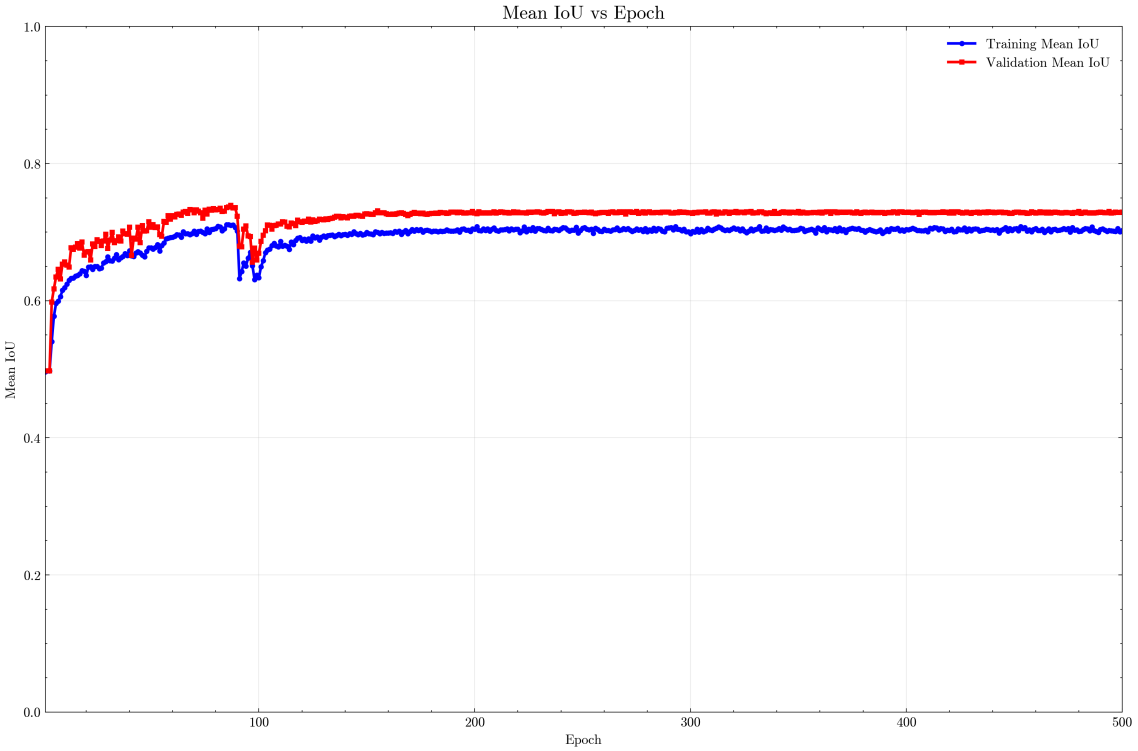

Figure 4.24: The Brain Stroke CT Dataset Training (IoU Vs Epoch)

The validation performance of the proposed MedFoundX model demonstrates exceptional effectiveness in both classification and segmentation tasks across a wide range of biomedical imaging datasets. As shown in Table 4.1, the model achieved near-perfect classification metrics in all four datasets. Specifically, it reached an accuracy of 0.9994 in the Brain MRI ND-5 dataset, 0.9993 on the Brain Tumor MRI dataset, and a perfect score of 1.0000 on both the Pancreas CT and Mammogram Mastery datasets. The corresponding precision, recall, and F1 scores mirror these results, indicating an outstanding ability to correctly identify pathological features with no significant trade-offs between sensitivity and specificity. The minimal loss values are also close to zero, further affirming the convergence of the model and the strong generalizability during training.

Table 4.1: Validation Metrics on Four Classification Datasets.

Dataset	Loss	Accuracy	Precision	Recall	F1-score
Brain MRI ND-5	0.0001	0.9994	0.9994	0.9994	0.9994
Brain Tumor MRI	0.0001	0.9993	0.9993	0.9993	0.9993
Pancreas CT	0.0001	1.0000	1.0000	1.0000	1.0000
Mammogram Mastery	0.0004	1.0000	1.0000	1.0000	1.0000

In parallel, the segmentation performance, summarized in Table 4.2, highlights the robustness and adaptability of the model in four distinct datasets. In particular, the model achieved a mean Dice score of 0.9638 and a mean IoU of 0.9322 on the MedSeg Liver CT dataset, as well as high scores on the Kvasir-SEG (Dice: 0.9262, IoU: 0.8668) datasets. Even in more challenging tasks, such as segmenting brain strokes, the model maintained a high accuracy of 0.9964 with a respectable

mean Dice score of 0.8152. Although the MedSeg COVID dataset presented more difficulty, as reflected by a lower mean Dice score (0.4953) and mean IoU (0.4907), the model still achieved a high accuracy of 0.9814, showcasing its resilience even under domain-specific variance and potential class imbalance.

Table 4.2: Validation Metrics on Four Segmentation Datasets.

Dataset	Loss	Accuracy	Mean IoU	Mean Dice
MedSeg COVID CT	0.0226	0.9814	0.4907	0.4953
MedSeg Liver CT	0.0060	0.9950	0.9322	0.9638
Kvasir-SEG	0.0046	0.9612	0.8668	0.9262
Brain Stroke	0.0017	0.9964	0.7293	0.8152

Overall, these results collectively validate the efficacy of the MedFoundX architecture in learning discriminative and generalizable features across diverse imaging modalities and tasks. Low loss values, consistently high accuracy, and superior overlap metrics (Dice and IoU) strongly indicate that the training process was highly stable, well-regularized, and effective in both the classification and segmentation pipelines. This level of performance underscores the readiness of the model for clinical translation and its potential to serve as a unified solution in real-world biomedical image analysis.

4.6 Performance Analysis of Unseen Datasets Through Finetuning

After completing multi-dataset pre-training, we consolidated the learned parameters into a single set of weights that constitute our foundation model, MedFoundX. This model now serves as a highly transferable initialization that can be rapidly fine-tuned for new biomedical imaging tasks and deployed for inference. To showcase its adaptability, we will fine-tune the model on two representative benchmarks: the PMRAM: Bangladeshi Brain-Cancer MRI Dataset for image classification and the CVC-ClinicDB Dataset for image segmentation.

4.6.1 Finetune for Classification: PMRAM Dataset

The loss curves begin in the range of 0.044 for training and 0.043 for validation, indicating a reasonable initial fit in Figure 4.25. Within a few dozen epochs, losses drop below 0.01 as the model learns the key patterns. By epoch 150, the validation loss stabilizes slightly below the training loss. The F1 score follows suit, initially at 0.27 and rising rapidly as class balance improves. In mid-training, it stabilizes, reaching approximately 0.97 in the training set and 0.99 on the validation set, indicating an improvement in both the hit rate and the balance between false positives and negatives in Figure 4.26. The precision and recall curves closely track each other, starting low due to class imbalance but quickly rising into the 0.90 range. They stabilize around 0.97 for training and converge at 0.997 for validation in Figure 4.27, indicating that the model neither systematically misses positives nor overcalls negatives. This alignment, together with a strong F1 score and accuracy

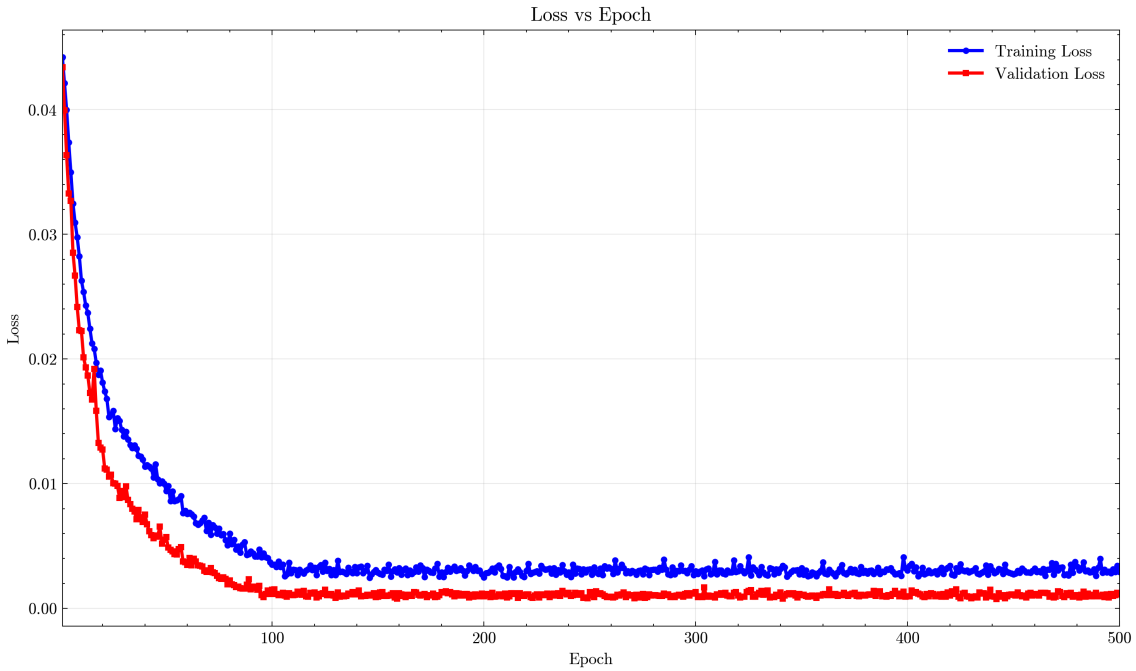

Figure 4.25: PMRAM Dataset Finetuning (Loss Vs Epoch)

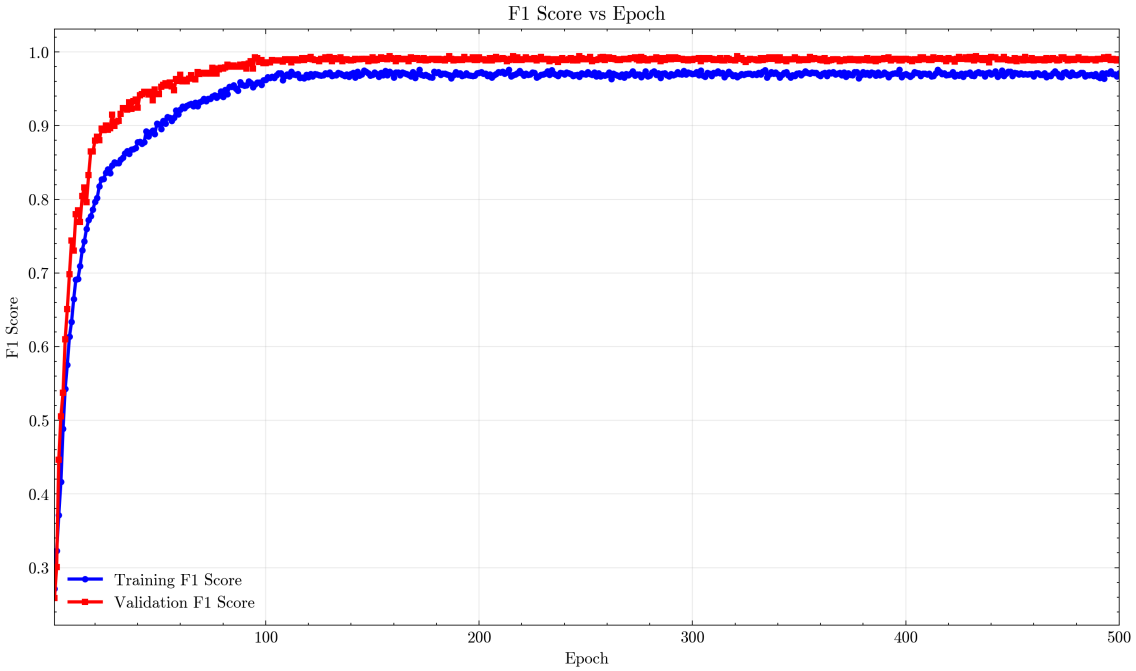

Figure 4.26: PMRAM Dataset Finetuning (F1 Score Vs Epoch)

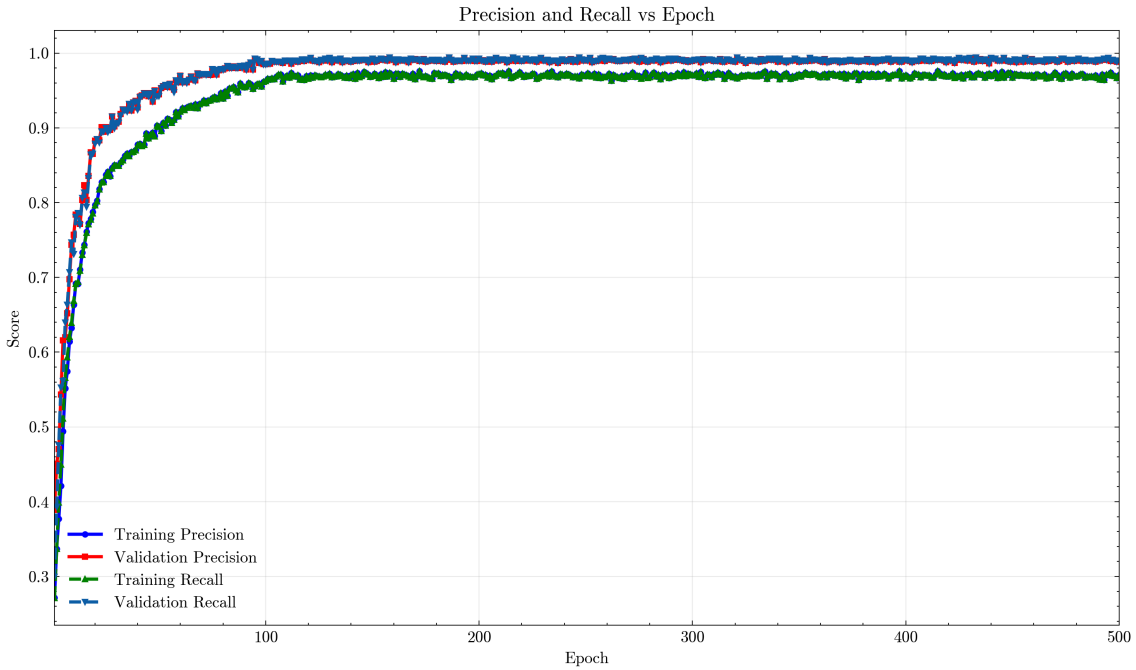

Figure 4.27: PMRAM Dataset Finetuning (Precision & Recall Vs Epoch)

trends, confirms that the final model achieves robust and clinically meaningful performance without sacrificing generalization. The finetuning performance metrics for the PMRAM Dataset are shown in Table 4.3.

Table 4.3: Finetune Metrics for Classification: PMRAM Dataset

Metric	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Overall Dataset				
All Classes	0.99	0.99	0.99	0.99
Class-wise Metrics				
Glioma		0.97	0.99	0.98
Meningioma		0.99	0.98	0.98
Normal		0.99	0.99	0.99
Pituitary		1.0	0.99	0.99

4.6.2 Finetune for Segmentation: CVC-ClinicDB Dataset

The loss curves begin with a gap between the training and validation losses, approximately 0.018 and 0.023 in the first epoch, indicating that the model initially struggled with unseen data. Over the following 40 epochs, both losses decline sharply and nearly coincide around epoch 60. After around epoch 120, the validation loss dips slightly below the training loss, reaching a low point of approximately 0.00234 around epoch 493, as shown in Figure 4.28. This overlap, combined with a smooth descent, suggests stable convergence with minimal overfitting. The mean Dice score, shown in Figure 4.29, starts at approximately 0.29 for validation and 0.58

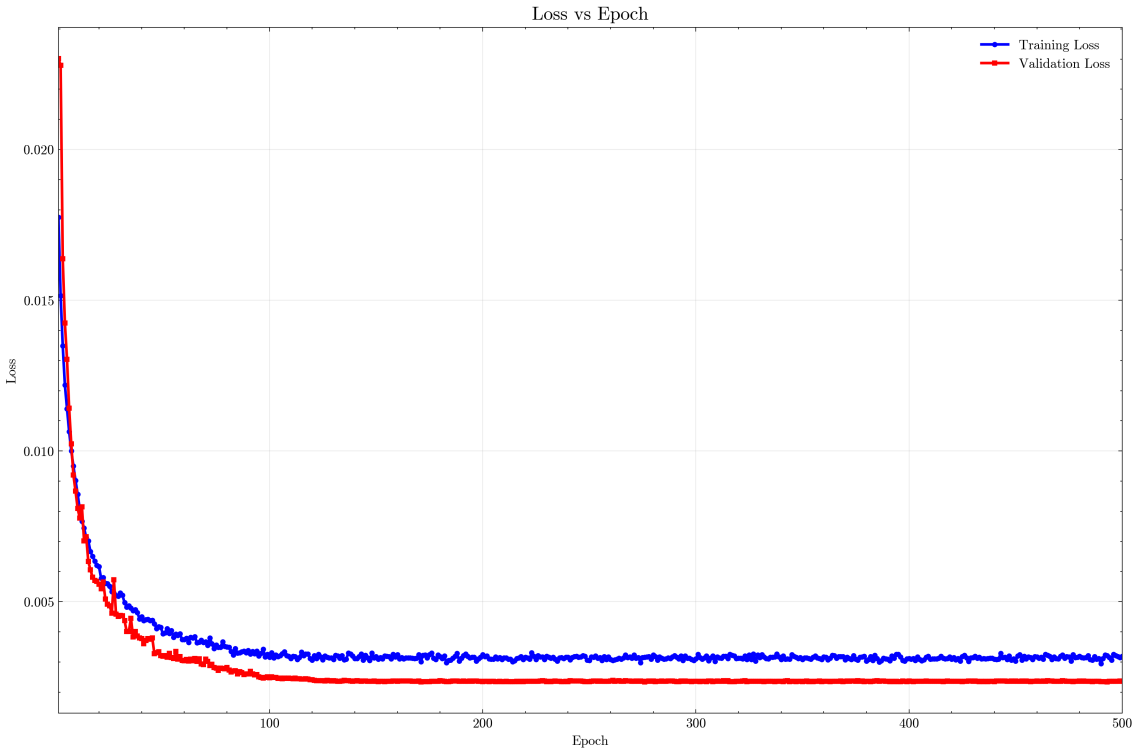

Figure 4.28: CVC-ClinicDB Dataset Finetuning (Loss Vs Epoch)

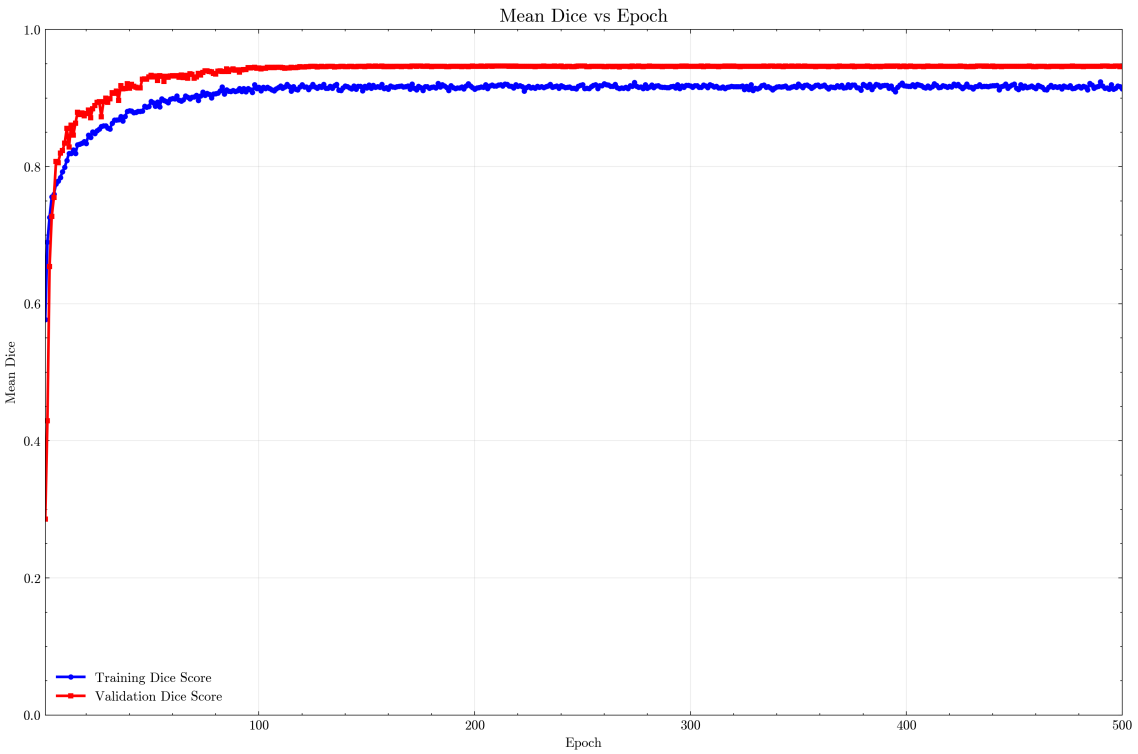

Figure 4.29: CVC-ClinicDB Dataset Finetuning (Dice Score Vs Epoch)

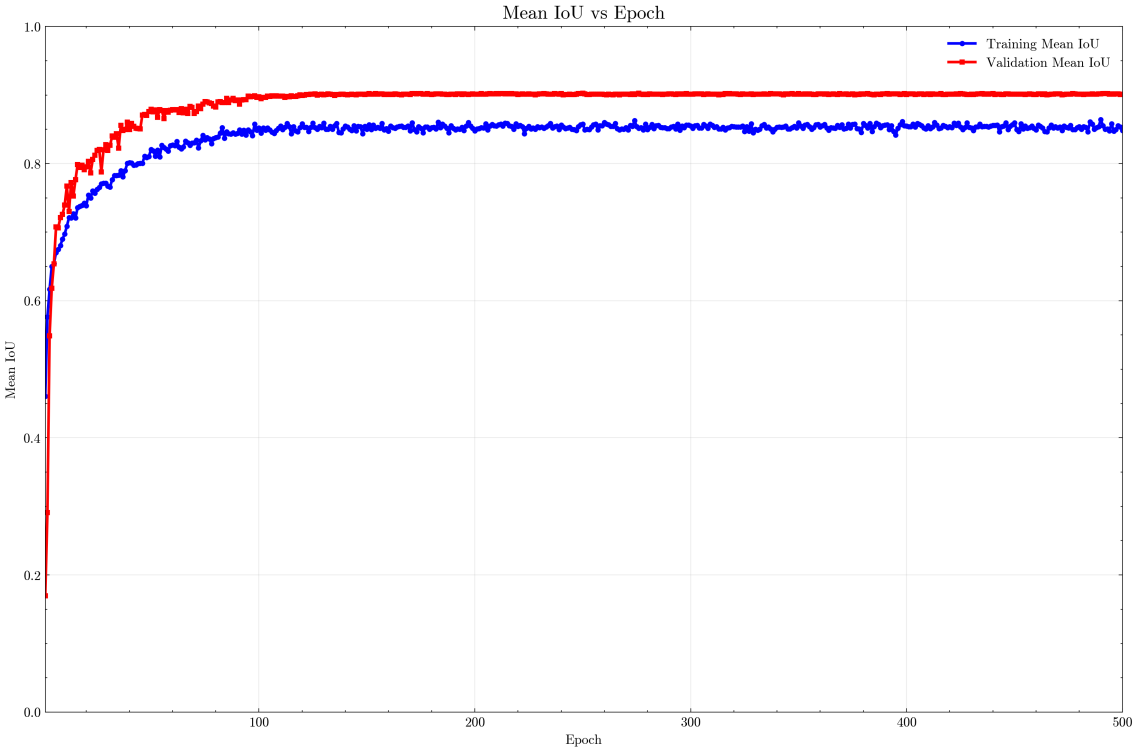

Figure 4.30: CVC-ClinicDB Dataset Finetuning (IoU Vs Epoch)

for training, increasing to over 0.90 after about 150 epochs, peaking at around 0.947 on the validation set by epoch 250. The validation score remains slightly higher than the training score, indicating that fine-tuning has improved boundary details and improved minority class predictions. The mean IoU follows a similar pattern, starting at approximately 0.17 for validation and 0.46 for training, and rising to 0.90 and 0.85, respectively, by the final epochs in Figure 4.30. Most minor errors appear in borderline regions, affirming the model's ability to generalize well to unseen scans. In summary, the plots illustrate rapid initial learning, ongoing refinement, and a final state where training and validation performances are nearly indistinguishable. The performance metrics for finetuning on the CVC-ClinicDB dataset are shown in Table 4.4.

Table 4.4: Finetune Metrics for Segmentation: CVC-ClinicDB Dataset

Metric	**Value**
Loss	0.002
Pixel Accuracy	0.98
Mean IoU	0.90
Mean Dice	0.95

4.7 Inference Analysis Using Test Datasets

This section provides a high-level overview of how the fine-tuned MedFoundX performs on unseen data. It first outlines classification inference on a brain-tumor

MRI test set. It then summarizes the segmentation inference on a colonoscopy dataset, describing the overlap and pixel-level metrics we track, as well as the insights they provide into boundary quality and clinical usability. Together, the two case studies demonstrate how MedFoundX's multi-dataset pre-training is effectively transferred to real-world tasks after domain-specific fine-tuning.

4.7.1 Inference for Classification: PMRAM Dataset

Inference performed on the PMRAM test set shows that the fine-tuned MedFoundX classifier generalizes almost perfectly to unseen data, as shown in Table 4.5. Out of 400 MRI volumes, the network correctly labeled 394, resulting in an overall accuracy of 0.985. Both precision and recall are equally high at 0.985, indicating that the model neither over-detects tumors nor fails to identify them; there are just six false positives and six false negatives, which is an impressive outcome considering the clinical variability of brain tumor appearances. The F1 score, which is the harmonic mean of precision and recall, also stands at 0.985. This confirms that the decision thresholds are well calibrated and that any class imbalance, if present, has been effectively addressed. The minimal performance gap between this task-specific fine-tuning and the original pre-training suggests that the latent representations learned during multi-dataset training transfer effectively to domain-specific tasks. In practical terms, the system misclassifies approximately one scan out of every 67, which is a low enough rate to support real-world triage or decision-support workflows while leaving room for further improvement through techniques such as ensembling or post-processing. Overall, the report confirms that MedFoundX maintains its characteristic generalizability after fine-tuning and provides clinically actionable accuracy on Bangladeshi brain cancer MRI data.

Table 4.5: Inference Results for Classification: PMRAM Dataset

Metric	**Value**
Accuracy	0.9850
Precision	0.9850
Recall	0.9850
F1 Score	0.9850

4.7.2 Inference for Segmentation: CVC-ClinicDB Dataset

The fine-tuned MedFoundX segmentation model exhibits solid and clinically significant performance on the CVC-ClinicDB held-out test set, as shown in Table 4.6. Across the 13 colonoscopy frames in this split, the network achieves an average Dice coefficient of 0.83 with a standard deviation of 0.10. A Dice score in the low 0.80s indicates that approximately four-fifths of the predicted polyp pixels overlap with the ground truth masks, which is a respectable result given the challenging variations in lighting, mucus, and tissue folds of the data set. The complementary IoU score stands at 0.72 ± 0.14, indicating a reliable but imperfect boundary delineation. The pixel-level accuracy is significantly higher, averaging 0.986, due to the dominance of the background class in each frame. This disparity between

overall accuracy and region-based metrics highlights that Dice and IoU are more informative indicators for medical image segmentation, particularly when capturing small foreground objects is crucial. In particular, the model exhibits excellent sensitivity (0.94 ± 0.04), indicating that it detects nearly all true polyp pixels. The specificity is similarly impressive at 0.988 ± 0.008, confirming that normal tissue is rarely misidentified as a polyp. The trade-off is primarily evident in precision, which is 0.75 ± 0.15; this means that when the model flags a pixel as abnormal, one of four predictions is a false positive. This mild tendency to over-segment is common in safety-oriented medical systems, where the risk of missing pathology is costlier than slightly enlarging masks. Overall, the metrics indicate that the model generalizes well from its multi-dataset foundation while effectively adapting to colon polyp morphology after fine-tuning. The high sensitivity and specificity make the model suitable for clinical decision support, as automated masks can guide endoscopists to regions of concern. Although precision is adequate, there is room for improvement; post-processing steps, such as conditional random fields or confidence-threshold calibration, could refine boundaries and reduce false positives. Finally, the modest standard deviations across Dice, IoU, and precision suggest consistent performance despite the very small test set. Enlarging the evaluation cohort would provide a clearer picture of the robustness of the real world. However, current figures already indicate that MedFoundX is a reliable starting point for gastrointestinal lesion segmentation workflows.

Table 4.6: Inference Results for Segmentation: CVC-ClinicDB Dataset

Metric	**Average**	**Standard Deviation**
Mean Dice	0.8292	0.1050
Mean IoU	0.7211	0.1441
Pixel Accuracy	0.9862	0.0070
Sensitivity	0.9412	0.0418
Specificity	0.9878	0.0078
Precision	0.7546	0.1535
F1 Score	0.8292	0.1050

4.8 Comparison with Baseline Models

This section summarizes how MedFoundX compares to well-known baselines. It demonstrates that, across four classification datasets, our model achieves or surpasses state-of-the-art accuracies while significantly reducing cross-entropy loss by orders of magnitude. On the four segmentation benchmarks, MedFoundX consistently posts the lowest losses and, in most cases, the highest Dice scores, demonstrating strong and confident mask predictions even on challenging modalities. Together, these head-to-head results highlight MedFoundX's superior accuracy, reliability, and adaptability compared to popular CNN, transformer, and hybrid architectures.

4.8.1 Classification Results Comparison

Table 4.7 highlights the classification accuracy (higher is better) and cross-entropy loss (lower is better) of various models across four biomedical imaging datasets. Our proposed model, MedFoundX, consistently outperforms baseline and state-of-the-art models in terms of both classification accuracy and loss.

Table 4.7: Comparison with Baseline Models for Classification Datasets

Model	Brain MRI ND-5		Brain Tumor MRI		Pancreas CT		Mammogram Mastery	
	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss
CNN	0.983	0.119	0.981	0.097	1.000	0.002	0.980	0.085
KAN	0.896	—	0.935	—	0.966	—	0.953	—
ResNet-50	0.998	0.012	0.999	0.002	1.00	0.018	0.993	0.055
SqueezeNet-1	0.992	0.036	0.999	0.021	1.000	0.017	**1.000**	0.008
Swin-T	0.313	1.373	0.594	1.013	0.813	0.541	0.926	0.294
VGG-16	0.993	0.035	0.992	0.049	1.000	0.006	0.940	0.206
MedFoundX	**0.999**	**3.76**$\times 10^{-5}$	**0.999**	**3.09**$\times 10^{-5}$	**1.000**	**6.49**$\times 10^{-7}$	0.987	**1.00**$\times 10^{-3}$

MedFoundX achieves the highest accuracy in three of the four datasets: Brain MRI ND-5 (0.999), Brain Tumor MRI (0.999), and Pancreas CT (1.0), demonstrating nearly perfect classification performance across these datasets. In the Mammogram Mastery dataset, although SqueezeNet achieves a perfect accuracy of 1.0, MedFoundX maintains a highly competitive accuracy of 0.987, significantly outperforming it in terms of cross-entropy loss. Across all datasets, MedFoundX achieves the lowest loss, often by several orders of magnitude. For example, in the Brain MRI ND-5 dataset, MedFoundX reports a loss of 3.76×10^{-5}, outperforming ResNet-50 (0.012) and the standard CNN (0.119). Similarly, in the Brain Tumor MRI dataset, the loss is reduced to 3.09×10^{-5}, and in the Pancreas CT dataset, the model achieves an almost negligible loss of 6.49×10^{-7}. Even in the Mammogram Mastery dataset, MedFoundX reports a loss of 1.00×10^{-3}, outperforming all other models. Transformer-based models such as Swin-T show noticeably inferior performance, with accuracies as low as 0.313 in the Brain MRI ND-5 dataset and consistently high loss values. This performance gap underlines the challenge of applying vanilla transformer architectures directly to biomedical imaging tasks without significant domain adaptation or architectural customization.

In conclusion, MedFoundX achieves state-of-the-art performance across multiple classification benchmarks, not only in terms of accuracy but also in terms of model confidence as reflected in its extremely low loss values. These results strongly support its potential as a generalizable and clinically viable foundation model for biomedical image classification.

4.8.2 Segmentation Results Comparison

Table 4.8 presents the segmentation performance of various models across four biomedical datasets using two key metrics: cross-entropy Loss (lower is better) and mean Dice Score (higher is better), which quantifies the overlap between predicted and ground-truth segmentations.

Table 4.8: Comparison with Baseline Models for Segmentation Datasets

Model	MedSeg COVID CT		MedSeg Liver CT		Kvasir-SEG		Brain Stroke CT	
	Loss	Dice	Loss	Dice	Loss	Dice	Loss	Dice
CNN	0.335	0.280	0.365	0.944	0.266	0.613	0.022	0.658
KAN	—	0.250	—	0.814	—	0.295	—	0.595
DeepLabV3	0.194	**0.668**	0.094	0.942	0.081	0.930	0.226	0.463
FCN	0.198	0.634	0.098	0.944	0.103	0.909	0.220	0.533
LRASPP	0.231	0.534	0.105	0.899	0.118	0.902	0.234	0.450
MedFoundX	**0.026**	0.493	**0.008**	**0.960**	**0.003**	**0.941**	**0.003**	**0.848**

MedFoundX outperforms all other models in three out of four datasets in terms of mean Dice Score while also consistently achieving the lowest loss values across all datasets. This demonstrates the model's ability to produce highly accurate segmentations with confident and well-calibrated predictions. In MedSeg COVID CT, MedFoundX achieves a remarkably low loss of 0.026, significantly outperforming CNN (0.335), DeepLabV3 (0.194), and FCN (0.198). While DeepLabV3 scores the highest mean Dice Score(0.668), MedFoundX yields a moderate Dice of 0.493. The slightly lower Dice score may be attributed to class imbalance or reduced lesion contrast; yet, the extremely low Loss suggests strong model confidence and pixel-level prediction certainty. In the MedSeg Liver CT dataset, MedFoundX achieves both the highest mean Dice score of 0.960 and the lowest Loss of 0.008, outperforming DeepLabV3 (0.942 Dice, 0.094 loss), FCN (0.944 Dice), and CNN (0.944 Dice, 0.365 loss). This indicates the model's superior ability to accurately and consistently delineate liver structures. In the polyp segmentation task of Kvasir-SEG, MedFoundX again leads with a mean Dice score of 0.941 and a remarkably low loss of 0.003. In comparison, DeepLabV3 and FCN achieve Dice scores of 0.930 and 0.909, respectively. The significant margin in Loss further reinforces the model's confidence in its segmentation masks. In Brain Stroke CT, MedFoundX delivers a strong performance with a mean Dice score of 0.848 and a minimal loss of 0.003, outperforming all baselines by a large margin. CNN and KAN achieve much lower Dice scores of 0.658 and 0.595, respectively, while DeepLabV3 and FCN perform even worse in this dataset (Dice of 0.463 and 0.533). Compared to other segmentation architectures, MedFoundX demonstrates high adaptability across diverse imaging modalities, including CT, endoscopy, and MRI. While DeepLabV3 and FCN perform well on select datasets, they consistently produce higher loss values, indicating less confident predictions. KAN, where available, underperforms in Dice scores, and CNN struggles particularly in datasets with complex boundaries or low-contrast regions. MedFoundX achieves state-of-the-art segmentation performance across multiple biomedical benchmarks, striking a balance between accuracy and confidence. Its consistently low Loss and high mean Dice scores underscore its potential as a reliable and generalizable segmentation backbone suitable for clinical deployment.

The performance gains of MedFoundX can be attributed to its architectural innovations, which include: A pre-trained EfficientNet-B3 backbone for efficient and robust feature extraction. Integration of CBAM (Convolutional Block Attention Module) to focus on the most relevant spatial and channel-wise features. Multi-

Head Attention to capture complex interdependencies in medical imaging data. These components work synergistically to enable MedFoundX to learn highly discriminative and generalizable representations from diverse biomedical imaging datasets.

4.9 Computation Analysis

Table 4.9 presents a comparative analysis of the computational complexity of MedFoundX against widely used classification and segmentation models. Three key metrics are considered: the number of trainable parameters, the number of floating-point operations per second (FLOPs), and the total model size on disk. For all metrics, lower values indicate greater computational efficiency, which is critical for resource-constrained real-time clinical applications.

Table 4.9: Computational Complexity Compared with Baseline Models

Model	#Params	FLOPs	Model Size
MedFoundX	**22.21 M**	**2.8 B**	**∼84.7 MB**
ResNet-50 [44]	25.6 M	4.1 B	∼98 MB
FCN (ResNet-50) [45]	32.7 M	5.0 B	∼131 MB
Swin-T [46]	28.3 M	4.5 B	∼113 MB
DeepLabV3 [19]	39.6 M	8.0 B	∼158 MB
ViT-B/16 [47]	86.6 M	17.6 B	∼345 MB
VGG-16 [48]	138 M	15.3 B	∼553 MB

MedFoundX demonstrates a balanced trade-off between performance and complexity, with 22.21 million parameters, 2.8 billion FLOPs, and a model size of approximately 84.7 MB. While not the smallest model in absolute terms, it remains significantly more compact than most other high-performing models in the comparison. For instance, DeepLabV3 with a ResNet-50 backbone requires 39.6 million parameters, 8.0 billion FLOPs, and occupies roughly 158 MB of disk space, nearly double the size and over triple the computational cost of MedFoundX. Similarly, ViT-B/16 and VGG-16 are among the most computationally expensive, requiring 86.6 million and 138 million parameters, respectively, and over 15 billion FLOPs each, making them unsuitable for deployment on edge devices or in real-time clinical settings. Compared to traditional architectures such as FCN (32.7 million parameters, 5.0 billion FLOPs) and ResNet-50 (25.6 million parameters, 4.1 billion FLOPs), MedFoundX achieves a notable reduction in model size and computational load while maintaining strong performance across both classification and segmentation tasks. Transformer-based models such as Swin-T and ViT-B/16, although powerful, exhibit higher computational demands. Swin-T, for example, has 28.3 million parameters and requires 4.5 billion FLOPs with a model size of 113 MB. These complexities pose practical challenges in low-power or latency-sensitive environments.

In summary, MedFoundX achieves a favorable balance between model performance and computational efficiency. Its moderate parameter count and model size make it more suitable for clinical deployment scenarios where computational resources may be limited. This efficiency, coupled with high accuracy and low loss across

tasks, reinforces the potential of MedFoundX as a scalable and deployable foundation model for biomedical imaging applications.

4.10 Discussion

This research aimed to design a unified and modality-agnostic foundation model that could perform both classification and segmentation across diverse biomedical imaging modalities while remaining computationally efficient. The experimental results demonstrate that these goals were met, providing evidence that the proposed MedFoundX architecture delivers both high accuracy and computational efficiency for clinical readiness.

4.10.1 Development of a Unified Foundation Model

The central aim of the thesis was to create a single architecture capable of performing both classification and segmentation across diverse biomedical imaging datasets. MedFoundX achieves this goal by consistently demonstrating high performance in classification tasks, while also excelling in segmentation. In the classification domain, the training curves for Brain MRI ND-5, Brain Tumor MRI, Pancreas CT, and Mammogram Mastery show rapid convergence, low loss values, and nearly identical training-validation trajectories, indicating a model that quickly learns discriminative features without overfitting. Table 4.1 corroborates this: MedFoundX achieves near-perfect accuracy, precision, recall, and F1 score in all four data sets, with Pancreas CT and Mammogram Mastery tasks reaching a full 1.0 in all metrics. The segmentation results echo these findings. On the MedSeg Liver and Kvasir-SEG datasets, the mean Dice scores exceed 0.93, with IoU values above 0.86, indicating a strong overlap between the predicted masks and the ground truth. Even in the more challenging Brain Stroke CT dataset, the model maintains a mean Dice of 0.82 and IoU of 0.73. Although performance in the MedSegCOVID set is lower (Dice = 0.495), the model still achieves 98% precision, suggesting that class imbalance, rather than architectural deficiencies, drives this modest plateau. Collectively, these results confirm that a unified end-to-end framework can effectively handle both classification and segmentation tasks, thus fulfilling the first objective.

4.10.2 Enhancement of Generalizability

A second goal was to ensure that the model could generalize across multiple imaging modalities and unseen datasets. Training sequences reveal that MedFoundX learned progressively from Brain MRI, Brain Tumor MRI, Pancreas CT, and Mammogram Mastery data by transferring weights between tasks, which enabled it to encode modality-agnostic characteristics. When fine-tuned for new datasets, this generalization proved to be robust. In the PMRAM Bangladeshi brain cancer MRI dataset, the fine-tuned model achieved an accuracy, precision, recall, and F1-Score of 0.99, with class-specific metrics ranging from 0.97 to 1.0. Training losses decreased drastically and validation losses remained slightly lower, signaling good regularization. Similarly, fine-tuning on CVC-ClinicDB for polyp segmentation produced a mean Dice score of 0.95, a mean IoU of 0.90, and a pixel-level accuracy of 0.98, with validation curves slightly surpassing the training curves during

optimization. Inference results on held-out test sets reinforce this adaptability: the PMRAM test set recorded an overall F1 score of 0.985 with only six false positives and six false negatives among 400 volumes, while the CVC-ClinicDB test set displayed a mean Dice coefficient of 0.83 and an IoU of 0.72, despite challenging lighting and tissue variability. These findings demonstrate that multi-dataset pre-training equips MedFoundX with representations that transfer effectively to new domains, thereby achieving the objective of enhanced generalizability.

4.10.3 Optimization for Computational Efficiency

The final objective was to balance high diagnostic performance with computational efficiency, thus facilitating deployment in resource-constrained clinical settings. The computational analysis in Table 4.9 demonstrates that MedFoundX uses 22.21 million parameters and 2.8 billion FLOPs, resulting in a model size of approximately 84.7 MB. Although not the smallest architecture available, it is significantly lighter than many baselines; for example, DeepLabV3 requires 39.6 million parameters and 8.0 billion FLOPs, while ViTB/16 demands 86.6 million parameters and 17.6 billion FLOPs. This efficiency does not compromise accuracy: In classification comparisons (Table 4.7), MedFoundX matches or exceeds state-of-the-art accuracies while achieving cross-entropy losses several orders of magnitude lower than competing models. In segmentation (Table 4.8), it consistently delivers the lowest loss values and high Dice scores across liver, polyp, and stroke datasets, outperforming CNN, KAN, FCN, LRASPP, and even transformer-based models. Training also converges within 100 to 150 epochs for most tasks, indicating that fine-tuning can be performed efficiently. Together, these attributes demonstrate that MedFoundX provides an effective trade-off between complexity and performance, making it suitable for deployment on edge devices without sacrificing diagnostic accuracy, and thus fulfilling the third objective.

The results provide strong evidence that MedFoundX fulfills the research objectives. A single architecture successfully performs both classification and segmentation with high accuracy in multiple modalities. The model generalizes well to diverse clinical imaging conditions and can be fine-tuned for new tasks with minimal performance degradation. A lightweight design achieves state-of-the-art performance while maintaining a manageable parameter count and FLOP budget, making it suitable for edge devices. These achievements validate the potential of unified and modality-agnostic foundation models in biomedical imaging and lay the groundwork for integrating MedFoundX into clinical workflows as a decision support tool or as a starting point for specialized downstream models.

Chapter 5

Conclusion

This thesis presents MedFoundX, a unified and modular deep learning architecture specifically designed for biomedical image analysis. It effectively handles both classification and segmentation tasks while ensuring high accuracy and computational efficiency. MedFoundX incorporates a pre-trained EfficientNet-B3 backbone along with Convolutional Block Attention Modules (CBAM) and Multi-Head Attention, enhancing feature representation across both local and global contexts, which is essential for detecting subtle pathological markers. The architecture includes task-specific heads for classification and segmentation, enabling multi-task learning while maintaining generalization across various imaging modalities. The model was trained on eight publicly available and clinically relevant datasets covering multiple imaging types such as MRI, CT, X-ray, and colonoscopy. It addressed a range of tasks, from binary disease detection to pixel-level lesion segmentation. A sequential weight-transfer protocol was employed during training, allowing MedFoundX to progressively learn from each dataset. This approach, combined with extensive data augmentation and weighted loss functions, resulted in rapid convergence and minimized overfitting. Fine-tuning on the PMRAM Bangladeshi brain tumor MRI dataset yielded exceptional results, achieving 99% accuracy, precision, recall, and F1-score, along with minimal loss and balanced performance across all tumor classes. Similarly, fine-tuning on the CVC-ClinicDB colonoscopy dataset resulted in a mean Dice score of 0.95 and a mean Intersection over Union (IoU) of 0.90, demonstrating stable convergence and low standard deviation in sensitivity and specificity. These findings confirm that MedFoundX generalizes effectively to unseen modalities and maintains high reliability in domain-specific applications. Inference analysis conducted on held-out test sets further validated the model's real-world applicability. On the PMRAM dataset, the model achieved an accuracy of 98.5% and an F1-score, misclassifying only six out of 400 MRI scans. For the CVC-ClinicDB segmentation inference, the average Dice score was 0.83 and the IoU was 0.72, with high pixel accuracy and robust sensitivity (0.94) and specificity (0.9878). These results illustrate not only strong feature retention during fine-tuning but also consistent and clinically meaningful performance during deployment. The model exhibited a slight over-segmentation bias, as indicated by its precision metrics. This tendency is typical in safety-critical applications, as it prioritizes sensitivity to prevent missed pathologies.

MedFoundX consistently outperformed a range of baseline models across all benchmarked datasets. In classification, it achieved near-perfect accuracies on

Brain MRI ND-5, Brain Tumor MRI, Pancreas CT, and Mammogram Mastery, with cross-entropy loss values several orders of magnitude lower than ResNet-50, CNN, VGG-16, and transformer-based models like Swin-T and ViT-B/16. Even when baseline models reached similar accuracy, they exhibited higher loss, indicating less confident predictions. In segmentation, MedFoundX delivered superior Dice scores on the MedSeg-Liver (0.964), Kvasir-SEG (0.941), and Brain Stroke (0.848) datasets, outperforming architectures such as FCN, DeepLabV3, and KAN in both accuracy and model confidence (low loss). Unlike transformer-based models, which struggled with medical imaging datasets due to architectural mismatch, MedFoundX effectively balanced inductive biases from CNNs with global attention mechanisms, resulting in improved spatial understanding and computational stability. From a deployment perspective, MedFoundX also demonstrated significant computational advantages. It maintained high performance while remaining efficient, with only 22.21 million parameters and 2.8 billion FLOPs, making it significantly lighter than ResNet-50 (25.6M), DeepLabV3 (39.6M), and ViT-B/16 (86.6M). The model's 84.7MB footprint enables it to be deployed in edge environments or integrated into portable diagnostic systems, where computational efficiency is crucial.

Overall, this work establishes MedFoundX as a reliable and adaptable foundation model for biomedical imaging. It not only excels across heterogeneous tasks and imaging modalities but also strikes a balance between performance and computational efficiency. Its successful application to unseen datasets underscores its transferability, and its architectural modularity supports flexible adaptation to new tasks. The results validate MedFoundX as a practical solution for bridging the performance-efficiency gap in clinical AI.

5.1 Limitations

Despite the promising results of MedFoundX, several limitations need to be addressed before it can be widely deployed.

- **Loss of fine detail due to low-resolution training:** To reduce computational costs we downsampled most images to 64×64. However, interpolation-based downsampling can fail to reproduce fine anatomical details and may blur or remove clinically important structures. Training on higher-resolution images or incorporating super-resolution techniques could improve the capture of subtle features.

- **Limited external validation:** Although MedFoundX was pre-trained on eight public datasets, it was fine-tuned and tested on only two. Differences in scanners, protocols and patient populations cause domain shifts, and models trained on one site often degrade on unseen institutions. Extensive evaluation across diverse, multi-institutional datasets is necessary to confirm generalizability.

- **Model efficiency:** Our architecture is smaller than many vision transformers but still larger than extreme lightweight models such as MobileNet or Tiny-YOLO. For ultra-low-resource or embedded applications, we must investigate further compression via pruning, quantization, or efficient

transformer variants that reduce computational complexity from quadratic to linear.

- **Narrow task coverage:** The present study focused on image classification and binary segmentation. However, medical imaging encompasses diverse tasks including multi-label segmentation, instance segmentation, registration, and enhancement. Foundation models currently emphasize segmentation and classification, while other clinical scenarios remain underexplored. Future work should extend MedFoundX to multi-task settings and temporal modeling.

- **Lack of clinical validation:** MedFoundX has not been evaluated in a prospective clinical workflow. Trustworthy AI requires that models be tested in the clinical context because AI systems can sometimes learn spurious correlations, such as hospital markers, and make incorrect inferences, undermining adoption and trust. Rigorous clinical trials with clinician feedback and explainability assessments are therefore essential before deployment.

Addressing these limitations—by training at higher resolutions, evaluating across multiple institutions, optimizing for resource-constrained hardware, broadening task coverage, and conducting clinical validation—will be crucial for MedFoundX to meet real-world and regulatory requirements.

5.2 Future Direction

To translate our MedFoundX prototype into a clinically reliable system, future efforts must address generalization, multi-modal integration, interpretability, and privacy within the medical imaging domain. Domain adaptation is crucial because imaging protocols and scanners can vary significantly across institutions, leading to a decline in model performance when tested on data from different sites. Instead of assuming access to labeled data from the target site, we can combine domain-generalization and meta-learning strategies with test-time adaptation (TTA). Meta-learning can help create a global initialization that can be quickly adjusted to fit a client's local distribution. At the same time, TTA enables a pretrained model to self-tune on unlabeled test samples using techniques like entropy minimization. This approach should result in a model capable of robust cross-site deployment without the need for retraining.

Since clinical decisions rarely rely solely on imaging, MedFoundX should evolve into a multi-modal foundation model. Intermediate-fusion networks that integrate representations across multiple modalities are more effective than simple early or late fusion methods, as they capture the hierarchical relationships among features. Future work may involve incorporating cross-modal transformers and graph-based fusion techniques, which have shown improvements in Alzheimer's disease diagnosis by jointly analyzing imaging and clinical data. This will allow us to integrate radiology with clinical notes, laboratory results, and genomic information. Graph-based fusion will enable us to treat each patient as a node with features from multiple sources, while cross-attention mechanisms will align shared and modality-specific

representations.

Interpretability should be an integral part of the model rather than an afterthought. Concept bottleneck models restrict predictions to human-understandable concepts and allow clinicians to intervene during the testing phase. A recent two-step approach employs a pretrained vision-language model to predict clinical concepts and a large language model to generate diagnoses based on those concepts. This enables clinicians to correct predictions without needing to retrain the model. We plan to combine such inherent interpretability with post-hoc tools (e.g., Grad-CAM) and calibration metrics like the Adaptive Expected Calibration Error, which assesses whether model confidence aligns with accuracy. Evidence suggests that evidential neural networks and uncertainty-aware training can enhance calibration and reliability, which we will explore to ensure MedFoundX's predictions are not overly confident.

Fairness and privacy are essential considerations as well. Bias can emerge during data collection, annotation, model development, or deployment. Recent multi-institutional studies emphasize the need for expanding fairness metrics and mitigation strategies throughout the entire process. Therefore, we will evaluate MedFoundX using fairness metrics stratified by demographic and clinical subgroups and investigate bias-mitigation strategies such as data augmentation and reweighting. Additionally, federated learning offers a practical approach for multi-center training without sharing raw data. Federated models avoid transferring data but may still expose intermediate parameters that could leak information. To address this, we are adopting privacy-enhancing techniques such as homomorphic encryption (which allows computations on encrypted data), secure multiparty computation, and differential privacy (which adds noise to parameters). We plan to experiment with these mechanisms, particularly partial homomorphic encryption, which balances privacy with computational efficiency, and will also address non-IID client data using meta-learning and clustered federated schemes.

Lastly, reproducibility and transparency will be vital to our evaluations. In addition to standard metrics, we will report on calibration, fairness, and robustness indices, publish model cards and containerized pipelines, and contribute to public leaderboards to support independent benchmarking. Together, these initiatives will extend our current work, guiding MedFoundX toward a robust, interpretable, and equitable foundation model for medical imaging.

Bibliography

[1] A. P. Dhawan, "A review on biomedical image processing and future trends," *Computer Methods and Programs in Biomedicine*, vol. 31, no. 3, pp. 141–183, 1990, ISSN: 0169-2607. DOI: https://doi.org/10.1016/0169-2607(90)90001-P. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016926079090001P.

[2] P. Singh, *Biomedical Image Analysis: Special Applications in MRIs and CT scans* (Brain Informatics and Health), en. Singapore: Springer Nature, 2024, ISBN: 9789819999385 9789819999392. DOI: 10.1007/978-981-99-9939-2. [Online]. Available: https://link.springer.com/10.1007/978-981-99-9939-2 (visited on 07/03/2025).

[3] S. Renukalatha and K. V. Suresh, "A REVIEW ON BIOMEDICAL IMAGE ANALYSIS," en, *Biomedical Engineering: Applications, Basis and Communications*, vol. 30, no. 04, p. 1 830 001, Aug. 2018, ISSN: 1016-2372, 1793-7132. DOI: 10.4015/S1016237218300018. [Online]. Available: https://www.worldscientific.com/doi/abs/10.4015/S1016237218300018 (visited on 07/03/2025).

[4] H. Sun and J. Wang, "Computational biomedical imaging: Ai innovations and pitfalls," *Medicine Plus*, vol. 2, no. 2, p. 100 081, 2025, ISSN: 2950-3477. DOI: https://doi.org/10.1016/j.medp.2025.100081. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S295034772500012X.

[5] M. Tounsi, E. Aram, A. T. Azar, A. Al-Khayyat, and I. K. Ibraheem, "A comprehensive review on biomedical image classification using deep learning models," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 19 538–19 545, Feb. 2025. DOI: 10.48084/etasr.8728. [Online]. Available: http://dx.doi.org/10.48084/etasr.8728.

[6] F. Wahid, Y. Ma, D. Khan, M. Aamir, and S. U. K. Bukhari, *Biomedical Image Segmentation: A Systematic Literature Review of Deep Learning Based Object Detection Methods*, arXiv:2408.03393, Aug. 2024. DOI: 10.48550/arXiv.2408.03393. [Online]. Available: http://arxiv.org/abs/2408.03393 (visited on 07/03/2025).

[7] Z. Yang, J. Zhang, X. Luo, Z. Lu, and L. Shen, *MedKAN: An Advanced Kolmogorov-Arnold Network for Medical Image Classification*, arXiv:2502.18416, Feb. 2025. DOI: 10.48550/arXiv.2502.18416. [Online]. Available: http://arxiv.org/abs/2502.18416 (visited on 07/03/2025).

[8] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, *MedViT: A Robust Vision Transformer for Generalized Medical Image Classification*, arXiv:2302.09462, Feb. 2023. DOI: 10.48550/arXiv.2302.09462. [Online]. Available: http://arxiv.org/abs/2302.09462 (visited on 07/03/2025).

[9] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," *BMC Medical Imaging*, vol. 22, p. 69, 2022, A comprehensive review of transfer-learning techniques for medical image classification. DOI: 10.1186/s12880-022-00793-7.

[10] G. K. Thakur, A. Thakur, S. Kulkarni, N. Khan, and S. Khan, "Deep learning approaches for medical image analysis and diagnosis," *Cureus*, vol. 16, no. 5, e59507, 2024, A review summarising deep learning techniques for medical image analysis and diagnosis. DOI: 10.7759/cureus.59507.

[11] V. van Veldhuizen, V. Botha, C. Lu, *et al.*, "Foundation models in medical imaging – a review and outlook," *arXiv preprint*, 2025, Comprehensive review of foundation models in medical imaging, covering architectures, pretraining strategies and adaptation methods. DOI: 10.48550/arXiv.2506.09095. arXiv: 2506.09095 [eess.IV].

[12] R. M. Haralick, K. Shanmugam, and I. h. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973, Introduces gray-level co-occurrence matrices for texture analysis in image classification. DOI: 10.1109/TSMC.1973.4309314.

[13] Y. Ding, Z. Yi, M. Li, *et al.*, "HI-MViT: A lightweight model for explainable skin disease classification based on modified mobilevit," *Digital Health*, vol. 9, p. 20552076231207197, 2023, Proposes HI-MViT, an efficient MobileViT variant for dermoscopic image classification. DOI: 10.1177/20552076231207197.

[14] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint*, 2022, Introduces MobileViT, a hybrid CNN-Transformer architecture for mobile applications. DOI: 10.48550/arXiv.2110.02178. arXiv: 2110.02178 [cs.CV].

[15] C. K. K. Reddy, H. I. Ahmed, M. Mohzary, *et al.*, "AlzheimerViT: Harnessing lightweight vision transformer architecture for proactive alzheimer's screening," *Frontiers in Medicine*, vol. 12, p. 1568312, 2025, Applies a MobileViT-based architecture for Alzheimer's disease classification from brain MRI. DOI: 10.3389/fmed.2025.1568312.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[17] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[18] O. Oktay, J. Schlemper, L. Folgoc, *et al.*, "Attention u-net: Learning where to look for the pancreas," in *Medical Imaging with Deep Learning*, 2018.

[19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 801–818.

[20] R. F. Khan, B.-D. Lee, and M. S. Lee, "Transformers in medical image segmentation: A narrative review," *Quantitative Imaging in Medicine and Surgery*, vol. 13, no. 12, pp. 8747–8767, 2023, Surveys transformer-based architectures for medical image segmentation, including hybrid designs and attention mechanisms. DOI: 10.21037/qims-23-542.

[21] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, "Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches," *Bioengineering*, vol. 11, no. 10, p. 1034, 2024, Provides a thorough review of segmentation techniques, from classic thresholding and region growing to modern deep-learning methods and hybrid models. DOI: 10.3390/bioengineering11101034.

[22] S. Noh and B.-D. Lee, "A narrative review of foundation models for medical image segmentation: Zero-shot performance evaluation on diverse modalities," *Quantitative Imaging in Medicine and Surgery*, vol. 15, no. 6, pp. 5825–5858, 2025, Reviews foundation models for medical image segmentation and assesses their zero-shot performance across modalities. DOI: 10.21037/qims-2024-2826.

[23] S.-C. Huang, M. Jensen, S. Yeung-Levy, M. P. Lungren, H. Poon, and A. S. Chaudhari, "A systematic review and implementation guidelines of multimodal foundation models in medical imaging," *Research Square Preprint*, 2025, Systematic review and guidelines for implementing multimodal foundation models in medical imaging. DOI: 10.21203/rs.3.rs-5537908/v1.

[24] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint*, 2016, Presents techniques for pruning and quantizing neural networks to reduce memory footprint and inference cost. arXiv: 1510.00149 [cs.NE].

[25] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint*, 2015, Introduces knowledge distillation, training a compact student network to mimic a larger teacher network. arXiv: 1503.02531 [stat.ML].

[26] M. Tan, B. Chen, R. Pang, *et al.*, "Mnasnet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2820–2828.

[27] M. N. Safwan, S. Rahman, M. H. Mahadi, T. M. Jabir, and I. Mobin, *Brain mri nd-5 dataset*, 2024. DOI: 10.21227/q2vt-tf46. [Online]. Available: https://dx.doi.org/10.21227/q2vt-tf46.

[28] J. Cheng, "brain tumor dataset," Apr. 2017. DOI: 10.6084/m9.figshare.1512427.v8. [Online]. Available: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427.

[29] S. Bhuvaji, A. Kadam, P. Bhumkar, S. Dedge, and S. Kanchan, *Brain tumor classification (mri)*, 2020. DOI: 10.34740/KAGGLE/DSV/1183165. [Online]. Available: https://www.kaggle.com/dsv/1183165.

[30] P. A. Abdalla, *Mammogram Mastery: A Robust Dataset for Breast Cancer Detection and Medical Education*, Apr. 2024. DOI: 10.17632/FVJHTSKG93.1. [Online]. Available: https://data.mendeley.com/datasets/fvjhtskg93/1 (visited on 07/23/2025).

[31] L. Fang, *Pancreas CT images*, Oct. 2023. DOI: 10.17632/WM8J6J2MPH.1. [Online]. Available: https://data.mendeley.com/datasets/wm8j6j2mph/1 (visited on 07/04/2025).

[32] P. Md Shahriar Mannan, M. Chowdhury, R. Rahman, A. U. Tamim, and M. M. Rahman, *PMRAM: Bangladeshi Brain Cancer - MRI Dataset*, Dec. 2024. DOI: 10.17632/M7W55SW88B.1. [Online]. Available: https://data.mendeley.com/datasets/m7w55sw88b/1 (visited on 07/04/2025).

[33] "Artificial Intelligence in Healthcare Competition (TEKNOFEST-2021): Stroke Data Set," *The Eurasian Journal of Medicine*, vol. 54, no. 3, pp. 248–258, Oct. 2022. DOI: 10.5152/eurasianjmed.2022.22096. [Online]. Available: https://www.eajm.org//en/artificial-intelligence-in-healthcare-competition-teknofest-2021-stroke-data-set-133405 (visited on 07/04/2025).

[34] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015, ISSN: 0895-6111. DOI: https://doi.org/10.1016/j.compmedimag.2015.02.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611115000567.

[35] D. Jha, P. H. Smedsrud, M. A. Riegler, *et al.*, *Kvasir-seg: A segmented polyp dataset*, 2019. arXiv: 1911.07069 [eess.IV]. [Online]. Available: https://arxiv.org/abs/1911.07069.

[36] MedSeg, H. B. Jenssen, and T. Sakinis, "MedSeg Covid Dataset 1," Jan. 2021. DOI: 10.6084/m9.figshare.13521488.v2. [Online]. Available: https://figshare.com/articles/dataset/MedSeg_Covid_Dataset_1/13521488.

[37] MedSeg, H. B. Jenssen, and T. Sakinis, "MedSeg Liver segments dataset," Jan. 2021. DOI: 10.6084/m9.figshare.13643252.v2. [Online]. Available: https://figshare.com/articles/dataset/MedSeg_Liver_segments_dataset/13643252.

[38] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2020. arXiv: 1905.11946 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1905.11946.

[39] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.

[40] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[41] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[43] Y. Xiao, J. Zhou, Y. Yu, and L. Guo, "Active jamming recognition based on bilinear EfficientNet and attention mechanism," en, *IET Radar, Sonar & Navigation*, vol. 15, no. 9, pp. 957–968, Sep. 2021, ISSN: 1751-8784, 1751-8792. DOI: 10.1049/rsn2.12089. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rsn2.12089 (visited on 07/24/2025).

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

[45] J. Long, E. Shelhamer, and T. Darrell, *Fully convolutional networks for semantic segmentation*, 2015. arXiv: 1411.4038 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1411.4038.

[46] Z. Liu, Y. Lin, Y. Cao, *et al.*, *Swin transformer: Hierarchical vision transformer using shifted windows*, 2021. arXiv: 2103.14030 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2103.14030.

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations*, 2021.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.