2025-12

Deep Learning-Based Stenosis Segmentation in X-ray Angiography: Vision Transformers vs. CNNs

Mekat, Md. Jahidul Hossain

IUB

Deep Learning-Based Stenosis Segmentation in X-ray Angiography: Vision Transformers vs. CNNs

Prepared by: Md. Jahidul Hossain Mekat (ID 2221395)
Kazi Samin Nawal (ID 2220256)
Jasper Oliver D Rozario (ID 2220841)

Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by
Saadia Binte Alam, PhD
Associate Professor
Department of Computer Science and Engineering
Independent University, Bangladesh

December 2025

Attestation

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

Author Name: _____ **Signature:** _____

Author Name: _____ **Signature:** _____

Author Name: _____ **Signature:** _____

Evaluation Committee

Supervision Panel

.....	.....
Supervisor	Co-supervisor

Examiners

.....	.....
.....	
External Examiner	

Office Use

.....	.....
Program Coordinator	Head of the Department
.....	
Director, Office of the Graduate Studies, Research and Industry Relations	
.....	
Library	

Acknowledgement

We would like to express sincere gratitude to our respected Professor, Saadia Binte Alam, PhD, Independent University, Bangladesh. Her dedicated guidance, insightful suggestions, remarkable patience, and the generous investment of her valuable time played a pivotal role in the successful completion of our project. Additionally, heartfelt thanks go to the Center for Computational Data Sciences (CCDS), Independent University, Bangladesh (IUB) for their extraordinary help and support. We acknowledge and appreciate the contributions of all individuals who supported us during this endeavor.

We also acknowledge the use of AI-assisted writing tools specifically OpenAI ChatGPT and Google Gemini for the sole purpose of refining sentence structures and improving the clarity of written expressions in this thesis. All intellectual content, analysis, findings, and conclusions presented herein are entirely our own.

Letter of Transmittal

To
Saadia Binte Alam, PhD
Associate Professor
Department of Computer Science & Engineering
Independent University, Bangladesh

Subject: Submission of Thesis Report on "Deep Learning-Based Stenosis Segmentation in X-ray Angiography: Vision Transformers vs. CNNs"

Dear Madam,

With due honor and respect, we, the undersigned students of the Department of Computer Science & Engineering at Independent University, Bangladesh (IUB), would like to submit our thesis report titled *"Deep Learning-Based Stenosis Segmentation in X-ray Angiography: Vision Transformers vs. CNNs,"* prepared in partial fulfillment of the requirements for the Bachelor of Science in Computer Science & Engineering degree.

This thesis was conducted under your esteemed supervision from February 2025 to December 2025. We have endeavored to make this report as comprehensive and thorough as possible, strictly adhering to the guidelines provided by the department. Throughout the research, we have applied our best efforts to ensure the accuracy, clarity, and academic integrity of every component of this work.

We hope you will find this report satisfactory and kindly accept it for evaluation.

Sincerely,

Md. Jahidul Hossain Mekat
ID: 2221395
Department of CSE, IUB

Kazi Samin Nawal
ID: 2220256
Department of CSE, IUB

.....

Jasper Oliver D Rozario
ID: 2220841
Department of CSE, IUB

Abstract

Millions of fatalities worldwide are attributed to coronary artery disease (CAD), which is caused by arterial stenosis. X-ray coronary angiography (XCA) remains the primary clinical method for diagnosing stenosis, yet manual reading is slow, subjective, and prone to inconsistency between observers. These practical shortcomings have motivated a push toward automated segmentation tools that can deliver consistent, reproducible assessments.

evaluations. Ten deep learning architectures for stenosis segmentation in 2D XCA images are thoroughly compared in this thesis, including two Transformer-based models (SegFormer MiT-B0 and MiT-B3) and eight CNN-based backbones (MobileNetV2, ResNet18/34/50/101, DenseNet121/201, ResNeXt50, and EfficientNetB3).

To guarantee a fair comparison, all models were trained using same preprocessing, augmentation, and training techniques using the publicly accessible ARCADE dataset. To solve the extreme class imbalance, a combination of Dice and Focal loss was used, and robustness was increased by using considerable data augmentation. The Dice Score Coefficient (DSC), Intersection over Union (IoU), sensitivity, specificity, and a bespoke efficiency score were used to assess performance. With a DSC of 0.632 and an IoU of 0.462, SegFormer MiT-B3 outperformed ResNet101, the strongest CNN baseline, to earn the best overall performance. MiT-B3 generated more continuous and anatomically coherent vessel borders, especially in faint or complex locations, according to qualitative analysis.Additionally, lightweight models such as MiT-B0 and MobileNetV2 demonstrated strong accuracy relative to their parameter counts, highlighting their suitability for real-time or resource-constrained clinical settings.

Overall, the results of the research show that the models used in the research for the segmentation of stenosis have a specific architectural benefit, which is based on the ability of the models to capture long-range relationships. The limitation of the research is based on the fact that the research is based on a single database, the use of static 2D images, and the lack of measurement of the severity of the stenosis, despite the positive results. For the development of more comprehensive clinical decision support systems, the research should be extended to include the measurement of the stenosis, the exploration of spatiotemporal models, and the use of multiple databases.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Background of the Study

1.1.1. Coronary Artery Disease (CAD) as a Major Health Issue

It develops when fatty plaque gradually accumulates along the inner walls of the coronary arteries, progressively narrowing the vessel lumen and restricting blood flow to the heart muscle. Left unchecked, this process substantially raises the risk of acute cardiac events. The disease carries an enormous global burden in terms of both death and long-term disability.

1.1.2. Role of X-ray Coronary Angiography (XCA) in Diagnosis

X-ray XCA is the primary method used to identify CAD and its severity. Angiograms are real-time pictures of the coronary arteries created by X-rays when a contrasting substance is injected into the arteries. These pictures provide a close-up view of the blood vessels, which aids in identifying whether stenosis is present. The visual evaluation of XCA pictures by radiologists and cardiologists has long been considered the gold standard [1] in CAD therapy decision-making.

1.2. Motivation

1.2.1. Challenges of Manual Stenosis Analysis

In spite of its importance in the diagnostic procedure, the conventional methodology of manual analysis of the XCA image suffers from various limitations. The most important limitations are that the procedure of interpretation is often time consuming and requires careful examination by highly skilled experts [2, 3].

1.2.2. Need for Automated Diagnostic Tools

These challenges of time consumption and variability in diagnosis emphasize the need to develop efficient tools for the automation of diagnosis [4]. This would be very helpful for doctors as an aid to provide an objective, consistent, and quick method for segmenting stenosis. It will also help medical experts diagnose diseases for patients all over the world.

1.3. Scope of the Study

The scope of this thesis is in a controlled environment where deep learning models are applied for a specific task which is segmenting the stenosis from 2D X-ray Coronary Angiography (XCA) images.

The research has these following boundaries:

- **Models:** The research is based on total 10 models where eight of them are convolutional neural networks and two vision transformer models which is Segformer with MiT B0 and MiT-B3 as backbones.
- **Dataset:**The whole research was conducted on the publicly available ARCADE dataset [2]. On this dataset training,validation,testing and all the other experiments were done. No private dataset were used in this research.
- **Task Focus:** Our task is only focused on stenosis segmentation where only the stenotic region is used. The research does not include other tasks like classification or coronary vessel segmentation.
- **Protocol:** The comparison is conducted under the same, standardized evaluation protocol [5] without the application of any external or different optimization methods for the individual models.
- **Deployment:** Although efficiency indicators are included in the study, the thesis does not address the creation or verification of a clinical deployment system that is ready for production [4].

1.4. Research Questions

This thesis aims to address the following core research questions:

1. **Performance Comparison:** When compared using the same standardized protocol [5], does the Vision Transformer (ViT) architecture, specifically SegFormer with the MiT backbone, show a statistically significant performance advantage in stenosis segmentation (as measured by DSC and IoU) over the Convolutional Neural Network (CNN) architectures tested?
2. **Robustness and Morphology Modeling:** How do the results of qualitative segmentation and the error patterns of the ViT and CNN models compare, especially with respect to dealing with the complex, small, and variable morphology of coronary stenosis in XCA images?
3. **Efficiency for Clinical Viability:** Which of the assessed deep learning models (ViT or CNN) is the best option for integration into a real-time clinical assistance system

[4] since it provides the best balance between high segmentation performance and computational efficiency (i.e., low parameter count)?

4. **Architectural Guideline:** What architecture of which model will perform better in this controlled environment study and what guidelines can be drawn from that which will be useful for future research on automated coronary artery disease detection ?

1.5. Research Objectives

1.5.1. Primary Objective

The main goal of this research is stenosis segmentation in XCA images and create a comparison of performance between conventional Convolutional Neural Network(CNN) architectures and modern Vision Transformer (ViT) model architectures. For a fair comparison all models were trained and tested under same experimental environement.

1.5.2. Secondary Objectives

Analyzing the trade-off between computational efficiency and model performance is a secondary goal. This will be accomplished by comparing the segmentation accuracy of each design to the total number of parameters, offering useful information for choosing models that may be used in clinical settings with limited resources [4].

1.6. Contributions of the Thesis

1.6.1. Evaluation of Deep Learning Model Performance

The prime motivation of this thesis is to tackle a crucial challenge in the automated segmentation of stenosis in images, which is that despite the immense progress in deep learning research, there is no clear, consistent, or standardized assessment of model performance on this specific task in medical imaging. The current state of research on this topic involves testing different models on various training settings, datasets, or preprocessing pipelines, which makes it very challenging to arrive at any conclusive findings about the performance of different architectures on this task. Consequently, there is a lack of a standardized reference for choosing the right standardized reference for selecting appropriate.

To tackle this challenge, the first contribution of this thesis is the systematic and controlled assessment of a wide range of deep learning architectures. The objective of this thesis is not only to compare CNNs and Vision Transformers but to systematically evaluate the performance of different families of architectures when exposed to the same dataset, the same training conditions, the same loss functions, and the same assessment metrics.

Performance was evaluated across several dimensions sensitivity, specificity [6], robustness under noisy conditions, and computational efficiency with the aim of building a picture that goes beyond surface-level accuracy metrics. Taken together, these measures offer a more grounded basis for understanding how deep learning models actually hold up when applied to something as clinically demanding as coronary stenosis segmentation, where the margin for error is narrow and the variability in presentation is considerable. The direct comparison across models is, arguably, where this work becomes most useful: rather than advocating for a single architecture, it surfaces what different approaches can and cannot do when the task is identifying and delineating stenotic regions.

1.6.2. Comparative Analysis of CNNs and ViTs

The research addresses the most important gap in the existing system of medical image segmentaion by doing a direct comparison of the most popular deep learning models of medical image segmentation which are CNNs and ViTs. Even though a lot of studies have been conducted on each of these models individually, few have compared them directly in a controlled study. As a result, there exist no current sound grounds in the literature to determine the more generally adapted model family to the task of coronary stenosis segmentation in X ray angiography.

In order to fill this gap, this research compares eight of the most prominent CNN models with two of the most popular SegFormer Transformer models. All these models are trained and evaluated on the same publicly available ARCADE dataset [2] with the same preprocessing, data augmentation techniques used [7], and with the same loss functions, optimization algorithms, and criteria applied all of them is the same [6]. In doing so, this study rules out confounding factors like variation in the training process, varied hyperparameters or varied data partitions [5] which do not affect the findings in any way other than that which is induced by the models themselves.

This research made a clear statement by addressing the pros and cons of each architectures through typical evaluation process. The findings show that in both qualitative and quantitative evaluations, Transformer-based models in particular, the MiT-B3 model perform better than CNNs. In addition to determining the modeling prowess of CNNs and ViTs, the comparative analysis of these two paradigms aids researchers, practitioners, and physicians in making evidence-based choices regarding the best architecture for stenosis segmentation.

1.6.3. Efficiency Analysis for Clinical Deployment

In addition to segmentation accuracy, there are also requirements of computational efficiency, memory, and near-real-time capability, as mentioned in [4]. In other words, it is not always possible to deploy models with high accuracy, especially in a clinical environment, as there are strict requirements of computational power. It is good to consider computational efficiency as a key aspect of medical image segmentation models. In light

of this, a separate section of efficiency analysis is included in the thesis, as it is essential to consider each model's efficiency for deployment-oriented scenarios.

To determine each model's efficiency, a separate efficiency score is introduced, based on a balance between segmentation accuracy and total model parameters. Using this approach, it is possible to compare the efficiency of heavy models like those based on the Transformer architecture with those based on CNN, as mentioned in [8]. Even though the highest accuracy was achieved by the MiT-B3 model, models like MiT-B0 and MobileNetV2 were found to have much higher efficiency, with a fraction of the parameters.

This Research gives a clear understanding of the accuracy of models and it also tells if the model is suitable to deploy in normal clinical hardware where high end devices is unavailable. This research opens up interesting directions for future research which particularly around model compression, lightweight architectures, and building deep learning systems that are actually practical to deploy. Beyond the immediate comparison, the findings feed into a larger conversation about what it would actually take to deploy computer-assisted diagnostic tools in real clinical environments where accuracy alone is rarely enough. Identifying which models manage to be both precise and computationally tractable is, in that sense, as practically significant as the performance results themselves.

1.7. Thesis Organization

The rest of thesis is organized as follows:

- **Chapter 2** covers clinical state of CAD, how deep learning models are used in medical image segmentation, and a detailed test of CNN and Vision Transformer architectures in a thorough assessment.
- **Chapter 3** describes the dataset that has been used in this research.
- **Chapter 4** explains all of the processes, model architectures, experiment setup, evaluation metrics and loss functions
- **Chapter 5** presents and evaluates the experimental findings, including discussion, efficiency analysis, and comparisons of both quantitative and qualitative data.
- **Chapter 6** summarizes the main conclusions, talks about the limits, and suggests areas for further research to wrap up the thesis.

2. Literature Review

This chapter discusses the interdisciplinary body of knowledge that provides a basis for this thesis. The chapter begins with a discussion that justifies the clinical importance of coronary artery stenosis, its pathophysiology, and the traditional diagnostic imaging modalities used to detect it. The chapter then proceeds to discuss the knowledge in the computational domain, where a detailed analysis of the two most prominent deep learning frameworks for image segmentation is presented. These two frameworks are CNNs and ViTs. A detailed examination of these two foundational frameworks is used to meticulously present how the evolution of these frameworks is a reflection of two prominent philosophies that have been in constant competition. Finally, this chapter presents a synthesis of these two domains by discussing the state-of-the-art in automated stenosis segmentation, which provides a precise basis for identifying a gap that thesis seeks to fix.

2.1. Fundamentals of Coronary Stenosis

2.1.1. Pathophysiology of Plaque Buildup

A cardiovascular disease termed CAD is the result of the gradual buildup of plaques in the walls of the coronary artery. This results in the "stenosis" of the coronary artery, or the narrowing of the lumen of the coronary artery. The reduced blood flow to the myocardial or the heart muscle is termed "ischemia." CAD can range from mild conditions such as "angina" and "fatigue" to life-threatening myocardial infarction. The reduced blood flow to the myocardium or the heart muscle brings about a number of symptoms.

CAD has a massive health impact on the world. The collective causes of death are cardiovascular diseases (CVDs) that are the major cause of death in the whole globe. As of 2021, an estimated 20.5 million deaths are caused by CVDs worldwide [9] with CAD alone causing an estimated 9 million of these deaths. This death toll does not exist in high-income countries alone; more than three-quarters of all deaths associated with CVDs are in low- and middle-income countries. This straightforward and direct cause-effect chain, including the accumulation of plaque to stenosis, and stenosis to CAD, the largest killer of the world, indicates the paramount significance of proper, timely, and dependable diagnosis of stenosis to successfully manage the patient.

2.1.2. Diagnostic Imaging Modalities

In view of the severity of CAD, its diagnosis is a cornerstone of clinical cardiology. CAD diagnosis is heavily dependent upon advanced medical imaging techniques to visualize the coronary arteries and measure the severity of stenosis. Two main medical imaging modalities are employed for this purpose: X-ray XCA and CCTA. XCA, or catheterization, is an invasive procedure in which a catheter is passed into the coronary arteries and a contrasting dye is injected. The image obtained from the X-ray is a high-resolution, 2D image of the vessel lumen and is "considered the gold standard" [1] for the assessment of stenosis. However, the standard clinical assessment of the XCA images using the manual approach is riddled with challenges. The manual assessment of the XCA images is a laborious task and, more importantly, "considerable interobserver variability" [2, 3] is an inherent problem with the standard manual assessment approach. Different cardiologists, or the same cardiologist on different occasions, may have different opinions on the presence or absence of a lesion in the coronary arteries. Thus, there is an urgent need for the development of an efficient and reliable diagnostic tool for the standardization of the manual assessment approach.

To facilitate the development of such automated tools, standardized benchmark datasets are essential. A key resource in this domain, and the one utilized in this thesis, is the ARCADE dataset. [2]. This is a publicly available resource composed of XCA images with corresponding ground-truth annotations, specifically curated to "facilitate the development and benchmarking of automated segmentation models".

In contrast to the invasive XCA, CCTA is a non-invasive imaging modality that uses computed tomography to generate detailed 3D volumetric reconstructions of the heart and coronary arteries [10]. Deep learning has also been extensively applied to CCTA data.

For example, Gupta et al. [11] created a deep learning model to detect the stenosis on CCTA scans. They have used EfficientNet and ResNet in their study and it showed great results. They achieved an Area Under the Curve for Receiver Operating Characteristic (AUROC) of up to 0.95 for detecting moderate to severe stenosis [12].

Automating the XCA analysis is currently very challenging .Automated CCTA analysis is currently the active field of research but XCA is still facing challenges. The 2D projection nature of XCA is a huge challenge along with vessel overlap and image artifacts which introduces new set of problems. This research is within the sub field of XCA based analysis which addresses the central clinical challenge. The challenge is "interobserver variability" inherent in the manual interpretation of this diagnostic procedure.

2.2. Deep Learning for Medical Image Segmentation

To address these challenges,CNNs have been recognized as the dominant technology in solving the problems related to automated medical image analysis [13]. "CNNs are now

recognized as a fundamental technology in solving a variety of problems in image analysis, following their revolutionary success [14]."

The strength behind CNNs is based on their "inherent architectural design, which allows CNNs to naturally integrate low/mid/high-level features," enabling a hierarchical representation of an image, "learning simple features like edges, textures, etc., in earlier layers, and complex, abstract features like objects, scenes, actions, etc., in later layers." This is extremely beneficial when it comes to image segmentation, as it is necessary to "identify complex structures like organs, tumors, or, in our case, blood vessels."

However, as scientists started to construct more complex networks to tackle more complex tasks, they faced a serious issue known as the "degradation problem". Contrary to expectations, adding more layers to a network led to the network's accuracy saturating and then dropping dramatically, resulting in a higher training error than its shallow counterpart. This was not due to overfitting [15], but rather the difficulty of optimizing such deep non-linear systems. The answer to this fundamental problem, proposed by He et al. [16], marked the start of the modern era of deep CNNs.

2.3. State-of-the-Art in Automated Stenosis Segmentation

This section summarizes all the reviews from general purpose model architectures to architectures which are only made to automate the task of stenosis segmentation. This sets the stage and gives us a clear performance baseline to measure the work in this thesis against.

2.3.1. Existing Methods and Models

For this specific task, ARCADE challenge is the central benchmark. This allowed the development of multiple high performing specialized models. A good example is StenUNet which is introduced by Lin et al. [17]. This model is a specialized architecture based on the highly successful UNet [18] framework, which is a gold standard for medical image segmentation. StenUNet was designed specifically for "automatic stenosis detection from X-ray coronary angiography" and was a top-performing entry in the ARCADE challenge, achieving an F1-score (equivalent to the Dice Score Coefficient, or DSC) of 0.5348.

Another leading solution from the same challenge is the SSASS framework by Lee et al. [19]. This work introduced a "Semi-Supervised Approach for Stenosis Segmentation". This represents a significant methodological contribution, as semi-supervised learning is highly valuable in the medical domain where acquiring large-scale, pixel-perfect annotations is a major bottleneck. The SSASS framework achieved a slightly higher F1-score of 0.536. Together, StenUNet and SSASS establish a strong performance baseline of approximately 0.536 DSC for specialized models on this public dataset.

More recently, the state-of-the-art has been advanced by the introduction of new architectural paradigms. Rostami et al. [20] explored the use of Mamba models for this

task. Mamba is a novel sequence model based on a State Space Model (SSM) architecture, which has been shown to "surpass transformers in terms of linear computational complexity". By integrating Mamba blocks into a U-Net-style architecture (termed "U-Mamba"), this work set the current state-of-the-art performance. The "U-Mamba BOT" model achieved a top F1-score of 0.6879, establishing the new performance ceiling for this task.

This review of the state-of-the-art (SOTA) is crucial for contextualizing the contribution of this thesis. The best-performing model in this work, SegFormer-MiT-B3, achieved a DSC of 0.632. While this is not superior to the current Mamba-based SOTA, it is a dramatic improvement over the high-performing specialized challenge benchmarks like StenUNet and SSASS (0.632 vs. 0.536). This demonstrates that the general-purpose, Transformer-based SegFormer is significantly more effective than specialized CNN-based approaches, a key scientific finding.

2.3.2. Identified Research Gap

What this literature review has tried to do, above all else, is trace a path from a real clinical problem to the architectures that have been proposed to solve it and in doing so, surface where the existing research falls short. That gap is not incidental. It is, in a sense, the entire reason this thesis exists. The architectural variety that emerged from the review is striking, and not always in a reassuring way. ResNet pursues depth as its primary strategy; ResNeXt introduces cardinality as a structural alternative; DenseNet takes aggressive feature reuse to its logical extreme; MobileNetV2 trades some representational capacity for efficiency; and EfficientNet attempts a principled scaling of all three dimensions at once. Then there is SegFormer, which represents a fundamentally different computational paradigm one built on attention mechanisms rather than local convolution. Each of these reflects a different hypothesis about what matters most in visual recognition tasks. The problem is that for stenosis segmentation specifically, no one has tested those hypotheses against each other in any systematic way. That is where the literature becomes genuinely frustrating. Most of these architectures have been evaluated in isolation, or under experimental conditions that differ enough to make cross-study comparison unreliable. Application-specific models have been built, some of them performing well on narrow benchmarks, but the underlying question which backbone actually generalises best to this task has been left largely unanswered. A researcher trying to make an informed architectural decision today has no fair comparison to draw on. It is a surprisingly basic gap for a field that has produced this much work. This thesis addresses it directly. There is, to the best of this review's knowledge, no existing study that places these CNN backbones alongside SegFormer on the same stenosis segmentation task, with identical data, preprocessing, and evaluation criteria [5]. The goal is not to claim a new state-of-the-art result that framing tends to obscure more than it reveals but to produce a standardised, reproducible benchmark that gives the field something it can actually use

[5]. Whether a ResNet from 2015 remains competitive, or whether Transformer-based global context modelling makes a meaningful practical difference, are questions worth answering properly. This thesis attempts to do that.

3. Dataset

3.1. ARCADE Dataset Description

The ARCADE (Automatic Region-based Coronary Artery Disease Diagnostics utilizing X-ray Angiography pictures) dataset [2] is a public dataset for coronary artery analysis which was used in this research. This dataset contains collection of grayscale X-ray Coronary Angiography (XCA) images which are important for training for CAD diagnosis. Semantic segmentation [13] of coronary artery stenosis regions is the main goal for which this dataset was utilized in the current study. Two main directories, "Syntax" and "Stenosis," comprise the dataset; only the "Stenosis" directory was used for this study because it includes the pertinent images and masks for the segmentation task.

3.2. Image Specifications and Annotations

The images in the ARCADE dataset are all in the Portable Network Graphics format (.png) and they all maintain uniform resolution of 512×512 pixel. The associated ground truth annotations for stenosis segmentation are provided following the Common Objects in Context (COCO) format [21]. From these annotations, binary ground truth masks [22] were created. The ARCADE dataset pictures are very consistent. The ground truth annotations for the ARCADE dataset are important, for stenosis segmentation. In these masks, pixels that represent the coronary artery are given an intensity value of 255 (white), whereas all background pixels are assigned a value of 0 (black). This binary format works especially well when training and evaluating segmentation models.

3.3. Data Distribution

The ARCADE dataset is designed in that way so machine learning models will learn from pre specified training, validation and test sets [5]. This also makes it easier to build machine learning models and reproduce research findings down the line.

Dataset Partition Composition

The image and corresponding mask pairs are divided into three subsets:

Table 3.1: Image-Mask Pairs Stenosis

Subset	**Pairs**	**Percentage**	**Primary Purpose**
Training Set	1,000	66.7%	Model learning and parameter optimization
Validation Set	300	20.0%	Hyperparameter tuning and overfitting monitoring [15]
Test Set	200	13.3%	Final unbiased performance evaluation [5]
Total		**1,500**	

Purpose and Rationale

Each subset plays a specific role in the machine learning pipeline:

- **Training Set (1,000 pairs):** The train set is the biggest set of the data collection and it is mainly used when the model is training or learning. The model will search trends and characteristics in these training set data and model will continously updates its internal parameters which are weights and biases. To make sure that the model can learn the actual job which is stenosis segmentation , a large data collection is needed.

- **Validation Set (300 pairs):** This portion of the data essentially stands in for all the real-world data the model never gets to see during training. It is utilized after each training epoch, or at a specified interval, to check the performance of the model. The main functions of this group of data are:

  1. **Hyperparameter Tuning:** Controlling the choice of training methods and model architecture parameters (such as learning rate and network depth).
  2. **Overfitting Monitoring:** Identifying the point at which the performance of the model on training set continues to doing better, but performance of the model on validation set starts to degrade, which means the model has begun to memorize noise instead of learning meaningful features [15].
  3. **Model Selection:** Choosing the best-performing model checkpoint across the training run.

- **Test Set (200 pairs):** This subset is reserved and only used once after the model has been completely trained and all decisions about hyperparameters are made. It is a final, unbiased measure of how the model will actually perform on its own in terms of segmentation [5].

4. Method

4.1. Data Preparation

4.1.1. Ground Truth Mask Generation

The original segmentation labels are available in JSON format, where the labels are in the form of polygonal coordinates that describe the boundary of the coronary arteries. To use these segmentation labels, a custom script has been developed that can read these JSON files and fill the polygons, thus creating a 512x512 pixel binary segmentation mask [22] where pixels corresponding to the arteries are white (255) and pixels corresponding to the background are black (0). Fig. 4.1 shows an image from the dataset and its corresponding segmentation mask.

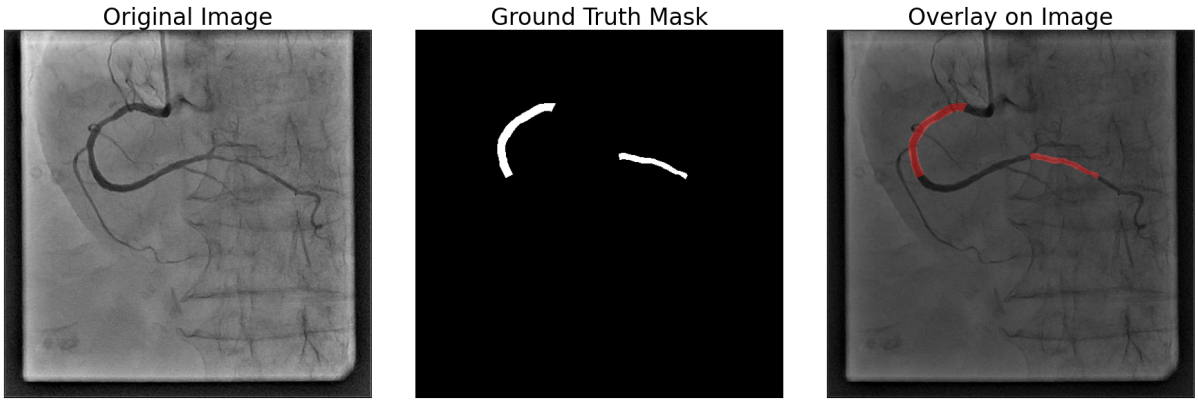

Figure 4.1: Annotated Mask

4.1.2. Data Cleaning

The training set of 1000 images data is verified before starting the actual training. This ensured the integrity of the dataset. From this verification we found out that there are two duplicate images which is redundant. Also three images do not have any JSON annotation files. So, those images were useless for supervised learning. We filtered out the dataset like this and finally we had the training set consisted of 995 image and their corresponding mask.

4.1.3. Data Augmentation Techniques

The training set is small so we applied a thorough data augmentation process [7, 23] to the training set. We used the Albumentations library [24] for that. It ensured decreased overfitting [15] and it made the models capacity for learning better. This process had multiple transformations:

- **Geometric Transforms:** This process makes horizontal flipping, shifting, scaling and rotation of images up to 15 degrees.[23].
- **Non-linear Deformations:** To simulate tissue like variations, elastic and grid distortions were applied
- **Intensity Variations:** Gaussian noise, random adjustments to brightness, contrast, and gamma values were applied.

No image augmentation were applied to validation and test sets as we want an objective assessment of the final models. Augmentation applied on a sample image is showed Fig. 4.3.

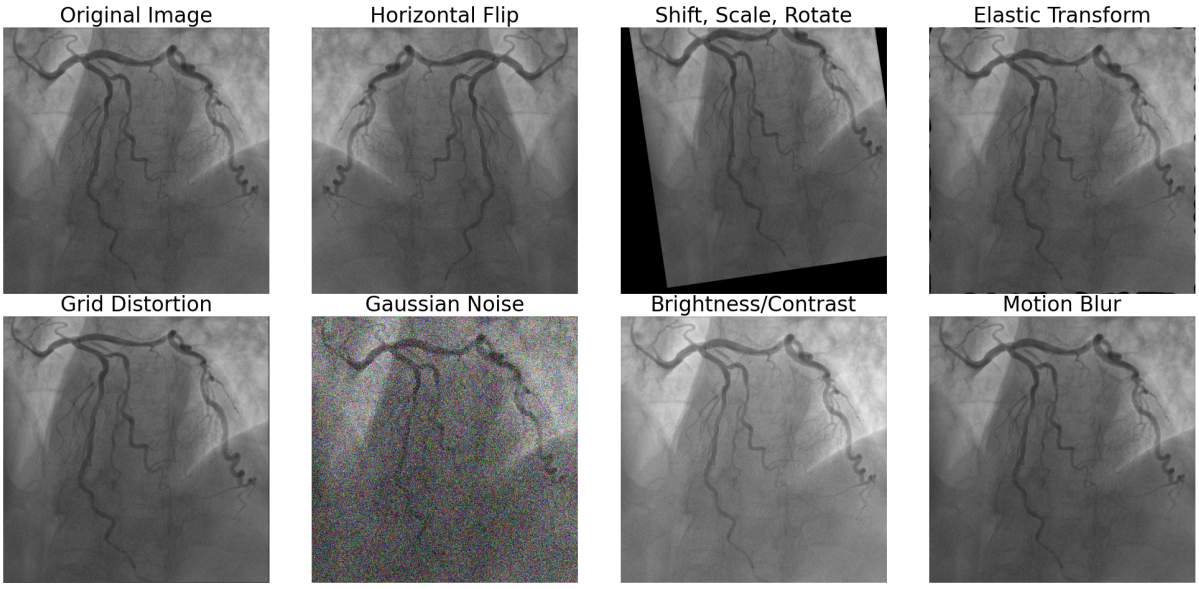

Figure 4.2: Data augmentation techniques examples

4.2. Model Architectures

We have chosen the models based on wide range of architectural differences, complexities and parameter counts. This made sure the thorough comparison between two dominant approach in segmentation task in computer vision which are CNN and ViT.

4.2.1. Key CNN Architectures

CNN architectures shined through 2015 to 2019 where there was a strong debate over how to build the most effective networks for higher performance. This Research mainly evaluated five of the most key and philosophically distinct architectures from that era.

1. ResNet (Deep Residual Learning)

He et al. [16] provided a direct solution ResNet introduced in 2015. The main innovation was "Deep Residual Learning." Instead of forcing a stack of layers to learn a desired underlying mapping $H(x)$, the framework reformulates the layers to learn a residual mapping defined as:

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

The original mapping is then recast as:

$$H(x) = F(x) + x \tag{2}$$

This is implemented via "shortcut" or "skip connections" [25] that bypass one or more layers, performing an identity mapping that is added element-wise to the output of the stacked layers.

Explanation of terms:

- x: Input to the residual block.
- $H(x)$: Original desired mapping the network would traditionally attempt to learn.
- $F(x)$: Residual function learned by the block.
- $F(x) + x$: Final output after combining residual and identity mapping.

The hypothesis, which proved correct, was that it is far easier to optimize the residual $F(x)$ than the original $H(x)$. In the extreme case where an identity mapping is optimal ($H(x) = x$), the network can simply learn to push the residual $F(x)$ to zero, a trivial task. This breakthrough allowed for the effective training of networks that were "substantially deeper" than any previous models, such as the 152-layer ResNet, which was 8x deeper than VGG [26] but had lower complexity [16]. This "depth-centric" philosophy became the new standard and is represented in this thesis by the ResNet18, ResNet34, ResNet50, and ResNet101 models.

2. ResNeXt (Aggregated Residual Transformations)

In 2016, Xie et al. [27] proposed an evolution of the ResNet philosophy. Their architecture, ResNeXt, introduced a new dimension of network design called "cardinality". ResNeXt is based on "aggregated residual transformations," which employs a "split-transform-merge" strategy. A network block is split into multiple (e.g., 32) parallel, low-dimensional transformation paths, each represented mathematically as $T_i(x)$. The aggregated output is:

$$F(x) = \sum_{i=1}^{C} T_i(x) \tag{3}$$

where C denotes cardinality.

The block output is then:

$$y = F(x) + x \tag{4}$$

Explanation of terms:

- C: Cardinality, number of parallel transformation paths.
- $T_i(x)$: Transformation applied by the i^{th} branch.
- $F(x)$: Aggregated residual output.
- y: Block output after adding identity.

The authors demonstrated that increasing cardinality was a more effective and efficient way to gain accuracy than increasing depth or width. For instance, a 101-layer ResNeXt achieved superior accuracy to a 200-layer ResNet while having only 50% of the computational complexity. This "parallelism-centric" philosophy represents a direct competitor to ResNet's depth-focused design and is evaluated in this thesis via the ResNeXt-50 model.

3. DenseNet (Densely Connected Convolutional Networks)

Also in 2016, Huang et al. [28] proposed taking the concept of shortcut connections to its logical extreme. In an L-layer block, this creates:

$$\frac{L(L+1)}{2} \tag{5}$$

direct connections.

Each layer receives the concatenation of all features from previous layers:

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \tag{6}$$

Explanation of terms:

- x_l: Output of the l^{th} layer.
- $[x_0, x_1, \ldots, x_{l-1}]$: Concatenation of all previous feature maps.
- $H_l(\cdot)$: Transformation applied at layer l.

DenseNet alleviates the vanishing-gradient problem [29], strengthens feature propagation, and encourages explicit feature reuse. In our research, DenseNet201 achieved the highest sensitivity which proves this property of DenseNet201. From this study it is proven that DenseNet201 excels at identifying faint stenotic structures .

4. MobileNetV2 (Inverted Residuals and Linear Bottlenecks)

Resource constrained environments often struggles with various architectures and sometimes fails to achieve high performance due to their resource limitation. Sandler et al. [30] introduced MobileNetv2 in 2018 which is a lightweight model that will work in resource constrained environments like mobile devices and it will deliver high performance.

MobileNetv2 works in these sequence of operations:

$$x_{\text{exp}} = PW_{\text{exp}}(x) \tag{7}$$

$$x_{\text{dw}} = DW(x_{\text{exp}}) \tag{8}$$

$$x_{\text{proj}} = PW_{\text{lin}}(x_{\text{dw}}) \tag{9}$$

When shapes match, a shortcut connection is added:

$$y = x + x_{\text{proj}} \tag{10}$$

Explanation of terms:

- PW_{exp}: 1×1 pointwise expansion convolution.
- DW: Depthwise convolution [31].
- PW_{lin}: Linear 1×1 projection convolution.
- x_{proj}: Compressed bottleneck output.

This architecture avoids nonlinear activations in low-dimensional bottlenecks, preventing information loss. This model has efficiency centric philosophy which is well suited for real time clinical deployment [4].

5. EfficientNet (Rethinking Model Scaling)

Tan and Le [8] addressed the network design problem in a different way in 2019. Their idea was that the best performance comes from scaling depth, width, and resolution together using a single compound coefficient ϕ:

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi} \tag{11}$$

subject to:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \tag{12}$$

Explanation of terms:

- d: Network depth.

- w: Network width.
- r: Input resolution.
- ϕ: Compound scaling coefficient.
- α, β, γ: Scaling constants determined via grid search.

This process produced the EfficientNet family (B0-B7). Those models achieved state-of-the-art results also they were smaller and faster models than current competing models. It represented the pinnacle of CNN design.

The summary of those five competing models are presented in Table 4.1. These comparison presents a great debate to find the optimal path to improve the performance of CNN architectures.

Table 4.1: Summary of Foundational CNN Encoder Architectures and Philosophies

Architecture (Citation)	**Core Innovation**	**Key Philosophy / Contribution**
ResNet [16]	Residual Learning (Skip Connections) [25]	Solved the degradation problem, enabling extreme depth (e.g., 100+ layers) to become the standard.
ResNeXt [27]	"Cardinality" (Aggregated Transformations)	Argued that scaling parallel paths ("cardinality") is more effective than scaling depth or width alone.
DenseNet [28]	Dense Connectivity (Feature Concatenation)	Maximized information flow and feature reuse by connecting every layer, leading to high parameter efficiency.
MobileNetV2 [30]	Inverted Residuals & Linear Bottlenecks	Designed for computational efficiency (low-latency) using lightweight depthwise convolutions [31].
EfficientNet [8]	Compound Scaling	Provided a principled method for scaling a network, proving that balancing depth, width, and resolution is key.

4.3. Vision Transformers (ViTs)

The second biggest approach in deep learning is for computer vision is the Vision Transformer (ViT).The details of architecture and its mechanism discuseed. Also it will outline the specific adaptation for semantic segmentation.

4.3.1. Transformer Architecture and Self-Attention Mechanism

The transformer architecture was introduced first in the Natural Language Processing (NLP) field. In that field it became the state-of-the-art for sequence-modeling tasks [32]. The main power of this architecture is the "self-attention mechanism" which provides a new way for model data dependencies.

The self attention mechanism represents a new principle which is totally different from the core principle of CNN architectures. A CNN operates on a fixed, local receptive field [33] (e.g., a 3x3 kernel). CNN progressivly stacks local operations and builds a global understanding of the whole image. Self attention works very differently. It has a global and dynamic receptive field right from the start. When it processes a single element, whether that is a token in language or a patch of pixels in an image, it looks at every other element in the sequence at the same time. It then learns a set of weights that tell the model how important each of those other elements is for understanding the current one.

The idea that this research puts to the test is that this built in ability to capture long range, global dependencies [34] should make Transformers a better fit for tasks that involve complex structures spread across the image. For the segmentation coronary artery, a CNN must stack many layers to connect one end of a long, curvilinear vessel to the other. A Transformer, however, can "model the entire curvilinear structure of a blood vessel" from the outset, potentially leading to more coherent and robust segmentations.

4.3.2. Adaptation for Computer Vision Tasks

The initial adaptation of the Transformer to computer vision, in the seminal ViT model [35], involved dividing an input image into a grid of non-overlapping patches (e.g., 16x16 pixels). These patches are "flattened" into vectors and fed to the Transformer as a sequence of tokens, analogous to words in a sentence.

While highly successful for image-level classification, this standard ViT adaptation is fundamentally ill-suited for dense, pixel-level segmentation. The standard ViT architecture processes the entire sequence of patches and outputs a single feature map at a very low resolution. This is sufficient for assigning one label to the entire image, but it discards the high-resolution spatial information required to assign a label to every pixel, which is the definition of segmentation [13].

4.3.3. SegFormer and MiT Backbone Overview

To overcome this limitation, Xie et al. [36] introduced SegFormer, a "simple and efficient design for semantic segmentation with Transformers". The SegFormer architecture, which is the Transformer model evaluated in this thesis, was purpose-built to solve the problem of adapting Transformers for dense prediction tasks. It consists of two key innovations: a hierarchical encoder and a lightweight decoder.

1. The Mix Transformer (MiT) Encoder: The core of SegFormer is its novel "hierarchical Transformer encoder," named the "Mix Transformer" (MiT). This encoder is explicitly designed to solve the single-scale problem of the standard ViT.

Hierarchical Features: The MiT encoder mimics the behavior of a CNN by producing "CNN-like multi-level features". Using a process of patch merging, it outputs a pyramid of multi-scale feature maps at 1/4, 1/8, 1/16, and 1/32 of the original image resolution.

This is the critical adaptation that makes it a powerful backbone for segmentation.

Efficiency: The design is also highly efficient. It uses fine-grained 4x4 patches for high-resolution detail and is "positional-encoding-free", which allows it to robustly handle input images of varying resolutions without the performance degradation associated with interpolating positional codes.

2. The Lightweight All-MLP Decoder: SegFormer pairs this powerful hierarchical encoder with a surprisingly "lightweight All-MLP decoder". Instead of requiring a complex and computationally heavy decoder (like the expansive path in a UNet), SegFormer's decoder simply takes the multi-level features from the MiT encoder and fuses them using simple Multi-Layer Perceptrons (MLPs).

This "simple and efficient design" proved highly effective, setting a new state-of-the-art in semantic segmentation while being significantly smaller and more efficient than previous Transformer-based methods. This review provides the direct technical explanation for this thesis's findings: the SegFormer-MiT model's superior performance is a direct result of its purpose-built design, which combines the global context-modeling of a Transformer with the multi-scale feature hierarchy of a CNN, making it ideal for segmenting long, complex vessel structures.

4.4. Experimental Setup

4.4.1. Training Protocol

All training experiments were conducted using the PyTorch Lightning framework [37] to ensure high reproducibility. The computational hardware was a Kaggle instance equipped with an NVIDIA P100 GPU with 16GB of VRAM. Models were trained for up to 50 epochs with a batch size of 8 [38]. The Dice Score Coefficient on the validation set was used as the primary metric to guide model selection and save the best-performing checkpoint.

4.4.2. Optimization

The AdamW optimizer [39, 40] was used for all models, which is a variant of the Adam optimizer with improved weight decay handling. It was configured with an initial learning rate of 1×10^{-4} [41]. The learning rate was dynamically adjusted during training using a Cosine Annealing scheduler [39, 42], which gradually decreases the learning rate following a cosine curve to help the model converge to a more optimal minimum.

4.4.3. Regularization Techniques

To further improve model generalization and prevent overfitting [15], two regularization techniques were employed:

- **Early Stopping:** The training process was monitored, and if the validation score did not improve for a set number of epochs, training was halted to prevent the model from overfitting to the training data [43].
- **Stochastic Weight Averaging (SWA):** This technique averages the model weights over the final epochs of training, which has been shown to lead to better generalization on unseen data [44].

In order to formalize these regularization techniques, the following mathematical formulations are provided.

Early Stopping (Formal Definition)

Early stopping prevents overfitting [15] by tracking the validation loss $L_{val}(t)$ at epoch t. Training is halted when the validation loss does not improve for a defined number of epochs, called *patience* (P).

$$L_{val}(t) \geq \min_{0 \leq k \leq t-P} L_{val}(k) \tag{13}$$

Explanation of terms:

- $L_{val}(t)$: Validation loss at epoch t
- P: Patience parameter (number of epochs without improvement allowed)
- Training is stopped at the first epoch t where the inequality holds

This formulation ensures that training stops before the model begins memorizing the training data, thus improving generalization.

Stochastic Weight Averaging (SWA)

SWA computes the average of model weights across the last K training checkpoints. Let w_t denote the model weights at epoch t. The SWA weights are computed as:

$$w_{\text{SWA}} = \frac{1}{K} \sum_{i=1}^{K} w_{t_i} \tag{14}$$

Explanation of terms:

- w_{t_i}: Model weights at selected epochs $t_1, t_2, \ldots, t_K$
- K: Number of weight snapshots included in the averaging
- w_{SWA}: Final averaged model weights used for inference

This averaging process leads to a smoother and wider local minimum in the loss landscape, which is known to improve generalization and stability.

Both techniques were crucial in ensuring that the evaluated deep learning models performed robustly on unseen angiographic data.

4.5. Loss Function

4.5.1. Composite Loss Function Rationale

A composite loss function was utilized in an effective manner to cope with the challenges of the stenosis segmentation task, especially in the presence of considerable class imbalance [45] between the foreground (stenosis) and background regions. Dice Loss [46] and Focal Loss [47] were combined in an effective manner in the loss function so that the benefits of both loss functions are exploited: Dice Loss handles the global shapes, whereas Focal Loss handles the boundaries.

4.5.2. Formulation: Dice Loss and Focal Loss

The final objective function is a linear combination of Dice Loss and Focal Loss:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Dice}} + \lambda \cdot \mathcal{L}_{\text{Focal}} \tag{15}$$

where the balancing coefficient λ was set to 1 in all experiments to give equal importance to both loss components.

The individual components are defined as follows:

- **Dice Loss ($\mathcal{L}_{\text{Dice}}$):** Dice Loss originates from the Dice Score Coefficient [46], which quantifies the spatial overlap between the predicted mask (A) and the ground truth mask (B). It is especially effective for highly imbalanced segmentation tasks [45] because it measures how well the predicted region matches the target region, independent of class frequency.

  The formal expression for Dice Loss is:

  $$\mathcal{L}_{\text{Dice}} = 1 - \frac{2|A \cap B|}{|A| + |B|} \tag{16}$$

  In binary segmentation, the above can be equivalently expressed using true positives (TP), false positives (FP), and false negatives (FN) [6]:

  $$\mathcal{L}_{\text{Dice}} = 1 - \frac{2TP}{2TP + FP + FN} \tag{17}$$

  Explanation of terms:

  - A: Predicted foreground (stenosis) mask

- B: Ground truth stenosis mask
- $|A \cap B|$: Number of pixels correctly predicted as stenosis (true positives)
- TP: True positives
- FP: False positives
- FN: False negatives

Dice Loss optimizes global region similarity, ensuring that the predicted stenosis shape matches the ground truth even when the region is very small.

- **Focal Loss ($\mathcal{L}_{\text{Focal}}$):** Focal Loss extends the standard cross-entropy loss by reducing the contribution from well-classified examples and emphasizing misclassified or ambiguous pixels [47]. This is particularly useful in stenosis segmentation, where boundaries are often faint and difficult to classify.

  The equation for Focal Loss is:

$$\mathcal{L}_{\text{Focal}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{18}$$

  Explanation of terms:

  - p_t: Predicted probability for the true class (foreground or background)
  - α_t: Class-balancing factor compensating for class imbalance [45]
  - γ: Focusing parameter that down-weights easy examples ($\gamma > 0$)
  - $(1 - p_t)^{\gamma}$: Modulation term; large for hard examples, small for easy ones

  A higher value of γ forces the loss to focus increasingly on challenging pixels, such as stenotic regions with weak boundaries or low contrast.

Combining Dice Loss and Focal Loss as in Eq. 15 yields a powerful objective function that simultaneously optimizes global region overlap and fine boundary accuracy—two requirements that are critical for reliable stenosis segmentation.

4.6. Evaluation Metrics

4.6.1. Confusion Matrix: Definition and Components

Understanding how a model's predictions relate to the ground truth is, in practical terms, the starting point for any honest evaluation of segmentation performance. The **confusion matrix** [6] provides a structured way to make that comparison explicit. For binary segmentation tasks where the image is divided into stenosis foreground and background four components define the matrix:

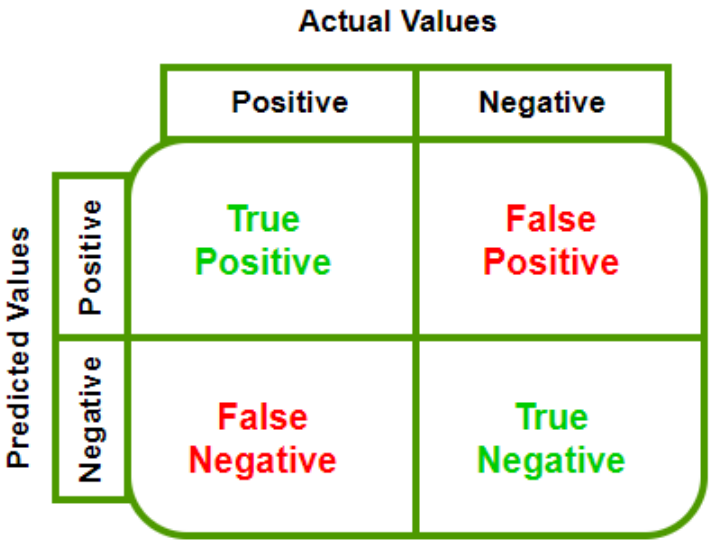

Figure 4.3: Graphical Representation of Confusion Matrix

Explanation of terms:

- **True Positive (TP):** Pixels correctly predicted as stenosis (foreground). These correspond to diseased regions the model successfully identified.
- **False Positive (FP):** Background pixels incorrectly labelled as stenosis. These reflect over-segmentation the model flagging regions that are not actually diseased.
- **True Negative (TN):** Background pixels correctly classified as background. These represent regions the model correctly left alone.
- **False Negative (FN):** Stenosis pixels misclassified as background. These are missed diseased regions and clinically, they are the most consequential type of error.

Each of these values feeds directly into the evaluation metrics used throughout this study. Broadly speaking, higher TP and TN counts reflect stronger model performance, while elevated FP and FN values point to segmentation failures of different kinds.

4.6.2. Metric Definitions and Formulas

Four established evaluation metrics were used to assess each model's segmentation accuracy. Taken together, they capture not just how often a model is correct, but whether it is failing in ways that matter clinically — missing stenosis regions, generating false alarms, or both.

- **Dice Score Coefficient (DSC):**
  The Dice Score, also referred to as the Sørensen–Dice coefficient [46], quantifies the degree of spatial overlap between the predicted and ground truth masks. It is

particularly well-suited to medical image segmentation tasks where the foreground class is small and imbalanced [45].

$$\text{Dice} = \frac{2TP}{2TP + FP + FN} \tag{19}$$

Explanation:

- Numerator: $2TP$ gives additional weight to correctly segmented stenosis pixels.
- Denominator: Accounts for all pixels contributing to overlap and errors.
- Value Range: 0 to 1; higher values indicate better segmentation.

- **Intersection over Union (IoU):**
  Also known as the Jaccard Index [48], IoU measures the ratio of overlap to total coverage between the predicted and ground truth masks.

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

Explanation:

- Captures strict overlap between predicted and true stenosis regions.
- Somewhat more penalising than Dice, since it does not double-count true positives.
- Value Range: 0 to 1.

- **Sensitivity (Recall):**
  Sensitivity reflects how reliably the model detects stenosis regions — that is, how few diseased pixels it fails to catch [6].

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

Explanation:

- Denominator: All pixels that are genuinely stenotic.
- High sensitivity means fewer missed diseased regions.
- Low sensitivity carries real clinical risk — a false negative here means undetected stenosis.

- **Specificity:**
  Specificity measures how well the model avoids misclassifying healthy background pixels as stenosis [6].

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

  Explanation:

  - Denominator: All pixels that are genuinely background.
  - High specificity means the model generates few false alarms.
  - This matters practically — over-segmentation can trigger unnecessary clinical concern.

Used in combination, these metrics offer a clinically grounded picture of model performance one that goes beyond raw accuracy to ask whether a model detects what it should, and refrains from flagging what it should not.

5. Results and Discussions

5.1. Quantitative Performance Analysis

Quantitative performance of all ten models was assessed on an unseen set of test images [5] for an unbiased measure of their segmentation accuracy. From the results provided in Table 5.1, it is clear that there is a distinct performance ranking of various approaches. Also, validation set results are provided in Table 5.2.

The Transformer-based SegFormer-MiT-B3 architecture was found to perform best among all models for all primary evaluation metrics. It recorded the best DSC of 0.632 [46] and IoU of 0.462 [48] among all models. Therefore, it is clear that the Transformer architecture has immense potential for this specific problem.

Among all eight CNN-based architectures, it was found that ResNet101 recorded the best DSC of 0.604. Therefore, it is clear that deeper and more complex residual networks are highly effective. However, they are still outperformed by the best Transformer-based model. It was found that the DenseNet201 model recorded the best sensitivity of all models at 0.670. However, this came at the expense of precision, resulting in a lower DSC than all other models.

Table 5.1: Test Set Performance Comparison

Model	**DSC**	**IoU**	**Loss**	**Sensitivity**	**Specificity**
MobileNet	0.576	0.434	0.470	0.632	0.996
ResNet50	0.595	0.4237	0.4122	0.6169	0.9955
DenseNet121	0.540	0.401	0.514	0.621	0.995
DenseNet201	0.541	0.396	0.521	**0.670**	0.994
EfficientNet-B3	0.555	0.413	0.496	0.668	0.994
ResNeXt	0.576	0.434	0.470	0.632	0.996
ResNet101	0.604	0.433	0.402	0.641	0.995
ResNet18	0.584	0.412	0.421	0.619	0.995
MiT-B3	**0.632**	**0.462**	**0.373**	0.637	0.996
MiT-B0	0.590	0.418	0.416	0.621	0.995

Table 5.2: Validation Set Performance Comparison

Model	**Dice**	**IoU**	**Loss**	**Sensitivity**	**Specificity**
MobileNetV2	0.611	0.447	0.427	0.623	0.996
ResNet101	0.622	0.452	0.381	0.614	**0.997**
ResNet18	0.595	0.423	0.410	0.579	**0.997**
MiT-B3	**0.655**	**0.487**	**0.347**	**0.651**	**0.997**
MiT-B0	0.590	0.418	0.412	0.627	0.996
ResNet50	0.616	0.446	0.391	0.624	**0.997**
ResNeXt-50	0.611	0.447	0.427	0.623	0.996
EfficientNet-B3	0.579	0.408	0.475	0.598	0.583
DenseNet121	0.583	0.412	0.464	0.611	0.996
DenseNet201	0.569	0.399	0.484	0.579	0.996

5.2. Qualitative Comparison Analysis

In order to provide an additional comparison of the models' segmentation outputs, a qualitative comparison was carried out using a subset of the test set, as shown in Figure

5.1. In this figure, the qualitative comparison of the segmentation outputs of the models was carried out, particularly in cases where the structures of the blood vessels, such as those that are faint, thin, and complex, are most difficult to segment using the models.

From the qualitative comparison of the segmentation outputs of the models, it can be noted that the segmentation masks generated by the best Transformer-based model, MiT-B3, have better continuity compared to the segmentation masks generated by the best CNN-based model, ResNet101, which had relatively more fragmented segmentation masks with clear discontinuity and less accurate boundaries in the segmentation of the blood vessel structures, compared to the Transformer-based models.

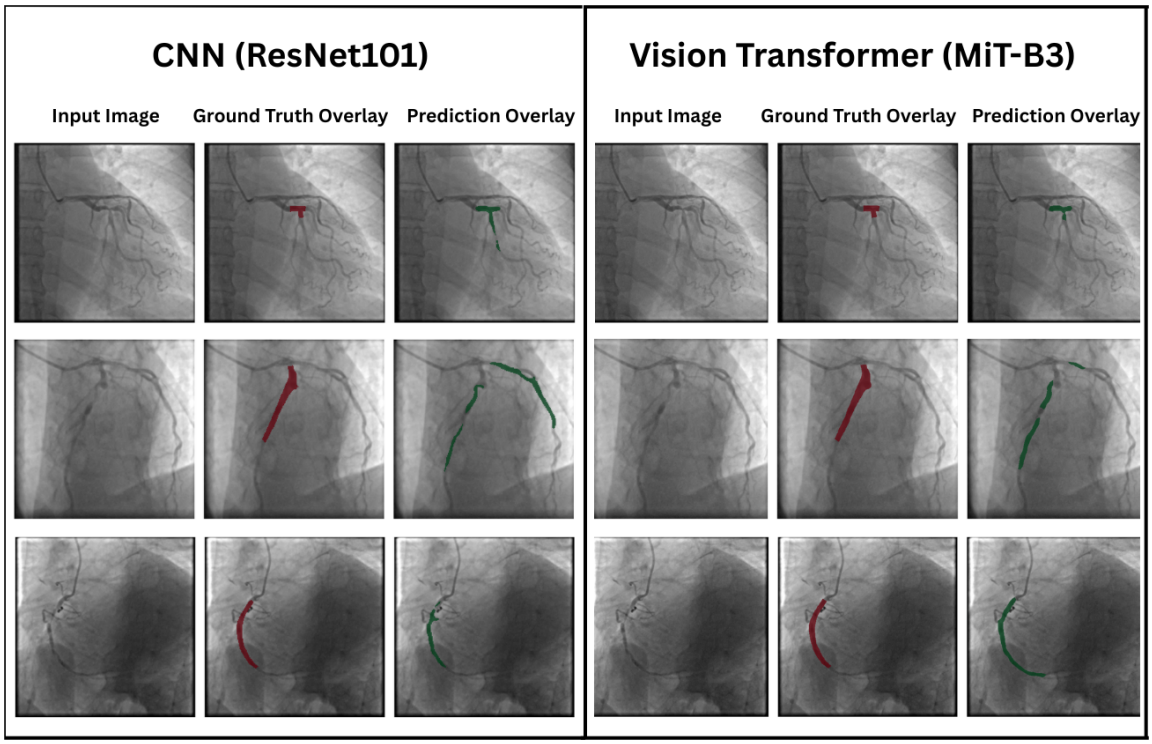

Figure 5.1: Qualitative comparison of segmentation results on the test set: CNN (ResNet101) vs. Transformer (MiT-B3). Predicted masks are shown alongside ground truth for visual evaluation.

5.3. Model Efficiency Analysis

To assess the practical viability of the evaluated models for potential clinical deployment [4], where computational resources may be limited, an analysis of the trade-off between segmentation accuracy and model size was conducted. Model size, used as a proxy for computational cost, was measured by the total number of trainable parameters. As detailed in Table 5.3, while the largest model, MiT-B3 (45.2M parameters), yielded the highest DSC, several lightweight models demonstrated notable efficiency. MiT-B0, the smallest Transformer variant, achieved a competitive DSC of 0.590 with only 3.7M parameters. MobileNetV2 performed comparably, scoring a DSC of 0.576 with just 3.5M parameters, which is a result that is difficult to dismiss given how little it costs compu-

tationally.

Table 5.3: Performance and efficiency comparison of all models.

Model	**Architecture**	**Params (M)**	**DSC**	**Efficiency Score***
CNN	3.5	0.576	1.06 ResNet18	CN
11.7	0.584	0.54 ResNet50	CNN	25.
0.595	0.42 ResNet101	CNN	44.5	0.60
0.37 ResNeXt-50	CNN	25.0	0.576	0.41 Dens
CNN	8.0	0.540	0.60 DenseNet201	CN
20.0	0.541	0.42 EfficientNet-B3	CNN	12.
0.555	0.50 heightMiT-B0	Transformer	3.7	0.59
1.04 MiT-B3	Transformer	45.2	0.632	0.38 h

To further quantify this relationship, a custom efficiency score was calculated for each model using the following formula:

$$\text{Efficiency Score} = \frac{\text{DSC}}{\log_{10}(\text{Parameters} \times 10^6)} \tag{23}$$

This analysis suggests that MiT-B0 and MobileNetV2 occupy a genuinely useful middle ground, delivering strong segmentation performance at a fraction of the computational cost demanded by larger architectures such as MiT-B3 and ResNet101. The broader implication is that real-time or near-real-time automated stenosis detection is feasible without requiring extensive computational infrastructure.

5.4. Training Dynamics and Generalization

The training and validation learning curves for the best-performing CNN (ResNet101) and Transformer (MiT-B3) models are presented in Figures **??** and **??** respectively. These plots offer a useful window into how each model learned and how well it generalised beyond the training data. For both models, the loss curves decreased consistently and smoothly before converging toward the later epochs, which points to stable and well-behaved learning. Notably, neither model showed significant divergence between its training and validation curves, whether for loss or F1 score. That degree of alignment between the two curves suggests the models generalised well to unseen validation data, and that the regularisation strategies employed, particularly data augmentation and early stopping [43], were effective in keeping overfitting in check [15]. The stable convergence observed in both top-performing models reflects the soundness of the overall training protocol.

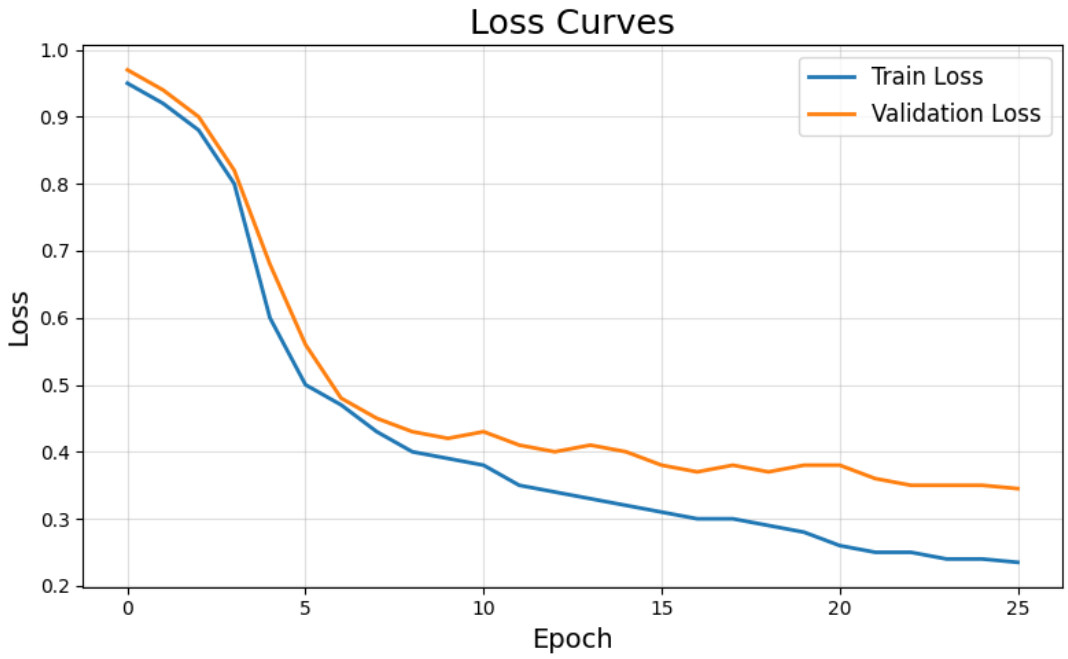

Figure 5.2: Training and validation loss trends across epochs for the best performing Transformer model (MiT-B3).

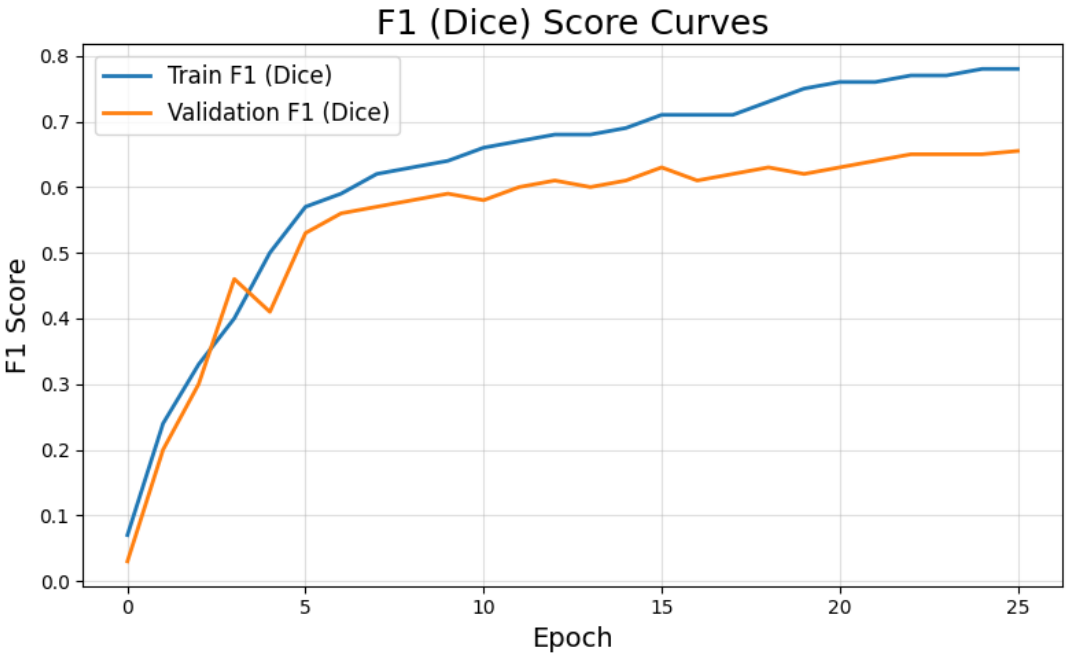

Figure 5.3: Training and validation Dice coefficient trends across epochs for the best performing Transformer model (MiT-B3).

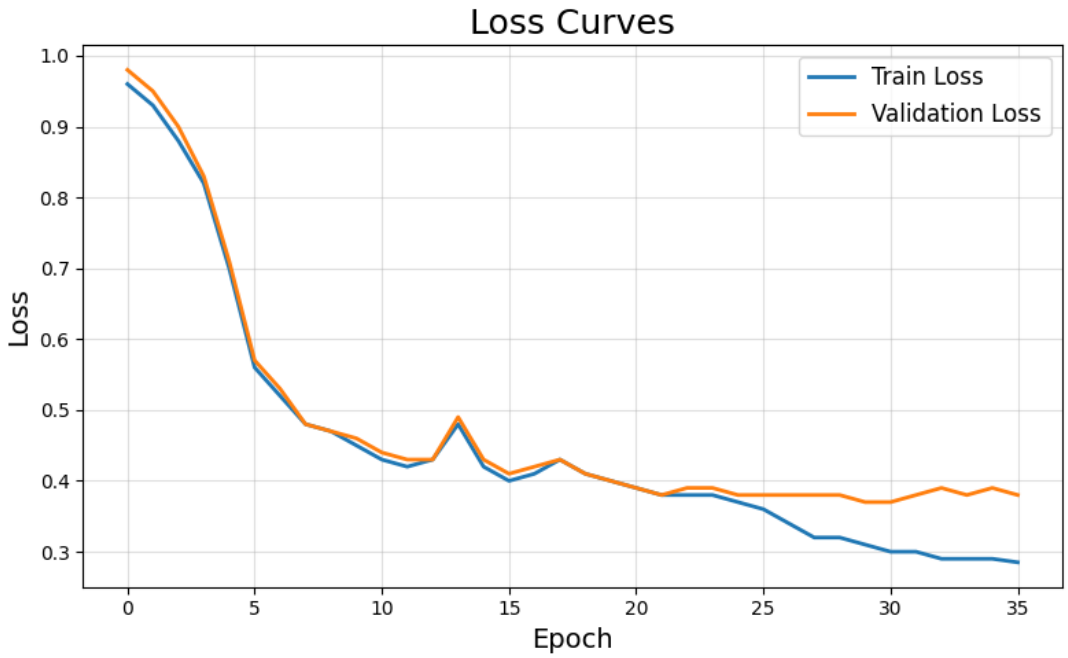

Figure 5.4: Training and validation loss trends across epochs for the best performing CNN model (ResNet101).

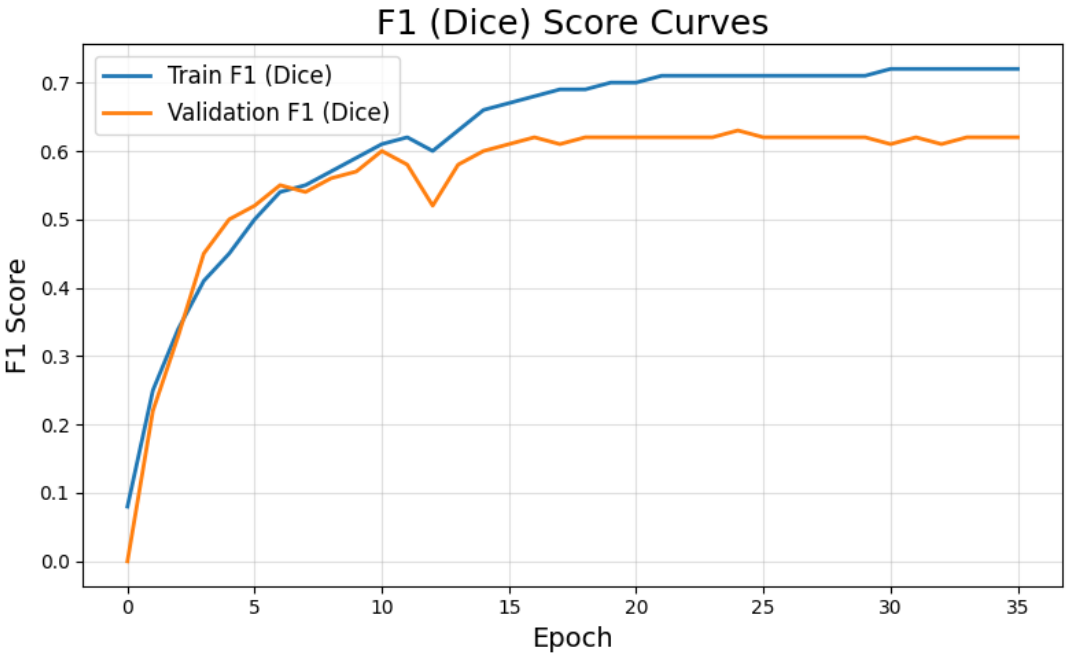

Figure 5.5: Training and validation Dice coefficient trends across epochs for the best performing CNN model (ResNet101).

5.5. Ablation Studies

To validate the effectiveness of key components of the experimental design, ablation studies were conducted on the top-performing MiT-B3 model. The results, shown in Table 5.4, confirm the contribution of the chosen training configurations.

- **Impact of Data Augmentation:** Removing the data augmentation pipeline [7, 23] produced a 4.0
- **Impact of Composite Loss Function:** Replacing the composite Dice+Focal loss with a standard Dice Loss led to a 2.5

Taken together, these results confirm that both the augmentation strategy and the composite loss function contributed meaningfully to the performance levels reported in this study. Neither was incidental to the outcome.

Table 5.4: Ablation study on the MiT-B3 model's training configuration.

Component Removed	**DSC**	**Absolute Change**	**Relative Change**
None (Full Model)	**0.632**	—	—
Data Augmentation	0.607	-0.025	-4.0%
Composite Loss (Focal)	0.616	-0.016	-2.5%

5.6. Discussion

The most consistent finding across the experiments is that Transformer-based architectures outperformed CNNs on the coronary artery stenosis segmentation task. This pattern is arguably explained by the self-attention mechanism that Transformers rely on [32], which allows the model to process the entire image simultaneously and capture long-range spatial dependencies. CNNs, by contrast, are constrained by fixed local receptive fields [33], which limits their ability to reason about structure beyond a small neighbourhood at a time. For a task that involves tracing the continuous geometry of a curved blood vessel, that global perspective appears to be a meaningful advantage, particularly under conditions of noise, imaging artefacts, and anatomical variation. Within the ResNet family, deeper models generally produced stronger results, which is consistent with the expectation that increased depth expands representational capacity. That said, even the deepest ResNet variants appeared to plateau, which may reflect the aggressive early downsampling built into the encoder, a design choice that discards fine-grained spatial detail before the network has had a chance to use it. A separate but related pattern worth noting is the sensitivity-specificity trade-off [6]. DenseNet201 illustrates this clearly. Its sensitivity of 0.670 was the highest recorded across all models, suggesting that its feature reuse mechanism makes it particularly adept at detecting faint or subtle stenosis signals. However, its specificity was among the lowest at 0.994, meaning this detection ability came with a higher rate of false positives. ResNet101, by comparison, presented a more balanced profile across both metrics. In a clinical context where false alarms carry real consequences, that balance is arguably as important as raw detection performance.

6. Conclusion

6.1. Summary

This thesis set out to compare deep learning architectures for coronary artery stenosis segmentation in X-ray angiography images in a controlled, systematic way. The central finding is that Transformer-based models consistently outperform traditional CNNs on this task. The SegFormer model with a MiT-B3 encoder achieved the strongest results overall, recording the highest Dice Score Coefficient (DSC) [46] and Intersection over Union (IoU) [48] across all architectures evaluated. Its segmentations were noticeably more faithful to anatomical structure, particularly in tracing the curved and branching geometry of coronary vessels. The study also looked beyond raw accuracy to consider what deployment in real clinical settings would actually require [4]. Lightweight models such as MiT-B0 and MobileNetV2 demonstrated that competitive segmentation performance is achievable with substantially fewer parameters, which points to a viable efficiency trade-off for resource-constrained environments. The results also reinforced the theoretical advantage of Transformers in capturing long-range spatial dependencies [34], which appears to translate into greater resilience against imaging noise and vessel appearance variability. Finally, the ablation studies validated the core experimental choices: aggressive data augmentation [7] and the composite Dice+Focal loss function [46, 47] each contributed meaningfully to final segmentation accuracy and generalisation.

6.2. Limitations

This study has several limitations that are worth addressing honestly, both to contextualise what the results can and cannot claim, and to identify where future work is most needed.

- **Only One Dataset Was Used:** All experiments were conducted on a single publicly available dataset, ARCADE [2]. While this supports reproducibility, it also means the reported performance may not transfer cleanly to data collected at other institutions using different equipment, imaging protocols, or patient demographics. Generalisation beyond this dataset remains an open question.
- **Restricted to 2D Images:** The study operates entirely on 2D image frames, even though coronary angiography is fundamentally a dynamic, three-dimensional

process. Working with static 2D snapshots means the models do not have access to temporal information such as blood flow dynamics or the spatial context that multi-angle acquisition could provide.

- **No Severity Estimation:** The models identify where a stenosis is located, but offer no information about its severity. Estimating the percentage of arterial narrowing is precisely what clinicians need to assess disease progression and guide treatment decisions, and this capability is absent from the current framework.

6.3. Future Work

The limitations outlined above point fairly directly toward the most productive directions for future research.

- **Testing on Larger, More Diverse Datasets:** Validating the best-performing models on multi-centre datasets drawn from different hospitals and patient populations would be an important next step. This kind of external validation is what separates models that perform well under controlled conditions from tools that could realistically be trusted in clinical practice.

- **Moving to 3D Spatio-Temporal Models:** Extending the current framework to handle full 3D angiographic sequences [49] represents a compelling research direction. Rather than processing isolated frames, such models could learn from both the spatial structure of vessels and the temporal dynamics of blood flow, potentially yielding more robust and informative stenosis detection.

- **Adding Stenosis Severity Estimation:** For this work to have genuine clinical utility, the ability to quantify stenosis severity would need to be incorporated alongside localisation. Attaching a regression head to predict arterial narrowing percentage from the learned feature representations is one plausible approach. This would shift the system from a segmentation-only tool [13] toward something closer to a complete diagnostic aid for Coronary Artery Disease.

Bibliography

[1] P. J. Scanlon, D. P. Faxon, A.-M. Audet, B. Carabello, G. J. Dehmer, K. A. Eagle, R. D. Legako, D. F. Leon, J. A. Murray, S. E. Nissen, C. J. Pepine, R. M. Watson, J. L. Ritchie, R. J. Gibbons, M. D. Cheitlin, K. A. Eagle, T. J. Gardner, A. Garson, R. O. Russell, T. J. Ryan, and S. C. Smith, "Acc/aha guidelines for coronary angiography123: A report of the american college of cardiology/american heart association task force on practice guidelines (committee on coronary angiography) developed in collaboration with the society for cardiac angiography and interventions," *Journal of the American College of Cardiology*, vol. 33, no. 6, p. 1756–1824, May 1999. [Online]. Available: http://dx.doi.org/10.1016/S0735-1097(99)00126-6

[2] A. Popov and others, "ARCADE: Automatic region-based coronary artery disease diagnostics using X-ray angiography images," *Journal of Medical Imaging*, vol. 10, no. 3, p. 034501, 2023.

[3] L. Quinn, K. Tryposkiadis, J. Deeks, H. C. De Vet, S. Mallett, L. B. Mokkink, Y. Takwoingi, S. Taylor-Phillips, and A. Sitch, "Interobserver variability studies in diagnostic imaging: a methodological systematic review," *The British Journal of Radiology*, vol. 96, no. 1148, Aug. 2023. [Online]. Available: http://dx.doi.org/10.1259/bjr.20220972

[4] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, p. 24–29, Jan. 2019. [Online]. Available: http://dx.doi.org/10.1038/s41591-018-0316-z

[5] P. Refaeilzadeh, L. Tang, and H. Liu, *Cross-Validation*. Springer US, 2009, p. 532–538. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-39940-9_565

[6] S. Sathyanarayanan, "Confusion matrix-based performance evaluation metrics," *African Journal of Biomedical Research*, p. 4023–4031, Nov. 2024. [Online]. Available: http://dx.doi.org/10.53555/AJBR.v27i4S.4345

[7] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, Jul. 2019. [Online]. Available: http://dx.doi.org/10.1186/s40537-019-0197-0

[8] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, 2019, pp. 6105–6114.

[9] World Health Organization, "Cardiovascular diseases (cvds) fact sheet," 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[10] S. Achenbach and W. G. Daniel, "Computed tomography of the coronary arteries," *Journal of the American College of Cardiology*, vol. 46, no. 1, p. 155–157, Jul. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.jacc.2005.05.006

[11] A. Gupta and others, "End-to-end deep-learning model for the detection of coronary artery stenosis on coronary CT images," *Open Heart*, vol. 12, no. 1, p. e002998, 2024.

[12] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve." *Radiology*, vol. 143, no. 1, p. 29–36, Apr. 1982. [Online]. Available: http://dx.doi.org/10.1148/radiology.143.1.7063747

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2014. [Online]. Available: https://arxiv.org/abs/1411.4038

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2009, p. 248–255. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2009.5206848

[15] D. M. Hawkins, "The problem of overfitting," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 1, p. 1–12, Dec. 2003. [Online]. Available: http://dx.doi.org/10.1021/ci0342472

[16] 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.

[17] H. Lin and others, "StenUNet: Automatic stenosis detection from X-ray coronary angiography," *Medical Image Analysis*, vol. 84, p. 102689, 2023.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[19] J. Lee and others, "SSASS: A semi-supervised approach for stenosis segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1345–1356, 2023.

[20] M. Rostami and others, "Segmentation of coronary artery stenosis in X-ray angiography using Mamba models," *arXiv preprint arXiv:2412.02568*, 2024.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft COCO: Common Objects in Context*. Springer International Publishing, 2014, p. 740–755. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10602-1_48

[22] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," 2016. [Online]. Available: https://arxiv.org/abs/1606.04797

[23] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017. [Online]. Available: https://arxiv.org/abs/1712.04621

[24] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020. [Online]. Available: http://dx.doi.org/10.3390/info11020125

[25] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015. [Online]. Available: https://arxiv.org/abs/1505.00387

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[27] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492–1500.

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

[29] Wikipedia contributors, "Vanishing gradient problem — Wikipedia, the free encyclopedia," 2025, [Online; accessed 5-December-2025]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Vanishing_gradient_problem&oldid=1314303370

[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[31] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2016. [Online]. Available: https://arxiv.org/abs/1610.02357

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[33] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1701.04128

[34] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1711.07971

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[36] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[37] W. Falcon and K. Cho, "A framework for contrastive self-supervised learning and designing a new approach," 2020. [Online]. Available: https://arxiv.org/abs/2009.00104

[38] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," 2018. [Online]. Available: https://arxiv.org/abs/1804.07612

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017. [Online]. Available: https://arxiv.org/abs/1711.05101

[41] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Mar. 2017, p. 464–472. [Online]. Available: http://dx.doi.org/10.1109/WACV.2017.58

[42] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," 2016. [Online]. Available: https://arxiv.org/abs/1608.03983

[43] L. Prechelt, "Early stopping - but when?" in *Neural Networks: Tricks of the Trade*. Springer, 1998, pp. 29–36.

[44] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," 2018. [Online]. Available: https://arxiv.org/abs/1803.05407

[45] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, p. 249–259, Oct. 2018. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2018.07.011

[46] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, p. 297–302, Jul. 1945. [Online]. Available: http://dx.doi.org/10.2307/1932409

[47] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017. [Online]. Available: https://arxiv.org/abs/1708.02002

[48] Wikipedia contributors, "Jaccard index — Wikipedia, the free encyclopedia," 2025, [Online; accessed 5-December-2025]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Jaccard_index&oldid=1311865137

[49] O. undefinediçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," 2016. [Online]. Available: https://arxiv.org/abs/1606.06650