Internship Project Report

Enhancing ROP Image Analysis through Synthetic Data and Attention Mechanisms

By:
Ashwath Vaithinathan Aravindan,
Department of Computer Science and Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
vaashwath@gmail.com

Guided by:
Dr. Raul Benitez,
Scientific Researcher
Biomedical Image analysis Laboratory,
Biomedical Engineering Research Center,
Universitat Politècnica de Catalunya
raul.benitez@upc.edu

BENITEZ IGLESIAS RAUL - 43735711S
Digitally signed by BENITEZ IGLESIAS RAUL - 43735711S
Date: 2024.07.19 14:56:31 +02'00'

Submitted on: 10/07/2024

1 Table of Contents

2 ACKNOWLEDGEMENT

I would like to express my sincerest gratitude to Dr. Raul Benitez, Scientific Researcher at the Biomedical Engineering Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain, for his invaluable guidance, mentorship, and support throughout my research internship. His expertise, encouragement, and dedication have been instrumental in shaping my understanding and enhancing my skills in the field of Explainable AI for medical applications.

I would also like to extend my thanks to the Computer Science and Engineering department, Amrita School of Computing, for their support during this internship.

I am deeply grateful for the opportunity to work under Dr. Benitez's supervision and to be a part of such a dynamic and supportive research environment. This experience has been enriching and transformative, and I am truly appreciative of all the knowledge and insights gained.

3 ABSTRACT

This project focuses on developing an interpretable deep learning system for the detection and diagnosis of Retinopathy of Prematurity (ROP), a critical condition affecting premature infants' retinas. The primary objectives include the creation of a synthetic dataset generator for ROP images capable of producing diverse and representative retinal samples. This generator aims to augment training datasets effectively, enhancing model robustness.

The study conducts a comparative analysis between Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to assess their efficacy in ROP identification. Special emphasis is placed on leveraging their attention mechanisms to improve accuracy and precision in localizing disease features within retinal images. Key metrics such as accuracy, sensitivity, and specificity are evaluated to determine each architecture's performance and interpretability.

The outcomes of this research not only contribute to advancing ROP detection technology but also provide valuable insights into the strengths of CNNs and ViTs in medical image analysis. By facilitating more accurate diagnoses through interpretable deep learning models, this project aims to support clinicians in making informed decisions about patient care, thus improving outcomes for infants at risk of ROP.

All related code is available at: https://github.com/Mystic-Slice/Enhancing-ROP-Image-Analysis-through-Synthetic-Data-and-Attention-Mechanisms

4 LIST OF FIGURES

5 LIST OF ABBREVIATIONS

ROP - Retinopathy of Prematurity

CNN - Convolutional Neural Network

GradCAM - Gradient-weighted Class Activation Mapping

ViT - Vision Transformer

RGB - Red, Green, Blue (color channels)

ReLU - Rectified Linear Unit

NLP - Natural Language Processing

6 INTRODUCTION

6.1 BACKGROUND

Retinopathy of Prematurity (ROP) is a potentially blinding eye disorder that primarily affects premature infants with low birth weight and low gestational age. Early detection and timely treatment are crucial for preventing vision impairment in these vulnerable patients. Traditionally, ROP screening and diagnosis have relied on manual examination by ophthalmologists using indirect ophthalmoscopy, a time-consuming and subjective process.

With recent advancements in deep learning and computer vision, automated systems for ROP diagnosis have shown promising results in accurately detecting and grading the disease from retinal images. These deep learning models, trained on large datasets of annotated retinal images, have demonstrated high accuracy and efficiency, offering the potential for scalable and consistent ROP screening.

However, despite their impressive performance, these deep learning models often operate as "black boxes," making it challenging to understand and interpret their decision-making processes. This lack of interpretability raises significant concerns in the medical domain, where transparency, accountability, and trust are paramount. Healthcare professionals, responsible for the well-being of infants, may be hesitant to rely solely on opaque models, especially in critical scenarios involving the diagnosis and management of ROP.

While existing deep learning models have shown promise in automating ROP detection and grading, their opaque nature poses challenges in gaining trust and acceptance among physicians and neonatal care providers. The inability to comprehend the underlying reasoning and decision-making processes of these models can hinder their adoption and integration into clinical workflows.

Therefore, there is a pressing need for interpretability techniques that can shed light on the inner workings of deep learning models for ROP diagnosis. By enhancing the interpretability of these models, we can provide healthcare professionals with insights into the visual patterns, features, and disease manifestations that the models rely on, fostering trust and enabling more informed decision-making in neonatal care settings.

6.2 PROBLEM STATEMENT

To build a synthetic dataset generator for ROP images, facilitating the creation of diverse and representative retinal image samples for research and development. Additionally, conduct a comparative study on Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), leveraging their attention mechanisms to enhance the accuracy and precision of localization. The goal is to not only assess the performance differences between these architectures but also to provide insights into their attention-based interpretability and effectiveness in pinpointing relevant features within the images.

6.3 SPECIFIC OBJECTIVE

Here are the specific objectives for the internship project on exploring an interpretable deep learning system for Retinopathy of Prematurity (ROP) detection and diagnosis:

1. Develop a synthetic dataset generator for ROP images that can create diverse and representative retinal image samples. This generator should be capable of producing high-quality synthetic images that capture the various manifestations of ROP, enabling researchers and developers to augment their training datasets effectively and enhance the robustness of their models.
2. Train and evaluate CNN and ViT architectures for the identification of ROP. This involves implementing and fine-tuning both types of models to assess their performance in detecting ROP from retinal images. The evaluation should focus on key metrics such as accuracy, sensitivity, and specificity to determine the effectiveness of each architecture in identifying the disease.
3. Conduct a comparative study on Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) using their attention mechanisms for localization. This study should assess the performance differences between these two architectures, leveraging their unique attention-based features to improve the accuracy and precision of localization within the retinal images. The goal is to provide insights into the interpretability and effectiveness of these models in highlighting relevant regions or features, aiding in the accurate detection and diagnosis of ROP.

By achieving these objectives, the proposed interpretable deep learning system aims to foster a collaborative environment where human experts and artificial intelligence can work in tandem, leveraging the model's predictions and explanations to make informed decisions in the diagnosis and management of Retinopathy of Prematurity.

6.4 LIMITATIONS

The project is limited to the determination of the presence/absence of ROP, but it doesn’t focus on the different types of ROP. It is left to the physician to determine the severity of the disease and decide on the appropriate treatment. The project is restricted to providing the physician with visual and quantifiable measures to help in their diagnosis.

7 LITERATURE SURVEY

7.1 IMAGE CLASSIFICATION

Image classification is a fundamental task in computer vision, which aims to assign a class label to an input image based on its visual content. Traditionally, image classification relied on hand-crafted feature extraction techniques, such as SIFT [1] and HOG [2], followed by classical machine learning algorithms like Support Vector Machines (SVMs) [3] or Random Forests [4]. However, with the advent of deep learning, data-driven approaches have revolutionized the field of image classification.

Convolutional Neural Networks (CNNs) [5] have been the most successful and widely used deep learning models for image classification tasks. CNNs are designed to automatically learn hierarchical representations of image features through a series of convolutional and pooling layers, followed by fully connected layers for classification. The groundbreaking work of Krizhevsky et al. [6] demonstrated the superior performance of CNNs on the ImageNet dataset[7], sparking a surge of research interest in developing more advanced CNN architectures.

Over the years, various CNN architectures have been proposed, including AlexNet, VGGNet , GoogLeNet, ResNet, and DenseNet, among others. These architectures have introduced innovative concepts, such as deeper network depths, skip connections, and dense connectivity, enabling the capture of more intricate and abstract image features, leading to improved classification performance.

While CNNs excel at capturing local spatial patterns and hierarchical features, they may struggle with modeling long-range dependencies and global context within images. To address this limitation, Vision Transformers (ViTs) [8] have emerged as a promising alternative, adapting the self-attention mechanism from the Transformer architecture, which was initially proposed for natural language processing tasks.

ViTs treat an input image as a sequence of flattened patches and apply self-attention mechanisms to capture global relationships between these patches. By leveraging the self-attention mechanism, ViTs can effectively model long-range dependencies and capture contextual information, which is crucial for tasks such as object detection and segmentation.

Since the introduction of the original ViT, various improvements and variations have been proposed, including the DeiT, Swin Transformer, and Pyramid Vision Transformer. These advancements aim to enhance the performance, efficiency, and scalability of Vision Transformers for image classification and other computer vision tasks.

Both CNNs and Vision Transformers have demonstrated impressive performance on various benchmark datasets, such as ImageNet, CIFAR, and MS-COCO. However, the choice between these architectures often depends on the specific task, computational resources, and dataset characteristics.

7.2 ATTENTION CAPTURE TECHNIQUES

Attention capture techniques have emerged as crucial tools for interpreting and understanding the decision-making processes of deep neural networks, particularly in the context of Convolutional Neural Networks (CNNs) and Vision Transformers. GradCAM[9] (Gradient-weighted Class Activation Mapping) is one such technique widely employed for visualizing and localizing the regions of input images that contribute most significantly to the model's predictions. Through gradient-based methods, GradCAM generates heatmaps that highlight the regions of high activation, providing valuable insights into the model's attention mechanism. For instance, in medical image analysis, GradCAM has been utilized to interpret CNN-based models for diagnosing diseases such as diabetic retinopathy. Similarly, Attention Rollout maps[10] offer a novel approach to visualizing attention patterns within Transformer models, facilitating a deeper understanding of their decision-making processes. Attention Rollout systematically analyzes the impact of each token on the model's output, thereby constructing a comprehensive picture of attention distribution across the input sequence. This technique has been applied in various natural language processing tasks, including machine translation and sentiment analysis, to interpret and analyze the attention mechanisms of Transformer models.

In addition to GradCAM and Attention Rollout, several other attention capture techniques have been proposed for CNNs and Vision Transformers. GradCAM++, an extension of GradCAM, improves the localization accuracy by incorporating both positive and negative gradients in the heatmap generation process. Excitation Backpropagation (EBP) is another method that highlights relevant image regions by backpropagating the excitation signals from the target class to the input image. On the other hand, for Transformers, approaches like Layer-wise Relevance Propagation (LRP) have been developed to attribute the model's predictions to specific input tokens, shedding light on the importance of different parts of the input sequence.

Overall, attention capture techniques such as GradCAM and Attention Rollout play a crucial role in enhancing the interpretability and transparency of deep learning models, enabling researchers and practitioners to gain valuable insights into their decision-making processes. By visualizing the regions of high activation and attention within CNNs and Vision Transformers, these techniques facilitate a deeper understanding of model behavior and support informed decision-making in various application domains.

8 PROPOSED METHODOLOGY

8.1 SYNTHETIC DATA GENERATOR

In the realm of medical image analysis, the availability of comprehensive datasets plays a pivotal role in the development and evaluation of machine learning models. However, certain medical conditions present unique challenges that hinder the acquisition of sufficient real-world data for robust model training and validation. One such condition is Retinopathy of Prematurity (ROP), a sight-threatening disorder affecting premature infants' retinas. ROP diagnosis relies heavily on the identification of characteristic features within retinal images, such as tortuous blood vessels, which signify disease progression.

Despite the critical importance of ROP diagnosis and management, the scarcity of large-scale, annotated ROP image datasets presents a formidable obstacle for researchers and practitioners alike. This scarcity stems primarily from stringent patient confidentiality regulations and the ethical imperative to protect sensitive medical information. Consequently, the limited availability of real-world ROP images impedes progress in developing accurate and generalizable machine learning algorithms for ROP detection and classification.

Recognizing the challenges posed by the scarcity of real-world ROP image datasets, a synthetic data generator was developed as a pragmatic solution to overcome these limitations. By harnessing computational techniques and mathematical models, this generator aimed to emulate the intricate features of retinal images, including the characteristic patterns of tortuous blood vessels indicative of ROP. The synthetic data generator facilitates experimentation and innovation in algorithm development by offering a flexible platform to explore different clinical scenarios and disease manifestations.

8.1.1 CHARACTERISTICS OF RETINAL IMAGES

The synthetic data generator was designed with meticulous attention to replicating key features observed in real-world retinal images, essential for accurate ROP diagnosis and classification. These features encompassed various aspects of the intricate blood vessel network within the retina, serving as critical indicators of ROP progression and severity.

1. **Wide-Spreading Network-like Structure:** One of the primary objectives of the synthetic data generator was to emulate the expansive and interconnected network-like structure of retinal blood vessels observed in real-life images. By incorporating algorithms that simulate the branching and interconnecting nature of blood vessels, the generator aimed to generate synthetic images that capture the complex spatial arrangement of retinal vasculature.
2. **Branching Pattern with Diminishing Thickness:** Another crucial aspect targeted by the synthetic data generator was the branching pattern of retinal blood vessels, characterized by a hierarchical structure with branches of diminishing thickness as they extend outward from the optic disc. Through careful calibration of branching algorithms, the generator endeavored to reproduce this characteristic pattern, ensuring fidelity to the anatomical features of real retinal vasculature.

3. **Tortuous Regions Indicating ROP:** Given the diagnostic significance of tortuous blood vessels in identifying ROP, special emphasis was placed on simulating tortuosity within the synthetic data. By introducing randomness and perturbations to the trajectory of blood vessels, the generator aimed to generate synthetic images with regions exhibiting the characteristic tortuousity associated with ROP-affected vasculature.
4. **Emergence and Convergence of Blood Vessels:** Furthermore, the synthetic data generator sought to replicate the dynamic process of blood vessel development within the retina, including the emergence of vessels from a central point (the optic disc) and their gradual convergence towards the peripheral retina. By modeling the directional movement of blood vessels and their spatial distribution, the generator aimed to mimic the physiological process of retinal vascularization observed in real-life images.

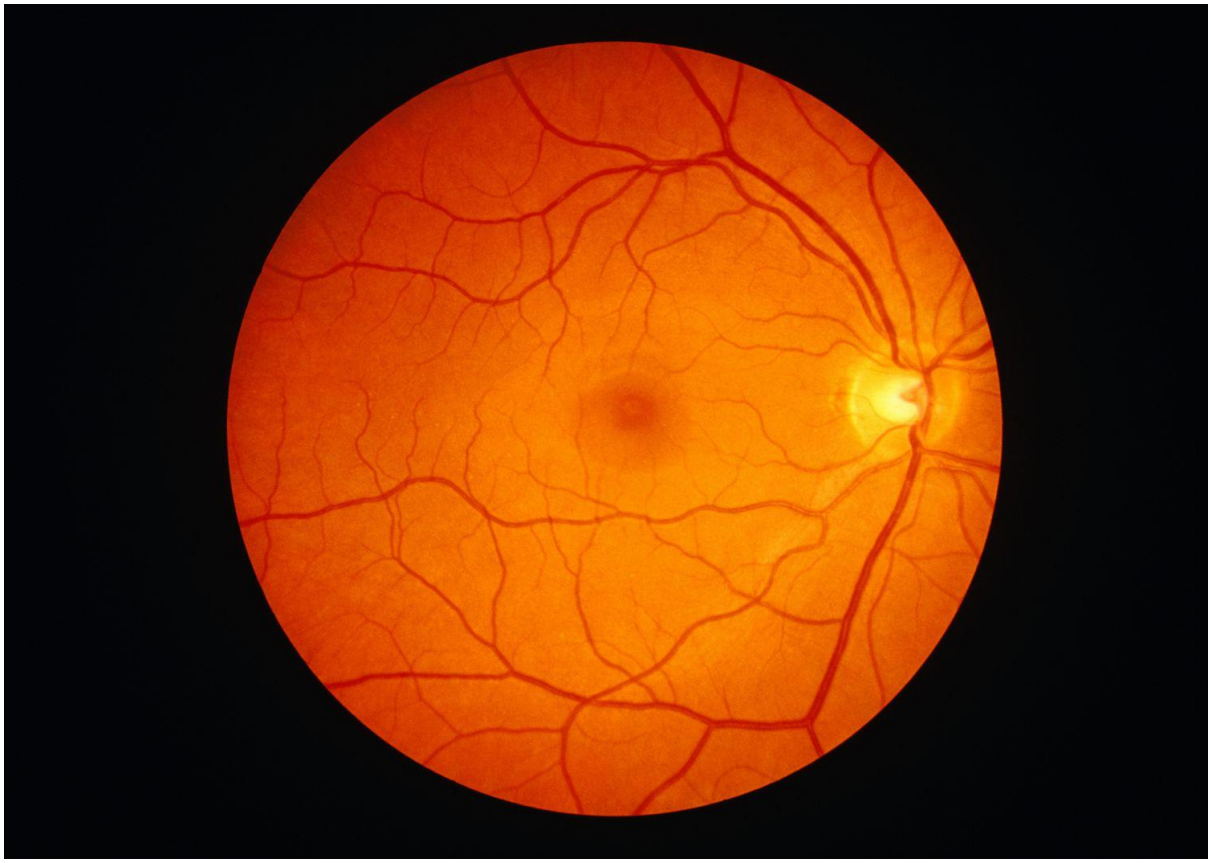


Figure 1: Retinal image of an healthy eye

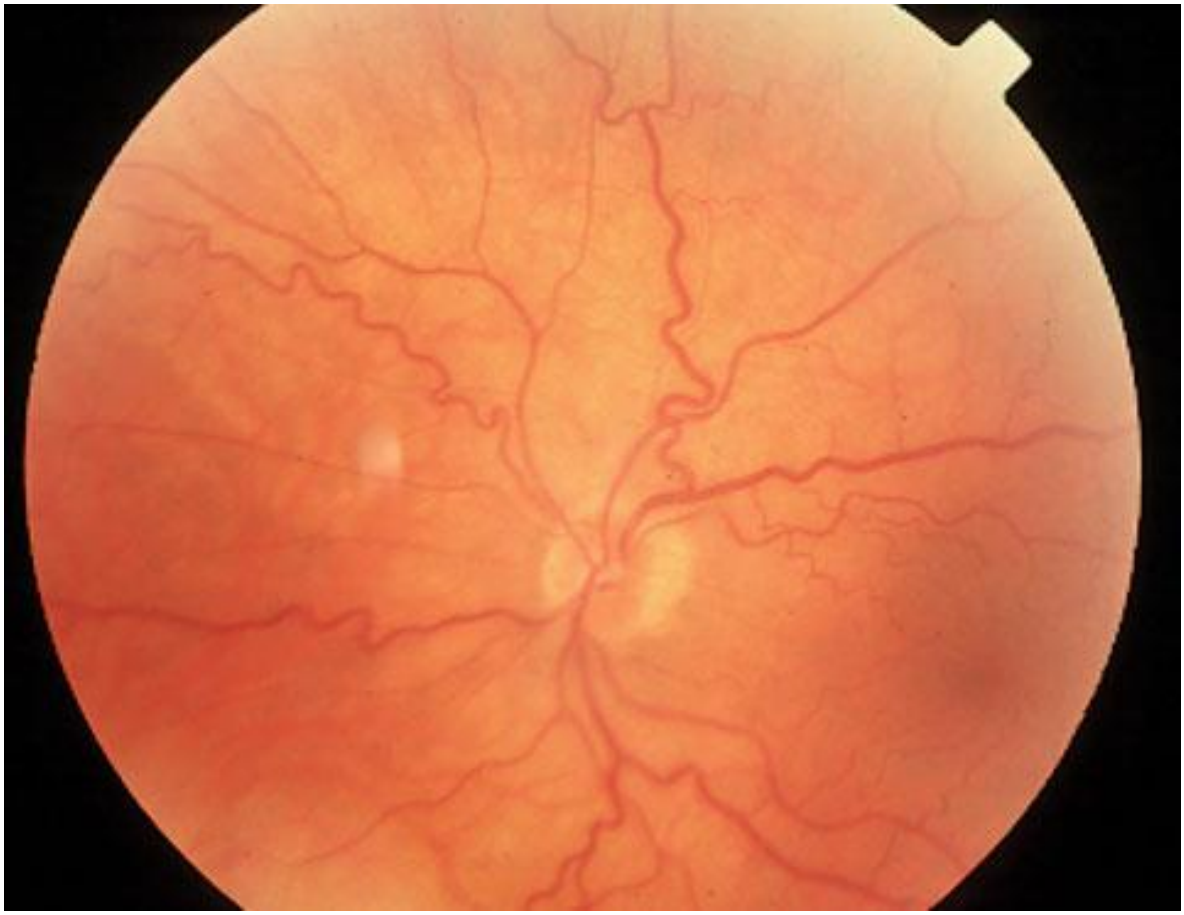


Figure 2: Retinal image of an eye with retinopathy indicated by the tortuous blood vessels

8.1.2 TECHNIQUES USED

To emulate the intricate network of retinal blood vessels observed in real-life images, a sophisticated approach utilizing random walkers was employed. Random walkers serve as virtual agents navigating through a simulated environment, determining their stride length and angle of deviation at each step based on predefined rules. The direction of movement of these walkers is determined by a weighted sum of two components: a vector field direction and a random deviation.

The vector field as shown in Figure 3 acts as a guiding force, steering the walkers along a circular path to mimic the natural trajectory of blood vessels observed in real retinal images. This directional guidance ensures that the synthetic blood vessels exhibit a coherent and anatomically plausible spatial arrangement, enhancing the realism of the generated images.

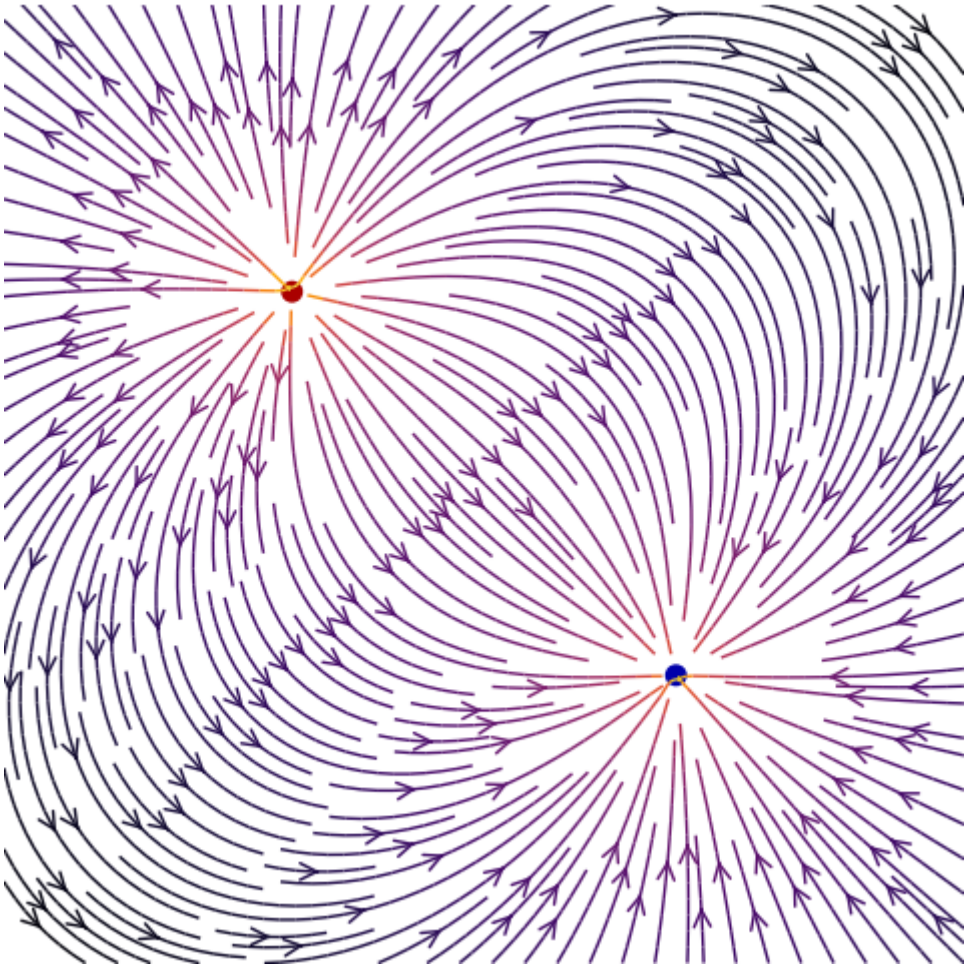


Figure 3: Vector field with a source and a sink

Moreover, the branching pattern of blood vessels was achieved through the concept of child walkers. At each step of the simulation, each walker has the potential to reproduce, controlled by a predefined probability. When a walker reproduces, its offspring become new walkers, effectively branching out from the parent vessel, and creating a new blood vessel segment. This branching mechanism mirrors the branching pattern observed in real retinal vasculature, contributing to the fidelity of the synthetic images.

Furthermore, to replicate the tapering effect observed in real blood vessels, the diameter of the vessels was dynamically adjusted as they propagated. This gradual decrease in thickness ensured that the synthetic blood vessels closely resembled their real-life counterparts, capturing the subtle nuances of vascular anatomy.

The synthetic images of the healthy eye look like Figure 4. As can be seen, it closely resembles blood vessel networks and can be used as a valid replacement in the absence of large amounts of real-life images.

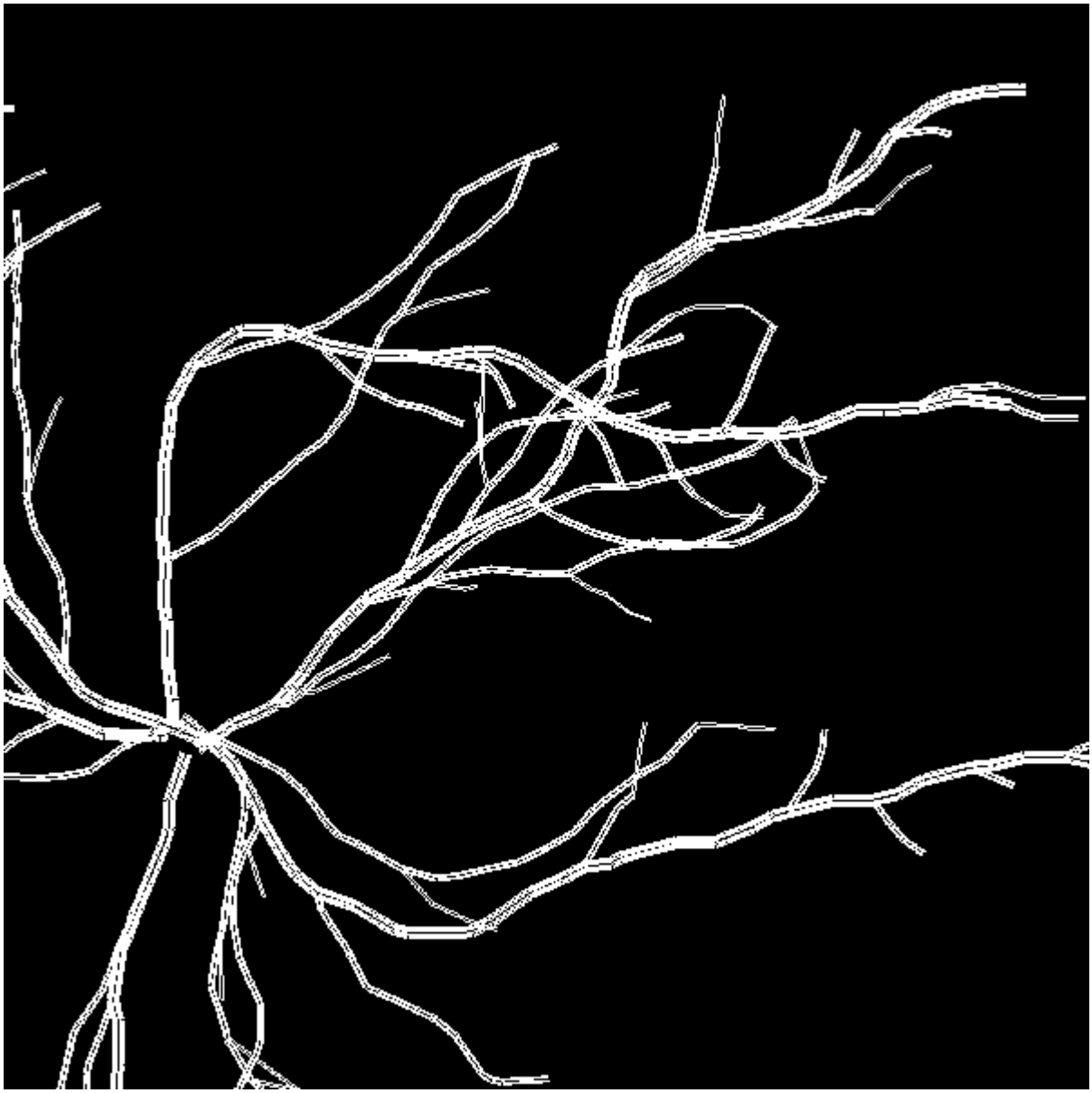


Figure 4: Synthetic Retinal Image – Healthy

To mimic the distinctive feature of Retinopathy of Prematurity (ROP) images – the winding and twisting of blood vessels known as tortuosity – we employed a special technique. We introduced random damped sine movements as shown in Figure 5 to our ROP walkers. Think of it like a wavy path that the walkers follow. These movements simulate the tortuous nature of blood vessels seen in ROP-affected retinas.

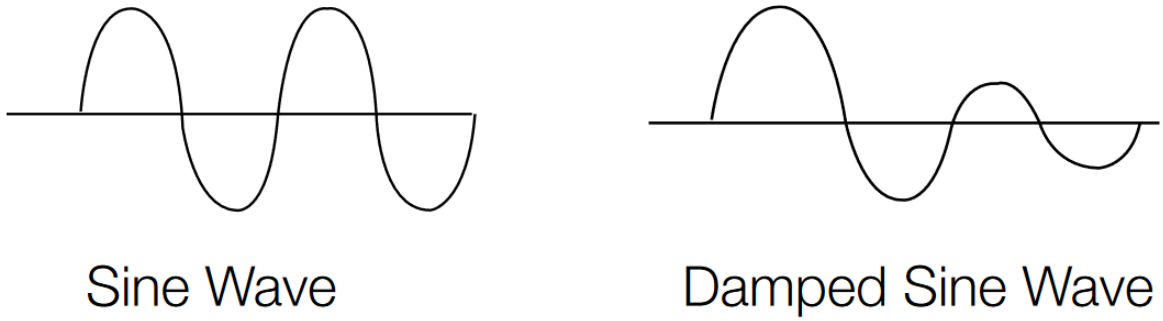


Figure 5: Sine and Damped sine wave structures

By adding these wavy motions, our synthetic blood vessels take on the characteristic irregularities seen in real ROP images. This helps make our synthetic data more realistic and useful for training computer algorithms to identify ROP accurately. In essence, this technique allows us to recreate the unique visual patterns associated with ROP, enhancing the authenticity and clinical relevance of our synthetic images. The synthetic images of the affected eye look similar to Figure 6. The tortuous regions are marked by a red bounding box.

8.1.3 IMPLEMENTATION DETAILS

The implementation of the synthetic data generator is developed in Python, utilizing libraries such as NumPy and Pillow. It efficiently generates synthetic images representative of Retinopathy of Prematurity (ROP) with a processing time of approximately 2-3 seconds per image. The generator offers extensive configurability through exposed parameters, enabling users to customize various aspects of the image generation process according to their specific requirements. This flexibility facilitates experimentation and adaptation to diverse research needs, contributing to the effectiveness of deep learning models in ROP diagnosis.

The pseudo code of the Walker is:

```
def move(self):
    for child in children:
        child.move()

    if not dead:
        if tortuous:
            make_tortuous_move()
        else:
            make_normal_move()
        try_reproduce()
        try_die_old_age()
```

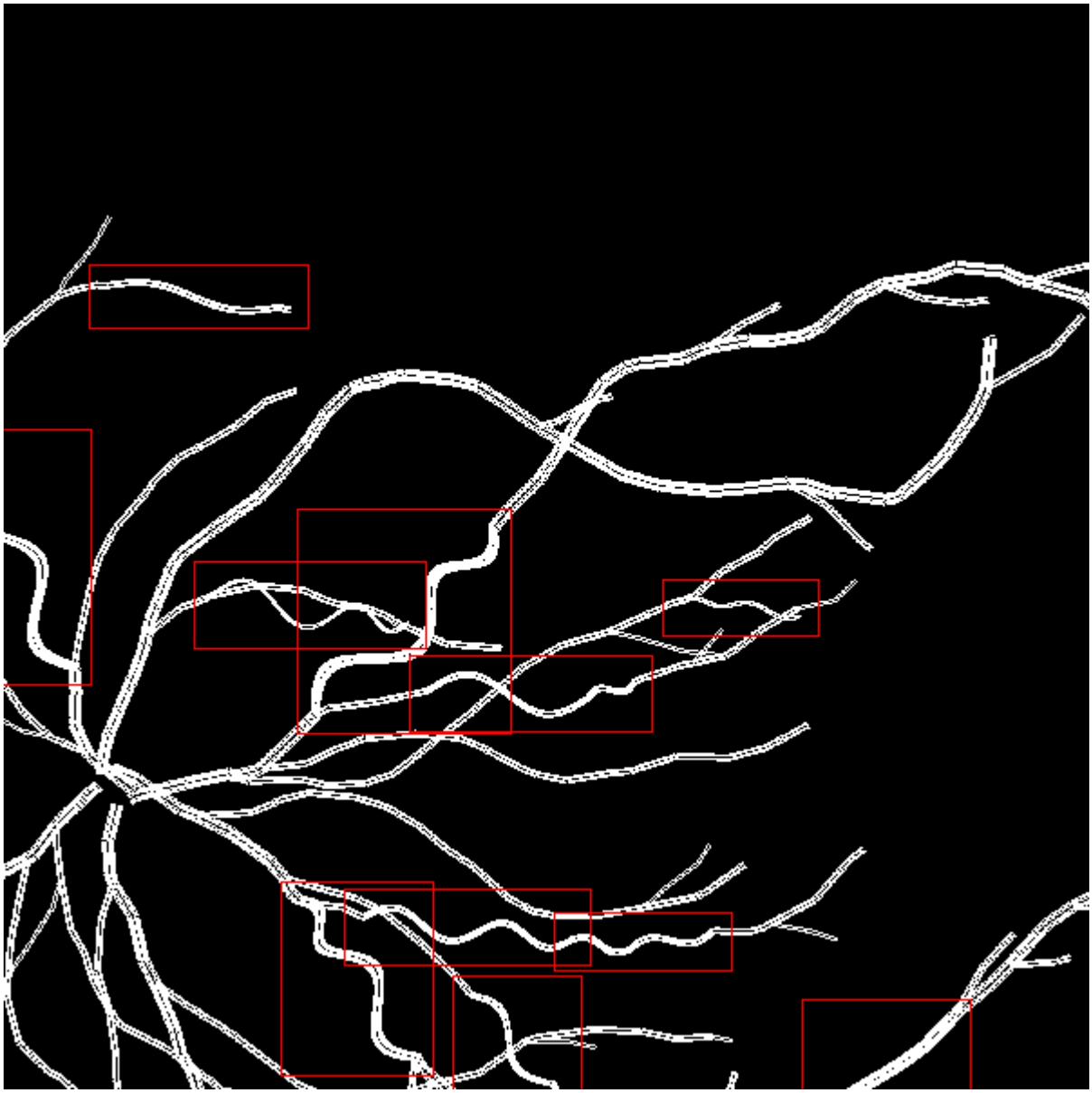


Figure 6: Synthetic Retinal Image - ROP

8.2 CLASSIFIERS

In this study, we employed two different types of deep learning architectures for the task of automatically diagnosing retinopathy of prematurity (ROP) from retinal image data: Convolutional Neural Networks (CNNs) and Vision Transformers. These architectures have demonstrated remarkable success in various computer vision tasks, including medical image analysis, and offer distinct advantages and capabilities for addressing the challenges posed by ROP diagnosis.

8.2.1 CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNNs) are a class of deep learning models specifically designed for processing grid-like data, such as images. They have been the predominant architecture for computer vision tasks, including image classification, object detection, and

segmentation. CNNs leverage a hierarchical structure of convolutional layers, which learn to extract increasingly complex and abstract features from the input data, enabling the model to capture local patterns and spatial relationships within the images effectively.

One of the key strengths of CNNs lies in their ability to automatically learn relevant features from raw image data, eliminating the need for manual feature engineering. This property makes them highly effective in dealing with complex and high-dimensional image data, where handcrafted features may fail to capture all relevant information. Additionally, CNNs are computationally efficient and suitable for real-time applications, making them attractive for deployment in clinical settings.

The Convolutional Neural Network (CNN) architecture employed for the retinopathy of prematurity (ROP) diagnosis task consists of an input layer designed to accept images of size 224 x 224 with three color channels (RGB). The subsequent layers include four convolutional layers, each followed by a max-pooling layer for spatial downsampling. The first convolutional layer has 16 filters with a 3x3 kernel size, utilizing the leaky ReLU activation function and 'same' padding to capture low-level features. The number of filters progressively increases to 32, 64, and 128 in the subsequent convolutional layers, allowing the network to learn more complex and abstract representations. After each convolutional layer, a 2x2 max-pooling operation is applied to reduce the spatial dimensions while retaining the most salient features. The output of the final convolutional layer is flattened and passed through a dropout layer with a rate of 0.5 to prevent overfitting. Finally, a fully connected dense layer with 2 units and a softmax activation function is used to produce the classification output, representing the probability distribution over the classes for ROP diagnosis.

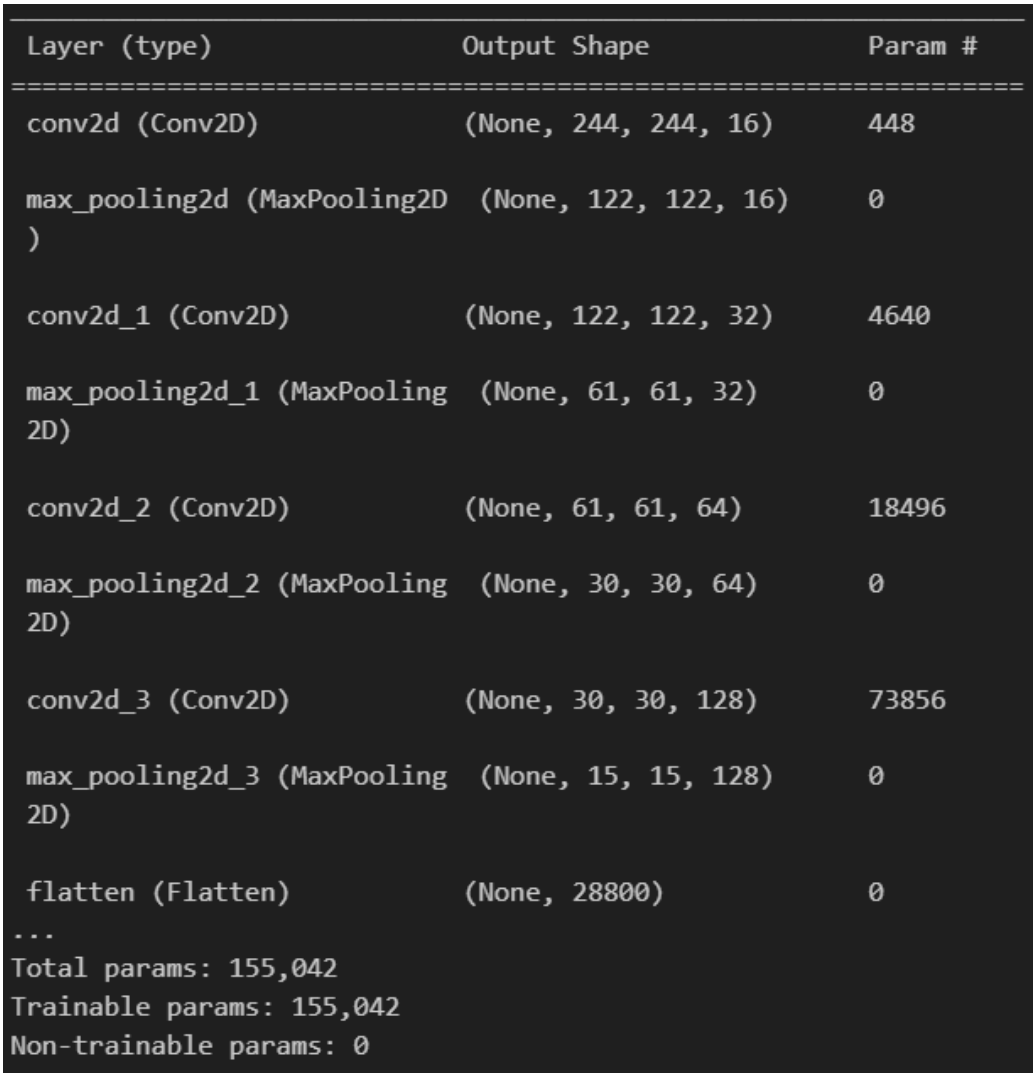


Figure 7: CNN model architecture

8.2.2 VISION TRANSFORMERS

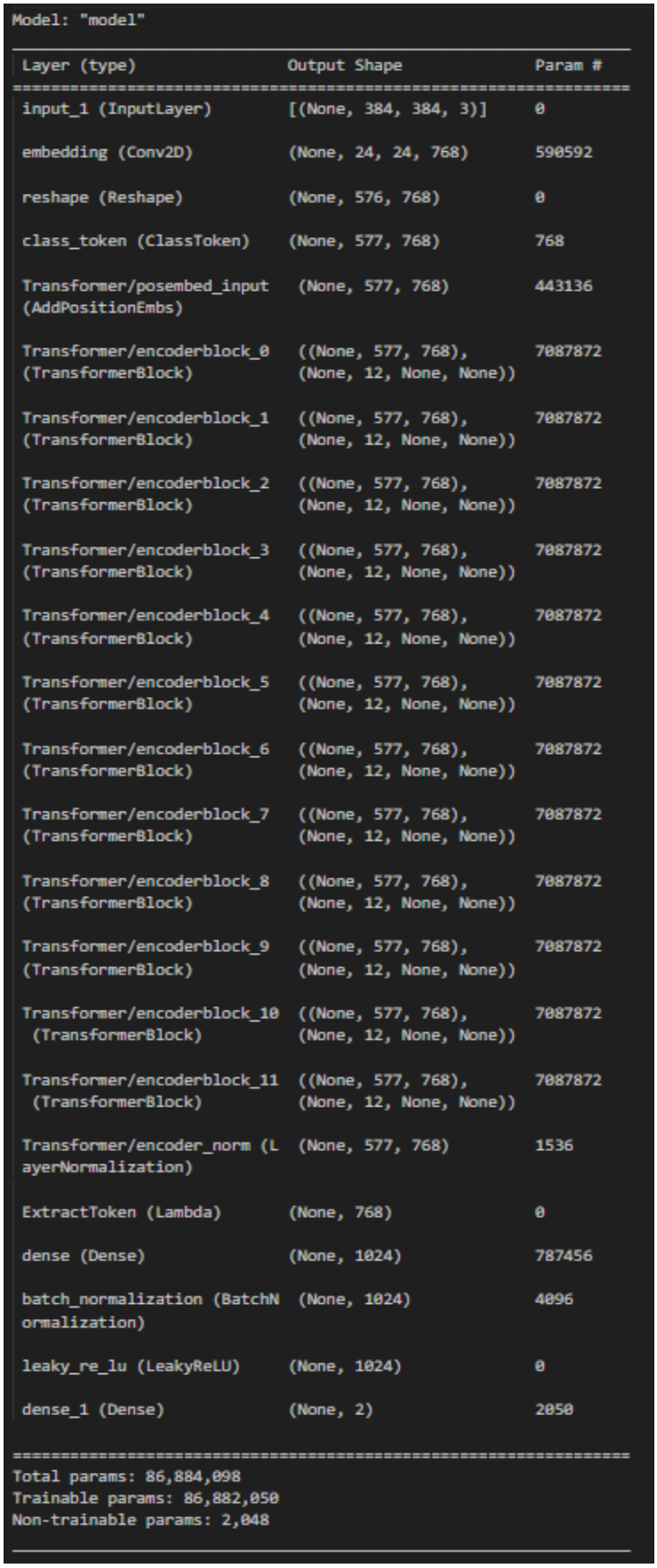
Vision Transformers are a relatively new class of deep learning models that have gained significant attention for their impressive performance in various computer vision tasks. Inspired by the success of Transformers in natural language processing, Vision Transformers adapt the self-attention mechanism to handle image data effectively. Unlike CNNs, which rely on local convolutions and pooling operations, Vision Transformers leverage global self-attention mechanisms to capture long-range dependencies and complex relationships within the input images.

The ability of Vision Transformers to model global context and long-range dependencies makes them particularly well-suited for tasks that require understanding intricate patterns and interactions within an image, such as object detection and segmentation. However, they tend to have higher computational requirements and may struggle with fine-grained local patterns compared to CNNs.

This model architecture integrates a Vision Transformer (ViT) backbone followed by additional classification layers, designed for effective image classification tasks. The

backbone_model variable encapsulates the ViT-B16 model, a variant of ViT equipped with 16 transformer layers. ViT-B16 is initialized with various parameters including the image size, activation function, and the choice of pre-trained weights. Notably, pre-trained weights are utilized for feature extraction while excluding the pre-trained weights for the classification head, indicating a customized approach for downstream tasks. Following the ViT backbone, the architecture incorporates supplementary layers for further feature refinement and classification. Specifically, a Dense layer with 1024 units and L2 regularization is introduced, providing a higher-level representation of the extracted features. Batch normalization and Leaky ReLU activation are subsequently applied to enhance the model's stability and non-linearity. Finally, a Dense layer with softmax activation serves as the classification head, producing probability distributions over the classes of interest.

The Vision Transformer (ViT) backbone, exemplified by ViT-B16, constitutes a pivotal component of this architecture, renowned for its prowess in capturing global contextual information from input images. ViT-B16 encompasses a sequence of transformer encoder layers, each comprising a Multi-Head Self-Attention Mechanism and a Feedforward Neural Network. This configuration enables the model to attend to various parts of the input image, facilitating the extraction of hierarchical features crucial for image classification tasks. Noteworthy is the patch-based approach employed by ViT-B16, wherein the input image is divided into fixed-size patches, subsequently linearly embedded to form token embeddings. These token embeddings are supplemented with positional encodings to provide spatial information, facilitating effective representation learning. By iteratively processing token embeddings through multiple transformer encoder layers, ViT-B16 adeptly captures intricate patterns and relationships within the input images. While traditionally accompanied by a classification head for image classification tasks, in this architecture, ViT-B16 serves primarily as a feature extractor, laying the groundwork for subsequent classification layers tailored to the specific task at hand.



```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 384, 384, 3)]     0

embedding (Conv2D)           (None, 24, 24, 768)       590592

reshape (Reshape)            (None, 576, 768)          0

class_token (ClassToken)     (None, 577, 768)          768

Transformer/posembed_input   (None, 577, 768)          443136
(AddPositionEmbs)

Transformer/encoderblock_0   ((None, 577, 768),        7087872
(TransformerBlock)           (None, 12, None, None))

Transformer/encoderblock_1   ((None, 577, 768),        7087872
(TransformerBlock)           (None, 12, None, None))

Transformer/encoderblock_2   ((None, 577, 768),        7087872
(TransformerBlock)           (None, 12, None, None))

Transformer/encoderblock_3   ((None, 577, 768),        7087872
(TransformerBlock)           (None, 12, None, None))

Transformer/encoderblock_4   ((None, 577, 768),        7087872
(TransformerBlock)           (None, 12, None, None))

Transformer/encoderblock_5   ((None, 577, 768),        7087872
(TransformerBlock)           (None, 12, None, None))

Transformer/encoderblock_6   ((None, 577, 768),        7087872
(TransformerBlock)           (None, 12, None, None))

Transformer/encoderblock_7   ((None, 577, 768),        7087872
(TransformerBlock)           (None, 12, None, None))

Transformer/encoderblock_8   ((None, 577, 768),        7087872
(TransformerBlock)           (None, 12, None, None))

Transformer/encoderblock_9   ((None, 577, 768),        7087872
(TransformerBlock)           (None, 12, None, None))

Transformer/encoderblock_10  ((None, 577, 768),        7087872
 (TransformerBlock)          (None, 12, None, None))

Transformer/encoderblock_11  ((None, 577, 768),        7087872
 (TransformerBlock)          (None, 12, None, None))

Transformer/encoder_norm (L  (None, 577, 768)          1536
ayerNormalization)

ExtractToken (Lambda)        (None, 768)               0

dense (Dense)                (None, 1024)              787456

batch_normalization (BatchN  (None, 1024)              4096
ormalization)

leaky_re_lu (LeakyReLU)      (None, 1024)              0

dense_1 (Dense)              (None, 2)                 2050

=================================================================
Total params: 86,884,098
Trainable params: 86,882,050
Non-trainable params: 2,048
_________________________________________________________________
```

Figure 8: Vision Transformer architecture

8.3 ATTENTION CAPTURE TECHNIQUES

Attention mechanisms have revolutionized the field of deep learning by enabling models to focus on relevant parts of input data while performing various tasks such as image classification and natural language processing. These mechanisms mimic human attention, allowing models to assign different weights to different parts of the input, thus improving their interpretability and performance. Two prominent attention capture techniques widely employed in modern deep learning architectures are GradCAM for Convolutional Neural Networks (CNNs) and Attention Rollout for Transformers.

8.3.1 GRADCAM

GradCAM (Gradient-weighted Class Activation Mapping) is a groundbreaking technique designed to elucidate the decision-making process of Convolutional Neural Networks (CNNs) by visualizing the regions of an input image that are most influential in predicting a specific class. At its core, GradCAM leverages the gradients flowing into the last convolutional layer of the CNN to understand which parts of the input image contribute most significantly to the final prediction. By computing the gradients of the target class score with respect to the feature maps of the last convolutional layer, GradCAM effectively identifies the importance of each feature map in making the prediction. These gradients are then globally averaged to obtain the importance score for each feature map, reflecting its contribution to the final decision.

Once the importance scores are computed, GradCAM generates a heatmap by weighting the feature maps with their corresponding importance scores and summing them. This heatmap highlights the regions of the input image that are most relevant to the CNN's decision-making process, providing valuable insights into the model's attention mechanism. By overlaying the heatmap on the input image, researchers and practitioners can visualize which parts of the image the model focuses on when making predictions. This visualization not only aids in understanding the model's reasoning but also helps in identifying potential biases, weaknesses, or areas for improvement.

GradCAM has found widespread applications across various domains, including medical imaging, where it is used to interpret the predictions of CNN-based models trained for tasks such as disease diagnosis and prognosis. By visualizing the regions of interest identified by the model, medical professionals can gain valuable insights into the diagnostic process, validate the model's decisions, and ultimately enhance patient care. Moreover, GradCAM facilitates model debugging, validation, and refinement by providing interpretable and actionable insights into the CNN's decision-making process.

Figure 9: Sample image with a GradCAM heatmap

8.3.2 ATTENTION ROLLOUT

Attention Rollout is a cutting-edge attention visualization technique tailored specifically for Transformers, a class of models renowned for their ability to capture long-range dependencies in sequential data. Unlike CNNs, which rely on convolutional layers for feature extraction, Transformers employ self-attention mechanisms to weigh the importance of different tokens in the input sequence. Attention Rollout leverages this self-attention mechanism to visualize how attention flows across different tokens in the input sequence, providing valuable insights into the model's inner workings and decision-making process.

The methodology of Attention Rollout involves iteratively masking out tokens in the input sequence and observing the resulting changes in the model's predictions. By systematically masking each token and analyzing its impact on the model's output, Attention Rollout constructs a comprehensive picture of the attention distribution within the Transformer model. This technique not only reveals which tokens the model attends to but also sheds light on the dynamics of attention propagation and allocation throughout the sequence.

Attention Rollout has emerged as a powerful tool for understanding and interpreting the decisions made by Transformer models across various tasks and domains. In natural language processing tasks such as machine translation and sentiment analysis, Attention Rollout provides valuable insights into how the model processes and attends to different parts of the input sequence, facilitating error analysis, model debugging, and performance optimization. Moreover, Attention Rollout enables researchers and practitioners to identify potential biases, limitations, or areas for improvement in Transformer-based models, ultimately enhancing their interpretability, reliability, and effectiveness in real-world applications.

Figure 10: Sample image with a Attention Rollout heatmap

9 RESULTS AND DISCUSSION

9.1 CLASSIFICATION

9.1.1 DATASET

A synthetic dataset consisting of 1000 images of each class: healthy, ROP. The images were generated using the synthetic data generated described in the previous section. All the images are resized to 224x224 before training the model.

9.1.2 CONVOLUTIONAL NEURAL NETWORK

The CNN model underwent extensive training over a span of 40 epochs, employing the Adam optimizer with a learning rate set to 1e-3. To enhance training stability and convergence, a dynamic learning rate adjustment strategy, known as "reduce on plateau," was incorporated into the training regimen. This strategy facilitated adaptive learning rate changes based on the model's validation performance, ensuring efficient optimization across the synthetic dataset. Throughout the training process, categorical crossentropy served as the primary optimization metric, guiding the model towards minimizing the disparity between predicted and ground-truth class labels.

The model was able to achieve 95% accuracy on the test data consisting of about 400 images in total.

GradCAMs were generated by the model for different images.

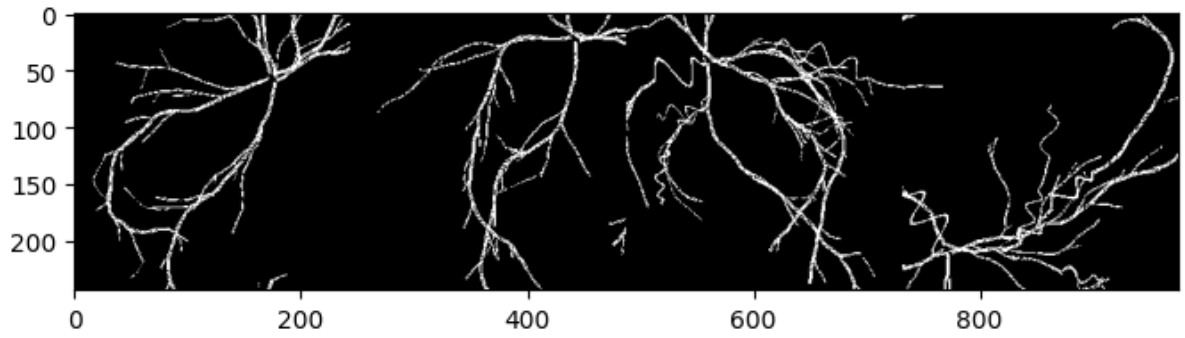


Figure 11: Sample images.1,2 – healthy, 3,4 - ROP

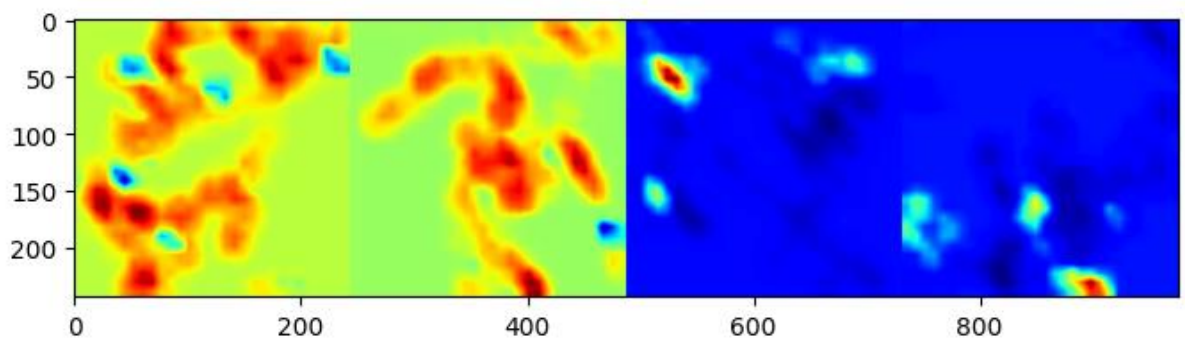


Figure 12: GradCAMs of the sample images. 1,2 – healthy, 3,4 - ROP

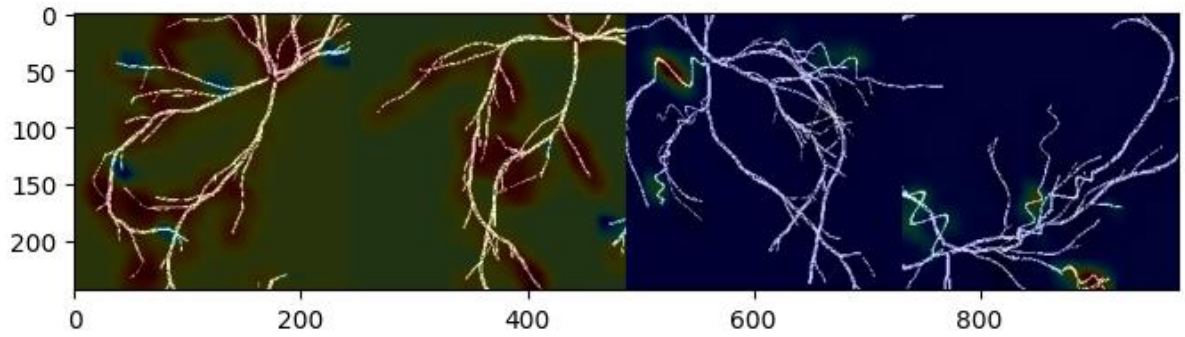


Figure 13: GradCAMs overlaid on the sample images. 1,2 – healthy, 3,4 - ROP

Interpreting a heatmap generated by GradCAM involves understanding the relative attention levels assigned by the model to different regions of the input image. In a GradCAM heatmap, regions depicted in shades of red indicate areas of high attention, where the model focuses most intensely during the prediction process. Conversely, regions shaded in blue represent areas of low attention, receiving comparatively less emphasis from the model. By analyzing the intensity and distribution of colors within the heatmap, one can discern which parts of the input image are deemed most crucial by the model for making its predictions. Higher intensity red regions signify strong model attention, suggesting significant relevance to the predicted class, while cooler blue regions denote lower relevance or negligible contribution to the prediction. Understanding these attention patterns facilitates the interpretation of model decisions and provides valuable insights into the features and characteristics influencing the classification outcome.

In the sample images show in Figure 11, the first two images belong to healthy class and the last two belong to ROP class. The GradCAMs overlaid on the images, as shown in Figure 13, show increased attention in the tortuous regions in the ROP class images. In the healthy images, the attention is more dispersed (displayed be the overall yellowish shade of the heatmap) and the high attention regions are usually branches and tips of the blood vessels.

To isolate the high attention regions in the ROP images, a threshold-based masking was applied. It is then compared to the tortuous regions in the ground truth masks. Figure 14 shows a sample image belonging to the ROP class and the associated ground truth mask of the tortuous regions. The mask is denoted by a red colour. The mask also includes a region of 10 pixels around the tortuous blood vessels since attention mechanisms like GradCAMs provide region-based attention scores.

Figure 15 shows the overlap between the tortuous regions in the ground truth masks and the high attention regions in the GradCAM. These images are for a threshold value of 95^{th} percentile i.e., the high attention region comprises of pixels with the top 5% attention scores in the attention map generated from the model. The image also shows the precision and recall values of each image. Precision denotes the ratio between the number of pixels that is correctly identified as tortuous to the total number of pixels marked as tortuous by the model's attention mechanism. Recall denotes the ratio between the number of pixels that is correctly identified

as tortuous by the model's attention mechanism to the total number of pixels in the ground truth mask.

The precision and recall for 50 test images is calculated. The average precision is 81.5% and average recall is 31.87% for the threshold of 95th percentile score.

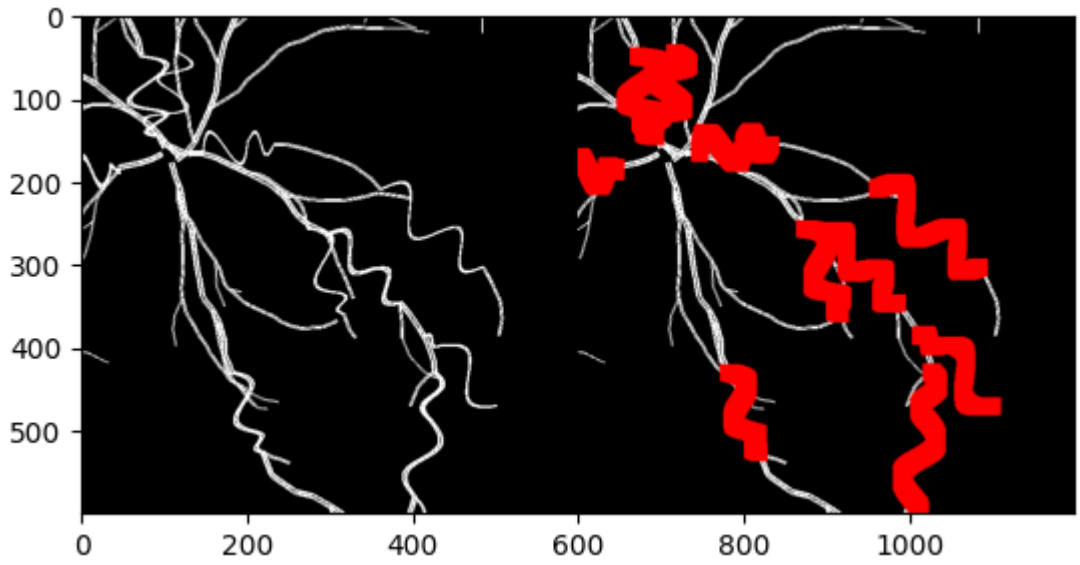


Figure 14: Sample ROP image and ground truth mask of tortuous regions

Similarly, this experiment is run for varying thresholds and the results are shown in Table 1. It shows an inverse relation between the precision and recall. When the threshold is increased, the model is able to focus on the tortuous regions but at the same time, it only focusses on few of the tortuous regions and not all.

Threshold Percentile	Precision	Recall
50	57.34	65.57
75	63.05	61.23
80	64.44	59.78
90	72.81	47.66
95	81.5	31.87

Table 1: Mean precision and recall scores for different threshold values for the CNN architecture

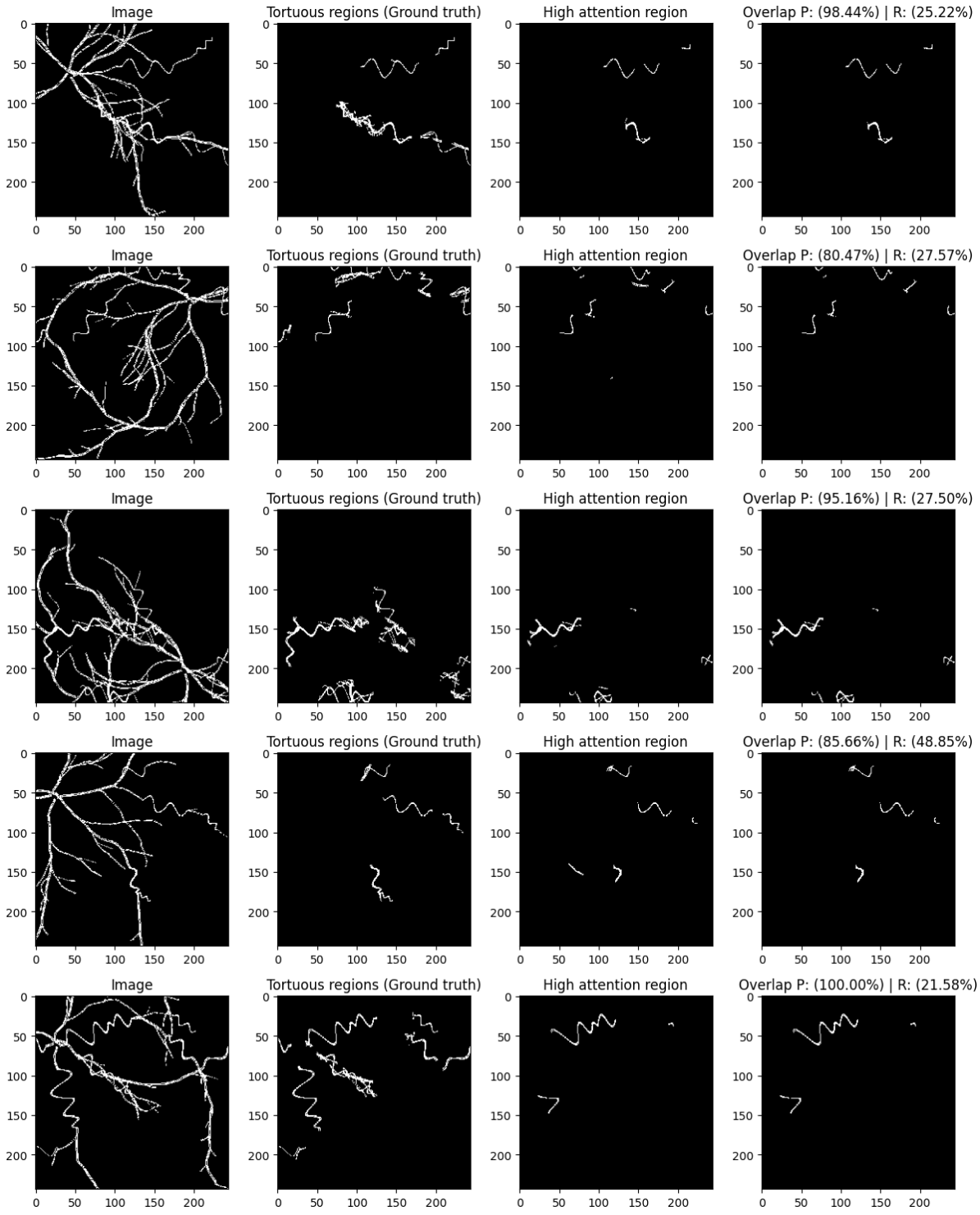


Figure 15: The overlap between ground truth masks and the high attention regions in the GradCAM for a threshold of 95th percentile for sample ROP images

9.1.3 VISION TRANSFORMER

The Vision Transformer model underwent training over a span of 10 epochs, utilizing the Adam optimizer with a fixed learning rate of 1e-3. To ensure stability and convergence, a "reduce on plateau" dynamic learning rate adjustment strategy was incorporated. This adaptive mechanism facilitated adjustments based on validation performance, optimizing efficiency across the

synthetic dataset. Throughout training, categorical crossentropy served as the primary optimization metric, minimizing the disparity between predicted and ground-truth class labels.

The vision transformer model also was a perfect classifier with 97.5% accuracy.

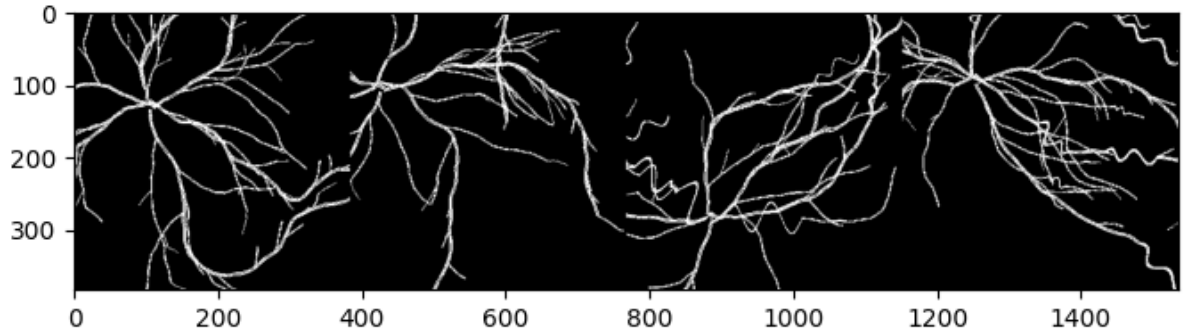


Figure 16: Sample images

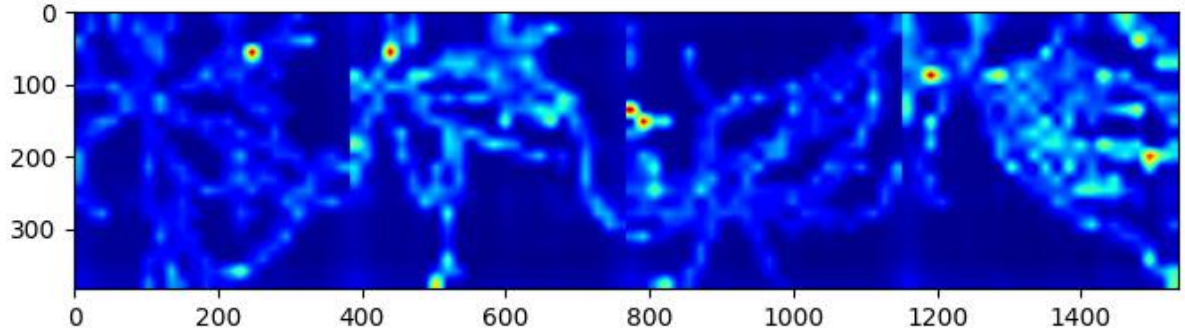


Figure 17: Attention Rollout maps

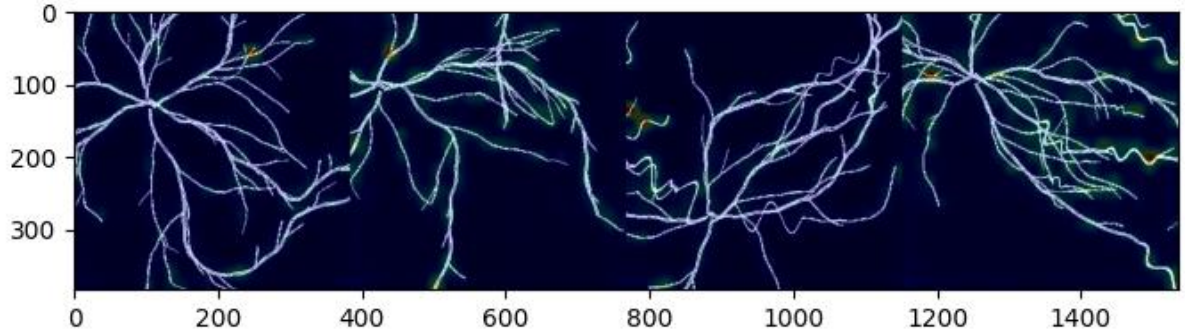


Figure 18: Attention Rollout maps overlaid on sample images

The sample images and Attention Rollout output is shown in Figure 16, Figure 17 and Figure 18. They also show localization of attention in the areas with tortuous blood vessels. The major observation is that the regions shown in attention rollout maps are tighter than the ones on GradCAMs allowing much better localization.

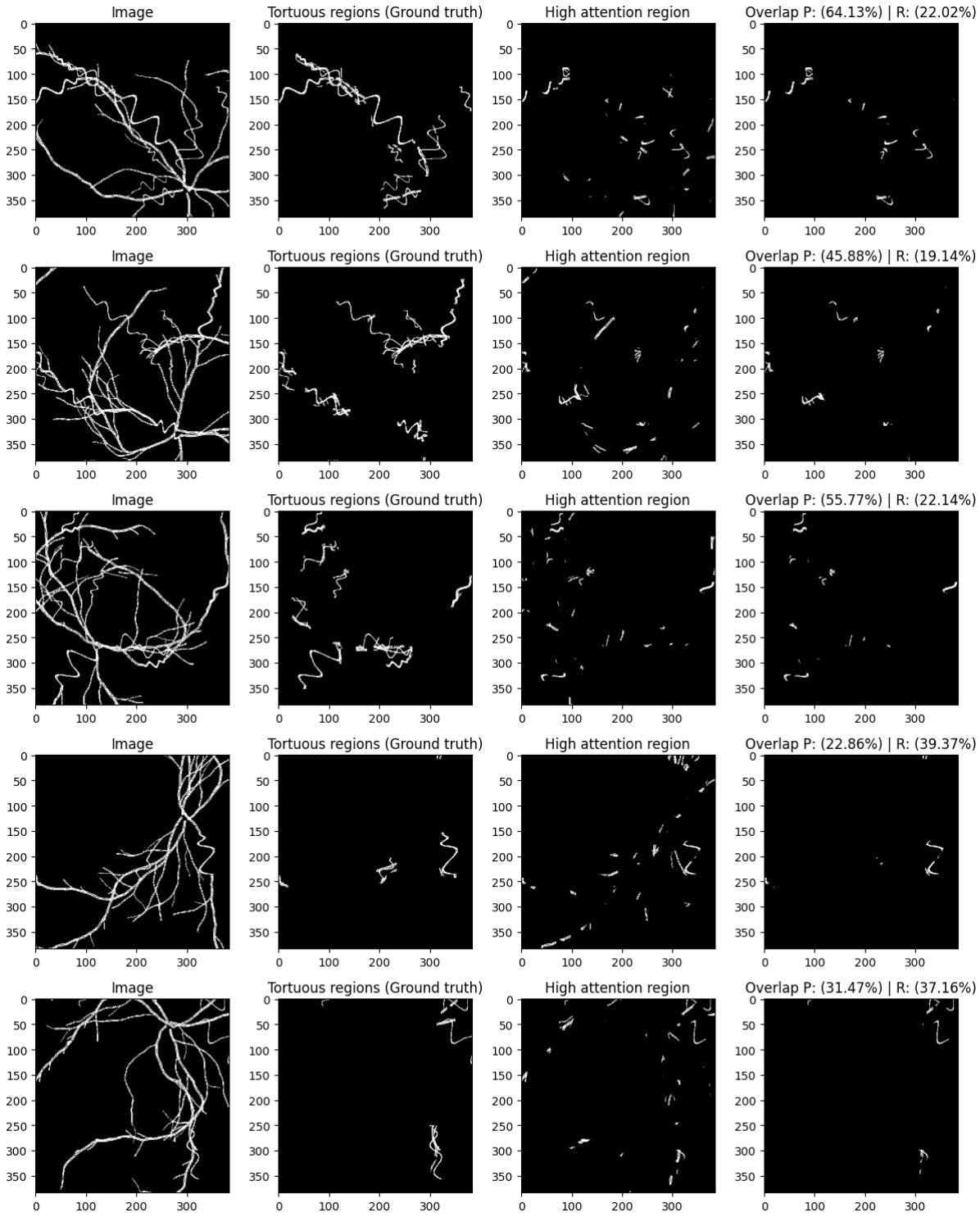


Figure 19: The overlap between ground truth masks and the high attention regions in the Attention Rollout map for a threshold of 95th percentile for sample ROP images

Figure 19 shows the overlap between the attention rollout map and the ground truth for the vision transformer architecture. It can be seen that meaningful features are harder to derive in the vision transformer architecture due to its very high localization. The precision and recall scores for different thresholds are shown in Table 2. It can be seen that the recall scores are much better than what is seen in CNN architecture. But it comes at the cost of very low precision values.

Threshold Percentile	Precision	Recall
50	18.29	99.63
75	21.23	86.06
80	22.78	77.25
90	28.92	53.24
95	34.79	31.18

Table 2: Mean precision and recall scores for different threshold values for the ViT architecture

10 FUTURE WORK

This project has explored the effectiveness of attention capture mechanisms in providing meaningful features to improve the classification of Retinopathy of Prematurity (ROP) images. Building on the insights gained, several directions for future work have been identified to further enhance the model's capabilities and reliability.

1. **Explore other methods of attention capture for these architectures:** Future research should investigate alternative methods of attention capture that can provide better localization within CNN and ViT architectures. Potential approaches include GradCAM++, Saliency maps for CNN architecture. These methods may enhance the model's ability to accurately identify and focus on the critical regions in retinal images, leading to improved diagnostic performance and reliability.
2. **Explore techniques to improve the attention maps and filter out the noise**: Enhancing the clarity and precision of attention maps is crucial for reliable model interpretation. Future work should focus on developing techniques to reduce noise and enhance the signal within attention maps. This could involve advanced filtering methods, noise-reduction algorithms, and the integration of contextual information to refine the attention focus. Improved attention maps will provide clearer insights into the model's decision-making process, facilitating better understanding and trust among healthcare professionals.
3. **Construct a quantifiable metric of tortuosity derived from the features obtained through attention capture:** An important area for future work is the development of a quantifiable metric of tortuosity based on the features identified through attention capture. This involves designing algorithms to measure the curvature and complexity of blood vessels in retinal images, leveraging the focused features highlighted by the attention mechanisms. A robust metric of tortuosity could serve as a valuable diagnostic indicator, aiding in the assessment and monitoring of ROP progression and severity.
4. **Expand and diversify the synthetic dataset generator:** While the initial development of a synthetic dataset generator has provided a valuable tool for augmenting training data, future efforts should aim to expand its capabilities. This includes generating a wider variety of retinal image scenarios and conditions, integrating advanced techniques such as generative adversarial networks (GANs) for more realistic image synthesis, and validating the effectiveness of these synthetic datasets in improving model performance across diverse clinical settings.

By addressing these future directions, the project can significantly advance the state of the art in ROP detection and diagnosis, ultimately contributing to better healthcare outcomes for preterm infants.

11 CONCLUSION

In conclusion, this study presents a comprehensive approach to addressing the challenges posed by the scarcity of real-world retinopathy of prematurity (ROP) image data. By developing a sophisticated synthetic data generator capable of emulating the intricate features of retinal vasculature, including the characteristic tortuosity associated with ROP, we have overcome a significant obstacle in the field of medical image analysis. Moreover, our comparative study of attention capture techniques—specifically CNN with GradCAM and ViT with Attention Rollout mechanisms—has provided valuable insights. The CNN architecture exhibited superior performance in capturing and localizing tortuous regions in ROP images, offering enhanced interpretability crucial for clinical decision-making. Conversely, attention maps generated by the ViT architecture demonstrated higher levels of noise, highlighting areas for potential refinement in future research. Overall, these findings underscore the efficacy of attention mechanisms in medical image analysis and pave the way for further advancements in ROP diagnosis and treatment planning.

12 REFERENCES

[1] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int J Comput Vis*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94/METRICS.

[2] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, vol. I, pp. 886–893, 2005, doi: 10.1109/CVPR.2005.177.

[3] C. Cortes, V. Vapnik, and L. Saitta, “Support-vector networks,” *Machine Learning 1995 20:3*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[4] T. K. Ho, “Random decision forests,” *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, vol. 1, pp. 278–282, 1995, doi: 10.1109/ICDAR.1995.598994.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998, doi: 10.1109/5.726791.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Adv Neural Inf Process Syst*, vol. 25, 2012, Accessed: Apr. 30, 2024. [Online]. Available: http://code.google.com/p/cuda-convnet/

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” pp. 248–255, Mar. 2010, doi: 10.1109/CVPR.2009.5206848.

[8] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *ICLR 2021 - 9th International Conference on Learning Representations*, Oct. 2020, Accessed: Apr. 30, 2024. [Online]. Available: https://arxiv.org/abs/2010.11929v2

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” *Int J Comput Vis*, vol. 128, no. 2, pp. 336–359, Oct. 2016, doi: 10.1007/s11263-019-01228-7.

[10] S. Abnar and W. Zuidema, “Quantifying Attention Flow in Transformers,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, May 2020, doi: 10.18653/v1/2020.acl-main.385.