Automated Detection of Retinal Diseases in OCT Scans and Improving Diagnosis of AMD, DME, and CNV

Shaunak Warty
Georgia Institute of Technology
Atlanta, GA 30332
shaunak.warty@gatech.edu

Michael Chian
Georgia Institute of Technology
Atlanta, GA 30332
mchian6@gatech.edu

Felicia Jamba
Georgia Institute of Technology
Atlanta, GA 30332
fjamba@gatech.edu

Samuel Sukendro
Georgia Institute of Technology
Atlanta, GA 30332
samsukendro@gatech.edu

Abstract

Optical Coherence Tomography (OCT) scans allow medical professionals to spot indicators of disease such as choroidal neovascularization (CNV), Diabetic Macular Edema (DME), and drusen. Manually interpreting OCT scans involves analyzing small irregularities such as tiny deposits between the retina and eye membrane. These signs can go unnoticed or be misinterpreted due to human error. Expanding on current research into creating deep learning models on OCT images, we have created a transfer learning model from ResNet-50, a model with U-Net architecture, and an Attention-based CNN to tackle this issue. There are varying results; the U-Net architecture and transfer learning models performed strongly and presented themselves as viable medical options due to their low false negative rates, while the Attention-based CNN struggled with achieving a high accuracy.

1. Introduction

Approximately 12.6% of Americans aged 40 and older suffered from age-related macular degeneration (AMD). These 19.8 million people struggle with blurred and distorted central vision [7]. Early detection of AMD and other retinal diseases is critical to preventing vision loss. Optical Coherence Tomography (OCT) scans allow medical professionals to spot indicators of disease such as choroidal neovascularization (CNV), Diabetic Macular Edema (DME), and drusen. Unfortunately, with millions of retinal scans to analyze, diagnosis of these retinal conditions require time and specialized knowledge. Manually interpreting OCT scans involves analyzing small irregularities such as tiny deposits between the retina and eye membrane [9]. These signs can go unnoticed or be misinterpreted due to human error.

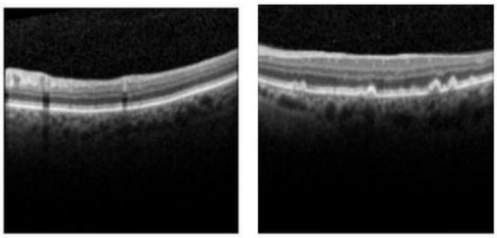

Figure 1. Comparison of healthy retina and retina with drusen deposits in OCT scans [1]

AI is becoming more and more prevalent in the medical field, with doctors' opinions being supplemented with the outputs of highly specialized and complex models. We believe retinal OCT scans are an area with potential for developing new useful models that can help doctors. This project aims to streamline the diagnosis process of retinal diseases by utilizing neural network image classification. The automated analysis would lead to a more accurate and efficient diagnosis by identifying potential disease indicators (CNV, DME, Drusen). Our goal is to make test different model architectures utilizing neural network image classification and evaluate the model's ability to identify retinal diseases indicators. The input for these models would be images of retinal OCT scans and the output would output diagnoses of potential disease indicators (CNV, DME, drusen, or normal/none) with at least accuracy, precision, recall, and F1 score of 0.9.

2. Related Works

There is previous work done in the identification of retinal abnormality signs using Deep Learning (DL) based methods on OCT images. This research has generally revolved around Convolutional Neural Networks (CNNs) such as VGG-16 because of their past effectiveness in other medical diagnoses purposes. From these pretrained models, transfer learning and fine-tuning techniques were used to learn specific features [4] [5] [8]. VGG-16 is a 16 layer deep neural network of convolutional layers followed by maxpool layers with two fully connected layers at the end, and it has been trained as an object detection and classification algorithm.

In our work, we will take a different approach towards this problem, combining approaches that have been described in literature before. Specifically, we will attempt to utilize a U-Net architecture along with Attention-Based CNNs to create better results.

3. Method / Approach

3.1. Transfer Learning with ResNet

For our first approach, we decided to use transfer learning in order to benefit from large pre-trained models trained with much more resources than we had access to. We chose to use the ResNet-50 model, which ships by default with the torchvision library. ResNet-50 is a deep convolutional neural network designed for image classification. It is 50 layers deep and was trained on the ImageNet dataset which includes millions of labeled images.

3.1.1 Architecture

ResNet-50 uses a series of stages, where each stage contains a convolutional block and an identity block. The identity block puts the input through some convolutional layers then adds the original input to the output. The convolutional block contains an additional 1x1 convolutional layer that is meant to reduce the number of filters before being sent to the identity block. ResNet also utilizes skip connections, in which certain values "skip forward" some layers before being put back into the model. This allows the model to learn deeper architectures without worrying about the vanishing gradient problem.

To adapt ResNet-50 to our classification problem, we had to adjust the final linear layer to have an output dimension of 4, representing the number of classes that we want to identify (CNV, DME, Drusen, Normal).

3.1.2 Training

During training, we used multiple strategies to optimize convergence and generalization within a small number of epochs. Our model minimized cross entropy loss which fits our multiclass classification for OCT scans. The logarithmic loss function results in high penalties for confidently incorrect predictions. The softmax is also applied to calculate negative log-likelihoods of the logits.

We used an Adam optimizer due to its robustness, adaptability, and fast convergence. We considered using SGD and RMSProp but chose Adam for our specific application. The OCT scans have many fine-grained, inconsistent features like drusen deposits that may lead to sparse gradient updates. Adam is able to handle this well and could converge to an optimization within 10 epochs.

3.2. U-Net Based Classifier

A U-Net classifier was implemented with the goal of creating a model that could outperform the baseline pretrained ResNet model. Introducing a U-Net creates a decoder pathway and skip connections that help preserve spatial information. This information is normally lost during the regular downsampling section, which is a reason CNNs struggle with image segmentation, but the change of adding the upsampling section helps mitigate these losses [3].

3.2.1 Architecture

The architecture involves a three stage encoder for downsampling, a bridge latent representation layer, and a 3 stage decoder for upsampling (Table 1).

Layer	Channels	Operations
Input	3	Input Image
Down1	32	DoubleConv + MaxPool
Down2	64	DoubleConv + MaxPool
Down3	128	DoubleConv + MaxPool
Bridge	256	DoubleConv
Up1	128	TransConv + Skip + DoubleConv
Up2	64	TransConv + Skip + DoubleConv
Up3	32	TransConv + Skip + DoubleConv
Output	1	Dropout(0.2) + Conv1x1

Table 1. UNet Architecture: Channel dimensions and operations at each layer.

The first step, data preprocessing of the OCT scans includes data augmentation and normalization. To enhance U-Net's generalization ability, a slight random rotation, flip, and affine translation was applied. The preprocessed images are passed into the encoder for downsampling and feature extraction, where spatial hierarchies and contextual representations are progressively learned through stacked convolutional layers.

Each stage of the encoder consists of a double block of a 3x3 convolution layer, batch normalization, and ReLU layers. The convolutional layers capture the critical complex features important while batch normalization stabilizes the training and ReLU introduces non-linearity. We perform max pooling to reduce spatial resolution and increase computational efficiency.

Then, the bridge captures the critical latent features, passing the high level information from the encoder to the decoder. Next, the decoder builds back the spatial information of the images while integrating context from the encoder using skip connections. The upsampling is done through double convolution blocks and transpose convolutions. We considered using bilinear upsampling but chose to use transpose convolutions instead because of its learnable parameters. Lastly, dropout was applied for better generalization and the final layer outputs a class probability prediction for each of the retinal disease indicators.

3.2.2 Training

Our training process was similar to that for our transfer learning model, in which we used the Adam optimizer to minimize cross-entropy loss. We did choose, however, to train for 40 epochs instead as we felt that a U-Net might take slightly longer to converge.

Additionally, we added a learning rate scheduler to stabilize training and weight decay (L2 regularization) to prevent overfitting. On the condition that the validation loss plateaus for 5 consecutive epochs, our learning rate scheduler reduced learning rate by a factor of 0.1. This fine tuning helps the model make small gradient steps towards the optimal loss. The weight decay was set to 0.0001 to help with regularization by penalizing large weights.

3.3. Attention-CNN Classifier

We implemented an Attention-Based CNN which augments a standard traditional fine-tuned pre-trained convolutional neural network (CNN) with channel-wise attention to classify macular diseases from OCT images. The attention mechanism recalibrates intermediate feature maps. This allows the network to emphasize relevant patterns while suppressing background noise [6].

3.3.1 Architecture

The Attention-CNN consists of three encoder stages followed by a global classification head. Each stage comprises a convolutional block, a channel-attention module, and spatial downsampling (Table 2).

Attention Modules. Each Squeeze-and-Excitation block begins with global average pooling over spatial dimensions, producing a channel-descriptor vector. Two fully-connected layers (reduction ratio 16) learn non-linear interactions, and a sigmoid gating produces per-channel weights. These weights rescale the block's convolutional outputs before pooling, ensuring the network adaptively focuses on clinically relevant features.

Layer	Channels	Operations
Input	1	256×256 grayscale OCT image
ConvBlock 1	32	Conv3×3 + BN + ReLU
Attention 1	32	Squeeze-Excitation block
Pool 1	—	MaxPool2×2
ConvBlock 2	64	Conv3×3 + BN + ReLU
Attention 2	64	Squeeze-Excitation block
Pool 2	—	MaxPool2×2
ConvBlock 3	128	Conv3×3 + BN + ReLU
Attention 3	128	Squeeze-Excitation block
Pool 3	—	MaxPool2×2
Global Pool	—	AdaptiveAvgPool
FC	4	Linear + Softmax

Table 2. Attention-CNN Architecture: channel dimensions and operations at each stage.

Unlike vanilla CNNs that treat all channels equally, the attention branch dynamically highlights discriminative feature maps—critical for detecting small, localized drusen deposits.Attention weights can be visualized to localize regions driving each class decision, supporting model explainability in a clinical setting.

3.3.2 Training

For our training process We used the Adam optimizer (initial learning rate 1×10^{-3}, weight decay 1×10^{-4}) to minimize cross-entropy loss. A ReduceLROnPlateau scheduler (`patience`=2, `factor`=0.5) lowered the learning rate whenever validation loss plateaued, allowing finer convergence in later epochs. To ensure that data variation would not impact our model we implemented the following augmentations to our data: random horizontal flip, random rotation (±15°), random affine translation (±10 pixels), Resize to 256×256, ToTensor, Normalize(mean=0.5, std=0.5).

A batch size of 32 was used for both training and evaluation. Early inspection of the training curves indicated convergence by epoch 8, confirming that the attention-enhanced backbone learns discriminative features rapidly.All convolutional layers use kernel size 3 with padding 1. Batch normalization follows each convolution to stabilize gradient flow. The global pooling head reduces spatial dimensions to 1×1 before a $128 \rightarrow 4$ linear classifier outputs softmax probabilities over the four disease categories.

4. Data

Our overarching task is image classifications, so as our primary dataset we used a repository of OCT images published by Mendeley Data [2]. This dataset contains 104809 images of retinal scans, labeled for CNV, DME, Drusen, and normal. 1000 of these images comprise the test dataset (250 in each category), while the remaining 103809 images comprise the training dataset. This dataset is large enough and contains images of the same size and perspective, reducing the preprocessing requirements and eliminating training size as a potential cause of inaccuracy. When preparing the dataset for model training, we performed the following preprocessing steps:

- Resizing to a uniform size
- Normalizing the pixel values
- Some data augmentation techniques such as RandomRotation and RandomHorizontalFlip

This dataset was procured by the University of California at San Diego in collaboration with the Guangzhou Women and Children's Medical Center. All images were validated before being labeled.

5. Experiments and Results

Our goal was to create and evaluate OCT image classification models that could identify retinal disease indicators. The goal was to achieve a model that obtained an accuracy, precision, recall, and F1 score of 0.9 or higher

5.1. Transfer Learning with ResNet

We tested the convolutional model using our test dataset of 1000 total OCT scans that not used during training. The test dataset contained 250 samples of each class (CNV, DME, DRUSEN, NORMAL). The data was normalized and resized to 224x224 and the model was tested with a batch size of 32. Our overall accuracy was 0.9310, which exceeded our 0.90 baseline. The precision, recall, and F1-Scores per class are displayed in Table 3.

Class	Precision	Recall	F1-Score
CNV	0.83	0.97	0.90
DME	0.95	0.92	0.94
DRUSEN	0.98	0.86	0.92
NORMAL	0.98	0.97	0.97

Table 3. Per-Class Performance of Pretrained ResNet-50 on OCT Images

Our transfer learning model met our goal metrics of an accuracy, precision, recall, and F1 score 0.9 or higher. On average, the model reached a precision of 0.9367, a recall of 0.9310, and an F1 score of about 0.9316 indicating balance between recall and precision. The model is most reliable for the NORMAL and DME cases. The precision of classification for DRUSEN, NORMAL, DME is generally very high, indicating strong confidence in the model diagnosis of positive cases.

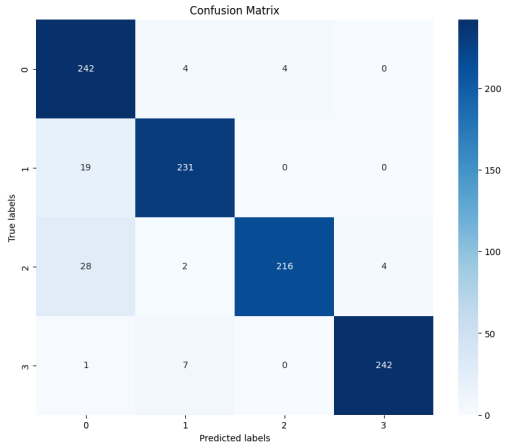

Figure 2. Confusion Matrix for Pretrained ResNet-50: visualizing the performance across the four classes

Looking at the confusion matrix (Figure 2), we can visualize the drop in precision for CNV. Several images were mislabeled as DME or DRUSEN (interestingly not NORMAL though). This also had an effect on the recall for DME and DRUSEN which are lower. Notably though, there were only four cases of false negatives, where diseased OCT scans were labeled as normal (this can also be seen by the 98% precision of NORMAL). This is quite important, as machine learning in the medical field often tends to prioritize minimizing false negatives. After all, it would be worse to leave a present condition untreated than to attempt to treat a condition that is actually not there.

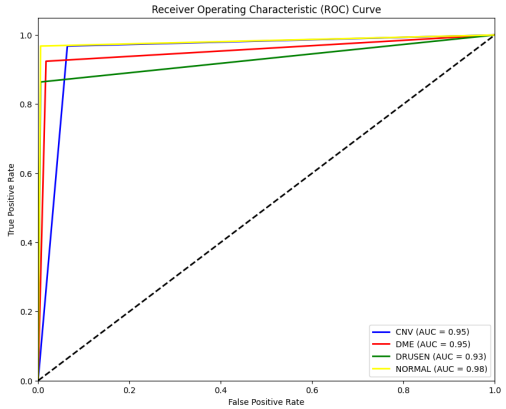

Figure 3. ROC curves for the Pretrained ResNet-50's multi-class classification, with corresponding Area Under the Curve (AUC) values.

We also plotted the ROC curve (Figure 3) for each class to visualize the model's ability to minimize false positives and maximize true positives. The area under the curve (AUC) gives us the model's ability to accurately determine a single class. All classes performed high, with only Drusen having an AUC under 0.95.

5.2. U-Net Based Classifier

We tested the U-Net model using the same 1000-image test dataset. The data was normalized and resized to 128x128 and the model was tested with a batch size of 32. The resulting accuracy was 0.9375, surpassing our 0.90 goal. Overall, the model was able to output correct classifications for the OCT scans (example Figure 4).

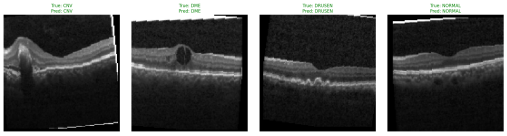

Figure 4. Example Classifications for U-Net Based Classifier: Optical coherence tomography (OCT) images with retinal disease indicators were classified at an average accuracy of 0.9375

The precision, recall, and F1-score metrics per a class are summarized in Table 4.

Class	Precision	Recall	F1-Score
CNV	0.88	0.98	0.93
DME	0.93	0.98	0.95
DRUSEN	0.99	0.81	0.89
NORMAL	0.95	0.97	0.96

Table 4. Per-Class Performance of U-Net on OCT Images

Our U-Net model met our goal metrics of an accuracy, precision, recall, and F1 score 0.9 or higher. On average, the model reached a precision of 0.9375, a recall of 0.935, and an F1 score of about 0.9325, indicating a balance between recall and precision. The model is most reliable for the NORMAL and DME cases. The precision of classification for DRUSEN, NORMAL, DME is generally very high, indicating strong confidence in the model diagnosis of positive cases.

The model slightly dips in precision for CNV which could be explained by the confusion matrix. According to the confusion matrix (Figure 5) it sometimes will diagnose CNV when the true label is Drusen. The recall for all classes is generally very high, but dips slightly for Drusen. This is likely due to the same confusion of misclassifying Drusen cases as CNV. For all classes, the U-Net model performs above 0.9 in all metric categories (precision, recall, F1-score, accuracy).

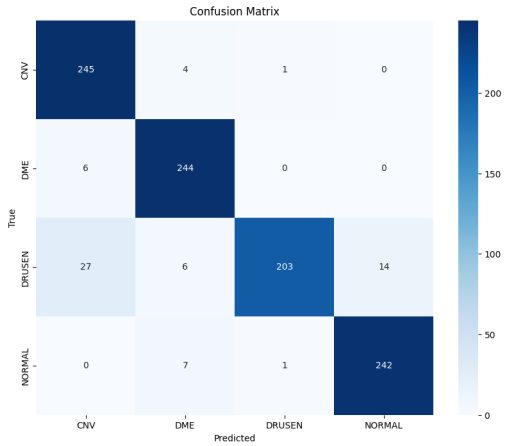

Figure 5. Confusion Matrix for U-Net Based Classifier: visualizing the performance across the four classes

We plotted the ROC (Receiver Operating Characteristic) curves per class to analyze the U-Net's performance in terms of false positives (Figure 6). The model's high AUC (area under the curve) values average to 0.9925, which indicates a high discrimination ability. The model is able to balance its ability to classify true positives while avoiding false positives. This risk mitigation is important for medical imaging classifications. These exceptional AUC scores validate that the model could be useful for an initial screening of OCT scans.

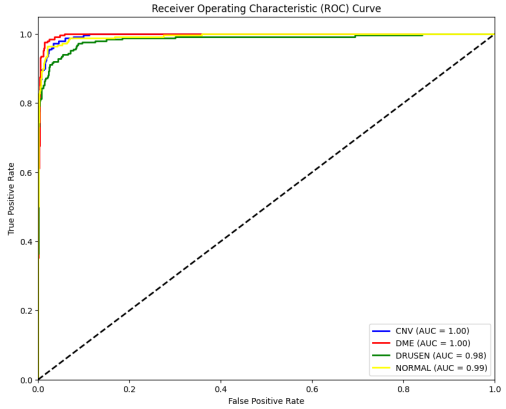

Figure 6. ROC curves for the U-Net model's multi-class classification, with corresponding Area Under the Curve (AUC) values.

5.3. Attention Based CNN

Our Attention-Based CNN model was tested in the same way as the U-Net model: a dataset of 1000 OCT scans split evenly among the four different types of classes. The data was normalized and resized to 256x256 with a batch size of 32. Our overall accuracy was 0.69 which was much lower

than our 0.90 target. Example correct and incorrect classifications are shown in Figure 7.

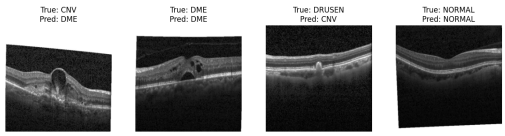

Figure 7. Example Outputs for Attention-CNN Classifier: OCT scans with true labels (top row) and predicted labels (bottom row). Note the frequent misclassification of DRUSEN as CNV.

The precision, recall, and F1-Scores per class are displayed in Table 5

Class	Precision	Recall	F1-Score
CNV	0.56	0.81	0.66
DME	0.86	0.74	0.80
DRUSEN	1.00	0.21	0.35
NORMAL	0.68	0.1	0.81

Table 5. Per-Class Performance of Attention CNN on OCT Images

Our Attention-Based CNN did not meet our goal metrics of an accuracy, precision, recall, and F1 score 0.9 or higher. On average, the model had a precision of 0.77, a recall of 0.69, an F1 score of 0.65. This is largely due to a strong misclassification between DRUSEN and CNV. Our model seemed to consistently misclassify DRUSEN as CNV cases. We noticed this in the U-Net and we hoped that using an attention based model would be able to overcome the slight dip in precision for CNV. However, according to the confusion matrix, (Figure 8) it is apparent that this weakness is not only still persistent but also much more prevalent.

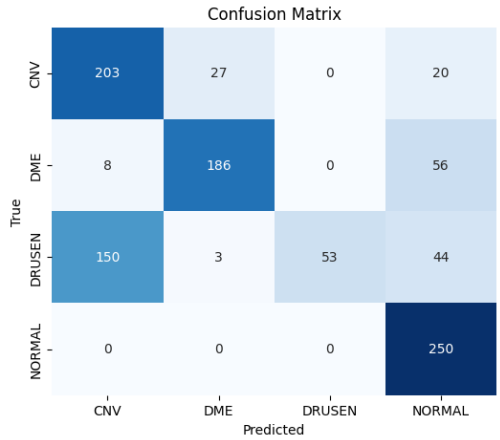

Figure 8. Confusion Matrix for Attention-CNN Based Classifier

Receiver Operating Characteristic curves (Figure 9) show AUCs of 0.88 (CNV), 0.97 (DME), 0.95 (DRUSEN), and 1.00 (NORMAL). Although ROC-AUC remains high—especially for DRUSEN—the extremely low recall for DRUSEN (0.21) implies our classification threshold or class weighting is ill-matched to that class's score distribution.

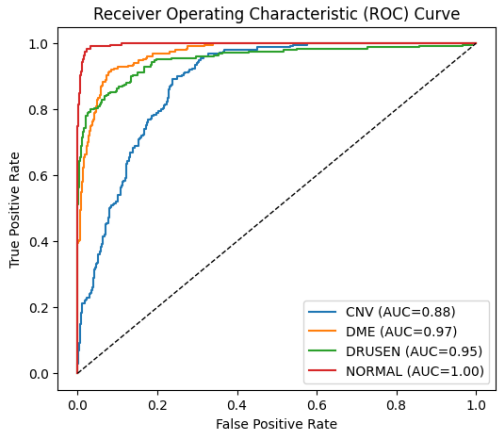

Figure 9. ROC Curves for the Attention-CNN Model's Multi-Class Classification

Reflecting on our results it seems that there were some flaws in our original architecture design. Since we implemented the attention CNN after implementing our U-Net, we assumed that downsampling would not drastically eliminate the drusen signal. However, since our model uses SE blocks, we are collapsing each feature-map to a single scalar which loses spatial granualrity. CNV lesions tend to be larger, higher-contrast, and drive strong global activations; drusen are small, low-contrast deposits that get "washed out."

Some measurable next steps are to replace our SE modules with Convolutional Block Attention Modules that compute channel and spatial attention maps. This way we will be properly encoding spatial information for our data. Ideally this will be able to help us dramatically increase our recall.

6. Conclusion

Ultimately, both our transfer learning with ResNet and our U-Net architecture yielded results above our baseline of precision, recall, and F1 scores above 0.9. Especially crucial to this is the strong precision score of 0.98 for normal images with transfer learning and 0.95 for normal images with U-Net, indicating a low rate of false negatives, which is important to minimize given our context of medical research and patient livelihoods. This demonstrates the strengths of the downsampling and upsampling nature of a U-Net architecture which helps build the bridge that holds the latent features of our OCT scan dataset. Similarily,

because we were using a pretrained model in ResNet-50, much of the initial feature detection in images has already been finetuned and optimized. As a result, because we were able to build off an already strong foundation of feature detection, we were able to specialize ResNet to fit our dataset by replacing the last layer, explaining the high scoring ability of our transfer learning model.

However, our attention-based CNN approach failed to reach the same criteria, with an over 20% drop in accuracy compared to the other models. There was evidence that this model was overfitting on the training data, which would have strongly contributed to a decrease in the test accuracy. Consequently, a future avenue of research could have involved tuning the Attention architecture and its hyperparameters such that it encounters less overfitting during its training procedure.

Given the strong precision and recall scores of our U-Net architecture, such a model can be trained to diagnose other such diseases or ailments which involve analyzing patient scans or images. For example, another closely related subject could be corneal OCT scans, which are scans for a different part of the eye, or MRI brain imaging scans. Doctors analyze these scans and charts for equally small wounds or symptoms as with a retinal OCT, which means it is just as easy to miss or misinterpret any one image. As a result, future work can be done to develop a robust dataset for these scans and then train a similar U-Net architecture.

7. Ethical Considerations

The integration of artificial intelligence for diagnosis of retinal diseases raises several ethical considerations. These ethical implications must be acknowledged to ensure responsible usage of such technology. One of the main ethical concerns is the risk of bias in image classification using neural networks. If the data used to train the deep learning model is not diverse and representative of all demographics, the model could be less accurate for certain populations. Secondly, the model's diagnosis is not a replacement for professional human diagnosis. The model is trained on common patterns in OCT scans and makes generalizations based on the training data. In practice, comprehensive diagnosis involves taking into account a patient's medical history and other necessary factors that the model does not consider. Ensuring that the technology is used to simply assist human analysis is crucial for accurate medical treatment.

8. Team Contributions

Shaunak Warty contributed towards the experimentation and implementation process for the Attention-Based CNN. He worked alongside Sam Sukendro to complete these tasks. He also worked on data collection and preprocessing to ensure bad data was not added and to add random augmentations to strengthen the dataset. Shaunak also helped to train and finetune the U-Net model and assisted in the implementation of the transfer learning.

Sam Sukendro was the primary contributor towards the experimentation, implementation, and training process for the Attention-Based CNN. He also tested various validation methods and determined the appropriate scores to use for our model result benchmarks. He also helped to train and finetune the U-Net model.

Felicia Jamba took the lead on the experimentation, implementation, and training process for the U-Net model. She also helped Shaunak on data collection and preprocessing the dataset while contributing towards the implementation of the transfer learning on ResNet.

Michael Chian worked on the implementation of the Transfer Learning with ResNet model and also contributed on the implementation and finetuning of the U-Net model. He also helped to train and debug the Attention-Based CNN.

References

[1] An illustration of oct images. from left to right: Healthy, drusen, ga, and wet amd. ResearchGate. Accessed 2025. 1

[2] Daniel Kermany, Michael Goldbaum, Wenjia Cai, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. 4

[3] Jongwoo Kim and Luan Tran. Retinal disease classification from oct images using deep learning algorithms. https://lhncbc.nlm.nih.gov/LHC-publications/PDF/IEEE-CIBCB-2021-Paper11-JongwooKim.pdf, 2021. Accessed 2025. 2

[4] I. Leandro, B. Lorenzo, M. Aleksandar, et al. Oct-based deep-learning models for the identification of retinal key signs. *Scientific Reports*, 13:14628, 2023. 2

[5] Yu-Ying Liu, Mei Chen, Hiroshi Ishikawa, Gadi Wollstein, Joel S. Schuman, and James M. Rehg. Automated macular pathology diagnosis in retinal oct images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding. *Medical Image Analysis*, 15(5):748–759, 2011. Special Issue on the 2010 Conference on Medical Image Computing and Computer-Assisted Intervention. 2

[6] S.S. Mishra, B. Mandal, and N.B. Puhan. Macularnet: Towards fully automated attention-based deep cnn for macular disease classification. *SN Computer Science*, 3:142, 2022. 3

[7] David B. Rein et al. Prevalence of age-related macular degeneration in the us in 2019. *JAMA Ophthalmology*, 140(12):1202–1208, 2022. 1

[8] Yibiao Rong, Dehui Xiang, Weifang Zhu, Kai Yu, Fei Shi, Zhun Fan, and Xinjian Chen. Surrogate-assisted retinal oct image classification based on convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, 23(1):253–263, 2019. 2

[9] A.M. VanDenLangenberg and M.P. Carson. Drusen bodies. In *StatPearls*. StatPearls Publishing, 2025. [Updated 2023 May 1]. 1