Enhancing Image Processing through Pretext Learning with Transformation Prediction in Self-Supervised Learning: A Case Study in Medical Imaging

Pramit Dutta

***Abstract*— In recent years with the rapid advancement in machine learning technologies, the complexity of the model is also increasing that requires vast labeled datasets for robust performance. This issue becomes even more significant in field of medical imaging where interpreting image demand specialized expertise and rigorous analysis. Considering to ensure feasibility and effectiveness of solution as well as reduce dependency on manually labeled data, in this study, a self-supervised learning (SSL) framework was implemented on an Optical Coherence Tomography (OCT) dataset where transformation-based pretext learning method was leveraged to capture the undermined geometry of the dataset as well as pattern of the images of different class. The OCT dataset is a fully labeled dataset where the labels were stripped off from 60% of the images which were used for pretext learning and later the acquired representations of the data were further passed to a model which fine tuned on the 40% downstream data and provided the prediction. The main idea behind the approach was to learn the undermined geometry of the data and using the feature to make our model robust in data-efficient manner. The quantitative analysis on this approach shows that this framework achieved an average classification accuracy of 87.11% with a precision of 0.84, recall of 0.83 and an F1 score of 0.82. These results highlight the capability of SSL approach to effectively capture intricate pattern and correlation of data points in a dataset which can help complex model to achieve robust performance using minimum labeled images.**

***Index Terms*— Image Processing, Optical Coherence Tomography, Pretext Learning, ResNet-50, Self-Supervised Learning**

I. Introduction

ACCORDING to a study of World Health Organization (WHO), in 2019, about 2.2 billion people experience various vision problems that significantly impact the quality of life [1]. One of the main reason of vision impairment is Age-related Macular Degeneration (AMD) which can impact the central area of the vision.

A. Background

AMD is also divided into several classes where every class has different types of characteristics and require different types of treatment protocol. One of the serious case of AMD is from the wet class which is Choroidal Neovascularization (CNV) that can cause hemorrhage [2]. As a result, the macula and the area dense with photoreceptor are impacted, potentially leading to blindness, as this region is essential for high-resolution vision. In a study, it was found that in USA the number of CNV patient is 2 million [3].

DRUSEN is a type of AMD that is generally considered as intermediate stage where the degeneration occurs in the center area of the foveolar. About 7 million people in the United States suffer from DRUSEN each year, a condition often linked to aging that can lead to vision problems [4]. On the contrary, diabetic macular edema (DME) is a condition linked to diabetes, often seen as a complication of diabetic retinopathy and involving changes in the retina. Among these disease DME is the most common one and a study shows each year about 7.5 million people diagnosed with this disease [5].

B. Problem Statement

Most of the AMD and DME can be treated effectively if detected earlier and for this purpose a special type of imaging method is used known as Optical Coherence Tomography (OCT). The treatment protocol of different types of AMD and DME often based on the evaluation of OCT which is basically an inside photo of the eyes and a two or three dimensional images is formed based on the reflection of tissue [6]. However, detecting these disease effectively using a traditional feature extraction method or using a supervised learning approach requires vast amount of labeled data which requires a lot of time as well as rigorous analysis and expertise which is always a problem considering the inadequate number of optometrists.

This article presents a project completed as part of the CIS*6020 course at the University of Guelph in Fall 2024, following IEEE Transactions format. This course, instructed by Professor Neil Bruce, was completed as part of the credit requirements for the Master of Applied Science in Engineering with a specialization in Artificial Intelligence (MASc. Engg+AI) program. The report will be submitted for grading in alignment with these academic requirements.

P. Dutta is a Master of Applied Science in Engineering candidate specializing in Artificial Intelligence with the School of Engineering, University of Guelph, Guelph, ON N1G 2W1, Canada (e-mail: pdutta@uoguelph.ca)

C. Scope of Study

Nowadays, in the realm of computer vision, one of the main obstacle in path of the feasibility of model is the need of labeled data. In case of medical imaging, it's become more prominent because unlike the other computer vision task, medical imaging is more critical to classify as this diagnostic often contain multiple complication where expert from different genre often has to provide their insight. That's why it requires a lot of time and rigorous analysis and often a lot of background history to associate an image with a specific task.

There are also cases where the because of faulty image acquisition technique it's becomes almost impossible for a human diagnostician to interpret the images. This highlights the necessity for data-efficient models that can effectively extract insights from constrained labeled datasets, reducing the burden of extensive annotation demands. These models hold the potential to streamline diagnostic workflows while maintaining reliability, even in the presence of incomplete or corrupted imaging data, thereby advancing the practicality of automated medical imaging solutions.

D. Study Significance

To address this challenge, this project proposes the implementation of a self-supervised learning (SSL) model utilizing pretext-learning techniques. One promising approach in SSL is transformation prediction, where the model learns to predict the transformations applied to an image (such as rotation, scaling, or flipping). This task, commonly referred to as pretext learning, allows the model to capture the geometric structure of the images, enhancing its ability to generalize across different data. Traditional deep learning models often depend on large, labeled datasets for image classification, which makes them less effective in scenarios where labeled data is scarce. Moreover, these models typically struggle with generalization across diverse image sets.

In the context of medical imaging—where accuracy is paramount—transformation prediction in self-supervised learning offers a compelling alternative. By leveraging unlabeled data, this approach helps the model learn robust features through geometric transformations, making it more adaptable and efficient in classifying unseen or rare medical conditions. Self-supervised learning model with a pretext learning approach has the potential to improve diagnostic accuracy, reduce the cost and time associated with data labeling, and make AI-driven diagnostics more accessible across healthcare systems.

E. Contribution

The main contribution in this paper is the developed framework for retinal disease detection which use Self-Supervised Learning (SSL) approach. This model use the approach using a technique called pretext learning. Pretext learning means using a task to understand the pattern of a data where the task is not the desired or final classification or desired prediction.

The pretext task in this case is the geometric transformation as this type of task prediction requires the model to learn geometric structure of an image. This type of feature generally works well as primary feature that helps the model to learn common element in the pictures which helps the model to differentiate between different classes. It also uses unlabeled data to learn important patterns in the dataset, thereby enabling its application to other medical imaging tasks and making it both feasible and useful.

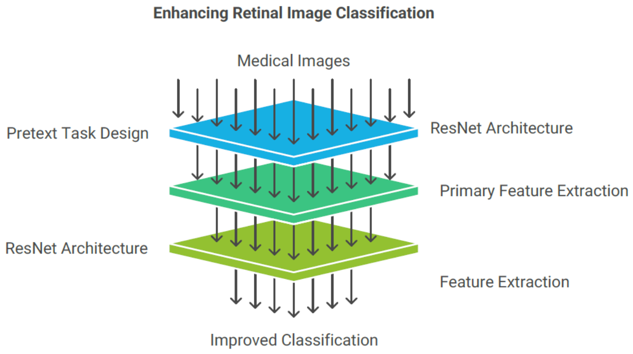


Fig. 1. Framework for Enhancing Retinal Image Classification Using Self-Supervised Learning (SSL). It illustrates the use of pretext tasks and ResNet architecture for improved retinal image classification.

The figure-1 shows the basic approach that is used for this project's implementation which shows how feature extracted from pretext learning can enhance classification efficiency of a model. This framework makes the model data efficient and reduce annotation dependency.

The contribution of the study can be listed as:

- Development of a novel framework for **retinal disease detection**.
- Utilizes Self-Supervised Learning (SSL) as the core approach.
- Employs a pretext learning technique to solve auxiliary tasks for learning data patterns.
- Pretext tasks aid in understanding the underlying structure of the data, even though they are not directly related to the final classification or prediction objective.

In this paper, Section II reviews the related work, providing an overview of prior advancements in medical imaging and self-supervised learning. Section III details the proposed methodology, including pretext learning and downstream tasks. Section IV describes the experimental setup and dataset utilized in this study. Section V presents the results and offers a detailed discussion on the findings. Finally, Section VI concludes the paper, highlighting the key contributions and future directions.

II. Related Work

Several studies have explored self-supervised learning through pretext tasks to improve the classification of medical images, focusing on learning key features from the data itself. Such research focuses on the importance of allowing models to learn critical features directly from the data itself, without relying on labeled datasets, and so tackling the problems created by limited labels in medical imaging.

For instance, Z. Tan et al. [7] proposed a self-supervised learning method using Masked Autoencoder (MAE) that learns by masking portion of the images and reconstructing the missing part of the messages. The model also used a self-distillation mechanism to transfer knowledge from the global decoder to the local encoder, improving the quality of feature extraction. The model achieves a recognition accuracy of 97.78% on SARS-COV-CT and experimental accuracy of 99.6% and 95.27% on these two datasets.

In one research study, H. Yu et. al [8], proposed a multitask learning approach that combines discriminative and generative approach where the model learn diverse and complementary representation from medical images without relying on labeled data. The framework incorporates multiple self-supervised tasks, including segmentation, registration, and reconstruction which tackles the obstacle of labeled data requirement to improve performance of the model. Finally, the approach was evaluated on various dataset where this approach showed improved performance compared to fully supervised methods.

In another research, a framework called DiRA was implemented by F. Haghighi et. al [9], that unifies discriminative, restorative, and adversarial learning for self-supervised medical image analysis. The discriminative component captures high level feature while the restorative component preserves local grained feature and adversarial component makes sure the generated features are different from original data. The ensembled learning approach was test for various task including classification and segmentation where the model outperformed fully supervised models achieving 97.58% accuracy in liver segmentation.

In one study, Ouyang et. al [10] introduced a framework called as SSL-ALPNet which integrates super pixel based self-supervised learning approach and adaptive prototype pooling network. The super pixel-based pseudo levels and ALP pooling are designed to capture fine grained local features and improved segmentation performance which is evident from the robust result found from the analysis that shows a 10-point increase in dice score.

In another study, X. B. Nguyen et al. [11] introduced a self-supervised learning framework based on spatial awareness for medical image analysis, utilizing Convolutional Neural Networks (CNNs) to extract spatial features from 3D medical images. The framework focuses on pretext tasks such as corrupted patch detection and spatial index prediction, allowing the model to learn spatial relationships in the data. The model achieved a 91.68% Dice score in organ-at-risk segmentation, outperforming traditional self-supervised methods.

S. Albelwi [12] presented a comprehensive survey on self-supervised learning (SSL), focusing on auxiliary pretext tasks and contrastive learning for imaging. The study explores various pretext tasks such as image rotation prediction and missing part reconstruction, which help models learn useful representations from unlabeled data. The survey concludes that SSL techniques can significantly enhance performance in imaging tasks without labeled data.

In another study, J. Dominic et al. [13] introduced a self-supervised learning framework based on inpainting-based pretext tasks for improving medical image segmentation. The approach involves tasks like context prediction and context restoration, where random image patches are masked or swapped, allowing the model to recover the original image. The proposed method was applied to MRI and CT scans, achieving a statistically significant improvement in segmentation performance in low-labeled data regime.

In one study, Thanaporn Viriyasaranon et al. [14] proposed a pretext learning based self-supervised learning approach for medical imaging that uses geometric pseudo-shape segmentation as a pretext task. In this study, the researcher added synthetic shapes to CT images trained a model to detect added shape in images, enabling it to learn meaningful features without manual labels. The framework was tested on liver, pancreatic, and breast cancer datasets, achieving significant improvements in performance. For instance, the proposed method improved classification accuracy by 4.2% and 2.41% for liver and pancreatic cancer datasets, respectively, when combined with a hierarchical transformer-based model. It also enhanced tumor segmentation performance with a mean IoU of 0.723 and a Dice score of 0.809 on the pancreatic cancer dataset.

In a study by Hosseini et al. [15], a domain adaptation framework was proposed to address class imbalance in histopathology images, using focal loss and a novel co-training approach. The method utilizes pseudo-labeled target samples alongside source domain data, with focal loss enhancing the learning of underrepresented classes. Experimental results demonstrated the framework's effectiveness in prostate cancer classification, achieving improved accuracy (up to 77.65%) and reducing majority-minority class differences across multiple datasets. This approach highlights the potential of domain adaptation and tailored loss functions in overcoming real-world challenges like domain shifts and imbalanced datasets in medical imaging.

In a study by Ren et al. [16], a novel semi-supervised learning framework, UKSSL, was proposed to address the challenge of limited labeled medical images in classification tasks. The framework combines MedCLR, a self-supervised contrastive learning model, with UKMLP, a fine-tuning module utilizing limited labeled data. By using underlying knowledge extracted from unlabeled datasets, the approach achieved superior performance, with precision, recall, and F1-scores reaching 98.9% on the LC25000 dataset using only 50% labeled data. This highlights the effectiveness of combining contrastive learning with supervised fine-tuning for robust medical image classification.

The studies reviewed show how self-supervised learning helps solve the problem of limited labeled data in medical imaging. By using tasks like masking, segmentation, and shape detection, these methods improve how well models can classify and segment medical images. Techniques like DiRA and SSL-ALPNet, along with new transformer-based models, show that self-supervised learning can make models better at learning features and performing well on different tasks.

III. Methodology

This section provides an explanation of the proposed methodology for classifying retinal diseases into four cat-

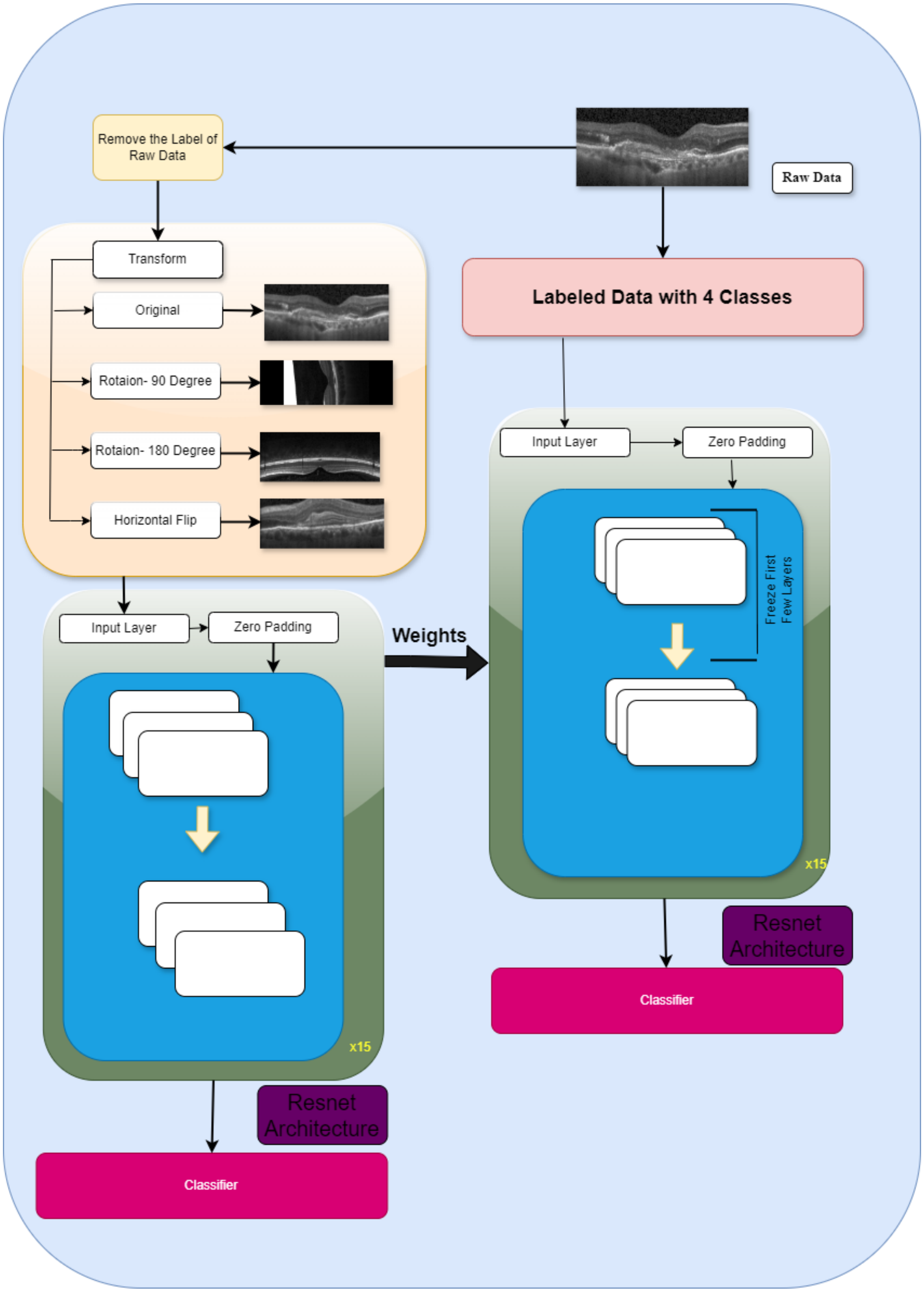


Fig. 2. Self-Supervised Learning Framework utilizes pretext tasks to improve feature extraction in retinal image classification with a ResNet-based architecture.

egories using the OCT image database which begins with removing the labels and data transformations, such as rotations and flips, to create diverse representations and enable the model to learn meaningful patterns. These features are extracted using a pre-trained ResNet architecture, which is fine-tuned for the specific classification task. The figure-2 contains a detailed workflow of this framework and how this framework is implemented. This learning approach allows the model to generate meaningful feature representations from unlabeled data. It reduces reliance on manual annotations, making the framework more efficient for real-world applications.

A. Geometric Transformation

In pretext learning based approach, the model learns different types of pattern via predicting the geometric transformation applied on the images. In this study, transformation like 90° and 180° rotation and horizontal flip was used (figure-2).

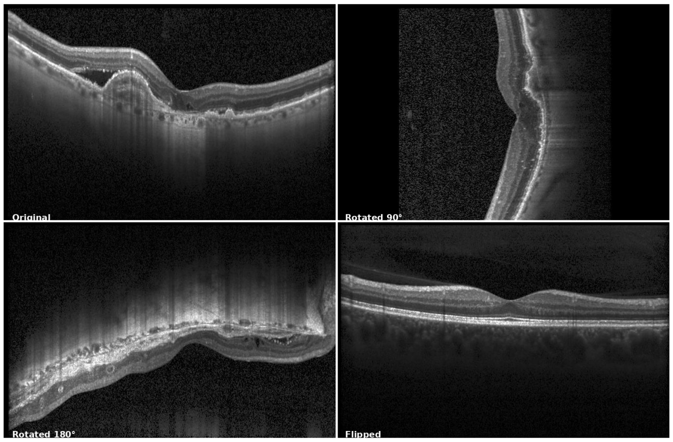


Fig. 3. Different geometric transformation

The main idea of this pretext learning is that the model use geometric transformation to learn primary features of the images such as edge or common pattern and later use limited annotation to fine-tune the model to differentiate between the patterns derived from the pretext method.

B. Modified CNN Architecture

In this framework, ResNet-50 was used as backbone architecture for both pretext and downstream task. ResNet-50 is used for image tasks because its design helps solve problems that come with very deep networks by focusing on learning the difference (residual) between the input and output of layers, instead of learning everything from scratch [17]. The ResNet-50 architecture was the foundation of this feature extraction process, utilizing its complete set of convolutional layers and residual connections. The whole feature extraction model can be represented as:

$$F_{\text{ResNet}} = \text{ResNet}(I) \tag{1}$$

where:

- F_{ResNet} is the output feature map produced by the ResNet model,
- I is the input image.

On the other hand, the residual connection within the ResNet architecture can be mathematically expressed as:

$$H = G(X, W) + X \tag{2}$$

where:

- H represents the output feature map of a residual block,
- $G(X, W)$ is the transformation applied to the input X using weights W,
- X is the input feature map that is directly added back via the skip connection.

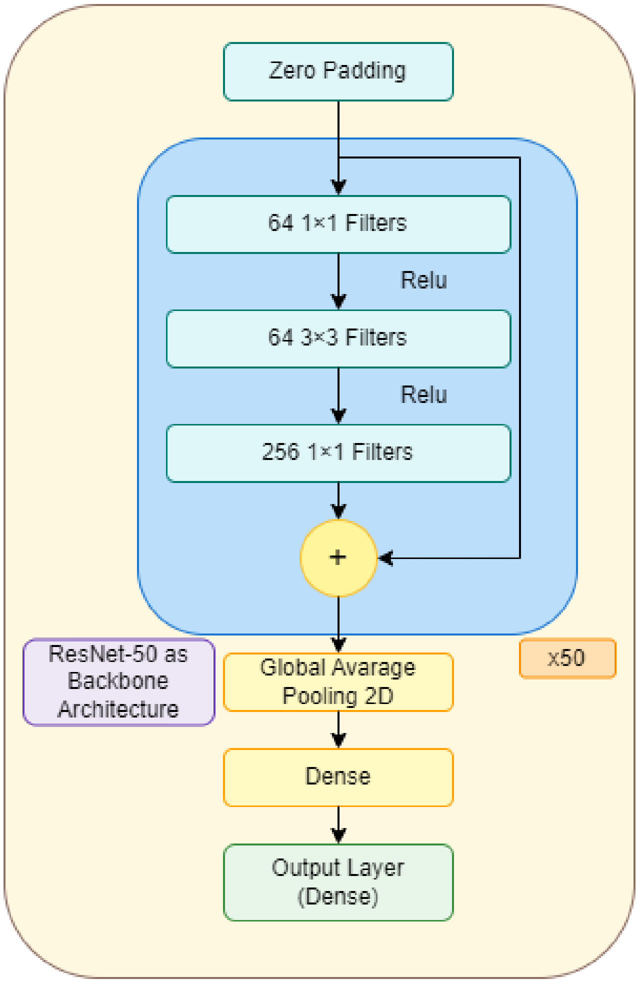


Fig. 4. This diagram depicts the ResNet-50 architecture repeats these residual blocks 50 times, includes residual connections, and concludes with global average pooling, a dense layer, and an output layer for classification.

In ResNet-50, fifty blocks of this residual architecture are stacked together for feature extraction (figure-4). In every block, there are three layers: the first layer has 64 1×1 filters, followed by 64 3×3 filters, and finally a layer with 256 1×1 filters each with ReLU activation [18]. These residual connections help preserve information through the network, and the architecture concludes with global average pooling, followed by a dense layer and an output layer for final predictions.

C. Pretext Learning

In this study, pretext learning involves training the model to recognize transformations applied to almost 60% of the OCT images after removing their labels, such as rotations and flips. By solving this auxiliary task, the model learns to extract meaningful and diverse feature representations from unlabeled data. These transformations encourage the model to focus on structural patterns within the images, which are essential for downstream classification. This approach not only reduces the reliance on annotated data but also enhances the model's ability to generalize across various retinal disease categories. The features learned during this stage form the foundation for fine-tuning the model on the labeled dataset in the downstream task.

D. Downstream Task

In this work, the downstream task involves fine-tuning a model initialized with weights transferred from the pretext learning stage. These weights, pre-trained on the transformation recognition task, are leveraged to initialize the downstream model, ensuring the transfer of meaningful features. To retain the robustness of the learned representations, the first 50 layers of the ResNet-50 backbone are frozen during fine-tuning. This allows the lower layers to retain their ability to capture generic patterns, while the remaining trainable layers are fine-tuned to adapt to the specific classification task.

IV. Experimentation

This section outlines the experimental setup and procedures employed to evaluate the proposed framework for classifying retinal diseases using OCT images. First of all, the dataset used in the experiments is described, including the preprocessing steps applied and the split of data for both pretext and downstream tasks. Then, the experimental setup or the details of training environment is detailed, covering both pretext learning and downstream task. Additionally, the training configuration, including hyperparameters and optimization techniques, as well as the evaluation metrics employed to assess the framework's performance, are also discussed and outlined in this section.

A. Dataset

In this project, a fully labeled dataset of Optical Coherence Tomography (OCT) images with Age-Related Macular Degeneration was used [19]. The images were obtained during routine checkups, and the dataset utilizes a foveal cut from the original images [20]. This dataset has 4 classes- CNV, DME, DRUSEN and Normal with more than 1,00,000 images. In the first stage, 60% for the images were seperated (randomly selected) for pretext task where the labels were removed and other 40 % were used for downstream task. The pie chart below provides a detailed breakdown of the dataset:

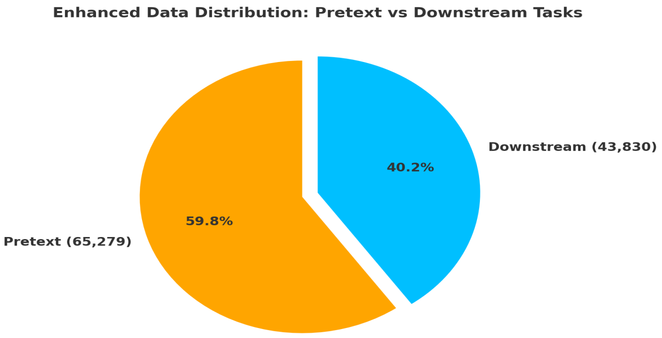


Fig. 5. Split of the dataset for pretext learning and downstream task

Later the downstream set of images are further classified into three subsets- train, validation and test set. Every class in the validation and test set contains 50 images and the rest of the images are used for fine-tuning the test set.

B. Evaluation Metrices

The performance of this model are evaluated based on both quantitative and qualitative analysis. In case of quantitative analysis Accuracy, Precision, Recall and F1 Score was used. The mathematical representations and their definitions are as follows:

- **Accuracy:** Accuracy measures the overall correctness of the model and is defined as the ratio of correctly predicted samples to the total number of samples.
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$
- **Precision:** Precision evaluates the correctness of positive predictions and is defined as the ratio of correctly predicted positive samples to the total predicted positive samples.
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$
- **Recall:** Recall, also known as sensitivity or true positive rate, measures the ability of the model to correctly identify all positive samples.
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$
- **F1 Score:** The F1 Score is the harmonic mean of Precision and Recall, providing a balanced metric when there is an uneven class distribution.
$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Here, TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively

C. Hyperparameter

For the training purpose, ADAM was used as the optimizer function with an learning rate of 0.001 and categorical crossentropy was used as the loss function. For hyperparameter tuning purpose ”Adaptive Learning Rate Scheduling” was used to tune the learning rate over 20 epoches when the model was trained.

D. Training Environment

The training and experimentation for this project were conducted using Google Colab, using its cloud-based infrastructure. The hardware utilized included an NVIDIA Tesla T4 GPU, which provided accelerated computational capabilities essential for handling the training process efficiently. The software environment consisted of Python 3 as the primary programming language and TensorFlow as the machine learning framework.

V. Result and Analysis

The pretext learning based self-supervised learning framework achieved an impressive performance by using only 40% of the labeled data which shows how such framework can be a potential solution for reducing annotation dependency in medical image classification task. This section presents quantitative and qualitative analysis of the model's performance and as well as comparison with other state of the art retinal disease classification model. Additionally, an ablation study is included to evaluate the impact of pretext learning by analyzing the model's performance without the pretext task, highlighting its critical role in feature extraction and classification efficiency.

A. Quantitative Analysis

For quantitative analysis, the class-wise performance and Area Under the Curve (AUC) are considered, with the evaluation conducted on the test set. It achieved an accuracy of 87.11% on the test set. The class-wise performance is detailed in Table 1, providing a comprehensive overview of the model's effectiveness across different retinal disease categories.

TABLE I
Class-Wise Performance Metrics of the SSL Approach

Class	**Precision**	**Recall**	**F1 Score**
CNV	0.79	0.98	0.88
DME	0.90	0.94	0.92
DRUSEN	0.89	0.50	0.64
NORMAL	0.78	0.90	0.83
Average	0.84	0.83	0.82

The results indicate that the model demonstrates strong performance in most classes, with particularly high Precision and Recall for DME and CNV with a F1 score of 0.92 and 0.88 respectively. However, the performance for DRUSEN is notably lower F1 Score of 0.64 which is basically because of the significant drop in Recall (0.50), suggesting challenges in correctly identifying this class. The overall average metrics reflect the model's robust capability to generalize across classes though improvement in DRUSEN class can significantly enhance the model's reliability.

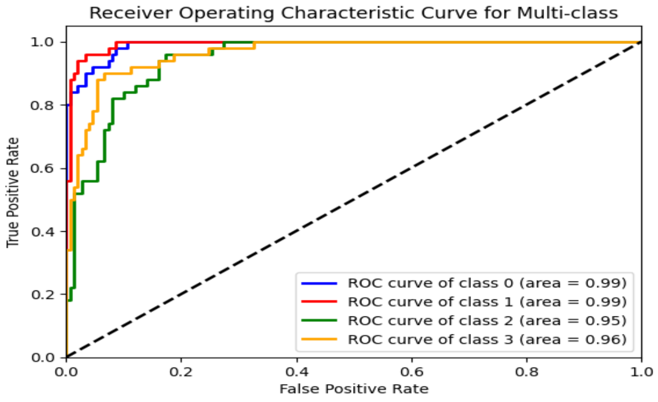
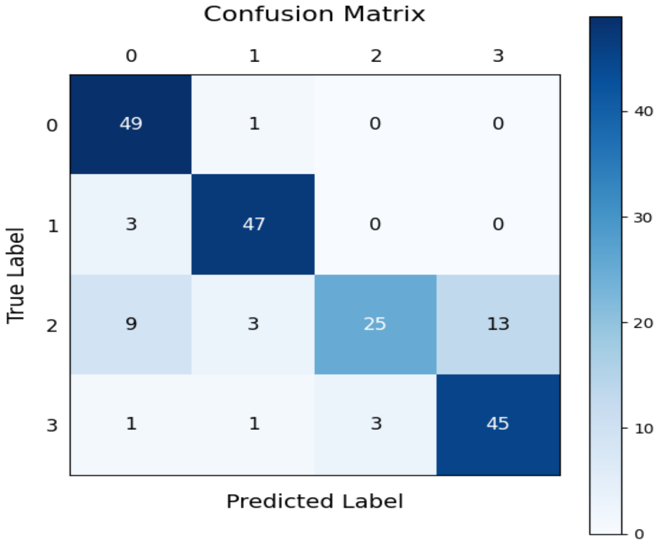


Fig. 6. ROC Curve for Multi-Class Classification Showing AUC Values for Each Retinal Disease Category

The Area Under the Curve is a performance metric derived from the Receiver Operating Characteristic curve. The ROC curve visualizes the trade-off between the True Positive Rate and the False Positive Rate at different classification thresholds. A higher AUC score indicates better discriminative power, meaning the model can distinguish between the presence and absence of a particular class.

From Figure-3, a high AUC value of 0.99 for Class 0 (CNV) and Class 1 (DME), indicate the model's robust performance while classifying these classes. While the AUC for Class 2 (DRUSEN) is 0.95 and Class 3 (NORMAL) is 0.96, which are lower than the other classes, they still indicate that the model performs well in classifying these classes.



	0	1	2	3
0	49	1	0	0
1	3	47	0	0
2	9	3	25	13
3	1	1	3	45

Fig. 7. Confusion Matrix Showing Class-Wise Prediction Performance of the Model

The confusion matrix illustrates the model's performance across the classes, where 0 represents CNV, 1 represents DME, 2 represents DRUSEN, and 3 represents Normal. Figure-7 indicates that the model performs well for CNV (49 correct predictions) and DME (47 correct predictions). However, the performance for DRUSEN is relatively weaker, with only 25 correct predictions and 13 misclassifications into Normal. Normal shows strong performance with 45 correct predictions but has minor misclassifications into other classes. This result indicates the model is making a lot of error just by classifying DRUSEN as normal.

B. Qualitative Analysis

In this section, the extracted features from both the pretext and downstream tasks are analyzed using Grad-CAM that stands for Gradient-weighted Class Activation Mapping. Grad-CAM is used to produce visual explanations for predictions by highlighting the important regions of an image that influence the model's decision. This visualization helps to show how pretext learning detect pattern and correlation of the dataset by extracting primary feature. The image in Figure-8, has been selected as an example. It is chosen to facilitate an in-depth analysis of feature extraction.

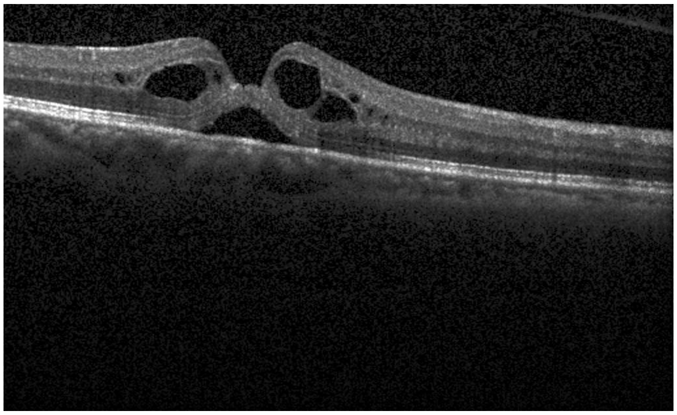
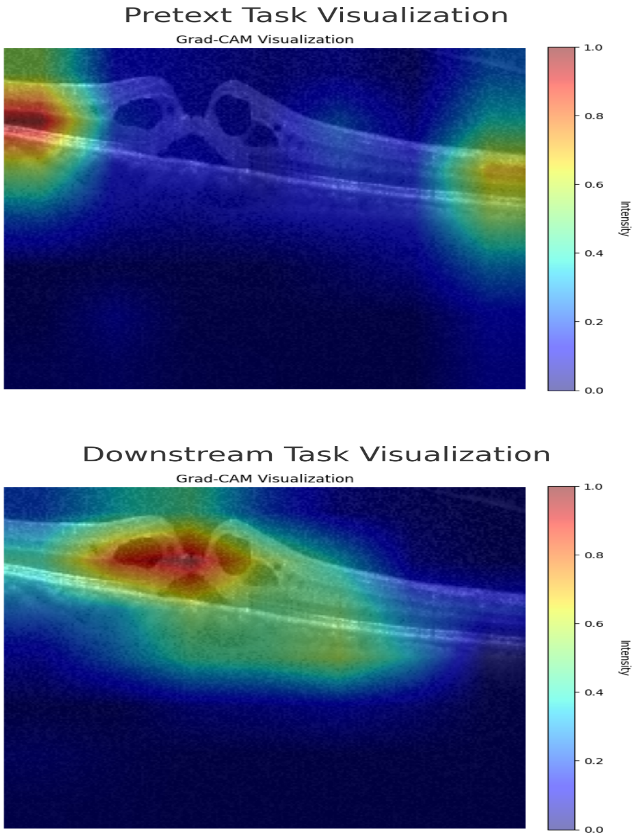


Fig. 8. Original input image selected for feature extraction analysis

The first image of figure 9 shows Grad-CAM visualization of the pretext task, indicates the primary structural regions of the image. These regions represent fundamental patterns learned during pretext training, such as basic textures or prominent features. The pretext task is essential for enabling the model to understand these primary features without relying on specific labels. The second image of figure 9 contains the visualization for the downstream task, demonstrates how fine-tuning builds upon the features extracted during the pretext task. The highlighted regions here reflect the model's refined focus on disease-specific patterns, achieving a higher level of feature discrimination.



Pretext Task Visualization

Grad-CAM Visualization

Downstream Task Visualization

Grad-CAM Visualization

Fig. 9. Grad-CAM visualization highlighting key regions from the original image

In the pretext phase, the model identifies general patterns, such as structural elements or primary features, that are common across the dataset. During fine-tuning, these initial features are refined and tailored to focus on more specific and discriminative patterns, such as disease-related anomalies. This transition from general to specific feature extraction ensures that the model achieves a deeper understanding of the task, enabling more precise predictions.

C. Ablation Study

In this section, a comparative analysis on the effect opf pretext learning is performed where the model was implemented with and without pretext to accurately signify the importance of pretext learning. From Table II, the use of pretext learning results in an increase in F1 scores for almost all classes, except for the Normal class, where a decrease is observed.

TABLE II

Ablation Study: Comparison between with and without Pretext Learning (Bold font indicates the best results).

Class	Model	Precision	Recall	F1 Score
CNV	Without Pretext	0.49	0.92	0.64
	With Pretext	**0.79**	**0.98**	**0.88**
DME	Without Pretext	**0.92**	0.90	0.91
	With Pretext	0.90	**0.94**	**0.92**
DRUSEN	Without Pretext	**0.93**	0.28	0.43
	With Pretext	0.89	**0.50**	**0.64**
Normal	Without Pretext	**1.00**	0.86	**0.92**
	With Pretext	0.78	**0.90**	0.83
Overall Accuracy	Without Pretext	0.78		
	With Pretext	**0.87**		

For CNV, the F1 Score rises from 0.64 to 0.88, and for DME, it improves from 0.91 to 0.92. In the case of DRUSEN, while the F1 Score increases from 0.43 to 0.64, challenges remain in recall improvement. However, for the Normal class, the F1 Score decreases from 0.92 to 0.83, indicating a drop in performance for this category. Overall accuracy sees a substantial improvement, increasing from 0.78 to 0.87, showing the effectiveness of pretext learning in improving classification results. Pretext learning helps the model generalize well by training it on a geometric transformation prediction task. This is fine-tuning after pretext learning significantly enhances the model's overall performance, indicating its effectiveness in improving classification metrics across most categories.

D. Model Comparison

For further evaluation of the proposed self-supervised learning approach with pretext learning, a comparative analysis is performed on the existing age related macular degradation grading. While it is ideal to compare models using the same

TABLE III
COMPARISON OF THE PROPOSED SSL APPROACH WITH STATE-OF-THE-ART MODEL FOR AGE RELATED MACULAR DEGRADATION

Reference	Database	Model	Predicted Class	Performance
[21]	NIH AREDS Dataset	CNN+Random Forest	• AREDS • AMD • Ungradeable	Accuracy: 63.3%
[22]	NIH AREDS dataset	OverFeat DCNN	• CNV • DME • DRUSEN • NORMAL	Accuracy:79.4%
[23]	NIH AREDS Dataset	DeepSeeNet	Severity of Age Related Macular Degradation • Score-0 • Score-1 • Score-2 • Score-3 • Score-4 • Score-5	Accuracy: 67.1%
[24]	SERI Dataset	VGG-16	• Diabetic Macular Edema (DME) • Normal	Accuracy: 87%
[25]	NIH AREDS Dataset	SVM+RF	• hemorrhages • geographic atrophy	Accuracy: (73.9%∼87.4%)
Proposed Framework	Mendeley Dataset	Pretext learning based Self-Supervised Learning approach	• CNV • DME • DRUSEN • NORMAL	**Accuracy: 87.11%**

dataset for consistency, our approach utilizes only 40% of the labeled data from the Mendeley dataset. Therefore, comparisons are made with other age-related macular degeneration grading datasets to demonstrate the robustness of our model.

Table-III contains a comparison of the proposed work with the previous work in the field of age-related macular degeneration grading. From this table, it can be observed that the accuracy varied from 63.3% to 87.4%. F. Grassmann et al. [21] proposed a combination of Convolutional Neural Network (CNN) and Random Forest to classify between AREDS, AMD and ungradeable where the model achieved an accuracy of 63.3%. On the same dataset, in two separate study Y. Peng et al. [23] and P. Burlina [22] proposed DeepSeeNet and OverFeat Deep Convolutional Neural Network (DCNN) respectively. The OverFeat DCNN performed better the former one with a accuracy of 79.4% which is an increase of 12.3% compared to DeepSeeNet. However, though they were using the same dataset, in study [22], they were classifying the class of CNV,DME, DRUSEN and Normal where in study [23] the researcher were focused on grading the age related macular degeneration. On the other hand, a combination of Support Vector Machine and Random Forest Classifier achieved a maximum of 87.4% which can go as low as 73.9% during classifying between hemorrhages and geographic atrophy [25]. On a different dataset (SERI Dataset), VGG-16 achieved an accuracy of 87% while classifying between DME and Normal[24]. Our porposed Framework with only 40% of the annoted label outperforms all of this state of the art model with an accuracy of 87.11 %.

VI. CONCLUSION

This study shows the potential of a self-supervised learning (SSL) framework utilizing transformation-based pretext learning to address challenges in medical imaging classification where often labeled data is limited. While supervised models may still outperform in fully labeled datasets, they require substantial annotation efforts that are costly and time-intensive. Our approach, using only 40% of the labeled data, achieved an accuracy of 87.11%, demonstrating robust performance and data efficiency.

One of the main strengths of this framework is how it extracts important features during the pretext learning phase. In this phase, the model is trained to recognize geometric transformations, it learns the key structural and spatial patterns in the dataset. These learned features provide a strong base for the next steps, helping the model perform well across different retinal disease categories. This approach not only reduces the need for labeled data to make the model more feasible and efficient but also helps the model better understand the shapes and patterns in medical images, making it more reliable for real-world use.

A. Limitations

While the framework performs well, particularly for CNV and DME classes, the performance for the DRUSEN category remains suboptimal likely due to the limited training setup. For resource-constrained experimentation, the model was trained for only 20 epochs, which might have restricted its capacity to fully learn complex patterns. However, even with these constraints, the framework achieved competitive results, showcasing the potential of SSL in data-efficient learning.

B. Future Directions

Future research could explore the integration of this SSL framework with vision-language models (VLMs) where VLM can provide more insights by associating image-based patterns with textual descriptions. Additionally, training for more epochs with higher computational resources can further optimize the model and enhance feature extraction and classification accuracy. Beyond retinal diseases, this framework could be extended to other domains of medical imaging, including tasks like segmentation, providing scalable and innovative solutions across various healthcare scenarios.

Acknowledgment

This project was completed as part of the CIS*6020: Artificial Intelligence course. I would like to thank the AI-Enabled Medical Imaging Analysis Lab for their support and resources throughout the work. The use of large language models and prompt engineering in this article was limited to grammar checking and the creation of Figure 1 to enhance the visualization.

References

[1] A. Ram and C. C. Reyes-Aldasoro, ”The relationship between fully connected layers and number of classes for the analysis of retinal images,” *arXiv*, 2020, arXiv:2004.03624. [Online]. Available: https://doi.org/10.48550/arXiv.2004.03624

[2] National Eye Institute, ”Age-Related Macular Degeneration (AMD).” [Online]. Available: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/age-related-macular-degeneration#section-id-7323. Accessed: Nov. 26, 2024.

[3] N. Ferrara, ”Vascular endothelial growth factor and age-related macular degeneration: From basic science to therapy,” *Nature Medicine*, vol. 16, pp. 1107–1111, 2010. [Online]. Available: https://doi.org/10.1038/nm1010-1107

[4] D. S. Friedman, B. J. O'Colmain, B. Muñoz, S. C. Tomany, C. McCarty, P. T. V. M. de Jong, B. Nemesure, P. Mitchell, and J. Kempen, ”Prevalence of age-related macular degeneration in the United States,” *Archives of Ophthalmology*, vol. 122, pp. 564–572, 2004. [Online]. Available: https://doi.org/10.1001/archopht.122.4.564

[5] R. Varma, N. M. Bressler, Q. V. Doan, M. D. Danese, C. M. Dolan, E. W. Gower, P. B. Greenberg, L. M. Jampol, J. L. Kinyoun, A. M. Kolomeyer, *et al.*, ”Prevalence of and risk factors for diabetic macular edema in the United States,” *JAMA Ophthalmology*, vol. 132, pp. 1334–1340, 2014. [Online]. Available: https://doi.org/10.1001/jamaophthalmol.2014.2854

[6] D. Wang and L. Wang, ”On OCT Image Classification via Deep Learning,” in IEEE Photonics Journal, vol. 11, no. 5, pp. 1-14, Oct. 2019, Art no. 3900714, doi: 10.1109/JPHOT.2019.2934484.

[7] Z. Tan, Y. Yu, J. Meng, S. Liu, and W. Li, ”Self-supervised learning with self-distillation on COVID-19 medical image classification,” Computer Methods and Programs in Biomedicine, vol. 243, 2024, Art. no. 107876. DOI: 10.1016/j.cmpb.2023.107876.

[8] H. Yu and Q. Dai, ”Self-supervised multi-task learning for medical image analysis,” Pattern Recognition, vol. 150, 2024, Art. no. 110327. DOI: 10.1016/j.patcog.2024.110327.

[9] F. Haghighi, M. R. H. Taher, M. B. Gotway, and J. Liang, ”Self-supervised learning for medical image analysis: Discriminative, restorative, or adversarial?,” Medical Image Analysis, vol. 94, 2024, Art. no. 103086. DOI: 10.1016/j.media.2024.103086.

[10] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu and D. Rueckert, ”Self-Supervised Learning for Few-Shot Medical Image Segmentation,” in IEEE Transactions on Medical Imaging, vol. 41, no. 7, pp. 1837-1848, July 2022, Doi: 10.1109/TMI.2022.3150682

[11] X. -B. Nguyen, G. S. Lee, S. H. Kim and H. J. Yang, ”Self-Supervised Learning Based on Spatial Awareness for Medical Image Analysis,” in IEEE Access, vol. 8, pp. 162973-162981, 2020, Doi: 10.1109/ACCESS.2020.3021469.

[12] S. Albelwi, ”Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging,” Entropy, vol. 24, no. 4, p. 551, 2022. DOI: 10.3390/e24040551.

[13] J. Dominic, N. Bhaskhar, A. D. Desai, A. Schmidt, E. Rubin, B. Gunel, G. E. Gold, B. A. Hargreaves, L. Lenchik, R. Boutin, et al., ”Improving data-efficiency and robustness of medical imaging segmentation using inpainting-based self-supervised learning,” Bioengineering, vol. 10, no. 2, p. 207, 2023. DOI: 10.3390/bioengineering10020207.

[14] T. Viriyasaranon, S. M. Woo, and J.-H. Choi, ”Unsupervised visual representation learning based on segmentation of geometric pseudo-shapes for transformer-based medical tasks,” IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 4, pp. 2003–2013, Apr. 2023, doi: 10.1109/JBHI.2023.3237596.

[15] S. M. Hosseini, A. Shafique, M. Babaie, and H. R. Tizhoosh, ”Class-imbalanced unsupervised and semi-supervised domain adaptation for histopathology images,” in *2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Sydney, NSW, Australia, Jul. 2023, pp. 20–23. DOI: 10.1109/EMBC40787.2023.10340049.

[16] Z. Ren, X. Kong, Y. Zhang, and S. Wang, “UKSSL: Underlying Knowledge Based Semi-Supervised Learning for Medical Image Classification,” IEEE Open Journal of Engineering in Medicine and Biology, vol. 5, pp. 459–466, 2024, doi: 10.1109/OJEMB.2023.3305190.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[18] L. Wen, X. Li, and L. Gao, ”A transfer convolutional neural network for fault diagnosis based on ResNet-50,” *Neural Computing and Applications*, vol. 32, pp. 6111–6124, 2020. DOI: 10.1007/s00521-019-04097-w.

[19] Optical Coherence Tomography (OCT), University of California San Diego, Guangzhou Women and Children's Medical Center, 2018. [Online]. Available: https://data.mendeley.com/datasets/rscbjbr9sj/3

[20] D. S. Kermany et al., ”Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning,” *Cell*, vol. 172, no. 5, pp. 1122-1131, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.

[21] F. Grassmann et al., ”A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography,” Ophthalmology, vol. 125, no. 9, pp. 1410-1420, Sep. 2018, doi: 10.1016/j.ophtha.2018.02.037.

[22] P. Burlina, K. D. Pacheco, N. Joshi, D. E. Freund, and N. M. Bressler, ”Comparing Humans and Deep Learning Performance for Grading AMD: A Study in Using Universal Deep Features and Transfer Learning for Automated AMD Analysis,” Computers in Biology and Medicine, 2017. doi: 10.1016/j.compbiomed.2017.01.018.

[23] Y. Peng, S. Dharssi, Q. Chen, T. D. Keenan, E. Agrón, W. T. Wong, E. Y. Chew, and Z. Lu, ”DeepSeeNet: A Deep Learning Model for Automated Classification of Patient-based Age-related Macular Degeneration Severity from Color Fundus Photographs,” Ophthalmology, 2018. doi: 10.1016/j.ophtha.2018.11.015.

[24] M. Awais, H. Müller, T. B. Tang, and F. Meriaudeau, ”Classification of SD-OCT Images Using a Deep Learning Approach,” in Proc. IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Malaysia, Sep. 2017, pp. 489–492, doi: 10.1109/ICSIPA.2017.8120646.

[25] T. V. Phan, L. Seoud, H. Chakor, and F. Cheriet, ”Automatic Screening and Grading of Age-Related Macular Degeneration from Texture Analysis of Fundus Images,” Journal of Ophthalmology, vol. 2016, Article ID 5893601, 2016, doi: 10.1155/2016/5893601.