OLIVES Dataset: Ophthalmic Labels for Investigating Visual Eye Semantics

Mohit Prabhushankar[1]**, Kiran Kokilepersaud**[1]*****, Yash-yee Logan**[1]*****,
Stephanie Trejo Corona[2]*****, Ghassan AlRegib**[1]**, and Charles Wykoff**[2]
[1]OLIVES at the Centre for Signal and Info. Processing, Georgia Tech, Atlanta, GA 30332, USA
[2]Retina Consultants Texas, Retina Consultants of America, Houston, Texas 77030, USA
{mohit.p, kpk6, ylogan3, alregib}@gatech.edu,
{stephanie.trejo, ccwmd}@retinaconsultantstexas.com

Abstract

Clinical diagnosis of the eye is performed over multifarious data modalities including scalar clinical labels, vectorized biomarkers, two-dimensional fundus images, and three-dimensional Optical Coherence Tomography (OCT) scans. Clinical practitioners use all available data modalities for diagnosing and treating eye diseases like Diabetic Retinopathy (DR) or Diabetic Macular Edema (DME). Enabling usage of machine learning algorithms within the ophthalmic medical domain requires research into the relationships and interactions between all relevant data over a treatment period. Existing datasets are limited in that they neither provide data nor consider the explicit relationship modeling between the data modalities. In this paper, we introduce the Ophthalmic Labels for Investigating Visual Eye Semantics (OLIVES) dataset that addresses the above limitation. This is the first OCT and near-IR fundus dataset that includes clinical labels, biomarker labels, disease labels, and time-series patient treatment information from associated clinical trials. The dataset consists of 1268 near-IR fundus images each with at least 49 OCT scans, and 16 biomarkers, along with 4 clinical labels and a disease diagnosis of DR or DME. In total, there are 96 eyes' data averaged over a period of at least two years with each eye treated for an average of 66 weeks and 7 injections. We benchmark the utility of `OLIVES` dataset for ophthalmic data as well as provide benchmarks and concrete research directions for core and emerging machine learning paradigms within medical image analysis.

1   Introduction

Ophthalmology refers to the branch of medical science that deals with the structure, functions, diseases, and treatments of the eye. A stylized version of the diagnostic and treatment process for a known disease is shown in Fig. 1. A patient's visit to a clinic is met with an assessment that includes visual acuity tests and collecting demographic information. This provides Best Corrected Visual Acuity (BCVA) scores, Patient ID, and Eye ID among other data. We term these as *clinical labels*. Next, the patient undergoes diagnostic imaging that includes Fundus and OCT scans. Finally, a trained practitioner interprets the diagnostic scans for known *biomarkers* for diseases. The authors in ([1]) describe biomarkers as objective indicators of medically quantifiable characteristics of biological processes which are often diseases. The biomarkers along with the scans and clinical labels are used to assess the presence and severity of a patient's disease and a recommendation of a treatment is provided. If the recommendation is yes, the patient is treated and asked to visit again after a gap.
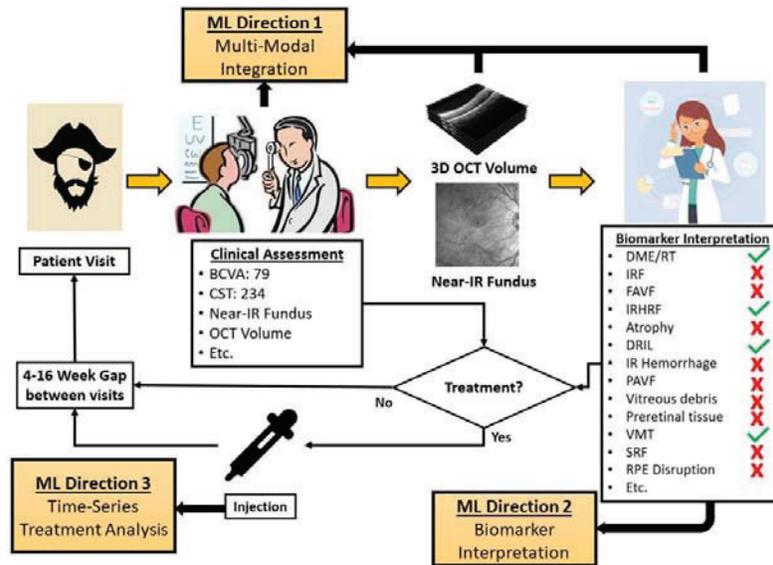
---

*Equal Contribution

Figure 1: Complete summary of data collection process for the OLIVES dataset and potential research directions for the machine learning community.

A number of Machine Learning (ML) techniques have sought to either automate or interpret individual processes within Fig. 1. We annotate three such ML research directions within the pipeline for clinically aiding and monitoring disease diagnosis and treatment. The first direction involves assessing multi-modal data for clinical applications including predicting disease states. The second direction is an interpretation of biomarkers. Biomarkers act as intermediary data between medical scans and disease diagnosis that aid clinical reasoning. The last direction is analyzing time-series treatment data across the treatment period. This direction aids initial treatment prescription and patient monitoring. To the best of our knowledge, no existing dataset provides access to data that promotes all three stated research directions for the clinical process from Fig. 1. In this paper, we introduce the Ophthalmic Labels for Investigating Visual Eye Semantics (OLIVES) dataset that provides structured and curated data to promote holistic clinical research in ML for ophthalmic diagnosis.

Clinical studies for OLIVES dataset The OLIVES dataset is derived from the PRIME (2) and TREX DME (3; 4; 5; 6) clinical studies. Both the studies are prospective randomized clinical trials that were run between December 2013 and April 2021 at the Retina Consultants of Texas (Houston, TX, USA). Prospective trials refer to studies that evaluate the outcome of a particular disease during treatment. PRIME evaluates Diabetic Retinopathy (DR) and TREX-DME evaluates Diabetic Macular Edema (DME). The trials provide access to near-IR fundus images and OCT scans along with de-identified Electronic Medical Records (EMR) data of 96 patients across an average of 66 weeks. Biomarkers are retrospectively added to this data by experienced graders upon open adjudication.

Challenging dataset for ML research While challenges in natural images are generally contrived by intervening on top of data (7; 8; 9), the complexities in ophthalmic datasets arise because of issues in data collection, inversion, representation and annotation. (10). OLIVES data modalities range from 1-dimensional numerical values (BCVA, Patient ID), vectorized biomarkers, 2-dimensional fundus images, and 3-dimensional scans (optical coherence tomography). Moreover, some of this data is objectively measured through instruments from patients (fundus, OCT), subjectively collected through eye tests (BCVA), while other data is interpreted and openly adjudicated through images (biomarkers). The variation within scans between visits can be minimal while the difference in manifestation of the same disease between patients may be substantial. This is shown in Fig. 2. The domain difference between OCT scans can arise due to pathology manifestation between patients (Fig. 2a and Fig. 2b), clinical labels (Fig. 2c), and the visit along the treatment process when the scan is taken (Fig. 2d). OLIVES provides access to these challenging data modalities that allow for innovative ML algorithms.
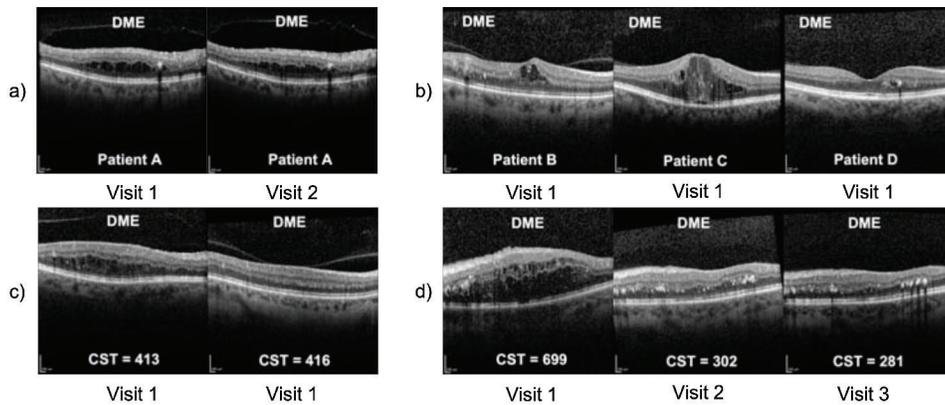
Figure 2: An illustration of some challenges within the dataset. In a) variation in OCT belonging to the same patient at different visits is minimal. In b) varied disease manifestations are among different patients. c) At times CST values are very similar but the OCT is visually dissimilar. d) Shows that CST values gradually decrease as time progresses.

Contributions and significance of the dataset The OLIVES dataset is curated to foster research in ophthalmic ML. The retrospective additions to the OLIVES dataset from its base clinical trials and its ensuing contributions include:

1. Sixteen biomarker labels are added to the OCT scans of every first and last visit of all patients. We experimentally validate the necessity of biomarkers and provide benchmarks in Sections 4.1 and 4.3. Along with biomarkers, OLIVES provides access to fundus, OCT scans, clinical labels and DR/DME diagnosis, thereby creating an ideal benchmarking mechanism for ophthalmic ML.

2. We curate the clinical labels that have known correlations between the four data modalities. These include Best Corrected Visual Acuity (BCVA), Central Subfield Thickness (CST), Patient ID, and Eye ID. We demonstrate its utility for medically-grounded contrastive learning where augmentations are based on contrasting between clinical labels in Section 4.2. Hence, OLIVES dataset promotes research in core and emerging ML paradigms.

3. The data and labels are made accessible to non-medical professionals. Biomarkers act as expert-annotated and interpretable visual indicators of diseases within OCT scans. The original labels from the clinical trials along with their data sheets are provided in Appendix B.5.2. Additionally, an ML-specific set of labels which is relevant to the three mentioned research directions in Fig. 1, is provided in Appendix D.3.

2   Related Works

Ophthalmology datasets   A number of publicly available ophthalmology datasets individually tackle each of the clinical modalities that exist in the OLIVES dataset. The authors in (11) provide a survey of 94 existing open access ophthalmic datasets. Among 54 of the 94 datasets, the underlying data is that of fundus images. 19 of the remaining datasets contain 3-dimensional OCT scans. The OCT scans provide structural information that enhances the performance of machine learning algorithms (11). Only three of the 94 considered open access datasets provide both OCT and fundus image modalities. The authors in (1) provide 650 OCT slices from a single volume. These are insufficient to leverage the data intensive machine learning algorithms to provide generalizable results. In contrast, the OLIVES dataset has 78, 189 slices taken from 1268 volumes. (12) provide OCT and fundus data from 50 healthy patients. However, these are all for healthy eyes and disease manifestation is not observed. Other datasets including (13) contains OCT scans for four OCT disease states: Healthy, Drusen, DME, and choroidal neovascularization (CNV). (14) and (15) introduced OCT datasets for age-related macular degeneration (AMD). (16) contains OCT scans labeled with segmentation of regions with DME. However, these datasets do not possess comprehensive clinical information or a wide range of expert-annotated biomarkers. A complete overview that considers clinical labels, biomarkers, disease labeling, and time-series analysis is provided in Tables 5 and 6.

We refer to ([11]) to compare other statistics including number of image scans and applicability of existing datasets against `OLIVES`.

Machine learning techniques on ophthalmic data   A number of works have separately addressed the research directions identified in Fig. [1]. The authors in ([17]) proposed transfer learning to screen for relative afferent pupillary defect due to lack of comprehensive data. ([18]) showed that transfer learning methods could be utilized to classify OCT scans based on the presence of key biomarkers. ([19]) introduced a dual-autoencoder framework with physician attributes to improve classification performance for OCT biomarkers. ([20]) expanded previous work towards segmentation of a multitude of different biomarkers and referred for different treatment decisions. Other work has demonstrated the ability to detect clinical information from OCT scans which is significant for suggesting correlations between different domains. ([21]) showed that a model trained entirely on OCT scans could predict the associated BCVA value. Similarly ([22]) showed that values such as retinal thickness could be learned from retinal fundus photos. The `OLIVES` dataset provides a standardized benchmark to conduct research across applications, data modalities and machine learning paradigms.

Table 1: High-level overview of the `OLIVES` Dataset. The modality column details the type of data. The columns "Per Visit" and "Per Eye" indicate the amount of data in each modality on a respective visit or eye. N_P is the number of visits that a patient P takes to the clinic. The statistics across all eyes across all visits are shown in the Total Statistics column. Biomarkers are binary values, clinical labels are integers, fundus are 2D images, and OCT are 3D slices.

OLIVES Dataset Summary				
Modality	Per Visit	Per Eye	Total Statistics	Overview
OCT	49	N_P*49	78189	**General:** 96 Eyes, Visits every 4-16 weeks, Average 16 visits and 7 injections/patient
Fundus	1	N_P	1268	**Clinical Labels obtained every visit:** BCVA, CST, Patient ID, Eye ID
Clinical	4	N_P*4	5072	**Biomarkers labeled:** IRHRF, FAVF, IRF, DRT/ME PAVF, VD,
Biomarker	16	1568	150528	Preretinal Tissue, EZ Disruption, IR Hemmorhages, SRF, VMT, Atrophy, SHRM, RPE Disruption, Serous PED

3   OLIVES Dataset

Statistics regarding the quantity of images and labels can be found in Table [1]. The `OLIVES` dataset is derived from the PRIME and TREX-DME trials. At every visit for each patient, ocular disease state data (DR/DME), clinical labels including BCVA, CST, Patient and Eye ID, and detailed ocular imaging including OCT, and fundus photography were obtained per the protocol in Section [B.5.2]. This procurement of data continues across N_P visits for every patient, where N_P is the number of visits by a patient P. For instance, 3D longitudinal scans of the eye provide 49 OCT scans per patient per visit. Across N_P visits where P can be any one of 96 patients, the total number of OCT scans in the dataset is 78, 189. Note that on every visit, each patient undergoes testing to determine the requirement of a treatment per the clinical protocol described in Appendix [D.1]. Biomarkers are retrospectively added to each slice in the OCT scans for the first and last visits. Table [1] also indicates the total number of eyes, average number of visits and injections, and the time between visits.

3.1   Biomarker Generation

After the clinical data collection process, we retrospectively provide additional insight into the OCT scans by providing corresponding biomarker labels. Biomarkers are quantifiable characteristics of biological processes in the eye. In this paper, the biological processes are diseases and biomarkers indicate the presence or absence of such diseases. Under limited circumstances, the authors in ([1]) suggest that biomarkers can be surrogate endpoints in clinical trials. However, they caution against doing so unless the underlying clinical trial is specifically meant for the study. In both the PRIME and TREX DME studies, biomarkers are retrospectively labeled. As such, biomarkers may indicate the presence of diseases, but are not causal to these diseases. Hence, biomarkers are different from visual causal features from ([23]) or causal question-based analysis in ([24]) or causal factor analysis in ([25]).
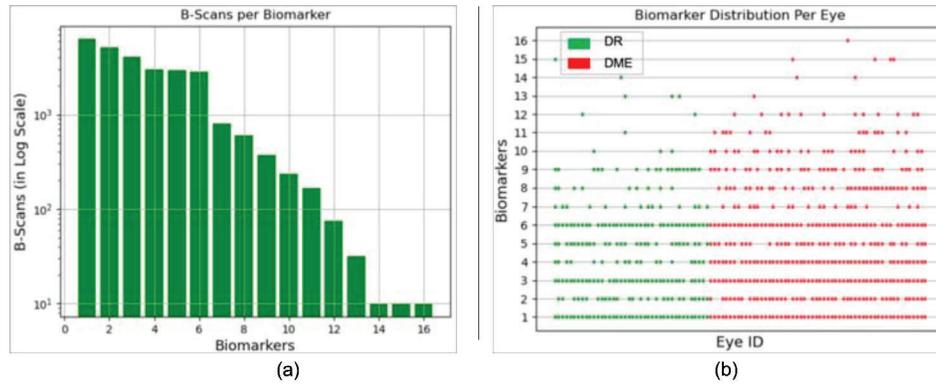
Figure 3: (a) Histogram of the number of scans per biomarker. (b) Unique biomarkers per Eye ID.
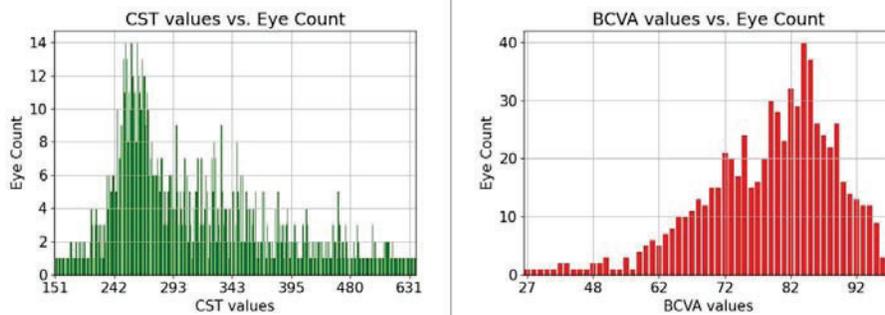

Figure 4: Distribution of CST and BCVA labels in `OLIVES` dataset based on number of eyes associated with each clinical value.

In the PRIME and TREX DME studies, images, clinical information, and biomarker labels were retrospectively collected at the Retina Consultants of Texas (Houston, TX, USA). This study was approved by the Institutional Review Board (IRB)/Ethics Committee and adheres to the tenets of the Declaration of Helsinki and Health Insurance Portability and Accountability Act (HIPAA). Informed consent was not required due to the retrospective nature of the study. A trained grader performed interpretation on OCT scans for the presence of 16 different biomarkers including: intraretinal hyperreflective foci (IRHRF), partially attached vitreous face (PAVF), fully attached vitreous face (FAVF), intraretinal fluid (IRF), and diffuse retinal thickening or macular edema (DRT/ME). A full list of the biomarkers as well as their characteristics is provided in Section B.5.1. The full form of the abbreviations are given in Table 7. These biomarkers are chosen because of their visual attributes that correlate with presence or absence of disease states. The trained grader was blinded to clinical information whilst grading each of 49 horizontal OCT B-scans of both the first and last study visit for each individual eye. Open adjudication was done with an experienced retina specialist for difficult cases. In total, there are 9408 OCT scans that consist of a 16×1 biomarker vector where 1 indicates the presence of the corresponding biomarker and a 0 indicates its absence. We provide a histogram of the number of scans (`y-axis`) against their respective biomarkers in Fig. 3a. Note that the `y-axis` is in log-scale. We also depict the eye ID against the biomarkers in Fig. 3b. The green dots are eyes that indicate the presence of the corresponding biomarker on the `y-axis` that are diagnosed with DR. The red dots are for DME. It can be seen that a number of eyes have overlapping biomarkers even between diseases. Hence, biomarkers in isolation are insufficient to diagnose disease states, strengthening the case for multi-modal data.

3.2 Clinical Labels

Within the `OLIVES` dataset, we have explicit clinical information regarding the Best Central Visual Acuity (BCVA), Central Subfield Thickness (CST), and identity of the eye. ETDRS best-corrected visual acuity (BCVA) is a visual function assessment performed by certified examiners where a standard vision chart is placed 4-meters away from the patient. The patient is instructed to read

the chart from left to right from top to bottom until the subject completes 6 rows of letters or the subject is unable to read any more letters. The examiner marks how many letters were correctly identified by the patient. Central subfield thickness (CST) is the average macular thickness in the central 1-mm radius of the ETDRS grid. Both BCVA and CST are coarse measurements over the eye as opposed to Biomarkers that exist for fine-grained longitudinal slices of the eye. BCVA can range from $0 - 100$ and CST from $100 - 1300$. We show in Fig. 4 the number of eyes (y-axis) that have the associated value (x-axis) for both BCVA and CST. This graph shows that our dataset has a wide variation in terms of range of clinical values across a multitude of eyes in the dataset. This is advantageous as it shows the dataset is not biased to any specific range of values or localized to single eye instances. A full list of all clinical labels present in PRIME and TREX-DME clinical trials are shown in Section B.5.2. No personally identifiable information was included in compliance with HIPAA regulations.
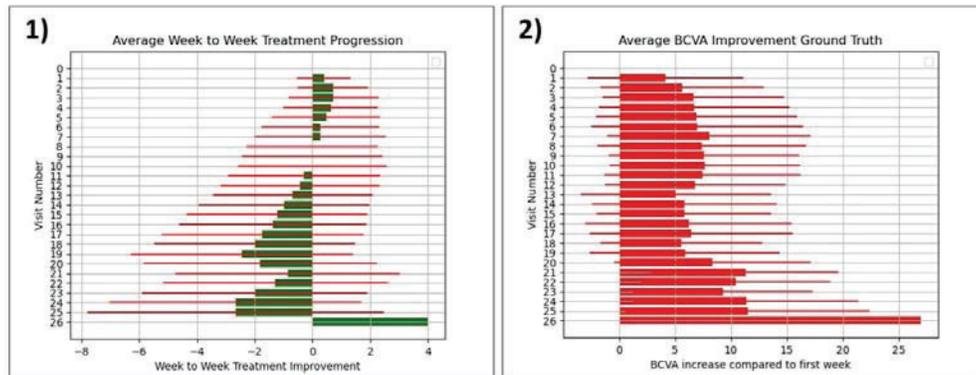

Figure 5: 1) A plot of average number of visits by patients that were an improvement or deterioration from previous week. Red bars indicate the standard deviation across all patients. 2) Plot of average change in BCVA with respect to the first week.

3.3 Time-series data

A core novelty of the dataset is that data exists for each patient visit across a defined period of time. As a result, it is possible to analyze trends in the collected imaging and clinical data over the visiting period of the patient. This is shown in Fig. 5 with an overall progression analysis shown by the bar graphs. Graph 1 indicates whether on average there was an improvement from the previous visit. This was computed by assigning a value of 1 for improvements and -1 for deterioration. This is accumulated every visit and the average across all patients is calculated on a per visit basis. From this plot, it can be observed that in the dataset, eyes generally improve on every visit until about the tenth visit. However, graph 2 n Fig. 5 shows that while visit to visit improvement declines, the overall improvement when compared with the first visit is generally substantial. This graph was computed by taking the difference between the current visit's BCVA and first visit's BCVA and averaging across all patients. The statistics of the number of patients every visit and visualizations of patient treatment is shown in Figs. 8 and 9 respectively.

3.4 Interaction between Data Modalities

Clinical labels correspond to measurements that pertain to the entire visual system, including the visual mechanism in the eye. These measurements give an overview of the health of the eye, but they do not enable fine-grained analysis of structures that exist within the eye. Biomarker labels exist at the longitudinal slice level. They are detailed labels for every slice of the eye and provide a fine-grained analysis of the biological structures that exist within the eye. Clinical studies such as (26) and (27) suggest that measured clinical labels can act as indicators of structural changes that manifest themselves in OCT scans and fundus images as well as the severity of disease associated with the patient. For example, visually, it can be observed that OCT scans with the same BCVA values exhibit more common structural characteristics than scans with different BCVA values. Furthermore, all data modalities exhibit visual, structural and clinical changes across the treatment period. OLIVES dataset

allows for exploiting these correlations between OCT, fundus, clinical labels, biomarkers, diseases, and treatment states.

4 Clinical Applications

With the multitude of modalities that exist within the `OLIVES` dataset, there is potential for research in a wide variety of ML applications. Within this section, we focus on applications, and benchmarks, that showcase key features of the dataset identified from Fig. 1, but acknowledge that other novel setups and formulations of the problem are possible and intended. These applications include multi-modal integration of OCT scans and biomarker/clinical labels, biomarker detection and interpretation using contrastive learning, and time-series treatment analysis.

4.1 Multi-Modal Integration Between OCT and Biomarkers/Clinical Labels

Table 2: Benchmark results for DR/DME detection.

Experiments	Model	Balanced Accuracy	Specificity	Sensitivity
OCT	R-18	70.15% ± 4.69	0.608	0.794
Clinical	MLP	75.49% ± 1.98	0.758	0.751
Biomarker	MLP	79.87% ± 3.03	0.826	0.771
OCT + Clinical	R-18 + MLP	75.92% ± 3.05	0.566	0.952
OCT + Biomarker	R-18 + MLP	82.33% ± 3.59	0.742	0.904

Baseline Detection of DR/DME with OCT   Since biomarkers are only available for the first and last clinical visits, we use the corresponding OCT at those visits for this baseline analysis. The entire dataset is partitioned by eyes into train, test and validation splits. Additional details about train/test/validation splits is in Appendix C.1. We evaluate performance with balanced accuracy, precision and recall performance metrics. The results for the baseline OCT model is shown in the first row of Table 2. Additional results showing specificity and sensitivity are in Table 9 in the Appendix. This and subsequent experiments are conducted using multiple random seeds for DR/DME detection and an average score and standard deviation is reported for balanced accuracy.

Supervised Learning with Clinical Labels   We aim to use clinical labels as an additional modality to aid the baseline model. However, to determine the suitability of this auxiliary data type, we first evaluate its impact on the classification of DR and DME. To do this we first find all unique clinical labels present in the dataset with their associated disease labels. Then, we create a training set with 70% of these clinical labels along with test and validation sets of 20% and 10% proportions respectively. This yields 1107 unique clinical labels for training, 306 for testing and 122 for validation. Within the test set, half the samples are DR and the remaining DME. The second row on Table 2 shows that CST and BCVA used as clinical features are more effective than the unimodal OCT baseline for DR/DME detection.

Supervised Learning with Biomarkers   We perform a similar analysis as described in supervised learning with clinical labels but using biomarkers as features. Hence, we substitute the clinical labels with biomarkers to characterize the diseases. There are 286 unique biomarker label features among which 200, 58, 28 samples are used for train, test and validation sets respectively. From the third row in Table 2, we observe that using biomarkers on their own leads to a 9.72% increase in DR and DME classification over baseline results.

Multi-Modal Learning with OCT and Clinical Labels   Having seen that clinical labels are more effective than the baseline model at DR/DME classification, we now investigate how to use the clinical label modality to aid the OCT model. Clinical labels and OCT are independently given as input to their models as described previously. We optimize both models jointly with a loss function that allows knowledge, in the form of logits, from the clinical model to guide the optimization of the OCT model. A detailed description of this optimization scheme can be seen in Appendix C.1. During testing, only the OCT model, having been optimized jointly with the other model, is used to classify the disease states. The fourth row of Table 2 shows that clinical labels also aid the OCT model at characterizing the diseases albeit not the most effectively.

Table 3: Benchmark of the performance of supervised contrastive training on images with clinical and biomarker data. The standard deviations are shown in Table 10.

| Method	Biomarkers										Metrics		
	IRF		DRT/ME		IRHRF		FAVF		PAVF		AUROC	Average Specificity	Average Sensitivity
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score			
PCL [28]	**76.50%**	0.717	80.11%	0.761	59.10%	0.683	76.30%	0.773	51.40%	0.165	0.767	0.741	0.604
SimCLR [29]	75.13%	0.716	80.61%	0.772	59.03%	0.675	75.43%	0.761	52.69%	0.249	0.754	0.747	0.614
Moco V2 [30]	76.00%	**0.720**	82.24%	0.793	59.60%	0.692	75.00%	0.784	52.69%	0.211	0.770	0.762	0.651
Eye ID	72.63%	0.674	80.20%	0.778	58.00%	0.674	74.93%	0.725	**65.56%**	**0.588**	0.767	**0.776**	0.656
CST	75.53%	0.720	**83.06%**	**0.811**	**64.30%**	**0.703**	76.13%	0.766	62.16%	0.509	**0.790**	0.772	**0.675**
BCVA	74.03%	0.701	80.27%	0.770	58.8%	0.672	**77.63%**	**0.785**	58.06%	0.418	0.776	0.713	0.645
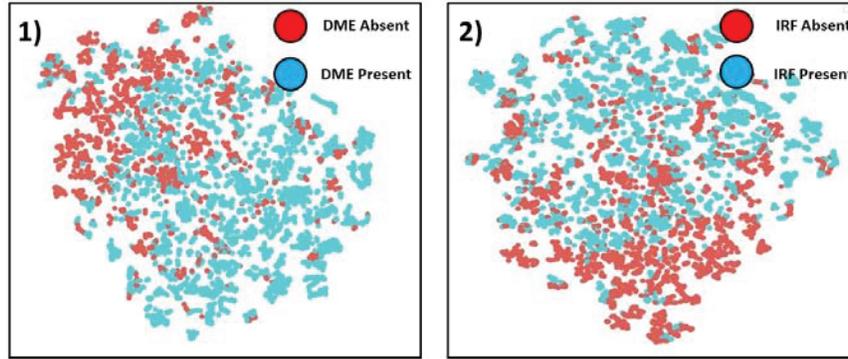
Figure 6: T-SNE visualization of the OLIVES Biomarker test set labeled by the presence or absence of DME and IRF. We can effectively achieve an embedding space that is separable with respect to biomarkers 1) DME and 2) IRF.

Multi-Modal Learning with OCT and Biomarkers   In like manner, we investigate the impact that biomarkers as features can have on aiding an OCT model to classify the two diseases. Each model is fed their input modality and optimized jointly using the same loss detailed in Appendix C.1. During testing, only the OCT model is used for inference. The final row in Table 2 shows that this is the most effective technique that significantly improves all baseline classification metrics.

4.2   Biomarker Interpretation with Contrastive Learning

Due to the prohibitive costs of expert-annotated biomarker labels, contrastive learning (31; 29; 32) approaches have garnered attention because of their state of the art self-supervised performance. These approaches generally create a representation space through minimizing the distance between positive pairs of images and maximizing the distance between negative pairs. Traditional approaches, like SimCLR (29), generate positives from augmentations of a single image and treat all other images in a batch as negatives. More modern approaches like Moco v2 (30) incorporate a queue system for additional negative samples while extensions of this include PCL (28) that introduce a clustering approach on the representation space. While these approaches have shown promising results on natural images, such augmentations are unrealistic for medical images that rely on fine-grained changes within OCT scans to detect diseases. Instead, we propose using clinically relevant labels as a means to better choose positive pairs. Since OLIVES provides a larger pool of clinical labels than biomarker labels, this task fits well within the scope of the dataset. Hence, OLIVES enables research into novel and multi-modal contrastive learning strategies. We implement one such strategy through reformulating the supervised contrastive loss (33) in a clinical context as discussed in (34) with related work located at (35; 36). Implementation details are provided in Section C.2.

We train a Resnet-18 (37) encoder with the clinically labeled data using the clinically aware supervised contrastive loss. After training the encoder with supervised contrastive loss, we freeze the weights of the encoder and append a linear layer to its output. This linear layer is trained using cross-entropy loss to distinguish between the presence or absence of the biomarker of interest in the OCT scan. Not all of these biomarkers exist in sufficiently balanced quantities to train a model to identify their presence or absence within an image. Hence, we use five biomarkers that fit this criteria in this study. The training details are presented in Section C.2. We compare this method with three state of the art self-supervised algorithms in Table 3. We evaluate performance in terms of individual biomarker accuracy and f1-score as well as in the setting where the goal is to simultaneously perform a multi-

label classification of biomarkers. Performance is measured by average AUROC, specificity, and sensitivity. We observe that a training strategy that chooses positives based on the clinical data Eye ID, BCVA, and CST outperforms baseline self-supervised methods in both a multi-label classification task as well as individual biomarker detection performance. While the results in Section 4.1 make use of correlations between the biomarkers and clinical labels with disease states, these results depict the correlations between the label modalities.

In Figure 6, we visualize the test set t-SNE embeddings of two different biomarkers from a model trained using the BCVA clinical label. We observe that even without any fine-tuning on the actual biomarker label of interest, we are able to get an embedding space where the absence and presence of DME and IRF form distinct clusters. This gives credence to the idea that there exists relationships between the biomarker and clinical label domains as training on only clinical labels leads to a separable space within the biomarker domain.

4.3   Time-Series Treatment Analysis

The multi-modal nature of `OLIVES` dataset allows for a large combination of experimental setups to analyze treatments. We present two experimental manifestations based off the temporal nature of the data: a) Predicting visit-by-visit successive treatment effects and b) Predicting the final ocular state using Biomarkers. A key metric used to evaluate treatment progression or regression is BCVA. At each visit to the clinic, patients' ocular disease states are evaluated and BCVA and other clinical labels are recorded. From a machine learning perspective, this motivates an analysis of treatment effect over consecutive weeks to predict how BCVA scores will change based on the state of the eye captured via OCT or Fundus. We detail the exact experimental procedure in Appendix C.3. We evaluate the performance of this strategy on both fundus images and 3D OCT volumes. We use a Resnet-18 (37), ResNet-50 (37), DenseNet-121 (38), EfficientNet (39), and Vision Transformer (40) (using a patch size of 32, 16 transformer blocks, 16 heads in multi-attention layer). For the OCT volumes, we utilize a version of each architecture that uses three-dimensional convolution layers. Performance in both modalities is reported in Table 4. We observe that the model is able to learn distinguishing features between the two classes, with better performance when using the OCT volumetric data. Additionally, we present results for predicting the final state of 16×1 biomarker vector given the initial biomarker vector for individual patients in Fig. 10. Similar to the week-wise case, these results indicate correlation among multiple modalities as well as the ability of ML algorithms to predict ocular states given treatment.

Table 4: Benchmark Performance of predicting treatment effects from time-series Fundus and OCT data.

Model	Image Modality	Accuracy	Precision	Recall
ResNet-18	Fundus	$55.19\% \pm 10.9$	0.256	0.343
	OCT Volume	$57.59\% \pm 9.51$	0.359	0.326
ResNet-50	Fundus	$48.73\% \pm 13.3$	0.372	0.3296
	OCT Volume	$57.70\% \pm 9.1$	0.301	0.1826
DenseNet-121	Fundus	$53.00\% \pm 8.9$	0.273	0.259
	OCT Volume	$54.75\% \pm 4.92$	0.219	0.188
EfficientNet	Fundus	$56.06\% \pm 4.85$	0.292	0.217
	OCT Volume	$60.65\% \pm 4.09$	0.3613	0.1633
ViT	Fundus	$55.01\% \pm 3.27$	0.285	0.350

5   Discussion and Conclusion

Domain Difference and Adaptation in Multi-Modal Data   The data in `OLIVES` is derived from two studies. As mentioned in Section 1, the domain difference in ophthalmic data can arise from sources such as treatment, disease manifestation, and clinical labels. In natural images, one source of domain difference is the equipment used for imaging. In PRIME and TREX studies, the same imaging and grading modalities, the Heidelberg Spectralis HRA+OCT software, is used in the same clinic. We provide extensive experiments in Appendix C.5 and C.6 to characterize possible domain differences on OLIVES. In Table 11, we show that the biomarker detection results when trained and tested on PRIME trial is lower than when trained with TREX and tested on PRIME. This is because a longer treatment period on TREX dataset provides more diverse data that is conducive for

https://doi.org/10.52202/068431-0669

training ML algorithms. Intuitively, this suggests that treatment causes domain shift in data, which is illustrated in Table 12. Training and testing within the first week data provides the best results for biomarker detection. This analysis is further expanded in Fig. 11. Rather than showing domain difference, we adapt between the first and last visit domains. Specifically, we use a part of the last visit data to train with the first visit data and show that: a) adapting between OCT scans before and after treatment is possible, and b) the addition of biomarkers increases the results for diagnosis of DR/DME. Hence, OLIVES provides data modalities that promotes research in treatment-based domain difference and adaptation in medical data.

Dataset Limitations, Societal Impact, and Ethical Concerns   The OLIVES dataset is derived from two clinical studies conducted from only one U.S. clinic. While there is a range in the age, ethnicity and racial demographics within the cohorts, this range is only limited to one geographical location. Hence, an end-to-end system can be biased. To mitigate this limitation, we provide links to existing open access ophthalmic datasets in Appendix B.4 that are collected from other parts of the world. While none of these datasets are as rich as our own in terms of numbers, modalities, or labels, they can be used to modularly test algorithms. We present one such result in Table 13 and show that combining datasets allows for higher results. The PRIME and TREX trials are randomized clinical studies with the goal of comparing different treatment regimens. These studies aim to find the best practices for how and when they should treat patients to get the most optimal outcomes. However, there are no control groups within the studies that did not receive treatment. While this is common in clinical trials (6), it adds a new challenge to ML-focused research of time-series analysis. We list datasets that provide healthy images in Appendix B.4 to complement OLIVES. We believe that a combination of datasets taken over multiple geographical regions, times, and disease states is essential to construct generalizable and ethical ML models. ML models can potentially amplify existing inequalities within healthcare access (41). For instance, the data in OLIVES is collected from December 2013 to April 2021, which implies the participants had the time and means to be part of these trials. This may not always be the case for disadvantaged groups. Hence, any benefit that machine learning could provide will be restricted to small subsets of society unless thought is put into preventing this disparity. Hence, a careful analysis of potential concerns is required to use OLIVES and any other dataset to enrich the functionality and adaptability of machine learning algorithms in everyday lives.

Conclusion   We introduce the OLIVES dataset to bridge the gap between existing ophthalmic datasets and the clinical diagnosis and treatment process. OLIVES provides curated and contained data that can be used for clinical interpretation of biomarkers, clinical reasoning regarding disease prediction, multi-modal integration of ophthalmic data and treatment monitoring through time-series analysis. Also, we propose and benchmark medically-grounded contrastive learning strategies that are possible because of the presence of correlated multi-modal data within the introduced dataset. The OLIVES dataset opens new frontiers for training holistic and medically-relevant ML frameworks that mimic the clinical diagnosis pipeline for ophthalmic studies.

References

[1] Marzieh Golabbakhsh and Hossein Rabbani, "Vessel-based registration of fundus and optical coherence tomography projection images of retina using a quadratic registration model," *IET Image Processing*, vol. 7, no. 8, pp. 768–776, 2013.

[2] J Yu Hannah, Justis P Ehlers, Duriye Damla Sevgi, Jenna Hach, Margaret O'Connell, Jamie L Reese, Sunil K Srivastava, and Charles C Wykoff, "Real-time photographic-and fluorescein angiographic-guided management of diabetic retinopathy: Randomized prime trial outcomes," *American Journal of Ophthalmology*, vol. 226, pp. 126–136, 2021.

[3] John F Payne, Charles C Wykoff, W Lloyd Clark, Beau B Bruce, David S Boyer, David M Brown, TREX-DME study group, et al., "Randomized trial of treat and extend ranibizumab with and without navigated laser for diabetic macular edema: Trex-dme 1 year outcomes," *Ophthalmology*, vol. 124, no. 1, pp. 74–81, 2017.

[4] John F Payne, Charles C Wykoff, W Lloyd Clark, Beau B Bruce, David S Boyer, David M Brown, John A Wells III, David L Johnson, Matthew Benz, Eric Chen, et al., "Randomized trial of treat and extend ranibizumab with and without navigated laser versus monthly dosing for

diabetic macular edema: Trex-dme 2-year outcomes," *American journal of ophthalmology*, vol. 202, pp. 91–99, 2019.

[5] John F Payne, Charles C Wykoff, W Lloyd Clark, Beau B Bruce, David S Boyer, and David M Brown, "Long-term outcomes of treat-and-extend ranibizumab with and without navigated laser for diabetic macular oedema: Trex-dme 3-year results," *British Journal of Ophthalmology*, vol. 105, no. 2, pp. 253–257, 2021.

[6] Charles C Wykoff, Muneeswar G Nittala, Brenda Zhou, Wenying Fan, Swetha Bindu Velaga, Shaun IR Lampen, Alexander M Rusakevich, Justis P Ehlers, Amy Babiuch, David M Brown, et al., "Intravitreal aflibercept for retinal nonperfusion in proliferative diabetic retinopathy: outcomes from the randomized recovery trial," *Ophthalmology Retina*, vol. 3, no. 12, pp. 1076–1086, 2019.

[7] Dogancan Temel, Gukyeong Kwon, Mohit Prabhushankar, and Ghassan AlRegib, "Cure-tsr: Challenging unreal and real environments for traffic sign recognition," *arXiv preprint arXiv:1712.02463*, 2017.

[8] Dogancan Temel, Jinsol Lee, and Ghassan AlRegib, "Cure-or: Challenging unreal and real environments for object recognition," in *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2018, pp. 137–144.

[9] Min-Hung Chen, Baopu Li, Yingze Bao, and Ghassan AlRegib, "Action segmentation with mixed temporal domain adaptation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 605–614.

[10] Ching-Yu Cheng, Zhi Da Soh, Shivani Majithia, Sahil Thakur, Tyler Hyungtaek Rim, Yih Chung Tham, and Tien Yin Wong, "Big data in ophthalmology," *The Asia-Pacific Journal of Ophthalmology*, vol. 9, no. 4, pp. 291–298, 2020.

[11] Saad M Khan, Xiaoxuan Liu, Siddharth Nath, Edward Korot, Livia Faes, Siegfried K Wagner, Pearse A Keane, Neil J Sebire, Matthew J Burton, and Alastair K Denniston, "A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability," *The Lancet Digital Health*, vol. 3, no. 1, pp. e51–e66, 2021.

[12] Tahereh Mahmudi, Rahele Kafieh, Hossein Rabbani, Mohammadreza Akhlagi, et al., "Comparison of macular octs in right and left eyes of normal people," in *Medical Imaging 2014: Biomedical Applications in Molecular, Structural, and Functional Imaging*. SPIE, 2014, vol. 9038, pp. 472–477.

[13] Daniel Kermany, Kang Zhang, Michael Goldbaum, et al., "Labeled optical coherence tomography (oct) and chest x-ray images for classification," *Mendeley data*, vol. 2, no. 2, 2018.

[14] Sina Farsiu, Stephanie J Chiu, Rachelle V O'Connell, Francisco A Folgar, Eric Yuan, Joseph A Izatt, Cynthia A Toth, Age-Related Eye Disease Study 2 Ancillary Spectral Domain Optical Coherence Tomography Study Group, et al., "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography," *Ophthalmology*, vol. 121, no. 1, pp. 162–172, 2014.

[15] Martina Melinščak, Marin Radmilović, Zoran Vatavuk, and Sven Lončarić, "Annotated retinal optical coherence tomography images (aroi) database for joint retinal layer and fluid segmentation," *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, vol. 62, no. 3-4, pp. 375–385, 2021.

[16] Stephanie J Chiu, Michael J Allingham, Priyatham S Mettu, Scott W Cousins, Joseph A Izatt, and Sina Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," *Biomedical optics express*, vol. 6, no. 4, pp. 1172–1194, 2015.

[17] Dogancan Temel, Melvin J Mathew, Ghassan AlRegib, and Yousuf M Khalifa, "Relative afferent pupillary defect screening through transfer learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 3, pp. 788–795, 2019.

[18] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[19] Yash Logan, Kiran Kokilepersaud, Gukyeong Kwon, Ghassan AlRegib, Charles Wykoff, and Hannah Yu, "Multi-modal learning using physicians diagnostics for optical coherence tomography classification," *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2022.

[20] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[21] Michael G Kawczynski, Thomas Bengtsson, Jian Dai, J Jill Hopkins, Simon S Gao, and Jeffrey R Willis, "Development of deep learning models to predict best-corrected visual acuity from optical coherence tomography," *Translational vision science & technology*, vol. 9, no. 2, pp. 51–51, 2020.

[22] Filippo Arcadu, Fethallah Benmansour, Andreas Maunz, John Michon, Zdenka Haskova, Dana McClintock, Anthony P Adamis, Jeffrey R Willis, and Marco Prunotto, "Deep learning predicts oct measures of diabetic macular thickening from color fundus photographs," *Investigative ophthalmology & visual science*, vol. 60, no. 4, pp. 852–857, 2019.

[23] Mohit Prabhushankar and Ghassan AlRegib, "Extracting causal visual features for limited label classification," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3697–3701.

[24] Ghassan AlRegib and Mohit Prabhushankar, "Explanatory paradigms in neural networks," *arXiv preprint arXiv:2202.11838*, 2022.

[25] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt, "Visual causal feature learning," *arXiv preprint arXiv:1412.2309*, 2014.

[26] Rosana Zacarias Hannouche, Marcos Pereira de Ávila, David Leonardo Cruvinel Isaac, Alan Ricardo Rassi, et al., "Correlation between central subfield thickness, visual acuity and structural changes in diabetic macular edema," *Arquivos brasileiros de oftalmologia*, vol. 75, no. 3, pp. 183–187, 2012.

[27] Jennifer K Sun, Michael M Lin, Jan Lammer, Sonja Prager, Rutuparna Sarangi, Paolo S Silva, and Lloyd Paul Aiello, "Disorganization of the retinal inner layers as a predictor of visual acuity in eyes with center-involved diabetic macular edema," *JAMA ophthalmology*, vol. 132, no. 11, pp. 1309–1316, 2014.

[28] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi, "Prototypical contrastive learning of unsupervised representations," *arXiv preprint arXiv:2005.04966*, 2020.

[29] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[30] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[31] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, 2020.

[32] Mohit Prabhushankar and Ghassan AlRegib, "Contrastive reasoning in neural networks," *arXiv preprint arXiv:2103.12329*, 2021.

[33] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*, 2020.

[34] Kiran P Kokilepersaud, Stephanie Trejo Corona, Mohit Prabhushankar, Ghassan Alregib, and Charles Wykoff, "Supervised contrastive learning on clinical labels for biomarker classification in oct," *Journal of Biomedical And Health Informatics*, Under Review.

[35] Kiran P Kokilepersaud, Stephanie Trejo Corona, Mohit Prabhushankar, Ghassan Alregib, and Charles Wykoff, "Gradient-based severity labeling for biomarker classification in oct," *IEEE International Conference in Image Processing*, 2022.

[36] Kiran Kokilepersaud, Mohit Prabhushankar, and Ghassan AlRegib, "Volumetric supervised contrastive learning for seismic semantic segmentation," in *Second International Meeting for Applied Geoscience & Energy*. Society of Exploration Geophysicists and American Association of Petroleum . . . , 2022, pp. 1699–1703.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[38] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[39] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[40] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020," *arXiv preprint arXiv:2010.11929*, 2010.

[41] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi, "Ethical machine learning in healthcare," *Annual review of biomedical data science*, vol. 4, pp. 123–144, 2021.

[42] Pratul P Srinivasan, Leo A Kim, Priyatham S Mettu, Scott W Cousins, Grant M Comer, Joseph A Izatt, and Sina Farsiu, "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," *Biomedical optics express*, vol. 5, no. 10, pp. 3568–3577, 2014.

[43] Stefan Maetschke, Bhavna Antony, Hiroshi Ishikawa, Gadi Wollstein, Joel Schuman, and Rahil Garnavi, "A feature agnostic approach for glaucoma detection in oct volumes," *PloS one*, vol. 14, no. 7, pp. e0219126, 2019.

[44] California Healthcare Foundation, "Diabetic retinopathy detection identify signs of diabetic retinopathy in eye images," https://www.kaggle.com/competitions/diabetic-retinopathy-detection/overview, 2015, Accessed: 2022-06-08.

[45] Liu Li, Mai Xu, Xiaofei Wang, Lai Jiang, and Hanruo Liu, "Attention based glaucoma detection: a large-scale database and cnn model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10571–10580.

[46] Ning Li, Tao Li, Chunyu Hu, Kai Wang, and Hong Kang, "A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection," in *International Symposium on Benchmarking, Measuring and Optimization*. Springer, 2020, pp. 177–193.

[47] Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, et al., "Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge," *Patterns*, p. 100512, 2022.

[48] Chi Liu, Xiaotong Han, Zhixi Li, Jason Ha, Guankai Peng, Wei Meng, and Mingguang He, "A self-adaptive deep learning method for automated eye laterality detection based on color fundus photography," *Plos one*, vol. 14, no. 9, pp. e0222025, 2019.

[49] Michael D Abràmoff, James C Folk, Dennis P Han, Jonathan D Walker, David F Williams, Stephen R Russell, Pascale Massin, Beatrice Cochener, Philippe Gain, Li Tang, et al., "Automated analysis of retinal images for detection of referable diabetic retinopathy," *JAMA ophthalmology*, vol. 131, no. 3, pp. 351–357, 2013.

[50] Kedir M Adal, Peter G van Etten, Jose P Martinez, Lucas J van Vliet, and Koenraad A Vermeer, "Accuracy assessment of intra-and intervisit fundus image registration for diabetic retinopathy screening," *Investigative ophthalmology & visual science*, vol. 56, no. 3, pp. 1805–1812, 2015.

[51] Antoine Rivail, Ursula Schmidt-Erfurth, Wolf-Dieter Vogl, Sebastian M Waldstein, Sophie Riedl, Christoph Grechenig, Zhichao Wu, and Hrvoje Bogunovic, "Modeling disease progression in retinal octs with longitudinal self-supervised learning," in *International Workshop on PRedictive Intelligence In MEdicine*. Springer, 2019, pp. 44–52.

[52] Yuting Hu, Zhiling Long, Anirudha Sundaresan, Motaz Alfarraj, Ghassan AlRegib, Sungmee Park, and Sundaresan Jayaraman, "Fabric surface characterization: assessment of deep learning-based texture representations using a challenging dataset," *The Journal of The Textile Institute*, vol. 112, no. 2, pp. 293–305, 2021.

[53] Yazeed Alaudah, Patrycja Michałowicz, Motaz Alfarraj, and Ghassan AlRegib, "A machine-learning benchmark for facies classification," *Interpretation*, vol. 7, no. 3, pp. SE175–SE187, 2019.

[54] Yuhan Zhang, Mingchao Li, Zexuan Ji, Wen Fan, Songtao Yuan, Qinghuai Liu, and Qiang Chen, "Twin self-supervision based semi-supervised learning (ts-ssl): Retinal anomaly classification in sd-oct images," *Neurocomputing*, 2021.

[55] Jiaming Qiu and Yankui Sun, "Self-supervised iterative refinement learning for macular oct volumetric data classification," *Computers in biology and medicine*, vol. 111, pp. 103327, 2019.

[56] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al., "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[57] Sven Holm, Greg Russell, Vincent Nourrit, and Niall McLoughlin, "Dr hagis—a fundus image database for the automatic extraction of retinal surface vessels from diabetic patients," *Journal of Medical Imaging*, vol. 4, no. 1, pp. 014503, 2017.

[58] Ashish Markan, Aniruddha Agarwal, Atul Arora, Krinjeela Bazgain, Vipin Rana, and Vishali Gupta, "Novel imaging biomarkers in diabetic retinopathy and diabetic macular edema," *Therapeutic Advances in Ophthalmology*, vol. 12, pp. 2515841420950513, 2020.

[59] Dominick A Rizzi and Stig Andur Pedersen, "Causality in medicine: towards a theory and terminology," *Theoretical Medicine*, vol. 13, no. 3, pp. 233–254, 1992.

[60] Arpan Guha Mazumder, Swarnadip Chatterjee, Saunak Chatterjee, Juan Jose Gonzalez, Swarnendu Bag, Sambuddha Ghosh, Anirban Mukherjee, and Jyotirmoy Chatterjee, "Spectropathology-corroborated multimodal quantitative imaging biomarkers for neuroretinal degeneration in diabetic retinopathy," *Clinical Ophthalmology (Auckland, NZ)*, vol. 11, pp. 2073, 2017.

[61] Hui-Zhuo Xu, Zhiming Song, Shuhua Fu, Meili Zhu, and Yun-Zheng Le, "Rpe barrier breakdown in diabetic retinopathy: seeing is believing," *Journal of ocular biology, diseases, and informatics*, vol. 4, no. 1, pp. 83–92, 2011.

[62] Laxmi Gella, Rajiv Raman, Padmaja Kumari Rani, and Tarun Sharma, "Spectral domain optical coherence tomography characteristics in diabetic retinopathy," *Oman journal of ophthalmology*, vol. 7, no. 3, pp. 126, 2014.

[63] Yuji Itoh, Ashleigh L Levison, Peter K Kaiser, Sunil K Srivastava, Rishi P Singh, and Justis P Ehlers, "Prevalence and characteristics of hyporeflective preretinal tissue in vitreomacular interface disorders," *British Journal of Ophthalmology*, vol. 100, no. 3, pp. 399–404, 2016.

[64] Jay S Duker, Peter K Kaiser, Susanne Binder, Marc D de Smet, Alain Gaudric, Elias Reichel, SriniVas R Sadda, Jerry Sebag, Richard F Spaide, and Peter Stalmans, "The international vitreomacular traction study group classification of vitreomacular adhesion, traction, and macular hole," *Ophthalmology*, vol. 120, no. 12, pp. 2611–2619, 2013.

[65] George Trichonas and Peter K Kaiser, "Optical coherence tomography imaging of macular oedema," *British Journal of Ophthalmology*, vol. 98, no. Suppl 2, pp. ii24–ii29, 2014.

[66] Mohit Prabhushankar, Gukyeong Kwon, Dogancan Temel, and Ghassan AlRegib, "Contrastive explanations in neural networks," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3289–3293.

[67] Yash-yee Logan, Mohit Prabhushankar, , and Ghassan AlRegib, "Decal: Deployable clinical active learning," in *International Conference on Machine Learning (ICML) Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2022.

[68] Yash-yee Logan, Ryan Benkert, Ahmad Mustafa, and Ghassan AlRegib, "Patient aware active learning for fine-grained oct classification," in *International Conference on Image Processing (ICIP)*. 2022, IEEE.

Checklist

1. For all authors...

   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

   (b) Did you describe the limitations of your work? [Yes] . Please see Section 5

   (c) Did you discuss any potential negative societal impacts of your work? [Yes] . Please see Section 5

   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [N/A]

   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] . Please see Section A.1.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] . Please see training details for each experiment in Section 4.1, 4.2, and 4.3. Further details for the contrastive experiments can be found in Section C.2.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Tables found at 2 and 3. Also, see Sections C.4, C.3, and C.2 for the standard deviation values associated with the time-series and contrastive learning experiments.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] . Please see Section C.8

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes] . Please see Section A.5.

   (b) Did you mention the license of the assets? [Yes] Please see Section A.2.

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Please see Section A.1.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] . Please see Section D.1, Line 138.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] . Please see Section 3.2, Line 173.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] The clinical procedure for both trials is discussed in Section D.

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] . Please see Section 3.1, Line 139 and Section D.2, Line 1086.

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]