iDetect: AI Powered Retinal Disease Detection from Fundus Images

A Project Report

Submitted by

Sl.No	STUDENT NAME	ID NUMBER
1	FAHAD ABDULRAHMAN ALOTHMAN	201673400
2	SHAHBAAZ AHMED SADIQ	202415720

Under the supervision of

Dr. Muzammil Behzad

Assistant Professor

Information & Computer Science Department

For the Course

ICS 504: DEEP LEARNING

COLLEGE OF COMPUTING & MATHEMATICS

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN

November 2025

ABSTARCT

We present an end-to-end pipeline for training and evaluating RET-CLIP on the ODIR-5K retinal fundus dataset. The system integrates binocular image inputs (left/right eyes) and clinical text to learn joint vision–language representations. Our implementation uses a ViT-B/16 vision encoder and medical-domain text encoders (primarily PubMedBERT), supports automated prompt generation from ODIR metadata, and builds RET-CLIP-compatible LMDB datasets. We evaluate two regimes: (1) zero-shot classification via image–text cosine similarity across disease keywords extracted from the test split, and (2) linear probing with a logistic-regression classifier trained on frozen image features. Key metrics are accuracy and F1 (macro/weighted). In a test configuration (≈100 patients, small number of epochs), we observe consistent zero-shot performance and further gains with linear probing, demonstrating that clinically grounded text prompts improve alignment and separability of retinal disease features. The pipeline is modular (data prep, prompt generation, pretraining/fine-tuning, evaluation), reproducible, and readily extensible to additional text encoders or training schedules.

Contents

1. INTRODUCTION

1.1. Problem Statement

Given paired left/right fundus images and corresponding clinical keywords, learn a joint image–text embedding space that enables (a) zero-shot disease classification from text prompts and (b) strong linear separability of visual features with minimal supervised labels. We target practical improvements in accuracy and F1 while keeping the workflow reproducible and extensible.

1.2. Objectives

- Build a unified ODIR-5K → RET-CLIP data pipeline (images + prompts → LMDB/JSONL).
- Train RET-CLIP with ViT-B/16 and medical text encoders (PubMedBERT by default).
- Evaluate zero-shot classification across disease keyword prompts (cosine similarity).
- Evaluate linear probing with logistic regression on frozen features.
- Report accuracy, macro-F1, and weighted-F1; visualize confusion matrices.
- Enable encoder/model swaps (e.g., BERT/BioBERT) and training schedule changes.

1.3. Scope of Study

We focus on ODIR-5K binocular images with English diagnostic keywords. The study evaluates zero-shot and linear-probe regimes; full supervised fine-tuning of downstream heads and multi-dataset generalization are out of scope but supported by the code structure.

2. LITERATURE REVIEW

2.1. Related Work

Contrastive vision language learning aligns images and text in a shared embedding space so that matched pairs are close and mismatched pairs are far apart. CLIP demonstrated that large-scale image–text pairing enables powerful zero-shot recognition, where class names or short prompts replace task-specific heads [1]. Follow-ups showed that prompt construction and lightweight adaptation (e.g., learnable prompts) further improve transfer, while strong vision backbones like ViT stabilize scaling [2] [3]. These ingredients contrastive loss, prompt design, and transformer-based encoders have become the de facto template for open-vocabulary perception.

Medical imaging work adapted this recipe to clinical data, where images can be paired with free-text reports or keywords. ConVIRT, BioViL, and GLoRIA trained on image report corpora to learn cross-modal representations that support retrieval and label efficient classification under limited supervision [4] [5] [6]. These methods consistently find that domain specific language encoders improve terminology coverage and disambiguation (e.g., differentiating “exudates” from “drusen”), which matters when prompts describe fine grained findings [7] [8]. Beyond global alignment, “local” text region interactions have also been explored to better ground lesions and anatomical structures.

In ophthalmology, most prior pipelines are vision only and supervised, tackling diabetic retinopathy grading, glaucoma assessment, and multi-disease tagging with CNN/ViT backbones trained on curated labels [9] [10] [11]. Recently, foundation style pretraining (e.g., masked autoencoders) and image–text contrast for medical domains have begun to reduce reliance on dense labels and improve long tail robustness [3], [12]. However, explicit open-vocabulary, prompt-based recognition in retinal imaging is less explored. Systems like RET-CLIP bridge this gap by pairing binocular fundus images with disease keywords to learn a joint space where clinically phrased prompts enable zero-shot predictions and provide a natural interface for linear probing and downstream finetuning.

2.2. Limitations in Existing Approaches

Vision backbones have shifted from CNNs to transformers, with ViT-B/16 emerging as a strong, widely reproduced baseline for both supervised and multimodal setups [3]. On the language side, biomedical pretraining (PubMedBERT, BioBERT) improves coverage of domain terms and abbreviations, reducing prompt sensitivity and synonym brittleness in zero-shot classification [7] [8]. Together, a ViT image tower and a biomedical LM text tower form a practical base for medical VLMs, especially when text is short (keywords, summaries) rather than full reports.

For retinal tasks, large supervised studies established strong CNN baselines for DR screening and related diseases, and community challenges like REFUGE advanced glaucoma assessment tooling and evaluation protocols [9] [10] [11]. Yet, labels are expensive, class distributions are long-tailed, and many findings are subtle or co-occurring, conditions under which zero-shot and linear-probe evaluations are attractive. VLMs trained on binocular inputs can, in principle, encode asymmetries between left and right eyes that carry diagnostic signal (e.g., unilateral lesions, cup-to-disc asymmetry), offering robustness when one eye is noisy or occluded.

Despite these advances, open issues persist. Zero-shot predictions derived from cosine similarity are often poorly calibrated, motivating calibration-aware evaluation and abstention in clinical settings [13]. Domain shift across devices, sites, and populations can degrade performance; prompt ensembling and small adapter-based finetuning help but add complexity and compute [2], [12]. Finally, dataset resources like ODIR-5K remain widely used for benchmarking open-set recognition and binocular modelling; consistent reporting of accuracy, macro-/weighted-F1, confusion matrices, and prompt sensitivity analyses would further clarify progress and failure modes [14].

In our implementation, the available RET-CLIP [15] checkpoint uses a Chinese-pretrained text encoder, which introduces a language domain mismatch when prompts and labels are written in English. This mismatch affects tokenization, vocabulary coverage (medical abbreviations, synonyms), and phrase-level semantics, often degrading zero-shot alignment and making predictions more sensitive to exact wording. Practical mitigations include swapping the text tower for a biomedical English encoder (e.g., PubMedBERT/BioBERT), translating prompts into the encoder's native language, or adapter-based finetuning of the text tower on English retinal terminology; in our experiments, English-domain biomedical encoders improved robustness to prompt phrasing and synonyms.

3. PROPOSED METHODOLOGY

3.1. Existing Model and Challenges

RET-CLIP jointly embeds binocular images and clinical text via a contrastive objective.

Challenges:

i) prompt sensitivity (wording, synonyms)
ii) class imbalance and multi-label presentations
iii) language mismatch if using a Chinese text tower with English prompts
iv) binocular noise from variable crops or laterality inconsistencies.

3.2. Proposed Enhancements

- Prompt sets & ensembling: multiple phrasings/abbreviations per class; aggregate similarities to reduce wording sensitivity.
- English biomedical text tower: prefer PubMedBERT/BioBERT for English prompts; keep adapter hooks for small, targeted tuning.
- Binocular consistency: add inter-eye constraints or augmentations (eye swap, single-eye occlusion) to encourage robust binocular features.
- Calibration: apply temperature scaling / confidence thresholds for zero-shot; log ECE along with F1/accuracy.

3.3. Algorithm and Implementation

i) Load ODIR metadata (patient ID, left/right labels/keywords).
ii) Generate prompts (left/right/patient), preserving clinical terms and synonyms.
iii) Preprocess images to 224×224, normalize; build LMDB + JSONL (paths, laterality, prompts).
iv) Train RET-CLIP with ViT-B/16 + selected text tower; checkpoint features.
v) Evaluate zero-shot (cosine to prompt sets) and linear probe (logistic regression).
vi) Save metrics, confusion matrix, and predictions for analysis/plots.

3.4. Loss Function and Optimization

- Loss Function: a CLIP-style contrastive loss with binocular (tripartite) structure
  i.e., it contrasts (left-eye ↔ text) and (right-eye ↔ text) and enforces inter-eye consistency (often described in the notebook as "Three-Level/Tripartite contrastive loss").

- Optimization: AdamW, cosine decay with warmup, gradient-norm clipping.

4. EXPERIMENTAL DESIGN AND EVALUATION

4.1. Dataset and Preprocessing

Ocular Disease Intelligent Recognition (ODIR) is a structured ophthalmic database of 5,000 patients with age, colour fundus photographs from left and right eyes and doctors' diagnostic keywords from doctors. This dataset is meant to represent ‘‘real-life’’ set of patient information collected by Shanggong Medical Technology Co., Ltd. from different hospitals/medical centres in China. In these institutions, fundus images are captured by various cameras in the market, such as Canon, Zeiss and Kowa, resulting into varied image resolutions. Annotations were labelled by trained human readers with quality control management.

They classify patient into eight labels including:

- Normal (N),
- Diabetes (D),
- Glaucoma (G),
- Cataract (C),
- Age related Macular Degeneration (A),
- Hypertension (H),
- Pathological Myopia (M),
- Other diseases/abnormalities (O)

In our pipeline, fundus images are uniformly resized to 224×224 (main path uses bicubic interpolation), converted to tensors, and normalized. Two normalization schemes appear depending on the branch: CLIP normalization with mean (0.4815, 0.4578, 0.4082) and std (0.2686, 0.2613, 0.2758) when features are extracted for RET-CLIP, and standard ImageNet normalization with mean (0.485, 0.456, 0.406) and std (0.229, 0.224, 0.225) in a separate block. No data augmentations (flips, color jitter, random crops) are applied in the current notebook—this keeps the preprocessing deterministic and directly comparable across runs.

For text, we clean and standardize prompts derived from ODIR metadata (left/right/patient keywords), including lowercasing and normalizing quotes. Tokenization is handled either by a Hugging Face AutoTokenizer for biomedical encoders (e.g., PubMedBERT/BioBERT) or by the RET-CLIP tokenizer for CLIP-style text towers. Dataset preparation removes duplicate patients, filters rows to images that exist on disk, writes train/test splits to CSV/TSV/JSONL, and packs image bytes plus annotations into LMDB databases for fast, consistent I/O during training and evaluation. This combination—side-aware prompt construction, strict split artifacts, and LMDB-backed loading—supports both zero-shot (prompt similarity) and linear-probe evaluation downstream.

4.2. Performance Metrics

We evaluate the model with top-1 accuracy, Macro-F1, Weighted-F1, and a confusion matrix. Accuracy reports overall correctness across all test images. Macro-F1 averages F1 scores equally over classes, so it highlights performance on rare/long-tail diseases. Weighted-F1 averages per-class F1 using class frequencies, reflecting how the model behaves under the dataset's natural skew. The confusion matrix visualizes where predictions go wrong (which classes are commonly confused) and is the basis for any per-class precision/recall you might quote. These metrics are computed for both tracks in the notebook—zero-shot (prompt similarity; implemented) and linear probe (logistic regression on frozen features; reported)—so you can compare alignment quality (zero-shot) with linear separability of the visual features (linear probe).

4.3. Experiment Setup

We use the ODIR-5K binocular fundus dataset with side-aware prompts derived from the metadata (left/right/patient keywords). Images are resized to 224×224, converted to tensors, and normalized with either CLIP mean/std (for RET-CLIP feature paths) or ImageNet mean/std (in a separate branch). No data augmentation is applied. Data is packed into LMDB with companion JSONL/TSV/CSV split artifacts for reproducible train/test evaluation and fast I/O.

The vision tower is ViT-B/16; the text tower is selectable among PubMedBERT and BioBERT for English prompts (the code also supports the original Chinese text encoder). Training follows a CLIP-style contrastive objective tailored to binocular inputs (left↔text, right↔text, optional inter-eye consistency); the notebook doesn't echo an optimizer instantiation in the visible cells, so it uses the RET-CLIP defaults from the training wrapper. Evaluation is run in two tracks: zero-shot (image–prompt cosine similarity; implemented in code but not printed in the saved run) and linear probing on frozen image features using scikit-learn LogisticRegression with multinomial loss, solver='lbfgs', max_iter=1000, and n_jobs=-1. Metrics computed/logged are top-1 accuracy, Macro-F1, Weighted-F1, plus a confusion matrix. In the captured run, linear-probe results were reported and the confusion matrix was generated; zero-shot cells are present and ready to rerun with your chosen prompt sets and text encoder.

Works Cited

[1] A. K. J. H. C. e. a. Radford, *Learning Transferable Visual Models From Natural Language Supervision,* arXiv, 2021.

[2] K. Y. J. L. C. &. L. Z. Zhou, "Learning to Prompt for Vision-Language Models," *International Journal of Computer Vision,* vol. 130, no. Springer Science and Business Media LLC, p. 2337–2348, 2022.

[3] A. B. L. K. A. e. a. Dosovitskiy, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,* arXiv, 2021.

[4] Y. Z. a. H. J. a. Y. M. a. C. D. M. a. C. P. Langlotz, *Contrastive Learning of Medical Visual Representations from Paired Images and Text,* arXiv, 2022.

[5] S.-C. a. S. L. a. L. M. P. a. Huang, "GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[6] B. U. N. B. S. e. a. Boecking, "Making the Most of Text Semantics to Improve Biomedical Vision--Language Processing," in *Springer Nature Switzerland*, 2022.

[7] Y. T. R. C. H. e. a. Gu, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Trans. Comput. Healthcare,* vol. 3, no. January 2022, p. 23, 2021.

[8] J. Y. W. K. S. e. a. Lee, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics,* vol. 36, p. 1234–1240, September 2019.

[9] V. P. L. C. M. e. a. Gulshan, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA,* vol. 316, pp. 2402-2410, 2016.

[10] "REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical Image Analysis,* vol. 59, p. 101570, January 2020.

[11] D. C. C. L. G. e. a. Ting, "Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes," *JAMA,* vol. 318, pp. 2211-2223, 2017.

[12] K. C. X. X. S. L. Y. D. P. &. G. R. He, *Masked Autoencoders Are Scalable Vision*

Learners, arXiv, 2021.

[13] C. G. a. G. P. a. Y. S. a. K. Q. Weinberger, *On Calibration of Modern Neural Networks,* arXiv, 2017.

[14] andrewmvd, "kaggle," 2019. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k.

[15] J. D. a. J. G. a. W. Z. a. S. Y. a. H. L. a. H. L. a. N. Wang, *RET-CLIP: A Retinal Image Foundation Model Pre-trained with Clinical Diagnostic Reports,* 2024.