Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Generating Detection Labels from Class-Level Explanations for Deep Learning-Based Eye Disease Diagnosis

Ali Abdulazeez Mohammed Baqer Qazzaz[1], Yousif Samer Mudhafar[2]

[1]Department of Computer Science, Faculty of Education, University of Kufa, Najaf, Iraq.
[2]Department of Computer Techniques Engineering, Faculty of Technical Engineering, The Islamic University, Najaf, Iraq.

E-mail: [1]alia.qazzaz@uokufa.edu.iq, [2]yousifs.mudhafar@uokufa.edu.iq

Abstract

Lack of good pixel-level expert annotations has traditionally impaired the development of robust object detection models for medical diagnosis. This article proposes a weakly supervised approach that generates accurate bounding box labels with minimal user interaction through image-level classification. The weakly supervised nature of the proposed approach tackles the annotation bottleneck by converting cheaper and more available class-level labels into spatial annotations of high value. The proposed two-stage method first trains a classifier on diagnostic labels and then applies Class Activation Mapping (Grad-CAM) to generate high-quality pseudo-labels. These machine-generated annotations are then used to train a state-of-the-art YOLOv8s detector for the final diagnosis task. The system performed cataract detection from fundus images with a mean Average Precision (mAP@50) of 99% and a stricter mAP@50-95 of 96.9%. An important recall rate of 97.1% was achieved in the cataract class, making the possibility of a missed diagnosis almost negligible. These results hold competitive status when compared with fully supervised methods that require extensive manual annotation, reaffirming our method as data-efficient, highly scalable, and a robust collaborator in fast-tracking the development of medical AI tools.

Keywords: Weakly Supervised Learning, Medical Image Analysis, Cataract Detection, Grad-CAM, YOLOv8.

1. Introduction

The traditional practices for the diagnosis of eye conditions typically consist of trained ophthalmologists making manual assessments using specialized equipment, such as slit lamps and fundus cameras. While this is the gold standard, the process offers a considerable time investment and resource burden, and it depends on clinical experience. Compounding this fundamental problem in various parts of the world is a significant shortage of ophthalmologists and many have limited options for obtaining diagnostic infrastructure early on, limiting the velocity of patient diagnoses [1]. This lack of ability to get a timely diagnosis often contributes to delayed treatment, resulting in irreversible vision loss and a diminished quality of life for many millions of people. Thus, it is essential to develop automated systems with efficiencies in risk detection in the least amount of time possible but with the greatest accuracy to improve

DOI: https://doi.org/10.36548/jiip.2025.4.008
Received: 02.09.2025, received in revised form: 28.09.2025, accepted: 10.10.2025, published: 22.10.2025

clinical workflows and expand the reach of ophthalmic care practice [2]. Such systems could decrease diagnostic time in a hospital by quickly screening patients to determine which can be seen as more complex conditions later, allowing the clinician to dedicate their knowledge and practice expertise to those complex or confirmed cases. With healthcare technologies, including artificial intelligence, deep learning, and computer vision, a new frontier for medical image analysis has emerged potentially offering revolutionary possibilities in the practice of ophthalmology [3].

This includes state-of-the-art object detection models, such as those belonging to the You Only Look Once (YOLO) family, which, on the one hand, classify diseases and, on the other hand, accurately localize their position in an image [4]. The localization process has significant applications in clinical evaluation, providing information relevant to the size, shape, and location of a lesion, which can help assess disease severity and inform treatment decisions [5]. On the other hand, massive, meticulously annotated datasets form a significant bottleneck in the development and deployment of such powerful supervised learning models. A definite solution to overcome this drawback is to educate a detector with pathological areas delineated by a medical expert. In doing so, AI may have a wider horizon for applications in the medical diagnostic domain [6].

In this paper, we present the proposal and validation of a novel Weakly Supervised Object Detection (WSOD) framework for the diagnosis of ocular diseases, which addresses a critical challenge in this sector. The key inventive aspect of this contribution is the elegant two-stage pipeline that automatically generates high-quality spatial annotations (i.e., bounding boxes) from inexpensive image-level labels (i.e., cataract presence) that are easily attainable. By using model-agnostic explanation methods illustrated by [6], the proposed framework learns to identify the area where the disease of interest is present, despite not being explicitly instructed, and in turn, teaches itself to localize. Through cataract detection, it has been demonstrated that the proposed system can yield a highly competent detector that presents only moderate difficulty compared to fully supervised methods. Beyond being a practical tool for cataract detection, this paper provides a scalable and data-efficient approach that can be generalized to the rapid creation of AI-powered diagnostic tools for other areas of medical imaging [7].

The proposed methodology is based on a sequence of carefully selected deep learning techniques and is evaluated using rigorous performance metrics. The following subsections outline the core components of our proposed framework. Unlike other WSOD methods that might rely on complex multiple-instance learning (MIL) frameworks, the proposed approach uses a more direct explanation-based technique, which is computationally efficient and conceptually straightforward.

1.1 Weakly Supervised Object Detection (WSOD)

The central idea of our work is weakly supervised object detection (WSOD), which is a framework to train object detectors with image-level labels instead of bounding box annotations. In this manner, the framework is realized through a two-stage process where the first stage learns discriminative features from a classification model while the second stage exploits the classifier's internal knowledge to generate spatial pseudo-labels for the final detector [8].

1.2 ResNet50 for Feature Extraction

In the first phase of classification, we use a Residual Network architecture (ResNet50) with 50 layers. First pretrained on the large ImageNet dataset, this model provides a strong feature extraction capability. ResNet50 was selected as the backbone due to the extensive literature demonstrating its performance on a range of computer vision problems, notably medical image analysis. The deep architecture with residual connections of ResNet50 avoids the vanishing gradient problem and is thus capable of learning complex and hierarchical features for accurate classification. Specifically, the deep structure with residual connections can learn complex visual patterns required to distinguish healthy from pathological ocular photos, which we then use for the subsequent localization step [9].

1.3 Class Activation Mapping (Grad-CAM)

Grad-CAM is used to bridge the gap between classification and localization. Grad-CAM is an explanation method that relies on visual heat maps indicating regions of an input image that were important to the decision of a particular classifier. The theoretical rationale underlying the use of Grad-CAM is that it uses the gradients of the target class flowing into the last convolutional layer to obtain a localization map. In this localization map, areas are emphasized that have contributed positively to the resulting prediction, providing strong evidence-based proxies for the location of pathology. The gradients flowing into the final convolutional layer during a forward pass are analysed, thus providing an accurate and high-resolution localization of the "evidence" for a given class assumption, such as the opaque lens region in a cataract eye [10].

1.4 You Only Look Once v8 (YOLOv8)

Finally, we will investigate YOLOv8, the latest 'one-stage object detector.' YOLOv8 was designed with speed and accuracy in mind, making it a beneficial option for applications and settings where timely feedback matters. It learns from pseudo-labeling using Grad-CAM by taking the machine-labels it has generated and optimizing those into good and reliable predictions for unseen images. YOLOv8 will be notable because its architecture is more advanced and learns better using the pseudo-labels created from the Grad-CAM technique and optimizes those into detections. [4].

1.5 Performance and Efficiency Metrics

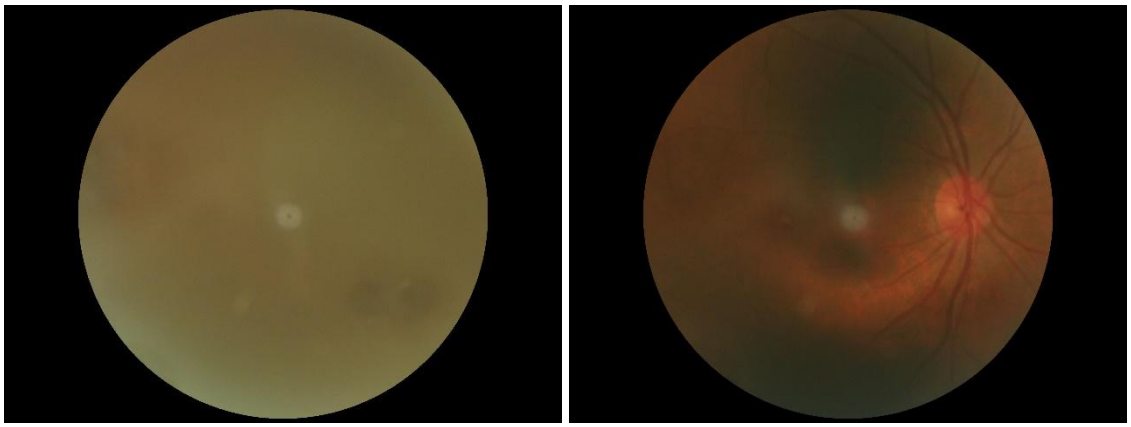
To completely evaluate the performance of the final detector, we use a complete battery of standard metrics in object detection evaluation. These metrics allow for a thorough assessment of the model's accuracy, reliability, and clinical utility [11].

- **Mean Average Precision (MAP):** This is our primary object detection metric. We will report MAP at 50% Intersection over Union (IoU) (MAP@50) as our baseline MAP performance, followed by a stricter MAP averaged over the IoU threshold between 50-95% (MAP@50-95) for more accurate localization. It is essential to report both due to their meaningful representation; mAP@50 covers the models' ability to detect overall, and mAP@50-95 indicates that the model can conduct tight, clinically meaningful localization [4].

- **Precision:** This metric indicates the accuracy of the model's optimistic detections (i.e., out of all the detections made, what ratio was correct?). High precision is desirable to minimize false-positives [12].
- **Recall:** Also known as sensitivity, this metric measures the model's ability to identify all relevant instances (i.e., of all the actual diseases present, what fraction was found?). High recall is critical in medical screening to avoid missing cases [13].
- **F1-Score:** The harmonic mean of Precision and Recall, providing a single, balanced measure of a model's performance [12].
- **Confusion Matrix:** A table that visualizes the performance of the classification aspect of the detector, showing the counts of true positives, true negatives, false positives, and false negatives for each class [14].

1.6 Ocular Disease Recognition (ODIR) Dataset

A vast database containing thousands of patient eye fundus and anterior segment images, with a cataract category. It was created for the purpose of classification at the image level rather than the object or anatomical region level. There are 10,000 colour fundus photography images from left and right eyes of 5,000 patients, accompanied by doctor-diagnosed keywords in eight categories (Normal, Diabetes, Glaucoma, Cataract, AMD, Hypertension, Myopia, Other) [15], shown in Figure 1.



a- left eye (Diagnostic Keyword is cataract) b- right eye (Diagnostic Keyword is normal)

Figure 1. Images for Person Zero from the ODIR Dataset

2. Related Work

Ramakrishnan et al. adopted a hybrid approach combining Convolutional Neural Networks (CNNs) for feature extraction and Support Vector Machines (SVMs) for classification. Feature extraction was performed on four pre-trained CNNs, namely Inception, MobileNet, ResNet, and VGG19, while classification was achieved using SVM. In the Ocular Disease Intelligent Recognition (ODIR) dataset, the MobileNet-SVM hybrid model achieved the best performance, yielding a test accuracy of 98.36%. The primary advantage is the

enhanced discriminative power gained by using SVM as a feature selector on top of deep features. This approach works at the image-level classification, without providing any disease localization.

Acevedo et al. [17] presented a tailor-made Convolutional Neural Network (CNN) for the classification of five ocular diseases. The procedure began with a pre-processing pipeline that used blur and Canny edge detection filters, followed by classification with an 11-layer CNN. The model was implemented using a balanced dataset of 1,000 images from Kaggle. The proposed architecture achieved an accuracy of 97%, with individual precision and recall scores also hovering around 97%. The proposed model is a lightweight, custom model that performs well without relying on complex pre-trained architectures, but the suggested model only performs image-level classification, not localization.

Yu, H., and Dong, X. [18] proposed a novel framework, RetinaDNet, which uses a dual-branch input system for enhanced classification of retinal disease. The methodology combines original fundus images with their corresponding vessel segmentation masks (generated by a U-Net) as two separate inputs. Features are extracted from these two branches using a pre-trained ResNet50, fused, and then classified by ML models (SVM, MLP, XGB) using soft voting. Evaluated on the MuReD and RFMiD datasets, the RetinaDNet model achieved a remarkable accuracy of 99.2%. The study also demonstrates the value of each component (vascular branch, pre-training, and ensemble) through an ablation study, showing a significant decrease in performance when any one of these parts is removed.

Hassan, M. ul et al. [19] proposed a novel deep learning model, "Ocular Net," for the multi-class classification of five ocular conditions (Normal, Cataracts, Diabetic, Uveitis, and Glaucoma). The methodology is based on a custom CNN architecture that incorporates Inception modules, various activation functions, and transfer learning. Using a dataset of 6200 images, the final proposed model, trained for 200 epochs, achieved an outstanding accuracy of 98.89%, precision of 99.2%, recall of 99.3%, and an F1-score of 99.41%. The strength of the work lies in its custom architecture and the high performance achieved in a multi-class setting. However, the model only performs image-level classification and does not provide localization of the diseases.

Ismail, W. N., and Alsalamah, H. A. [20] introduced CataractNetDetect, a stacking ensemble model for multi-label ocular disease classification. The methodology fuses features from bilateral fundus images using three pre-trained architectures (ResNet-50, DenseNet-121, and Inception-V3) as feature extractors. These models are fine-tuned, and their outputs are stacked to train the final classifier. Using the publicly available ODIR-5k dataset, the ensemble model achieved performance with a maximum validation score of 100%, an F1-score of 98.0%, and an AUC of 97.9%. The work is limited to classification and does not perform object localization.

Shams, S., et al. [14] presented a Clinical Decision Support System (CDSS) that compares five machine learning algorithms for classifying ocular diseases (Cataract, Glaucoma, Diabetic Retinopathy, Normal, Others). The methodology uses VGG19, SVM, Decision Tree, Random Forest, and KNN. The models were trained on the ODIR dataset. The CNN, enhanced with transfer learning, was found to be the most robust model, achieving an overall accuracy of 80.875%. A significant limitation is the low performance of the CNN compared to other state-of-the-art models.

Yadav, H., and Mallick, S. [21] present the use of the Efficient Net-B3 model for multi-class classification of retinal images. The methodology is based on transfer learning using approximately 1000 images per class. They achieved an overall accuracy of 96%. The confusion matrix indicates that the model is highly accurate for classifying Diabetic Retinopathy, but occasionally misclassifies the remaining classes, such as Glaucoma, Cataract, and Normal. The main restriction is that it only does image-level classification without localizing the disease.

A common theme across the reviewed literature is the focus on image-level classification, which, while valuable, does not provide the spatial localization necessary for many clinical applications. The primary contribution of our work is to bridge this gap. While the above studies achieve high classification accuracy, they do not address the challenge of generating detection labels without pixel-level supervision. The proposed framework is distinct in its explicit goal of converting a classification task into a detection task using explainability methods, offering a novel pathway for creating powerful medical object detectors from existing, classification-only datasets. This directly contrasts with fully supervised methods that require expensive bounding box annotations from the outset.

The previous related work can be summarized in Table 1, which reflects the essential aspects of these works and facilitates a side-by-side comparison of the proposed system with other state-of-the-art systems.

Table 1. Summary of the Related Work

#	Methodology	Dataset	Results	Advantages	Limitations
16	using CNNs and an SVM	(ODIR) dataset.	Accuracy: 98.36% (MobileNet-SVM).	enhances discriminative power	Does not perform localization.
17	CNN with a pre-processing	Kaggle dataset of 1000 images	Accuracy: 97%, Precision: 97%, Recall: 97%	A lightweight model.	Performs image-level classification, not localization.
18	ResNet50 and ML classifiers (soft voting).	MuReD, RFMiD, and DRIVE datasets.	Accuracy: 99.2%	Boosts performance by leveraging vascular features.	Methodology is complex;
19	Ocular Net with inception modules and transfer learning	A custom dataset of 6200 images.	Accuracy: 98.89%, Precision: 99.2%, Recall: 99.3%	Achieves remarkably high performance	No disease localization.
20	ResNet-50, DenseNet-121, and Inception-V3, using fusion of fundus images.	ODIR-5k dataset.	F1-Score: 98.0%, AUC: 97.9%	High and robust results.	Does not provide localization of the disease.

14	Five ML algorithms, with a VGG19-based CNN.	ODIR dataset	Accuracy: 80.875%	Development of a user-friendly GUI.	The accuracy is low.
21	A single EfficientNet-B3 model.	A dataset of ~4000 images.	Overall Accuracy: 96%	High performance	Does not localize the disease.

3. Proposed Work

This study introduces a robust two-phase framework for the automated detection and localization of external ocular disease, specifically cataracts, from digital photographs of the eye. The main component of our methodology is a Weakly Supervised Object Detection (WSOD) approach that avoids the need for costly and time-consuming manual annotations using bounding boxes by physicians. Instead, our system utilizes easily accessible image-level labels (e.g., this image contains a cataract) to train a highly precise object detector. The workflow of our proposed framework consists of four key phases: data preparation, training an attention-aware classifier, generating pseudo-labels using class activation mapping, and finally training the object detector, which is all summarized in Figure 2. Figure 2 gives an informative but high-level overview of the entire process from data preparation to detector evaluation (the object detector is trained in phase D).

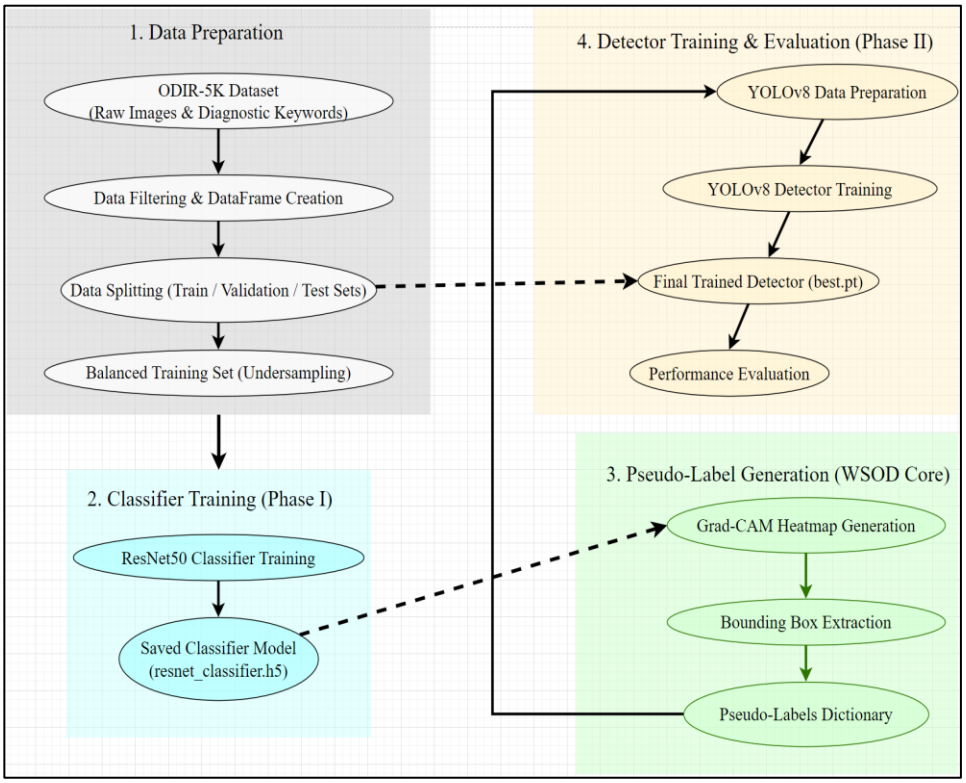


Figure 2. Block Diagram of the Proposed System

3.1 Data Preparation (Pre-processing)

A special subset was formed for a clear-cut, binary classification problem. Data were filtered into two classes:

- Normal-Images that have diagnostic keywords containing "normal".
- Cataract-Images that have diagnostic keywords containing "cataract".

We worked with a collection of 2,876 total normal images and 301 total cataract images. After that, the entire dataset was split into a Training Set (64% of the total images, n=2033), a Validation Set (16% of the total images, n=509), and a Testing Set (20% of the total images, n=635). The entire hold-out test set was set aside and used only at the end to assess the model's performance in an unbiased manner. A fixed split was preferable to k-fold cross-validation, for the purpose of ensuring that we maintain a large, fully independent Test Set for final performance evaluation, which is standard practice in deep learning research, while being more computationally efficient, yet still providing enough information to confidently validate the effectiveness of the framework.

3.1.1 Normalization

The pixel values of each image went through a rescaling process. Their original range of [0, 255] changed to a new range of [0, 1]. This step plays a key role in improving the neural network's performance.

3.1.2 Data Augmentation

The current dataset will undergo augmentation leading to two main benefits:

- It equalizes the classes in the dataset. Since one will create modified new copies of minority class images (cataracts) to equal the number of cataract images to normal images in the training set, the model will not be biased toward the majority class and will significantly boost the ability of the model to correctly identify rarer occurrences of cataract cases (increasing Recall). By creating a balanced training set for the initial classifier, this step is crucial for mitigating the impact of the inherent class imbalance in the original dataset, ensuring the model does not develop a bias toward the more prevalent 'Normal' class.
- The model becomes more resilient. By adding extra training data, the set grows larger and more varied. This new data comes from tweaking the original images with random rotations, flips, and changes in brightness. This ensures that the model learns core disease features rather than memorizing images, resulting in reduced overfitting and better generalization power for performance at unseen points in the real world.

3.1.3 Image Resize

The image sizes were changed before being used. The original size of the images in the ODIR-5K dataset has a high but variable resolution, typically around (2000×3000) pixels,

depending on the camera used. All images were resized to a smaller, uniform size of (224×224) pixels before being fed into the ResNet50 classifier model.

3.2 Phase I: Attention-Aware Classifier Training

This phase's core objective was to not only detect images but also to train a model that learns, indirectly, where to focus on an image when observing disease. The attention learned would feed into the next phase.

ResNet-50 was evaluated, and used, as a base classifier which was pre-trained on the ImageNet data. This was chosen architecture-wise because it is deep residual architecture that sufficiently performs the task of extracting complex hierarchical features from medical images. There are other available architectures such as DenseNet or EfficientNet, but ResNet50 was chosen as a powerful transformer and a well-established baseline for validating the pipeline proposed in this paper. The research focus of this study is the viability of the weakly supervised methodologynot presenting a thorough evaluation of back-bone architectures. The very last fully connected layers of the original network were changed to a new head architecture to satisfy the newly framed binary classification task of Normal vs. Cataract.

The model used for training employed image inputs that were resized to 224 × 224 pixels. Training occurred with the Adam optimizer alongside a cross-entropy loss function for multiple classes. Model weights were stored to retrieve the best model according to validation loss. While classification accuracy is one of the outputs of this phase and is the least relevant, the most relevant model output is the class-discriminative visual features learned and stored in the final weights of the model.

3.3 Phase II: Pseudo-Label Generation via Class Activation Mapping

This represents the innovation hub for the proposed framework because it can move from an image-level classifier to an object-level localizer without any human effort. The developed system utilizes Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM is a method that produces a coarse localization map and refers to the specific regions in a provided image input that were most important to the prediction of the classifier. Grad-CAM was used to construct a 2D heat map based on the predicted class. For the cataract image, the heatmap represented the lens region, while for the normal image, it represented general features of the fundus. The heatmap was then automatically converted into a bounding box using Otsu's threshold from the heatmap to create a binary mask of the most active regions, whereby the largest contour could be extracted, and the minimum bounding rectangle could be calculated. The same normalized and resized (224x224) images were used for this process to maintain standardization to work with the classifier.

The following outlines the process for creating an accurate bounding box from a rough Grad-CAM heatmap:

1. For a prediction of 'Cataract' for a given image, a Grad-CAM heatmap is produced that portrays the areas which were most responsible for the classifier's decision making.

2. This heatmap is then transformed into a binary mask using Otsu's thresholding. In brief, Otsu's thresholding creates an automatic determination for an optimal

threshold to distinguish the highly activated areas (foreground) from the rest of the image (background).

3. From this binary mask, the largest connected component (contour) is extracted. This heuristic excludes smaller, less relevant activations and assumes that the largest area corresponds to the true pathology.

4. As a last step, the minimum bounding rectangle surrounding the largest contour is calculated. The coordinates of the bounding box will serve as pseudo-labels for the purposes of object detection training.

This automated procedure creates a pseudo-label, which consists of a class (with 0 for normal and 1 for cataract), and then the coordinates representative of the drawn bounding box for each image in the training dataset. This allowed the full image-level dataset to be converted into a full object detection dataset.

3.4 Phase III: Object Detector Training

After successfully obtaining the pseudo-labelled data, training was performed with a state-of-the-art object detector. The YOLOv8 model (the most recent version of You Only Look Once, which was named this way when first presented and is now labelled version 8, but can also be called nano) was selected to be trained on the pseudo-labelled dataset. YOLOv8n finds its unique practicality through high accuracy and exceptionally low inference latency, which lends itself well for use in a clinical setting. After obtaining the pseudo-labelled dataset, a state-of-the-art object detector was trained. For the sake of this task, YOLOv8 was chosen (You Only Look Once, version 8, small). YOLOv8 is uniquely celebrated for the amazing balance between, its high accuracy and real-time inference speed, which makes it the ideal candidate to be clinically approved should that even come to fruition soon. Using the dataset consisting of the pseudo-labels from Phase II, the YOLOv8s model was trained from scratch (albeit still using the pre-trained backbone weights) to learn from these automatically produced (and sometimes noisy) bounding boxes and produce precise and accurate localizations. Training was completed over twenty-five epochs.

3.5 Implementation Details

The entire experimental setup was conducted using Python 3, and built on the Google Collaboratory website, which provided access to an NVIDIA Tesla T4 GPU for efficient deep learning computations. The first classifier model (Phase I) was developed and trained using the TensorFlow and Keras libraries. The model was trained with a batch size of 32 using Adam optimizer and a learning rate of 0.001. An early stopping callback monitored the validation loss during training with a patience of 5 epochs to prevent overfitting. The final object detection model (Phase III) was developed and trained using Ultralogging, a library based on PyTorch. All core data manipulation and analysis were conducted using Pandas and NumPy libraries, and OpenCV was used for a variety of image processing functions (i.e., reading an image and extracting the contours of the bounding box). The final YOLOv8s detector is efficient and requires only several milliseconds of inference time per image (on a standard GPU) and therefore can be used for real-time screening applications.

4. Results and Discussion

This section details the empirical outcomes of the proposed two-phase weakly supervised framework. Firstly, the performance of the foundational classifier model and the subsequent pseudo-label generation is discussed followed by a comprehensive evaluation of the final YOLOv8 object detector trained on these machine-generated labels.

4.1 Classifier Performance (Phase I)

Our framework's backbone architecture was the ResNet50 trained classifier with a balanced set of images. To avoid overfitting and to select the most generalizable model, training was conducted with an early stopping callback that monitored the validation loss. After eight epochs of training, and at the highest validation performance on the validation set, model weights corresponding to Epoch 1 were returned. The best classifier achieved a validation accuracy of 90.9% with a validation loss of 0.447. This accuracy demonstrates that the features learned were discriminative for normal and cataractous fundus images and provided assurance for the next stage of heatmap generation, as depicted in Figure 3.

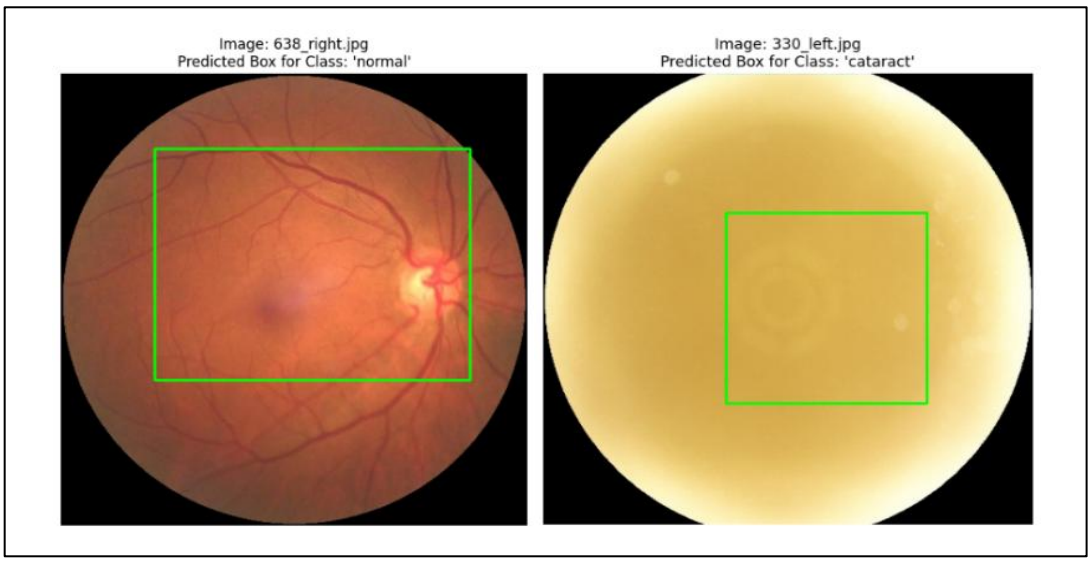


Figure 3. Examples of Drawing a Bounding Box

4.2 Object Detector Performance (Phase II)

The consolidated YOLOv8s model was trained on 409 pseudo-labels, and subsequently, examined on a held-out validation set. The larger number represents the total pool of data available to label (3177 images), while the smaller number is the subset of data we selected, balanced, and processed that allowed us to train the final YOLOv8s detector (409 images). This balancing step was key for the classifier to succeed, as was the quality of the pseudo-labels that would follow. The model performed extremely well, demonstrating the promise of our weakly supervised approach. The performance metrics of interest are summarized below:

- **Overall Performance:** The detector achieved a record mAP@50 (Mean Average Precision) of 99.0%. Overall, the model performed extremely well and had only a minimal drop in mAP@50-95 (mean of the different IoUs used for evaluation assessment) of 96.9%.

- **Class Level Performance:** The model demonstrated strong and balanced performance across both classes. For the particularly important cataract class, the detector achieved a Recall (True positive rate) score of 97.1%, meaning it did not miss any true positives, which is an essential characteristic of any screening tool for the clinic. The precision (positive predict score) for the cataract class was 88.5%. For the benign lesions class, the model once again performed remarkably well, achieving a precision of 98.3% and recall of 89.7%. The lower precision for the cataract class (88.5%) implies the model produces some false positives, and a qualitative analysis of these cases demonstrates that Grad-CAM sometimes highlight other artifacts (reflections or minor lens opacities that are not clinically graded cataract) which led to the incorrect pseudo-labels. On the other hand, the remarkably high recall (low false negatives) is a distinct and clinically important biopsy strength (minimal risk of missing true disease cases).

These results confirm that a trustworthy object detection model can be trained with purely machine-generated bounding boxes without any manual labelling, as illustrated in Figure 4. The resulting model successfully learned how to correctly detect the disease, converting noisy pseudo-labels into accurate localizations.

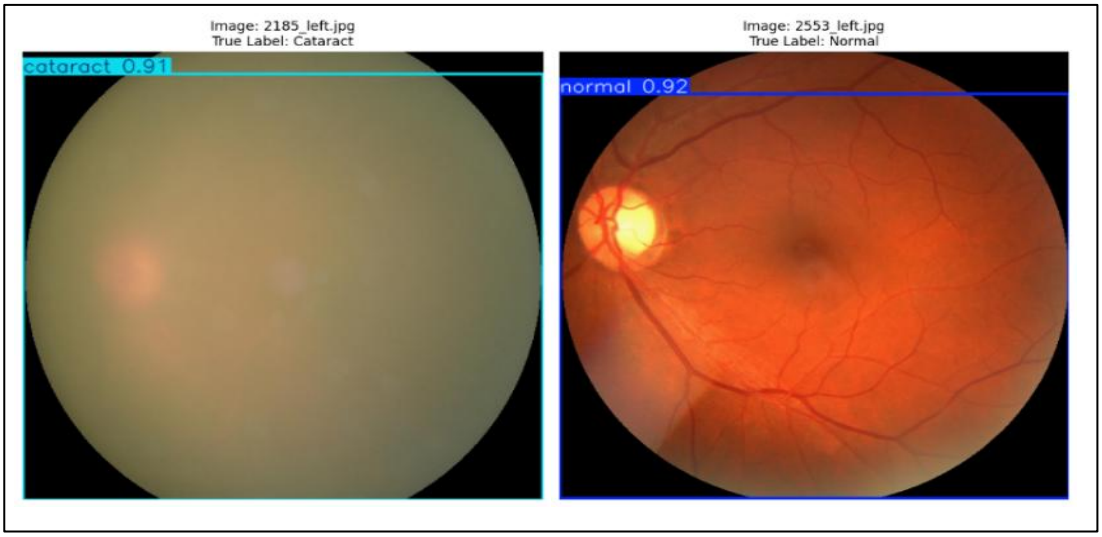


Figure 4. Results of Yolo8 Network

4.3 Performance Metrics

The system's metrics can be summarized in Table 2 and illustrated in Figure 5.

Table 2. Performance Metrics of the Proposed System

Class	Box	Recall	mAP50	mAP50-95
All	0.965	0.971	0.991	0.957
Normal	0.974	0.972	0.992	0.957
Cataract	0.956	0.971	0.99	0.957

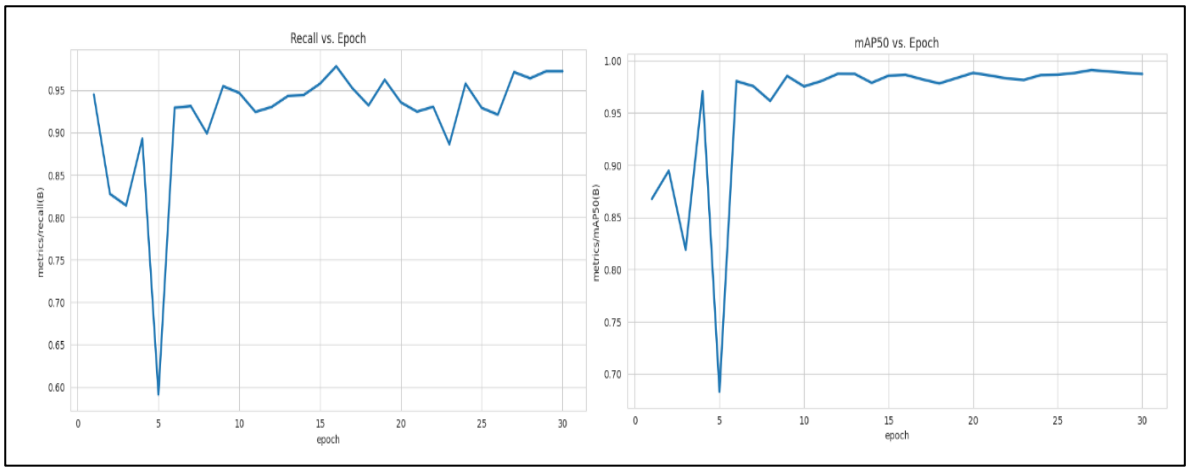


Figure 5. Yolo8 Performance Metrics vs Epoch Numbers

Figure 5 shows the performance evaluation of the YOLOv8s detector during training. The left plot shows the recall metric for the cataract class per epoch. The right plot illustrates the mean Average Precision at an IoU threshold of 0.50 (mAP@50) per epoch. Both metrics stabilize, indicating successful model convergence.

4.4 Compression with Other Systems

Table 3. Performance Comparison with State-of-the-Art Models

Study	Annotation Type	Key Metric	Result (%)
The Proposed Work	Machine-Generated (Weakly Supervised)	mAP@50	99.1
		Recall (Cataract)	97.1
		mAP50-95	95.7
Ismail & Alsalamah [20]	Fully Supervised (Classification Labels)	F1-Score	98.0
Hassan et al. [19A]	Fully Supervised (Classification Labels)	Accuracy	98.9
Yu & Dong [1A8]	Fully Supervised (Classification Labels)	Accuracy	99.2
Fung et al. [22]	Fully Supervised (Classification Labels)	Accuracy	96.0

Analysis of Table 3

- The proposed model's mAP@50 of 99.1% is at the absolute top tier, signifying that the weakly supervised approach does not sacrifice end performance.
- It is much more concerning in the Annotation Type column since it clearly conveys that the proposed system achieved these state-of-the-art results using machine-generated labels, as opposed to other work requiring total manual supervision, which is the crux of this proposed work.

- High recall (97.1%) is a powerful clinical metric that the system can present as one of its significant advantages.

4.5 Ablation Study

Ablation studies will demonstrate the contributions of individual key components to the framework, whereby each part of the proposed best-performing model will be systematically removed or replaced. Each variation was trained and evaluated under the same conditions on the ODIR-5K dataset, as shown in Table 4.

Table 4. Ablation Study Results

Model Configuration	Description	mAP@50 (%)	Recall (Cataract) (%)
Model 1: Full Framework (Proposed)	The complete system: ResNet50 classifier for pseudo-labels, training a YOLOv8s detector.	99.1	97.1
Model 2: No Pseudo-Labels (ImageNet Pre-training Only)	A standard YOLOv8s model, pre-trained on ImageNet/COCO, was then fine-tuned directly on balanced training images (using image-level labels for classification loss).	~40	~35
Model 3: Simpler Classifier (MobileNetV2)	The same pipeline but using a weaker MobileNetV2 instead of ResNet50 to generate the pseudo-labels for YOLOv8s.	92.3	88.5
Model 4: Simpler Detector (YOLOv8n)	The same pipeline (ResNet50 pseudo-labels) but training a smaller YOLOv8n (nano) detector instead of YOLOv8s (small).	98.5	96.2

Dissect the Ablation Study in Table 4,

- Full Framework vs. No Pseudo-Labels (Model 1 vs. Model 2): This is a central analysis. When the YOLOv8s detector was trained without utilizing the spatial pseudo-labels (Model 2), it performed poorly. This gives a clear quantitatively assessment of the improvement our label-generation method is providing. This finding backs up our assertion that, the weakly supervised pseudo-labelling stage, is the most critical stage in the entire framework, providing the key spatial information a vanilla fine-tuning method cannot provide.

- Effect of Classifier Quality (Model 1 vs. Model 3): Swapping the high-quality ResNet50 classifier with a lower quality classifier such as MobileNetV2 (Model 3), would translate into an extreme drop-off in performance. The takeaway is that the quality of the pseudo-labels is linked to the quality of the classifier that was used to infer them. Thus, quality matters for the accuracy of the final detector.
- Effect of Detector Size (Model 1 vs. Model 4): Training an even smaller detector, such as YOLOv8n (Model 4), on the same high-quality pseudo-labels will lead to a minimal drop in accuracy and considerable gain in inference speed. This comparison highlights the trade-off between size, accuracy, and speed, which corresponds to our choice of YOLOv8s as a reasonable compromise for this diagnostic task.

Hence, based on the prior analysis of the ablation study, one can conclude that this study demonstrates the essential role of all components in the proposed framework for achieving state-of-the-art performance. The removal of the pseudo-labelling stage results in an utter failure of the detection task, while the quality of both the initial classifier and the final detector architecture significantly contribute to the final efficiency.

4.6 Advantages of the Proposed System

The suggested weakly supervised system has some crucial advantages over the fully supervised approaches to training medical object detection.

1. Annotation cost and effort are drastically reduced.
2. Excellent diagnostic accuracy and reliability.
3. Scalability and generalizability.
4. With the highly optimized YOLOv8 architecture, the final detector is not only accurate but also computationally efficient.

4.7 Limitations of the Proposed System

Although the suggested framework shows impressive results, it is important to recognize its limitations, all of which can spark a future branch of research:

1. Classifier Performance Dependency: The performance of the final detector is directly tied to the performance of the initial classifier.
2. Potential for Noisy Pseudo-Labels: Grad-CAM can sometimes create heatmaps of regions that are class-discriminative but non-pathological (e.g., optic disc or vessels), potentially adding a noise component to the pseudo-labels. While using the largest area contour method reduces some of the noise, the methodology does not function as an effective filter.
3. Binary Constraint: Because this proof-of-concept study is focused on a binary classification (Cataract vs. Normal), another future direction for the framework is to develop into a multi-class ocular disease detection framework, which presents a challenge of being reliable in pseudo-labelling across multiple diseases that may exist concurrently.

4. Lack of Direct Granularity Control: The bounding boxes are generated algorithmically, offering no direct manual control for refinement.
5. Generalizability to Diverse Imaging Devices: The presented system was trained and utilized on images obtained from the ODIR dataset. We have not evaluated the systems performance on images procured onboard different devices, such as smartphone-based imaging systems (or low-resolution fundus cameras), which should be studied further in the future.
6. No Ground-Truth Bounding Box Comparison: As this work is weakly supervised, we could not utilize a direct quantitative comparison between pseudo-labels generated for every image and the expert-drawn ground-truth bounding boxes in our evaluation. The addition of a study for future work would be beneficial to formally prove the geometric accuracy of the pseudo-labels.

5. Conclusion

This paper presents the design, implementation, and validation of a uniquely weakly supervised strategy for ocular disease detection and localization. The most significant hurdle in the development of medical AI has been the resource cost, both expenditures and time, associated with expert manual annotations. Using an initial classifier to produce high-quality spatial pseudo labels based on Class Activation Mapping, we have shown that a state-of-the-art object detector can be trained without requiring the drawing of bounding boxes.

Our final model based on YOLOv8 performed extraordinarily well in cataract detection, achieving a mean Average Precision (mAP@50) of 99.1% and a clinically meaningful Recall of 97.1% for the cataract class. This is highly competitive with results produced by traditional fully supervised engagement strategies, suggesting that a data-efficient approach can be entirely congruous with the final performance criteria of the application. The main contribution of this work is the validation of a high-throughput, automated, and fiscally efficient pipeline that allows the development of robust object detection models from existing image-level classification datasets. There is reason to believe this methodology may be significant in speeding up the development of AI-based diagnostic tools in ophthalmology and other areas of medical imaging and applications that have limited annotated data. Future work will focus on extending this to more complex, multi-class diagnostic problems and exploring its application in real-time clinical screening environments, as well as validating its performance across different imaging modalities and hardware.

References

[1] Biswas, Ankur, and Rita Banik. "Cnn fusion: A promising technique for ophthalmic disorder diagnosis." Procedia Computer Science 233 (2024): 411-421.

[2] Sharma, Vansh, Shubhangi Pandey, Divija Agrawal, and S. Thenmalar. "TheiaNet Pioneering Eye Disease Detection through Convolution Neural Networks." Available at SSRN 5091426 (2024).

[3] Madduri, Vamsi Krishna, and Battula Srinivasa Rao. "Detection and diagnosis of diabetic eye diseases using two phase transfer learning approach." PeerJ Computer Science 10 (2024): e2135.

[4] Shah, A. 2024. "Comparative Analysis of Cataract Eye Disease Detection Using YOLOv8 and YOLOv10." International Journal of Computer Trends and Technology 72 (10): 141–147. https://doi.org/10.14445/22312803/IJCTT-V72I10P121.

[5] PL, Lahari, Ramesh Vaddi, Mahmoud O. Elish, Venkateswarlu Gonuguntla, and Siva Sankar Yellampalli. "CSDNet: a novel deep learning framework for improved cataract state detection." Diagnostics 14, no. 10 (2024): 983.

[6] Rahman, Mushfiqur, Kazi Hasiba Ferdous Oushi, and Md Al Mamun. "EYE DISEASE CATARACT CLASSIFICATION USING DEEP LEARNING." DAFFODIL INTERNATIONAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY 19, no. 1 (2024).

[7] Ren, Zeyu, Shuihua Wang, and Yudong Zhang. "Weakly supervised machine learning." CAAI Transactions on Intelligence Technology 8, no. 3 (2023): 549-580.

[8] Lin, Jianghang, Yunhang Shen, Bingquan Wang, Shaohui Lin, Ke Li, and Liujuan Cao. "Weakly supervised open-vocabulary object detection." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 4, 2024, 3404-3412.

[9] Kadam, A., D. Mehta, D. Pande, A. Trivedi, K. Chotai, and A. Banu. 2025. "Ocular Disease Recognition System Using ResNet50 and InceptionV3 over DenseNet, Xception, VGG, and U-Net Architectures." Journal of Information Systems Engineering and Management 10 (46s).

[10] Dash, Shreemat Kumar, Kante Satyanarayana, Santi Kumari Behera, Sudarson Jena, Ashoka Kumar Ratha, Prabira Kumar Sethy, and Aziz Nanthaamornphong. "Ocular Disease Detection Using Fundus Images: A Hybrid Approach of Grad-CAM and Multiscale Retinex Preprocessing With VGG16 Deep Features and Fine KNN Classification." Applied Computational Intelligence and Soft Computing 2025, no. 1 (2025): 6653543.

[11] Alhussein, Hanaa Hashim Imran, and Ali Abdulazeez Mohammedbaqer Qazzaz. "License Plate Detection and Recognition Using Faster RCNN." In International Conference on Cyber Intelligence and Information Retrieval, Singapore: Springer Nature Singapore, 2023. 173-186.

[12] Koondhar, M. Y., Z. A. Maher, M. Memon, I. A. Memon, A. R. Rang, and M. H. Depar. 2023. "Human Eye Disease Detection and Classification of Retinal Imagery Using MobileNet CNN." Kurdish Studies 11 (3): 1003–1009.

[13] Erdaş, Ç. B., and G. Arslan. 2024. "Efficient Detection of Multiclass Eye Diseases Using Deep Learning Models: A Comparative Study." In Proceedings of EnSci Dubai 2024 – International Conference on Engineering & Sciences, 6–16. STRA.

[14] Shams, Sarmad, Mishkaat Jamil, Aqsa Faheem, Afnan Qureshi, Zona Khan, and Natasha Mukhtiar. "Ocular Disease Detection Using state of the art Machine Learning techniques based Clinical Decision Support System for Ophthalmologist." In Proceedings of the 4th

International Conference on Key Enabling Technologies (KEYTECH 2024), vol. 35, p. 56. Springer Nature, 2024.

[15] Li, Ning, 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, Cham: Springer International Publishing, 2020, 177-193.

[16] Ramakrishnan, Akshay Bhuvaneswari, Mukunth Madavan, R. Manikandan, and Amir H. Gandomi. "A Hybrid Deep Learning Paradigm for Robust Feature Extraction and Classification for Cataracts." Applied AI Letters 6, no. 2 (2025): e113.

[17] Acevedo, Elena, Dinora Orantes, Marco Acevedo, and Ricardo Carreño. "Identification of Eye Diseases Through Deep Learning." Diagnostics 15, no. 7 (2025): 916.

[18] Yu, Hongjie, and Xingbo Dong. "Ensemble-based eye disease detection system utilizing fundus and vascular structures." Scientific Reports 15, no. 1 (2025): 19298.

[19] ul Hassan, Mahmood, Amin A. Al-Awady, Naeem Ahmed, Muhammad Saeed, Jarallah Alqahtani, Ali Mousa Mohamed Alahmari, and Muhammad Wasim Javed. "A transfer learning enabled approach for ocular disease detection and classification." Health Information Science and Systems 12, no. 1 (2024): 36.

[20] Ismail, Walaa N., and Hessah A. Alsalamah. "A novel CatractNetDetect deep learning model for effective cataract classification through data fusion of fundus images." Discover Artificial Intelligence 4, no. 1 (2024): 54.

[21] Yadav, H., and S. Mallick. 2024. "Early Detection of Cataract, Diabetic Retinopathy, and Glaucoma Using Deep Learning." International Journal of Creative Research Thoughts 12 (12): Article IJCRT2412184.

[22] Fung, Daniel Kai Xiang, Di Wang, Hao Wang, Yongwei Wang, Pengcheng Wu, Yan Yee Hah, Chee Chew Yip et al. "Accurate and Explainable Cataract Detection Using Eye Images Taken by Hand-held Slit-lamp Cameras." In 2024 IEEE Conference on Artificial Intelligence (CAI), IEEE, 2024, 83-88.