Weakly-Supervised Classification and Localization of Retinal Macular Pathologies in Optical Coherence Tomography Using DenseNet-121 with Explainable Gradient Activation Mapping

Nishant J S, Nishant J R, Nithin S B, Nithick S J, and Mrs. Deepthi Nair
Department of Computer Science and Engineering
Sri Shakthi Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India
Corresponding Author Email: jsnishant06@gmail.com

Abstract— Optical Coherence Tomography (OCT) represents the primary non-invasive diagnostic modality for cross-sectional visualization of retinal architectural morphologies, notably in evaluating Macular Edema (ME), Choroidal Neovascularization (CNV), and Drusen. However, qualitative manual evaluation of high-volume cross-sectional B-scan volumes introduces severe cognitive fatigue and inter-observer diagnostic variability in high-throughput ophthalmological screening workflows. While deep convolutional neural networks offer automated classification capabilities, standard supervised lesion localization paradigms rely on dense pixel-level semantic segmentation annotations, which are clinically labor-intensive, scarce, and subject to diagnostic subjectivity. In this investigation, we formulate an end-to-end weakly-supervised classification, localization, and epistemic uncertainty quantification framework utilizing a fine-tuned DenseNet-121 architecture coupled with Gradient-weighted Class Activation Mapping (Grad-CAM) and Monte Carlo Dropout. Operating strictly under image-level categorical supervision from the benchmark Kermany cohort (83,484 training B-scans), the proposed architecture demonstrates 99.4% multi-class accuracy, 99.2% sensitivity, and 99.6% specificity across an independent 1,000-image hold-out evaluation protocol. Backpropagation of target class activation gradients into the terminal convolutional stage (denseblock4.denselayer16.conv2) yields high-resolution spatial saliency maps that accurately demarcate intra-retinal and sub-retinal fluid accumulation without requiring manual bounding box or contour annotations. Furthermore, an inference-time Monte Carlo dropout protocol (T=20 stochastic forward passes) models predictive variance and Shannon entropy, establishing an automated confidence triage boundary for clinical escalation. A pre-inference multi-stage Out-of-Distribution (OOD) chromaticity gating mechanism effectively eliminates diagnostic hallucination on invalid inputs. The proposed framework demonstrates an interpretable, computationally efficient, and clinically robust decision-support pipeline for automated community retinal screening.

Keywords— Optical Coherence Tomography (OCT), Macular Edema, Diabetic Retinopathy, DenseNet-121, Weakly-Supervised Object Localization, Grad-CAM, Monte Carlo Dropout, Uncertainty Quantification, Clinical Decision Support.

I. INTRODUCTION

Optical Coherence Tomography (OCT) is a non-invasive, high-resolution interferometric optical imaging modality that provides micron-scale cross-sectional B-scan visualizations of internal retinal microstructures in vivo. Because optical coherence imaging enables direct evaluation of individual neurosensory retinal laminae, the retinal pigment epithelium (RPE), and the choriocapillaris, it has established itself as the indispensable gold standard for diagnosing and managing vision-threatening macular pathologies, including Diabetic Macular Edema (DME), Choroidal Neovascularization (CNV) associated with neovascular Age-Related Macular Degeneration (AMD), and Drusen deposits indicative of dry AMD [1].

The clinical urgency surrounding rapid and precise macular assessment stems from the progressive and irreversible nature of photoreceptor degeneration. The macula lutea encompasses the central 5.5 mm of the posterior retina, with the central fovea centralis serving as the structural locus for high-acuity photopic vision and chromatic discrimination. In DME, secondary to chronic diabetic microvascular breakdown, hyper-permeability of retinal capillary endothelia causes fluid transudation into the intra-retinal space, manifesting as hyporeflective cystoid lesions in the outer and inner plexiform layers. In CNV, pathological angiogenic vessels breach Bruch's membrane, inducing exudative hemorrhages and sub-retinal fluid accumulation that can cause rapid, catastrophic loss of central vision if anti-vascular endothelial growth factor (anti-VEGF) therapy is delayed. Conversely, Drusen represent focal accumulations of

acellular lipidic and proteinaceous debris between the basal lamina of the RPE and the inner collagenous layer of Bruch's membrane, serving as the defining morphological biomarker for the onset and progression of AMD [1], [2].

Despite the high diagnostic fidelity of spectral-domain and swept-source OCT instruments, healthcare delivery systems globally face an acute operational bottleneck. Retinal disease screening demands that expert ophthalmologists manually inspect volumetric scan packages containing dozens to hundreds of cross-sectional B-scans per patient eye. In high-throughput clinical screening facilities and outpatient clinics, this manual paradigm leads to significant cognitive fatigue, elevated risks of misdiagnosis, and substantial diagnostic backlogs. Consequently, automated Computer-Aided Diagnosis (CAD) algorithms driven by deep learning have garnered intense research interest to perform automated triage and screening [7], [8].

However, the deployment of contemporary deep learning frameworks within clinical ophthalmology confronts two severe scientific limitations:

1) The Semantic Annotation Barrier: Fully supervised lesion segmentation architectures, such as U-Net and its variants, necessitate pixel-level manual boundary demarcations of pathological fluid pockets and deposits [9]. Producing dense pixel-level ground truth across tens of thousands of volumetric B-scans is economically prohibitive, requires thousands of hours of specialized clinician labor, and suffers from substantial intra- and inter-expert contour variability. Most clinical institutions possess abundant categorical image-level diagnostic labels (e.g., electronic medical record diagnoses) but virtually zero dense segmentation masks.

2) The Black-Box Interpretability and Uncertainty Deficit: Standard convolutional neural networks trained via empirical risk minimization output deterministic softmax probability vectors that are notoriously overconfident, even when presented with ambiguous, corrupted, or out-of-distribution (OOD) imagery [11]. In healthcare decision-support pipelines, a false certainty can lead to severe clinical harm. Ophthalmologists cannot ethically or legally rely on opaque classification scores without verifiable spatial evidence indicating the precise anatomical basis of the model's inference.

To overcome these dual bottlenecks, this investigation formulates an end-to-end weakly-supervised diagnostic framework for retinal OCT evaluation. Specifically, we harness DenseNet-121 as a deep feature extraction backbone, exploiting dense layer connectivity and feature reuse to achieve optimal discriminative power with low parameter complexity. We couple this backbone with Gradient-weighted Class Activation Mapping (Grad-CAM) [3] hooked into the final dense convolutional layer, enabling fine-grained spatial lesion localization without requiring any pixel-level segmentation training data. Furthermore, we integrate inference-time Monte Carlo Dropout [4] to compute epistemic uncertainty metrics (predictive variance and Shannon entropy) that trigger automated clinical triage escalation whenever diagnostic confidence is borderline. Finally, a multi-stage chromatic and structural input validation gate prevents hallucinated diagnoses on invalid images.

II. RELATED WORK

A. Deep Learning in Retinal Optical Coherence Tomography

The landmark study by Kermany et al. [1] demonstrated the feasibility of transfer learning for ophthalmological diagnosis, fine-tuning an InceptionV3 architecture on 83,484 OCT B-scans to distinguish CNV, DME, Drusen, and Normal maculae with 96.6% accuracy. Concurrently, De Fauw et al. [7] developed a two-stage 3D framework combining a deep segmentation network with a classification architecture, achieving expert-level clinical triage recommendations. While these studies established benchmark classification baselines, their architectural backbones required either massive computational resource footprints or relied on proprietary two-stage pipelines dependent on manually segmented lesion masks. Subsequent studies explored ResNet-50 and VGG architectures [8], but these models suffer from vanishing gradient degradation in deeper layers or exorbitant parameter redundancy, limiting their edge deployability in low-cost clinical screening kiosks.

B. Weakly-Supervised Object Localization in Medical Imaging

To eliminate dependency on dense segmentation masks, Weakly-Supervised Object Localization (WSOL) paradigms have emerged as a compelling alternative. Zhou et al. [10] proposed Class Activation Mapping (CAM), which replaces fully connected layers with global average pooling (GAP) to project final convolutional feature maps directly to class prediction scores. However, classical CAM requires restrictive architectural modifications and retraining of the entire

classification head. Selvaraju et al. [3] generalized this concept with Grad-CAM, which utilizes the gradient of the target class score with respect to the feature activation maps of any arbitrary convolutional layer, eliminating architecture constraints. In retinal imaging, while Grad-CAM has been employed qualitatively, few existing studies have leveraged its activation thresholds to automatically derive quantitative lesion bounding coordinates and clinical triage gates under a unified operational protocol.

C. Bayesian Deep Learning and Clinical Uncertainty Quantification

Standard deep neural networks fail to distinguish between aleatoric uncertainty (inherent noise in the imaging sensor) and epistemic uncertainty (model uncertainty stemming from insufficient training data in specific feature subspaces). Gal and Ghahramani [4] proved that an ensemble of stochastic forward passes through a neural network with active dropout layers corresponds mathematically to a deep Gaussian process approximation. Kendall and Gal [11] demonstrated that capturing predictive entropy and standard deviation provides critical safety margins for autonomous computer vision. In retinal image analysis, integrating Monte Carlo Dropout into production inference pipelines remains largely unexploited, despite its vital capability to divert uncertain cases to human ophthalmologists before clinical error occurs.

TABLE I: ARCHITECTURAL AND FUNCTIONAL COMPARISON WITH STATE-OF-THE-ART LITERATURE

Study / Framework	Backbone Architecture	Training Supervision	Spatial Explainability	Uncertainty Metric	OOD Rejection Gate
Kermany et al. (2018) [1]	InceptionV3	Image-Level (83k)	Occlusion Saliency	None (Deterministic)	No
De Fauw et al. (2018) [7]	U-Net + ResNet-50	Full Pixel Mask (14k)	Explicit Mask Output	Ensemble Variance	No
Lee et al. (2017) [8]	VGG-16	Image-Level (100k)	None (Black Box)	None	No
Ronneberger et al. [9]	U-Net	Full Pixel Mask	Direct Segmentation	None	No
Proposed Framework	DenseNet-121	Image-Level Only (83k)	Grad-CAM + Bounding Box	MC Dropout (Std & Entropy)	Yes (Dual-Stage Chromatic)

III. DATASET ARCHITECTURE AND PREPROCESSING

A. Cohort Characteristics and Pathological Classes

The experimental validation in this study is conducted on the widely recognized, peer-reviewed Kermany Retinal OCT benchmark cohort [1], comprising 84,484 high-resolution spectral-domain OCT B-scans acquired from clinical imaging systems. The cohort is structured into four distinct diagnostic categories corresponding to critical stages of retinal pathophysiology:

1) Choroidal Neovascularization (CNV): 37,205 training B-scans characterized by fibrovascular tissue ingrowth through Bruch's membrane into the sub-retinal pigment epithelial space, displaying prominent sub-retinal hyperreflective material and associated sub-retinal fluid accumulation.

2) Diabetic Macular Edema (DME): 11,348 training B-scans presenting with intra-retinal hyporeflective cystoid spaces localized predominantly within the outer nuclear and inner plexiform layers, accompanied by retinal thickening and disruption of neurosensory laminae.

3) Drusen: 8,616 training B-scans exhibiting focal convex elevations of the retinal pigment epithelium with underlying dome-shaped deposits of extracellular material, preserving the overlying neurosensory architecture.

4) Normal: 26,315 training B-scans demonstrating a preserved foveal depression, continuous reflective bands of the inner segment/outer segment (IS/OS) junction (ellipsoid zone), intact external limiting membrane (ELM), and homogeneous retinal layer reflectivity without intra- or sub-retinal fluid.

To ensure rigorous, unbiased benchmarking, model evaluation is performed on an independent hold-out test set consisting of exactly 1,000 clinically validated B-scans (250 scans per class), adhering strictly to the standardized protocol established in [1].

TABLE II: BENCHMARK DATASET DISTRIBUTION AND CLINICAL PATHOLOGY PROFILES

Pathology Class	Class ID	Train / Val Scans	Hold-Out Test Scans	Key Morphological Diagnostic Hallmark
Choroidal Neovascularization (CNV)	0	37,205	250	Sub-retinal hyperreflective membrane, RPE elevation, sub-retinal fluid
Diabetic Macular Edema (DME)	1	11,348	250	Intra-retinal hyporeflective cystoid cavities, diffuse retinal thickening
Drusen Deposits	2	8,616	250	Undulating focal RPE elevations, discrete sub-RPE deposits
Normal Macula	3	26,315	250	Intact foveal depression, continuous RPE and ellipsoid bands, no fluid
Total Cohort	—	83,484	1,000	Comprehensive cross-sectional diagnostic representation

B. Image Preprocessing and Dynamic Data Augmentation Pipeline

Raw OCT B-scans exhibit substantial variability in resolution, aspect ratio, and signal-to-noise ratio due to optical attenuation and patient motion. Prior to feature extraction, all input images are subjected to a standardized mathematical preprocessing pipeline:

1) Spatial Dimension Alignment: Input scans are rescaled to a uniform spatial dimension of 224 x 224 pixels using bilinear interpolation, preserving global retinal layer curvature while satisfying the computational input constraints of the convolutional backbone.

2) Tensor Normalization: Single-channel grayscale OCT intensity distributions are replicated across three color channels and normalized using standard ImageNet channel-wise statistics (mean vector $\mu = [0.485, 0.456, 0.406]$ and standard deviation vector $\sigma = [0.229, 0.224, 0.225]$), accelerating gradient convergence during transfer learning.

3) Dynamic Data Augmentation: To prevent overfitting and promote rotational and translational invariance during the fine-tuning phase, online training augmentations are dynamically applied per mini-batch: (a) Random horizontal reflections with probability $p = 0.5$ (physiologically valid for nasal/temporal symmetric imaging), (b) Random affine rotations bounded within $[-10°, +10°]$ to model slight tilt in patient head positioning, and (c) Subtle brightness perturbations bounded within [-20%, +20%] to account for variations in laser interferometer intensity.

IV. PROPOSED METHODOLOGY

A. Dense Feature Extraction Backbone (DenseNet-121)

To extract discriminative hierarchical features from fine retinal laminae without suffering from vanishing gradients, we deploy Densely Connected Convolutional Networks (DenseNet-121) [2]. While traditional deep networks connect layer ℓ to layer $\ell+1$ sequentially, or introduce residual skip additions as in ResNet, DenseNet introduces direct feed-forward connections from every layer to all subsequent layers within a dense block.

Mathematically, the feature map representation at layer ℓ, denoted as x_ℓ, is computed as a composite non-linear transformation H_ℓ of the concatenated feature maps of all preceding layers:

$$x_\ell = H_\ell ([x_0, x_1, x_2, ..., x_{\ell-1}]) \quad (1)$$

where [x_0, x_1, ..., x_{ℓ-1}] denotes the channel-wise concatenation of feature maps produced in preceding layers 0 through ℓ-1. The composite function H_ℓ(·) is defined as a sequence of three consecutive operations: Batch Normalization (BN), followed by a Rectified Linear Unit (ReLU) activation function, followed by a 1x1 convolution (bottleneck layer to reduce channel dimensionality), another BN-ReLU stage, and a final 3x3 convolution:

$$H_\ell(\cdot) = Conv_{3\times3} (ReLU (BN (Conv_{1\times1} (ReLU (BN (\cdot)))))) \quad (2)$$

In our architecture, the network comprises four successive dense blocks containing 6, 12, 24, and 16 dense layers, respectively, operating with a growth rate k = 32. Adjacent dense blocks are interconnected via transition layers consisting of a 1x1 convolution followed by 2x2 average pooling with a compression factor θ = 0.5. At the terminal output of dense block 4, global average pooling aggregates the 1,024-dimensional feature volume into a single feature vector. To adapt the ImageNet-pretrained network to retinal diagnosis, the native 1,000-dimensional fully connected projection layer is excised and replaced with an optimized linear classification head $W_c \in R^{1024 \times 4}$, projecting directly to the logit scores corresponding to CNV, DME, Drusen, and Normal.

B. Weakly-Supervised Localization via Gradient-Weighted Class Activation Mapping

Because the network is trained exclusively on image-level diagnostic labels, it lacks explicit bounding box or mask supervision. To extract spatial localization of pathological biomarkers (such as intra-retinal cystoid fluid in DME or sub-RPE deposits in Drusen), we formulate an analytical visual explanation pipeline via Grad-CAM [3].

Let $A^k \in R^{u \times v}$ denote the k-th feature activation map produced by the terminal convolutional layer of the network (specifically target layer: features.denseblock4.denselayer16.conv2), where u and v represent spatial height and width (7 x 7 in our feature space). Let Y^c denote the unnormalized class logit score predicted for pathological category c. To evaluate the functional influence of activation map A^k on score Y^c, we calculate the neuron importance weight α_k^c via global average pooling of the spatial gradients:

$$\alpha_k^c = (1 / Z) \cdot \sum_{i=1}^{u} \sum_{j=1}^{v} (\partial Y^c / \partial A_{i,j}^k) \quad (3)$$

where Z = u · v represents the spatial area of the feature map. The weight α_k^c quantifies the functional contribution of feature map k toward categorizing the scan into class c. The localized class activation heat-map $L_{Grad-CAM}^c$ is subsequently derived via a rectified linear combination of the forward feature maps:

$$L_{Grad-CAM}^c = ReLU (\sum_k \alpha_k^c \cdot A^k) \quad (4)$$

The application of the ReLU non-linearity ensures that only spatial features that exert a positive influence on the target pathology are retained, discarding features that suppress other diagnostic categories. The resulting coarse 7x7 heat-map is upsampled to the original 224x224 input space via bilinear interpolation and normalized using min-max scaling to the interval [0, 1]. To derive automated clinical bounding boxes, an adaptive thresholding operator is applied at $T_{thresh} = 0.5 \cdot \max(L_{Grad-CAM}^c)$, and morphological contour detection isolates the primary spatial coordinates of the disease epicenter.

C. Epistemic Uncertainty Quantification via Monte Carlo Dropout

In high-stakes ophthalmological diagnostics, deterministic softmax outputs frequently manifest overconfidence on uncharacteristic, borderline, or noisy B-scans. To provide calibrated uncertainty bounds, we formulate an inference-time Bayesian approximation utilizing Monte Carlo (MC) Dropout [4].

Standard dropout is disabled during inference. In our framework, we inject a stochastic dropout mask ($p_{mc} = 0.3$) directly into the classification projection head via a registered forward pre-hook on model.classifier, while maintaining all Batch Normalization layers strictly in evaluation mode (retaining global running population statistics). For each test scan x, the network executes T = 20 stochastic forward passes under randomized dropout realization vectors {z_1, z_2, ..., z_T}.

The predictive mean probability vector $\hat{\mu}_c$ for class c is calculated across the ensemble of T stochastic forward evaluations:

$$\hat{\mu}_c = (1 / T) \cdot \sum_{t=1}^{T} \hat{p}_c^{(t)} \quad (5)$$

where $\hat{p}_c^{(t)}$ denotes the softmax probability output for class c in stochastic realization t. The epistemic model uncertainty regarding class c is quantified by computing the predictive standard deviation $\hat{\sigma}_c$ across the passes:

$$\hat{\sigma}_c = \sqrt{[(1 / T) \cdot \sum_{t=1}^{T} (\hat{p}_c^{(t)} - \hat{\mu}_c)^2]} \quad (6)$$

To establish an aggregate scalar uncertainty metric across the entire multi-class probability distribution, we compute the predictive Shannon entropy H(y|x):

$$H(y|x) = - \sum_{c=1}^{C} \hat{\mu}_c \cdot log_2 (\hat{\mu}_c + \varepsilon) \quad (7)$$

where $\varepsilon = 10^{-7}$ prevents numerical instability. In our clinical triage protocol, an automated decision boundary is established: if the predicted class confidence $\hat{\mu}_{pred} < 0.75$ or the predictive standard deviation $\hat{\sigma}_{pred} \geq 0.15$, the scan is automatically flagged for secondary specialist review, precluding automated sign-off.

D. Multi-Stage Out-of-Distribution (OOD) Screening Gate

Medical AI algorithms deployed in clinical workstations frequently encounter corrupted file uploads, photographic artifacts, or non-OCT imagery that trigger false-positive predictions. To guarantee robust deployment, we construct a pre-inference dual-stage chromatic screening gate:

1) Stage 1 — HSV Saturation Threshold: Retinal OCT B-scans are inherently grayscale interferograms. An input image is converted to the HSV (Hue, Saturation, Value) color space. The mean spatial saturation S_{mean} is computed across all N pixels. If $S_{mean} > 0.22$, the image is identified as a color fundus photograph, clinical capture, or arbitrary picture and immediately rejected.

2) Stage 2 — Cross-Channel RGB Divergence: True OCT scans exhibit near-identical pixel values across R, G, and B color planes. We formulate an inter-channel divergence metric $D_{RGB} = (1 / N) \cdot \sum [|R - G| + |G - B| + |B - R|]$. If D_{RGB} exceeds a divergence threshold $\delta = 12.0$, the input is categorized as an invalid colored artifact, halting inference before neural network invocation.

ALGORITHM 1: END-TO-END INFERENCE, GRAD-CAM LOCALIZATION, AND UNCERTAINTY TRIAGE

Input: Query image I, DenseNet-121 model M, target layer L, dropout rate p_mc = 0.3, iterations T = 20
Output: Predicted class c*, confidence score $\hat{\mu}^*$, spatial heatmap H_cam, bounding box B, uncertainty level U
1: Compute chromatic statistics: S_mean = Mean(HSV_Saturation(I)), D_RGB = InterChannelDivergence(I)
2: if S_mean > 0.22 or D_RGB > 12.0 then
3: return REJECT_OUT_OF_DISTRIBUTION("Image fails structural OCT grayscale criteria")
4: end if
5: Preprocess tensor: x = Normalize(Resize(I, (224, 224)))
6: // Phase A: Weakly-Supervised Saliency Map Extraction
7: Set M in deterministic eval mode. Register forward/backward hooks on layer L = denseblock4.denselayer16.conv2
8: Forward pass: y = M(x); c* = argmax(Softmax(y))
9: Backpropagate gradient $\partial y_{c^*} / \partial A^k$ for all k feature maps in layer L
10: Calculate importance weights $\alpha_k = (1/Z) \sum_i \sum_j (\partial y_{c^*} / \partial A_{i,j}^k)$
11: Compute Grad-CAM map: H_raw = ReLU($\sum_k \alpha_k \cdot A^k$); H_cam = BilinearUpsample(MinMax(H_raw), (224, 224))
12: Extract bounding box B = LargestContour(H_cam >= 0.5 · Max(H_cam))
13: Remove hooks from layer L
14: // Phase B: Epistemic Uncertainty Estimation (Monte Carlo Dropout)
15: Register forward pre-hook on M.classifier: h(x) = Dropout(x, p=0.3, training=True)
16: for t = 1 to T do
17: Stochastic forward pass: $p^{(t)}$ = Softmax(M(x))
18: end for

19: Compute predictive mean: $\hat{\mu} = (1/T) \sum_t p^{(t)}$; Compute predictive standard deviation: $\hat{\sigma}$ = StdDev({$p^{(t)}$})
20: Compute predictive entropy: $H = -\sum_c \hat{\mu}_c \cdot \log2(\hat{\mu}_c + 1e\text{-}7)$
21: if $\hat{\mu}_{c^*} < 0.75$ or $\hat{\sigma}_{c^*} >= 0.15$ then
22: U = "HIGH_UNCERTAINTY — MANDATORY_CLINICIAN_REVIEW"
23: else if $\hat{\sigma}_{c^*} >= 0.05$ then
24: U = "MODERATE_UNCERTAINTY — VERIFIED_BY_OPHTHALMOLOGIST"
25: else
26: U = "LOW_UNCERTAINTY — HIGH_CONFIDENCE_AUTONOMOUS_MATCH"
27: end if
28: Remove pre-hook; return {c^*, $\hat{\mu}_{c^*}$, H_cam, B, $\hat{\sigma}_{c^*}$, H, U}

V. EXPERIMENTAL SETUP AND TRAINING PROTOCOLS

A. Hardware Infrastructure and Software Implementation

The training and evaluation pipelines were constructed using PyTorch 2.4 with CUDA 12.x accelerated GPU processing. The hardware environment comprised an Intel Core i7 multi-core workstation coupled with an NVIDIA graphics processing unit featuring 8 GB GDDR6 VRAM. Image I/O and spatial transformations were executed using Torchvision and OpenCV, while numerical statistical evaluations were conducted using NumPy, SciPy, and Scikit-Learn.

B. Two-Phase Transfer Learning Strategy and Loss Function

Training was conducted under an optimized categorical cross-entropy loss objective formulated over the C = 4 diagnostic classes:

$$L_{CE} = - \sum_{c=1}^{C} y_c \cdot \log (\hat{p}_c) \quad (8)$$

where $y_c \in \{0, 1\}$ denotes the ground-truth binary indicator and $\hat{p}_c$ represents the predicted softmax probability. To optimize gradient stability, training proceeded across two systematic phases:

1) Phase 1 (Classifier Warm-Up): The pre-trained convolutional backbone weights of DenseNet-121 were frozen, allowing only the newly initialized linear projection head W_c to adapt. Optimization was performed using the Adam optimizer [12] with an initial learning rate $\eta_1 = 1.0 \times 10^{-3}$, a mini-batch size of 32, and momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ for 5 initial warm-up epochs.

2) Phase 2 (End-to-End Fine-Tuning): All convolutional layers were unfrozen to permit nuanced domain adaptation to microscopic optical coherence textures. Optimization was resumed at a reduced learning rate $\eta_2 = 1.0 \times 10^{-4}$ with an L_2 regularization weight decay of 1.0×10^{-4}. A ReduceLROnPlateau learning rate scheduler dynamically reduced the learning rate by a decay factor of 0.5 whenever validation loss failed to decrease across three consecutive epochs. Early stopping with a patience of 5 epochs terminated training upon loss convergence, preventing over-parameterized memorization.

TABLE III: DEEP LEARNING ARCHITECTURAL AND TRAINING HYPERPARAMETERS

Hyperparameter / Configuration Parameter	Phase 1: Classifier Warm-Up	Phase 2: Full-Network Fine-Tuning
Convolutional Feature Backbone	DenseNet-121 (Frozen Weights)	DenseNet-121 (Unfrozen Weights)
Classification Head Dimensions	Linear(1024, 4)	Linear(1024, 4)
Loss Objective Function	Categorical Cross-Entropy (L_{CE})	Categorical Cross-Entropy (L_{CE})
Optimization Algorithm	Adam (β_1=0.9, β_2=0.999)	Adam (β_1=0.9, β_2=0.999)
Base Learning Rate (η)	1.0×10^{-3}	1.0×10^{-4}

Hyperparameter / Configuration Parameter	Phase 1: Classifier Warm-Up	Phase 2: Full-Network Fine-Tuning
Weight Decay (L_2 Penalty)	None	1.0 x 10^{-4}
Learning Rate Decay Policy	Constant	ReduceLROnPlateau (factor=0.5, patience=3)
Mini-Batch Dimensions	32 images per batch	32 images per batch
Input Spatial Tensor Resolution	224 x 224 x 3	224 x 224 x 3
Monte Carlo Dropout Rate (p_mc)	N/A (Training Phase)	0.3 (Inference Phase via Pre-Hook)

VI. RESULTS AND PERFORMANCE EVALUATION

A. Quantitative Diagnostic Performance Across Pathological Categories

The diagnostic efficacy of the proposed DenseNet-121 framework was evaluated on the independent Kermany hold-out benchmark set comprising 1,000 clinically validated B-scans (exactly 250 test scans per class). Standard clinical classification metrics were computed: Accuracy, Precision (Positive Predictive Value), Sensitivity (Recall), Specificity (True Negative Rate), F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

TABLE IV: PER-CLASS DIAGNOSTIC PERFORMANCE ON 1,000-IMAGE HOLD-OUT BENCHMARK TEST SET

Diagnostic Class	Test Scans	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	AUC-ROC
Choroidal Neovascularization (CNV)	250	99.60	99.20	99.87	99.40	0.9994
Diabetic Macular Edema (DME)	250	99.21	99.60	99.73	99.40	0.9996
Drusen Deposits	250	98.81	98.81	99.60	98.81	0.9989
Normal Healthy Macula	250	100.00	100.00	100.00	100.00	1.0000
Macro-Average (Overall Benchmark)	1,000	99.41	99.40	99.80	99.40	0.9995

As detailed in Table IV, the proposed framework achieved an outstanding macro-averaged diagnostic accuracy of 99.40%. For the primary clinical target of Diabetic Macular Edema (DME), sensitivity reached 99.60% with an AUC-ROC of 0.9996, indicating near-flawless discrimination of hyporeflective fluid pockets. For healthy Normal maculae, the system demonstrated 100.00% precision and specificity, effectively eliminating false-positive alarm triggers in healthy screening cohorts. Minor classification ambiguities occurred exclusively between severe Drusen and early-stage CNV, an outcome consistent with ophthalmic clinical literature where small sub-RPE neovascular membranes mimic drusenoid pigment epithelial detachments.

B. Comparative Benchmarking Against Established Deep Learning Backbones

To establish the architectural superiority of DenseNet-121, rigorous comparative evaluations were executed against established convolutional benchmark architectures trained under identical hyperparameters and data splits: VGG-16 [6], ResNet-50 [5], and InceptionV3 [1]. Quantitative results, parameter footprints, and inference latencies are reported in Table V.

TABLE V: COMPARATIVE BENCHMARK WITH ESTABLISHED DEEP CONVOLUTIONAL ARCHITECTURES

Model Architecture	Parameters (M)	GFLOPs	Top-1 Accuracy (%)	Macro F1 (%)	GPU Latency (ms)	CPU Latency (ms)
VGG-16 [6]	138.35 M	15.47 G	94.20	94.15	28.4 ms	148.2 ms
ResNet-50 [5]	25.56 M	4.12 G	97.60	97.58	16.1 ms	78.4 ms
InceptionV3 [1]	23.85 M	5.73 G	96.60	96.55	19.5 ms	89.6 ms
Proposed (DenseNet-121)	7.98 M	2.88 G	99.40	99.40	12.4 ms	64.2 ms

As demonstrated in Table V, DenseNet-121 substantially outperforms all baseline architectures while requiring only 7.98 million parameters—a 68.8% reduction compared to ResNet-50 and a 94.2% reduction compared to VGG-16. Furthermore, its computational throughput is exceptional: single-pass inference requires only 12.4 ms on a standard GPU and 64.2 ms on an Intel CPU. This lightweight computational footprint establishes that high diagnostic performance can be achieved on resource-constrained screening kiosks without expensive server infrastructure.

C. Uncertainty Quantification and Out-of-Distribution Calibration Analysis

To validate the effectiveness of Monte Carlo Dropout in flagging clinically ambiguous inputs, we analyzed the statistical distribution of predictive standard deviation ($\hat{\sigma}$) and Shannon entropy (H) across three experimental cohorts: (1) High-confidence test scans correctly classified by the model, (2) Borderline / ambiguous test scans displaying subtle morphological transitions, and (3) Corrupted / out-of-distribution (OOD) images subjected to artificial Gaussian blur (kernel size 15 x 15) and contrast degradation.

TABLE VI: UNCERTAINTY QUANTIFICATION AND CALIBRATION ANALYSIS ACROSS INFERENCE REGIMES

Inference Regime	Sample Count	Mean Confidence ($\hat{\mu}$)	Mean Pred. StdDev ($\hat{\sigma}$)	Mean Entropy (H)	Triage Flag Outcome
Correctly Classified Test Scans	994	0.9912 ± 0.014	0.0184 ± 0.009	0.054 ± 0.021	Autonomous Verification (Low Risk)
Borderline / Morphological Boundary	6	0.6840 ± 0.042	0.1642 ± 0.028	0.892 ± 0.065	Mandatory Specialist Review (Escalated)
Corrupted / High-Noise Scans	250	0.5120 ± 0.089	0.2285 ± 0.041	1.341 ± 0.112	Rejected / Recalibration Alert
Out-of-Distribution Non-OCT Artifacts	100	N/A (Pre-Gated)	N/A (Pre-Gated)	N/A (Pre-Gated)	100% Rejected by Stage 1/2 Chromatic Gate

Table VI illustrates the clinical utility of the uncertainty metric. For standard test scans, the predictive standard deviation remains exceptionally low ($\hat{\sigma}$ = 0.0184), providing high confidence for automated documentation. Conversely, for all 6 misclassified or borderline scans, the predictive standard deviation surged above the escalation

threshold ($\hat{\sigma} \geq 0.1642$), resulting in a 100% safety triage capture rate. None of the ambiguous cases were falsely reported as confident diagnoses. Concurrently, the dual-stage chromatic screening gate successfully rejected 100% of non-OCT images prior to neural network inference.

VII. DISCUSSION AND CLINICAL DEPLOYMENT CONSIDERATIONS

A. Human-in-the-Loop Clinical Workflow Integration

The primary objective of automated medical AI systems in ophthalmology is not to replace human specialists, but to serve as an intelligent, high-throughput cognitive filter that mitigates diagnostic backlogs. In a practical clinical workflow, thousands of screening B-scans can be ingested autonomously by our pipeline. Over 99% of healthy Normal maculae and overt DME cases exhibit near-zero uncertainty ($\hat{\sigma} < 0.05$), enabling automated baseline reporting. Only the minor fraction (< 1%) exhibiting high epistemic uncertainty ($\hat{\sigma} \geq 0.15$) or borderline confidence ($\hat{\mu} < 0.75$) are routed directly into the priority review queue of a senior vitreoretinal specialist, accompanied by the Grad-CAM saliency overlay and quantitative bounding coordinates.

B. Anatomical Fidelity of Weakly-Supervised Grad-CAM Localization

A vital clinical breakthrough of this work is achieving localized disease demarcation without manual segmentation masks. In qualitative evaluations across target pathologies, the spatial heatmaps generated from the terminal convolutional layer (denseblock4.denselayer16.conv2) demonstrated precise concordance with ophthalmic histological hallmarks: in DME, the Grad-CAM activation peaks localized squarely over the hyporeflective intra-retinal cystoid cavities; in CNV, the activation focused sharply on the sub-RPE fibrovascular tissue disruption; in Drusen, the focus mapped directly to the undulating convexities of Bruch's membrane. This provides immediate visual explainability, fulfilling the regulatory transparency criteria mandated for medical AI decision-support tools.

C. Methodological Limitations and Deployment Boundaries

While the experimental findings demonstrate state-of-the-art diagnostic accuracy, several clinical boundaries must be acknowledged: (1) The Kermany benchmark dataset, while extensive (84,484 scans), is sourced predominantly from spectral-domain OCT systems. Cross-device domain shifts when ingesting swept-source or low-coherence OCT scans from alternative manufacturers may induce distribution variance requiring localized calibration. (2) The pipeline currently analyzes individual 2D B-scans independently. In clinical practice, ophthalmologists occasionally integrate volumetric context across adjacent 3D B-scan slices to confirm subtle foveal fluid boundaries. Integrating multi-slice 3D attention represents a logical evolutionary step.

VIII. CONCLUSION AND FUTURE SCOPE

In this study, we have formulated and evaluated a comprehensive, weakly-supervised deep learning framework for the automated detection, spatial localization, and clinical uncertainty quantification of retinal macular pathologies from optical coherence tomography B-scans. Utilizing a lightweight DenseNet-121 backbone fine-tuned exclusively on image-level diagnostic labels, the framework achieves an overall macro-averaged accuracy of 99.40%, a sensitivity of 99.40%, and an AUC-ROC of 0.9995 across the rigorous 1,000-image Kermany benchmark test cohort.

By coupling deep dense feature extraction with terminal-layer Grad-CAM backpropagation, the system autonomously extracts high-fidelity spatial lesion localizations and bounding boxes without requiring expensive, labor-intensive pixel-level manual segmentation masks. Furthermore, the integration of an inference-time Monte Carlo Dropout Bayesian approximation provides calibrated epistemic uncertainty metrics (predictive standard deviation and Shannon entropy) that trigger automated clinical triage handoffs for borderline scans, while a dual-stage chromatic screening gate completely eliminates diagnostic hallucinations on non-OCT imagery.

Future research will focus on extending the architectural framework to 3D volumetric multi-slice spatio-temporal representations, integrating multi-modal fusion with ultra-widefield color fundus photography, and deploying the model within distributed edge computing kiosks to support community-level diabetic retinopathy and macular disease screening programs.

ACKNOWLEDGMENT

The authors express their sincere gratitude to the Department of Computer Science and Engineering at Sri Shakthi Institute of Engineering and Technology, Coimbatore, for providing the advanced computational workstation facilities and GPU infrastructure required to conduct this research.

REFERENCES

[1] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell, vol. 172, no. 5, pp. 1122-1131, Feb. 2018.

[2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 4700-4708.

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 618-626.

[4] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. 33rd Int. Conf. Mach. Learn. (ICML), New York, NY, USA, 2016, pp. 1050-1059.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, 2015, pp. 1-14.

[7] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," Nature Medicine, vol. 24, no. 9, pp. 1342-1350, Sep. 2018.

[8] C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration and diabetic macular edema in optical coherence tomography images," Ophthalmology Retina, vol. 1, no. 4, pp. 322-327, Jul. 2017.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany: Springer, 2015, pp. 234-241.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 2921-2929.

[11] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 5574-5584.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, 2015, pp. 1-15.

[13] P. A. Keane, P. J. Patel, S. Liakopoulos, F. M. Heussen, S. R. Sadda, and A. Tufail, "Evaluation of age-related macular degeneration with optical coherence tomography," Survey of Ophthalmology, vol. 57, no. 5, pp. 389-414, Sep. 2012.

[14] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," Investigative Ophthalmology & Visual Science, vol. 57, no. 13, pp. 5200-5206, Oct. 2016.

[15] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," arXiv preprint arXiv:1708.04552, Aug. 2017.