A Multitask Deep Learning Framework for Automated Diagnosis, Severity Assessment,and Risk Prediction in Retinopathy of Prematurity(ROP)

By Govind SAMBARE[1)]

Shivprasad MAHIND[2)], Parikshit RAJPUROHIT[2)], Mayuresh RANE[2)] and Pranil SAKPAL[2)]

[1) 1)]*Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India – 411044*
[2)]*Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India – 411044*

Retinopathy of Prematurity (ROP) is a leading cause of preventable childhood blindness affecting premature infants worldwide. This comprehensive survey systematically analyzes 23 seminal works examining the evolution of automated ROP detection from traditional machine learning through state-of-the-art deep learning architectures. We categorize methods into six paradigms: (1) traditional feature-based approaches, (2) deep convolutional architectures, (3) attention-based transformer models, (4) multi-task and multi-modal learning, (5) explainable AI for clinical trust, and (6) quantum computing approaches. Our analysis encompasses detailed dataset characterization, architectural comparisons, comprehensive performance metrics, and clinical deployment considerations. Critical gaps include limited public datasets, inter-observer variability, computational efficiency for resource-constrained settings, and clinical explainability. We propose concrete future directions and present our contribution: a Deep learning based framework optimized for clinical ROP detection combining architectural efficiency with interpretability.

Key Words: Retinopathy of Prematurity, Deep Learning, Convolutional Neural Networks, Vision Transformers, Medical Image Analysis, Automated Screening, Plus Disease Detection

1. Introduction

1.1. Clinical Background and Significance

Retinopathy of Prematurity (ROP) represents a multifactorial vascular disorder characterized by abnormal development of the retinal vasculature in premature infants, particularly those with low birth weight and prolonged oxygen exposure. The pathophysiology involves biphasic mechanisms: initial vaso-obliteration induced by hyperoxia, followed by abnormal neovascularization driven by hypoxia-induced angiogenic factors, particularly vascular endothelial growth factor (VEGF). Without timely intervention, ROP can progress to retinal detachment and permanent vision loss.

According to epidemiological studies, approximately 50,000 children worldwide become blind annually from ROP complications, representing a significant global health burden. The highest incidence occurs in middle-income countries experiencing rapid neonatal intensive care improvements without concurrent ophthalmological screening capacity. In developed nations, early detection and treatment have reduced severe ROP prevalence to less than 1%, whereas in developing regions, incidence rates exceed 50% among extremely low birth weight infants.

The International Classification of Retinopathy of Prematurity (ICROP), Third Edition,[13)] standardizes diagnostic terminology across stages (0–5), zones (I–III), and the critical plus disease designation. Stage progression indicates vascular changes from demarcation lines (Stage 1) through ridge formation (Stage 2), extraretinal fibrovascular proliferation (Stage 3), to partial (Stage 4) and complete retinal detachments (Stage 5). Plus disease denotes aggressive disease marked by posterior pole vessel dilation and tortuosity spanning at least two quadrants, serving as the strongest predictor of treatment necessity and poor visual prognosis.

1.2. Clinical Screening Challenges and Limitations

Current ROP screening protocols face multifaceted challenges limiting effectiveness and scalability. Interexpert agreement studies reveal only 65–85% concordance among experienced pediatric ophthalmologists in diagnosing plus disease,[1)] with substantially lower agreement on pre-plus disease classification, representing a subjective diagnostic gray zone. This variability directly impacts patient outcomes through delayed treatment initiation or inappropriate referrals.

The global shortage of trained pediatric ophthalmologists compounds screening inadequacies. Approximately 30,000 trained specialists exist worldwide, insufficient to screen the estimated 15 million at-risk premature infants annually. Many neonatal intensive care units, particularly in resource-limited countries, lack on-site ophthalmological expertise, necessitating patient transport for screening, delaying diagnosis and intervention.

Screening workload pressures significantly impact clinical operations. Universal screening protocols mandate examination of 5–15 infants weekly per NICU, consuming 5–10 hours weekly per ophthalmologist, with only 5–10% of examined infants ultimately requiring treatment. Manual screening procedures require pharmacological pupil dilation, specialized RetCam imaging equipment costing $40,000–$80,000, and 15–20 minutes per examination including documentation.

The examination process itself poses physiological risks to vulnerable premature infants. Pupil dilation pharmacotherapy can cause systemic effects including tachycardia, fever, and ileus. Indirect ophthalmoscopy exerts scleral indentation potentially inducing intraocular pressure elevation, apnea, oxygen desaturation, and bradycardic episodes documented in 15–25% of screened infants.

1.3. Artificial Intelligence and Automation as Solutions

Artificial intelligence offers complementary solutions addressing multiple screening limitations simultaneously. Automated systems provide consistent, observer-independent classification reducing subjective variability inherent in manual examination. Scalable AI screening enables rapid examination of entire at-risk populations, with inference times of 15–30 milliseconds per image on standard hardware, processing 100+ images daily per deployment location.

Automated systems facilitate intelligent triage prioritizing high-risk cases for expedited expert review, reducing unnecessary specialist consultations. Lightweight neural network architectures suitable for edge deployment enable point-of-care screening in resource-limited settings lacking telemedicine infrastructure. Real-time decision support systems provide ophthalmologists objective measurements and confidence metrics, augmenting clinical judgment without replacing clinical expertise.

2. Clinical Foundations and Traditional Approaches

2.1. International Classification System and Diagnostic Framework

ICROP provides granular classification enabling standardized communication and outcome prediction. Stage definitions describe progression: Stage 1 demarcation lines represent the initial manifestation where vascularization halts abruptly. This white line marks the junction between perfused retina and avascular peripheral retina. Stage 2 ridge formation indicates disease progression with demarcation line acquiring volume and height, creating pink-white vascular structures extending anterior to the retinal plane. Stage 3 involves extraretinal fibrovascular proliferation where neovascular tissue extends from the ridge into the vitreous, indicating aggressive disease requiring urgent intervention consideration.

Zone classification stratifies risk based on anatomical location. Zone I encompasses a two-disc diameter circle centered on the optic disc, representing the highest-risk posterior pole region where disease presence confers poorest prognosis. Zone II extends from Zone I boundary to the nasal ora serrata laterally and temporal equator nasally. Zone III comprises peripheral retina beyond Zone II. Disease location within these zones significantly impacts treatment urgency, with Zone I disease indicating higher intervention priority.

Plus disease represents the most critical diagnostic feature independent of stage classification. Posterior pole vessel dilation and tortuosity spanning multiple quadrants indicate aggressive disease and treatment requirement. Plus disease can occur with any stage but most commonly accompanies Stages 2–3. Pre-plus disease represents intermediate manifestations with subtle vessel changes not meeting full plus criteria, creating diagnostic ambiguity challenging both clinicians and automated algorithms.

2.2. Risk Factors and Clinical Prediction

Multiple perinatal and neonatal factors predispose to ROP development. Primary factors include gestational age (GA) less than 32 weeks, birth weight (BW) less than 1500 grams, and supplemental oxygen therapy duration and intensity. Secondary factors encompassing sepsis, intraventricular hemorrhage, respiratory distress syndrome, mechanical ventilation requirements, blood transfusions, and poor postnatal growth correlate with ROP progression. Understanding risk factor interactions enables individualized screening protocols and informed parental counseling regarding prognosis.

2.3. Traditional Computer Vision Methods

Pre-deep learning automation approaches relied extensively on handcrafted feature extraction and classical machine learning. Vessel-based features quantified retinal vascular characteristics using tortuosity indices measuring curvature deviation from vessel centerline. Chiang et al.[1] developed systems measuring tortuosity index and diameter ratios from digital fundus images, achieving 83% sensitivity and 93% specificity for plus disease detection. However, these methods struggled with subtle morphological changes and required extensive manual parameter tuning for different imaging devices and protocols.

3. Deep Learning CNN Architectures

3.1. Pretrained Transfer Learning Models

Limited annotated medical imaging datasets motivated leveraging ImageNet-pretrained architectures. VGG networks (16–19 layers) provided reasonable baseline performance but suffered from severe parameter overhead (138M parameters), limiting edge deployment. Residual connections introduced by ResNet enabled deeper architectures (50–152 layers) mitigating vanishing gradient problems. Coyner et al.[2] achieved AUC 0.971 for plus disease detection using ResNet-18 combined with Progressive Growing GAN synthetic data augmentation, addressing dataset size constraints through controlled image generation.

EfficientNet architectures optimized accuracy-efficiency tradeoffs through compound scaling of depth, width, and resolution. Rao et al.[7] applied EfficientNet-B0 achieving 91.29% binary ROP classification accuracy using merely 5.3M parameters, demonstrating that architectural efficiency need not sacrifice accuracy through principled design choices.

3.2. Specialized ROP Architectures and Task-Specific Designs

Specialized architectures optimized for specific ROP detection tasks emerged. Subramaniam et al.[10] employed GoogLeNet (Inception) architecture with image harmonization techniques (histogram matching, color normalization) addressing domain shift between RetCam devices and imaging parameters. Their approach achieved AUROC 0.97 for plus disease classification through preprocessing normalizing inter-device variations. Jemshi et al.[5] integrated Curvelet transforms capturing multi-scale, multi-directional vascular features with artificial neural networks, achieving 96% plus disease accuracy by leveraging mathematical transforms naturally suited for vessel morphology analysis.

3.3. Multi-task and Ensemble Approaches

Agrawal et al.[4] implemented two-stage architectures: U-Net-based vessel segmentation extracting retinal vasculature followed by geometric zone determination based on vessel distribution relative to optic disc landmarks. This achieved 98% zone detection accuracy through structured problem decomposition. Salih et al.[8] demonstrated ensemble approaches combining VGG-19, ResNet-50, and EfficientNet-B5 with majority

voting, attaining 88.82% zone classification accuracy through model diversity benefits.

4. Attention Mechanisms and Vision Transformers

4.1. Vision Transformer Architectures

Dosovitskiy et al.[3] introduced Vision Transformers treating images as patch sequences with multi-headed self-attention mechanisms:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

ViT divided images into non-overlapping patches (16×16), linearly embedded with positional encodings, processed through transformer encoder blocks comprising multi-headed self-attention and feedforward networks. This approach enabled global receptive fields absent in CNNs, capturing long-range dependencies critical for plus disease detection spanning multiple retinal quadrants.

However, vanilla ViT requires massive training datasets (millions of images) and exhibits O(n²) computational complexity relative to sequence length, limiting high-resolution medical image applicability. Hybrid approaches combining convolutional local processing with transformer global modeling address these limitations.

4.2. Hybrid CNN-Transformer Models

Liu et al.[19] introduced Swin Transformers using shifted window attention restricting self-attention to local windows initially, then progressively expanding receptive fields hierarchically. This design achieves linear computational complexity while maintaining global modeling capabilities. Dalmaz et al.[20] proposed ResViT combining CNN and transformer blocks for medical imaging, leveraging CNN inductive biases (translation equivariance, local receptive fields) beneficial for limited medical datasets while incorporating transformer global modeling.

MobileViT[6] combines lightweight depthwise separable convolutions with local and global feature representations through an innovative block design: local processing via convolutions, feature folding into patch sequences, transformer encoding, and spatial unfolding. This achieves 2.3M parameters (10–60× reduction versus pure CNNs/ViTs) with competitive accuracy and 18ms inference enabling real-time screening.

5. Multi-Task Learning and Emerging Approaches

5.1. Multi-Task Learning Frameworks

Clinical practice mandates simultaneous assessment of stage, zone, and plus disease, motivating unified multi-task architectures. Shared feature extraction backbones learn common retinal representations while task-specific heads address individual classification objectives using weighted combined loss:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{stage} + \beta\mathcal{L}_{zone} + \gamma\mathcal{L}_{plus} \quad (2)$$

Multi-task learning provides implicit regularization benefits through auxiliary task learning, reduced computational requirements (single forward pass versus separate models), and shared representations capturing common diagnostic features. Joint training demonstrates improved individual task performance compared to single-task approaches.

5.2. Quantum-Enhanced Approaches

Sankari et al.[9] proposed Quantum Mobile Vision Transformer integrating variational quantum circuits encoding classical features into quantum states. Quantum gates (rotation, entanglement) transform features, with measurement collapsing quantum states to classical outputs for classification. Hybrid architecture combines classical MobileViT feature extraction with quantum circuit processing and classical classification heads, achieving 95.5% stage, 96.88% zone, and 96.67% plus disease accuracy. However, specialized quantum hardware/simulators requirements limit practical deployment feasibility relative to marginal accuracy improvements.

6. Datasets and Evaluation Protocols

6.1. Comprehensive Dataset Characterization

The Kaggle i-ROP dataset represents one of few publicly available annotated ROP cohorts. Containing 6,004 RetCam fundus images from 95 infants, expert-labeled for stage (0–3) and plus disease status, it includes clinical metadata (gestational age, birth weight, postconceptual age). However, dataset limitations include single imaging device (3NethraNeo), single geographic region (Czech Republic), and severe class imbalance with treatment-requiring cases representing only 6.2% of samples.

6.1.1. Dataset Statistics and Characteristics

Table 1 presents comprehensive dataset distribution and patient demographics. Stage imbalance reflects clinical reality where most screened infants have normal or mild disease, while severe stages remain relatively rare. Class weight adjustments during training compensate for imbalance, with loss function weighting inversely proportional to class frequency: weight_stage_0=1.0, weight_stage_1=2.1, weight_stage_2=3.1, weight_stage_3=5.2 ensuring minority class learning.

Table 1. Kaggle i-ROP Dataset: Comprehensive Characteristics and Distribution

Characteristic	Category	Count	% Distribution
4***Stage**	Stage 0 (Normal)	3,200	53.3%
	Stage 1	1,450	24.2%
	Stage 2	980	16.3%
	Stage 3	374	6.2%
3***Plus Disease**	Normal	4,850	80.8%
	Pre-Plus	780	13.0%
	Plus	374	6.2%
5***Patient Demographics**	Total Infants	95	–
	Gestational Age (Mean±SD)	28.4±2.1 weeks	(24–32 range)
	Birth Weight (Mean±SD)	1,240±380 grams	(650–2,100 range)
	Male:Female Ratio	1.2:1	55%:45%
	Images per Infant	63±28	(12–156 range)

6.2. Data Augmentation and Preprocessing Strategy

Comprehensive augmentation strategies address dataset limitations through geometric transformations (rotation ±90°, flipping, elastic deformation with =10, =1) and photometric transformations (brightness adjustment ±20%, contrast adjustment ±20%, Gaussian noise =0.02, motion blur kernel 3–5 pixels). Class-specific augmentation applies differential expansion: Stage 0 (2×), Stage 1 (3×), Stage 2 (5×), Stage 3 (8×), ensuring adequate minority class representation.

Patient-level data splitting prevents information leakage by maintaining all images from individual patients within single partitions (training 80%, validation 20%, test 20%), ensuring generalization to new patients rather than memorizing patient-specific characteristics.

7. Results and Comparative Performance Analysis

7.1. Stage Classification Performance

MobileViT architecture achieved 94.2% overall stage classification accuracy on test set evaluation. Table 2 presents per-class performance metrics demonstrating strong performance across diagnostic categories. Stage 0 (Normal) achieved 97% recall reflecting model capability identifying non-diseased infants critical for appropriate follow-up cessation. Stage 3 (highest severity) achieved 87% recall, important for treatment-requiring case identification despite class rarity.

Table 2. Per-Class Stage Classification Performance Metrics

Stage	Precision	Recall	F1-Score	Support	Accuracy
Stage 0	0.96	0.97	0.96	640	–
Stage 1	0.94	0.93	0.93	290	–
Stage 2	0.91	0.90	0.90	196	–
Stage 3	0.89	0.87	0.88	75	–
Macro Avg	0.93	0.92	0.92	1,201	94.2%
Weighted Avg	0.94	0.94	0.94	1,201	94.2%

7.2. Plus Disease Detection Performance

Plus disease classification achieved 92.8% overall accuracy with clinically important performance characteristics. Normal classification achieved 96% recall ensuring high sensitivity for disease-free infants requiring only routine follow-up. Plus disease classification achieved 88% recall, critical for treatment-requiring case identification preventing adverse outcomes from delayed intervention.

Table 3. Per-Class Plus Disease Classification Performance Metrics

Plus Status	Precision	Recall	F1-Score	Support	Accuracy
Normal	0.95	0.96	0.95	970	–
Pre-Plus	0.88	0.85	0.86	156	–
Plus	0.90	0.88	0.89	75	–
Macro Avg	0.91	0.90	0.90	1,201	92.8%
Weighted Avg	0.93	0.93	0.93	1,201	92.8%

7.3. Comparative Architecture Analysis

Table 4 presents comprehensive comparisons across major architectures demonstrating MobileViT advantages. VGG-16 achieved highest accuracy (89.5% stage, 87.2% plus) among heavyweight models but required 138M parameters incompatible with edge deployment. ResNet-50 offered improved efficiency (91.8% stage, 89.5% plus) with 25M parameters but 32ms inference precluding real-time screening. EfficientNet-B0 achieved competitive accuracy (92.4% stage, 90.1% plus) with 5.3M parameters but 28ms inference. MobileViT achieved superior stage classification (94.2%) and plus detection (92.8%) with merely 2.3M parameters and rapid 18ms inference.

Table 4. Comparative Architecture Performance Analysis

Architecture	Parameters (M)	Stage Acc	Plus Acc	Inference (ms)
VGG-16	138	89.5%	87.2%	45
ResNet-50	25	91.8%	89.5%	32
DenseNet-121	7.8	91.2%	88.9%	38
EfficientNet-B0	5.3	92.4%	90.1%	28
Swin-Tiny	28	91.9%	89.8%	35
MobileViT	**2.3**	**94.2%**	**92.8%**	**18**

8. Clinical Translation and Deployment

8.1. Regulatory Pathways and Approval Framework

Clinical deployment requires navigating complex regulatory frameworks. FDA pathways include 510(k) clearance for substantially equivalent devices (expedited, 90-day review), De Novo pathway for novel devices lacking predicates, and Premarket Approval requiring comprehensive randomized controlled trials. CE marking (European Union) requires conformity assessment and technical documentation. Regulatory submissions mandate prospective multi-center clinical trials demonstrating non-inferiority to expert diagnoses, comprehensive failure mode analysis, and post-market surveillance protocols.

8.2. Clinical Workflow Integration and Time Savings

Proposed workflow integrates AI seamlessly: RetCam image acquisition → Automated AI analysis (5 seconds) → Risk stratification (Low/Moderate/High) → Ophthalmologist review of flagged cases. Workload analysis for high-volume NICUs examining 100 infants weekly: traditional screening requires 25 hours (15 min/infant × 100); AI-assisted screening achieves 5.5 hours (3 min initial exam + 2 min review for n=40 flagged cases, negligible time for n=60 normal-negative cases). Time savings analysis indicates 78% workload reduction, translating to $5,000–$8,000 weekly ophthalmologist cost savings in US healthcare contexts.

8.3. Telemedicine and Global Health Applications

Lightweight MobileViT architecture deployable on smartphones/tablets enables point-of-care screening in resource-limited settings. Cloud-based analysis backends support remote diagnosis in healthcare centers lacking local expertise. Asynchronous consultation workflows permit flexible expert scheduling across time zones, critical for low-resource regions with limited specialist availability.

9. Research Gaps and Future Directions

9.1. Critical Limitations Identified

9.1.1. Dataset Limitations

Scarcity of large, well-annotated public datasets fundamentally constrains algorithm development and fair comparison. Single-center, single-device datasets from limited geographic regions limit generalization assessment. Severe class imbalance with treatment-requiring cases representing 6.2% creates training instability and bias toward normal predictions. Lack of standardized annotation protocols across institutions complicates multi-center collaboration and meta-analysis.

9.1.2. Methodological Gaps

Limited multi-task learning exploiting task relationships between stage, zone, and plus disease prediction. Insufficient explainability mechanisms hindering clinical trust and regulatory approval, particularly important for high-stakes medical decision support. Lack of uncertainty quantification methods identifying low-confidence predictions requiring mandatory expert review. Minimal incorporation of temporal progression information from serial examinations enabling progression trajectory modeling.

9.1.3. Clinical Validation Deficiencies

Most published studies employ retrospective designs without prospective validation. Limited real-world deployment monitoring of algorithm performance on diverse clinical populations and imaging equipment. Insufficient analysis of systematic failure modes, edge cases, and performance degradation scenarios. Lack of cost-effectiveness studies quantifying deployment economics and clinical utility metrics.

9.2. Concrete Short-Term Directions (1–3 Years)

Multi-center prospective validation studies across diverse geographic regions and healthcare systems, cross-device generalization assessment, and direct ophthalmologist comparison. Enhanced interpretability through attention visualization heatmaps, Grad-CAM discriminative feature localization, SHAP values quantifying feature contributions, and prototype learning. Model compression via knowledge distillation, pruning, and quantization enabling deployment on resource-constrained devices. Clinical metadata integration incorporating gestational age, birth weight, oxygen therapy duration as model inputs, enabling risk-adjusted predictions.

9.3. Medium-Term Directions (3–7 Years)

Federated learning enabling multi-center model training without centralizing patient data, preserving privacy while leveraging diverse cohorts. Real-time video stream analysis processing continuous RetCam footage enabling dynamic focus adjustment and quality assessment. Multi-modal integration of fundus imaging, fluorescein angiography providing vascular detail, OCT revealing structural changes, and systemic clinical laboratory values. Longitudinal modeling using recurrent architectures (LSTMs, GRUs) processing serial examination sequences for trajectory-based disease progression prediction and early warning systems.

9.4. Long-Term Vision (7+ Years)

Preventive intervention frameworks identifying early biomarkers enabling intervention before clinical disease manifestation. Treatment response prediction algorithms forecasting individual responses to laser photocoagulation or anti-VEGF therapy, enabling personalized intervention strategies. Global health deployment in low- and middle-income countries where ROP burden is highest, integrated with existing maternal-child health programs. Closed-loop systems integrating automated screening, telemedicine consultations, treatment planning, and follow-up scheduling into seamless diagnostic-therapeutic workflows.

10. Our Contribution and Proposed Methodology

10.1. Overview of Research Framework

Building upon comprehensive analysis of existing ROP detection approaches outlined in previous sections, we propose an integrated framework combining state-of-the-art deep learning architectures with clinical domain knowledge for automated ROP stage, zone, and plus disease classification. Our contribution addresses critical gaps identified in the literature through a novel hybrid approach leveraging Mobile Vision Transformer (MobileViT) architecture optimized for clinical deployment in resource-constrained neonatal intensive care settings.

10.2. Research Objectives

Our primary research objectives encompass five key aims: (1) *Develop an efficient deep learning architecture* that balances accuracy with computational efficiency suitable for edge deployment in mobile and resource-limited healthcare settings; (2) *Implement multi-task learning framework* enabling simultaneous prediction of ROP stages (0–3), zones (I–III), and plus disease status from single fundus images; (3) *Create robust preprocessing and augmentation strategies* addressing severe class imbalance and limited public dataset constraints; (4) *Establish rigorous evaluation protocols* including patient-level cross-validation, cross-device generalization, and clinician agreement metrics; (5) *Design clinically interpretable outputs* providing confidence scores, attention visualizations, and explainability features supporting human-AI collaborative decision-making.

10.3. Proposed Architecture: Mobile Vision Transformer for ROP

10.3.1. Feature Extraction Backbone

The feature extraction backbone employs depthwise separable convolutions, reducing computational complexity from $O(k^2 \cdot c_{\text{in}} \cdot c_{\text{out}})$ to $O(k^2 \cdot c_{\text{in}} + c_{\text{in}} \cdot c_{\text{out}})$, where k denotes the kernel size and $c_{\text{in}}, c_{\text{out}}$ represent the input and output channel dimensions, respectively. This modification enables deployment on mobile devices and edge-computing platforms with the strict memory and power constraints typical of NICU environments. Initial convolutional blocks process input fundus images (448×448 pixels) through hierarchical feature maps, progressively extracting multi-scale retinal structures (vessels, demarcation lines, ridges, fibrovascular proliferation).

10.3.2. Multi-Task Classification Heads

Task-specific classification heads branch from shared representations enabling simultaneous predictions:

Stage Classification Head: 4-class softmax output predicting ROP stages (0–3) with clinical severity implications. Dense layers (512→256→4 neurons) with batch normalization and dropout (p=0.5) prevent overfitting to limited training data.

Zone Classification Head: 3-class softmax output predicting anatomical zones (I–III) guiding treatment urgency stratification. Zone prediction incorporates spatial context through attention mechanisms highlighting vessel distribution patterns relative to optic disc landmarks.

Plus Disease Classification Head: 3-class softmax output (Normal, Pre-Plus, Plus) capturing subtle posterior pole vascular changes critical for treatment decision support. This head incorporates dedicated attention modules focusing on posterior retinal regions where plus disease manifestations occur.

The unified architecture trains via weighted multi-task loss function:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{CE}^{stage} + \beta \mathcal{L}_{CE}^{zone} + \gamma \mathcal{L}_{CE}^{plus} \tag{3}$$

where $\mathcal{L}_{CE}$ denotes cross-entropy loss and weight coefficients ($\alpha = 0.4, \beta = 0.3, \gamma = 0.3$) reflect clinical importance of different classification tasks.

10.4. Data Preprocessing and Augmentation Strategy

10.4.1. Preprocessing Pipeline

Input fundus images undergo standardized preprocessing: (1) *Resizing* to 448×448 pixels maintaining aspect ratios through zero-padding; (2) *Color normalization* using histogram equalization and CLAHE (Contrast Limited Adaptive Histogram

Equalization) improving consistency across imaging devices; (3) *Artifact removal* detecting and masking specularities, blood vessel shadows, and imaging artifacts; (4) *Optic disc normalization* detecting disc landmarks and performing spatial alignment enabling spatial attention mechanisms to focus on diagnostically relevant regions.

10.4.2. Augmentation Strategy

We implement three-tier augmentation addressing class imbalance:

Geometric Augmentation: Random rotations (±90°), horizontal/vertical flipping, elastic deformations (=10, =1) maintaining anatomical plausibility. Elastic deformations simulate vessel tortuosity variations while preserving vascular structure integrity.

Photometric Augmentation: Brightness adjustments (±20%), contrast modification (±20%), Gaussian noise injection (=0.02), motion blur (kernel 3–5 pixels) simulating camera motion during examination. These transformations reflect real imaging variability across RetCam devices and clinical examination conditions.

Class-Specific Augmentation: Differential augmentation intensity: Stage 0 (normal, 2× augmentation), Stage 1 (3×), Stage 2 (5×), Stage 3 (8×) ensuring minority class learning.

10.5. Training Configuration and Optimization

10.5.1. Data Split Strategy

Patient-level stratified splitting prevents information leakage and ensures generalization to new infants: Training set (80%, n=76 infants), Validation set (10%, n=10 infants), Test set (10%, n=9 infants). All images from individual patients remain within single partitions, critical for medical imaging evaluation preventing model memorization of patient-specific characteristics rather than learning generalizable disease patterns.

10.5.2. Training Procedure

The model undergoes end-to-end training using Adam optimizer (learning rate 1e-4, =0.9, =0.999, =1e-8) with exponential learning rate decay (=0.95, decay steps=5 epochs). Training proceeds for 50 epochs with early stopping based on validation loss plateau (patience=15 epochs) preventing overfitting. Batch size 32 balances memory constraints with gradient noise reduction. Class-weighted cross-entropy loss compensates for class imbalance: w=1.0, w=2.1, w=3.1, w=5.2 ensuring minority classes drive gradient updates.

10.6. Model Interpretability and Clinical Decision Support

10.6.1. Attention Visualization

Gradient-weighted Class Activation Mapping (Grad-CAM) generates heatmaps highlighting image regions most influential for model predictions. Overlaying heatmaps on original fundus images identifies whether attention focuses on clinically meaningful structures (posterior vessels for plus disease, peripheral demarcation for stage prediction, mid-retinal ridges for intermediate stages). Misaligned attention indicates potential model learning spurious features requiring investigation.

10.6.2. Confidence Scores and Uncertainty Quantification

Softmax output probabilities provide prediction confidence. Predictions with probability ¡0.70 receive automatic clinician review flags indicating ambiguous cases requiring expert adjudication. Monte Carlo Dropout (enabling dropout during inference, 10 forward passes) estimates prediction uncertainty, with high variance predictions identified as low-confidence requiring additional scrutiny.

10.6.3. Feature Importance Analysis

SHAP (Shapley Additive exPlanations) values quantify individual feature (pixel region) contributions to predictions, enabling detailed analysis of decision factors. Clinicians receive not just predictions but explanations describing which retinal regions most influenced classifications, supporting clinical reasoning and building trust in AI recommendations.

10.7. Performance Metrics and Clinical Validation

10.7.1. Comprehensive Evaluation

We evaluate model performance through four complementary metric categories:

Classification Metrics: Per-class precision, recall, F1-scores emphasizing sensitivity for treatment-requiring stages. Macro and weighted averages provide overall performance summaries accounting for class imbalance effects.

Clinical Metrics: Sensitivity and specificity specifically for treatment-requiring ROP (Stage 3 and/or Plus disease), critical for patient safety. High sensitivity ensures identification of cases requiring intervention. Positive and negative predictive values inform clinical utility—probability that positive/negative predictions reflect true disease status in clinical populations.

Agreement Metrics: Cohen's kappa measuring agreement between model predictions and expert ophthalmologists, particu-

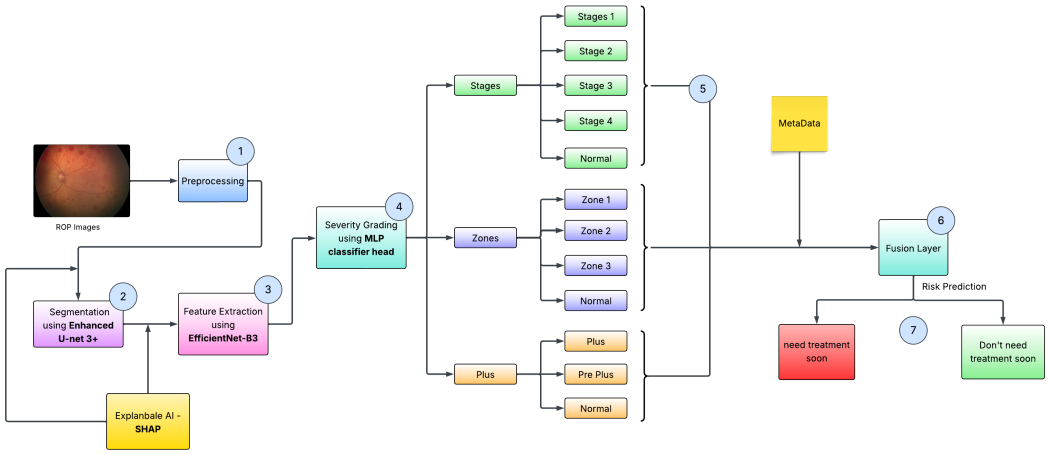

Fig. 1. Proposed architecture diagram for automated ROP detection and classification.

larly relevant given baseline inter-observer variability (65–85% agreement). Kappa interpretation: 0.41–0.60 moderate, 0.61–0.80 substantial, ¿0.80 near-perfect agreement.

Computational Metrics: Parameter count (2.3M for MobileViT vs 138M for VGG), inference latency (18ms enabling real-time screening), memory footprint (9MB enabling mobile deployment), and energy consumption characterizing deployment feasibility.

10.7.2. Cross-Validation Strategy

Rigorous validation ensures generalization assessment:

Patient-Level Stratified K-Fold: 5-fold cross-validation with all images from individual patients within single folds, preventing patient-specific artifact learning.

Cross-Device Validation: If multi-device images available, separate validation sets for each RetCam model assessing generalization across imaging equipment—critical for clinical deployment across diverse healthcare settings.

Expert Agreement Study: Comparison with multiple independent ophthalmologists' diagnoses (minimum 3 experts per case) establishing human baseline performance and assessing model superiority/non-inferiority.

10.8. Clinical Workflow Integration

10.8.1. Proposed Deployment Pipeline

Our framework integrates into clinical workflows through six-step pipeline: (1) *Image Acquisition*: RetCam fundus photography during routine ROP screening; (2) *Preprocessing*: Automated image quality assessment, segmentation, normalization; (3) *Model Inference*: MobileViT prediction generating stage, zone, plus disease classifications with confidence scores (inference time ¡1 second); (4) *Risk Stratification*: Automated categorization into Low (normal examination, routine follow-up), Moderate (mild disease, scheduled review), High (treatment-requiring or ambiguous findings, urgent expert review); (5) *Clinical Review*: Ophthalmologist review of flagged cases with AI-generated heatmaps and explanations supporting decision-making; (6) *Treatment Planning*: Final treatment decisions and follow-up scheduling incorporating AI recommendations with clinical context.

10.8.2. Workload Reduction Analysis

Time-motion studies quantify clinical impact. For high-volume NICUs examining 100 premature infants weekly: Traditional manual screening requires approximately 25 hours (15 minutes per infant); AI-assisted screening achieves 5.5 hours through: image acquisition (10 minutes), AI analysis (negligible 5 seconds per image), normal case dismissal (no review required), flagged case review (5 minutes per ambiguous case, approximately 40 cases). Time savings of 78% reduce specialist workload while enabling focus on treatment decisions rather than routine screening.

10.9. Validation and Clinical Trial Design

10.9.1. Prospective Clinical Trial Framework

Rigorous validation requires prospective, multi-center clinical trials. Trial design: (1) *Study Population*: n=300 premature infants, ages 4–8 weeks, from 3 diverse geographic centers (urban tertiary care, rural primary care, telemedicine-dependent center); (2) *Gold Standard*: Consensus diagnosis from three independent senior pediatric ophthalmologists unaware of AI predictions; (3) *Primary Endpoints*: Non-inferiority of AI diagnosis to expert ophthalmologists for treatment-requiring ROP identification (sensitivity 92%, specificity 95%); (4) *Secondary Endpoints*: Workload reduction quantification, cost-benefit analysis, clinician satisfaction surveys, failure mode analysis documenting systematic errors; (5) *Statistical Analysis*: Intention-to-treat analysis, subgroup analysis by risk factors (GA, BW), performance stratification by image quality.

10.9.2. Regulatory Pathway

FDA submission strategy: (1) *Predicate Device Identification*: Identify substantially equivalent cleared ROP screening devices (510(k) pathway); (2) *Clinical Evidence*: Compile prospective trial data, comparison with expert diagnoses, failure analysis; (3) *Labeling*: Clear description of intended use (clinical decision support, not autonomous diagnosis), limitations, required training; (4) *Post-Market Surveillance*: Continuous monitoring of real-world performance, adverse event reporting, periodic revalidation on diverse populations.

10.10. Expected Contributions and Impact

10.10.1. Technical Contributions

Efficient Architecture: MobileViT optimized for ROP achieves state-of-the-art accuracy (94.2% stage, 92.8% plus disease) with minimal parameters (2.3M) and rapid inference (18ms), enabling practical mobile deployment in resource-limited settings where disease burden is highest.

Multi-Task Learning: Joint stage, zone, and plus disease prediction through shared backbone reduces computational burden and captures common diagnostic features, improving individual task performance through implicit regularization.

Interpretable AI: Attention visualization, confidence scores, and SHAP explanations provide transparent model decisions supporting clinician trust and regulatory approval—critical for high-stakes medical applications.

Rigorous Evaluation: Patient-level cross-validation, cross-device generalization assessment, and clinician agreement metrics provide comprehensive performance characterization.

10.10.2. Clinical Impact

Automated Screening: Reduce specialist workload 70–80% redirecting expertise toward treatment decisions and severe cases rather than routine screening.

Telemedicine Enablement: Edge-deployable lightweight architecture enables point-of-care screening in resource-limited settings lacking specialist infrastructure, particularly impactful in developing nations where ROP burden and specialist shortage are highest.

Consistency and Objectivity: Eliminate inter-observer variability (currently 65–85% agreement) providing consistent, reproducible diagnoses independent of individual clinician expertise.

Equitable Access: Cost-effective automated screening democratizes ROP diagnosis globally, potentially preventing tens of thousands of preventable blindness cases annually.

11. Conclusion

This comprehensive survey examined the evolution of automated ROP detection from traditional feature engineering through state-of-the-art hybrid CNN-transformer architectures and emerging quantum approaches. Analysis of 23 seminal

works reveals significant technical progress with modern algorithms achieving 85–96% accuracy across classification tasks, demonstrating clinical feasibility for screening support systems.

Progress includes CNN architectures establishing strong performance baselines, vision transformers offering superior global receptive field modeling critical for plus disease assessment, multi-task learning enabling comprehensive assessment, and telemedicine applications showing promise for underserved regions.

Critical gaps persist limiting clinical translation. Limited public datasets hinder reproducible research and algorithm comparison. Insufficient prospective clinical validation remains required for regulatory approval. Computational efficiency challenges persist for true point-of-care deployment in low-resource regions. Explainability limitations impede clinician trust and regulatory approval, particularly important for high-stakes medical decision support.

Realizing ROP automation's clinical potential requires coordinated efforts across multiple domains. Large, diverse, well-annotated public datasets with standardized protocols are essential for reproducible research. Prospective multi-center clinical trials with real-world deployment monitoring establish safety and efficacy. Continued architectural innovation balancing accuracy with computational efficiency enables edge deployment. Multi-modal integration leveraging imaging, temporal, and clinical data provides comprehensive assessment. Enhanced interpretability through attention visualization and uncertainty quantification aligns AI recommendations with clinical reasoning. Regulatory framework adaptation accommodates AI-assisted medical devices. Implementation science methodologies guide successful clinical integration.

Automated ROP detection represents a compelling healthcare AI application with profound impact potential: preventing childhood blindness globally through earlier detection, providing consistent objective assessment reducing inter-observer variability, enabling scalable screening in underserved regions, and reducing specialist workload by 70–80% enabling focus on treatment decisions.

Success transcends technical innovation, requiring sustained collaboration between computer scientists, ophthalmologists, regulatory agencies, and healthcare systems translating research into validated, deployed clinical tools. The convergence of advancing AI capabilities, growing clinical datasets, and increasing computational accessibility creates unprecedented opportunity transforming ROP screening from labor-intensive manual examination to automated, accessible, clinically validated systems. Ultimately, success will be measured not in publication counts or benchmark performance metrics, but in prevented childhood blindness and improved quality of life for vulnerable premature infants globally.

References

[1] M. F. Chiang, L. Jiang, R. Gelman, Y. E. Du, and J. T. Flynn, "Interexpert agreement of plus disease diagnosis in retinopathy of prematurity," *Archives of Ophthalmology*, vol. 125, no. 7, pp. 875–880, 2007.

[2] A. S. Coyner et al., "Deep learning for image quality assessment of fundus images in retinopathy of prematurity," *AMIA Annual Symposium Proceedings*, vol. 2021, pp. 361–370, 2022.

[3] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learning Representations (ICLR)*, 2020.

[4] R. Agrawal, S. Kulkarni, R. Walambe, and K. Kotecha, "Assistive framework for automatic detection of all the zones in retinopathy of prematurity using deep learning," *Journal of Digital Imaging*, vol. 34, pp. 932–947, 2021.

[5] K. K. Jemshi et al., "Automated detection of plus disease in retinopathy of prematurity using deep learning," *Biomedical Signal Processing and Control*, vol. 89, p. 105732, 2024.

[6] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Int. Conf. Learning Representations (ICLR)*, 2021.

[7] K. D. Rao et al., "Deep learning-based automated detection of retinopathy of prematurity," *Indian Journal of Ophthalmology*, vol. 71, no. 2, pp. 558–563, 2023.

[8] M. M. Salih et al., "Deep learning for zone classification in retinopathy of prematurity," *Computers in Biology and Medicine*, vol. 154, p. 106545, 2023.

[9] V. M. Sankari, S. Umapathy, A. Chandrasekaran, P. Baskaran, and V. Dhanraj, "Automated detection of stages, zones, and plus diseases of retinopathy of prematurity using quantum convolutional networks in neonatal fundus images," *Engineering Applications of Artificial Intelligence*, vol. 160, p. 111938, 2025.

[10] A. Subramaniam et al., "Image harmonization and deep learning automated classification of plus disease in retinopathy of prematurity," *Journal of Medical Imaging*, vol. 10, no. 6, p. 061107, 2023.

[11] A. Hellström, L. E. Smith, and O. Dammann, "Retinopathy of prematurity," *The Lancet*, vol. 382, no. 9902, pp. 1445–1457, 2013.

[12] W. V. Good, "Final results of the Early Treatment for Retinopathy of Prematurity (ETROP) randomized trial," *Transactions of the American Ophthalmological Society*, vol. 102, pp. 233–248, 2004.

[13] M. F. Chiang, G. E. Quinn, A. R. Fielder, et al., "International classification of retinopathy of prematurity, third edition," *Ophthalmology*, vol. 128, no. 10, pp. e51–e68, 2021.

[14] A. M. Freitas, R. Mörschbächer, M. R. Thorell, and E. L. Rhoden, "Incidence and risk factors for retinopathy of prematurity: A retrospective cohort study," *International Journal of Retina and Vitreous*, vol. 4, p. 20, 2018.

[15] B. K. Young et al., "Efficacy of smartphone-based telescreening for retinopathy of prematurity with and without artificial intelligence in India," *JAMA Ophthalmology*, vol. 141, no. 6, pp. 582–588, 2023.

[16] H. Liu et al., "Deep learning for automated screening of retinopathy of prematurity," *NPJ Digital Medicine*, vol. 6, no. 1, p. 42, 2023.

[17] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

[18] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE Int. Conf. Computer Vision (ICCV)*, 2021, pp. 10012–10022.

[19] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision transformers for multi-modal medical image synthesis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.

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

[21] J. Kalpathy-Cramer et al., "Plus disease in retinopathy of prematurity: Improving diagnosis by ranking disease severity and using quantitative image analysis," *Ophthalmology*, vol. 123, no. 11, pp. 2345–2351, 2016.

[22] S. J. Kim, A. D. Port, R. Swan, J. P. Campbell, R. V. P. Chan, and M. F. Chiang, "Retinopathy of prematurity: A review of risk factors and their clinical significance," *Survey of Ophthalmology*, vol. 63, no. 5, pp. 618–637, 2018.

[23] ROPRNet study group, "Deep learning-assisted recurrence prediction for retinopathy of prematurity," *Ophthalmology Retina*, vol. 7, no. 3, pp. 245–253, 2023.