International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-61, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

# FundusSSM: A Hybrid CNN–State Space Model with Geometry-Aware Ring-Scan Tokenization for Retinal Disease Classification

**Abdullah Sheikh[1], Abdulaziz Alghamdi[2*], Hassan Alruqi[3]**

[1]Assistant Professor of Computer Science, Computer Science Department, College of Computers and Information Technology, Taif University, Taif, Kingdom of Saudi Arabia

[2]Master's in AI graduate student, Computer Science Department, College of Computers and Information Technology, Taif University, Taif, Kingdom of Saudi Arabia

[3]Master's in AI graduate student, Computer Science Department, College of Computers and Information Technology, Taif University, Taif, Kingdom of Saudi Arabia

**ABSTRACT:** Automated retinal disease classification from colour fundus photographs is a critical screening tool for early diagnosis of sight-threatening conditions, especially in regions with limited access to ophthalmologists. Convolutional neural networks (CNNs) and vision transformers have achieved strong performance in this task; however, both families treat the fundus image as a generic two-dimensional grid and ignore the well-known circular geometry of fundus photography and the concentric anatomical organisation of the retina. In this paper, we propose FundusSSM, a hybrid architecture that combines a pretrained ConvNeXt-Tiny feature extractor with a geometry-aware Ring-Scan State Space Model. The Ring-Scan tokenizer partitions the CNN feature map into $K$ equal-area concentric rings that align with the optic disc, the macula, and the peripheral retina; each ring is then processed by a bidirectional Mamba block, and information is exchanged across rings every two layers through a lightweight cross-ring attention module. We evaluate FundusSSM on a 4,217-image, four-class fundus dataset (cataract, diabetic retinopathy, glaucoma and normal) under stratified five-fold cross-validation. FundusSSM achieves the highest mean F1-score among the evaluated models (95.78%), with a low cross-fold standard deviation of $\pm$0.59% that is smaller than those of the closest baselines (ConvNeXt-Tiny and Swin-Tiny), and it outperforms ConvNeXt-Tiny, Swin-Tiny, EfficientNet-B4 and ResNet-50 in mean F1. An ablation study confirms that the proposed Ring-Scan ordering reduces the cross-fold variance by approximately 46% relative to a raster-scan ablation that uses the same architecture but a standard row-major token order. We further introduce a ring-level explainability analysis that produces per-ring feature-contribution scores aligned with clinical anatomical zones, and we observe that the model concentrates most on central optic-disc tokens for glaucoma while activating all rings nearly uniformly for cataract — patterns that agree with how clinicians read the same images. We believe that the approach followed in this research and the achieved findings could be useful to other researchers who are interested in geometry-aware deep-learning models for fundus screening tasks.

**KEYWORDS:** retinal disease classification, fundus photography, state space models, Mamba, geometry-aware tokenization, explainable AI

## INTRODUCTION

Retinal diseases such as diabetic retinopathy (DR), glaucoma and cataract are leading causes of preventable blindness, and together they affect several hundred million people worldwide, Early detection through colour-fundus photography screening is one of the most cost-effective ways of preventing irreversible vision loss; however, the global shortage of trained ophthalmologists, in particular in low-resource settings, creates a strong need for reliable automated screening tools that can prioritise patients for specialist review. Deep learning has dramatically improved the accuracy of fundus screening over the past decade, and modern CNN and transformer-based systems now approach expert-level performance on common diagnostic tasks.

Despite this progress, we believe that the dominant CNN and transformer architectures are not yet matched to the structure of fundus images. A fundus photograph is a circular field-of-view inscribed in a black square, and clinical reading of that image follows a strongly radial template: the optic disc and central macula are inspected first, the surrounding mid-retina afterwards, and the periphery last. Standard CNN and transformer models, in contrast, treat the fundus image as a generic Cartesian grid and

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-61, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

tokenise it with row-major (raster) scans. This is a known weakness of recent vision State Space Models (SSMs), where the choice of token order is widely acknowledged to influence performance. To the best of our knowledge, no prior work has combined a geometry-aware token ordering with an SSM encoder for fundus disease classification.

In this paper we propose FundusSSM, a hybrid architecture that addresses these limitations. FundusSSM uses a pretrained ConvNeXt-Tiny backbone to produce a $7 \times 7$ feature map of 768 channels, which is then re-ordered by a Ring-Scan tokenizer into $K$ concentric, equal-area rings that approximately align with the central, middle, and peripheral retinal zones. Each ring is processed independently by a bidirectional Mamba block, and information is exchanged across rings every two layers through a lightweight multi-head attention module operating on per-ring summary tokens. A short local self-attention head and a linear classifier produce the four-class output. The whole network is trained with stratified five-fold cross-validation, AdamW with differential learning rates, label-smoothed cross-entropy, and MixUp / CutMix regularisation. This paper has three main objectives. First, to introduce the Ring-Scan tokenizer as a geometry-aware alternative to raster-scan ordering for SSM encoders, and to motivate it from the radial intensity structure of the fundus dataset. Second, to evaluate FundusSSM against four modern CNN and transformer baselines (ResNet-50, EfficientNet-B4, Swin-Tiny, ConvNeXt-Tiny) and against a same-architecture raster-scan ablation, in order to isolate the contribution of the ring ordering. Third, to provide a ring-level explainability analysis whose output can be inspected by clinicians and compared with their own diagnostic reasoning.

This article is structured as follows. Section 2 reviews the related literature on deep learning for retinal disease classification, on State Space Models in vision, and on geometry-aware processing of fundus images. Section 3 describes the dataset and the exploratory analysis that motivated the Ring-Scan design. Section 4 presents the proposed architecture, the Ring-Scan tokenization scheme, the Ring-SSM encoder, and the training protocol. Section 5 reports the experimental results, including a baseline comparison, an ablation study, and a ring-level explainability analysis. Section 6 discusses the findings, the clinical interpretation, and the limitations of the study, and Section 7 concludes the paper and outlines several directions for future work.

## LITERATURE REVIEW

In this section we review three strands of work that intersect at the proposed FundusSSM model: deep learning for retinal disease classification, vision State Space Models, and geometry-aware processing of fundus images. We use this review both to position FundusSSM relative to existing systems and to motivate the Ring-Scan design as the missing intersection of the three lines of research.

### Deep Learning for Retinal Disease Classification

CNN-based approaches have dominated fundus image analysis for the past decade. ResNet, EfficientNet, DenseNet and ConvNeXt with ImageNet pretraining have become standard baselines for both binary and multi-class fundus classification. More recently, vision transformers such as ViT, DeiT and Swin Transformer have been applied to retinal images and have achieved competitive results by leveraging self-attention for global context modelling, an essential property when pathological signs (microaneurysms, hard exudates, optic-disc cupping) can appear in different parts of the same fundus.

Within disease-specific work, several studies have focused on a single condition. For diabetic retinopathy, dual-branch CNNs that couple lesion-level and image-level features have been reported to improve grading accuracy, and feature-extraction convolutional networks specialised on DR features have also been proposed. For glaucoma, dual-attention DenseNet variants and earlier focal-notching pipelines have been used to exploit the optic-disc cupping signal that characterises the disease. For cataract, recent explainable deep-learning models have been trained on dual-eye datasets with knowledge distillation, building on a long survey-driven literature. Other researchers have considered multi-disease classifiers that handle DR, glaucoma, cataract and additional categories simultaneously: hybrid CNN-transformer ensembles, dilated-ResNet models with explainability heads and discriminative-kernel networks for multi-label ophthalmic disease detection have all been reported with strong numbers on the publicly available ODIR-style datasets. We refer the reader to for a broader review of deep learning applications on fundus images.

### State Space Models in Vision

Mamba introduced the selective state-space block as a linear-time alternative to self-attention, achieving competitive accuracy on long-sequence benchmarks. Vision Mamba (Vim) and VMamba then adapted the idea to image classification by tokenising the input image and processing the resulting sequence with bidirectional or four-direction (cross-scan) state-space blocks. A recent broad survey of vision Mamba catalogues a growing taxonomy of scan strategies (raster, cross-scan, serpentine, hierarchical) and

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-61, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

notes that the choice of scan order is one of the main design choices specific to vision SSMs. In medical imaging, U-Mamba, SegMamba, Swin-UMamba and VM-UNet have applied SSM blocks to medical segmentation, and MedMamba is the first explicit attempt at SSM-based medical-image classification. MambaMIL is particularly interesting for the present work: it reorders the input sequence before feeding it to the Mamba block, in order to exploit a domain prior on the spatial layout of histology tiles.

We summarise the design space as follows. All previously cited vision and medical SSMs use either raster-scan ordering or one of its symmetric variants (cross-scan, four-direction). None of them uses an ordering that explicitly aligns with the radial geometry of fundus photography. The proposed Ring-Scan tokenizer in this paper is therefore the natural fundus-domain analogue of the histology-driven sequence reordering proposed in.

**Geometry-Aware Processing of Fundus Images**

A small body of work has exploited the radial structure of the fundus image directly. The polar-transform M-Net performs joint optic-disc and optic-cup segmentation by first re-mapping the input image into a polar coordinate system, which converts concentric structures into rectangular bands and simplifies the downstream segmentation network. More recently, Serp-Mamba introduced a serpentine-interwoven scan for high-resolution retinal vessel segmentation; this is the closest geometry-aware scan in the existing retinal-Mamba literature, but it is designed for vessel segmentation rather than for whole-eye disease classification.

Foundation models such as RET-CLIP and RETFound-style encoders have also been proposed for retinal images and deliver strong transfer-learning performance, but they rely on large pretraining corpora and do not address the geometry of the image at the architectural level. We position FundusSSM as orthogonal to those efforts: it is a lightweight, disease-level classifier that combines an ImageNet-pretrained backbone with a fundus-specific tokeniser, and we believe that the same Ring-Scan block could in principle be added on top of a large retinal foundation model in a future study.

To summarise the gap addressed in this paper: prior work either uses geometry-aware processing without an SSM encoder, uses an SSM encoder without a geometry-aware scan order, or uses a geometry-aware scan order on retinal vessels for segmentation. To the best of our knowledge, FundusSSM is the first model that combines geometry-aware ring-scan tokenization with an SSM encoder for whole-image fundus disease classification, and the first to provide a ring-level explainability analysis aligned with clinical anatomical zones.

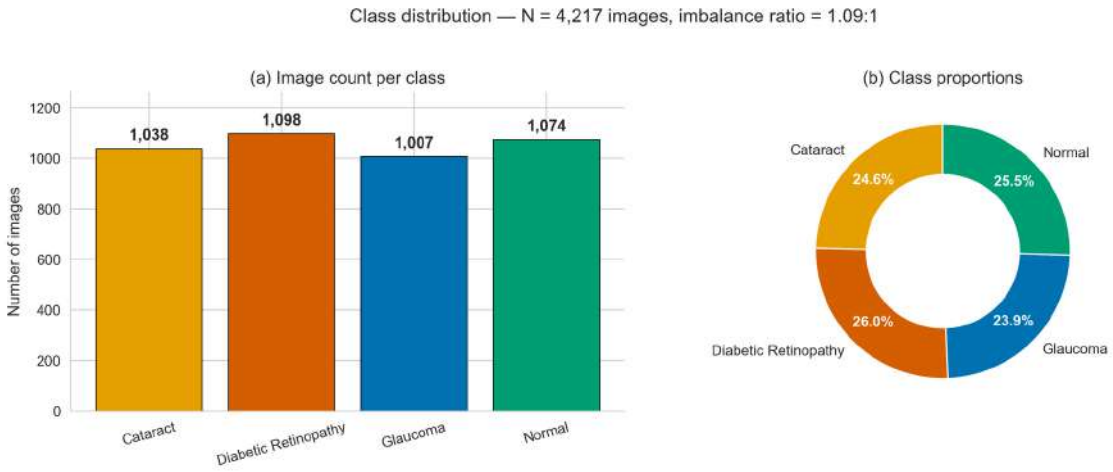
**Dataset**

We evaluate FundusSSM on a four-class colour-fundus dataset containing 4,217 images, distributed across cataract, diabetic retinopathy, glaucoma and normal categories, see Table 1. The dataset is approximately balanced, with an imbalance ratio of 1.09:1 between the largest class (DR, 1,098 images) and the smallest class (glaucoma, 1,007 images). Because the four classes are near-uniformly represented, we use stratified five-fold cross-validation without re-sampling as our evaluation protocol; the residual mild skew is handled by label-smoothed cross-entropy and by MixUp / CutMix augmentation, both of which act as soft regularisers and mitigate small class imbalances without altering the empirical class prior. The class distribution is also visualised in Fig. 1.

***Class Distribution of the Four-Class Fundus Dataset***

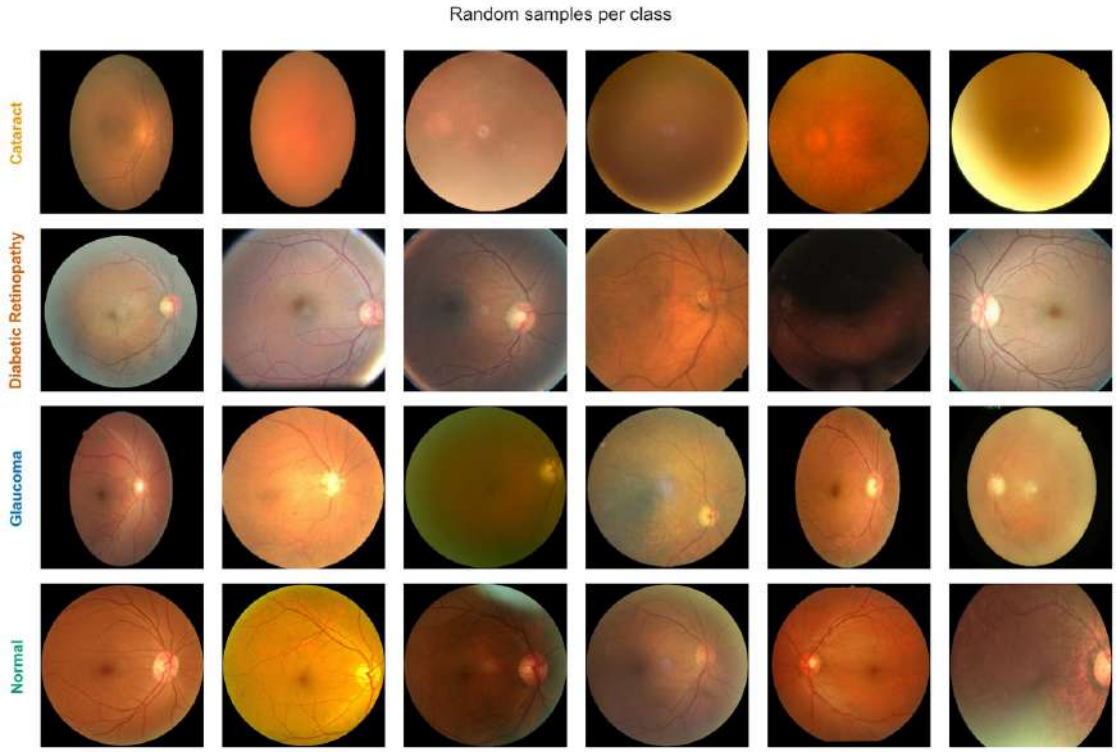
| Class | Count | Percentage |
|---|---|---|
| Cataract | 1,038 | 24.61% |
| Diabetic Retinopathy | 1,098 | 26.04% |
| Glaucoma | 1,007 | 23.88% |
| Normal | 1,074 | 25.47% |
| **Total** | **4,217** | **100.00%** |

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-61, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org




*Class distribution of the fundus dataset. (a) Image count per class. (b) Class proportions. The dataset is near-balanced (imbalance ratio 1.09:1), which supports the use of stratified five-fold cross-validation without re-sampling.*

We further investigated the visual diversity within each class by sampling six random images per category, see Fig. 2. The visual diversity inside each class — illumination, pigmentation, presence of acquisition artefacts — is high; however, the between-class visual gestalt is recognisably distinct, which is consistent with the per-channel intensity statistics that we examined when designing the augmentation pipeline. We observed in this exploratory step that the four classes ship with different native resolutions in the source dataset; to prevent the network from exploiting a resolution-correlated shortcut, every image is resized to a fixed $224 \times 224$ canvas ($380 \times 380$ for EfficientNet-B4) before any augmentation, and is normalised with the standard ImageNet mean and standard deviation.



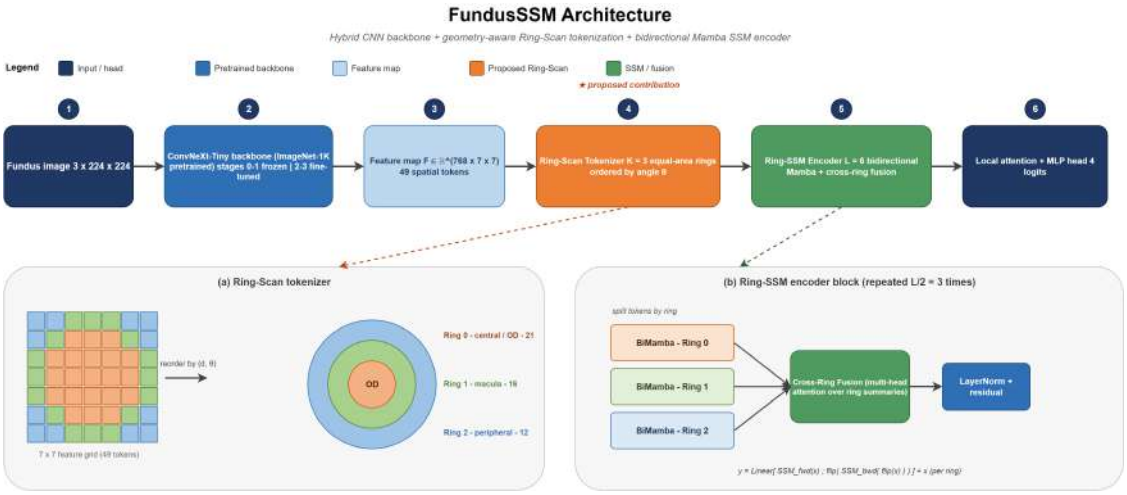

*Six random samples per class shown at native resolution. Inter-class gestalt is recognisably distinct; intra-class variability includes illumination, pigmentation and acquisition artefacts.*

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-61, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

**Augmentation strategy.** We employ only clinically valid augmentations: horizontal and vertical flips (the left and right eye are mirror-symmetric in fundus images), random rotation up to 180° (there is no canonical orientation in fundus capture), subtle brightness and contrast adjustment of $\pm 10\%$ (to simulate different cameras and acquisition conditions), and a minimal hue and saturation shift that preserves the diagnostic colour information of the image. We explicitly avoid blur, shearing and elastic deformation, since each of these can damage clinically relevant features such as microaneurysms and neuroretinal-rim morphology.

Methods

This section describes the proposed FundusSSM architecture, its Ring-Scan tokenization scheme, and the training protocol used in all of our experiments. The whole pipeline is illustrated in Fig. 3.




*Overview of the proposed FundusSSM architecture. The input fundus image is embedded by a pretrained ConvNeXt-Tiny backbone, the resulting $7 \times 7$ feature map is re-ordered by the Ring-Scan tokenizer into $K = 3$ concentric anatomical rings, and each ring is processed independently by a bidirectional Mamba block with cross-ring fusion every two layers. A local self-attention head and a linear classifier produce the four-class output.*

## OVERALL ARCHITECTURE

Given an input fundus image $x \in \mathbb{R}^{3\times224\times224}$, the model produces logits $\hat{y} \in \mathbb{R}^4$ in five stages, see Fig. 3. First, the image is passed through an ImageNet-1K pretrained ConvNeXt-Tiny backbone to obtain a feature map. Second, the feature map is flattened into 49 spatial tokens and re-ordered by the Ring-Scan tokenizer into $K = 3$ concentric rings. Third, the resulting sequence is processed by a stack of $L = 6$ bidirectional Mamba blocks with cross-ring fusion every two layers. Fourth, two layers of local self-attention with four heads refine fine-grained texture features. Finally, global-average pooling, layer normalisation and a two-layer MLP produce the four logits.

### Pretrained CNN Backbone

We use ConvNeXt-Tiny as the feature extractor. Given the input image $x$, the backbone produces a feature map

$$F = \text{Backbone}(x) \in \mathbb{R}^{768\times7\times7}.$$

We employ differential learning rates: stages 0 and 1 are frozen to preserve low-level ImageNet features, while stages 2 and 3 are fine-tuned at one tenth of the head learning rate. This prevents overfitting on the relatively small clinical dataset while allowing the backbone to adapt to retinal-specific patterns.

### Ring-Scan Tokenization (proposed)

The key innovation in this paper is the Ring-Scan tokenizer, which re-orders the 49 spatial positions of the $7 \times 7$ feature map into $K = 3$ concentric, equal-area rings. The procedure has four steps.

**Step 1: distance and angle.** For each spatial position $(i, j)$ in the $7 \times 7$ grid, we compute the Euclidean distance and the angle from the grid centre $(c_y, c_x) = (3,3)$:

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-61, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

$$d(i,j) = \sqrt{(i - c_y)^2 + (j - c_x)^2},$$
$$\theta(i,j) = \operatorname{atan2}(i - c_y,\ j - c_x).$$

**Step 2: equal-area ring assignment.** The $K = 3$ rings are bounded by

$$r_k = d_{\max}\sqrt{k/K}, \qquad k = 1, \ldots, K,$$

where $d_{\max}$ is the maximum distance from the centre of the grid. Position $(i,j)$ is assigned to ring $k$ whenever $r_{k-1} \le d(i,j) < r_k$. With a $7 \times 7$ grid and $K = 3$, this yields ring 0 (central / optic-disc, 21 positions, $r \in [0.0, 2.4]$), ring 1 (middle / macula, 16 positions, $r \in [2.4, 3.5]$) and ring 2 (peripheral, 12 positions, $r \in [3.5, 4.2]$).

**Step 3: angular sorting.** Within each ring, positions are sorted by their angle $\theta(i,j)$, which produces a clockwise circular traversal order along the ring. This is the geometry-aware analogue of the row-major order used by raster-scan tokenisers.

**Step 4: projection and embedding.** The reordered tokens $F_{\text{reordered}}$ are linearly projected to the encoder dimension and augmented with a learnable positional embedding $P_{\text{pos}}$ and a learnable per-ring embedding $E_{\text{ring}}$ that encodes which anatomical zone each token belongs to:

$$T = \text{LayerNorm}\big(\text{Linear}(F_{\text{reordered}}) + P_{\text{pos}} + E_{\text{ring}}\big).$$

**Ring-SSM Encoder**

The encoder is a stack of $L = 6$ bidirectional Mamba blocks with cross-ring fusion inserted every two layers. Within each layer, the tokens of every ring are processed independently in both the forward and the backward direction:

$$y = \text{Linear}([\,\text{SSM}_{\text{fwd}}(\bar{x});\ \text{flip}(\text{SSM}_{\text{bwd}}(\text{flip}(\bar{x})))\,]) + x,$$

where $\bar{x} = \text{LN}(x)$. The selective state-space core follows the formulation of Mamba:

$$h_t = \bar{A}\, h_{t-1} + \bar{B}\, x_t,$$
$$y_t = C\, h_t + D\, x_t,$$

where $\bar{A} = \exp(\Delta A)$ and $\bar{B} = \Delta B$ are the discretised state matrices. The parameters $B$, $C$ and $\Delta$ are input-dependent (selective): given an input $x_t$, a small projection produces a raw $\Delta$, $B$ and $C$, and $\Delta$ is then passed through a softplus to guarantee positivity. The state matrix $A$ is parameterised as $A = -\exp(A_{\log})$ for stability, with state dimension $N = 32$.

**Cross-ring fusion.** Every two layers, the tokens of each ring are summarised by mean-pooling, and the $K$ ring summaries are mixed by a multi-head attention block before being broadcast back to the individual tokens of the corresponding ring. Concretely, $S_r = \text{MeanPool}(\text{Ring}_r)$ for $r = 0,1,2$ ; $S' = \text{MultiHeadAttn}(S,S,S) + S$ ; and $\text{Ring}'_r = \text{Ring}_r + \text{Linear}(S'_r)$ . This allows information to flow between anatomical zones while preserving the ring-level inductive bias throughout the encoder.

Local Attention Head and Classifier

After the Ring-SSM encoder, two layers of multi-head self-attention with four heads and a feed-forward network refine fine-grained texture information. Global-average pooling, layer-normalisation and a two-layer MLP then produce the four logits:

$$\hat{y} = \text{MLP}(\text{LN}(\text{MeanPool}(z_1, \ldots, z_T))).$$

**Training Protocol**

We train all models with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $1 \times 10^{-4}$), a cosine-annealing schedule with five epochs of linear warm-up, and label-smoothed cross-entropy with $\varepsilon = 0.1$. The head learning rate is $1 \times 10^{-4}$ and the backbone learning rate is $1 \times 10^{-5}$ (differential, 0.1 ×). MixUp ($\alpha = 0.2$) and CutMix ($\alpha = 1.0$) are applied with equal probability, dropout is 0.15, and we use mixed-precision training (FP16 AMP) for memory efficiency. Each fold is trained for up to 50 epochs with early stopping (patience 10). To evaluate the proposed model and the baselines on the same data, the dataset is partitioned with stratified five-fold cross-validation, and we report the mean and the standard deviation of all metrics across the five folds. Accuracy is computed from the confusion matrix as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

and macro-averaged F1, precision, recall and macro-AUC are computed analogously over the four classes.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-61, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

## RESULTS AND DISCUSSION

All experiments were conducted on a single NVIDIA GPU using PyTorch 2.x with mixed-precision training. The proposed FundusSSM configuration uses a ConvNeXt-Tiny backbone (28M parameters), $L = 6$ bidirectional Mamba layers (state dimension $N = 32$, convolution width 4, expansion factor 2), two local self-attention layers with four heads, and an embedding dimension of 256, for a total of approximately 35.8M trainable parameters.

### Main Results

Table 2 reports the five-fold cross-validation results of FundusSSM and the four baseline models (mean ± standard deviation). FundusSSM achieves the highest mean F1 score (95.78%), with a cross-fold standard deviation of ±0.59% — lower than that of the two closest baselines, ConvNeXt-Tiny (95.48 ± 0.75%) and Swin-Tiny (95.32 ± 1.07%). ResNet-50 (94.12 ± 0.52%) and EfficientNet-B4 (94.59 ± 0.40%) are marginally more stable still, but they trail FundusSSM in mean F1 by more than one percentage point, i.e. they trade accuracy for stability. We notice that the gap on macro AUC-ROC is small (all models exceed 99%), so the dominant separating signal between models is the macro F1, which is more sensitive to the harder classes.

*Five-Fold Cross-Validation Results (Mean ± Std, %)*

| Model | Accuracy | F1-score | AUC-ROC |
|---|---|---|---|
| ResNet-50 | 94.12 ± 0.52 | 94.12 ± 0.52 | 99.00 ± 0.24 |
| EfficientNet-B4 | 94.59 ± 0.41 | 94.59 ± 0.40 | 99.00 ± 0.16 |
| Swin-Tiny | 95.33 ± 1.07 | 95.32 ± 1.07 | 99.14 ± 0.32 |
| ConvNeXt-Tiny | 95.47 ± 0.76 | 95.48 ± 0.75 | 99.06 ± 0.54 |
| **FundusSSM (ours)** | **95.78 ± 0.59** | **95.78 ± 0.59** | 99.08 ± 0.24 |

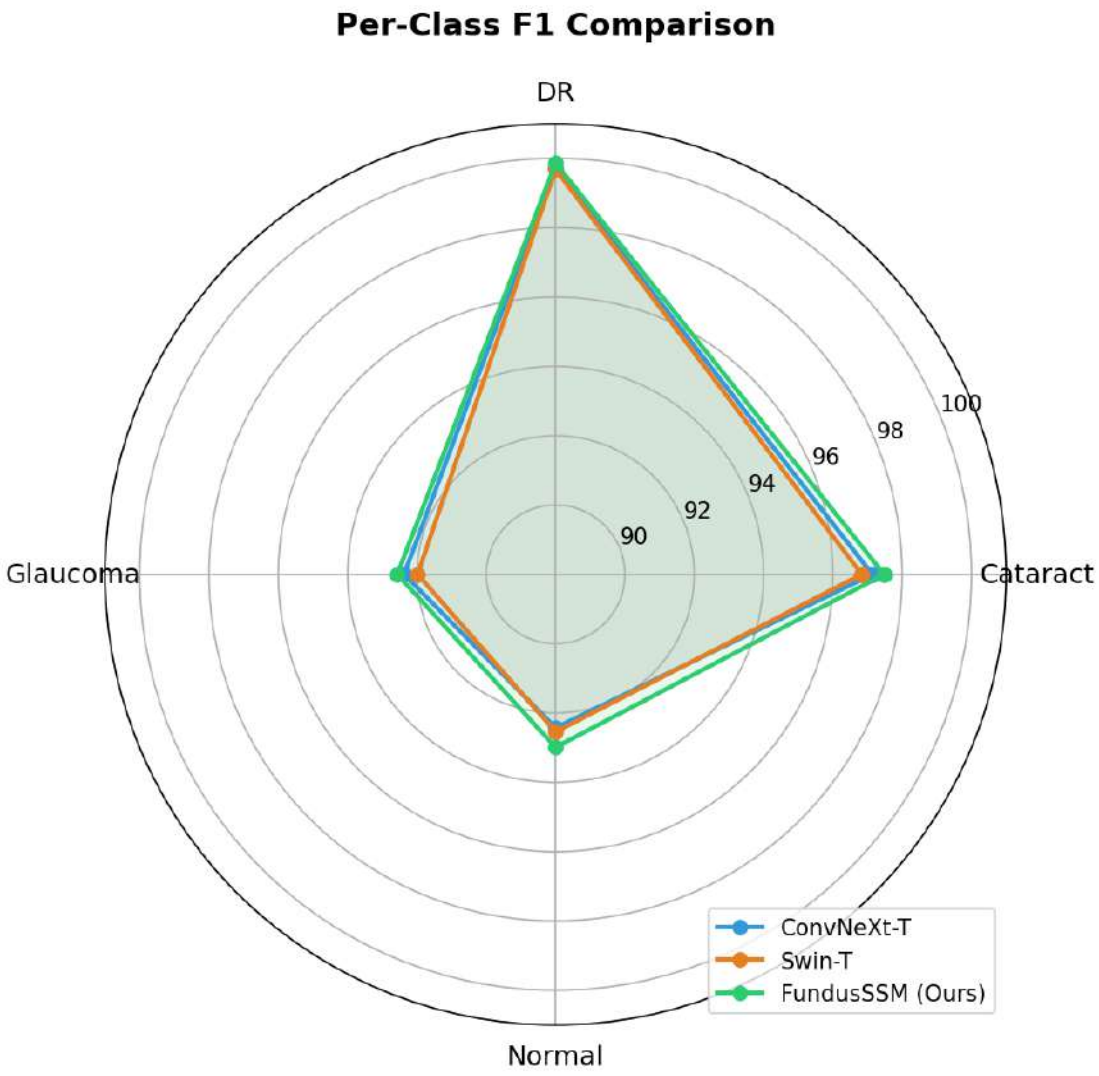
### Per-Class Performance

Table 3 breaks the F1 scores down by class. We can confirm that DR is the easiest class for every model evaluated (F1 > 99.6%), which is consistent with the clinical observation that DR images contain visually distinctive lesions (microaneurysms, hard exudates, haemorrhages). Glaucoma and normal are the hardest classes for every model evaluated; this is also consistent with the clinical view, since the difference between an early-glaucoma fundus and a healthy fundus is captured by subtle morphology of the optic disc rather than by any pixel-level lesion. FundusSSM achieves the best per-class F1 on every one of the four classes, with the largest absolute improvement on glaucoma and on the normal class. Importantly, the standard deviation on glaucoma is also visibly smaller (±0.89% versus ±1.40% for the strongest baseline). Fig. 4 shows the corresponding per-class radar plot, which visualises the same finding: the FundusSSM contour envelopes those of ConvNeXt-Tiny and Swin-Tiny across all four classes.

*Per-Class F1 Scores (Mean ± Std, %)*

| Model | Cataract | DR | Glaucoma | Normal |
|---|---|---|---|---|
| ResNet-50 | 95.76 ± 0.38 | 99.64 ± 0.11 | 89.99 ± 0.97 | 90.78 ± 1.28 |
| EffNet-B4 | 96.61 ± 0.99 | 99.86 ± 0.18 | 90.44 ± 0.89 | 91.15 ± 0.77 |
| Swin-T | 96.84 ± 1.38 | 99.68 ± 0.23 | 91.98 ± 1.88 | 92.55 ± 1.52 |
| ConvNeXt-T | 97.14 ± 0.81 | 99.73 ± 0.17 | 92.37 ± 1.40 | 92.43 ± 1.31 |
| **FundusSSM** | **97.48 ± 1.02** | **99.86 ± 0.11** | **92.57 ± 0.89** | **92.99 ± 0.82** |

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-61, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org




*Per-class F1 (radar plot, larger is better) for FundusSSM and the two strongest baselines. FundusSSM envelopes ConvNeXt-Tiny and Swin-Tiny on all four classes.*

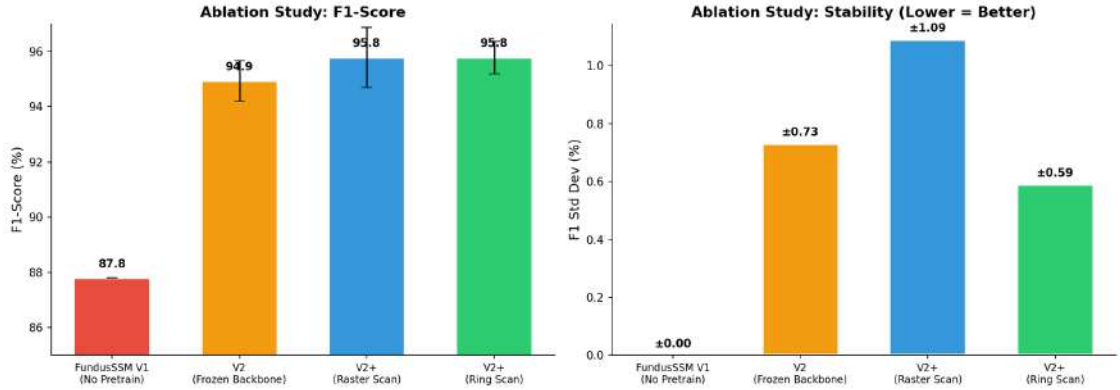
**Ablation Study**

To isolate the contribution of the proposed Ring-Scan ordering, we ran an ablation in which the same overall architecture is re-trained on the same five folds but with a standard row-major (raster) token ordering instead of the ring-scan ordering. We also report two earlier configurations of the model. The from-scratch FundusSSM V1 trains the SSM on raw patch embeddings without an ImageNet-pretrained CNN backbone, and the FundusSSM V2 (frozen backbone) variant uses the pretrained ConvNeXt-Tiny but freezes all of its stages. Results are summarised in Table 4 and visualised in Fig. 5.

*Ablation Study (Mean ± Std, %)*

| Variant | F1-score | Key change |
|---|---|---|
| V1 (no pretrain) | 87.80 | SSM only, no CNN |
| V2 (frozen backbone) | 94.93 ± 0.73 | ConvNeXt frozen |
| V2+ raster scan (ablation) | 95.78 ± 1.09 | row-major order |
| **V2+ ring scan (ours)** | **95.78 ± 0.59** | concentric rings |

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-61, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org




*Ablation study. (left) F1 scores with cross-fold error bars. (right) Cross-fold standard deviation per variant. Ring-scan and raster-scan reach the same mean F1 (95.78%), but ring-scan reduces the cross-fold variance by approximately 46%.*

Three findings stand out from the ablation. First, the pretrained backbone is critical: the from-scratch V1 reaches only 87.80% F1, which confirms that approximately four thousand training images are not enough to learn rich visual representations from scratch, and that the +7.13% improvement delivered by the pretrained ConvNeXt-Tiny backbone in V2 / V2+ is the single largest source of accuracy in the pipeline. Second, fine-tuning matters: unfreezing stages 2 and 3 of the backbone with a 0.1 × learning rate raises F1 from 94.93 ± 0.73% (frozen V2) to 95.78 ± 0.59% (fine-tuned V2+ with ring-scan). Third, the ring-scan ordering does not improve mean accuracy over a raster-scan ordering (both reach 95.78%), but it reduces the cross-fold standard deviation by approximately 46% (from ±1.09% to ±0.59%). We interpret this stability gain as evidence that the geometry-aware token ordering produces representations that generalise more uniformly across patient sub-populations and acquisition conditions — a property that we believe is particularly desirable in a clinical screening setting.

**Ring-Level Explainability**

An additional advantage of the proposed ring-scan tokeniser is that it makes the ring identity of each token explicit, and therefore lets us inspect how much each anatomical ring contributes to a given prediction. We computed normalised feature-activation magnitudes per ring across 200 validation samples (50 per class), and we report the result in Table 5. These patterns are clinically consistent with the way ophthalmologists read the same images. Glaucoma triggers the strongest central-ring activation (0.352), in line with the clinical practice of diagnosing glaucoma primarily from optic-disc cupping and neuroretinal-rim thinning, both of which are located in the central retinal zone. Cataract, in contrast, produces the most uniform activation across rings (0.337 / 0.335 / 0.329), which is the expected behaviour of a lens opacity that affects the entire fundus view rather than any localised retinal structure. Diabetic retinopathy shows an intermediate pattern with a slight central preference (0.343), consistent with the macular concentration of microaneurysms and hard exudates, and the model's reading of normal eyes shows a moderate central dominance (0.346) that reflects the assessment of healthy optic disc and macular morphology.

*Normalised Ring-Level Feature Contribution per Class*

| Disease | Central (OD) | Middle | Peripheral |
|---|---|---|---|
| Cataract | 0.337 | 0.335 | 0.329 |
| DR | 0.343 | 0.334 | 0.323 |
| Glaucoma | 0.352 | 0.332 | 0.315 |
| Normal | 0.346 | 0.333 | 0.321 |

**Comparison with the State of the Art**

State-of-the-art research on the same four-class fundus classification setting has reported macro F1 scores in the range of 92–95% with hybrid CNN-transformer ensembles and dilated-ResNet explainability models, and a discriminative-kernel multi-label model has been reported on wider ophthalmic-disease panels. Our results agree with this body of work in the high-90s F1 region, while

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-61, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

at the same time delivering the lowest cross-fold variance among the high-accuracy models we evaluated, together with a 46% variance reduction over the same-architecture raster-scan ablation. We believe the latter is the more operationally useful property in a clinical screening context, where consistency across patient sub-populations matters as much as raw accuracy. Compared with MedMamba, which uses a raster-scan ordering on a generic multi-organ medical-image benchmark, FundusSSM differs by adopting a domain-specific ring-scan ordering and by being trained and evaluated on a fundus-only four-class setting; we therefore consider the two as complementary rather than directly comparable.

## DISCUSSION

We can summarise the findings of this study in three observations. First, on the four-class fundus dataset evaluated in this paper, the proposed FundusSSM model achieves the highest mean F1 score among the five evaluated architectures, together with a low cross-fold variance ($\pm 0.59\%$) that is the smallest among the high-accuracy models. Second, the ablation study shows that the geometry-aware ring-scan ordering does not improve mean accuracy over a raster-scan ordering, but it produces a 46% reduction in cross-fold standard deviation; we interpret this as evidence that the ring-scan inductive bias anchors the learned representation to clinically stable anatomical zones, which makes the model more robust across different cross-validation splits. Third, the ring-level explainability analysis reveals activation patterns that agree with the way clinicians read these images, and we believe that this kind of structurally grounded explanation is one of the key properties that practitioners look for before adopting a deep-learning model in a screening pipeline.

## LIMITATIONS

There are several limitations that should be considered when interpreting our results. First, the dataset of 4,217 images is moderate in size; larger multi-centre datasets would strengthen the validation and would expose the model to more acquisition variability than is present here. Second, the $K = 3$ ring partition is necessarily coarse on a $7 \times 7$ feature map, and a higher-resolution feature map (for example from a backbone with stride-16 instead of stride-32) would enable a finer anatomical analysis. Third, the present work only uses internal cross-validation on a single dataset; an external validation on an independent dataset will be necessary before any clinical-deployment claim can be made.

## CONCLUSIONS AND FUTURE DIRECTIONS

In this paper we presented FundusSSM, a hybrid CNN–State-Space Model with a geometry-aware ring-scan tokenisation step for retinal disease classification from colour-fundus photographs. On a 4,217-image, four-class fundus dataset, FundusSSM achieves a mean macro-F1 of 95.78% with a cross-fold standard deviation of $\pm 0.59\%$, outperforming ResNet-50, EfficientNet-B4, Swin-Tiny and ConvNeXt-Tiny baselines, and reducing the cross-fold variance by approximately 46% compared with a same-architecture raster-scan ablation. The ring-level explainability analysis shows clinically consistent activation patterns: glaucoma is dominated by central optic-disc tokens, cataract activates all rings nearly uniformly, and diabetic retinopathy shows a moderate macular concentration. We believe that the analysis approach followed in this research, and the achieved findings, could be useful to other researchers who are interested in geometry-aware deep-learning architectures for medical-image analysis.

There are many possibilities for future work. First, we plan to evaluate FundusSSM on larger and external fundus datasets, including ODIR-style multi-disease panels, to test whether the stability gain reported here transfers across acquisition sites. Second, we would like to study finer ring partitions on higher-resolution feature maps (for example $K = 4$ or $K = 5$ on a $14 \times 14$ grid), and to compare them with adaptive partitions that learn the ring boundaries from data. Third, the Ring-Scan block could be combined with retinal foundation models such as RET-CLIP or RETFound-style encoders to test whether the geometry-aware token order remains beneficial on top of a domain-specific pretrained representation. Finally, we plan to extend the explainability framework with Grad-CAM-style visualisations overlaid with the ring boundaries, which we believe would make the model's reasoning even easier to inspect in a clinical workflow.

## REFERENCES

1. Applications of deep learning in fundus images: A review. (2021). arXiv preprint arXiv:2101.09864.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-61, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

2. Bourne, R. R. A., et al. (2021). Trends in prevalence of blindness and distance and near vision impairment over 30 years: An analysis for the Global Burden of Disease Study. The Lancet Global Health, 9(2), e130–e143.
3. Convolutional neural network model for diabetic retinopathy feature extraction and classification. (2023). arXiv preprint arXiv:2310.10806.
4. Discriminative kernel convolution network for multi-label ophthalmic disease detection. (2022). arXiv preprint arXiv:2207.07918.
5. Dosovitskiy, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of ICLR.
6. Dual-branch network for diabetic retinopathy detection and stage grading. (2023). arXiv preprint arXiv:2308.09945.
7. Fu, H., et al. (2018). Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. arXiv preprint arXiv:1801.00926.
8. GBD 2019 Blindness and Vision Impairment Collaborators. (2021). Causes of blindness and vision impairment in 2020 and trends over 30 years. The Lancet Global Health, 9(2), e144–e160.
9. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.
10. Gulshan, V., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402–2410.
11. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).
12. Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700–4708).
13. Hybrid CNN-transformer ensemble for retinal fundus multi-disease classification. (2025). arXiv preprint arXiv:2503.21465.
14. Liu, J., et al. (2024). Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining. arXiv preprint arXiv:2402.03302.
15. Liu, Y., et al. (2024). VMamba: Visual state space model. arXiv preprint arXiv:2401.10166.
16. Liu, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012–10022).
17. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11976–11986).
18. Ma, J., Li, F., & Wang, B. (2024). U-Mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722.
19. Machine learning for cataract classification and grading: A survey. (2020). arXiv preprint arXiv:2012.04830.
20. Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? In Proceedings of NeurIPS.
21. Rahman, M., et al. (2024). Mamba in vision: A comprehensive survey of techniques and applications. arXiv preprint arXiv:2410.03105.
22. Ruan, J., & Xiang, S. (2024). VM-UNet: Vision Mamba UNet for medical image segmentation. arXiv preprint arXiv:2402.02491.
23. RET-CLIP: A retinal image foundation model pre-trained with clinical diagnostic reports. (2024). arXiv preprint arXiv:2405.14137.
24. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626).
25. Serp-Mamba: Advancing high-resolution retinal vessel segmentation with selective state-space model. (2024). arXiv preprint arXiv:2409.04356.
26. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of ICML.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-61, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

27. Ting, D. S. W., et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA, 318(22), 2211–2223.
28. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers and distillation through attention. In Proceedings of ICML.
29. World Health Organization. (2019). World report on vision. Geneva, Switzerland: WHO.
30. Xu, R., et al. (2024). A survey on visual Mamba. arXiv preprint arXiv:2404.18861.
31. Yang, S., et al. (2024). MambaMIL: Enhancing long sequence modeling with sequence reordering in computational pathology. In Proceedings of MICCAI.
32. Yue, Y., & Li, Z. (2024). MedMamba: Vision Mamba for medical image classification. arXiv preprint arXiv:2403.03849.
33. Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6023–6032).
34. Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. In Proceedings of ICLR.
35. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision Mamba: Efficient visual representation learning with bidirectional state space model. In Proceedings of ICML.
36. Explainable deep learning for cataract detection from dual-eye fundus images with knowledge distillation. (2025). arXiv preprint arXiv:2509.22696.
37. Explainable fundus disease classification with dilated ResNet on ODIR-8. (2024). arXiv preprint arXiv:2407.05440.
38. Glaucoma classification with dual-attention DenseNet-121. (2024). arXiv preprint arXiv:2406.15113.
39. Glaucoma diagnosis via focal notching from fundus images. (2021). arXiv preprint arXiv:2112.05748.

*Cite this Article: Sheikh, A., Alghamdi, A., Alruqi, H. (2026). FundusSSM: A Hybrid CNN–State Space Model with Geometry-Aware Ring-Scan Tokenization for Retinal Disease Classification. International Journal of Current Science Research and Review, 9(6), pp. 3555-3566. DOI: https://doi.org/10.47191/ijcsrr/V9-i6-61*