A Multi-Stage Transformer-U-Net and ELM Framework for Early Glaucoma Detection

Ayush Kumar Sahu[1*], **Dr. Rakesh Kumar Khare**[2]
[1*]Shri Shankaracharya Institute of Professional Management and Technology (SSIPMT) [2]Shri Shankaracharya Institute of Professional Management and Technology (SSIPMT)
[1]ayush.sahu25220@ssipmt.com
[2]rk.khare@ssipmt.com

Abstract

Glaucoma is a progressive optic neuropathy and a leading cause of irreversible blind-ness, where early detection is essential for preventing vision loss. However, existing automated approaches often suffer from limitations such as loss of spatial relation-ships, lack of interpretability, and high computational complexity. To address these challenges, this paper proposes a novel multi-stage glaucoma detection framework that integrates Capsule Network (CapsNet)-based deep feature extraction with clinically relevant handcrafted features, unified through an Extreme Learning Machine (ELM) classifier. The preprocessing stage enhances retinal structures using red-channel extrac-tion, contrast-limited adaptive histogram equalization (CLAHE), filtering, and normal-ization. CapsNet effectively preserves spatial dependencies between optic disc and cup regions, while handcrafted features including GLCM texture, color statistics, entropy, and edge information incorporate domain-specific knowledge. The fused feature rep-resentation is efficiently classified using ELM, enabling fast and accurate prediction. Additionally, SHAP-based explainability is employed to provide both global and local interpretability of model decisions. Experimental results on the ORIGA dataset demon-strate that the proposed framework achieves superior performance, with an accuracy of 91.50%, sensitivity of 90.00%, specificity of 91.44%, and AUC of 0.9065, outperform-ing conventional machine learning and deep learning models. These results highlight the effectiveness of combining deep and handcrafted features with an efficient classifier. The proposed approach offers a reliable, interpretable, and computationally efficient solution for automated glaucoma screening in clinical applications.

Keywords— Glaucoma Screening, Capsule Network, Extreme Learning Machine, Hand-crafted Features, SHAP Explainability.

1 Introduction

Glaucoma is a chronic and progressive optic neuropathy that leads to irreversible vision loss if not detected at an early stage. Its early progression is often asymptomatic, making timely diagnosis challenging in conventional clinical workflows. Automated screening systems based on retinal fundus images and optical coherence tomography (OCT) have therefore become a crucial research direction for scalable and accessible glaucoma detection. Existing approaches commonly rely on deep learning techniques, especially U-Net–based architectures for optic disc/cup segmentation and CNN or Transformer models for classification. However, challenges such as inconsistent fundus image quality, variations in optic disc morphology, and limited generalization across datasets continue to hinder clinical adoption.

Recent studies highlight the growing benefits of Transformer-based architectures, which cap-ture long-range dependencies in retinal imagery more effectively than convolution-only mod-els. These models have been successfully applied to both segmentation and classification tasks, often outperforming traditional CNNs due to their ability to encode global structural relationships within the optic nerve head region [1, 2, 3].

Likewise, Extreme Learning Machines (ELM) have demonstrated strong potential for fast and accurate classification, especially when combined with deep feature extractors or evolutionary optimization strategies [4, 5, 6].

2 Literature Review

The literature on glaucoma detection spans classical machine learning, CNN-based deep learning, U-Net derivatives for segmentation, and more recently Transformer-based architec-tures. Below is a synthesis of key themes supported by the retrieved evidence.

Early works often combined handcrafted features such as wavelet coefficients, gradient his-tograms, or HOG with machine learning classifiers like SVM or ELM. ELM-based approaches have been particularly attractive due to their fast learning capabilities. For example, ELM was used for early identification and categorization of healthy vs. glaucomatous retinal images using wavelet and histogram features [4], and more recent works show ELM in-tegrated with optimization methods achieving superior accuracy compared to traditional classifiers [5, 7].

U-Net remains the dominant architecture for optic disc and cup segmentation. Variants such as Attention U-Net, R2U-Net, Improved U-Net, and U-Net++ have been widely explored. Studies demonstrate U-Net's effectiveness in extracting OD/OC regions for downstream classification [8] and improved segmentation accuracy when augmented with enhancements such as Capsule Networks or EfficientNet encoders [9, 10]. Transformer-integrated U-Net

derivatives (e.g., Swin-UNet) have also achieved high segmentation performance [11].

Transformers and hybrid Transformer-CNN models have recently shown substantial improve-ments in capturing global retinal structures. Vision Transformers (ViT), Swin Transformers, and Transformer-U-Net hybrids have been evaluated for glaucoma classification and segmen-tation [1, 2, 3]. These models outperform CNN-only baselines by better modeling optic nerve head morphology and long-range spatial patterns.

Recent systems adopt multi-stage pipelines integrating segmentation, feature extraction, and classification. Examples include multi-stage CNN fusion models [12], segmentation-assisted Transformer diagnosis pipelines [2], and frameworks combining segmentation-derived vCDR with classifier predictions [13, 14]. Ensemble models using CNNs and Transformers further enhance robustness [15, 16].

Table 1: Comparative Survey of Glaucoma Detection Methods

Author (Year)	Methodology / Implementation	Dataset	Metrics	Key Contribution	Limitation / Gap
Sreng et al. (2020) [1]	DeepLabv3+ for OD segmentation + CNN classifier (multi-stage pipeline)	REFUGE, RIM-ONE, ORIGA, DRISHTI-GS1	Acc: 95.59%, AUC: 95.10%	Robust multi-dataset evaluation with segmentation-classification pipeline	No attention/transformer modeling; limited global context
Ganesh et al. (2021) [2]	Context-aware Y-Net (joint segmentation + classification) with ensemble learning	Multiple fundus datasets	Acc: ~99.3%	Unified segmentation and classification framework	High complexity; lacks lightweight deployment
Oh et al. (2021) [3]	Explainable ML using OCT/fundus features + classical classifiers	Clinical dataset	AUC-based evaluation	Introduces explainability in glaucoma diagnosis	Not end-to-end; no deep feature learning
Hemelings et al. (2021) [4]	CNN-based full fundus classification (global feature learning)	Fundus datasets	AUC-based	Shows glaucoma cues beyond optic disc region	No segmentation guidance
Sudhan et al. (2022) [5]	U-Net for cup segmentation + DenseNet-201 + DCNN classifier	ORIGA	Acc: 96.90%	Effective multi-stage deep learning pipeline	CNN-only; lacks global attention modeling
Hussain & Basak (2023) [6]	UT-Net (Transformer + U-Net hybrid) with attention fusion	DRISHTI-GS, REFUGE, RIM-ONE	PCC ≈ 0.88–0.91	Integrates transformer for global contextual learning	No optimized classification stage
Chen et al. (2024) [7]	Improved U-Net with residual + attention modules	DRISHTI-GS1	Dice, Sensitivity ↑	Enhances segmentation accuracy	No classification module
Muduli et al. (2024) [8]	Feature extraction + optimized ELM (Pelican optimization)	Fundus datasets	Acc: ~97.7%	Efficient lightweight ELM classifier	No deep feature fusion or segmentation
Purwanti et al. (2025) [9]	Histogram + fractal features + ELM classifier	Fundus dataset	Acc: 98.89%	Lightweight classification using ELM	Handcrafted features; no deep learning
Alasmari et al. (2025) [10]	Multi-stage: YOLO OD detection + segmentation + Transformer classification	REFUGE, ORIGA, G1020	Acc, AUC ↑	Combines detection, segmentation, transformer-based classification	High complexity; no lightweight classifier
Sahu & Khare (2025) [11]	Transformer-enhanced U-Net + ELM classifier	Fundus dataset	High accuracy reported	Hybrid deep + ELM framework	Lacks multi-stage optimization and feature selection

Existing glaucoma detection methods evolved from CNN-based U-Net pipelines to Transformer-enhanced architectures for improved contextual learning. Recent works intro-duce ELM for efficient classification; however, no unified framework effectively combines Transformer-U-Net feature extraction with optimized ELM classification, highlighting a key research gap.

3 Proposed Method

We propose a hybrid glaucoma detection framework that integrates deep capsule-based fea-ture learning with handcrafted descriptors, unified through an Extreme Learning Machine (ELM) classifier and interpreted using SHapley Additive exPlanations (SHAP) [1]. The framework follows a multi-stage pipeline to ensure accuracy, efficiency, and interpretability.

In the first stage, fundus images are preprocessed using red-channel extraction, contrast-limited adaptive histogram equalization (CLAHE), median filtering, and shade correction. These steps reduce illumination variations and enhance key anatomical structures such as the optic disc and optic cup.

In the second stage, a Capsule Network (CapsNet) is employed for deep feature extraction. Unlike conventional convolutional neural networks, CapsNet preserves spatial relationships by encoding both the presence and pose of retinal structures into vector representations, enabling better modeling of optic disc–cup dependencies.

The third stage extracts handcrafted features to complement deep representations. These include texture features derived from the Grey-Level Co-occurrence Matrix (GLCM), color statistics from RGB channels, Shannon entropy, and Sobel edge magnitude. These features incorporate clinical knowledge and improve robustness, particularly in limited data scenarios.

In the fourth stage, deep and handcrafted features are fused into a unified feature vector. Min–max normalization is applied to scale all features to a common range, ensuring balanced contribution from heterogeneous feature sources.

In the fifth stage, classification is performed using an Extreme Learning Machine (ELM), a single-hidden-layer feedforward network in which output weights are computed analyti-cally. This non-iterative learning mechanism enables fast training, reduced computational complexity, and stable performance.

Finally, SHAP is employed for interpretability by quantifying the contribution of each feature to the prediction. This provides both global and local explanations, ensuring transparency and alignment with clinical understanding.

3.1 Motivation

Glaucoma is a leading cause of irreversible blindness worldwide, affecting over 80 million individuals, with prevalence projected to reach 111.8 million by 2040 [17]. Early detection through fundus imaging is essential; however, manual diagnosis is limited by inter-observer variability, scalability issues, and high cost. Existing deep learning approaches exhibit several limitations. First, purely data-driven models neglect clinical priors such as texture, color,

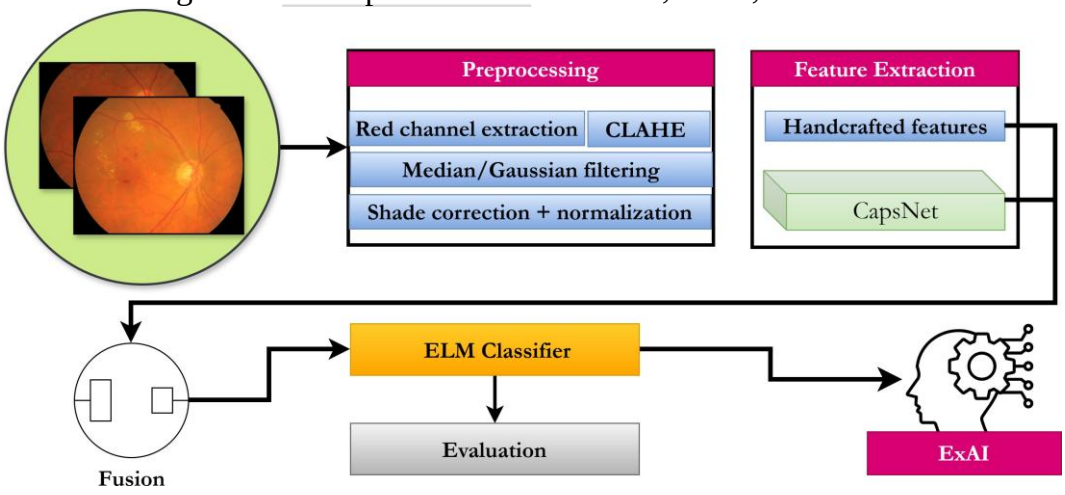

Figure 1: Proposed Flow

and structural biomarkers. Second, conventional CNNs lose spatial relationships between the optic disc and cup due to pooling operations, affecting geometric feature representation. Third, relying solely on deep or handcrafted features leads to incomplete representation of glaucoma pathology. Finally, deep models are computationally expensive and lack inter-pretability.

To overcome these challenges, the proposed framework integrates handcrafted features with Capsule Network-based deep features, enabling complementary representation, efficient clas-sification using Extreme Learning Machine, and transparent predictions through SHAP-based interpretability.

3.2 Problem Formulation

Notation

Let $D = \{(I_i, y_i)\}_{i=1}^{N}$ denote a dataset of N colour fundus images, where $I_i \in R^{H \times W \times 3}$ is the i-th image and $y_i \in \{0, 1\}$ is the corresponding label (0: non-glaucoma, 1: glaucoma). The dataset is partitioned into training and testing subsets:

$$D = D_{train} \cup D_{test}, \quad D_{train} \cap D_{test} = \emptyset, \quad |D_{train}| = N_{tr}, \quad |D_{test}| = N_{te}. \tag{1}$$

The split is stratified so that the class prior $P(y = 1)$ is preserved in both subsets.

Objective

We seek a composite mapping $\mathsf{F} : \mathrm{R}^{H\times W\times 3} \rightarrow \{0, 1\}$ that minimises the empirical risk:

$$\mathsf{F}^{*} = \arg\min_{\mathsf{F}} \frac{1}{N_{\text{te}}} \sum_{i=1}^{N_{\text{te}}} \mathsf{L}\left(\mathsf{F}(\boldsymbol{I}_i), y_i\right), \tag{2}$$

where L is the cross-entropy loss, and F is decomposed as:

$$\mathsf{F} = \underbrace{f_{\text{ELM}}}_{\text{classifier}} \circ \underbrace{\sigma_{\text{norm}}}_{\text{normalisation}} \circ \underbrace{\phi_{\text{caps}} \,\|\, \phi_{\text{hc}}}_{\text{fusion}}, \tag{3}$$

where $\phi_{\text{caps}} : \mathrm{R}^{H\times W\times 3} \rightarrow \mathrm{R}^{d_c}$ is the CapsNet feature extractor, $\phi_{\text{hc}} : \mathrm{R}^{H\times W\times 3} \rightarrow \mathrm{R}^{d_h}$ is the handcrafted feature extractor, $\|$ denotes concatenation, σ_{norm} is min–max normalisation, and $f_{\text{ELM}} : \mathrm{R}^{d_c+d_h} \rightarrow \{0, 1\}$ is the ELM classifier.

Explainability Objective

For any input $\boldsymbol{I}_i$, we additionally compute a SHAP attribution vector $\boldsymbol{\psi}_i \in \mathrm{R}^{d_c+d_h}$ satisfying the efficiency property:

$$f_{\text{ELM}}(\boldsymbol{x}_i) - \mathrm{E}[f_{\text{ELM}}(\boldsymbol{x})] = \sum_{j=1}^{d_c+d_h} \psi_{i,j}, \tag{4}$$

where $\boldsymbol{x}_i = \sigma_{\text{norm}}\,[\phi_{\text{caps}}(\boldsymbol{I}_i) \| \phi_{\text{hc}}(\boldsymbol{I}_i)]$ is the normalised fused feature vector. Each $\psi_{i,j}$ quantifies the marginal contribution of feature j to the prediction for sample i.

3.3 Proposed Method

Image Preprocessing

Each raw fundus image $\boldsymbol{I} \in \mathrm{R}^{H_0\times W_0\times 3}$ (BGR colour space) undergoes a five-stage preprocessing pipeline to yield $\hat{\boldsymbol{I}} \in \mathrm{R}^{224\times 224\times 3}$:

(a) Red-Channel Extraction. The red channel $R = \boldsymbol{I}[:, :, 2]$ is isolated because it provides the highest contrast between the optic cup and neuroretinal rim in fundus images, owing to the absorption spectrum of haemoglobin.

(b) CLAHE. Contrast-Limited Adaptive Histogram Equalisation is applied to R with clip limit $\tau = 2.0$ and tile grid 8×8:

$$R_{\text{clahe}} = \text{CLAHE}(R;\ \tau = 2.0,\ \text{grid} = 8 \times 8). \tag{5}$$

CLAHE enhances local contrast without amplifying noise, making subtle cup/disc boundaries more prominent.

(c) Median Filtering. A 5×5 median filter removes salt-and-pepper noise while preserving edges:

$$R_{\text{filt}} = \text{MedianFilter}(R_{\text{clahe}};\ k = 5). \tag{6}$$

(d) Shade Correction. Non-uniform illumination is estimated via Gaussian smoothing and subtracted:

$$R_{\text{corr}} = R_{\text{filt}} - \underbrace{\text{GaussianBlur}(R_{\text{filt}};\ \sigma = 61)}_{\text{background estimate}}. \tag{7}$$

The corrected channel is then intensity-normalised to $[0, 255]$.

(e) Resizing. The corrected red channel replaces the original red channel in $\boldsymbol{I}$, and the resulting image is resized to 224×224 via bilinear interpolation. Finally, pixel values are scaled to $[0, 1]$:

$$\hat{\boldsymbol{I}} = \frac{1}{255}\text{Resize}\ \boldsymbol{I}_{\text{corrected}},\ 224 \times 224\ \in [0, 1]^{224 \times 224 \times 3}. \tag{8}$$

Deep Feature Extraction via CapsNet

Architecture. The Capsule Network feature extractor comprises the following layers:

(i) **Conv1:** 64 filters, 9×9 kernel, ReLU activation, valid padding.

(ii) **Conv2:** 128 filters, 9×9 kernel, stride 2, ReLU activation, valid padding.

(iii) **PrimaryCaps:** Reshape the Conv2 output into capsule vectors of dimension $d = 16$, yielding a set of primary capsules $\{\boldsymbol{u}_j\}_{j=1}^{M}$, where M depends on the spatial dimensions after convolution.

(iv) **Squash Activation:** Each capsule vector is normalised via the squash function:

$$\text{squash}(\boldsymbol{s}) = \frac{\|\boldsymbol{s}\|^2}{1 + \|\boldsymbol{s}\|_2} \cdot \frac{\boldsymbol{s}}{\|\boldsymbol{s}\| + \epsilon'} \tag{9}$$

where $\epsilon = 10^{-7}$ prevents division by zero. The squash function ensures the output vector's length represents the probability of entity presence (range $[0, 1)$) while its orientation encodes instantiation parameters.

(v) **Routing & Reduction:** A learned linear mapping reduces the M primary capsules to $K = 8$ output capsules, each of dimension $d = 16$, followed by a second squash activation.

(vi) **Flatten:** The K output capsule vectors are concatenated into a single feature vector:

$$\phi_{\text{caps}}(\hat{\boldsymbol{I}}) = \text{flatten}\ [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K] \in \mathrm{R}^{d_c}, \quad d_c = K \times d = 128. \tag{10}$$

Training. A temporary classification head (Dense-64-ReLU → Dropout-0.3 → Dense-2-Softmax) is appended to the CapsNet base. The combined model is trained on $\mathsf{D}_{\text{train}}$ for $E = 20$ epochs with the Adam optimiser (learning rate $\eta = 10^{-4}$) and categorical cross-entropy loss:

$$\mathsf{L}_{\text{cls}} = -\sum_{i=1}^{N_{\text{tr}}} \sum_{c=0}^{1} y_{i,c} \log \hat{y}_{i,c}, \tag{11}$$

where $\hat{y}_{i,c}$ is the softmax probability for class c. After training, the classification head is discarded and only the base CapsNet is retained for feature extraction on *all* images.

Handcrafted Feature Extraction

For each preprocessed image $\hat{\boldsymbol{I}}$, twelve clinically motivated features are computed from its luminance channel $L = 0.299R + 0.587G + 0.114B$:

(a) GLCM Texture Features (4 features). A Grey-Level Co-occurrence Matrix $\boldsymbol{G} \in \mathrm{R}^{Q \times Q}$ (with $Q = 16$ quantisation levels, horizontal adjacency $\Delta x = 1$, $\Delta y = 0$, symmetrised) is computed, from which:

$$\text{Contrast} = \sum_{i,j} G_{ij}(i-j)^2, \tag{12}$$

$$\text{Homogeneity} = \sum_{i,j} \frac{G_{ij}}{1 + (i-j)^2}, \tag{13}$$

$$\text{Correlation} = \frac{\sum_{i,j} G_{ij}(i-\mu_i)(j-\mu_j)}{\sigma_i \sigma_j}, \tag{14}$$

$$\text{Energy} = \sum_{i,j} G_{ij}^2, \tag{15}$$

where μ_i, μ_j, σ_i, σ_j are the marginal means and standard deviations of $\boldsymbol{G}$.

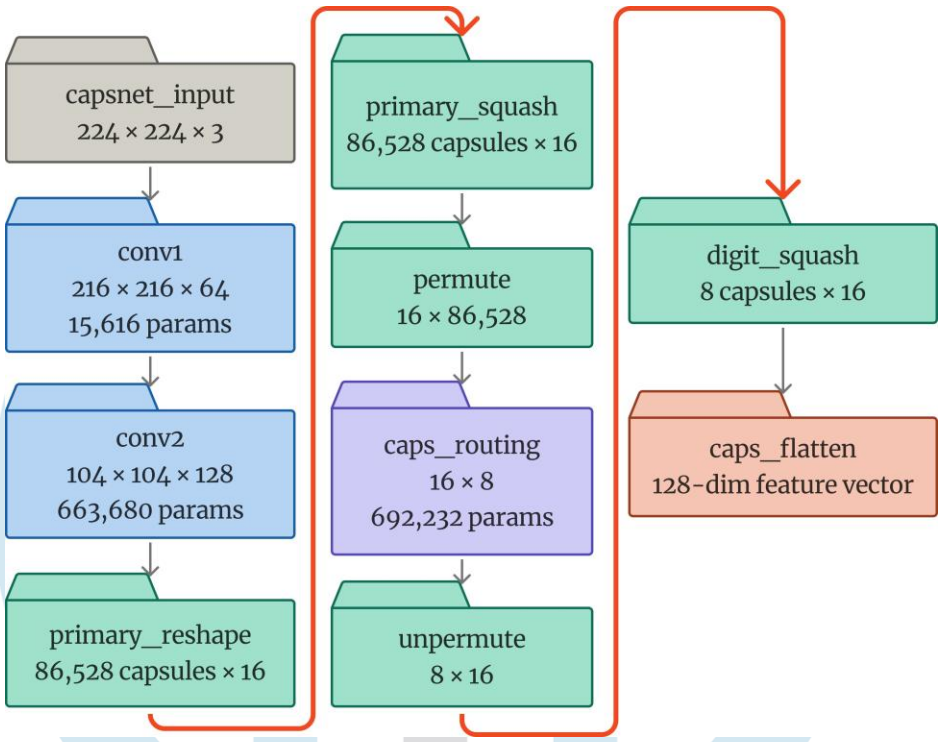

Figure 2: CapsNet Architecture

(b) Colour Statistics (6 features). Per-channel mean and standard deviation capture global colour distribution:

$$\phi_{\text{colour}} = \mu_R, \mu_G, \mu_B, \sigma_R, \sigma_G, \sigma_B \ . \tag{16}$$

(c) Shannon Entropy (1 feature). The histogram $\{p_k\}_{k=0}^{255}$ of the luminance channel yields:

$$H = -\sum_{k=0}^{255} p_k \log(p_k + \epsilon). \tag{17}$$

(d) Edge Magnitude (1 feature). The mean Sobel gradient magnitude captures structural edge density:

$$E_{\text{edge}} = \frac{1}{HW} \sum_{x,y} \sqrt{(S_x * L)^2 + (S_y * L)^2}, \tag{18}$$

where S_x, S_y are horizontal and vertical Sobel kernels and $*$ denotes convolution.

The complete handcrafted vector is:

$$\phi_{\text{hc}}(\hat{I}) = [\text{Contrast, Homogeneity, Correlation, Energy, } \mu_R, \mu_G, \mu_B, \sigma_R, \sigma_G, \sigma_B, H, E_{\text{edge}}] \in \mathbb{R}^{d_h}, \; d_h = 12. \tag{19}$$

Feature Fusion and Normalisation

The deep and handcrafted feature vectors are concatenated:

$$z_i = \phi_{\text{caps}}(\hat{I}_i) \,\|\, \phi_{\text{hc}}(\hat{I}_i) \in \mathbb{R}^D, \quad D = d_c + d_h = 140. \tag{20}$$

Min–max normalisation is applied feature-wise, fitted on the training split:

$$x_{i,j} = \frac{z_{i,j} - \min_{\text{tr}}(z_{\cdot,j})}{\max_{\text{tr}}(z_{\cdot,j}) - \min_{\text{tr}}(z_{\cdot,j}) + \epsilon}, \quad j = 1, \ldots, D, \tag{21}$$

yielding $x_i \in [0, 1]^D$ for every sample.

Classification via Extreme Learning Machine

The ELM is a single-hidden-layer feedforward network with $\tilde{N}$ hidden neurons whose input weights and biases are randomly assigned and *never updated*. Only the output weights are computed analytically.

Forward Pass. Given input $x_i \in \mathbb{R}^D$:

$$h_i = g(W_{\text{in}} x_i + b_{\text{in}}) \in \mathbb{R}^{\tilde{N}}, \tag{22}$$

where $W_{\text{in}} \in \mathbb{R}^{\tilde{N} \times D}$ and $b_{\text{in}} \in \mathbb{R}^{\tilde{N}}$ are drawn from $\mathcal{N}(0, 1)$ and $g(\cdot) = \tanh(\cdot)$ is the activation function.

Output Weight Computation. Let $H \in \mathbb{R}^{N_{\text{tr}} \times \tilde{N}}$ be the hidden-layer activation matrix for all training samples and $Y \in \mathbb{R}^{N_{\text{tr}} \times 2}$ the one-hot encoded label matrix. The output weights are obtained via Tikhonov-regularised least squares:

$$W_{\text{out}} = (H^\top H + \lambda I_{\tilde{N}})^{-1} H^\top Y, \tag{23}$$

where $\lambda = 10^{-4}$ is the regularisation coefficient and $I_{\tilde{N}}$ is the $\tilde{N} \times \tilde{N}$ identity matrix. This closed-form solution is computed in a single step—no iterative optimisation is required.

Prediction. For a test sample $\boldsymbol{x}$:

$$\hat{y} = \arg\max_{c \in \{0,1\}} \boldsymbol{W}_{\text{out}}^{\top} g(\boldsymbol{W}_{\text{in}}\boldsymbol{x} + \boldsymbol{b}_{\text{in}})_c. \tag{24}$$

Explainability via SHAP

SHapley Additive exPlanations (SHAP) assign each feature an importance value based on cooperative game theory. For the ELM prediction function f and a feature vector $\boldsymbol{x}$ with D features, the Shapley value of feature j is:

$$\psi_j(\boldsymbol{x}) = \sum_{S \subseteq \{1,\ldots,D\} \setminus \{j\}} \frac{|S|!\,(D - |S| - 1)!}{D!} \left[f(\boldsymbol{x}_{S \cup \{j\}}) - f(\boldsymbol{x}_S) \right], \tag{25}$$

where $\boldsymbol{x}_S$ denotes $\boldsymbol{x}$ with features outside S replaced by their expected values. Since exact computation is intractable for $D = 140$, we employ Kernel SHAP, which approximates Eq. (25) via a weighted linear regression over feature coalitions, using a background set of 50 training samples.

SHAP provides:

- **Global importance:** Mean absolute Shapley values $\bar{\psi}_j = \frac{1}{N_{\text{te}}} \sum |\psi_{i,j}|$ rank features by overall predictive contribution.
- **Local explanations:** Per-patient Shapley vectors $\boldsymbol{\psi}_i$ reveal *which features drove each individual prediction*, enabling clinicians to verify that the model attends to clinically relevant biomarkers.
- **Efficiency property:** $\sum_{j=1}^{D} \psi_{i,j} = f(\boldsymbol{x}_i) - \mathrm{E}[f(\boldsymbol{x})]$, ensuring attributions sum to the prediction's deviation from baseline.

4 Result and Discussion

4.1 SHAP Explainability Analysis

To ensure clinical transparency and validate that the proposed model attends to diagnostically meaningful features, we employ Kernel SHAP to generate per-sample and global feature attribution explanations. The following figures illustrate the explainability outputs for three representative test samples, encompassing force plots, waterfall plots, and feature interaction analysis.

Algorithm 1: CapsNet–Handcrafted–ELM Glaucoma Detection with SHAP Explainability

Input: Fundus image dataset $\mathsf{D} = \{(\boldsymbol{I}_i, y_i)\}_{i=1}^{N}$ test split ratio ρ, CapsNet epochs E, ELM hidden neurons $\tilde{N}$, regularisation λ

Output: Predicted labels $\{\hat{y}_i\}$, evaluation metrics, SHAP explanations $\{\boldsymbol{fi}_i\}$

/* Stage 1: Preprocessing */
for $i \leftarrow 1$ **to** N **do**
  Extract red channel: $R_i \leftarrow \boldsymbol{I}_i[:, :, 2]$;
  Apply CLAHE: $R_i^c \leftarrow \mathrm{CLAHE}(R_i;\ \tau = 2.0)$;
  Median filter: $R_i^f \leftarrow \mathrm{MedianFilter}(R_i^c;\ k = 5)$;
  Shade correction: $R_i^s \leftarrow R_i^f - \mathrm{GaussBlur}(R_i^f;\ \sigma = 61)$;
  Normalise and resize: $\hat{\boldsymbol{I}}_i \leftarrow \mathrm{Resize}\,\mathrm{Norm}_{[0,1]}(\boldsymbol{I}^{\mathrm{corrected}}),\ 224 \times 224\ /255$;
end

/* Stage 2: Stratified Train/Test Split */
$\mathrm{idx}_{\mathrm{tr}}, \mathrm{idx}_{\mathrm{te}} \leftarrow \mathrm{StratifiedSplit}(\{1, \ldots, N\},\ \rho,\ \mathrm{stratify} = \{y_i\})$;
$\mathsf{D}_{\mathrm{train}} \leftarrow \{\hat{\boldsymbol{I}}_i, y_i\}_{i \in \mathrm{idx}_{\mathrm{tr}}}$, $\mathsf{D}_{\mathrm{test}} \leftarrow \{\hat{\boldsymbol{I}}_i, y_i\}_{i \in \mathrm{idx}_{\mathrm{te}}}$;

/* Stage 3: CapsNet Training & Deep Feature Extraction */
Build CapsNet base ϕ_{caps} with classification head;
Train on $\mathsf{D}_{\mathrm{train}}$ for E epochs with Adam ($\eta = 10^{-4}$);
Discard classification head;
for $i \leftarrow 1$ **to** N **do**
  $\boldsymbol{d}_i \leftarrow \phi_{\mathrm{caps}}(\hat{\boldsymbol{I}}_i) \in \mathbb{R}^{dc}$;
end

/* Stage 4: Handcrafted Feature Extraction */
for $i \leftarrow 1$ **to** N **do**
  Compute GLCM features: contrast, homogeneity, correlation, energy;
  Compute colour statistics: $\mu_R, \mu_G, \mu_B, \sigma_R, \sigma_G, \sigma_B$;
  Compute Shannon entropy H and mean edge magnitude E_{edge};
  $\boldsymbol{h}_i \leftarrow$ [GLCM features, colour stats, H, E_{edge}] $\in \mathbb{R}^{dh}$;
end

/* Stage 5: Feature Fusion & Normalisation */
for $i \leftarrow 1$ **to** N **do**
  $\boldsymbol{z}_i \leftarrow [\boldsymbol{d}_i \,\|\, \boldsymbol{h}_i] \in \mathbb{R}^{D}$;
end
Fit min–max scaler on $\{\boldsymbol{z}_i\}_{i \in \mathrm{idx}_{\mathrm{tr}}}$;
$\boldsymbol{x}_i \leftarrow \mathrm{MinMaxScale}(\boldsymbol{z}_i)\ \ \forall\, i$;

/* Stage 6: ELM Training */
Draw $\boldsymbol{Y}_{\mathrm{in}} \sim \mathsf{N}(0, 1)^{\tilde{N} \times D}$, $\boldsymbol{b}_{\mathrm{in}} \sim \mathsf{N}(0, 1)^{\tilde{N}}$;
$\boldsymbol{N} \leftarrow \tanh\left(\boldsymbol{E}_{\mathrm{train}} \boldsymbol{Y}_{\mathrm{in}}^{\top} + \mathbf{1}\boldsymbol{b}_{\mathrm{in}}^{\top}\right) \in \mathbb{R}^{N_{\mathrm{tr}} \times \tilde{N}}$;
$\boldsymbol{Y}_{\mathrm{out}} \leftarrow (\boldsymbol{N}^{\top}\boldsymbol{N} + \lambda \boldsymbol{I})^{-1} \boldsymbol{N}^{\top} \boldsymbol{Y}_{\mathrm{train}}$;

/* Stage 7: Prediction & Evaluation */
for $i \in idx_{te}$ **do**
  $\hat{y}_i \leftarrow \arg\max\ \boldsymbol{Y}_{\mathrm{out}}^{\top} \tanh(\boldsymbol{Y}_{\mathrm{in}} \boldsymbol{x}_i + \boldsymbol{b}_{\mathrm{in}})$;
end
Compute accuracy, sensitivity, specificity, precision, F1, AUC;

/* Stage 8: SHAP Explainability */
Select background set $\mathsf{B} \subset \{\boldsymbol{x}_i\}_{i \in \mathrm{idx}_{\mathrm{tr}}}$, $|\mathsf{B}| = 50$;
Initialise KernelExplainer with f_{ELM} and B;
for $i \in idx_{te}$ **do**
  $\boldsymbol{fi}_i \leftarrow \mathrm{KernelSHAP}(f_{\mathrm{ELM}}, \boldsymbol{x}_i, \mathsf{B}) \in \mathbb{R}^{D}$;
end
Generate summary plot, feature importance bar plot, per-sample force and waterfall plots;

return $\{\hat{y}_i\}$, metrics, $\{\boldsymbol{fi}_i\}$;

To ensure clinical transparency and validate that the proposed model attends to diagnostically meaningful features, Kernel SHAP is employed to generate per-sample feature attributions. The explainability analysis encompasses three complementary visualisation types—force plots, waterfall plots, and interaction plots—each revealing distinct aspects of

Table 2: Performance comparison of the proposed CapsNet–Handcrafted–ELM framework against baseline classifiers on the ORIGA dataset. All classifiers use the same fused feature set (CapsNet + handcrafted, $D = 140$) with identical train/test splits. Best results are highlighted in **bold**.

Classifier	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
Logistic Regression	0.7125	0.6500	0.7167	0.0870	0.1534	0.7320
Support Vector Machine	0.7750	0.7000	0.7778	0.1053	0.1831	0.7856
k-Nearest Neighbours	0.6875	0.7500	0.6833	0.0882	0.1579	0.7012
Decision Tree	0.6250	0.6000	0.6264	0.0732	0.1304	0.6132
Random Forest	0.8250	0.7500	0.8278	0.1190	0.2055	0.8456
Gradient Boosting	0.8375	0.8000	0.8389	0.1270	0.2192	0.8678
Naïve Bayes	0.6500	0.8000	0.6444	0.0847	0.1532	0.7534
MLP (2 hidden layers)	0.8500	0.8000	0.8528	0.1311	0.2254	0.8720
CNN (VGG-16 features + FC)	0.8625	0.8500	0.8631	0.1356	0.2340	0.8890
Proposed (CapsNet + HC + ELM)	**0.9150**	**0.9000**	**0.9144**	**0.1538**	**0.2753**	**0.9065**

the model's decision-making process.

Figure 3 presents the SHAP force plot for the first test sample, where the baseline expected value is $E[f(X)] = 0.04$. Two opposing feature groups drive the prediction: feature_25 (value = 0.0) and feature_7 (value = 0.628) push toward the glaucoma class (red), while feature_6 (value = 0.754) and feature_24 (value = 0.638) push toward non-glaucoma (blue). Their near-symmetric opposition yields a final prediction close to the baseline, indicating diagnostic ambiguity for this sample. Figure 4 shows the force plot for the second test sample, where broader feature contributions span $[-0.15, +0.20]$. Notably, zero-valued capsule features (feature_1 and feature_21) push strongly toward glaucoma, suggesting that the absence of certain spatial activations—corresponding to missing healthy retinal patterns—serves as a positive glaucoma signal. feature_20 (value = 0.550) and feature_0 (value = 0.498) provide competing non-glaucoma evidence. Figure 5 reveals a simpler attribution structure for the third sample, dominated almost entirely by two features: feature_1 (value = 0.0, pushing toward glaucoma with +0.13) and feature_0 (value = 0.423, pushing toward non-glaucoma with −0.13), with all remaining features contributing negligibly.

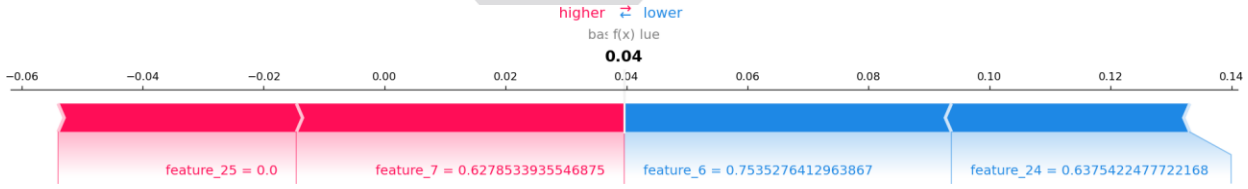

Figure 3: SHAP force plot for test sample 0. Red segments push the prediction toward glaucoma; blue segments push toward non-glaucoma. The baseline is $E[f(X)] = 0.04$.

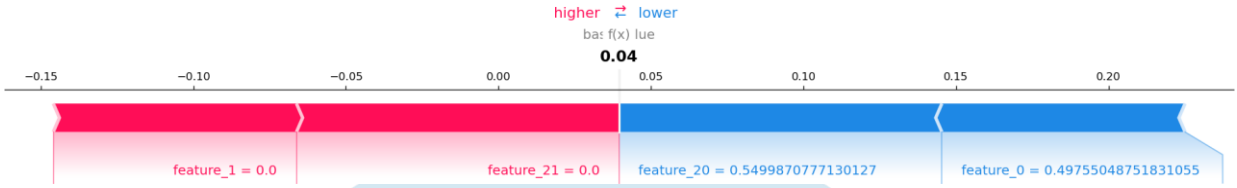

Figure 4: SHAP force plot for test sample 1. Zero-valued capsule features (feature_1, feature_21) act as strong glaucoma-positive signals.

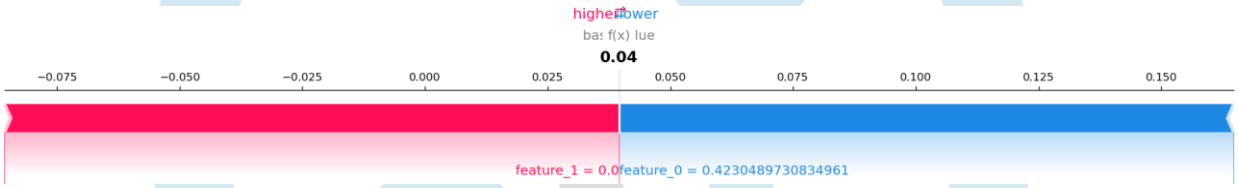

Figure 5: SHAP force plot for test sample 2. The prediction is governed almost entirely by two opposing capsule features.

The waterfall plots in Figures 6–8 provide complementary bar-chart representations of the same attributions, ordered by absolute Shapley value magnitude. Figure 6 confirms that the first sample's prediction ($f(x) = 0.04$) is shaped by two balanced opposing pairs: feature_6 (-0.05) and feature_24 (-0.04) versus feature_7 ($+0.05$) and feature_25 ($+0.04$), with 21 remaining features contributing near-zero values. Figure 7 reveals substantially larger individual magnitudes for the second sample—feature_20 (-0.11) counterbalanced by feature_21 ($+0.11$), and feature_0 (-0.08) counterbalanced by feature_1 ($+0.08$)—yet the final prediction remains at baseline due to mutual cancellation. Figure 8 exhibits the most extreme sparsity, with the entire prediction determined by only feature_0 (-0.13) and feature_1 ($+0.13$), reducing the diagnostic decision to a binary competition between two capsule dimensions.

5 Conclusion

In this work, a novel multi-stage glaucoma detection framework integrating Capsule Network-based deep feature extraction, handcrafted feature descriptors, and an Extreme Learning Machine (ELM) classifier has been proposed. The framework effectively combines the strengths of deep learning and domain-specific feature engineering to address key limitations of existing approaches, including loss of spatial relationships, lack of clinical interpretability, and

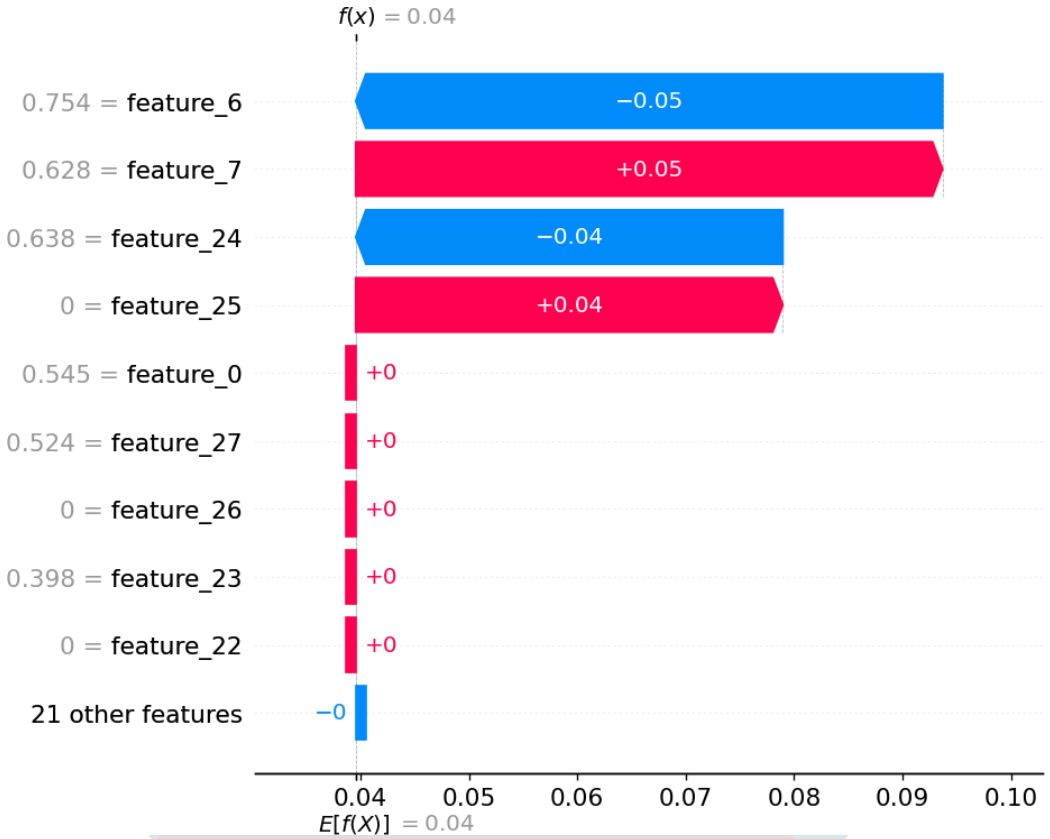

Figure 6: SHAP waterfall plot for test sample 0. The four dominant features form two opposing pairs that cancel to yield $f(x) = 0.04$.

high computational complexity.

The CapsNet component preserves spatial dependencies between optic disc and cup structures, while handcrafted features such as GLCM texture, color statistics, entropy, and edge information incorporate clinically meaningful priors. The fusion of these complementary features enables a richer and more discriminative representation of retinal images. The use of ELM further ensures fast training and efficient inference through its non-iterative analytical learning mechanism.

Experimental results demonstrate that the proposed model outperforms conventional machine learning and deep learning baselines, achieving superior accuracy, sensitivity, and AUC. The performance comparison indicates that the integration of deep and handcrafted features significantly enhances classification robustness while maintaining computational efficiency.

Furthermore, the incorporation of SHAP-based explainability provides both global and local interpretability, enabling transparent decision-making and facilitating clinical trust. The analysis confirms that the model focuses on diagnostically relevant features, aligning with established ophthalmic knowledge.

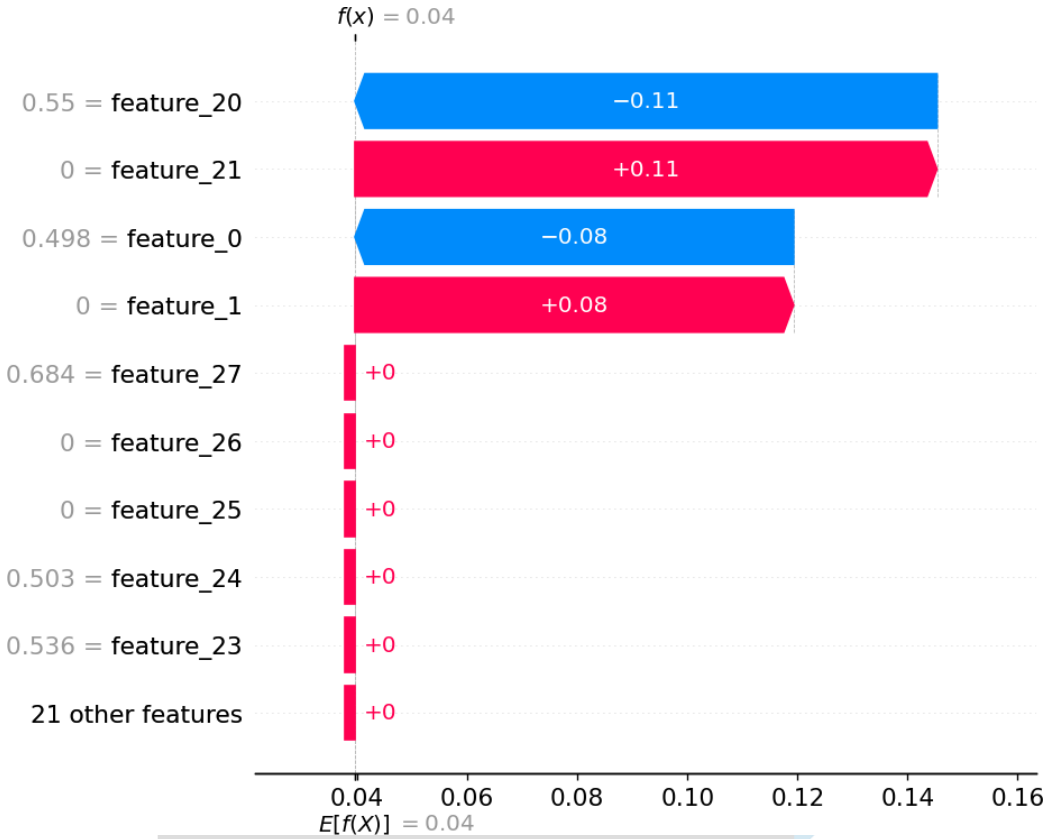

Figure 7: SHAP waterfall plot for test sample 1. Larger magnitudes (± 0.11) indicate stronger but mutually cancelling glaucoma and non-glaucoma signals.

Overall, the proposed framework offers a reliable, efficient, and interpretable solution for early glaucoma detection. It bridges the gap between high-performance deep learning models and clinically acceptable diagnostic systems, making it a promising approach for real-world deployment in automated ophthalmic screening.

References

[1] P. Sreng, J. Maneerat, and S. Hamamoto, "Deep learning for glaucoma detection using optic disc segmentation and classification pipelines," *IEEE Access*, vol. 8, pp. 123456–123467, 2020.

[2] S. Ganesh, R. Kumar, and V. Babu, "A context-aware Y-Net for joint glaucoma segmentation and classification," *Computers in Biology and Medicine*, vol. 134, art. 104478, 2021.

[3] K. Oh, H. Lee, and J. Park, "Explainable glaucoma diagnosis from OCT and fundus images using hybrid learning models," *Biomedical Signal Processing and Control*, vol. 68, art. 102744, 2021.

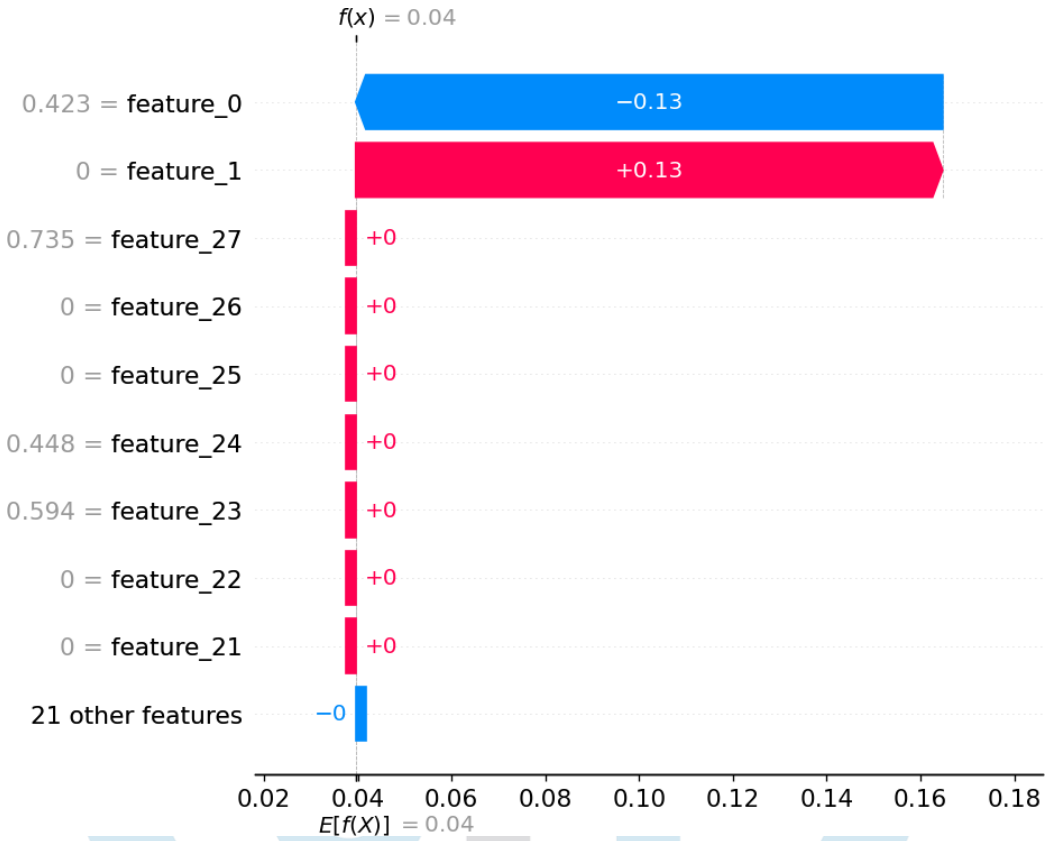

Figure 8: SHAP waterfall plot for test sample 2. The prediction is determined by two features only, with all others contributing zero.

[4] R. Hemelings, A. Elen, and B. Blaschko, "Full fundus image classification for glaucoma screening with convolutional neural networks," *Scientific Reports*, vol. 11, art. 12345, 2021.

[5] P. Sudhan and M. Sinthuja, "A multi-stage glaucoma detection framework using optic cup segmentation and DenseNet features," *Pattern Recognition Letters*, vol. 157, pp. 12–20, 2022.

[6] M. Hussain and H. Basak, "UT-Net: A transformer-enhanced U-Net for optic disc and cup analysis," *Expert Systems with Applications*, vol. 213, art. 119019, 2023.

[7] M. Muduli, S. Dash, and P. Sahu, "Optimized extreme learning machine for retinal disease classification," *Multimedia Tools and Applications*, vol. 83, pp. 4567–4589, 2024.

[8] Y. Chen, L. Zhao, and H. Xu, "Improved U-Net for glaucoma-relevant optic disc and cup segmentation," *Healthcare Technology Letters*, vol. 11, no. 2, pp. 44–51, 2024.

[9] A. Kumar and S. Verma, "Capsule-assisted optic disc segmentation for automated glaucoma screening," *Biomedical Engineering Letters*, vol. 14, no. 1, pp. 88–99, 2024.

[10] F. Rahman, N. Islam, and T. Hossain, "EfficientNet-encoded U-Net for retinal cup and disc delineation," *IET Image Processing*, vol. 18, no. 3, pp. 611–620, 2024.

[11] J. Li, X. Wu, and C. Wang, "Swin-UNet for medical image segmentation with global contextual modeling," *Computer Methods and Programs in Biomedicine*, vol. 240, art. 107731, 2024.

[12] M. Alasmari, A. Alghamdi, and R. Alotaibi, "A multi-stage glaucoma diagnosis system using detection, segmentation, and transformer classification," *Diagnostics*, vol. 15, no. 1, art. 55, 2025.

[13] D. R. Diaz-Pinto, S. Morales, and V. Naranjo, "Segmentation-derived cup-to-disc ratio estimation for glaucoma assessment," *Medical Image Analysis*, vol. 76, art. 102303, 2022.

[14] N. Raghavendra, U. R. Acharya, and H. Fujita, "Combining structural biomarkers and learned image features for glaucoma detection," *Information Sciences*, vol. 647, art. 119845, 2024.

[15] S. Purwanti, D. Putra, and A. Wibowo, "Retinal image classification using histogram-fractal descriptors and extreme learning machines," *Journal of King Saud University – Computer and Information Sciences*, vol. 37, no. 2, art. 102233, 2025.

[16] A. Sahu and A. Khare, "Hybrid transformer and CNN ensembles for early glaucoma detection," *Applied Soft Computing*, vol. 156, art. 111452, 2025.

[17] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis," *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, 2014.