IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 August 2025, accepted 1 September 2025,
date of publication 4 September 2025, date of current version 11 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3606334

RESEARCH ARTICLE

A Novel Transformer-CNN Hybrid Deep Learning Architecture for Robust Broad-Coverage Diagnosis of Eye Diseases on Color Fundus Images

CELINA RIECK, CHRISTOPHER MAI, LUCA EISENTRAUT, AND RICARDO BUETTNER, (Senior Member, IEEE)
Chair of Hybrid Intelligence, Helmut-Schmidt-University/University of the Federal Armed Forces Hamburg, 22043 Hamburg, Germany

Corresponding author: Ricardo Buettner (buettner@hsu-hh.de)

This work was supported in part by the Open Access Publication Fund of the Helmut-Schmidt-University/University of the Federal Armed Forces Hamburg; and in part by the High Performance Computing (HPC) Cluster HSUper through the project hpc.bw, funded by dtec.bw—Digitalization and Technology Research Center of the Bundeswehr and European Union—NextGenerationEU.

ABSTRACT We present a novel and robust deep-learning architecture that takes into account the pathological characteristics of eye diseases on color fundus images. The proposed hybrid architecture is capable of accurately recognizing the local features on the one hand and detecting the peripheral abnormalities of the specific eye disease on the other, and diagnosing the respective eye disease on this basis. The hybrid architecture provides robust local feature extraction and enhances global contextual understanding via hierarchical self-attention mechanisms. In contrast to previous models, which were limited to binary or poorly differentiated classification tasks, our approach considers the most clinically relevant eye diseases (here 9 diseases), including central serous chorioretinopathy (CSCR), diabetic retinopathy (DR), disc edema, glaucoma, macular scar, myopia, pterygium, retinal detachment, and retinitis pigmentosa (RP). The architecture was validated using a peer-reviewed real-world dataset and stratified 5-fold cross-validation. It achieved an average balanced accuracy of 81.91%, an average accuracy and True Positive Rate (TPR) of 76.40%, and a True Negative Rate (TNR) of 95.60% across all classes. For individual diseases, the model reached an TPR of 88.93% for DR, 89.93% for RP, and 61.23% for Glaucoma. The results show that the architecture is powerful and can distinguish between the 10 different classes (9 disease classes + healthy class). This makes it particularly suitable for pre-screening applications in telemedicine, for screening programs at the population level, and for integration into clinical decision support systems.

INDEX TERMS Eye disease, color fundus images, deep learning, convolutional neural network, EfficientNet, Swin Transformer V2, multiclass classification, fundus images, hybrid architecture.

I. INTRODUCTION

EYE diseases such as diabetic retinopathy (DR), glaucoma, macular degeneration, and retinal detachment are among the leading causes of vision impairment worldwide [1]. According to the World Health Organization, more than 2.2 billion people are affected by some form of visual impairment, many of which are preventable through early diagnosis and timely treatment [1], [2]. However, manual assessment of retinal fundus images is not only time-consuming and highly dependent on specialist expertise, but also prone to interobserver variability and diagnostic inconsistency [3], [4]. Recent studies reveal significant limitations in human diagnostic accuracy, particularly in the context of image-based screening tasks. For instance, in a large-scale comparison of DR screening, human graders

The associate editor coordinating the review of this manuscript and approving it for publication was Amin Zehtabian.

achieved a sensitivity of only 73.4% for referable disease and as low as 62% for diabetic macular edema (DME), resulting in a substantial false negative rate up to 14.1% in cases of proliferative DR [3]. Similarly, in the field of neuro-ophthalmology, nearly half of the patients were misdiagnosed prior to referral, and 26.00% of those affected experienced preventable harm due to diagnostic errors [2]. These findings underscore the inherent vulnerability of manual assessments, especially in scenarios requiring structured evaluation and comprehensive differential diagnosis. Recent advances in deep learning (DL) have enabled the development of systems based on artificial intelligence (AI) that could assist clinicians in ophthalmic diagnostics [2]. These models have demonstrated high performance in detecting several retinal diseases [3]. Although several approaches leveraged advanced architectures, such as residual networks or attention mechanisms, their diagnostic scope often remained restricted, and they were commonly trained on small-scale or non-peer-reviewed datasets. Furthermore, some approaches do not use cross-validation to evaluate the results. These limitations raise concerns regarding their robustness and generalizability in real-world clinical environments. In response to these challenges, we propose a hybrid DL architecture that combines the local feature extraction capabilities of EfficientNet-B4 [5] with the global contextual understanding of Swin Transformer V2 [6]. Swin Transformers are a new type of AI model that has shown strong results in medical image analysis [7], [8], [9]. This design is also particularly suited for ophthalmology, where both fine-grained patterns and broader structural changes must be jointly assessed for reliable diagnosis [1]. Performance is validated through stratified 5-fold cross-validation, ensuring robust and generalizable outcomes. So, our approach offers practical clinical value since a reliable automated classification system can support ophthalmologists by increasing diagnostic throughput, reducing error rates, and enabling earlier intervention to reduce preventable blindness [1], [3]. Figure 1 illustrates one potential application of our DL approach for detecting eye diseases. The main contributions of this work are as follows:

1) With our hybrid architecture, we establish a new benchmark for the broad detection of nine different eye diseases (+healthy class), validated through robust cross-validated results.

2) We show strong screening performance with high TNR (95.60%) and NPV (94.06%), highlighting the model's reliability for safely excluding non-diseased cases, which is crucial for telemedicine and large-scale screening.

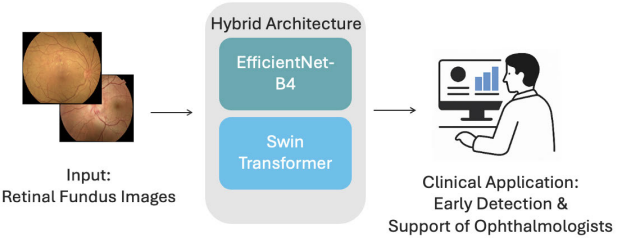

FIGURE 1. Conceptual overview illustrating a potential real-world application of the hybrid approach for the automated detection of various eye diseases in clinical settings.

II. RELATED WORK

A. LIMITED COVERAGE OF EYE DISEASES

Eye diseases are diverse, varying in shape, color, and severity. The range of conditions is broad, with some diseases appearing more frequently in the population than others. Several studies have concentrated on classifying only a limited number of these diseases. This section highlights studies that focused on the classification of fewer than six eye diseases. One example is Abbas [10], who introduced Glaucoma-Deep, a hybrid DL architecture designed specifically for glaucoma detection. The model integrates unsupervised feature extraction via CNNs, deep belief networks for feature selection, and a softmax classifier, which achieved an accuracy of 99%. With a hybrid approach, Glaret Subin and Muthukannan [18] designed a CNN model fine-tuned using the flower pollination algorithm, combined with entropy-based preprocessing and support vector machine-based classification. Their system yielded an accuracy of 95.27% and an F1-score of 93.3% across several disease categories. Babaqi et al. [16] applied transfer learning for classifying fundus images into four categories: normal, DR, glaucoma, and cataract. Their fine-tuned CNN model, trained on 4,200 labeled images, reached an overall accuracy of 94%. Hossain et al. [12] developed an automated cataract detection system based on Deep Convolutional Neural Networks (DCNNs) using a modified ResNet50 architecture, trained directly on fundus images without prior preprocessing, achieving an accuracy of 95.77% and an AUC of 0.982 on the test dataset. Bernabé et al. [11] extended the binary classification paradigm to differentiate between DR and glaucoma. Their CNN-based system, evaluated using K-fold cross-validation, reached a classification accuracy of 99.89%. Focusing on classification across DR, glaucoma, and age-related macular degeneration (AMD), Chea and Nam [23] developed a residual network optimized through iso-luminance histogram equalization and extensive data augmentation. Their model achieved a peak accuracy of 91.16% and an average of 85.79% for AMD detection. Focusing on early DR, Oh et al. [13] explored the use of ultra-wide-field (UWF) fundus imaging, which captures up to 200Â° of the retina. Their study demonstrated that employing ETDRS 7-field images, as opposed to optic disc-centered views, significantly enhanced diagnostic performance, with the model achieving an AUC of 0.9150. In the context of glaucoma detection, Xu et al. [14] proposed TIA-Net (Transfer Induced Attention Network), a transfer learning-based architecture that incorporates channel-wise attention and reduces domain discrepancy using maximum mean discrepancy. The model achieved classification accuracies of 85.7% and 76.6% across two benchmark datasets, highlighting its adaptability across

TABLE 1. Overview of studies on automated eye disease detection. Included are the authors, whether the work covers a broad coverage of relevant eye diseases, and whether peer-reviewed datasets and cross validation were used.

Author(s)	Peer-reviewed	Broad coverage of relevant eye diseases	Cross-validated
Abbas et al. [10]	✗	✗	✓
Bernabé et al. [11]	✗	✗	✓
Hossain et al. [12]	✗	✗	✗
Oh et al. [13]	✗	✗	✓
Xu et al. [14]	✗	✗	✓
Wahab Sait and Rahaman [15]	✗	✗	✗
Babaqi et al. [16]	✗	✗	✗
Biswas et al. [17]	n/a	✗	✗
Glaret subin and Muthukannan [18]	✗	✗	✓
He et al. [19]	✗	✓	✓
Al-Fahdawi et al. [20]	✗	✓	✗
Grassmann et al. [21]	✓	✗	✗
Tomar et al. [22]	✓	✗	✗
Chea and Nam [23]	✓	✗	✓
This study	✓	✓	✓

domains. Tomar et al. [22] proposed a DL pipeline using a multi-layer Fire Hawk Convolutional Neural Network. This system incorporated image enhancement, U-net-based segmentation, and classification layers optimized with the Fire Hawk Optimizer and achieved strong results in detecting DR. A complementary method by Biswas et al. [17] combined CNNs and artificial neural networks to both classify and localize diseases such as DR, glaucoma, and cataract from fundus images. Their system demonstrated a maximum accuracy of 93%, emphasizing real-time diagnostic applicability. Grassmann et al. [21] focused on classifying various stages of age-related macular degeneration (AMD) using the AREDS dataset. Their ensemble of six convolutional neural networks, based on architectures such as Inception, ResNet, and VGG, was trained on more than 120,000 fundus images. In external validation, the model achieved a weighted kappa of 0.92, correctly identifying 94.3 percent of healthy cases and 84.2 percent of AMD cases. Wahab Sait and Rahaman [15] developed a lightweight DL model integrating denoising autoencoders, SSD-based feature extraction, and ShuffleNet V2, further optimized using the Whale Optimization Algorithm. The model achieved an accuracy of 99.4% and a Cohen's Kappa of 96.5% on the ODIR and EDC datasets, highlighting its efficiency for deployment in resource-constrained settings. In addition to focusing on a limited number of diseases, most of the studies also share the limitation of relying on non-peer-reviewed image data. Non-peer-reviewed datasets may contain incorrect labels, inconsistent image quality, and non-standardized acquisition conditions, leading to biased training data, reduced model accuracy, limited generalizability, and an increased risk of clinically unreliable predictions. A classification of the work can be seen in table 1.

B. BROAD COVERAGE OF EYE DISEASES

The advantage of broad coverage of eye diseases is that it closely reflects the complexity of real-world clinical settings. For DL models to be truly effective in the automated detection of eye conditions, they must be capable of recognizing a wide range of diseases, which is essential for practical use in diverse patient populations. Models trained on only two or three specific diseases are limited in scope and may struggle when confronted with unfamiliar or coexisting conditions. This represents a significant drawback in realistic medical applications. This section highlights studies that focus on broad coverage of at least six different eye diseases. He et al. [19] addressed multi-label classification with DCNet, a densely connected CNN architecture enhanced by a spatial correlation module and patient-level feature fusion. By leveraging the relationships between left and right eye images, DCNet outperformed conventional CNNs on benchmark datasets. One of the most comprehensive frameworks, Fundus-DeepNet, was introduced by Al-Fahdawi et al. [20]. Their model integrates HRNet, attention mechanisms, SENet blocks, and discriminative restricted Boltzmann machines, alongside a robust preprocessing pipeline and feature-level fusion of binocular fundus data. The approach achieved high AUC scores with F1-Scores exceeding 88%. As is clearly evident, the number of studies is very limited, suggesting that only a small number of works have so far addressed the classification of multiple eye diseases. An additional limitation is that the studies rely on

non-peer-reviewed datasets, which raises concerns about data quality and reliability. Another shortcoming is the lack of robustness in their evaluation methods, as either no validation is used or it is limited to a small number of folds. Moreover, there is a lack of mechanisms capable of jointly capturing both local and global image patterns with high precision, which is particularly critical in ophthalmology, where accurate diagnosis often depends on integrating fine-grained pathological cues with broader anatomical context. To address these limitations, we propose a hybrid architecture that utilizes the strengths of EfficientNet-B4 in capturing local features with the capabilities of Swin Transformer V2 in modeling global features. To demonstrate the robustness of our approach, we validate the results through stratified 5-fold cross-validation on a peer-reviewed dataset.

III. METHODOLOGY

A. MODEL ARCHITECTURE

1) HYBRID ARCHITECTURE

The hybrid architecture integrates convolutional and transformer based components, leveraging EfficientNet-B4 as a locally focused backbone and enhancing it with Swin Transformer V2 modules to capture global contextual information, an essential capability for eye images where fine grained local structures and broader spatial patterns often appear simultaneously. EfficientNet-B4 serves as the foundational component of the model, selected for its balance between computational efficiency and classification accuracy [5]. The network is initialized with pretrained ImageNet weights, which promote faster convergence and improved generalization [24]. Swin Transformer was selected to leverage its strength in capturing long-range dependencies [6], which is advantageous in detecting diseases of varying size and shape. The hybrid architecture is shown in Figure 2. Both EfficientNet-B4 and the Swin Transformer independently generate deep feature maps from the input image [6]. Let $\mathbf{F}_{\text{Eff}}$ and $\mathbf{F}_{\text{Swin}}$ denote the resulting feature maps. These are subsequently passed through adaptive average pooling (AAP) operations, which reduce their spatial dimensions to a fixed size by computing the mean value within each pooling region, regardless of the input resolution [25]. The output is then flattened into one-dimensional feature vectors $\mathbf{f} \in \mathbb{R}$:

$$\mathbf{f}_{\text{Eff, 1D}} = \text{Flatten}(\text{AAP}(\mathbf{F}_{\text{Eff}})) \in \mathbb{R}^{d_1} \quad (1)$$

$$\mathbf{f}_{\text{Swin, 1D}} = \text{Flatten}(\text{AAP}(\mathbf{F}_{\text{Swin}})) \in \mathbb{R}^{d_2} \quad (2)$$

Here, d_1 and d_2 denote the resulting feature vector dimensions, which typically correspond to the number of output channels in the final convolutional or transformer stages of EfficientNet-B4 and the Swin Transformer, respectively. To stabilize and normalize the intermediate representations, both vectors are individually processed through Batch Normalization layers [26]. The normalized vectors $\hat{\mathbf{f}}_{\text{Eff, 1D}}$ and $\hat{\mathbf{f}}_{\text{Swin, 1D}}$ then fused to form a joint representation.

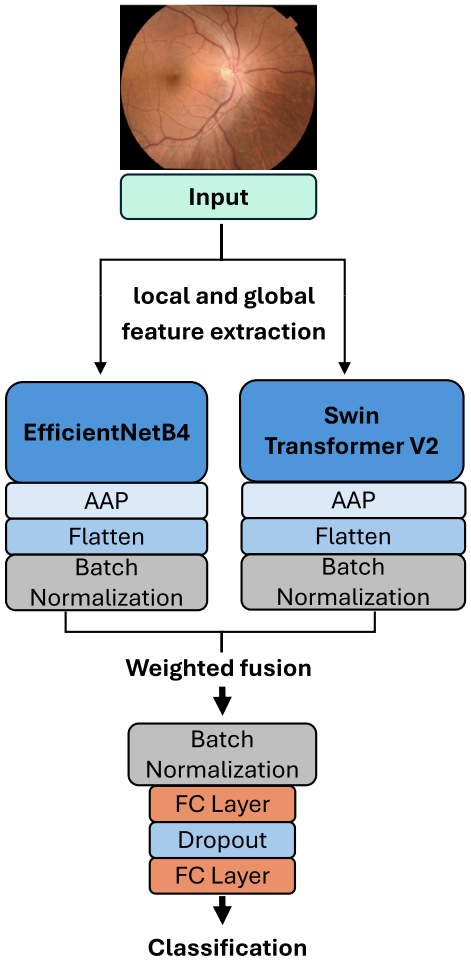

FIGURE 2. Structure of the hybrid architecture. EfficientNetB4 extracts local features from fundus images, while Swin Transformer V2 captures both local and predominantly global features. The resulting 1D vectors are fused and passed through two fully connected layers to make the final classification.

The importance of each vector is taken into account through the learnable scalar parameters α and β, which assign a weight to each vector such that $\alpha + \beta = 1$.

$$\mathbf{f}_{\text{fused}} = \alpha \cdot \hat{\mathbf{f}}_{\text{Eff, 1D}} + \beta \cdot \hat{\mathbf{f}}_{\text{Swin, 1D}} \quad (3)$$

The fused vector $\mathbf{f}_{\text{fused}}$ is then batch-normalized again to ensure numerical stability before entering the classification head [27], and it serves as the input to two fully connected (FC) layers, with a dropout layer between them to reduce overfitting. By maintaining separate processing pipelines up to the fusion stage, the architecture retains fine-grained local information captured by EfficientNetB4 [5], as well as long-range contextual dependencies captured by the Swin Transformer V2 [6]. This design ensures that the classifier benefits from complementary representational strengths, improving its robustness on complex image classification tasks.

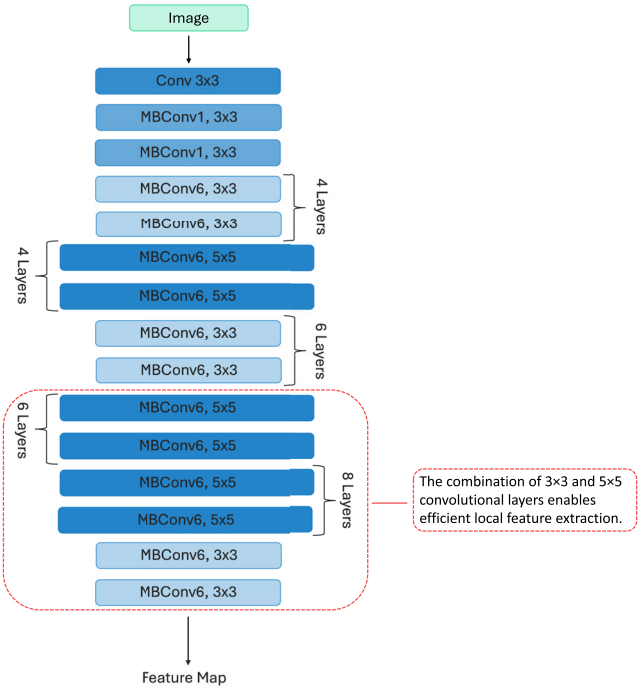

FIGURE 3. Layer-wise structured EfficientNet-B4 architecture, [28], [29]. The combination of 3 × 3 and 5 × 5 convolutional layers is optimal for local feature extraction.

2) EFFICIENTNET-B4

EfficientNet, introduced by Tan and Le [5], comprises a series of CNN architectures that achieve state-of-the-art accuracy with substantially improved computational efficiency over previous models. The key innovation of EfficientNet lies in its compound scaling method, which uniformly scales the network's depth, width, and input resolution using a set of fixed scaling coefficients [5]. This avoids the manual and often suboptimal scaling of individual dimensions seen in prior architectures. The base model, EfficientNet-B0, is built using the Mobile Inverted Bottleneck Convolution (MBConv) blocks introduced in MobileNetV2, which combine depthwise separable convolutions with squeeze-and-excitation optimization to enhance channel-wise feature recalibration [5]. It also employs Swish activations for improved nonlinear expressiveness [5]. Subsequent models in the EfficientNet family (B1-B7) are derived by systematically scaling up the base architecture with the compound coefficient, preserving the balance between accuracy and efficiency. The compound scaling method at the heart of EfficientNet simultaneously scales the network's depth, width, and input resolution using a set of fixed coefficients [5]. Rather than arbitrarily adjusting one dimension, this approach seeks to optimize all dimensions in concert, governed by the following equations [5]:

$$\text{depth} = \alpha^{\phi}, \quad \text{width} = \beta^{\phi}, \quad \text{resolution} = \gamma^{\phi} \tag{4}$$

subject to the constraint:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \quad \text{with} \quad \alpha \geq 1,\ \beta \geq 1,\ \gamma \geq 1 \tag{5}$$

where ϕ is the compound coefficient that controls the overall scaling of the network. Through extensive grid search, Tan and Le empirically determined the optimal values for the scaling constants as [5]:

$$\alpha = 1.2, \quad \beta = 1.1, \quad \gamma = 1.15 \tag{6}$$

EfficientNet-B4 corresponds to a compound coefficient of $\phi = 4$, resulting in the following scaled factors [5]:

$$\begin{aligned} \text{depth scale} &\approx 1.2^4 = 2.07 \\ \text{width scale} &\approx 1.1^4 = 1.46 \\ \text{resolution scale} &\approx 1.15^4 = 1.75 \end{aligned} \tag{7}$$

Thus, EfficientNet-B4 is derived by uniformly scaling up EfficientNet-B0 by these factors, leading to a model with approximately 19 million parameters and an input resolution of 380 × 380 pixels. This principled scaling maintains the balance between model capacity and computational cost. As shown in Figure 3, EfficientNet-B4 consists of an initial convolutional layer followed by a series of MBConv blocks with varying kernel sizes (3×3 and 5×5) and expansion factors. Each block group is annotated with the number of layers, progressively building hierarchical feature representations down to the final feature map.

3) SWIN TRANSFORMER V2

The Swin Transformer (Shifted Window Transformer), introduced by Liu et al. [6], is a hierarchical vision transformer architecture specifically designed to enable scalable and efficient visual representation learning; visualized in Figure 4. Unlike traditional vision transformers that apply global self-attention, swin limits attention computations to non-overlapping local windows, thereby reducing the computational complexity from quadratic to linear with respect to image size [6]. To facilitate interaction between these local windows and capture long-range dependencies, swin employs a shifted windowing scheme. In this design, the positions of the local windows are shifted between consecutive layers, allowing information to flow across window boundaries without incurring the cost of full self-attention [6]. This mechanism enhances representational capacity while maintaining high efficiency. Another key feature of the Swin Transformer is its hierarchical structure, which gradually merges image patches to build multi-scale feature representations, closely resembling the pyramidal design of CNNs [6]. This enables seamless integration into dense prediction tasks such as object detection, image segmentation, and, in the medical domain, disease localization. By combining localized self-attention, hierarchical processing, and cross-window communication, the Swin Transformer is well-suited for tasks requiring both fine-grained detail and global contextual awareness. This makes it particularly advantageous for analyzing medical images, where disease markers may be subtle, small-scale, or spread across different spatial regions.

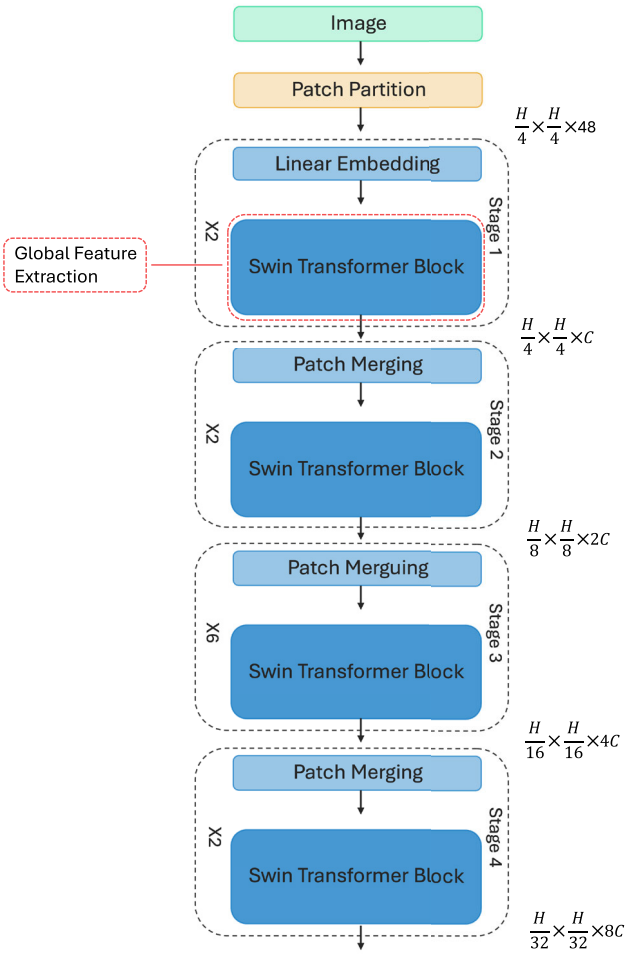

FIGURE 4. Swin Transformer: hierarchical vision model with local attention, as proposed by Liu et al. [6]. Its architecture also enables the extraction of global features, which is advantageous for disease diagnosis in fundus images.

B. PROCESS OF TRAINING

The entire training process is shown in Figure 5. Before model training, a stratified 5-fold cross-validation was applied using the scikit-learn library [30], to divide the dataset into five equally sized subsets. This method was chosen to ensure that each fold maintains the original class distribution, unlike standard k-fold cross-validation, which may produce biased results due to class imbalance. In each iteration, four folds (80%) are used for training, and one fold (20%) is held out for testing. This process is repeated five times, rotating the test fold each time to ensure robust performance evaluation. Within the training set of each fold, 10% is further reserved for validation and hyperparameter tuning. All images are resized to 256 × 256 pixels with three RGB channels and normalized using the ImageNet mean and standard deviation. To address the imbalance between the individual disease classes, the class weights are integrated into the loss function (*CrossEntropy*(weight=class_weights)) in order to also take rarer classes into account.

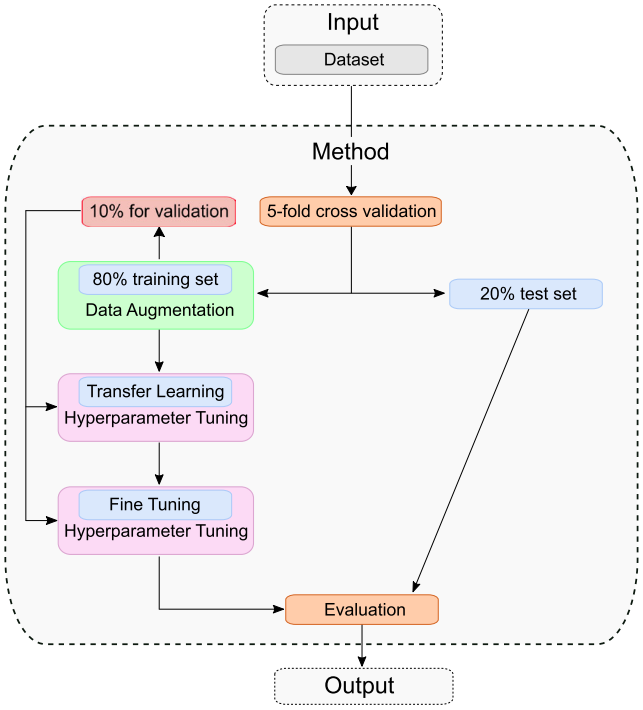

FIGURE 5. Training and Evaluation approach: The dataset is split into a training set and a test set. Data augmentation is applied to the training set, followed by model training using transfer learning and fine-tuning. Finally, the resulting model is evaluated on the test set.

TABLE 2. Overview of the hyperparameters used for hyperparameter tuning. TL = Transfer Learning; FT = Fine-Tuning.

Hyperparameter	Minimum Value	Maximum Value	Step	TL or FT
Dropout	0.1	0.5	0.05	TL, FT
Learning Rate (1)	10^{-4}	10^{-2}	calculated logarith-mically	TL
Learning Rate (2)	10^{-6}	10^{-4}	calculated logarith-mically	FT
Weight decay	10^{-5}	10^{-3}	calculated logarith-mically	TL, FT
1. FC layer units	128	1024	128	TL, FT
Batch size	-	-	8, 16 or 32	FT
Optimizer	-	-	AdamW, SGD	TL

To reduce overfitting and improve generalization, the training data is augmented using RandomRotation(±10°) and RandomResizedCrop(scale=(0.9, 1.0)) as well as RandomAffine(degrees=0, translate=(0.1, 0.1)). These transformations account for variability in image scale and

position, making the model more robust to input variations. Each model is trained for a maximum of 100 epochs with a batch size of 16. The training consists of two phases: transfer learning and fine-tuning. During transfer learning, the pretrained feature extraction layers (based on ImageNet weights [24]) are frozen, while the fully connected layers remain trainable. In the fine-tuning phase, the entire model is updated. Hyperparameter optimization is performed separately for both training phases using Optuna with a Tree-Structured Parzen Estimator (TPE) approach. For each fold, 20 trials are conducted, and the configuration that yields the lowest validation loss is selected. Each trial is allowed to run up to 100 epochs. To avoid overfitting and reduce computational overhead, we implemented early stopping, which terminates training if the validation loss does not decrease for ten consecutive epochs.

After identifying the optimal hyperparameters, a final transfer learning model is trained using the same configuration as during optimization. The model with the lowest validation loss is saved and used as the base for the subsequent fine-tuning phase. In fine-tuning, the hyperparameters; see Table 2, are re-optimized, and additional feature extraction layers are unfrozen to enable further domain adaptation. Again, the best-performing model based on validation loss is stored. After training, the final model is evaluated on the 20% test set using several following performance indicators.

C. EVALUATION METRICS

To evaluate and interpret the model's performance, we employ the following performance indicators: Accuracy, Balanced accuracy, True positive rate (TPR), True negative rate (TNR), Positive predictive value (PPV), Negative predictive value (NPV), Cohen's Kappa, and F1-Score. These metrics refer to the multi-class case, taking into account the class weighting. N is the total number of classes, n_i represents the number of samples per class, and S with $S = \sum_{i=1}^{N} n_i$ denotes the total number of samples. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) represent the four fundamental outcomes of a classification model.

- **Accuracy** indicates the ratio of correctly predicted instances to the total number of predictions [31]. Although it provides a general sense of model performance, it can be misleading in the presence of class imbalance.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

- **Balanced Accuracy** calculates the average recall over all classes, providing a more robust measure for imbalanced datasets [32].

$$\text{Balanced Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i} \tag{9}$$

- **Cohen's Kappa** (κ) evaluates the agreement between predicted and true labels, taking into account the agreement occurring by chance [33]. Here, P_o represents the observed agreement, and P_e the expected agreement by chance.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{10}$$

- **PPV** is a value that describes how many of the instances that are assigned to a class actually belong to that class. This means that you can also see the ratio of how many instances were incorrectly assigned to a class [34]. PPV is calculated by [30], [35]:

$$\text{PPV} = \sum_{i=1}^{N} \frac{n_i}{S} \cdot \frac{TP_i}{TP_i + FP_i} \tag{11}$$

- **NPV** refers to the proportion of true negative predictions among all predicted negatives [36].

$$\text{NPV} = \sum_{i=1}^{N} \frac{n_i}{S} \cdot \frac{TN_i}{TN_i + FN_i} \tag{12}$$

- **TPR** reflects the ratio of instances assigned to the correct class by the model compared to the total number of instances of this class. Therefore, this value also shows how often a class was incorrectly assigned to another class [34]. TPR is calculated by [30], [35]:

$$\text{TPR} = \sum_{i=1}^{N} \frac{n_i}{S} \cdot \frac{TP_i}{TP_i + FN_i} \tag{13}$$

- **TNR** denotes the proportion of actual negative instances correctly identified as such [30], [36].

$$\text{TNR} = \sum_{i=1}^{N} \frac{n_i}{S} \cdot \frac{TN_i}{TN_i + FP_i} \tag{14}$$

- **F1-Score** is the harmonic mean of precision and recall [31]. It is particularly valuable in scenarios where both false positives and false negatives carry significant consequences. The formula used for the multiclass case is provided by [37]:

$$\text{F1-Score} = \sum_{i=1}^{N} \frac{n_i}{S} \, 2 \cdot \frac{PPV_i \cdot TPR_i}{PPV_i + TPR_i} \tag{15}$$

D. DATASET

This study makes use of the Eye Disease Image Dataset, introduced by Sharmin et al. [4], which offers a comprehensive and diverse set of high-quality real-world images designed for the detection and classification of different eye diseases. The dataset comprises a total of 5,335 original images, collected over an eight-month period from the Anawara Hamida Eye Hospital and the B.N.S.B. Zahurul Haque Eye Hospital, both located in the Faridpur district of Bangladesh. To illustrate

the visual diversity across different disease classes, Figure 6 shows representative fundus images of Glaucoma, Healthy and Myopic eyes, Central Serous Chorioretinopath (CSCR), Retinitis Pigmentosa (RP), and Disc Edema out of the dataset [4]. The images were acquired using Topcon TRC-50DX and TL-211 fundus cameras attached to Nikon DSLR cameras, with resolutions ranging from 2,004 × 1.690 to 5,600 × 3.728 pixels, and are provided in jpg format. Table 3 presents two different properties of the dataset images to illustrate the model's robustness to variations. The signal-to-noise ratio was determined by estimating noise with the Immerkaer method and expressing it relative to the signal [38]. The dataset is categorized into ten distinct classes, covering both retinal diseases and anterior segment conditions. Specifically, it includes 1,509 images of DR, 1,349 images of Glaucoma, 444 images of Macular Scar, 127 images of Optic Disc Edema, 101 images of CSCR, 125 images of Retinal Detachment, 139 images of RP, 500 images of Myopia, and 1,024 images of healthy eyes. Additionally, there are 17 images of Pterygium, representing the anterior segment category. Overall, the importance of high-quality and diverse datasets for the training of robust DL models has also been emphasized in recent literature, for example by Gan et al. [39], highlighting the dataset [4] as one of the key resources available for research in this domain.

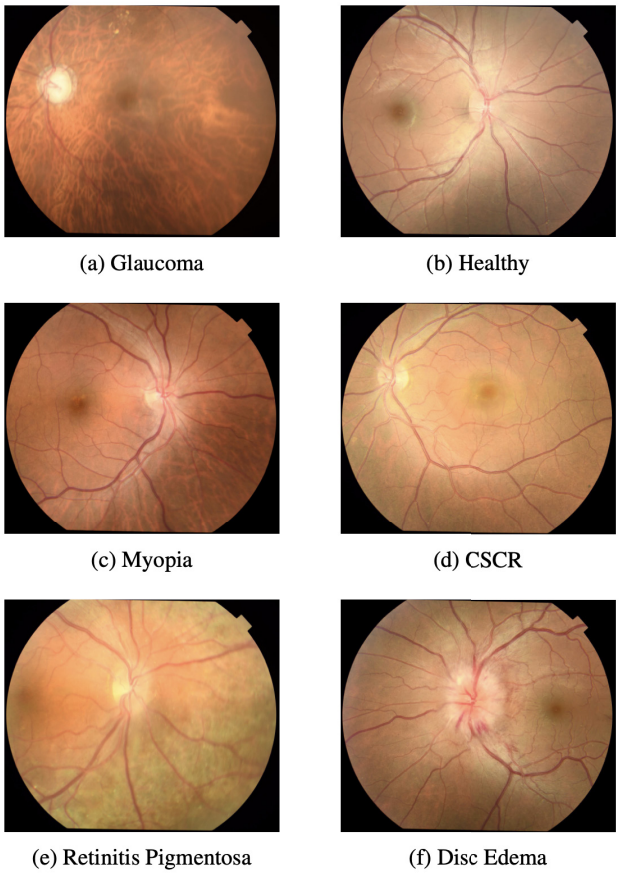

FIGURE 6. Sample fundus images from six different classes [4].

TABLE 3. Noise and Brightness properties of the images contained in the dataset.

Property	Value
Brightness	Mean: 0.3642; Std.: 0.0784
Signal to noise ratio (Immerkaer method)	Mean: 30.04 dB; Std.: 2.70 dB

E. SETUP

For training and testing the architecture, an NVIDIA L40S GPU with 48 GB of memory with PyTorch 2.5.0 is employed. Furthermore, Python version 3.11.7 and CUDA version 12.4.1 were used. The hybrid architecture was trained for a maximum of 100 epochs using either AdamW or SGD, depending on which optimizer achieved the best performance during hyperparameter tuning. Training was performed in phases, including transfer learning and fine-tuning. To identify the optimal parameters for transfer learning and fine-tuning, hyperparameter tuning was performed in 20 trials using Optuna (version 4.2.1). To avoid overfitting and save computation time, the callback function *early stopping* with a *delta* = 0.001 was used, which stops the training after 10 consecutive epochs in which the validation loss has not decreased. Scikit-learn (version 1.5.2) was utilized for stratified cross-validation and the computation of performance indicators. Throughout the entire training and validation process, the images were converted to a resolution of 256 × 256 pixels.

IV. RESULTS

The evaluation results of the proposed hybrid architecture, summarized in Table 4 and shown in Figure 8, reflect a robust and well-calibrated classification performance across all ten disease categories. Table 6 shows the results of the individual classes. The overall accuracy ranged from 73.29% to 78.26%, with an average of 76.40%, indicating reliable general performance in multiclass retinal classification. Balanced accuracy, which compensates for class imbalance by averaging sensitivity across classes, varied between 79.29% and 84.23%, yielding a mean value of 81.91%. TPR followed the same range and average as overall accuracy, with a minimum of 73.29% and a maximum of 78.26%. The PPV achieved values between 75.81% and 80.25%, averaging 78.15%. The corresponding F1-Score varied from 73.81% to 78.31%, with an average of 76.65%, which provides a useful summary of classification quality, particularly for uneven class distributions. This suggests that the model maintains a good trade-off between identifying true positives and avoiding false positives. Of particular importance are the NPV and the TNR, with average values of 94.06% and 95.60% respectively. A high TNR indicates that the model is effective at correctly excluding irrelevant classes for a given prediction, reducing the risk of false alarms in clinical practice. Last, the Cohen's Kappa score of 71.00% reflects

substantial agreement between the model's predictions and the ground truth beyond chance, supporting the statistical reliability of the results. Across all metrics, the results show moderate fluctuations between individual runs, suggesting consistent behavior of the model under repeated evaluation conditions. Figure 7 compares the accuracy and balanced accuracy of the hybrid architecture with the baseline models EfficientNet-B4 and Swin Transformer V2, showing that the hybrid architecture achieves higher values in both cases. In addition to the chosen backbones, the fusion strategy also influences the architecture's performance. Table 5 compares four different fusion strategies, showing that weighted fusion achieves the highest balanced accuracy.

TABLE 4. Evaluation results of the hybrid architecture across five-fold cross-validation. The table reports key performance metrics for each fold (I-V) and the corresponding average (Avg.) with standard deviation, given in percent (%).

Metric	I	II	III	IV	V	Avg.
Acc.	77.79	73.29	74.70	78.26	77.98	76.40 ± 2.26
Bal. Acc.	81.14	80.87	79.29	84.04	84.23	81.91 ± 2.15
TPR	77.79	73.29	74.70	78.26	77.98	76.40 ± 2.26
TNR	95.82	95.37	95.05	95.89	95.85	95.60 ± 0.37
PPV	79.25	75.81	76.21	79.22	80.25	78.15 ± 2.00
NPV	94.78	93.05	93.60	94.43	94.43	94.06 ± 0.71
Kappa	72.59	67.41	68.85	73.24	72.90	71.00 ± 2.68
F1-Score	78.07	73.81	75.07	78.31	78.01	76.65 ± 2.07

TABLE 5. Overview of the achieved balanced accuracy (in %) with four different fusion strategies. The bold value indicates the highest score.

Fusion strategy	Concatenation	Weighted fusion	Gating	Attention-based
Bal. Acc.	80.64	**81.91**	80.58	80.96

Figure 8 presents the average confusion matrix across all five folds for the ten-class retinal disease classification task. It displays the mean prediction counts alongside class-wise accuracies in percent, providing a detailed view of the model's discriminative capabilities. The pronounced diagonal structure reflects a generally high classification performance, with the majority of samples correctly assigned to their respective categories. The model performed particularly well in detecting Retinal Detachment (class 9), corresponding to a strong TPR of 94.40%. Retinitis Pigmentosa (class 10) was detected with high reliability, achieving the third-highest TPR

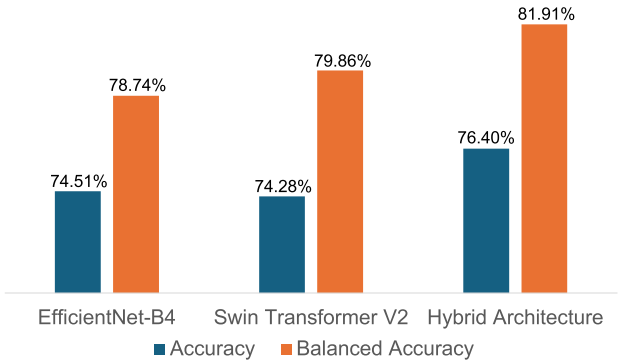

FIGURE 7. Comparison of accuracy and balanced accuracy between the baseline architectures EfficientNet-B4 and Swin Transformer V2 and the proposed hybrid architecture.

at 89.93%, followed by Disc Edema (class 3) at 89.76% and DR at 88.93%. Pterygium (class 8) was classified without error, albeit on a small sample size (100.00%). The classification of healthy eyes (class 5) has a TPR of 75.78%, whereby this class was most frequently misclassified as Glaucoma (class 4) with 15.72%. Myopia (class 7) achieved a correct classification rate of 76.60%, yet shared feature similarities with Glaucoma (16.40%) led to notable misclassifications. Performance deteriorated in more complex or visually overlapping categories. Glaucoma (class 4), for instance, was correctly classified with 61.23%, with substantial confusion toward Myopia (15.27%) and Healthy eyes (14.97%), suggesting limitations in distinguishing subtle optic nerve changes from normal anatomical variance. Macular Scar (class 6) yielded moderate results. 67.34% were correctly identified, with errors mainly relating to misclassifications with the diseases CSCR, Healthy, Glaucoma, Myopia and DR.

V. DISCUSSION

A. EVALUATION OF THE HYBRID ARCHITECTURE AND LIMITATIONS

The use of EfficientNet-B4 offered a strong foundation due to its optimized compound scaling and pretrained weights, enabling robust feature extraction with limited overfitting. Swin Transformers enriched this by providing multiscale attention, capturing global structural cues often overlooked by CNNs. This combination enhances the model's ability to discern both fine-grained and contextual features, critical in retinal imaging where pathologies appear at varying scales and locations. The proposed hybrid architecture demonstrated strong classification performance on the ten-class, peer-reviewed eye disease dataset [4], as shown in Table 4 and Figure 8. With an average balanced accuracy of 81.91%, and a Cohen's Kappa of 71.00%, the model not only shows predictive strength but also statistical reliability. Furthermore, a TNR of 95.60% and an NPV of 94.06% underline its effectiveness in excluding non-target conditions, an essential

TABLE 6. Per-class evaluation (averaged over five folds), including the mean true-class confidence score ± standard deviation. Values are given in percent (%).

Class	Acc.	Bal. Acc.	TPR	TNR	PPV	NPV	F1-Score	Kappa	Confidence Score
Central Serous Chori-orentinpathy	97.58	86.63	75.25	98.01	42.22	99.52	54.09	52.95	67.19 ± 35.95
Diabetic Retinopathy	95.71	93.66	88.93	98.38	95.58	95.75	92.14	89.19	86.91 ± 28.73
Disc Edema	98.82	94.40	89.76	99.04	69.51	99.75	78.35	77.75	88.31 ± 25.94
Glaucoma	84.65	76.90	61.23	92.57	73.62	87.59	66.86	56.98	52.08 ± 26.97
Healthy	90.29	84.76	75.78	93.74	74.19	94.22	74.98	68.95	63.01 ± 27.19
Macular Scar	94.81	82.32	67.34	97.30	69.37	97.04	68.34	65.52	63.01 ± 37.35
Myopia	92.48	85.36	76.60	94.13	57.42	97.49	65.64	61.52	68.60 ± 30.59
Pterygium	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00 ± 0.00
Retinal Detachment	99.42	96.97	94.40	99.54	83.10	99.87	88.39	88.09	92.72 ± 19.77
Retinitis Pigmentosa	99.04	94.61	89.93	99.29	77.16	99.73	83.06	82.57	89.47 ± 26.00

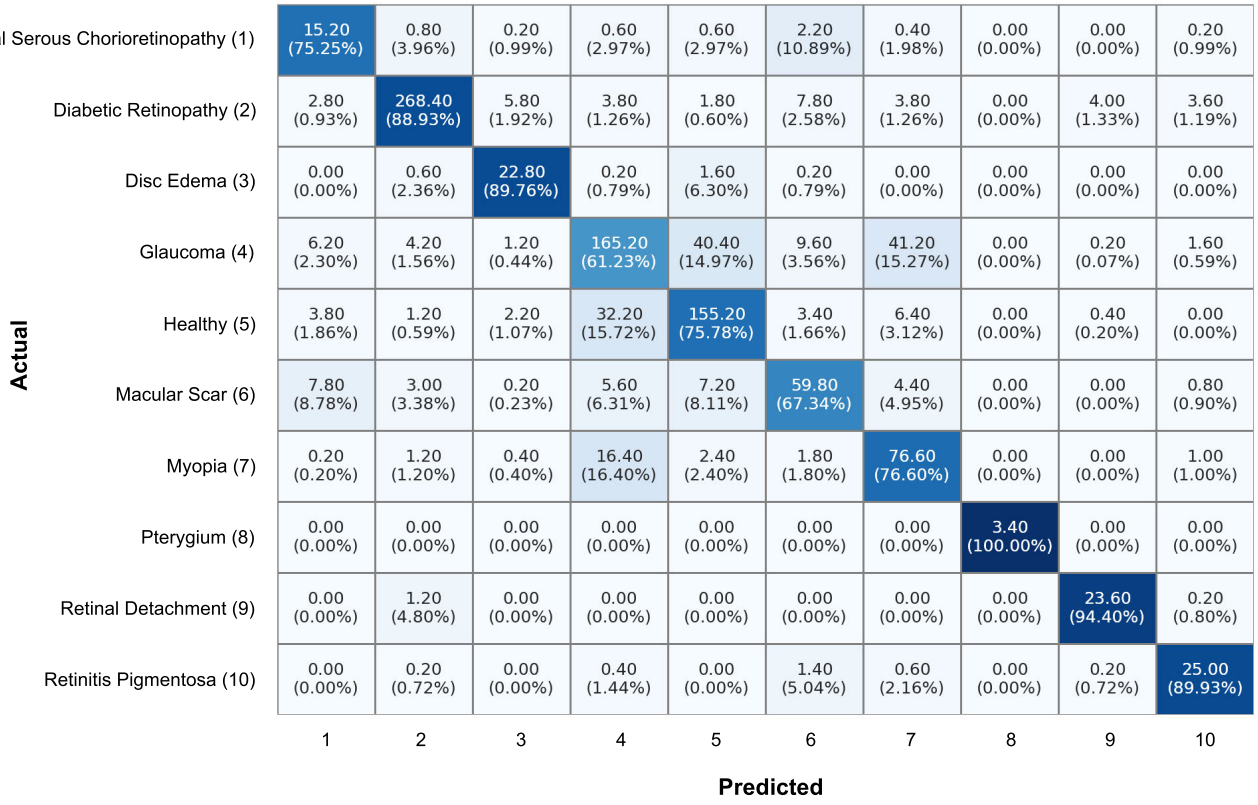

FIGURE 8. Row-normalized average confusion matrix across all five folds. The absolute values are shown, with their corresponding percentage shares displayed below, rounded to two decimal places.

factor for screening use, where false positives can lead to unnecessary interventions. Figure 7 compares the achieved average balanced and overall accuracy of the baseline models EfficientNetB4 and Swin Transformer V2 with those of the proposed hybrid architecture. It can be observed that the hybrid architecture outperforms both baselines, indicating that the fusion of CNN- and Transformer-based features is advantageous compared to relying on a single architecture. The largest improvement is seen when comparing EfficientNetB4 to the hybrid model, with gains of +3.17 percentage points (pp) in balanced accuracy and +1.89 pp in accuracy. In direct comparison to Swin Transformer V2, the hybrid architecture achieves improvements of +2.05 pp in balanced accuracy and +2.12 pp in accuracy. These results suggest

that, in this case, Transformer-based features contribute more significantly to performance than CNN-based features. In this study, we used a real-world dataset containing clinically relevant diseases. As shown in Table 3, the images have a signal-to-noise ratio of 30.04 dB with a standard deviation of 2.70 dB, indicating that our model was trained on images with varying levels of noise. Similarly, the mean brightness is 0.3642 with a standard deviation of 0.0784, meaning that different brightness levels were also considered during training. Considering that we employed stratified 5-fold cross-validation for evaluation, our results demonstrate that the proposed approach already exhibits strong robustness and generalization. Nevertheless, comparability with other datasets may be limited. The dataset we used contains ten classes, providing broader disease coverage than many other datasets and thereby increasing the complexity of the classification task. At the same time, this diversity offers a more realistic representation of clinical conditions, where even rare diseases must be accurately identified. Consequently, our hybrid architecture may yield different results for diseases such as glaucoma or DR when applied to other datasets, due to variations in image quality, disease stage, or the number of disease classes. In future work, we plan to evaluate the hybrid architecture on additional external fundus image datasets to further assess its generalizability.

B. CONTEXTUALIZATION WITHIN EXISTING LITERATURE

Compared to the studies in Table 1, our work stands out by employing a hybrid architecture that integrates EfficientNet-B4 with Swin Transformer V2 to unify local and global feature processing. This combination strategically addresses the limitations of conventional CNN-only models, which often excel in capturing localized patterns but struggle with long-range contextual dependencies; an essential factor in retinal disease diagnostics, where pathologies can manifest diffusely and across scales [1], [4]. While previous studies focusing on limited coverage of eye diseases reported high accuracy rates, the models were trained on small, non-peer-reviewed datasets with a narrow diagnostic focus. Using models for the broad coverage of eye diseases is a step closer to reality. These models reflect the complexity of the diseases and are more useful in practice. Binary models only indicate the presence of a disease, not which disease is present. This means that at least one human worker must be available for manual classification, which does not reduce the workload or save time. Although multi-class approaches, such as those of Babaqi et al. [16], and Wahab Sait and Rahman [15], are technically sophisticated, they are usually based on proprietary or unvalidated datasets and lack rigorous cross-validation. Even though cross-validation was used, there are differences in our work that need to be considered. Glaret subin and Muthukannan [18] used a non-peer-reviewed dataset consisting of 3,000 images (600 per class), which ensures balance within the classes. In contrast, we used 5,335 images and ten different disease classes. Additionally, there is no balance within the classes. Compared to the non-peer-reviewed multiclass approaches listed in Table 1, we use a peer-reviewed dataset and 5-fold cross-validation, which ensures robustness and transparency. Additionally, the ten-class classification problem we address here covers a broad clinical spectrum. This diversity contrasts strongly with the three to eight class settings commonly used in related work, making our model more reflective of real-world diagnostic challenges. If the quality criterion of broad coverage of eye diseases is applied, only two studies can be considered comparable to the present work. The study by He et al. [19] also utilizes a peer-reviewed dataset. However, the dataset includes only eight classes, one of which groups together several undefined diseases. Effectively, this results in six specific disease classes, one ''healthy'' class, and one broad ''other'' category-though this grouping is not further discussed in the study. Notably, the ''other'' class contains the third-highest number of images, which may influence model behavior. The evaluation was conducted using a 3-fold cross-validation, with no indication that stratification was applied. This raises the risk of uneven class distribution across the folds, potentially biasing the results. In contrast, our approach employs stratified 5-fold cross-validation to mitigate this risk and ensure more balanced evaluation. Fahdawi et al. [20] use a dataset similar to that in [19], consisting of eight classes: one for healthy cases, six for specific diseases, and one for other unspecified conditions. However, their study does not apply a robust evaluation strategy. Instead, the model is assessed using a single, fixed split of training and test data. No alternative data-splitting scenarios were explored. This limited evaluation setup does not account for potential variability in data distribution, making it difficult to assess the true robustness and generalizability of the proposed model across different clinical scenarios. Overall, this work is among the few studies that focus on the classification of multiple eye diseases using a peer-reviewed dataset. A total of ten classes are analyzed, including nine distinct disease classes and one representing healthy cases, making it the study with the highest number of disease-specific categories compared to existing literature. The proposed approach, which combines local and global feature representations in a novel way, demonstrates robust performance and is evaluated using stratified five-fold cross-validation, ensuring a reliable comparison with related studies.

C. ERROR TYPES

The misclassifications, observed in Figure 8, reveal several recurrent error patterns that point to underlying structural and data-related limitations of the hybrid architecture. These errors can be broadly categorized into inter-class ambiguities, anatomical overlaps, and effects stemming from class imbalance and low prevalence. A frequent and significant confusion occurred between Glaucoma, Myopia, and Healthy eyes. These categories share structural similarities in the optic nerve head and peripapillary area, particularly in early

disease stages [4]. In the confusion matrix, Glaucoma was misclassified as Myopia in 15.27% of cases and as Healthy in 14.97% of cases, while class Healthy was mistakenly classified as Glaucoma with 15.72%. This asymmetrical pattern, where healthy eyes are more often predicted as diseased than the other way around, suggests that the model lacks specificity and tends to overdiagnose disease. Another reason for the misclassification of healthy eyes as Glaucoma could be that the dataset contains images of glaucoma in its early stages, which may closely resemble the Healthy class. Another confusion arose with Macular Scar, which were often incorrectly classified as CSCR, Healthy or Glaucoma, which ultimately resulted in the second lowest TPR of 67.34% due to their frequency. This may be attributed to the relatively low contrast of atrophic or fibrotic lesions in standard fundus imaging, especially in the absence of hemorrhages or exudates [4]. Similarly, CSCR was most frequently confused with Macular Scar (10.89%), followed by DR (3.96%), Healthy (2.97%), and Glaucoma (2.97%). This confusion is likely due to overlapping central macular features such as shallow serous detachments or focal pigment abnormalities, which are difficult to distinguish in 2D fundus imagery [4]. Furthermore, depending on the stage, CSCR can be very similar to the Healthy class, which leads to difficulties in classification. Myopia reached an average TPR of 76.60%. Mainly because Myopia is misclassified as Glaucoma with 16.40%, a high value. Similarities of Myopia to Glaucoma, such as distorted parapapillary and macular structures [40], are mainly the reason. Confusions between DR, RP, and Disc Edema are likely due to overlapping vascular abnormalities such as hemorrhages, vessel attenuation, or microaneurysms [4]. One possible reason for the confusion between myopia and healthy eyes is the early stages of myopia, which show only subtle changes and are therefore difficult to distinguish. It is also possible that at a later stage of Myopia, this can cause damage to the macula, which resembles a scar and can therefore be confused with Macular Scar. Although these conditions differ etiologically, their late-stage funduscopic manifestations might appear similar when pigmentary patterns are subtle or partially obscured by image artifacts. Finally, certain classes, such as Myopia, appear to function as misclassification ''sinks'' for other categories, possibly due to their more visually salient features, such as a tessellated background or peripapillary atrophy. In addition to these confusions, the TPR of two classes was particularly low: Glaucoma with 61.23% and Macular Scar with 67.34%. This represents a significant imbalance in performance compared to other classes, such as Pterygium (100% TPR) or Retinal Detachment (94.40%). Therefore, both the low TPR of the first two and the relatively high TPR of the second two should be discussed in order to explain this imbalance. The high TPR of Pterygium is a logical consequence: images of this disease show a different aspect of the eyeball, which makes them easy to differentiate. Retinal Detachment involves the detachment of large parts of the retina, which is very clearly visible in retinal fundus images [41]. Since, unlike in other disease classes, it is not necessary to identify diseases on the retina, but rather its detachment as a whole, this classification task is also relatively easy for the model. Glaucoma and Macular Scar have a relatively low TPR of less than 70%. Since early forms of glaucoma are only indicated by subtle changes in the cup-to-disc ratio, i.e., the ratio of excavation to the outer edge of the optic nerve, this disease is particularly challenging to classify [42]. Macular Scar, on the other hand, is a ''collective class'' for scars that can develop on the retina for various reasons. Since the characteristics of this clinical picture are not necessarily consistent, this disease is equally difficult for the model to classify [43].

D. MEDICAL IMPLICATIONS

While the overall performance of the model demonstrates clinical potential, the observed misclassification patterns carry significant medical implications, particularly in terms of diagnostic reliability, risk stratification, and patient triage. For instance, frequent confusion between CSCR, Myopia, and Glaucoma may lead to incorrect triage of patients with macular detachment as low-priority or stable optic disc conditions, thereby delaying appropriate intervention. Similarly, Macular Scar was often mistaken for Healthy or Myopic eyes, potentially leading to a failure in identifying patients with chronic structural damage that could benefit from low-vision rehabilitation or further diagnostics. The misidentification of Disc Edema as DR or Glaucoma poses more acute risks, given that optic disc swelling may reflect serious underlying conditions such as increased intracranial pressure or inflammatory processes [2]. If left undetected, such misclassifications could delay life-saving neurological referrals [2]. Likewise, confusion with Retinal Detachment or RP might obfuscate the urgency of care. These examples underscore that even clinically plausible misclassifications can have serious consequences, particularly in screening or telemedicine settings where DL output may be interpreted with limited specialist oversight [44], [45]. Therefore, beyond accuracy metrics, it is essential to assess the model's failure patterns in terms of their potential clinical consequences, especially for high-risk conditions or categories that may initially present with subtle fundus changes. To mitigate these risks, any deployment in real-world settings must include well-defined fallback mechanisms, such as second-level human review, flagging of low-confidence predictions, and integration into structured clinical workflows.

E. PRACTICAL IMPLICATIONS

The proposed hybrid architecture shows strong potential for integration into clinical ophthalmology workflows, particularly in resource-limited settings where access to trained ophthalmic specialists is scarce. While an overall balanced accuracy level of 81.91% is insufficient for use as a

standalone diagnostic tool, our model has a quite different objective: to cover as broad a range of eye diseases as possible. This makes our architecture suitable not as an expert but as a generalist, hence it should be used in such settings. Studies such as those by Bernabé et al. [11] and Abbas [10] are highly specialized binary classifiers that achieve high performance for a single disease and can therefore be used as an aid for specific diagnoses, but this requires an initial suspicion of a particular disease. Our model provides exactly this initial suspicion and can therefore be seen as an assigning generalist that identifies suspected cases of a specific disease, which can then be confirmed by specialists. Ideal systems would deliver both the highest performance and broad coverage of eye disease, but these goals conflict from a technical point of view. It is therefore necessary to develop not only specialized highest performance systems but also generalized functional, broad diagnostic systems. In this role as a generalist, our system can be used as a decision-making aid to identify and prioritize patients with a variety of eye diseases at an early stage. Unlike approaches limited to binary or narrowly defined classifications, this system is capable of distinguishing among nine distinct disease categories plus healthy cases, making it particularly valuable for pre-screening applications in environments such as telemedicine platforms [45], public health screening initiatives, and general practitioner offices [44]. This broader diagnostic capability is especially advantageous for non-specialist healthcare providers, offering a reliable triage solution that helps bridge the gap created by the lack of immediate specialist access [44]. In particular, it demonstrates a very high capability to reliably rule out certain conditions, as reflected by a TNR of 95.60% and a NPV of 94.06%. This high exclusion accuracy is of great practical relevance, as it reduces unnecessary follow-up examinations, lowers the clinical workload, and enables physicians to focus their attention on cases that are more likely to be pathological and therefore require further diagnostic investigation. Consequently, the model can help make screening processes more efficient while simultaneously enhancing patient safety. The model's modular structure, open-source implementation, and compatibility with commonly available imaging equipment (e.g., fundus cameras) further support its adaptability across a range of clinical settings. These features not only lower technical barriers to entry but also enable customization by healthcare providers, medical device manufacturers, and digital health startups. For meaningful real-world adoption, however, several prerequisites must be addressed. Chief among them are rigorous external validation across demographically and technically heterogeneous datasets, compliance with regulatory standards, and the development of intuitive, low-threshold user interfaces tailored to clinicians, technicians, and primary care personnel. Additionally, medical informatics departments and hospital IT managers could play a crucial role in ensuring interoperability with existing electronic health record systems and safeguarding data security. Beyond clinical contexts, the implications of such a model extend to adjacent fields. High-resolution images of the human eye are increasingly acquired in non-medical settings, ranging from eye-tracking in consumer electronics to routine optical measurements at optometrists or within augmented and virtual reality platforms. When processed through DL systems, these images may offer diagnostic signals for early-stage diseases, providing new avenues for preventive care. In this regard, stakeholders such as technology developers, public health agencies, insurance providers, and patient advocacy groups must be engaged to ensure responsible and equitable deployment.

F. THEORETICAL IMPLICATIONS

Our work has various implications that arise from the composition of the model and its performance. First, our hybrid model shows that **hybrid approaches** are particularly valuable in multiclass domains. Since certain diseases are best captured by local features, while others are reflected in global patterns, multiclass tasks are best approached by more flexible models. A model that can process both types of information is essential in medical applications. With this work, we contribute to the advancement of theory by highlighting the relevance of hybrid approaches, which are already applied in medical settings [46], [47], [48], within this specific context.

Second, our results indicate that the model is **robust to imbalanced data**. When examining the confusion matrix, no systematic biases in favor of the more frequent classes are exhibited. Even the most underrepresented classes, such as class 8 (pterygium) and class 3 (disc edema), are identified with high accuracy. This suggests that multiclass models can perform reliably even under realistically imbalanced conditions, providing a meaningful contribution to the discussion on generalizability in medical informatics.

Another point that should be discussed from a theoretical perspective concerns the fundamental **objective of models in medical settings**. What should an AI model ideally achieve in medicine? The key question is whether a model should be primarily optimized for correct positive classifications or whether its ability to confidently rule out certain conditions should also be seen as a distinct and valuable contribution. A multiclass model with a broad range of classes, for example ten, can serve an important purpose in this context. Even if it does not always provide the correct diagnosis positively, it can still rule out other diseases with high certainty. This ability to exclude specific diagnoses is especially important in clinical reality. If a model can state with high confidence that a presented case is, e.g., not retinal detachment, not pterygium, and not DR, then that alone offers substantial added value, even when the positively given prediction is somewhat uncertain. This opens the door to a new theoretical framework for evaluating medical AI. Instead of focusing solely on top one accuracy, we should move toward

more nuanced performance dimensions such as diagnostic exclusion capability and coverage across a broad range of disease classes.

VI. CONCLUSION

With the proposed hybrid architecture approach, we set a new benchmark for the broad classification of eye diseases with a state-of-the-art validation technique. The findings of this study underline a central insight: the effective diagnosis of complex, visually heterogeneous eye diseases cannot be achieved through monolithic model architectures alone. By utilizing the proposed hybrid architecture, this work demonstrates how combining convolutional inductive biases with attention-based contextual modeling can provide a superior solution to the challenges of multiclass classification in ophthalmology. The demonstrated classification accuracy reflects technical feasibility within a controlled dataset context; however, it also provokes critical reflection on the practical boundaries of such systems. Several misclassification patterns identified during the error analysis suggest that, despite the maturity of the current architecture, it remains sensitive to certain classes that may exhibit similar disease patterns. This is a phenomenon that is not uncommon in real-world diagnostics. Moreover, the results of this study carry a dual implication. On a methodological level, they call for further development in multimodal learning and interpretability. At a conceptual level, they highlight the importance of embedding DL systems into socio-technical infrastructures that take into account clinical workflows, regulatory requirements, and practitioner trust.

A. LIMITATIONS

Despite the promising results, this study is subject to several limitations that warrant consideration. The model was exclusively trained on high-quality color fundus images, limiting both modality diversity and robustness to real-world image artefacts. Pathologies better captured by optical coherence tomography may be missed, and degraded images in routine settings could reduce performance. Although the dataset used in this study reflects the realistic distribution of eye diseases in clinical practice, including the relative rarity of certain conditions, this imbalance may have introduced biased learning dynamics. Nevertheless, achieving robust classification performance on these minority classes is essential, since real-world models must be able to handle such distributions reliably. Furthermore, the results are limited to the ten disease categories included in the dataset, which already cover a broad spectrum; however, no conclusions can be drawn about other diseases that may occur in real-world settings. Lastly, the accuracy of DL models depends on the quality of the labeled data. Since our dataset is labeled based on human diagnoses, mislabeling due to errors is possible, affecting model performance. Correcting this would require expert re-evaluation, which was beyond the scope of this study.

B. FUTURE WORK

The hybrid model presented demonstrated robust classification performance on a high-quality, peer-reviewed dataset. However, there are starting points for further research, particularly regarding the model's performance and integration into real-world clinical application scenarios. Some studies [49], [50] have used the Gaussian filter as a preprocessing step, demonstrating a positive effect on accuracy. This suggests that further investigation of this and other preprocessing filters would be worthwhile. Future investigations may focus on extending the current approach by incorporating multimodal input such as OCT imaging and clinical information to further enhance accuracy and model robustness. In the future, we plan to further develop the hybrid architecture so that heat maps can be used effectively to visualize the model's decision-making process. To verify the model's external validity, it is essential to include diverse datasets with different geographic and technological origins. Including different imaging systems, patient groups, and care contexts reduces the risk of overfitting to site-specific characteristics and increases the model's validity for real-world medical care. At the same time, the potential use of the model in clinical decision support should be considered further, for example, in the context of primary care or telemedical screening initiatives [44], [45]. Not only technical aspects are relevant here, but also questions of user acceptance, interoperability, and prospective effectiveness in everyday clinical practice. Future studies should therefore empirically investigate how collaboration between human specialists and DL systems affects diagnostic accuracy, process throughput times, and patient-related endpoints.

REFERENCES

[1] World Health Organization. (2019). *World Report on Vision*. [Online]. Available: https://www.who.int/docs/default-source/documents/publications/world-vision-report-accessible.pdf

[2] L. Stunkel, R. A. Sharma, D. D. Mackay, B. Wilson, G. P. Van Stavern, N. J. Newman, and V. Biousse, "Patient harm due to diagnostic error of neuro-ophthalmologic conditions," *Ophthalmology*, vol. 128, no. 9, pp. 1356–1362, Sep. 2021.

[3] P. Ruamviboonsuk et al., "Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program," *NPJ Digit. Med.*, vol. 2, Apr. 2019, Art. no. 25.

[4] S. Sharmin, M. R. Rashid, T. Khatun, M. Z. Hasan, M. S. Uddin, and Marzia, "A dataset of color fundus images for the detection and classification of eye diseases," *Data in Brief*, vol. 57, Dec. 2024, Art. no. 110979.

[5] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, Jun. 2019, pp. 6105–6114.

[6] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12009–12019.

[7] J. Xin, A. Wang, R. Guo, W. Liu, and X. Tang, "CNN and Swin-transformer based efficient model for Alzheimer's disease diagnosis with sMRI," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105189.

[8] S. Hao, L. Zhang, Y. Jiang, J. Wang, Z. Ji, L. Zhao, and I. Ganchev, "ConvNeXt-ST-AFF: A novel skin disease classification model based on fusion of ConvNeXt and Swin transformer," *IEEE Access*, vol. 11, pp. 117460–117473, 2023.

[9] Y. Chen, J. Feng, J. Liu, B. Pang, D. Cao, and C. Li, "Detection and classification of lung cancer cells using Swin transformer," *J. Cancer Therapy*, vol. 13, no. 7, pp. 464–475, Jul. 2022.

[10] Q. Abbas, "Glaucoma-deep: Detection of glaucoma eye disease on retinal fundus images using deep learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, pp. 41–45, 2017.

[11] O. Bernabe, E. Acevedo, A. Acevedo, R. Carreno, and S. Gomez, "Classification of eye diseases in fundus images," *IEEE Access*, vol. 9, pp. 101267–101276, 2021.

[12] M. R. Hossain, S. Afroze, N. Siddique, and M. M. Hoque, "Automatic detection of eye cataract using deep convolution neural networks (DCNNs)," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Nov. 2020, pp. 1333–1338.

[13] K. Oh, H. M. Kang, D. Leem, H. Lee, K. Y. Seo, and S. Yoon, "Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images," *Sci. Rep.*, vol. 11, Jan. 2021, Art. no. 1897.

[14] X. Xu, Y. Guan, J. Li, Z. Ma, L. Zhang, and L. Li, "Automatic glaucoma detection based on transfer induced attention network," *Biomed. Eng. OnLine*, vol. 20, no. 1, Apr. 2021, Art. no. 39.

[15] A. R. Wahab Sait, "Artificial intelligence-driven eye disease classification model," *Appl. Sci.*, vol. 13, no. 20, Oct. 2023, Art. no. 11437.

[16] T. Babaqi, M. Jaradat, A. E. Yildirim, S. H. Al-Nimer, and D. Won, "Eye disease classification using deep learning techniques," in *Proc. IISE Annu. Conf. Expo*, 2023.

[17] J. Biswas, S. S. Hossain, S. M. Mustaquim, and I. M. Siddique, "Instantaneous classification and localization of eye diseases via artificial intelligence," *Eur. J. Adv. Eng. Technol.*, vol. 11, no. 3, pp. 45–53, Mar. 2024.

[18] P. Glaret subin and P. Muthukannan, "Optimized convolution neural network based multiple eye disease detection," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105648.

[19] J. He, C. Li, J. Ye, Y. Qiao, and L. Gu, "Multi-label ocular disease classification with a dense correlation deep neural network," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102167.

[20] S. Al-Fahdawi, A. S. Al-Waisy, D. Q. Zeebaree, R. Qahwaji, H. Natiq, M. A. Mohammed, J. Nedoma, R. Martinek, and M. Deveci, "Fundus-DeepNet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images," *Inf. Fusion*, vol. 102, Feb. 2024, Art. no. 102059.

[21] F. Grassmann, J. Mengelkamp, C. Brandl, S. Harsch, M. E. Zimmermann, B. Linkohr, A. Peters, I. M. Heid, C. Palm, and B. H. F. Weber, "A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography," *Ophthalmology*, vol. 125, no. 9, pp. 1410–1420, Sep. 2018.

[22] M. S. Tomar, A. K. Jhapate, R. Dronawat, R. Chaure, and M. Jhapate, "Automatic diabetic retinopathy detection in fundus images using multi-level fire hawk convolution neural network," *Res. Square*, Jul. 2024, doi: 10.21203/rs.3.rs-4506963/v1.

[23] N. Chea and Y. Nam, "Classification of fundus images based on deep learning for detecting eye diseases," *Comput., Mater. Continua*, vol. 67, no. 1, pp. 411–426, 2021.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[25] L. Zhao and Z. Zhang, "A improved pooling method for convolutional neural networks," *Sci. Rep.*, vol. 14, no. 1, Jan. 2024, Art. no. 1589.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, Jul. 2015, pp. 448–456.

[27] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3559–3568.

[28] A. Pak, A. Ziyaden, K. Tukeshev, A. Jaxylykova, and D. Abdullina, "Comparative analysis of deep learning methods of detection of diabetic retinopathy," *Cogent Eng.*, vol. 7, no. 1, Jan. 2020, Art. no. 1805144.

[29] C.-Y. Zhu, Y.-K. Wang, H.-P. Chen, K.-L. Gao, C. Shu, J.-C. Wang, L.-F. Yan, Y.-G. Yang, F.-Y. Xie, and J. Liu, "A deep learning based framework for diagnosing multiple skin diseases in a clinical environment," *Frontiers Med.*, vol. 8, Apr. 2021, Art. no. 626369.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Apr. 2012.

[31] D. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *Int. J. Mach. Learn. Technol.*, vol. 2020, no. 1, pp. 37–63, Apr. 2020.

[32] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3121–3124.

[33] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[34] M. Buckland and F. Gey, "The relationship between recall and precision," *J. Amer. Soc. Inf. Sci.*, vol. 45, no. 1, pp. 12–19, Jan. 1994.

[35] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices," *Remote Sens.*, vol. 16, no. 3, Jan. 2024, Art. no. 533.

[36] T. F. Monaghan, S. N. Rahman, C. W. Agudelo, A. J. Wein, J. M. Lazar, K. Everaert, and R. R. Dmochowski, "Foundational statistical principles in medical research: Sensitivity, specificity, positive predictive value, and negative predictive value," *Medicina*, vol. 57, no. 5, May 2021, Art. no. 503.

[37] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance metrics for multilabel emotion classification: Comparing micro, macro, and weighted F1-scores," *Appl. Sci.*, vol. 14, no. 21, Oct. 2024, Art. no. 9863.

[38] J. Immerkær, "Fast noise variance estimation," *Comput. Vis. Image Understand.*, vol. 64, no. 2, pp. 300–302, Sep. 1996.

[39] H.-S. Gan, M. H. Ramlee, Z. Wang, and A. Shimizu, "A review on medical image segmentation: Datasets, technical models, challenges and solutions," *WIREs Data Mining Knowl. Discovery*, vol. 15, no. 1, Jan. 2025, Art. no. e1574.

[40] J.-A. Kim, H. Yoon, D. Lee, M. Kim, J. Choi, E. J. Lee, and T.-W. Kim, "Development of a deep learning system to detect glaucoma using macular vertical optical coherence tomography scans of myopic eyes," *Sci. Rep.*, vol. 13, no. 1, May 2023, Art. no. 8040.

[41] N. G. Ghazi and W. R. Green, "Pathology and pathogenesis of retinal detachment," *Eye*, vol. 16, no. 4, pp. 411–421, Jul. 2002.

[42] R. N. Weinreb, T. Aung, and F. A. Medeiros, "The pathophysiology and treatment of glaucoma: A review," *JAMA*, vol. 311, no. 18, pp. 1901–1911, 2014.

[43] E. Daniel, C. A. Toth, J. E. Grunwald, G. J. Jaffe, D. F. Martin, S. L. Fine, J. Huang, G.-S. Ying, S. A. Hagstrom, K. Winter, and M. G. Maguire, "Risk of scar in the comparison of age-related macular degeneration treatments trials," *Ophthalmology*, vol. 121, no. 3, pp. 656–666, Mar. 2014.

[44] A. L. De Araujo, T. D. C. Moreira, D. R. V. Rados, P. B. Gross, C. G. Molina-Bastos, N. Katz, L. Hauser, R. Souza Da Silva, S. D. Gadenz, R. G. D. Moro, F. C. Cabral, L. Matturro, C. G. M. Pagano, A. G. Faria, M. Falavigna, A. C. Da Silva Siqueira, P. Schor, M. R. Gonçalves, R. N. Umpierre, and E. Harzheim, "The use of telemedicine to support Brazilian primary care physicians in managing eye conditions: The TeleOftalmo project," *PLoS ONE*, vol. 15, no. 4, Apr. 2020, Art. no. e0231034.

[45] R. S. Meshkin, G. W. Armstrong, N. E. Hall, E. J. Rossin, M. B. Hymowitz, and A. C. Lorch, "Effectiveness of a telemedicine program for triage and diagnosis of emergent ophthalmic conditions," *Eye*, vol. 37, no. 2, pp. 325–331, Jan. 2022.

[46] J. R. Dixon, O. Akinniyi, A. Abdelhamid, G. A. Saleh, M. M. Rahman, and F. Khalifa, "A hybrid learning-architecture for improved brain tumor recognition," *Algorithms*, vol. 17, no. 6, 2024, Art. no. 221.

[47] O. Chibuike and X. Yang, "Convolutional neural network–vision transformer architecture with gated control mechanism and multi-scale fusion for enhanced pulmonary disease classification," *Diagnostics*, vol. 14, no. 24, Dec. 2024, Art. no. 2790.

[48] M. Kaddes, Y. M. Ayid, A. M. Elshewey, and Y. Fouad, "Breast cancer classification based on hybrid CNN with LSTM model," *Sci. Rep.*, vol. 15, Feb. 2025, Art. no. 4409.

[49] R. Buettner, C. Mai, and P. Penava, "Improvement of deep learning models using retinal filter: A systematic evaluation of the effect of Gaussian filtering with a focus on industrial inspection data," *IEEE Access*, vol. 13, pp. 43201–43217, 2025.

[50] L. Fischer-Brandies, L. Müller, J. J. Riegger, and R. Buettner, "Fresh or rotten? Enhancing rotten fruit detection with deep learning and Gaussian filtering," *IEEE Access*, vol. 13, pp. 31857–31869, 2025.

CELINA RIECK received the B.A. degree in political sciences from the Helmut-Schmidt-University/University of the Federal Armed Forces Hamburg, Germany, where she is currently pursuing the M.A. degree in digital governance and administration. She is also a Student Research Assistant with the Chair of Hybrid Intelligence, Helmut-Schmidt-University/University of the Federal Armed Forces Hamburg.

CHRISTOPHER MAI received the B.Sc. and M.Sc. degrees in mechanical engineering from Brandenburg University of Technology Cottbus-Senftenberg, Germany, in 2019 and 2023, respectively. He is currently pursuing the Ph.D. degree with the chair of Hybrid Intelligence, Helmut-Schmidt-University/University of the Federal Armed Forces Hamburg. His research interests include machine learning and deep learning.

LUCA EISENTRAUT received the B.Sc. degree in industrial engineering, the M.Sc. degree in business administration, and the M.Sc. degree in mechanical engineering from the University of Bayreuth, in 2020, 2023, and 2024, respectively. He is currently pursuing the Ph.D. degree with the Chair of Hybrid Intelligence, Helmut-Schmidt-University/University of the Federal Armed Forces Hamburg, Hamburg, Germany. His research interests include the development and application of deep learning techniques, especially in manufacturing and deepfakes.

RICARDO BUETTNER (Senior Member, IEEE) received the Dipl.-Inf. degree in computer science and the Dipl.-Wirtsch.-Ing. degree in industrial engineering and management from the Technical University of Ilmenau, Germany, the Dipl.-Kfm. degree in business administration from the University of Hagen, Germany, the Ph.D. degree in information systems from the University of Hohenheim, Germany, and the Habilitation (venia legendi) degree in information systems from the University of Trier, Germany. He is currently a Chaired Professor of Hybrid Intelligence with the Helmut-Schmidt-University/University of the Federal Armed Forces Hamburg, Germany. He has published more than 170 peer-reviewed articles, including articles in *Electronic Markets*, *AIS Transactions on Human-Computer Interaction*, *Personality and Individual Differences*, *European Journal of Psychological Assessment*, *PLOS One*, and IEEE Access. He has received 21 international best paper, best reviewer, and service awards and award nominations, including the Best Paper Awards by *AIS Transactions on Human-Computer Interaction*, *Electronic Markets*, and *HICSS*, for his work.