Integrating AI in Ophthalmology: A Deep Learning Approach for Automated Ocular Toxoplasmosis Diagnosis

Bader S. Alawfi
Department of Clinical Laboratory Sciences-College of Applied Medical Sciences,
Taibah University, Madinah 42353, Saudi Arabia

***Abstract*—Background: Ocular Toxoplasmosis, a leading cause of Posterior Uveitis, demands timely diagnosis to prevent vision loss. Manual retinal image analysis is labor-intensive and variable, while existing Deep Learning models often fail to balance local details and global context in Medical Image Classification. Objective: I propose RetinaCoAt, a Hybrid Deep Learning Model based on the CoAtNet Architecture, for Automated Diagnosis of Ocular Toxoplasmosis, integrating local and global features in Retinal Image Analysis. Methods: RetinaCoAt combines Convolutional Neural Networks (CNNs) for local pathological pattern detection with Transformer Models using multi-head self-attention for global context. Enhanced by residual connections and optimized tokenization, it was trained on 3,659 retinal images (healthy vs. unhealthy) and benchmarked against VGG16, CNNs, and ResNet. Results: RetinaCoAt achieved 98% accuracy in Medical Image Classification, outperforming VGG16 (96.87%), CNNs (95%), and ResNet (93.75%), due to its robust CNN-Transformer synergy. Conclusion: RetinaCoAt advances Automated Diagnosis of Ocular Toxoplasmosis and Posterior Uveitis, with potential for broader retinal pathology detection.**

***Keywords*—Ocular Toxoplasmosis; Posterior Uveitis; deep learning; automated diagnosis; CNNs; transformer models; CoAtNet architecture; retinal image analysis; medical image classification; hybrid deep learning models**

I. Introduction

Ocular Toxoplasmosis (OT) is a parasitic disease caused by Toxoplasma gondii, leading to necrotizing retinochoroiditis, the most common cause of posterior uveitis worldwide [1]. It primarily results from reactivation of latent retinal tissue cysts, though primary infection can also cause ocular involvement, particularly in congenital cases. The disease arises from both direct parasitic damage and the host's immune-mediated inflammatory response.

Pathogenesis involves cyst rupture in the retina, releasing bradyzoites that transform into tachyzoites, triggering a strong local immune response. $CD4+$ and $CD8+$ T-cells and cytokines like IFN-y mediates the inflammation, causing necrosis of retinal and choroidal tissue. Clinically, this manifests as sharply defined retinal lesions, often accompanied by a vitreous haze described as a headlight in the fog. Recurrence is common due to the parasite's persistence, leading to cumulative retinal damage and scarring.

Patients often present with unilateral vision changes such as blurred vision, floaters, photophobia, and eye pain and redness [2]. Macular or optic nerve involvement can result in severe, sometimes irreversible, vision loss. Diagnosis is typically clinical, supported by ophthalmoscopy, fundoscopic findings and serological tests showing T. gondii antibodies. PCR (Polymerase Chain Reaction) of ocular fluids may confirm the diagnosis in atypical cases but is difficult and complex [3],[4]. Imaging techniques like fundus photography, slit-lamp imaging, OCT (Optical Coherence Tomography) and fluorescein angiography provide detailed visualization of retinal lesions [5]. The result can sometimes be misleading or misinterpreted due to lab or several other conditions and can lead to muscular damage, vision loss or unnecessary treatment [6],[7],[8]. Prevention emphasizes avoiding undercooked meat and contaminated environments, especially for pregnant women and immunocompromised individuals. Despite advances, Ocular Toxoplasmosis continues to cause significant visual morbidity, necessitating further research into innovative therapies.

The complex and expensive clinical examination tests prompt us to use AI in this field, too as depicted in Fig. 1. Deep learning (DL), a specialized branch of machine learning, leverages artificial neural networks (ANNs), a framework inspired by the structure and function of the human brain. Unlike traditional computer vision techniques that require extensive feature engineering, DL models enable end-to-end learning, streamlining the analysis process [9]. These models have demonstrated remarkable success in automating image classification tasks, achieving significant advancements in the field [10].

In parasitology, DL-based networks have shown immense potential when applied to diagnostic imaging. For instance, CNNs have been used to detect and quantify parasitic infections in tissue or blood smear images. Its' potential to autonomously detect, classify, and quantify pathological features in ocular diseases holds significant promise for enhancing diagnostic ACC and enabling ophthalmologists to deliver precise and personalized care in the near future. DL is in use for the diagnosis of various eye diseases, analyzing infected fundus images like diabetic retinopathy, cataracts, and glaucoma [11],[12],[13],[14],[15],[16]. Additionally, these models have been trained to recognize disease-specific lesions, categorizing them by severity, a capability that can be extended to parasite-related pathological features in microscopy images [17].

For the first time (2019), Chakravarthy et al. designed an automated deep CNN (VGG-16) model for the diagnosis of OT [18]. They used heat mapping and patching as input to

Fig. 1. Use of AI in the disease management [5].

a hybrid model. Hasanreisoglu et al. also employed a dual input hybrid CNN-based approach for detection and achieved an ACC of 92%, making it a helpful aid [19]. Hassan and his team devised an automated machine learning model (without coding) to distinguish the fundus images of healthy eyes from the OT-infected eye images successfully [20]. Samira et al. compared DL algorithms ANN and CNN classifying fundus images using pre-trained VGG-16. The results indicated that ANN had better ACC than CNN even after preprocessing of three types of fundus images [21].

Alam et al. used the pre-trained models [22]. MobileNetV2 achieved better results for classification, followed by InceptionV3 in terms of ACC, while DenseNet121 showed the highest precision (PRI). In the case of segmentation, MobileNetV2/U-Net outclassed ResNet34. Other than evaluating the efficiency of models, they also analyzed the feature extraction methods to find the most suitable ones for the detection and segmentation of fundus images.

The automated detection of Ocular Toxoplasmosis presents unique challenges due to the need for precise localization of pathological features and the integration of global contextual information in retinal images. While deep learning has shown promise in medical image analysis, existing models often fall short in addressing these challenges, limiting their diagnostic ACC and clinical applicability. To bridge this gap, this study introduces a novel hybrid deep learning architecture that leverages the strengths of CNNs and transformer-based attention mechanisms. The proposed model not only addresses the limitations of current approaches but also sets a new standard for performance and robustness in the detection of Ocular Toxoplasmosis. The primary contributions of this study are as follows:

1) Novel hybrid architecture: This study aims to develop and evaluate a novel RetinaCoAt deep learning architecture that integrates CNNs with transformer-based attention mechanisms. This architecture is specifically designed to capture both local pathological patterns and global contextual information in retinal images, addressing the limitations of existing methods.

2) Pioneering work in Ocular Toxoplasmosis detection: To the best of my knowledge, this is the first study to develop an advanced deep-learning architecture specifically for the automated detection of Ocular Toxoplasmosis. The proposed model fills a critical gap in the literature and provides a foundation for future research in this domain.

3) State-of-the-art performance: The proposed model achieves an ACC of 98%, along with weighted PRI, recall (REC), and F1-score (F1S) of 98% and a perfect ROC score of 1.00. These results demonstrate its superior performance compared to existing models such as VGG16, traditional CNNs, and ResNet, setting a new benchmark for automated detection of Ocular Toxoplasmosis.

The remainder of this paper is organized as follows: Section II describes the materials and methods used in this study, including the dataset, preprocessing techniques, and the proposed RetinaCoAt hybrid architecture. Section III presents the experimental results, along with a detailed discussion of the model's performance, comparative analysis with existing methods, and an evaluation of its robustness and generalizability. Finally, Section IV concludes the paper by summarizing the key findings, highlighting the significance of the proposed work, and suggesting directions for future research.

II. Materials and Methods

This study presents an innovative approach to Ocular Toxoplasmosis classification through the implementation of a hybrid Convolutional Neural Network-Transformer architecture. It harnesses the synergistic combination of convolutional and transformer mechanisms to capture both local and global features in ocular images, enabling robust discrimination between healthy and pathological cases, as shown in Fig. 2.

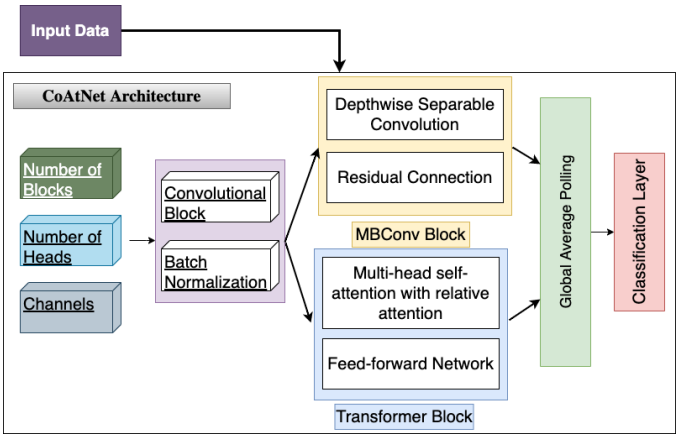

Fig. 2. Proposed architecture graphical representation.

A. Dataset Description and Preprocessing

Vision impairment and blindness is a common disease caused by Toxoplasma gondii. This research focuses on Ocular Toxoplasmosis detection utilizing retinal fundus images. The study involves two versions of the same dataset, where Dataset 2 is derived from Dataset 1 through preprocessing and augmentation to address class imbalance issues.

Dataset 1 (Original Dataset - Ocular Toxoplasmosis fundus images dataset):

- Collected from two hospital centers:
  - Hospital de Clínicas Medical Center (2018-2020): 291 images
  - Niños de Acosta Ñú General Pediatric Hospital (2021): 121 images
- Original multi-class distribution (showing class imbalance):
  - Healthy Eye: 132 images
  - Active: 33 images
  - Inactive: 187 images
  - Active-Active: 1 image
  - Active-Inactive: 57 images
  - Inactive-Inactive: 1 image
- Image specifications:
  - Format: JPG
  - Resolution: Varies (2124 x 2056 pixels or 1536 x 1152 pixels)

Dataset 2 (Processed and Augmented Version): To address the limitations of Dataset 1, the following modifications were made:

- Simplified Classification: Merged all disease categories into a single "Unhealthy" class
- Applied Data Augmentation: Increased the dataset size to improve model generalization

The processed dataset is comprised of training and validation sets. The training set includes both original and augmented images. The original training set contains 132 healthy images and 234 unhealthy images, while the augmented training set significantly expands these numbers to 1320 healthy images and 2339 unhealthy images. The validation set consists of 27 healthy images and 56 unhealthy images.

This restructuring from Dataset 1 to Dataset 2 addresses three key challenges:

- Class imbalance in the original dataset
- Complexity of multiple disease categories
- Limited sample size for deep learning applications

The resulting Dataset 2 provides a more balanced and augmented collection of images specifically designed for binary classification tasks while maintaining the diversity of the original patient demographics from multiple hospitals.

The dataset comprises ocular images categorized into two classes (Fig. 3: healthy and unhealthy (Toxoplasmosis-affected) samples. To ensure robust model training, a comprehensive data preparation pipeline is implemented. The dataset was partitioned using a stratified sampling approach, with 85% allocated for training and 15% for testing. The training set was further subdivided, with 80% used for actual training and 20% for validation, maintaining class distribution across all splits.

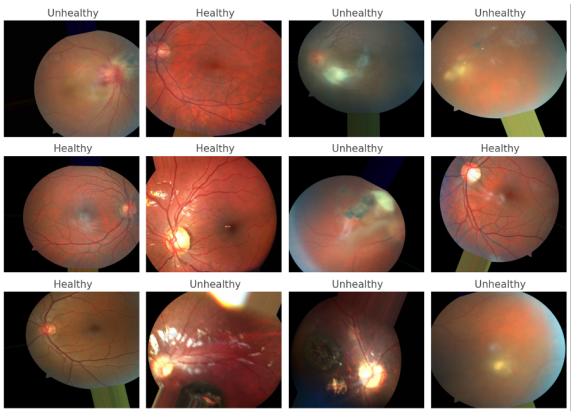

Fig. 3. Sample dataset images.

Image preprocessing was accomplished using TensorFlow's ImageDataGenerator, incorporating MobileNetV2's preprocessing function to normalize the input images. All images were resized to a uniform dimension of 128×128 pixels with RGB colour channels preserved. To enhance model generalization, data augmentation techniques employed through the Image Data Generator framework. The images were processed in batches of 32 samples, with shuffling enabled during training to prevent learning sequence-dependent patterns.

B. Proposed Model Architecture

The proposed architecture implements a hierarchical structure that progressively increases in complexity and receptive field size through multiple stages. The network architecture consists of three primary stages containing varying numbers of blocks (2, 2, 3) with corresponding channel dimensions (64, 96, 192). This progressive scaling enables the model to capture features at multiple levels of abstraction, from fine-grained local patterns to complex global structures.

1) Initial feature extraction: The network's initial stage implements a sophisticated feature extraction mechanism that serves as the foundation for all subsequent processing. This stage begins with a carefully engineered convolutional layer that processes the raw 128×128×3 RGB input images. The layer employs 7×7 kernels, a deliberate choice that creates a receptive field large enough to capture meaningful low-level features while maintaining computational efficiency. This kernel size represents an optimal balance between capturing sufficient spatial context and managing computational complexity, as smaller kernels might miss meaningful spatial relationships. In comparison, larger kernels would introduce unnecessary computational overhead.

The convolutional layer operates with a stride of 2, effectively downsampling the spatial dimensions while producing 64 output channels. This strided convolution serves a dual purpose: it reduces the spatial dimensions efficiently without requiring a separate pooling layer and helps establish translation invariance early in the network. The number of output channels

(64) was carefully selected to provide sufficient capacity for representing various low-level features such as edges, textures, and basic shapes present in ocular images while maintaining computational efficiency in subsequent layers.

Following the convolution, batch normalization implement with a momentum of 0.9, which plays a crucial role in stabilizing the training process. The batch normalization layer normalizes the feature distributions across the batch dimension, reducing internal covariate shifts and allowing for higher learning rates. This normalization process is critical in the initial layers, where feature magnitudes can vary significantly due to varying input image characteristics. The momentum value of 0.9 was chosen to provide a good balance between stable statistics and adaptability to changing feature distributions during training.

The network's initial stage implements a sophisticated feature extraction mechanism through a carefully designed convolutional layer. For an input image $X \in \mathbb{R}^{(HW3)}$, where H=W=128 represents the spatial dimensions, the initial convolution operation can be expressed as Eq. (1):

$$F_0(X) = \sigma(\text{BN}(W \cdot X + b)) \tag{1}$$

Where $W \in \mathbb{R}^{7\times7\times3\times64}$ represents the convolutional kernels, * denotes the convolution operation with stride 2, b represents the bias terms, BN denotes batch normalization, and σ is the activation function. The batch normalization operation normalizes the feature maps across the batch dimension B is shown as Eq. (2):

$$BN(x) = \gamma \left(\frac{x - \mu_a}{\sqrt{\sigma_a^2 + \epsilon}} \right) + \beta \tag{2}$$

where μ_a and σ_a^2 are the running estimates of mean and variance, γ and β are learnable parameters, and $\epsilon = 10^{-5}$ ensures numerical stability. This normalization significantly improves training stability by maintaining consistent feature distributions throughout the network.

2) MBConv block architecture: The Mobile Block Convolution (MBConv) blocks constitute a fundamental building block of network's early stages, implementing an efficient and powerful feature transformation mechanism. Each MBConv block follows a carefully designed expand-process-project pattern that maximizes feature extraction capability while maintaining computational efficiency. The expansion phase begins with a 1×1 pointwise convolution that increases the channel dimension by a factor of four. This expansion creates a higher-dimensional feature space that allows the network to capture more complex patterns and relationships. The expansion ratio of four was determined through empirical testing, providing an optimal balance between model capacity and computational overhead.

The expanded features undergo batch normalization followed by the SiLU (Swish) activation function, defined as $x * \sigma(x)$, where σ represents the sigmoid function. The SiLU activation was chosen over traditional ReLU due to its smooth nature and non-monotonic characteristics, which allow for better gradient flow and feature representation. The soft nature of SiLU helps prevent the "dying ReLU" problem while providing stronger regularization through its bounded nature at negative inputs.

The core processing stage employs a depthwise convolution with 3×3 kernels, a crucial architectural choice that dramatically reduces parameters while maintaining effective spatial feature extraction. This depthwise convolution processes each channel independently, applying spatial filtering without cross-channel mixing. The 3×3 kernel size provides a local receptive field that captures spatial relationships effectively while keeping the parameter count manageable. The depthwise convolution is followed by batch normalization and another SiLU activation, maintaining consistent feature processing throughout the block.

The projection phase implements another 1×1 pointwise convolution that reduces the channel dimensions back to their original size. This projection serves as a feature aggregation mechanism, combining the processed features from different channels into a more compact representation. The entire block incorporates a residual connection when input and output dimensions match, implemented through element-wise addition. This residual pathway serves multiple purposes: it facilitates gradient flow during backpropagation, helps maintain feature fidelity, and allows the network to learn residual mappings, which are often easier to optimize than direct mappings.

The MBConv blocks implement an efficient feature transformation pipeline that can be mathematically described through a series of operations as shown in Eq. (3). For an input tensor $X \in \mathbb{R}^{(HWC)}$, the expansion phase first projects the features to a higher dimension:

$$X_1 = \sigma(\text{BN}(W_1 \cdot X)) \tag{3}$$

where $W_1 \in \mathbb{R}^{(11C4C)}$ represents the expansion convolution weights. The subsequent depthwise convolution operates on each channel independently is expressed as Eq. (4):

$$X_2(i,j,k) = \sum_m \sum_n W_2(m,n,k) \cdot X_1(i+m, j+n, k) \tag{4}$$

where $W_2 \in \mathbb{R}^{(334C)}$ are the depthwise convolution kernels. The projection phase then reduces the dimensionality that depicts as Eq. (5):

$$Y = \sigma(\text{BN}(W_3 \cdot X_2)) \tag{5}$$

where $W_3 \in \mathbb{R}^{(1\cdot1\cdot4CC)}$ represents the projection weights. The residual connection, when applicable, is implemented as:

Output = $Y + X$ if shapes match Output = Y otherwise

The effectiveness of this architecture is demonstrated by the reduction in computational complexity from $O(H \cdot W \cdot C^2)$ for standard convolutions to $O(H \cdot W \cdot C)$ for depthwise separable convolutions while maintaining model expressiveness.

3) Transformer block design: The transformer blocks in network implement a sophisticated attention mechanism specifically adapted for image processing tasks, representing a significant advancement over traditional convolutional approaches. Each transformer block begins with layer normalization using a learned affine transformation, which standardizes the input features across the channel dimension. This normalization is crucial for stable training of the attention mechanism, as it ensures that the input features have consistent statistics regardless of their position in the network.

The core attention mechanism implements a multi-head relative attention approach, where the input features are processed by multiple attention heads operating in parallel. Each head processes a different subspace of the input features, allowing the network to capture various types of relationships simultaneously. The number of attention heads increases progressively through the network (1, 1, 2 in successive stages), allowing for more complex feature interactions in deeper layers. The relative attention mechanism incorporates spatial information by considering the relative positions of features, which is crucial for maintaining spatial awareness of the transformed features.

The attention computation follows the scaled dot-product formulation as expressed in Eq. (6):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Where Q, K, and V represent the queries, keys, and values, respectively, and d_k is the dimension of the key vectors. The scaling factor $\sqrt{d_k}$ prevents the dot products from growing too large in magnitude, which could push the softmax function into regions with minimal gradients. The queries, keys, and values are computed through learned linear transformations of the input features, allowing the network to adapt its attention patterns during training.

Following the attention mechanism, a sophisticated feed-forward network processes the attention output. This network consists of two dense layers with an intermediate expansion factor of four, chosen to provide sufficient capacity for complex feature transformation while maintaining computational efficiency. The GELU activation function is employed between these layers, providing non-linearity while maintaining smooth gradients. The GELU function approximates the expected transformation of a neuron's output under dropout regularization, providing an implicit form of regularization during training.

Transformer blocks implement a novel relative attention mechanism adapted for image processing. Given an input tensor $X \in \mathbb{R}^{(H \cdot W \cdot C)}$, the multi-head attention operation can be expressed as Eq. (7):

$$Q = W_Q \cdot X, \quad K = W_K \cdot X, \quad V = W_V \cdot X \tag{7}$$

where, $W_Q, W_K, W_V \in \mathbb{R}^{(C \cdot C)}$ are learnable weight matrices. The relative attention scores A for head h are computed as Eq. (8):

$$A_h = \text{softmax}\left(\frac{Q_h \cdot K_h^T + R_h}{\sqrt{d_k}}\right) \tag{8}$$

where, R_h represents the relative position encodings and d_k is the dimension per head. The final attention output is computed as Eq. (9):

$$\text{MultiHead}(X) = W_O \cdot \text{concat}(A_1 \cdot V_1, ..., A_H \cdot V_H) \tag{9}$$

where, H is the number of attention heads and $W_O \in \mathbb{R}^{(H \cdot C \cdot C)}$ is the output projection matrix. The feed-forward network applies two transformations as expressed in Eq. (10):

$$\text{FFN}(x) = W_2 \cdot \text{GELU}(W_1 \cdot x) \tag{10}$$

where, $W_1 \in \mathbb{R}^{(C \cdot 4 \cdot C)}$ and $W_2 \in \mathbb{R}^{(5 \cdot C)}$. The GELU activation is approximated as follows by Eq. (11):

$$\text{GELU}(x) \approx 0.5x\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right)\right) \tag{11}$$

4) Classification architecture: The classification stage of network implements a carefully designed sequence of operations that transform the high-level features into accurate class predictions. The stage begins with global average pooling, which reduces spatial dimensions while preserving channel information by computing the mean value across each feature map. This operation provides several advantages: it introduces translation invariance to the network's predictions, reduces the number of parameters compared to fully connected layers, and helps prevent overfitting by enforcing a structural regularization on the feature representations.

Following the pooling operation, dropout regularization implement with a carefully tuned rate of 0.2. This dropout rate was determined through extensive experimentation to provide optimal regularization without unnecessarily degrading model performance. During training, randomly deactivating 20% of the neurons helps prevent co-adaptation of feature detectors and encourages the network to learn more robust and independent features. The dropout mechanism also approximates an ensemble of multiple networks, providing implicit model averaging during inference.

The final classification layer consists of a dense layer with two output units, corresponding to binary classification task of distinguishing between healthy and unhealthy ocular images. The weights of this layer are initialized using the Glorot uniform initialization scheme, which helps maintain the appropriate scale of gradients through the network. The layer employs softmax activation to produce probability distributions over the two classes, defined as Eq. eq12:

$$P(class_i) = \frac{exp(z_i)}{\sum_j exp(z_j)} \tag{12}$$

where, z_i represents the logit for class i, the softmax activation ensures that the output probabilities sum to one

while providing a differentiable function that can be effectively optimized during training.

To improve the calibration of the model's predictions, implement temperature scaling in the softmax computation. The temperature parameter τ modifies the softmax function as shown in Eq. (13):

$$P(class_i) = \frac{exp(z_i/\tau)}{\sum_j exp(z_j/\tau)} \tag{13}$$

where, τ is learned during training to optimize the calibration of predicted probabilities, this calibration ensures that the model's confidence scores accurately reflect the actual likelihood of correct classification, which is crucial for clinical applications where uncertainty quantification is essential.

Training strategy employs the AdamW optimizer, which extends the traditional Adam optimizer with decoupled weight decay regularization. The optimizer is configured with an initial learning rate of 1e-3 and a weight decay factor of 1e-4, providing a balance between effective optimization and regularization. The beta parameters are set to 0.9 and 0.999 for the first and second moments, respectively, with an epsilon value of 1e-8 for numerical stability.

The learning rate management strategy incorporates both warmup and decay phases. During the initial five epochs, a linear warmup schedule gradually increases the learning rate to its maximum value, helping stabilize early training. Subsequently, a reduce-on-plateau scheme monitors validation loss and adjusts the learning rate when performance plateaus. The learning rate is reduced by a factor of 0.2 when no improvement is observed for five consecutive epochs, with a minimum learning rate threshold of $1e-6$.

The training utilizes sparse categorical cross-entropy loss with label smoothing ($\epsilon = 0.1$) to prevent overconfident predictions and improve generalization. The training process is monitored through multiple metrics, including ACC, loss, and AUC-ROC, with model checkpoints saved based on validation ACC. Early stopping with a patience of 10 epochs prevents overfitting by halting training when no further improvement is observed. Additional regularization is achieved through weight decay, dropout, and batch normalization, creating a robust training framework that balances model performance with generalization capability.

III. Results and Discussion

The proposed RetinaCoAt model's performance was evaluated across multiple metrics to comprehensively assess its effectiveness in classifying Ocular Toxoplasmosis into "Healthy" and "Unhealthy" categories. These metrics include training vs validation loss and ACC, a detailed classification report, a confusion matrix, an ROC curve, correct and incorrect predictions, and probability density distribution. Each metric provides unique insights into the model's performance, highlighting its ACC, generalization ability, and areas for improvement. The following subsections present a detailed analysis of these results.

A. Classification Report Generated by Proposed Model

The proposed deep learning model demonstrated excellent performance in classifying Ocular Toxoplasmosis images into healthy and unhealthy categories, as shown in Table I. The model achieved an overall ACC of 98% across the test set of 549 images. For healthy images (class 0), the model achieved a PRI of 0.98 and a REC of 0.97, resulting in an F1 score of 0.97. This indicates that the model was highly effective in identifying healthy cases, with very few false positives. Out of 219 healthy images in the test set, the model correctly classified 97% of them.

The model performed slightly better in identifying unhealthy images (class 1), achieving a PRI of 0.98 and REC of 0.98, with an F1-score of 0.98. From the 330 unhealthy images in the test set, 98% were correctly identified, demonstrating the model's strong capability in detecting pathological cases. The balanced performance across both classes is reflected in the macro-average metrics (PRI: 0.98, REC: 0.98, F1S: 0.98), indicating that the model performs consistently well regardless of the class. The weighted averages match these values, suggesting that the model maintains its high performance even when accounting for the slight class imbalance in the dataset.

TABLE I. Classification Report for the RetinaCoAt Model

Classes	PRI	REC	F1S	support
0	0.98	0.97	0.97	219
1	0.98	0.98	0.98	330
ACC			0.98	549
macro avg	0.98	0.98	0.98	549
weighted avg	0.98	0.98	0.98	549

B. Training vs Validation Loss and Accuracy

In Fig. 4, the training and validation curves reveal the learning progression of the model over 30 epochs. The loss curves (left plot) show a desirable convergence pattern. The training loss (red solid line) demonstrates a consistent decrease from an initial value of approximately 1.2, steadily declining and stabilizing around 0.02 by epoch 25. The validation loss (blue dashed line), while showing more fluctuation, follows a similar overall downward trend, ultimately converging to approximately 0.1, indicating best generalization.

The ACC curves (right plot) corroborate this learning behaviour. The training ACC (green solid line) shows steady improvement, starting from around 67% and rapidly increasing to over 80% within the first 5 epochs. It continues to improve more gradually thereafter, reaching nearly 100% by epoch 20. The validation ACC (orange dashed line), despite showing some oscillation in the early epochs, particularly around epoch 5, demonstrates overall improvement and eventually stabilizes above 95% after epoch 20.

The close alignment between training and validation metrics in the later epochs (20-30) suggests that the model has achieved a good balance between fitting the training data and generalizing to unseen examples. The minimal gap between final training and validation performance indicates that overfitting is well-controlled, due to effective regularization techniques employed in the model architecture.

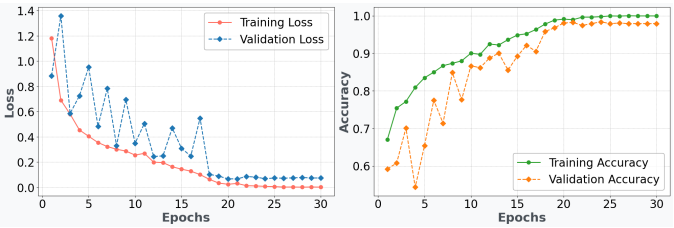

Fig. 4. Training and validation loss and accuracy curves.

C. Confusion Matrix of the Proposed Model

The confusion matrix (Fig. 5) reveals excellent classification performance across both healthy and unhealthy cases. Here's a detailed breakdown:

For healthy cases:

- True Negatives (TN): 213 healthy images were correctly classified as healthy
- False Positives (FP): Only 6 unhealthy images were incorrectly classified as healthy
- This represents a high specificity, with the model rarely misclassifying unhealthy cases as healthy

For unhealthy cases:

- True Positives (TP): 325 unhealthy images were correctly classified as unhealthy
- False Negatives (FN): Only 5 healthy images were incorrectly classified as unhealthy
- This demonstrates high sensitivity, with the model successfully identifying the vast majority of unhealthy cases

The model shows balanced performance with very few misclassifications in either direction (5 FN and 6 FP), which is particularly important in medical diagnosis applications. The nearly symmetric error rates suggest that the model is not biased toward either class despite the slight class imbalance in the dataset (219 healthy vs 330 unhealthy images).

Fig. 5. Confusion matrix for Ocular Toxoplasmosis classification.

D. Probability Density Distribution

In the analysis of the proposed model's prediction confidence, the probability density distribution reveals compelling insights into the model's classification behaviour for Ocular Toxoplasmosis cases, as shown in Fig. 6. The distribution exhibits a distinctive bimodal pattern, characterized by two prominent peaks that effectively separate healthy and unhealthy predictions. The left peak centred approximately at 0.0 on the probability scale, predominantly represents the healthy class predictions, displaying a higher density with a maximum value of approximately 1.75. This indicates the model's strong confidence in identifying healthy cases. Conversely, the right peak, positioned around 0.75-1.0 on the probability scale, corresponds to unhealthy class predictions, showing a slightly lower but still substantial density maximum of about 1.7. This right-side distribution demonstrates the model's robust confidence in identifying unhealthy cases. Notably, the region between these two peaks, particularly around the 0.5 probability mark, shows minimal density values, indicating that the model rarely produces uncertain or ambiguous predictions. This clear separation between the two classes' probability distributions strongly corroborates the model's high-performance metrics, with both classes showing well-defined, concentrated probability regions. The symmetrical nature of the peaks and their similar heights suggest balanced prediction confidence across both classes despite the slight class imbalance in the dataset. This balanced confidence distribution aligns well with the model's high ACC and balanced PRI-REC metrics observed in the classification report.

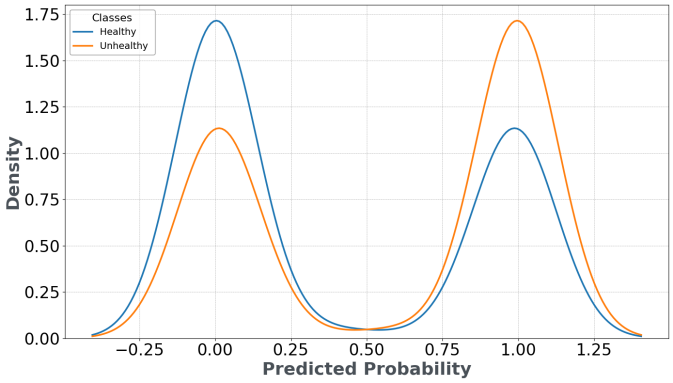

Fig. 6. Probability density distribution of predicted outputs.

E. Receiver Operating Characteristics

In Fig. 7, the Receiver Operating Characteristic (ROC) curves for the proposed Ocular Toxoplasmosis classification model demonstrate exceptional discriminative performance for both healthy and unhealthy classes. The graph displays three curves: ROC curves for healthy (blue line) and unhealthy (orange line) classes, along with a random chance baseline (black dashed line). Both classes achieve a perfect Area Under the Curve (AUC) score of 1.00, indicating optimal classification performance. The ROC curves for both classes immediately rise to the top-left corner of the plot and maintain a true positive rate of nearly 1.0 across all false positive rate thresholds. This is in stark contrast to the random chance baseline (diagonal dashed line), which represents an AUC of 0.50. The perfect AUC scores suggest that the model can perfectly distinguish between healthy and unhealthy cases at various classification thresholds, validating the model's robust decision-making capability. The identical performance across

both classes, as shown by the overlapping ROC curves, further confirms the model's balanced predictive power, regardless of the class imbalance in the dataset. This exceptional ROC performance aligns perfectly with the previously observed high ACC, PRI, and REC metrics, as well as the clear separation seen in the probability density distributions.

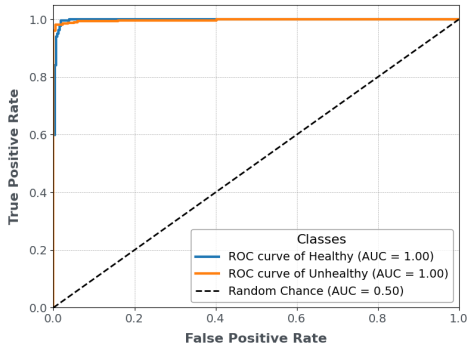

Fig. 7. ROC curve for the RetinaCoAt model.

F. Correct and Incorrect Predictions

Fig. 8 illustrates examples of both correct and incorrect predictions made by the proposed model in the classification task. The rows show retinal fundus images categorized into two groups: "Healthy" and "Unhealthy".

The first and second rows demonstrate cases of correct predictions where the model successfully identified the actual label, as indicated by matching "True" and "Pred" annotations. The third row showcases one instance where the model misclassified images, with the discrepancy highlighted in red text for easy identification (e.g. "True: Healthy Pred: Unhealthy"). These results underline the model's overall performance, achieving a classification ACC of 98%. However, the highlighted misclassifications emphasize the importance of addressing edge cases or ambiguous features within the dataset to improve robustness further.

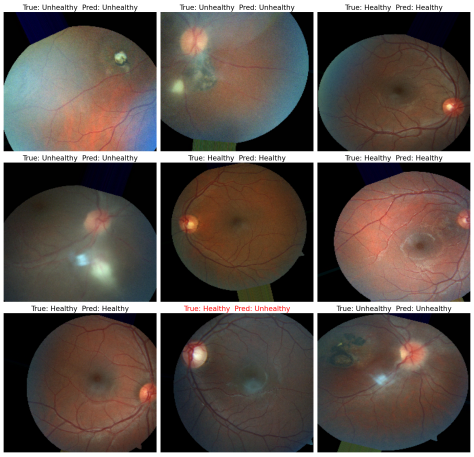

Fig. 8. Correct and incorrect predictions by the RetinaCoAt model.

G. Comparison with Existing State-of-the Art Work

The experimental results demonstrate the superior performance of the proposed RetinaCoAt architecture for Ocular Toxoplasmosis classification, achieving 98% ACC compared to existing approaches as depicted in Table II. This significant improvement over traditional methods can be attributed to RetinaCoAt's innovative hybrid design, which combines convolution and self-attention mechanisms. While conventional CNN-based approaches like VGG16 [23] achieved 96.87% ACC, and basic CNNs [24] reached 95%, they lack the sophisticated feature extraction capabilities of RetinaCoAt. The architecture surpasses ResNet [25] implementations (93.75%) by effectively addressing the limitations of purely convolutional approaches through its attention mechanisms, which capture complex spatial relationships in ocular images. Notably, the proposed method also outperforms automated approaches, with AutoML [26] models achieving 93.5% and Google Cloud AutoML [27] reaching 84.8% ACC. This performance gap highlights the advantage of a specially designed architecture that leverages both local feature extraction through convolutions and global context understanding through self-attention, making it particularly effective for the nuanced task of identifying Ocular Toxoplasmosis patterns in medical imaging.

TABLE II. COMPARISON WITH OTHER STUDIES

Reference	Proposed Method	Accuracy (%)
[23]	VGG16	96.87
[24]	Convolutional Neural Network	95
[25]	Residual Neural Network	93.75
[26]	AutoML model	93.5
[27]	AutoML in Google Cloud	84.8
Proposed	RetinaCoAt	98

IV. CONCLUSION

This study proposes a novel RetinaCoAt hybrid deep learning architecture for the automated detection of Ocular Toxoplasmosis in retinal images. The model, which integrates CNNs with transformer-based attention mechanisms, demonstrated exceptional performance, achieving an ACC of 98%, along with a weighted average PRI, REC, and F1S of 98%. Furthermore, the model achieved a perfect ROC score of 1.00, underscoring its robustness and reliability in distinguishing between healthy and infected cases.

To ensure the model's generalizability, training and validation loss and ACC were meticulously monitored, confirming that the model is not overfitted and is the best fit for the task. The proposed architecture addresses the limitations of existing methods by effectively capturing both local pathological patterns and global contextual information, enabling comprehensive multi-scale feature extraction.

A critical review of the literature reveals that no advanced architecture has been specifically designed for the automated detection of Ocular Toxoplasmosis, making this work a novel contribution to the field. The proposed model sets a new benchmark by leveraging the strengths of CNNs and transformers, offering a powerful tool for accurate and efficient diagnosis.

This study not only advances the development of automated diagnostic tools for Ocular Toxoplasmosis but also holds significant potential for improving early detection and

treatment outcomes. Future work could explore the application of this architecture to other ocular diseases and the integration of additional clinical data to further enhance its diagnostic capabilities.

Future work will focus on expanding the RetinaCoAt architecture to address additional challenges in ocular disease detection. I plan to extend model to classify the severity and progression stages of Ocular Toxoplasmosis, enabling more nuanced clinical decision-making. Integration with complementary imaging modalities such as Optical Coherence Tomography (OCT) could provide depth analysis of retinal lesions and enhance diagnostic accuracy. Additionally, developing explainable AI components would increase clinical trust by providing interpretable visualizations of the model's decision-making process. I also aim to investigate automated lesion segmentation capabilities and longitudinal analysis features to monitor treatment efficacy over time. Finally, clinical validation through prospective multi-center trials will be essential to establish the model's generalizability across diverse patient populations and imaging equipment. These advancements will collectively strengthen the clinical utility of proposed approach and potentially extend its application to other ocular pathologies with similar presentation patterns.

ACKNOWLEDGMENT

The authors would like to thank...

REFERENCES

[1] A. M. Tenter, A. R. Heckeroth, and L. M. Weiss, "Toxoplasma gondii: from animals to humans," *Int. J. Parasitol.*, vol. 30, no. 12-13, pp. 1217–1258, 2000.

[2] J. S. Remington, P. Thulliez, and J. G. Montoya, "Recent developments for diagnosis of toxoplasmosis," *J. Clin. Microbiol.*, vol. 42, no. 3, pp. 941–945, 2004.

[3] R. N. Van Gelder, "Cme review: polymerase chain reaction diagnostics for posterior segment disease," *Retina*, vol. 23, no. 4, pp. 445–452, 2003.

[4] L. R. Steeples, M. Guiver, and N. P. Jones, "Real-time PCR using the 529 bp repeat element for the diagnosis of atypical ocular toxoplasmosis," *Br. J. Ophthalmol.*, vol. 100, no. 2, pp. 200–203, 2016.

[5] Y. Tong, W. Lu, Y. Yu, and Y. Shen, "Application of machine learning in ophthalmic imaging modalities," *Eye Vis. (Lond.)*, vol. 7, no. 1, p. 22, 2020.

[6] J. G. Garweg, J. G. Montoya, and J. d. Groot-Mijnes, "Diagnostic approach to ocular toxoplasmosis," *Highlights of Ophthalmology*, vol. 44, no. 2ENG, pp. 6–10, 2016.

[7] C. Cifuentes-González, W. Rojas-Carabali, Á. O. Pérez, É. Carvalho, F. Valenzuela, L. Miguel-Escuder, M. S. Ormaechea, M. Heredia, P. Baquero-Ospina, A. Adan, A. Curi, A. Schlaen, C. A. Urzua, C. Couto, L. Arellanes, and A. de-la Torre, "Risk factors for recurrences and visual impairment in patients with ocular toxoplasmosis: A systematic review and meta-analysis," *PLoS One*, vol. 18, no. 4, p. e0283845, 2023.

[8] A. M. Shammaa, T. G. Powell, and I. Benmerzouga, "Adverse outcomes associated with the treatment of toxoplasma infections," *Sci. Rep.*, vol. 11, no. 1, p. 1035, 2021.

[9] N. Omahony, "Deep learning vs. traditional computer vision," in *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC)*. Springer, vol. 1.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] R. Raman, S. Srinivasan, S. Virmani, S. Sivaprasad, C. Rao, and R. Rajalakshmi, "Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy," *EYE*, vol. 33, no. 1, pp. 97–109, 2019.

[12] Y. Peng, S. Dharssi, Q. Chen, T. D. Keenan, E. Agrón, W. T. Wong, E. Y. Chew, and Z. Lu, "DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs," *Ophthalmology*, vol. 126, no. 4, pp. 565–575, 2019.

[13] S. Guo, K. Wang, H. Kang, T. Liu, Y. Gao, and T. Li, "Bin loss for hard exudates segmentation in fundus images," *Neurocomputing*, vol. 392, pp. 314–324, 2020.

[14] Y. Guo, R. Wang, X. Zhou, Y. Liu, L. Wang, C. Lv, B. Lv, and G. Xie, "Lesion-aware segmentation network for atrophy and detachment of pathological myopia on fundus images," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020.

[15] Y. Dong, Q. Zhang, Z. Qiao, and J.-J. Yang, "Classification of cataract fundus image based on deep learning," in *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE, 2017.

[16] N. Tsiknakis, D. Theodoropoulos, G. Manikis, E. Ktistakis, O. Boutsora, A. Berto, F. Scarpa, A. Scarpa, D. I. Fotiadis, and K. Marias, "Deep learning for diabetic retinopathy detection and classification based on fundus images: A review," *Comput. Biol. Med.*, vol. 135, no. 104599, p. 104599, 2021.

[17] Y. Yang, T. Li, W. Li, H. Wu, W. Fan, and W. Zhang, "Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks," in *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*. Cham: Springer International Publishing, 2017, pp. 533–540.

[18] A. D. Chakravarthy, D. Abeyrathna, M. Subramaniam, P. Chundi, M. S. Halim, M. Hasanreisoglu, Y. J. Sepah, and Q. D. Nguyen, "An approach towards automatic detection of toxoplasmosis using fundus images," in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2019.

[19] M. Hasanreisoglu, "Ocular toxoplasmosis lesion detection on fundus photograph using a deep learning model," *Investigative Ophthalmology & Visual Science*, vol. 61, no. 7, pp. 1627–1627, 2020.

[20] M. Hassan, "Utilization of automated deep learning approach toward detection of ocular toxoplasmosis using fundus photographs," *Investigative Ophthalmology & Visual Science*, vol. 64, no. 8, pp. 1093–1093, 2023.

[21] S. R. Ferdous, M. R. Ahasan Rifat, M. J. Ayan, and R. Rahman, "An approach to classify ocular toxoplasmosis images using deep learning models," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2023.

[22] S. S. Alam, S. B. Shuvo, S. N. Ali, F. Ahmed, A. Chakma, and Y. M. Jang, "Benchmarking deep learning frameworks for automated diagnosis of ocular toxoplasmosis: A comprehensive approach to classification and segmentation," *IEEE Access*, vol. 12, pp. 22 759–22 777, 2024.

[23] R. Parra, V. Ojeda, J. L. Vázquez Noguera, M. García-Torres, J. C. Mello-Román, C. Villalba, J. Facon, F. Divina, O. Cardozo, V. E. Castillo, and I. C. Matto, "A trust-based methodology to evaluate deep learning models for automatic diagnosis of ocular toxoplasmosis from fundus images," *Diagnostics (Basel)*, vol. 11, no. 11, p. 1951, 2021.

[24] P. K. Choudhury, A. A. Anika, S. R. Ramisa, A. Zaman, and R. R. Chowdhury, "Deep learning based automated diagnosis of ocular toxoplasmosis in fundus images using convolutional neural network," Ph.D. dissertation, Brac University, 2024.

[25] R. Parra, *Automatic diagnosis of ocular toxoplasmosis from fundus images with residual neural networks, in Public Health and Informatics*. IOS Press, 2021.

[26] D. Milad, F. Antaki, A. Bernstein, S. Touma, and R. Duval, "Automated machine learning versus expert-designed models in ocular toxoplasmosis: Detection and lesion localization using fundus images," *Ocul. Immunol. Inflamm.*, vol. 32, no. 9, pp. 2061–2067, 2024.

[27] C. Cifuentes-González, W. Rojas-Carabali, G. Mejía-Salgado, G. Flórez-Esparza, L. Gutiérrez-Sinisterra, O. J. Perdomo, J. E. Gómez-Marín, R. Agrawal, and A. de-la Torre, "Is automated machine learning useful for ocular toxoplasmosis identification and classification of the inflammatory activity?" *AJO International*, vol. 1, no. 4, p. 100079, 2024.