Vision Transformer for Cataract Detection in Anterior Eye Images: Achieving Optimal Performance with Limited Data

Nina Sevani[1*], Westlee Matthew Agustinus[2], and Edy Kristianto[3]
[1]Department of Informatics, Faculty of Smart Engineering, Universitas Kristen Krida Wacana
Jakarta, Indonesia 11470
[2]Research Center for Knowledge Management and Collaborative Innovation,
Universitas Kristen Krida Wacana
Jakarta, Indonesia 11470
[3]Research Center for Connected Intelligence and Security, Universitas Kristen Krida Wacana
Jakarta, Indonesia 11470
Email: [1]nina.sevani@ukrida.ac.id, [2]westlee.412021012@civitas.ukrida.ac.id,
[3]edy.kristianto@ukrida.ac.id

***Abstract*—Convolutional Neural Networks (CNNs) continue to face several challenges in cataract classification using image data, particularly due to limitations in dataset size and variability in color, shape, and position. These constraints arise because CNNs primarily process local features within each layer. The research uses Vision Transformer (ViT) to capture global spatial relationships in front-eye images. ViT can detect cataracts more accurately than CNN-based owing to its capability of modeling long-range dependencies. The research aims to evaluate ViT's performance on four publicly available cataract datasets using training time, accuracy, precision, recall, and F1-score. To assess the model's reliability when working with a small number of training datasets, ViT's performance is also be compared with ResNet-50 and EfficientNet-B7. The results indicate that ViT outperforms ResNet-50 in accuracy by 10%–31% and exceeds EfficientNet-B7 by 30%–41%. However, ViT requires approximately 16 seconds longer to train than ResNet-50 and EfficientNet-B7 due to its deeper architecture. Although ViT's accuracy is 2.47% lower than previous studies using hybrid deep learning approaches, its parameter structure is simpler. Overall, the findings indicate that, compared with CNN-based models, ViT performs more effectively when trained on smaller datasets. Future research may incorporate segmentation techniques to further validate cataract detection in anterior eye images.**

***Index Terms*—Cataract Detection, Vision Transformer (ViT), Anterior Eye Images**

Received: May 14, 2025; received in revised form: Nov. 30, 2025; accepted: Dec. 12, 2025; available online: Aug. 13, 2026.
*Corresponding Author

I. INTRODUCTION

CATARACT is one of the eye diseases which is marked by the cloudiness or greyish colour of the eye's lens. These diseases become one of the main causes of blindness in the world, due to the obstruction of the path of light to the lens [1]. Hence, early detection of cataract disease is important to prevent the severe vision disruption. In recent years, the development of image processing methods and deep learning become the focus in automatic cataract detection [2].

Deep learning, a branch of Artificial Intelligence (AI), utilizes Artificial Neural Networks (ANN) to enable automated learning. Within image processing, Convolutional Neural Networks (CNNs) are often used due to their ability to hierarchically extract the features. However, CNNs are inherently limited in their capacity to capture global spatial relationships, such as interactions among different regions of an image [3]

Several deep learning architectures have been widely applied in disease classification, including ResNet-50, VGG-16, and EfficientNet [4–11]. Experimental results demonstrate that these these models achieve strong performance across various tasks. In ocular disease detection, ResNet-50 has been particularly prominent, producing robust results with diverse image datasets. Studies employing ResNet-50 have reported accuracies of 78% [9], 92% [11], 97% [10], and 98% [6], highlighting the model's effectiveness in feature extraction [8, 12]. Meanwhile, EfficientNet a more recent architecture, has evolved from version B0 to B7 and

has consistently demonstrated superior results in image classification tasks, often surpassing earlier architectures [13]. Compared with ResNet-50, EfficientNet incorporates a larger number of parameters and a more complex architecture, which enables higher accuracy. However, these advantages come with limitations, including increased computational demands and longer training times [6]. Prior research on ocular disease classification using EfficientNet has reported accuracies of up to 94% [14].

A study on automatic cataract detection using ResNet-50 has reported sensitivity and specificity values of 92% and 83.85%, respectively. This research focuses on developing algorithms based on slit-lamp images and fundus images [15]. Another study utilized deep learning for cataract detection with retinal images has achieved an Area Under the Receiver Operating Characteristic (AUROC) value of 96.6% in internal testing, with external testing results ranging from 91.6% to 96.5% [16]. While these findings are promising, there are limitations in the previous research. A major issue concerns image quality, which is strongly influenced by dataset acquisition methods. Moreover, many algorithms developed in previous studies rely on CNN structures, which tend to be less effective at understanding complex spatial relationships within images. In addition, several studies face challenges in establishing subjective "ground truth" and often lack robust external validation [16].

Adopting a CNN-based deep learning approach still poses challenges in handling variations in the object's color [17], position, texture, and shape. This limitation affects the ability to effectively extract and aggregate features from images [18, 19]. Furthermore, the reliance on high-quality images in many previous deep learning studies has contributed to achieving strong accuracy results [20].

One deep learning model that shows great potential in overcoming these limitations is the Vision Transformer (ViT). Introduced by Dosovitskiy et al. in 2021, ViT features a different architecture compared to traditional CNNs. While CNNs rely on convolutional operations to extract features from images, ViT uses a streamlined architecture that maintains a consistent resolution across feature maps [21]. Specifically, it divides an image into fixed-size patches, converts these patches into vector representations, and processes them using a self-attention mechanism. Unlike CNNs, which primarily capture local features layer by layer, this mechanism enables ViT to model spatial relationships across the entire image more effectively [21–23].

ViT has demonstrated excellent performance in medical image segmentation and classification tasks, as evidenced by several studies [19, 23, 24]. Different versions of ViT have achieved superior results with smaller training datasets and without requiring additional preprocessing [25]. Compared with traditional CNNs, the Visformer model, an optimized extension of ViT for image processing, has shown notable accuracy gains. For instance, Visformer-S outperforms DeiT-S and ResNet-50 in ImageNet classification by margins of up to 2.12% and 3.46%, respectively, despite having comparable model complexity. Moreover, Visformer maintains higher accuracy than ResNet-50 and DeiT-S even when trained on only 10% of the dataset [26].

Previous studies have also indicated that ViT is more resilient to variations in data distribution and external disturbances compared to CNNs, while maintaining greater accuracy. The Robust ViT (RVT) model, for instance, surpasses ResNet-50 by 2.9% on ImageNet, achieving a top-1 accuracy of 81.9% with the same number of floating-point operations (FLOPs). In addition, RVT demonstrates greater robustness against adversarial attacks and image corruptions, making it a more reliable option in highly variable environments. With its improved accuracy, computational efficiency, and adaptability to data changes, ViT is emerging as a leading approach for various computer imaging tasks, including medical image segmentation and classification [26, 27].

Research on Transformer-based Knowledge Distillation Networks (TKD-Net) further illustrates that Transformer models can significantly improve cataract grading accuracy. Compared to traditional CNN techniques, TKD-Net increases prediction accuracy by 20%–30% by employing zone decomposition methodologies and multi-modal Transformers [2]. The findings from the TKD-Net implementation show that zone division and the inclusion of clinical subscores can increase prediction accuracy. This method also proves the model's reliability in the presence of insufficient data.

The potential of ViTs in cataract surgery image processing has been highlighted by studies exploring AI applications in this field, particularly for surgical instrument segmentation and process classification. While CNNs remain widely used, ViTs are attracting increasing interest due to their ability to capture complex spatial correlations in medical images [28]. In essence, ViTs provide an alternative to CNN-based models. Previous research has also shown that ViTs can achieve higher accuracy than pre-trained CNN-ResNet models while requiring fewer resources [29].

A key factor influencing the performance of the ViT model during cataract classification training is the optimization algorithm employed. Previous research has proposed the Adaptive Moment Estimation (Adam) algorithm, which has become one of the most widely used optimization methods in deep learning. This al-

Cite this article as: N. Sevani, W. M. Agustinus, and E. Kristianto, "Vision Transformer for Cataract Detection in Anterior Eye Images: Achieving Optimal Performance with Limited Data", CommIT Journal 20(2), 297–311, 2026.

TABLE I
LIST OF DATASETS.

Dataset	Number of Data	Before Cleaning		After Cleaning	
		Cataract	Normal (Dataset 1–3) or Immature (Dataset 4)	Cataract	Normal (Dataset 1–3) or Immature (Dataset 4)
1	672	355	317	155	210
2	410	214	196	60	70
3	612	304	306	132	250
4	578	403	184	55	32

gorithm adaptively adjusts the learning rate for each model parameter by combining the benefits of both Momentum and RMSprop techniques. Compared to traditional methods like Stochastic Gradient Descent (SGD), Adam enables faster convergence [30].

Adam has been widely applied in medical image processing research to improve training stability and address challenges such as the vanishing gradient problem. Additionally, its ability to manage parameters with varying gradient magnitudes is particularly important for large models such as ViT [22]. Therefore, the use of Adam in ViT-based cataract classification is a crucial factor in improving the precision and effectiveness of autonomous cataract detection systems.

Although ViT has demonstrated superiority in certain medical image classification tasks, several research limitations remain. Its application to anterior eye images for cataract classification remains limited, as many existing works have focused only on glaucoma, diabetic retinopathy, and fundus image analysis [31, 32]. Even when trained on relatively small datasets, ViT excels at capturing spatial relationships and can achieve high accuracy. However, this performance comes at the cost of higher computational requirements compared with CNNs [25].

Previous research on cataract detection using ViT and CNN-based models has primarily relied on fundus images from large datasets, typically employing hybrid approaches with multiple preprocessing steps to enhance image quality and model performance [10, 11, 24]. However, such techniques demand substantial resources and extended time to obtain the results. Accordingly, the research addresses the question of whether ViT alone, without the integration of additional preprocessing methods, can achieve effective cataract detection using anterior eye images from a limited dataset. The aim is to evaluate the effectiveness of ViT in diagnosing cataracts from anterior eye images, particularly in small-data scenarios. A key advantage of this approach is that it does not require specialized cameras, thereby making cataract diagnosis more accessible for medical professionals.

The research offers two primary contributions. First, it proposes a reliable approach for cataract detection using anterior eye images that is not dependent on patient pose, lighting conditions, or specific camera devices. Second, it presents empirical findings on the application of ViT for cataract detection in limited-dataset settings. The research results are expected to serve as a preliminary reference for future research on ViT applications in data-constrained medical imaging scenarios.

II. RESEARCH METHOD

The research is conducted in several phases, beginning with data collection and preprocessing, which involve dividing the dataset into training and testing sets. The next step is the development of the Transformer-based model, which is optimized through hyperparameter tuning to achieve the best classification performance. Finally, the model is evaluated using the testing data, with results presented in the form of a confusion matrix and training time.

A. Data Collection and Pre-Processing

The anterior eye images used in the research are obtained from several publicly available datasets, including Roboflow [33] and Kaggle [34–36]. These publicly accessible datasets are used to evaluate the model's ability to handle a wide range of eye image conditions. This variability is evident in both image quality and data volume. Some of these datasets, as shown in Table I, have also been used in prior studies on eye illnesses.

This selection of anterior eye images is made because, these images can be obtained more easily with a standard camera, eliminating the need for specialized equipment like the Zeiss FF 450plus [37] for fundus photography (see Fig. 1). It makes the approach more practical and user-friendly, increasing the likelihood of data collection in settings with limited medical infrastructure. The ViT-based classification algorithm is trained and evaluated using this dataset, which encompasses a wide range of cataract conditions.

Each dataset differs in image conditions and data volume. Representative examples from each class are

Cite this article as: N. Sevani, W. M. Agustinus, and E. Kristianto, "Vision Transformer for Cataract Detection in Anterior Eye Images: Achieving Optimal Performance with Limited Data", CommIT Journal 20(2), 297–311, 2026.

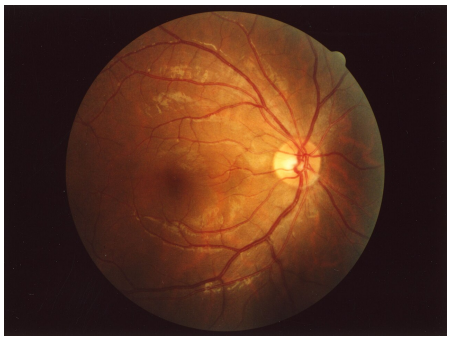

Fig. 1. Example of fundus image.

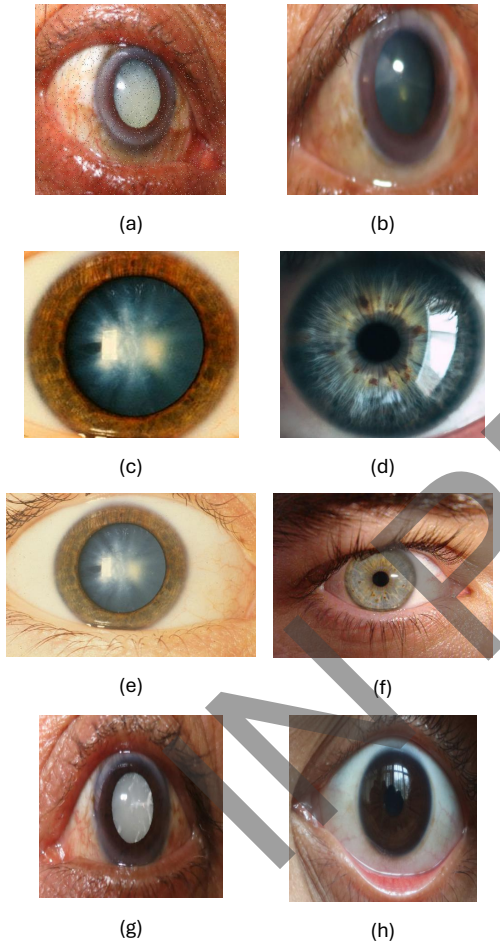

Fig. 2. Image example for each dataset.

provided, along with an overview of the dataset contents. Three classes are identified in the fourth dataset: normal, immature, and mature. As shown in Fig. 2, which presents examples of anterior eye images for each class in each dataset, the immature and mature classes are merged into a single category (cataract) to maintain consistency in the number of testing classes.

Examples of images from the four datasets utilized in this research are displayed in Fig. 2. Dataset 1 [35] contains image pairings (a) and (b), where image (a) depicts an adult cataract in the eye. Dataset 2 [34] contains pairs (c) and (d), where image (c) depicts a cataracted eye. Images (e) and (f) from Dataset 3 [36] are shown, with image (e) depicting a cataract. Lastly, Dataset 4 [33] provides pairs (g) and (h), where image (g) depicts a cataract. The diversity of eye images is reflected in the range of illumination, conditions, and viewing angles across the datasets.

Clinical data are not used in the research, as the primary aim is to evaluate model performance under small-dataset conditions. In addition, the use of clinical data entails significant ethical and regulatory constraints related to patient confidentiality, which must be carefully addressed before dissemination. Although it is possible that using clinical data will yield more diverse outcomes, allowing for broader analysis.

The original datasets contain duplicate images, making data cleaning necessary. The purpose of this cleaning process is to ensure that the training and testing sets do not include duplicates. Duplicate images can cause overfitting, whereby the model learns from repeated patterns without achieving proper generalization. Because ResNet-50 has sufficient depth to capture intricate information while minimizing overfitting through residual connections, it is used in the research for data cleaning. Previous research has shown that ResNet-50 can identify common patterns in images using pretrained weights from ImageNet [38], enabling more precise detection and filtering of low-quality inputs (noise or duplicates). Furthermore, its effectiveness in feature extraction makes it suitable for ensuring cleaner and more representative data prior to use in subsequent machine learning models [8].

Since the original data consist of various image conditions, data normalization and scaling are performed after data cleaning. No data augmentation is applied in the research, as the primary objective is to evaluate the capability of ViT in handling small-scale datasets. The images are resized to 224×224 pixels and normalized using mean = [0.5, 0.5, 0.5] and standard deviation = [0.5, 0.5, 0.5]. These values are chosen because they allow the input images to be represented as centroid features, which improve representation quality and contribute to greater stability and faster convergence during training [39]. This preprocessing step aims to enhance model performance and data homogeneity across a range of imaging angles and lighting conditions. The goal is to ensure that the developed model can identify cataract patterns more reliably and accurately under varying circumstances [40].

Additionally, the dataset is split into two parts, with

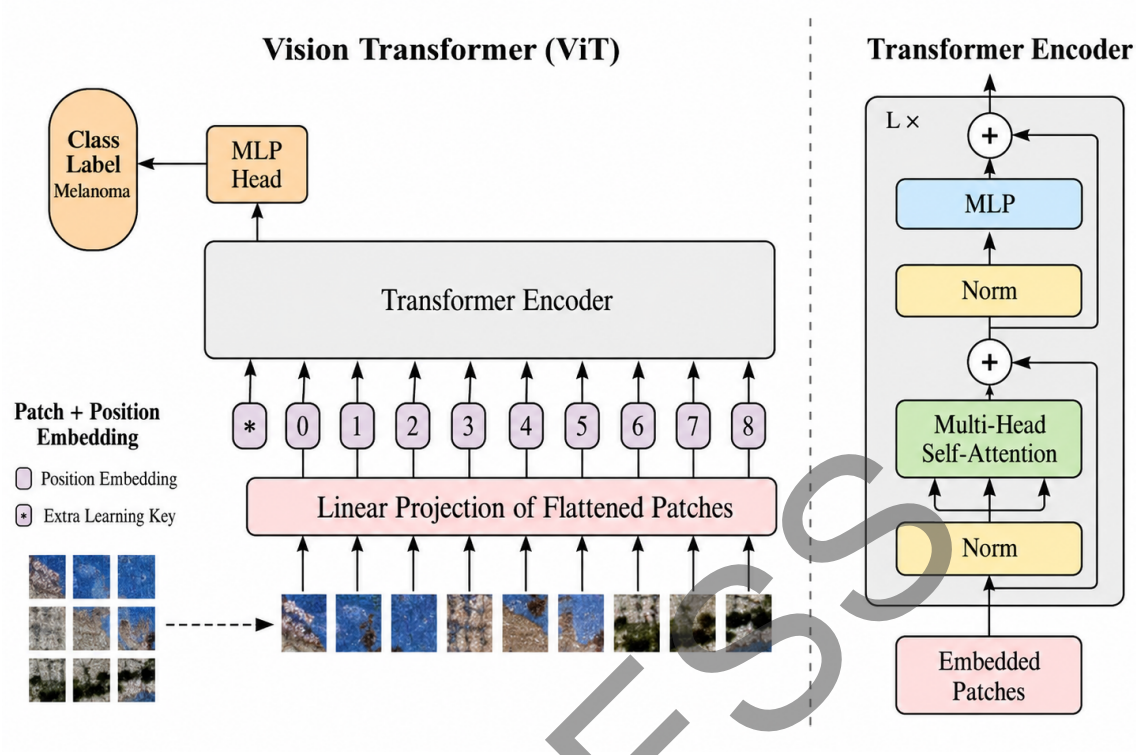

Fig. 3. Vision Transformer (ViT) illustration by Dosovitsky et al. [22]. Note: Multi-Layer Perceptron (MLP).

20% allocated for testing and 80% for training. This method is chosen to ensure that the model has sufficient data to learn patterns from cataract images while providing an unbiased assessment of performance on previously unseen data [41]. It is expected that this balanced division more accurately reflect the model's capacity for generalization in correctly diagnosing cataracts under real-world conditions.

B. Model Set-up

The main model used in this research, ViT, is based on one of the Google-released versions, ViT-Base-Patch16-224. This model is trained using both large and small datasets. Experimental results show the model performs well on large datasets but is less than optimum on small datasets. However, the ViT architecture has demonstrated efficacy in pattern recognition tasks involving medical images [22].

As shown in Fig. 3, ViT processes an image by dividing it into fixed-size patches and applying the Transformer encoder's self-attention mechanisms. To enable the model to capture spatial relationships broadly, the self-attention mechanism computes the similarity between pairs of patch tokens and assigns weights based on their similarity scores [22, 23]. Each patch (P) is incorporated into the Transformer encoder as an input token. Then, Transformer decoder up-samples a set of contextual tokens from the encoder for each pixel class score [42]. Equation (1) is used to project the patches linearly as follows:

$$z_0 = [x_{class}; x_p^1 E, x_p^2 E; ...; x_p^N E] + E_{pos}. \quad (1)$$

These patches are then flattened and linearly projected into a fixed-dimensional embedding space using a learnable matrix E. The x_p^i represents the flattened vector of the i-th patch, E is the embedding projection matrix, and E_{pos} is the positional embedding added to preserve spatial information lost during patch flattening. A special learnable vector is represented as x_{class}.

Let N denote the number of patches, formulated as:

$$N = \frac{H \times W}{P^2}, \quad (2)$$

where H and W in Eq. (2) are the height and width of the input image, respectively. This equation ensures that the transformer receives a fixed-length sequence of embeddings regardless of the original image size, provided the patch dimensions are fixed.

In the Transformer encoder, a Multi-Head Self-Attention (MSA) mechanism analyzes the relationships

among the patches, formulated as:

$$A = \text{softmax}(\frac{QK^T}{\sqrt{D_h}}), \tag{3}$$

where Q and K in Eq. (3) are the query and key matrices derived from the input, and D_h is the dimension of each attention head. This scaled dot-product attention computes similarity scores between tokens, allowing the model to attend selectively to different parts of the image based on learned relevance. The softmax function then normalizes these scores into a probability distribution, enabling effective and stable training.

During the data processing step, the images are fed into the ViT model, whose final layer is adjusted to accommodate two classes: normal and cataract. This figure is based on the number of classes in the dataset. Since the research uses a dataset with two classes, the final adjustment layer is confined to two classes.

The training process is carried out using the Cross-Entropy loss function, which is suitable for classification tasks with multiple classes. The Cross-Entropy loss function, which is effective for multi-class classification tasks, is used for training. Compared with SGD, the Adam optimizer, which integrates Momentum and RMSprop, optimizes prediction errors from Cross-Entropy, enabling more stable and faster learning [30].

The first moment in the Adam optimizer is the average (mean) gradient estimate, formulated in Eq. (4) as follows:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \tag{4}$$

where β_1 is the decay rate for the first moment (typically close to 1, such as 0.9). Then, m_{t-1} is the previous first moment estimate. Meanwhile, g_t is the gradient of the loss with respect to the model parameters at time step t.

The second moment, representing the variance of the gradient estimates (i.e., the average of the squared gradients), is expressed in Eq. (5) as follows:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, \tag{5}$$

where β_2 is the decay rate for the second moment (typically around 0.999), and g_t^2 is the element-wise square of the gradient. To prevent the moment estimates at the start of training from being skewed toward zero due to insufficient data, bias correction is applied to both the first and second moments. The corrected estimates are defined using the following bias-corrected forms:

$$\widehat{m}_t = \frac{m_t}{1 - \beta_1^t}, \tag{6}$$

$$\widehat{v}_t = \frac{v_t}{1 - \beta_2^t}, \tag{7}$$

where $\widehat{m}_t$ at Eq. (6) and $\widehat{v}_t$ at Eq. (7) are the bias-corrected first and second moment estimates, respectively. Then, t is the time step (iteration index).

Finally, the model parameters were updated to optimize the ViT prediction results, as defined in Eq. (8) as follows:

$$\theta_t = \theta_{t-1} - \frac{\alpha \widehat{m}_t}{\sqrt{\widehat{v}_t} + \in}. \tag{8}$$

Here, α is the learning rate (step size). Then, $\in$ is a small constant (e.g., 10^{-8}) added to the denominator to prevent division by zero.

To update the step that determines how much the model parameters change based on previous gradient information, the learning rate is multiplied by the result of the first-moment bias correction in the parameter update. The learning rate used in the research is 2×10^{-5}, which has been shown to improve training stability, prevent catastrophic forgetting (where the model loses previously learned information when acquiring new information), and maintain convergence stability [30, 43, 44]. To improve efficiency and speed up computation, model training is conducted with a batch size of 16 using a GPU (CUDA). Because the dataset is relatively small, training for an extended period can cause the model to learn unfavorable patterns [45]. Therefore, the number of epochs is set to 10.

These equations form the mathematical foundation of ViT, illustrating how the selected parameters and hyperparameter configurations influence its performance. Specifically, they describe the process by which the input image is partitioned into fixed-size patches, an essential tokenization step before being transformed and modeled through the self-attention mechanism. The self-attention mechanism is a core component of ViT's representation learning pipeline.

C. Model Evaluation

Several criteria are used to assess cataract classification performance. The primary metric for evaluating performance is accuracy, calculated as the proportion of correct predictions to all test data. In addition, each class's prediction error pattern is examined using a confusion matrix to evaluate classification performance [46]. The distribution of predictions for each class in the confusion matrix is defined according to the following terms [47]: True Positive (TP) which the positive class is correctly predicted; False Positive (FP) which the negative class is incorrectly predicted as positive; False Negative (FN) which the positive class is incorrectly predicted as negative; and True Negative (TN) which the negative class is correctly predicted.

In addition, accuracy, precision, recall, and F1-score are used for evaluation [31]. The final results for each

Cite this article as: N. Sevani, W. M. Agustinus, and E. Kristianto, "Vision Transformer for Cataract Detection in Anterior Eye Images: Achieving Optimal Performance with Limited Data", CommIT Journal 20(2), 297–311, 2026.

TABLE II
EXPERIMENT RESULTS BEFORE CLEANING WITH ORIGINAL DATASET.

Scores	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Precision	0.975	1.00	0.975	1.00
Recall	0.980	1.00	0.975	1.00
F1-score	0.980	1.00	0.975	1.00
Data testing	135	82	122	118
Accuracy	97.78%	100%	97.54%	100%
Time	83 s	46 s	210 s	62 s

score are then compared between the cleaned dataset and the original source dataset. Accuracy, defined in Eq. (9), represents the percentage of the number of correct predictions compared to total images

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%. \quad (9)$$

Equation (10) defines precision, which quantifies how accurate a class's predictions are. On the other hand, recall evaluates how well the model captures all samples that are actually part of a class. The formula for recall is found in Eq. (11). When there is an imbalance in the amount of data between classes, the F1-score is used to provide a balance between precision and recall. Equation (12) presents the F1-score formula [5]. Here are the equations used:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (10)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (11)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (12)$$

III. RESULTS AND DISCUSSION

The experiment for cataract detection is conducted using four public datasets of anterior eye images. The model employed is ViT, with performance evaluated through confusion matrix metrics and training duration. The ViT model's ability to effectively capture spatial relationships in anterior eye images is expected to yield improved performance, particularly with smaller datasets.

Given the initial condition of the datasets, which include a significant number of duplicate images, the first phase of the research involves a thorough data-cleaning process. This step aims to reduce dataset size and optimize model training time. Subsequent experiments evaluate the performance of the ViT model in classifying cataracts from eye images. The evaluation is carried out under two conditions: before data cleaning and after the cleaning process, with ResNet-50 used for comparative analysis.

A. Classification Results Before Cleaning.

The precision, recall, and F1-score results for each dataset are reported along with the average scores for the two classes (cataract and normal). Table II presents the results of 10 epochs of classification testing for each dataset before cleaning. Additional experiments are also conducted with 20, 30, 40, and 50 epochs to assess their impact on accuracy and loss, as shown in Fig. 4. The results indicate that 10 epochs yield the most optimal performance. Beyond this point, accuracy and loss become unstable, likely due to the model capturing irrelevant features and noise, as no preprocessing techniques are applied to the standalone ViT. In addition, larger epoch values increase computational cost and training time, whereas 10 epochs achieve comparable accuracy with significantly shorter training duration.

Figure 5 illustrates the confusion matrices for Datasets 1–4 prior to the cleaning process. Figure 5a illustrates the confusion matrix for Dataset 1. It correctly identifies all 61 normal cases, while 3 cataract cases are misclassified as normal, indicating a cataract sensitivity of 95.95% and a specificity of 100%. These results demonstrate the model's strong classification performance, although further improvement is needed to reduce false-negative predictions for cataract cases. Figure 5b presents the confusion matrix for the classification of normal and cataracts for Dataset 2. The proposed model correctly classifies all 45 immature and 37 mature cataract images without any misclassification, resulting in 0 false-positive and false-negative predictions. The perfect diagonal distribution indicates that the learned feature representations effectively capture the discriminative characteristics of cataract maturity, demonstrating the model's excellent capability and reliability in distinguishing between normal and cataract eyes within the evaluated dataset. Figure 5c presents the confusion matrix for cataract and normal eye classification for Dataset 3. The proposed model correctly classifies 62 cataract and 57 normal images, with only 2 false-negative and 1 false-positive predictions. The limited number of misclassifications indicates that the model effectively learned discriminative features for distinguishing cataract from normal eyes, resulting in high classification accuracy and balanced performance across both classes. Figure 5d presents the confusion matrix for Dataset 4. The proposed model correctly classifies all 75 cataract and 43 normal images, with no false-positive or false-negative predictions. The complete concentration of predictions along the main diagonal demonstrates that the extracted feature representations effectively distinguish cataract from normal eyes, yielding perfect classification performance on the

Cite this article as: N. Sevani, W. M. Agustinus, and E. Kristianto, “Vision Transformer for Cataract Detection in Anterior Eye Images: Achieving Optimal Performance with Limited Data”, CommIT Journal 20(2), 297–311, 2026.

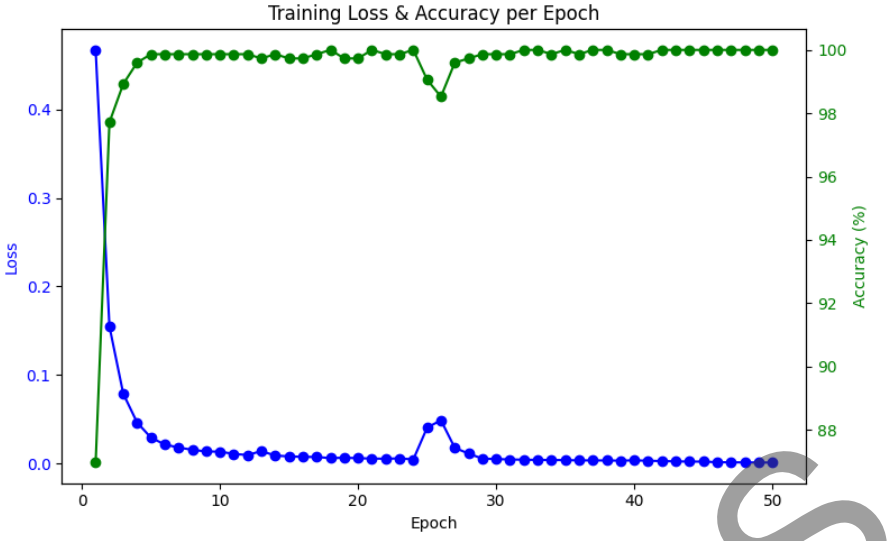

Fig. 4. Training loss & accuracy per epoch graph.

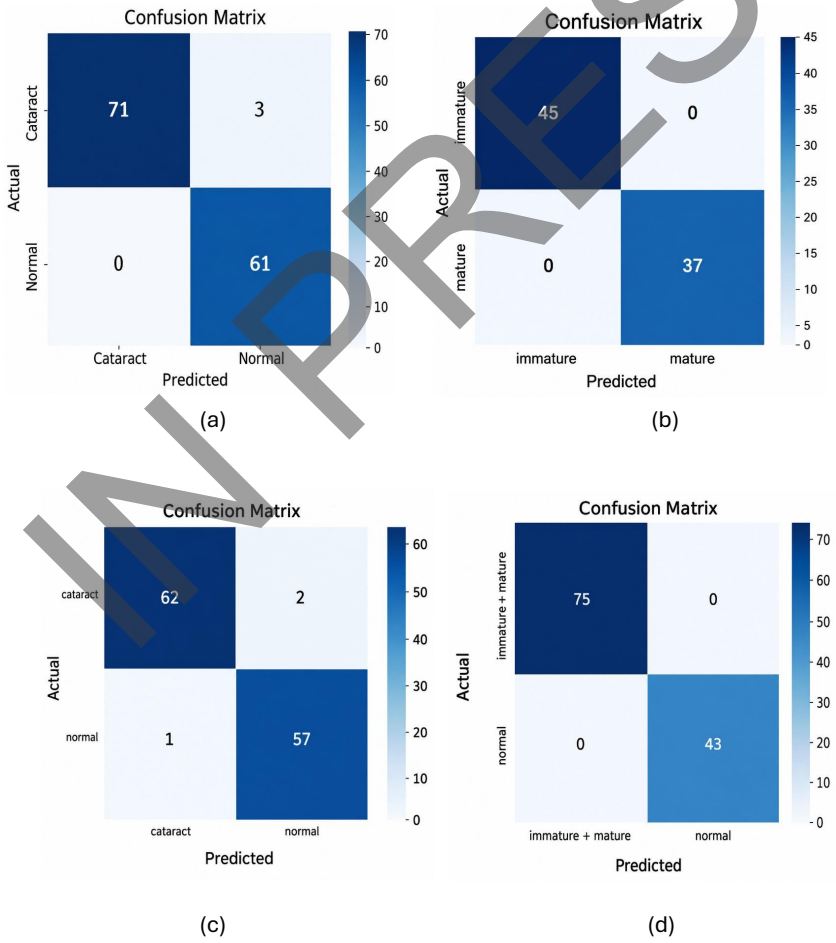

Fig. 5. Confusion matrix of Dataset 1–4 before cleaning.

evaluated dataset. All of these results highlight the robustness and reliability of the proposed model for automated cataract screening when cataract severity is considered as a single diagnostic category.

As demonstrated in Table II, accuracy results ranged from 97% to 100% within a relatively short timeframe. The experimental results prior to data cleaning indicate that ViT achieves the highest accuracy of 100% with training times of 46 seconds for Dataset 2 and 62 seconds for Dataset 4. These results are obtained using 82 and 118 test data points, respectively. The accuracy values for Dataset 1 and Dataset 3 are 97.78% and 97.54%, with training times of 83 seconds and 210 seconds. These findings suggest that Datasets 2 and 4 produce the highest accuracy with minimal training time, despite the small amount of training data.

B. Classification Results After Cleaning.

The cleaning process is performed using the ResNet-50 model, which identifies duplicate images within each dataset. The data-cleaning procedure using ResNet-50 is illustrated in Fig. A1 in Appendix. Initially, the input images are resized and converted into tensor format after feature vectors are extracted for each image. Pairwise comparisons are then carried out using a similarity threshold of 0.95, serving as the criterion for duplicate detection. Any image with a similarity score greater than 0.95 is classified as a duplicate and eliminated to prevent bias in subsequent cataract detection tasks. This procedure is applied iteratively until all image pairs have been evaluated.

Previous research has suggested that a high threshold should be used to ensure that only images that are genuinely similar and likely to be duplicates are flagged [48]. After trials with threshold values ranging from 0.9 to 0.99, a similarity threshold of 0.95 is found to be the most effective, yielding optimal results [48]. In Table III, the removal of duplicate data significantly reduces the number of test samples. Despite the smaller datasets and shorter testing period, ViT model demonstrates strong performance in classifying cataract and normal classes, achieving accuracy levels between 94% and 100%. The result indicates that ViT is an efficient model capable of maintaining high accuracy with limited training data. After data cleaning, Datasets 2 and 4 yield the highest results in accuracy, precision, recall, and F1-score. Although accuracy decreases slightly across all four datasets, the reduction is not significant, as accuracy remains above 94%. In addition, training time decreases as the amount of data the model processes is reduced. The Dataset 4, containing only 18 test samples, facilitates cataract classification. It results in the highest accuracy of 100% and the fastest processing time of just 10 seconds.

TABLE III
EXPERIMENTAL RESULTS AFTER THE DATASET CLEANING.

Scores	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Precision	0.960	0.965	0.965	1
Recall	0.955	0.960	0.920	1
F1-score	0.955	0.960	0.935	1
Data testing	73	26	77	18
Accuracy	95.89%	96.15%	94.81%	100%
Time	43 s	10 s	127 s	10 s

The results indicate that Dataset 2 consistently achieves higher accuracy than both Datasets 1 and 3, before and after data cleaning. Similarly, the processing time for Dataset 2 is notably fast, comparable to Dataset 4, likely due to its limited size of only 65 data points. These results show that ViT can handle small datasets.

The lowest accuracy is observed for Datasets 1 and 3, with the same trend evident before and after data cleaning. Dataset 3 is affected by excessive noise, such as Salt-and-Pepper interference, while Dataset 1 contains wide variation in eye color. Variations in eye color introduce differences in global image patterns that may affect the robustness of cataract detection. Specifically, while cataracts are generally characterized by a gray appearance, individuals with naturally gray or grayish irises present a case of inter-class similarity, where the pixel-level distribution of healthy eyes strongly overlaps with cataract-affected regions. This color-based feature ambiguity increases the likelihood of misclassification in automated cataract detection systems. Combined with the limited number of datasets available for training, it adversely affects ViT's ability to learn spatial relationships in the images, ultimately impacting model accuracy.

In Fig. 6a, the confusion matrix for Dataset 1 illustrates the total number of classes predicted correctly by ViT compared with the original labels. The results show that ViT predicts 28 out of 28 data points in the cataract class correctly and 44 out of 45 data points in the normal class accurately. ViT achieves high accuracy even with limited data and increased variation in the test set after cleaning. Figure 6b illustrates the confusion matrix for the cleaned Dataset 2. The results show that ViT correctly classifies all 16 normal and 10 cataract images. It achieves perfect classification without any misclassified samples. Figure 6c shows the confusion matrix for the cleaned Dataset 3. The results show that ViT correctly classifies 21 out of 25 cataract images and 52 out of 52 normal images. It misclassifies only 4 cataract samples as normal. Figure 6d illustrates the confusion matrix for the cleaned Dataset 4. The results show that ViT correctly classifies all 12 cataract and 6 normal images, yielding perfect classification

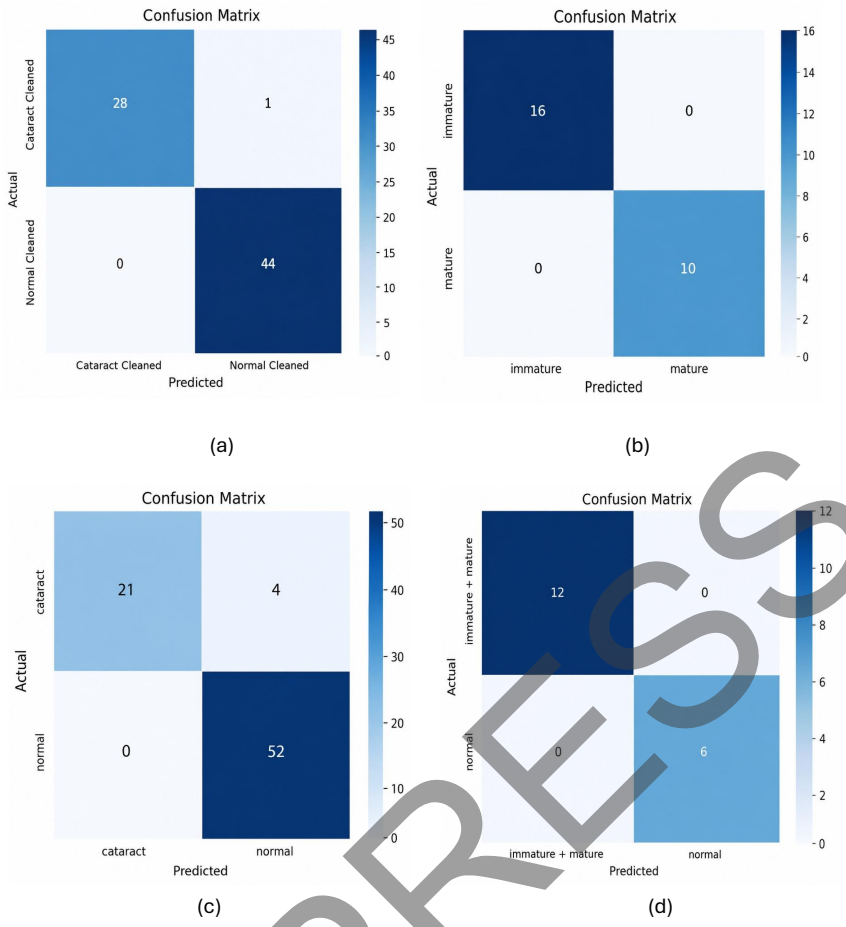

Fig. 6. Confusion matrix of Dataset 1–4 after cleaning.

with no false-positive or false-negative predictions. These results are similarly consistent across all other datasets, despite having variable ocular image quality. These data demonstrate that ViT is sufficiently trustworthy for use on a wide range of eye types.

C. Results Comparison with Other Models

The research compares the performance of the ViT in cataract classification with that of commonly used CNN models, such as ResNet-50 and EfficientNet-B7. The goal of this comparison is to assess the advantages of the ViT model in recognizing cataract patterns, particularly in terms of accuracy, sensitivity, and specificity. It also evaluates how effectively ViT handles variations in cataract shapes in images. This analysis aims to highlight the strengths and limitations of each model. Performance is evaluated in terms of classification accuracy and computational time, two widely recognized metrics that capture both predictive effectiveness and computational efficiency in deep learning applications.

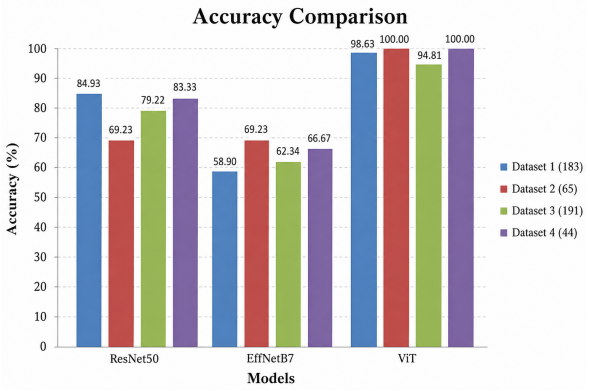

Fig. 7. Results of accuracy comparison across several models.

Figure 7 presents the accuracy results of three model architectures tested on four equally preprocessed datasets. The results indicate that ViT consistently achieves substantially higher accuracy than the two CNN models. Specifically, across Datasets

Cite this article as: N. Sevani, W. M. Agustinus, and E. Kristianto, "Vision Transformer for Cataract Detection in Anterior Eye Images: Achieving Optimal Performance with Limited Data", CommIT Journal 20(2), 297–311, 2026.

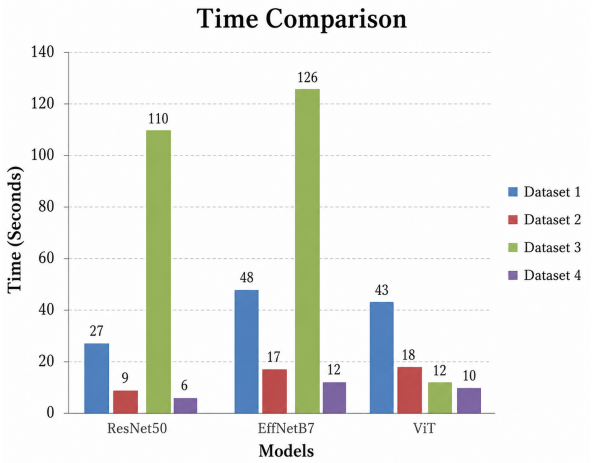

Fig. 8. Results of time comparison across several models.

1 to 4, ViT achieves accuracies of 95.89%, 96.15%, 94.81%, and 100%, respectively. In contrast, ResNet-50 records scores of 84.93%, 69.23%, 79.22%, and 83.33%, while EfficientNet-B7 achieved accuracies of 58.9%, 69.23%, 62.34%, and 66.67%. These results demonstrate that ViT offers stronger classification performance.

Furthermore, ViT's accuracy remains stable even when applied to relatively small datasets, as seen in Datasets 2 and 4. The results confirm that Transformer-based models can learn more effectively from limited yet clean data. The comparative analysis shows that ViT consistently outperforms ResNet-50 and EfficientNet-B7 across datasets of different scales, highlighting its scalability and generalization capability.

Despite ViT's strong accuracy, Fig. 8 shows that its training and testing times are competitive with CNN models. For Dataset 1, ViT completes the inference in 43 s, outperforming EfficientNet-B7 (48 s) but remaining slower than ResNet50 (27 s). On Dataset 2, ViT requires 18 s, slightly longer than ResNet50 (9 s) and EfficientNet-B7 (17 s). For Dataset 3, ViT achieves the shortest execution time of only 12 s, substantially outperforming both ResNet50 (110 s) and EfficientNet-B7 (126 s). Similarly, on Dataset 4, ViT completes the inference in 10 s, compared with 6 s for ResNet50 and 12 s for EfficientNet-B7. Overall, ViT demonstrates the fastest execution only on Dataset 3, while ResNet50 achieves the shortest execution times on Datasets 1, 2, and 4. Overall, these findings confirm that ViT not only delivers high accuracy but also maintains efficiency in terms of computation time, particularly when handling small- to medium-scale cleaned datasets.

ViT's ability to sustain high accuracy under limited data conditions stems from its self-attention mechanism [25], which effectively captures long-range spatial dependencies and global context within images. This capability fundamentally distinguishes it from conventional deep learning architectures used for comparison. Making ViT recognize and capture global patterns in images to distinguish between classes [3, 22].

TABLE IV
DATASET 3 RESULT COMPARISON WITH VGG16 (%).

Model	Accuracy	Precision	Recall	F1-Score
Full Model (VGG16, Siamese, Linear Regression (LR), Explainable Artificial Intelligence (XAI))	100.00	100.00	100.00	100.00
Proposed Model (ViT)	97.54	97.50	97.50	97.50

D. Results Comparison with Previous Research.

To evaluate the effectiveness of the ViT model in cataract classification based on anterior eye images, a comparison is made with the results of previous studies using the conventional CNN architecture VGG-16 on the same dataset, specifically Dataset 3 [36] in the research. Among the datasets, Dataset 3 is selected for further evaluation because it preserves the largest number of test samples after the data cleaning process. A larger testing set is expected to provide a more statistically reliable performance assessment. Moreover, its sample size closely matches those adopted in previous studies, allowing for a more consistent comparison with the existing literature. This comparison aims to assess the performance of the Transformer-based approach in terms of accuracy, precision, recall, and F1-score, without model merging as in previous studies [49]. As shown in Table IV, VGG-16 achieves 2.46% higher accuracy than ViT, although the dataset contains duplicate images. Furthermore, the number of ViT parameters is up to 35% fewer than VGG-16, making it more efficient in terms of computing resources.

IV. CONCLUSION

The research evaluates the performance of the ViT for cataract detection with small datasets. ViT achieves accuracies between 94% and 100%, outperforming ResNet-50 and EfficientNet-B7 by 10% to 41%, although it sometimes requires longer training times. Performance is influenced by noisy data and variations in eye color, yet ViT consistently demonstrates strong generalization across four datasets, highlighting its ability to capture global patterns even with limited and

cleaned data. These findings suggest that Transformer-based architectures hold significant promise for medical imaging applications such as cataract detection from anterior eye images.

Despite these strengths, the research does not incorporate image segmentation, demographic diversity, or advanced evaluation metrics beyond standard classification scores, which may limit interpretability and generalizability. Future work should explore segmentation techniques (such as U-Net or Mask R-CNN) to enhance interpretability, hybrid approaches combining CNNs and ViTs to leverage both local and global features. Clinical data can also be used for designing experiments to ensure the robustness of the model in real-world settings.

ACKNOWLEDGEMENT

The research was supported by a grant from Lembaga Penelitian dan Pengabdian Masyarakat from Krida Wacana Christian University No 07/UKKW/LPPM-FTIK/LIT/VI/2024.

AUTHOR CONTRIBUTION

Conceived and designed the analysis, N. S.; Collected the data, W. M. A.; Contributed data or analysis tools, N. S.; Performed the analysis, N. S. and E. K.; Wrote the paper, N. S., W. M. A., and E. K.; and Conducted LaTeX typesetting & figure refinement, E. K.

DATA AVAILABILITY

The data that support the findings of the research are available in several repositories at https://universe.roboflow.com/cataract/cataract-v01/dataset/7, https://www.kaggle.com/datasets/akshayramakrishnan28/cataract-classification-dataset, https://www.kaggle.com/datasets/kershrita/cataract, and https://www.kaggle.com/datasets/nandanp6/cataract-image-dataset/data, reference number [33–36]. These data were derived from the following resources available in the public domain.

REFERENCES

[1] L. Khan, N. Shaheen, Q. Hanif, S. Fahad, and M. Usman, "Genetics of congenital cataract, its diagnosis and therapeutics," *Egyptian Journal of Basic and Applied Sciences*, vol. 5, no. 4, pp. 252–257, 2018.

[2] J. Wang *et al.*, "A transformer-based knowledge distillation network for cortical cataract grading," *IEEE Transactions on Medical Imaging*, vol. 43, no. 3, pp. 1089–1101, 2023.

[3] R. Azad *et al.*, "Advances in medical image analysis with vision transformers: A comprehensive review," *Medical Image Analysis*, vol. 91, 2024.

[4] A. F. Adinegoro, G. N. Sutapa, A. A. N. Gunawan, N. K. N. Anggarani, P. Suardana, and I. G. A. Kasmawan, "Classification and segmentation of brain tumor using EfficientNet-B7 and u-net," *Asian Journal of Research in Computer Science*, vol. 15, no. 1, pp. 1–9, 2023.

[5] J. Wang *et al.*, "Prediction of postoperative visual acuity in patients with age-related cataracts using macular optical coherence tomography-based deep learning method," *Frontiers in Medicine*, vol. 10, pp. 1–10, 2023.

[6] I. Santoso, A. M. Manurung, and E. R. Subhiyakto, "Comparison of ResNet-50, EfficientNet-B1, and VGG-16 algorithms for cataract eye image classification," *Journal of Applied Informatics and Computing*, vol. 9, no. 2, pp. 284–294, 2025.

[7] A. K. Khanra, M. Kumar, A. Mandal, and S. Chatterjee, "EfficientNetB0 vs. VGG16 vs. ResNet-50: Classification of various skin diseases using deep learning," in *2025 International Conference on Next Generation of Green Information and Emerging Technologies (GIET)*. Gunupur, India: IEEE, Aug. 8–9, 2025, pp. 1–5.

[8] B. V. K. Golusu, N. Rajana, R. R. Arji, N. Killamsetty, and P. Naveena, "Revolutionizing image duplicate detection: Harnessing ResNet-50 with spatial transformers for unprecedented precision," *International Journal of Novel Research and Development*, vol. 9, no. 4, pp. 521–530, 2024.

[9] S. B. Thakare, "Knowledge distillation of the ResNet50 model for ocular diseases analysis," Master's thesis, National College of Ireland, 2022.

[10] S. S. Mahmood, S. Chaabouni, and A. Fakhfakh, "Improving automated detection of cataract disease through transfer learning using ResNet50," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17 541–17 547, 2024.

[11] H. Imaduddin, I. C. Utomo, and D. A. Anggoro, "Fine-tuning ResNet-50 for the classification of visual impairments from retinal fundus images," *International Journal of Electrical & Computer Engineering*, vol. 14, no. 4, pp. 4175–4182, 2024.

[12] R. Comle, "Fusion of vision transformer, Inception-V3 and ResNet50 for efficient eye disease detection," *International Journal of Engineering & Management Informatics*, vol. 2, no. 2, pp. 10–22, 2026.

[13] M. Tan and Q. Le, "EfficientNet: Rethinking

model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, California, USA: PMLR, June 9–15, 2019, pp. 6105–6114.

[14] A. A. Rakib, M. M. Billah, A. S. Ahamed, H. M. Imamul, and M. S. A. Masum, "EfficientNet-based model for automated classification of retinal diseases using fundus images," *European Journal of Computer Science and Information Technology*, vol. 12, no. 8, pp. 48–61, 2024.

[15] J. H. L. Goh *et al.*, "Artificial intelligence for cataract detection and management," *Asia-Pacific Journal of Ophthalmology*, vol. 9, no. 2, pp. 88–95, 2020.

[16] Y. C. Tham *et al.*, "Detecting visually significant cataract using retinal photograph-based deep learning," *Nature Aging*, vol. 2, no. 3, pp. 264–271, 2022.

[17] A. C. I. Ardison, M. J. R. Hutagalung, R. Chernando, and T. W. Cenggoro, "Observing pre-trained Convolutional Neural Network (CNN) layers as feature extractor for detecting bias in image classification data," *CommIT (Communication and Information Technology) Journal*, vol. 16, no. 2, pp. 149–158, 2022.

[18] D. Philippi, K. Rothaus, and M. Castelli, "A vision transformer architecture for the automated segmentation of retinal lesions in spectral domain optical coherence tomography images," *Scientific Reports*, vol. 13, no. 1, pp. 1–14, 2023.

[19] B. Hassan, T. Hassan, R. Ahmed, N. Werghi, and J. Dias, "SIPFormer: Segmentation of multiocular biometric traits with transformers," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–14, 2023.

[20] J. W. Kusno and A. Chowanda, "Modeling emotion recognition system from facial images using convolutional neural networks," *CommIT (Communication and Information Technology) Journal*, vol. 18, no. 2, pp. 251–259, 2024.

[21] B. Zhang *et al.*, "SegViT: Semantic segmentation with plain vision transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4971–4982, 2022.

[22] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *ICLR 2021: The Ninth International Conference on Learning Representations*, Virtual, May 3–7, 2021.

[23] Q. Liu, C. Kaul, J. Wang, C. Anagnostopoulos, R. Murray-Smith, and F. Deligianni, "Optimizing vision transformers for medical image segmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, June 4–10, 2023, pp. 1–5.

[24] D. Kumar, B. Bakariya, C. Verma, and Z. Illes, "Cataract disease identification using transformer and convolution neural network: A novel framework," in *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*. Tashkent, Uzbekistan: IEEE, Nov. 1–3, 2023, pp. 1230–1235.

[25] T. Zhang, W. Xu, B. Luo, and G. Wang, "Depth-wise convolutions in vision transformers for efficient training on small datasets," *Neurocomputing*, vol. 617, pp. 1–11, 2025.

[26] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, "Visformer: The vision-friendly transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Virtual: Computer Vision Foundation, Oct. 11–17, 2021, pp. 589–598.

[27] X. Mao *et al.*, "Towards robust vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, Louisiana: Computer Vision Foundation, 2022, pp. 12 042–12 051.

[28] S. Müller *et al.*, "Artificial intelligence in cataract surgery: A systematic review," *Translational Vision Science & Technology*, vol. 13, no. 4, 2024.

[29] I. Sonata, Y. Heryadi, A. Wibowo, and W. Budiharto, "End-to-end steering angle prediction for autonomous car using vision transformer," *CommIT (Communication and Information Technology) Journal*, vol. 17, no. 2, pp. 221–234, 2023.

[30] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations 2015*, San Diego, California, United States, 2015.

[31] P. B. Singh *et al.*, "Glaucoma classification using light vision transformer," *EAI Endorsed Trans Pervasive Health and Technology*, vol. 9, pp. 1–17, 2023.

[32] J. Vuppalapati, T. R. Kotha, K. P. Battula, and P. Kavali, "Eye disease classification using vision transformer: A deep learning approach," 2024. [Online]. Available: http://bit.ly/4xcWG7N

[33] Roboflow, "Cataract v01 computer vision dataset," 2022. [Online]. Available: https://universe.roboflow.com/cataract/cataract-v01/dataset/7

[34] A. Ramakrishnan, "Cataract classification dataset," 2024. [Online]. Available: https://www.kaggle.com/datasets/akshayramakrishnan28/

cataract-classification-dataset

[35] A. Abdulkhaliq, "Cataract," 2023. [Online]. Available: https://www.kaggle.com/datasets/kershrita/cataract

[36] N. Padia, A. Hirpara, D. Jani, and S. Patel, "Cataract dataset," 2023. [Online]. Available: https://www.kaggle.com/datasets/nandanp6/cataract-image-dataset/data

[37] S. Park *et al.*, "Features of long-standing Korean type 2 diabetes mellitus patients with diabetic retinopathy: A study based on standardized clinical data," *Diabetes & Metabolism Journal*, vol. 41, no. 5, 2017.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, June 27–30, 2016, pp. 770–778.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*. Lille, France: PMLR, July 7–9, 2015, pp. 448–456.

[40] A. K. Annamraju, "Dataset pre-processing and artificial augmentation, network architecture and training parameters used in appropriate training of convolutional neural networks for classification based computer vision applications: A survey," *International Journal of Advanced Engineering, Management and Science*, vol. 2, no. 9, 2016.

[41] A. Nazarkar, H. Kuchulakanti, C. S. Paidimarry, and S. Kulkarni, "Impact of various data splitting ratios on the performance of machine learning models in the classification of lung cancer," in *Proceedings of the Second International Conference on Emerging Trends in Engineering (ICETE 2023)*, vol. 223. Hyderabad, India: Atlantis Press, April 28–30, 2023, pp. 96–104.

[42] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Virtual: Computer Vision Foundation, Oct. 11–17, 2021, pp. 7262–7272.

[43] E. Simon, "Fine-tuning a vision transformer model for image classification into IAB taxonomy categories," 2023. [Online]. Available: https://www.blend360.com/thought-leadership/fine-tuning-a-vision-transformer

[44] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *China National Conference on Chinese Computational Linguistics*. Kunming, China: Springer, Oct. 18–20, 2019, pp. 194–206.

[45] R. M. Zur, Y. Jiang, L. L. Pesce, and K. Drukker, "Noise injection for training artificial neural networks: A comparison with weight decay and early stopping," *Medical Physics*, vol. 36, no. 10, pp. 4810–4818, 2009.

[46] J. Zhang, F. Li, X. Zhang, H. Wang, and X. Hei, "Automatic medical image segmentation with vision transformer," *Applied Sciences*, vol. 14, no. 7, pp. 1–18, 2024.

[47] I. Markoulidakis and G. Markoulidakis, "Probabilistic confusion matrix: A novel method for machine learning algorithm generalized performance analysis," *Technologies*, vol. 12, no. 7, pp. 1–23, 2024.

[48] O. Gorokhovatskyi and O. Peredrii, "Image pair comparison for near-duplicates detection," *International Journal of Computing*, vol. 22, no. 1, pp. 51–57, 2023.

[49] J. Olaniyan, D. Olaniyan, I. C. Obagbuwa, B. M. Esiefarienrhe, and M. Odighi, "Transformative transparent hybrid deep learning framework for accurate cataract detection," *Applied Sciences*, vol. 14, no. 21, pp. 1–20, 2024.

APPENDIX

The Appendix can be seen in the next page.

Cite this article as: N. Sevani, W. M. Agustinus, and E. Kristianto, "Vision Transformer for Cataract Detection in Anterior Eye Images: Achieving Optimal Performance with Limited Data", CommIT Journal 20(2), 297–311, 2026.

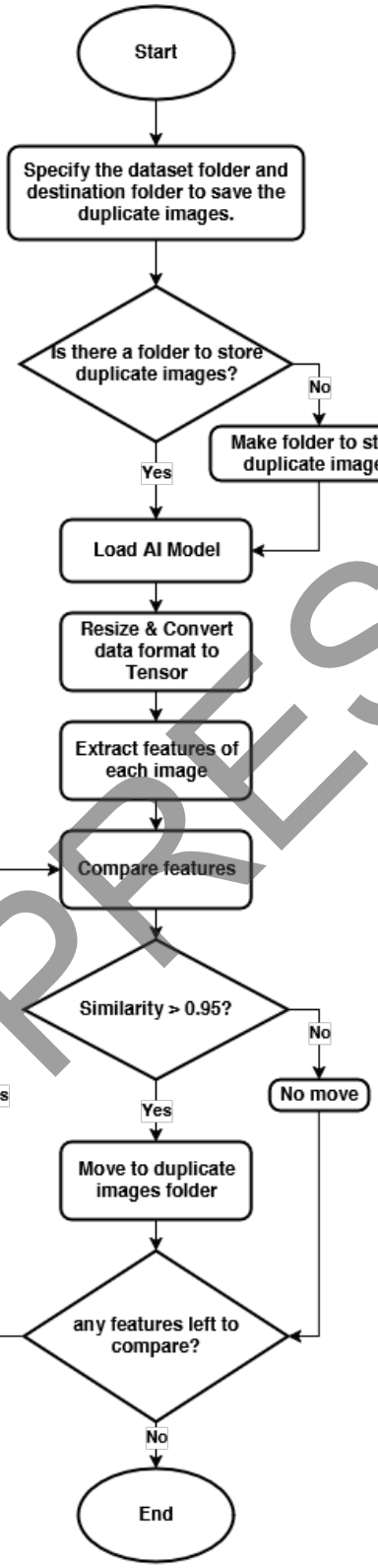

Fig. A1. Cleaning process flowchart.