# Attentional Mechanisms in Deep Learning for Ocular Disease Classification: A Comprehensive Analysis and Methodological Survey

## Chapter 1: Clinical Foundations of Ocular Disease Diagnosis from Fundus Imaging

### 1.1 The Role of Fundus Photography in Ophthalmology

Retinal fundus photography is a cornerstone of modern ophthalmic practice, providing a non-invasive method to capture high-resolution images of the posterior segment of the eye.1 This technique allows clinicians to visualize critical structures, including the retina, optic nerve head, macula, fovea, and the intricate network of retinal blood vessels.2 By projecting light through the pupil to illuminate the fundus, a specialized camera captures the reflected light, forming a detailed photographic record of the eye's interior. Its role as a primary screening, diagnostic, and documentary tool is indispensable for managing a wide spectrum of ocular and systemic diseases.1

The clinical utility of fundus photography is particularly evident in the longitudinal monitoring of chronic conditions. For diseases such as diabetic retinopathy, glaucoma, and age-related macular degeneration (AMD), serial fundus photographs allow ophthalmologists to track disease progression, evaluate the efficacy of treatments, and make informed clinical decisions.1 The visual record serves as an objective baseline against which future changes can be compared, facilitating the detection of subtle pathological developments that might otherwise be missed.

Despite its widespread use, the fundus imaging process is susceptible to a variety of technical challenges and artifacts that can degrade image quality and complicate diagnostic interpretation. These issues are common in real-world clinical collections and necessitate the development of robust automated analysis systems capable of handling such imperfections. Common quality issues include:

* **Uneven Illumination:** Fluctuations in the light source or ambient lighting can lead to images that are either underexposed (too dark) or overexposed (too bright), obscuring fine details.
* **Obstructions and Artifacts:** Particulate matter such as dust on the camera lens or anatomical obstructions like eyelashes can cast shadows or create spot artifacts on the image, potentially mimicking or masking pathological signs.4
* **Media Opacities:** Opacities in the eye's optical media, most notably cataracts, scatter and attenuate the light passing through the lens. This results in fundus images that appear generally blurry, have low contrast, and exhibit a hazy quality, which can severely impede the visibility of retinal structures.4
* **Focusing Errors:** Involuntary patient movement or improper focusing by the operator can lead to defocusing blur, rendering fine structures like small vessels or drusen indistinct.

### 1.2 Pathophysiology and Visual Biomarkers in Fundus Images

Automated classification of ocular diseases from fundus images depends on the system's ability to recognize specific visual biomarkers associated with the underlying pathophysiology of each condition. The diseases targeted in this analysis—Glaucoma, Cataract, AMD, Hypertensive Retinopathy, and Myopia—present distinct and varied visual characteristics.

* **Glaucoma (G):** A progressive optic neuropathy characterized by the gradual loss of retinal ganglion cells and their axons. The primary structural changes are visible at the optic nerve head. Key fundus biomarkers include an increase in the vertical cup-to-disc ratio (optic disc cupping), asymmetry in this ratio between a patient's two eyes, and a violation of the "ISNT" rule, which describes the normal thickness hierarchy of the neuroretinal rim (Inferior > Superior > Nasal > Temporal).1 Recent deep learning studies have also suggested that subtle alterations in the retinal vasculature may serve as early biomarkers that precede clinically apparent structural changes to the optic disc.
* **Cataract (C):** A clouding or opacification of the crystalline lens, which is located anterior to the retina.11 Since fundus photography images the posterior segment, cataracts are not directly visualized. Instead, their presence is inferred from the degradation they cause in the resulting image. Visual signs in a fundus photograph indicative of a cataract include a generalized blurring or haziness, reduced overall image contrast, loss of sharp clarity at the optic disc margin, and diminished visibility of the retinal vasculature.7 Different types of cataracts can produce distinct patterns of obscuration; for example, nuclear cataracts tend to cause a blur that radiates from the optic disc region, whereas posterior subcapsular cataracts often cause blurring centered on the macula. This positions cataract detection from fundus images as a unique challenge of artifact classification rather than direct pathology identification.
* **Age-related Macular Degeneration (AMD):** A degenerative disease affecting the macula, the central part of the retina responsible for sharp, detailed vision.15 The most common form, "dry" AMD, is characterized by the presence of yellowish sub-retinal deposits known as drusen. These can be small and well-defined ("hard") or larger with indistinct borders ("soft").16 Other biomarkers include pigmentary changes in the retinal pigment epithelium (RPE) and, in advanced stages, well-demarcated areas of RPE and photoreceptor loss known as geographic atrophy.16 The less common "wet" AMD is identified by signs of choroidal neovascularization, which can lead to subretinal fluid or hemorrhage visible on the fundus image.16
* **Hypertensive Retinopathy (H):** A condition involving damage to the retinal microvasculature caused by chronically elevated blood pressure.17 The visual biomarkers are direct indicators of systemic vascular pathology. Early signs include generalized or focal narrowing of the retinal arterioles and arteriovenous (AV) nicking, where a hardened arteriole compresses an underlying venule at a crossing point.18 More moderate to severe signs include flame-shaped or blot hemorrhages, microaneurysms, cotton-wool spots (infarctions of the nerve fiber layer), and, in cases of malignant hypertension, optic disc swelling (papilledema).19
* **Pathologic Myopia (M):** A severe, degenerative form of nearsightedness characterized by excessive axial elongation of the eyeball, leading to mechanical stretching and thinning of the posterior structures.20 This is distinct from simple high myopia. Key fundus biomarkers include a posterior staphyloma (a localized outpouching of the posterior wall of the eye), diffuse or patchy chorioretinal atrophy due to the thinning of the choroid and RPE, and lacquer cracks, which are fine, irregular breaks in the Bruch's membrane.20

The diverse nature of these biomarkers, ranging from morphological changes in a specific structure (optic disc cupping in glaucoma), to fine textural deposits (drusen in AMD), to subtle geometric alterations in vessels (AV nicking in hypertensive retinopathy), and even global image artifacts (blur from cataracts), presents a significant challenge for a unified automated diagnostic system. A successful model must be capable of learning features at multiple scales and localizations. This inherent complexity provides a strong rationale for exploring advanced deep learning architectures, particularly those incorporating attention mechanisms that can dynamically focus on different feature types and spatial regions relevant to each specific pathology.

## Chapter 2: Foundational Deep Learning Architectures for Medical Vision

### 2.1 The Evolution of Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have become the dominant paradigm in computer vision, including medical image analysis. Their architecture is inspired by the organization of the animal visual cortex and is designed to automatically and adaptively learn spatial hierarchies of features. The foundational building blocks of a CNN are the convolution operator, which applies learnable filters to input images to create feature maps; pooling layers, which downsample feature maps to reduce computational complexity and create a degree of translation invariance; and activation functions, which introduce non-linearity into the model.

A key reason for the success of CNNs in image-based tasks is their inherent **inductive biases**: locality and translation invariance.2 The convolutional operation processes data in small local receptive fields, assuming that spatially close pixels are more related than distant ones. The sharing of filter weights across different spatial locations provides translation invariance, meaning the network can detect a feature regardless of its position in the image. This architectural prior is highly effective for visual tasks. The history of CNNs is marked by architectural milestones that progressively increased their depth and representational power, from early models like LeNet and AlexNet, to the deep VGG networks, and culminating in the development of Residual Networks (ResNet), which introduced skip connections to enable the training of networks hundreds of layers deep.22

### 2.2 EfficientNet: A Paradigm of Model Scaling

Prior to the development of EfficientNet, scaling up CNNs to achieve better accuracy was typically done along one of three dimensions: increasing network **depth** (adding more layers), **width** (increasing the number of channels), or using higher **resolution** input images.23 However, this one-dimensional scaling resulted in diminishing returns, and manually balancing these factors was a tedious and often suboptimal process.23

The core innovation of the EfficientNet architecture is the principle of **compound scaling**.23 The authors demonstrated empirically that there is an optimal balance between the network's depth, width, and resolution. Scaling all three dimensions in a coordinated and principled manner leads to superior performance and greater computational efficiency. This is achieved by using a simple compound coefficient, $\phi$, to uniformly scale the network dimensions according to a fixed set of ratios. The scaling relationship can be formulated as:

$$\text{depth}: d = \alpha^{\phi}$$

$$\text{width}: w = \beta^{\phi}$$

$$\text{resolution}: r = \gamma^{\phi}$$

subject to $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$, where $\alpha, \beta, \gamma$ are constant coefficients determined by a small grid search on an initial baseline model.23 The intuition is that as input resolution increases, the network needs more layers to increase its receptive field and more channels to capture finer-grained patterns on the larger image.23

The baseline architecture, **EfficientNet-B0**, was designed using a multi-objective neural architecture search that optimized for both accuracy and floating-point operations per second (FLOPS). Its primary building block is the **Mobile Inverted Bottleneck (MBConv)**, which includes an expansion layer, a depthwise convolution, a Squeeze-and-Excitation (SE) optimization block, and a projection layer.24 By fixing the coefficients $\alpha, \beta, \gamma$ and scaling the baseline B0 model with different values of $\phi$, the authors created the family of models from **EfficientNet-B0 to B7**. Each successive model in this family achieves a new state-of-the-art accuracy on benchmarks like ImageNet while being significantly smaller and faster than previous CNNs with similar performance.22 The project under review utilizes EfficientNet-B0 and considers scaling to B3, leveraging its strong performance-to-cost ratio.2

### 2.3 The Rise of Transformers in Vision (ViT)

While EfficientNet represents a highly optimized evolution of CNN principles, the Vision Transformer (ViT) introduced a fundamentally different paradigm for image recognition.25 Inspired by the success of the Transformer architecture in natural language processing (NLP), ViT was the first work to demonstrate that a pure transformer model, with minimal modifications, could achieve state-of-the-art results on image classification tasks, challenging the long-held dominance of CNNs.25

The ViT architecture processes images as follows:

1. **Patching and Embedding:** The input image is first divided into a sequence of fixed-size, non-overlapping patches (e.g., 16x16 pixels). Each patch is flattened into a 1D vector and then linearly projected into a high-dimensional embedding space. These patch embeddings are analogous to the word token embeddings used in NLP transformers.25
2. **Positional Embeddings:** Because the transformer architecture is permutation-invariant, it has no inherent understanding of the spatial arrangement of the patches. To address this, learnable 1D positional embeddings are added to the patch embeddings to encode their original location in the image grid.29
3. **Transformer Encoder:** The resulting sequence of embeddings is fed into a standard Transformer encoder. This encoder is composed of a stack of identical layers, each containing a Multi-Head Self-Attention (MHSA) block and a position-wise feed-forward MLP block. Layer normalization and residual connections are used throughout.28
4. **Classification Head:** A special, learnable [class] token embedding is prepended to the sequence of patch embeddings. The state of this token at the output of the final transformer encoder layer serves as the aggregate image representation. This output token is then passed to a simple MLP head for the final classification.29

The fundamental departure of ViT from CNNs lies in its approach to inductive bias. CNNs are built on the strong priors of locality (pixels in a neighborhood are strongly related) and translation invariance. ViT, by contrast, has a much weaker inductive bias. Its self-attention mechanism allows every patch to interact directly with every other patch from the very first layer, enabling it to model global, long-range dependencies across the entire image. This flexibility, however, comes at a cost: ViTs are notoriously "data-hungry" and generally require pre-training on massive datasets (e.g., the private JFT-300M dataset with 300 million images) to learn the fundamental properties of images that are hard-coded into CNNs.25 When trained on smaller datasets like ImageNet, they often underperform their convolutional counterparts.

This distinction provides a clear rationale for the architectural choices made in the project under review. For a mid-sized, specialized medical dataset like ODIR-5K, an architecture like EfficientNet, which leverages the powerful and data-efficient inductive biases of convolutions, is a more pragmatic and theoretically sound choice than a pure ViT. The CNN's inherent ability to learn hierarchical local features is well-suited to the task without requiring the vast pre-training data that a ViT would need to achieve comparable performance.

## Chapter 3: A Theoretical Survey of Attention Mechanisms in Computer Vision

### 3.1 Fundamental Concepts of Visual Attention

Attention mechanisms in deep learning are inspired by the human visual system's ability to selectively focus on the most salient parts of a complex scene while filtering out irrelevant information.31 In a computational context, an attention mechanism can be viewed as a process that dynamically adjusts the weights of features based on their importance for a given task. This allows a neural network to allocate more of its computational resources to the most informative signals in the input data.

A general and powerful formulation of attention, popularized by the original Transformer paper, is based on the concepts of **queries ($Q$), keys ($K$), and values ($V$)** 2]. For a given query (representing the current focus), the mechanism computes a score or similarity with a set of keys. These scores are then normalized (typically via a softmax function) to produce attention weights. Finally, a weighted sum of the corresponding values is computed, yielding an output that emphasizes the values associated with the most relevant keys. The scaled dot-product attention is formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d\_k}}\right)V$$

While this formulation is central to self-attention in Transformers, the underlying principle of re-weighting features is adapted in various forms within convolutional networks.

### 3.2 A Taxonomy of Attention in Convolutional Networks

In the context of CNNs, attention mechanisms are typically designed as lightweight modules that can be inserted into existing architectures to enhance feature representations. They are often categorized by the domain over which they operate: channel, spatial, or a hybrid of both.

#### 3.2.1 Channel Attention ("What to attend to")

Channel attention mechanisms focus on modeling the interdependencies between the channels of a convolutional feature map, adaptively recalibrating channel-wise feature responses.

* **Squeeze-and-Excitation (SE) Networks:** The SE block was a pioneering work that introduced an effective and computationally efficient form of channel attention 2]. The process involves three steps:
  1. **Squeeze:** The feature map $F \in \mathbb{R}^{H \times W \times C}$ is spatially squeezed into a channel descriptor $z \in \mathbb{R}^{1 \times 1 \times C}$ using global average pooling. This step aggregates global spatial information for each channel.33
  2. **Excitation:** The channel descriptor $z$ is passed through a two-layer Multi-Layer Perceptron (MLP) with a bottleneck structure (a dimensionality-reducing hidden layer followed by a dimensionality-increasing layer). This small network learns a non-linear, cross-channel interaction to produce a set of channel-wise weights or activations, $s \in \mathbb{R}^{1 \times 1 \times C}$.33
  3. **Recalibration:** The final output of the SE block is obtained by rescaling the original feature map $F$ with the learned activations $s$. This adaptively emphasizes informative feature channels while suppressing less useful ones.33
* **Efficient Channel Attention (ECA-Net):** ECA-Net builds directly upon SE-Net, arguing that SE's dimensionality reduction is both unnecessary and potentially detrimental to learning effective channel attention.35 To improve efficiency and effectiveness, ECA-Net proposes two key modifications:
  1. **No Dimensionality Reduction:** It completely avoids the bottleneck MLP, preserving the direct correspondence between a channel and its weight.
  2. **Local Cross-Channel Interaction:** Instead of modeling dependencies across all channels (as the MLP does), ECA captures *local* cross-channel interactions. This is achieved efficiently by applying a 1D convolution with a kernel of size $k$ across the channel dimension. The kernel size $k$ is adaptively determined based on the channel dimension $C$, allowing the model to flexibly adjust the scope of local interaction.35

#### 3.2.2 Spatial and Hybrid Attention ("Where to attend to")

While channel attention focuses on *what* is important, spatial attention focuses on *where* the informative features are located. Hybrid mechanisms combine both.

* **Bottleneck Attention Module (BAM):** BAM computes channel and spatial attention maps along two *parallel* pathways, which are then summed and passed through a sigmoid function to create a unified 3D attention map.36 This map is then applied to the input feature map using a residual connection. The spatial attention branch in BAM notably uses dilated convolutions to increase the receptive field without adding significant computational cost 2]. BAM is designed to be placed at the bottlenecks of a network where feature maps are downsampled.37
* **Convolutional Block Attention Module (CBAM):** As used in the primary study, CBAM proposes a *sequential* arrangement of channel and spatial attention modules, finding it empirically superior to a parallel arrangement.38
  1. **Channel Attention Module:** CBAM refines the SE block's approach by using both **global average pooling** and **global max pooling** across the spatial dimensions. The two resulting channel descriptors are processed through a shared MLP, and their outputs are summed before being passed through a sigmoid function. The use of max pooling allows the module to capture distinctive object features in addition to the average response, providing richer channel statistics.38 Given an input feature map $F \in \mathbb{R}^{C \times H \times W}$, the channel attention map $M\_c \in \mathbb{R}^{C \times 1 \times 1}$ is computed as:  
     $$M\_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F)))$$  
     where $\sigma$ denotes the sigmoid function and the MLP is a two-layer network with a shared bottleneck 2; 38].
  2. **Spatial Attention Module:** This module takes the channel-refined feature map $F'$ as input. It first applies average pooling and max pooling along the *channel* axis, creating two 2D feature maps ($1 \times H \times W$) that summarize the channel-wise information at each spatial location. These two maps are concatenated and then processed by a single convolutional layer with a relatively large kernel (typically 7x7) to generate the final 2D spatial attention map $M\_s \in \mathbb{R}^{1 \times H \times W}$.38 The formulation is:  
     $$M\_s(F') = \sigma(f^{7 \times 7}([\text{AvgPool}(F'); \text{MaxPool}(F')]))$$  
     where $f^{7 \times 7}$ represents the convolution operation 2]. The final refined feature map is then computed by element-wise multiplication of $F'$ with $M\_s(F')$.

The progression from SE to ECA and CBAM illustrates a logical refinement in the design of attention mechanisms for CNNs. SE established the value of channel re-weighting. CBAM extended this by demonstrating the complementary importance of spatial attention and the effectiveness of a sequential fusion strategy. Concurrently, ECA revisited the original channel attention mechanism itself, proposing a more efficient way to capture cross-channel interactions. The choice of CBAM in the primary study represents a selection of a mature and well-validated hybrid attention model that addresses both "what" and "where" to focus.

### 3.3 Comparative Analysis: Convolutional vs. Self-Attention

The attention mechanism in CBAM and the self-attention in ViT, while sharing the same high-level goal of feature re-weighting, operate on fundamentally different principles.

* **Scope of Attention:**
  + **CBAM (Convolutional Attention):** Operates on a structured grid of features (the feature map). Its attention is computed based on aggregated information. Channel attention is derived from spatially-pooled features, and spatial attention is derived from channel-pooled features. The scope of interaction is implicitly local, constrained by the effective receptive field of the underlying CNN.2
  + **ViT (Self-Attention):** Operates on an unstructured set of tokens (the patch embeddings). Each token directly computes an attention score with every other token in the sequence, regardless of their spatial distance. This allows for the modeling of global, long-range dependencies from the very first layer of the network.25
* **Computational Complexity:**
  + **CBAM:** Extremely lightweight. Its components (pooling, small MLPs, a single convolution) add negligible computational overhead and parameters to the base CNN. The complexity is linear with respect to the number of pixels in the feature map.38
  + **ViT:** Computationally intensive. The complexity of the self-attention mechanism is quadratic with respect to the sequence length (i.e., the number of patches), $O(N^2)$, where $N$ is the number of patches. This makes it computationally prohibitive for very high-resolution images without using more complex, hierarchical transformer designs.28
* **Inductive Bias:**
  + **CBAM:** Designed as an add-on module for CNNs, it fully inherits and leverages the strong inductive biases of convolution, namely locality and translation invariance. It aims to enhance, not replace, these priors.
  + **ViT:** Possesses a much weaker inductive bias. It does not assume locality, treating all patches as equidistant in the attention computation. This makes it more flexible and expressive but requires it to learn the basic properties of images from data, necessitating large-scale pre-training.26

## Chapter 4: State-of-the-Art in Automated Ocular Disease Classification

### 4.1 The ODIR-5K Dataset: A Real-World Challenge

The Ocular Disease Intelligent Recognition (ODIR-5K) dataset is a publicly available, structured ophthalmic database designed to reflect the complexities of real-world clinical data.41 It contains data from 5,000 patients, including color fundus photographs of both the left and right eyes, patient age, and diagnostic keywords provided by clinicians.41 This structure presents several significant challenges for the development of automated diagnostic systems, making it an important benchmark for robust model evaluation.

* **Class Imbalance:** The dataset exhibits a severe class imbalance, a common issue in medical datasets. The number of "Normal" fundus images far exceeds the number of images for specific pathologies. For instance, classes like Hypertension are represented by a very small number of single-label samples, making it difficult for models to learn their distinguishing features without being biased towards the majority classes.2
* **Multi-Label Nature:** Ocular diseases often co-occur. A single patient may present with both cataracts and diabetic retinopathy, for example. The ODIR-5K dataset captures this reality, with many images associated with multiple diagnostic keywords.2 This requires models to be capable of multi-label classification, where each image can be assigned a set of non-exclusive labels.
* **Label Ambiguity:** The diagnostic information is provided as free-text keywords (e.g., "hypertensive retinopathy," "htn") rather than clean, one-hot encoded labels 2]. This necessitates a careful and consistent label parsing and mapping strategy to convert the raw text into a structured format suitable for training a supervised model.

### 4.2 Systematic Literature Review of Deep Learning on ODIR-5K (post-2020)

Research on the ODIR-5K dataset has explored various strategies to tackle its inherent challenges. The methodologies can be broadly categorized by how they handle class imbalance, their problem formulation, and their architectural choices.

* Methodological Theme 1: Addressing Class Imbalance  
  A primary focus of research has been to mitigate the effects of the dataset's long-tailed class distribution. One common strategy involves simplifying the problem by converting the multi-class task into a series of binary classification problems (e.g., Glaucoma vs. Normal). This allows researchers to manually create balanced subsets for training and testing each binary classifier, often leading to high reported accuracies for the specific binary task but losing the broader clinical context of co-occurring diseases.42 More sophisticated approaches applied within a multi-class or multi-label framework include data-level techniques like oversampling minority classes and undersampling majority classes, as demonstrated in the DKCNet paper 43, or algorithm-level techniques like using class weights during training to give higher importance to errors on minority classes, a method employed in the primary study.2
* Methodological Theme 2: Handling Multi-Label Classification  
  The multi-label nature of the data is another critical challenge. A straightforward approach, adopted by the primary study, is to simplify the problem into a single-label classification task. This is done by defining a priority rule, such as assigning 'Hypertension' as the primary label if it is present among multiple diagnoses, to ensure its representation in the evaluation sets.2 While pragmatic, this approach discards information about comorbidities. More advanced works tackle the problem directly as a multi-label classification task. Models like DKCNet 43 and the work by Li et al. (2021) are designed with architectures and loss functions suitable for predicting a vector of binary labels for each image.44
* Methodological Theme 3: Architectural Choices and Fusion  
  A diverse range of CNN backbones has been applied to the ODIR-5K dataset. While the primary study focuses on the highly efficient EfficientNet architecture, other works have successfully employed VGG-19 42, various ResNet models 45, and InceptionResNet.43 A notable trend in recent high-performing models is the use of model fusion and attention mechanisms. For example, a 2024 study proposed a fusion of ResNet50 and ResNet101 outputs using Dempster-Shafer (D-S) evidence theory to improve robustness and reduce decision bias.45 The incorporation of attention is also prevalent, with models like AUB-Net using dedicated attention mechanisms for left and right eye features 45, and others integrating spatial attention modules to enhance feature representation.43 This body of work validates the primary study's focus on attention as a relevant and promising direction for improving classification performance on this challenging dataset.

The following table summarizes key findings from recent literature, providing a comparative landscape for the work presented in this report.

| **Reference (Year)** | **Methodology/Key Contribution** | **Backbone** | **Problem Formulation** | **Reported Metrics (AUC/F1/Kappa)** | **Handling of Imbalance** |
| --- | --- | --- | --- | --- | --- |
| Islam et al. (2022) 42 | Converted multi-class to binary classification | VGG-19 | Binary | N/A (reports high binary accuracy) | Manual balancing for binary tasks |
| He et al. (cited in 45) | AUB-Net with left/right eye attention | Custom | Multi-label | AUC: 0.934, F1: 0.913, Kappa: 0.640 | Not specified |
| Sun et al. (2020) 45 | AEye Doctor system with saliency heatmap | ResNet | Multi-label | AUC: 0.90 | Not specified |
| Zhou et al. (2020) 45 | Inductive transfer learning (MTC-DSAA) | Custom | Multi-label | AUC: 0.938, F1: 0.929, Kappa: 0.697 | Domain adaptation |
| Gour & Khanna (2022) 44 | DKCNet with attention and SE blocks | InceptionResNet | Multi-label | AUC: 0.961, F1: 0.943, Kappa: 0.810 | Oversampling & undersampling |
| Ou et al. (2022) 43 | Two-input CNN with multi-scale attention | ResNet-50 | Multi-label | AUC: 0.903, F1: 0.886, Kappa: 0.826 | Not specified |
| Shang et al. (2024) 45 | Fusion of models with D-S evidence theory | ResNet50, ResNet101 | Multi-label | AUC: 0.987, F1: 0.914, Kappa: 0.878 | Model fusion to reduce bias |

## Chapter 5: Experimental Framework: Benchmarking EfficientNet and CBAM on ODIR-5K

### 5.1 Data Curation and Preprocessing Pipeline

The foundation of a reproducible and robust machine learning experiment lies in a well-defined data handling pipeline. The methodology employed in this study was designed to address the specific challenges of the ODIR-5K dataset, from raw label parsing to balanced data splitting and image augmentation.

* **Dataset:** The study utilizes the Ocular Disease Intelligent Recognition (ODIR-5K) dataset, a public collection of 5,000 patient cases with paired fundus images and diagnostic keywords 2].
* **Label Parsing and Mapping:** The initial step involved converting the unstructured, free-text diagnostic keywords into a set of discrete class labels. A keyword-matching algorithm was implemented to map various diagnostic terms (e.g., "hypertensive retinopathy", "htn") to one of the five target classes: Glaucoma (G), Cataract (C), Age-related Macular Degeneration (A), Hypertension (H), and Myopia (M).2
* **Problem Formulation - Hypertension Priority:** Recognizing the multi-label nature of the dataset, a pragmatic decision was made to simplify the task into a single-label, multi-class classification problem. This was done to facilitate clearer analysis and evaluation, particularly with confusion matrices. To address the severe underrepresentation of the Hypertension class, a priority-based labeling rule was established: if the diagnostic keywords for an image included any term mapping to 'H', that image was assigned the single label 'H', regardless of other co-occurring pathologies. If 'H' was not present, the label was assigned based on the first match in a fixed order (G, C, A, M). This strategy was explicitly chosen to increase the prevalence of the 'H' class in the validation and test sets, thereby enabling a more meaningful evaluation of the model's performance on this difficult minority class.2
* **Robust Stratified Splitting:** To ensure that the model's performance could be reliably evaluated across all classes, a robust data splitting procedure was implemented. Standard random splits could easily result in validation or test sets that lack samples from one or more of the minority classes. To prevent this, the StratifiedShuffleSplit function was used within a loop. This process was repeated until a split was generated where all five target classes were confirmed to be present in both the validation and test sets, guaranteeing a representative evaluation environment.2
* **Image Preprocessing and Augmentation:** All images were subjected to a standardized preprocessing pipeline. They were first resized to an input dimension of $224 \times 224$ pixels to match the input requirements of the chosen backbone. The pixel values were then normalized according to the standard procedure for models pretrained on ImageNet. During the training phase, on-the-fly data augmentation was applied to increase the diversity of the training data and reduce overfitting. The augmentations included basic geometric and photometric transformations: random horizontal flips, small rotations, minor zooming, and contrast adjustments.2

### 5.2 Model Design and Implementation

Two architectures were benchmarked under identical conditions: a strong CNN baseline and an attention-augmented variant.

* **Baseline Architecture (EfficientNet-B0):** The baseline model employed an ImageNet-pretrained **EfficientNet-B0** as its feature extraction backbone.2 This choice leverages the model's high efficiency and strong performance derived from its compound scaling and MBConv blocks. A custom classification head was appended to the backbone. This head consists of a Batch Normalization layer, a $1 \times 1$ convolutional layer to reduce feature dimensionality, a Global Average Pooling (GAP) layer to produce a feature vector, followed by a sequence of Dropout, a Dense layer with ReLU activation, another Dropout layer for regularization, and a final Softmax layer to output class probabilities.2
* **Attention-Augmented Architecture (EfficientNet-B0 + CBAM):** The experimental model was identical to the baseline, with the sole addition of a **Convolutional Block Attention Module (CBAM)**.2
  + **Integration Strategy:** The CBAM block was strategically inserted immediately after the final convolutional block of the EfficientNet backbone, just before the custom classification head (see Figure 4.4 in the original paper).2
  + **Justification:** This late-stage placement is a deliberate design choice. It allows the pretrained EfficientNet backbone to extract a rich hierarchy of low-level and mid-level features using its ImageNet-learned weights. The attention module then operates on these high-level, semantically rich feature maps, enabling the model to adaptively refine them by focusing on the most discriminative channels and spatial regions just prior to classification.2
  + **CBAM Implementation Details:** The CBAM module was implemented following the specifications of the original paper 2]. The channel attention MLP used a reduction ratio of $r=16$, providing a balance between parameter efficiency and representational capacity. The spatial attention module employed a convolutional kernel of size $7 \times 7$, which allows for a reasonably large receptive field to capture spatial context.2

### 5.3 Training and Evaluation Protocol

To ensure a fair and reproducible comparison, both models were trained and evaluated using an identical protocol.

* **Training Environment:** All experiments were conducted on the Kaggle platform using a Tesla P100 GPU. The models were implemented in TensorFlow/Keras. **Mixed-precision training** was enabled, which uses 16-bit floating-point numbers for most computations while maintaining 32-bit precision for critical components like the final logits. This technique significantly reduces GPU memory consumption and can accelerate training without a substantial loss in accuracy.2
* **Optimizer and Schedule:** The models were trained using the **Adam optimizer** with an initial learning rate of $3 \times 10^{-4}$ and a batch size of 16.2 The training process was governed by a set of callbacks to manage the learning rate and prevent overfitting:
  + ModelCheckpoint: Saved the model weights only when the validation accuracy improved, ensuring that the final model represents the best-performing iteration.
  + ReduceLROnPlateau: Automatically reduced the learning rate if the validation accuracy stagnated, allowing for finer-grained optimization in later stages of training.
  + EarlyStopping: Terminated the training process if the validation accuracy did not improve for a set number of epochs, preventing wasted computation and overfitting.2
* **Loss Function:** The models were optimized by minimizing the **categorical cross-entropy** loss function. To counteract the class imbalance inherent in the dataset, **class weighting** was applied during training. This technique assigns a higher weight to the loss calculated for samples from minority classes (like Hypertension), effectively forcing the model to pay more attention to correctly classifying them.2
* **Evaluation Metrics and Their Justification:** Model performance was assessed using a suite of metrics chosen to provide a comprehensive view, especially in the context of an imbalanced dataset.
  + **Standard Metrics:** Accuracy, Macro F1-score, and Weighted F1-score were reported. While accuracy can be misleading in imbalanced scenarios, Macro F1 (the unweighted average of per-class F1 scores) provides a balanced view of performance across all classes, treating each class equally regardless of its size.
  + **Metrics for Imbalanced Data:** The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) and the Precision-Recall Area Under the Curve (PR-AUC) were chosen as primary evaluation metrics. In medical diagnosis, where correctly identifying a rare disease (a true positive) is often more critical than correctly identifying a healthy case, these metrics are essential. ROC-AUC measures the model's ability to discriminate between positive and negative classes across all possible thresholds and is robust to class imbalance.46 PR-AUC is even more informative when there is a large skew in the class distribution, as it focuses on the performance on the positive class and is not influenced by the large number of true negatives.46 The inclusion of these metrics reflects a rigorous approach to evaluating model performance in a realistic clinical context.

## Chapter 6: Results, In-depth Analysis, and Model Interpretability

### 6.1 Quantitative Performance Analysis

The quantitative results demonstrate a clear and consistent performance improvement when augmenting the EfficientNet architecture with the CBAM module. A direct comparison of the overall test metrics reveals that the attention-augmented model surpasses the baseline across all evaluated criteria.

| **Model** | **Acc** | **Macro F1** | **Weighted F1** | **ROC-AUC (macro)** | **PR-AUC (macro)** |
| --- | --- | --- | --- | --- | --- |
| EfficientNet (baseline) | 0.840 | 0.835 | 0.841 | 0.970 | 0.905 |
| EfficientNet + CBAM | 0.855 | 0.854 | 0.858 | 0.974 | 0.921 |

Table: Overall test metrics from the primary study. All metrics are higher for the CBAM-augmented model.2

A deeper, per-class analysis derived from the row-normalized confusion matrices (Tables 6.2 and 6.3 in the original paper) provides more granular insights into where these improvements originate.2 The recall (or true positive rate) for several key classes shows marked improvement with the CBAM model. Most notably, the recall for **Hypertension** sees a dramatic increase from 62.1% to 82.8%. This is a significant finding, as Hypertension was the most challenging minority class, and the data curation strategy was specifically designed to enable its robust evaluation. Similarly, recall for **Glaucoma** improves from 83.3% to 87.5%, and for **Cataract** from 87.2% to 89.4%.

However, the confusion matrices also highlight persistent challenges. Both models exhibit confusion between certain classes. For instance, the baseline model misclassifies 10.4% of Glaucoma cases as AMD, a confusion that is slightly reduced to 8.3% in the CBAM model. The confusion between AMD and Myopia is also a notable error pattern for both architectures.2 These results suggest that while attention helps to better separate classes, some pathologies may share subtle visual features that remain difficult for the models to distinguish.

### 6.2 Qualitative Analysis of Model Behavior

Qualitative artifacts from the training and evaluation process provide further evidence of the CBAM model's superior performance and generalization capabilities.

* **Training Dynamics:** The training and validation curves for accuracy and loss (Figure 6.1 in the original paper) illustrate the learning behavior of both models.2 The EfficientNet+CBAM model consistently trends towards a higher validation accuracy and, more importantly, a lower final validation loss compared to the baseline. This smoother convergence and reduced gap between training and validation performance suggest that the attention mechanism acts as a form of regularization, helping the model to generalize better to unseen data rather than overfitting to the training set.
* **Class Separability:** The ROC and PR curves (Figures 6.2 and 6.3 in the original paper) offer a visual representation of the models' discriminative power.2 The larger area under both the overall ROC and PR curves for the CBAM model indicates its enhanced ability to distinguish positive samples from negative samples across all classes and at various decision thresholds. The per-class curves further highlight that this improvement is not uniform but is particularly pronounced for certain classes, corroborating the quantitative findings from the confusion matrices.

### 6.3 Unpacking the "Black Box": Advanced Model Interpretability

While quantitative metrics and performance curves demonstrate *that* the CBAM model is better, they do not explain *why*. Model interpretability techniques, also known as Explainable AI (XAI), are essential for understanding the decision-making process of these complex "black box" models, which is particularly crucial in high-stakes domains like medical diagnosis.

* **Revisiting Grad-CAM:** The primary study employed Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize which regions of an input image were most influential for a given prediction.2 The visualizations presented (Figures 6.5 and 6.6) suggest that the CBAM-augmented model produces heatmaps that are more tightly focused on clinically relevant anatomical structures. For instance, in an image of Glaucoma, the CBAM model's heatmap might concentrate more on the optic disc, whereas the baseline's might be more diffuse.2 While insightful, Grad-CAM is a gradient-based method that can sometimes produce noisy or coarse visualizations.

To build a more rigorous and quantifiable understanding of the models' reasoning, more advanced XAI techniques can be employed.

* **Introduction to LIME and SHAP:**
  + **LIME (Local Interpretable Model-agnostic Explanations):** LIME is a model-agnostic technique that explains an individual prediction by learning a simple, interpretable model (e.g., a linear model) in the local neighborhood of that prediction.48 For image classification, LIME works by segmenting the image into "super-pixels" (contiguous patches of similar pixels). It then generates a set of perturbed images by turning some of these super-pixels off (e.g., graying them out). By observing how the model's prediction changes for these perturbed images, LIME fits a local linear model to identify which super-pixels had the most positive or negative influence on the original prediction.48 LIME answers the question: "For this specific image, which regions pushed the prediction towards 'Glaucoma'?"
  + **SHAP (SHapley Additive exPlanations):** SHAP is a unified framework for model interpretation based on Shapley values, a concept from cooperative game theory.50 It assigns each feature (in this case, each super-pixel) an importance value representing its contribution to pushing the model's output from a baseline value to the final prediction. SHAP values have desirable theoretical properties, including consistency and local accuracy, providing a fair and robust attribution of importance to each feature.52 SHAP can generate both local explanations for a single image (force plots) and global explanations that summarize feature importance across the entire dataset (summary plots).50
* Proposed Application for Deeper Analysis:  
  By applying LIME and SHAP, the interpretability analysis can be elevated from a qualitative observation to a quantitative investigation. For a set of representative images, one can generate explanations from both the baseline and CBAM models. This allows for direct hypothesis testing. For example, one could test the hypothesis: "For images correctly classified as Glaucoma, does the CBAM model assign significantly higher positive SHAP values to super-pixels covering the optic nerve head compared to the baseline model?" For misclassifications, these techniques can reveal whether the model was distracted by image artifacts or focused on a clinically irrelevant region. This approach provides a functional link between the architectural component (CBAM) and the performance outcome (higher accuracy). The improved metrics of the CBAM model are not merely a statistical artifact but are a direct consequence of the model learning to assign greater importance to the correct clinical biomarkers. This moves the conclusion from a simple correlation ("CBAM helps") to a more causal explanation ("CBAM helps by enabling the model to more reliably identify and weight known clinical biomarkers"), a significantly stronger and more clinically relevant claim.

## Chapter 7: Synthesis, Limitations, and Future Research Directions

### 7.1 Synthesis of Findings

This report has provided a comprehensive analysis of the application of attention mechanisms for ocular disease classification, centered on the benchmark comparison of an EfficientNet baseline against an attention-augmented variant using the Convolutional Block Attention Module (CBAM) on the ODIR-5K dataset. The core finding of the primary study is that the integration of a lightweight, sequential channel-spatial attention module provides a consistent and statistically significant improvement in classification performance across multiple metrics, including those specifically designed for imbalanced datasets like ROC-AUC and PR-AUC.2

This result is situated within a broader body of literature that confirms the utility of deep learning for ophthalmic diagnosis. The challenges posed by the ODIR-5K dataset—namely class imbalance, multi-label complexity, and label ambiguity—are representative of real-world medical imaging tasks. The success of the EfficientNet+CBAM model validates the strategy of enhancing a strong, data-efficient convolutional backbone with attention mechanisms as a practical and effective approach for mid-sized medical datasets. The analysis further suggests that these performance gains are attributable to the model's improved ability to focus on clinically relevant spatial regions and feature channels, a conclusion supported by both qualitative visualizations and the proposed application of advanced XAI techniques.

### 7.2 Addressing Current Limitations and Proposing Advanced Extensions

While the primary study establishes a robust baseline and demonstrates the value of attention, its methodology contains several limitations that open clear and compelling avenues for future research. The following sections outline these limitations and propose concrete extensions based on state-of-the-art techniques in machine learning and medical imaging.

#### 7.2.1 Limitation 1: Data Scarcity and Basic Augmentation

The study's reliance on basic geometric and photometric augmentations (e.g., flips, rotations) is a common practice but provides limited diversity, which is a significant constraint when working with finite medical datasets.2 To build more robust and generalizable models, more advanced data augmentation strategies are necessary.

* **Proposed Extension: Advanced Data Augmentation Techniques**
  + **Elastic Deformations:** This technique applies random, non-linear local distortions to an image, simulating the natural, non-rigid variability of biological tissues.54 Unlike rigid transformations like rotation, elastic deformations create a "rubber-sheet" effect, which is highly effective for medical imaging where anatomical structures can vary in shape and size due to physiological differences or imaging angles.57 Implementation involves generating a random displacement field, smoothing it with a Gaussian filter to ensure coherence, and then warping the image according to this field. Key parameters include sigma (the standard deviation of the Gaussian filter, controlling smoothness) and alpha (a scaling factor for the displacement intensity).56
  + **Generative Adversarial Networks (GANs):** GANs offer a powerful method for generating entirely new, synthetic-but-realistic fundus images.59 A GAN consists of two competing neural networks: a Generator, which creates new images from random noise, and a Discriminator, which tries to distinguish between real and generated images. Through this adversarial training process, the Generator learns to produce images that are indistinguishable from the real data distribution.60 This technique can be used to generate an almost infinite number of training samples, and is particularly valuable for augmenting minority classes like Hypertension, thereby helping to balance the dataset.62 Common architectures like Deep Convolutional GANs (DCGAN) and CycleGAN (for image-to-image translation) have been successfully applied in medical imaging.63

#### 7.2.2 Limitation 2: Single-Label Problem Formulation

The study's simplification of the inherently multi-label ODIR-5K dataset into a single-label classification problem is a pragmatic choice for simplifying analysis but comes at the cost of losing valuable clinical information about disease comorbidities.2 A patient with both diabetes and glaucoma presents a different clinical picture than a patient with only one of these conditions, and a model trained to recognize these co-occurrences would be more clinically useful.

* **Proposed Extension: Multi-Label Classification Framework**
  + **Theory and Implementation:** Future work should address the problem in its native multi-label formulation, where the model outputs a vector of probabilities, one for each potential disease.65 This is typically achieved by replacing the final Softmax activation function with a Sigmoid function and training the model to predict the presence or absence of each disease independently.
  + **Advanced Loss Functions:** The standard loss function for this setup is the Binary Cross-Entropy (BCE) loss, applied to each class output.67 However, BCE is susceptible to the extreme imbalance between positive and negative samples for each label (most images do not have a given disease). More advanced loss functions are better suited for this task. **Focal Loss**, for example, modifies the BCE loss to down-weight the contribution of easy-to-classify negative samples, forcing the model to focus on harder-to-classify positive samples.68 **Asymmetric Loss (ASL)** further refines this by applying different focusing parameters for positive and negative samples, providing even greater control in scenarios with high imbalance.68

#### 7.2.3 Limitation 3: Reliance on Supervised Pre-training on Natural Images

The models in this study leverage pre-training on ImageNet, a dataset of natural images.2 While this is a standard and effective transfer learning practice, the visual features of natural images (e.g., cars, animals) are fundamentally different from those in medical fundus images. Pre-training on a more domain-relevant dataset could lead to a more effective feature extractor.

* **Proposed Extension: Self-Supervised Learning (SSL) for Domain-Specific Pre-training**
  + **Principles:** SSL provides a way to learn powerful feature representations from large, *unlabeled* datasets.70 Instead of using human-provided labels, the model is trained to solve a "pretext task." A dominant SSL paradigm is **contrastive learning** (e.g., SimCLR, MoCo), where the model learns to pull different augmented views of the same image closer together in an embedding space while pushing views from different images apart.70
  + **Application:** A powerful extension would be to gather a large, unlabeled collection of fundus images (which are more readily available than labeled ones) and pre-train an EfficientNet backbone using an SSL method. This pre-trained model would learn features that are specific to the domain of retinal imaging. The resulting backbone could then be fine-tuned on the smaller, labeled ODIR-5K dataset, potentially leading to significantly improved performance and robustness.70

#### 7.2.4 Limitation 4: Generalization and Data Privacy

The study is conducted on a single, publicly available dataset. Models trained on data from one specific source (a particular set of hospitals, cameras, and patient demographics) often fail to generalize well to data from other sources due to domain shift. Furthermore, sharing medical data between institutions to create larger, more diverse datasets is severely restricted by privacy regulations like HIPAA and GDPR.71

* **Proposed Extension: Federated Learning (FL) for Privacy-Preserving, Multi-Institutional Training**
  + **Concept:** FL is a decentralized machine learning paradigm that enables collaborative model training across multiple institutions without them having to share their raw, private data.72 In a typical FL setup, a central server coordinates the training process. It distributes a global model to participating institutions (clients), each of which then trains the model locally on its own private data. The clients then send their updated model parameters (not the data) back to the server, which aggregates them to create an improved global model. This process is repeated iteratively.72
  + **Application:** A forward-looking research agenda would involve establishing a federated learning network across multiple ophthalmology clinics. This would allow for the training of a single, robust ocular disease classifier on a much larger and more diverse dataset than any single institution could provide. This approach directly addresses the critical challenges of model generalization and data privacy, paving the way for the development of AI tools that are truly viable for real-world clinical deployment.71

### Conclusion

This comprehensive report has systematically deconstructed the task of ocular disease classification, beginning with the clinical foundations of fundus imaging and the specific visual biomarkers of key pathologies. It has provided a rigorous theoretical survey of the deep learning architectures and attention mechanisms relevant to this task, situating the primary study's use of EfficientNet and CBAM within a logical progression of scientific inquiry. The experimental framework was detailed, justifying the methodological choices made to handle the real-world challenges of the ODIR-5K dataset.

The analysis of results was extended beyond a simple performance summary to include a deeper dive into per-class metrics and a proposal for advanced model interpretability using LIME and SHAP, aiming to build a causal link between architectural improvements and performance gains. Finally, the report has transformed the original study's brief mention of future work into a substantial research agenda, proposing concrete, state-of-the-art extensions—including advanced data augmentation, multi-label classification frameworks, self-supervised pre-training, and federated learning—that address the primary limitations of the current work. This roadmap not only suggests incremental improvements but also points toward the development of more robust, clinically relevant, and deployable AI systems for ophthalmology.

#### Works cited

1. The Ultimate Guide to Identifying Retinal Disease on Fundus ..., accessed October 19, 2025, <https://eyesoneyecare.com/resources/ultimate-guide-to-identifying-retinal-disease-on-fundus-photography/>
2. main.pdf
3. Retinal Disease Detection Using Deep Learning Techniques: A ..., accessed October 19, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10145952/>
4. Fundus Image Quality Assessment and Enhancement: a Systematic Review - arXiv, accessed October 19, 2025, <https://arxiv.org/html/2501.11520v1>
5. Errors in Fundus Photography, accessed October 19, 2025, <https://cdn.ymaws.com/www.opsweb.org/resource/resmgr/boc_resourses_pdf/07-2-09.pdf>
6. Fundus Photography - Medical Clinical Policy Bulletins - Aetna, accessed October 19, 2025, <https://www.aetna.com/cpb/medical/data/500_599/0539.html>
7. Fundus photograph-based cataract evaluation network ... - Frontiers, accessed October 19, 2025, <https://www.frontiersin.org/journals/physics/articles/10.3389/fphy.2023.1235856/full>
8. Effect of Cataract Grade according to Wide-Field Fundus Images on Measurement of Macular Thickness in Cataract Patients - PMC - PubMed Central, accessed October 19, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC5990639/>
9. Cataracts: Signs, Symptoms & Treatment - Cleveland Clinic, accessed October 19, 2025, <https://my.clevelandclinic.org/health/diseases/8589-cataracts-age-related>
10. Cataract - Wikipedia, accessed October 19, 2025, <https://en.wikipedia.org/wiki/Cataract>
11. "Explainable Deep Learning for Cataract Detection in Retinal Images: A Dual-Eye and Knowledge Distillation Approach" - arXiv, accessed October 19, 2025, <https://arxiv.org/html/2509.22696v1>
12. Improvement of Retinal Images Affected by Cataracts - MDPI, accessed October 19, 2025, <https://www.mdpi.com/2304-6732/9/4/251>
13. Detecting age-related macular degeneration (AMD) biomarker images using MFCC and texture features, accessed October 19, 2025, <https://par.nsf.gov/servlets/purl/10301904>
14. Age-Related Macular Degeneration (AMD) by Gregory S. Hageman ..., accessed October 19, 2025, <https://www.webvision.pitt.edu/book/part-xii-cell-biology-of-retinal-degenerations/age-related-macular-degeneration-amd/>
15. Automatic Detection and Classification of Hypertensive Retinopathy with Improved Convolution Neural Network and Improved SVM - PMC - PubMed Central, accessed October 19, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10813404/>
16. Fundus camera image showing signs of hypertensive retinopathy, i.e.... | Download Scientific Diagram - ResearchGate, accessed October 19, 2025, <https://www.researchgate.net/figure/Fundus-camera-image-showing-signs-of-hypertensive-retinopathy-ie-arteriovenous-nicking_fig4_263289543>
17. Hypertensive retinopathy signs as risk indicators of cardiovascular ..., accessed October 19, 2025, <https://academic.oup.com/bmb/article/73-74/1/57/332371>
18. IMI Pathologic Myopia - PMC, accessed October 19, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8083114/>
19. Morphological and clinical characteristics of myopic posterior staphyloma in Caucasians, accessed October 19, 2025, <https://www.researchgate.net/publication/301581303_Morphological_and_clinical_characteristics_of_myopic_posterior_staphyloma_in_Caucasians>
20. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks - SciSpace, accessed October 19, 2025, <https://scispace.com/papers/efficientnet-rethinking-model-scaling-for-convolutional-2jsibrxy0c>
21. EfficientNet: Rethinking Model Scaling for Convolutional ... - arXiv, accessed October 19, 2025, <https://arxiv.org/pdf/1905.11946>
22. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks - ResearchGate, accessed October 19, 2025, <https://www.researchgate.net/publication/333444574_EfficientNet_Rethinking_Model_Scaling_for_Convolutional_Neural_Networks>
23. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, accessed October 19, 2025, <https://research.google/pubs/an-image-is-worth-16x16-words-transformers-for-image-recognition-at-scale/>
24. Vision Transformers Explained: The Future of Computer Vision?, accessed October 19, 2025, <https://blog.roboflow.com/vision-transformers/>
25. (PDF) An Image is Worth 16x16 Words: Transformers for Image ..., accessed October 19, 2025, <https://scispace.com/papers/an-image-is-worth-16x16-words-transformers-for-image-v85s5ahlww>
26. Vision Transformer (ViT) Architecture - GeeksforGeeks, accessed October 19, 2025, <https://www.geeksforgeeks.org/deep-learning/vision-transformer-vit-architecture/>
27. A Deep Dive into the Code of the Visual Transformer (ViT) Model - Medium, accessed October 19, 2025, <https://medium.com/data-science/a-deep-dive-into-the-code-of-the-visual-transformer-vit-model-1ce4cc05ca8d>
28. Vision transformer - Wikipedia, accessed October 19, 2025, <https://en.wikipedia.org/wiki/Vision_transformer>
29. (PDF) Attention mechanisms in computer vision: A survey, accessed October 19, 2025, <https://www.researchgate.net/publication/359253090_Attention_mechanisms_in_computer_vision_A_survey>
30. Attention mechanisms in computer vision: A survey, accessed October 19, 2025, <https://d-nb.info/1261706951/34>
31. Squeeze-and-Excitation Networks - arXiv, accessed October 19, 2025, <https://arxiv.org/pdf/1709.01507>
32. Squeeze-and-Excitation Networks - CVF Open Access, accessed October 19, 2025, <https://openaccess.thecvf.com/content_cvpr_2018/papers/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.pdf>
33. [1910.03151] ECA-Net: Efficient Channel Attention for Deep ..., accessed October 19, 2025, <https://ar5iv.labs.arxiv.org/html/1910.03151>
34. BAM: Bottleneck Attention Module | Request PDF - ResearchGate, accessed October 19, 2025, <https://www.researchgate.net/publication/326459629_BAM_Bottleneck_Attention_Module>
35. BAM: Bottleneck Attention Module, accessed October 19, 2025, <https://arxiv.org/pdf/1807.06514>
36. CBAM: Convolutional Block Attention Module - CVF Open Access, accessed October 19, 2025, <https://openaccess.thecvf.com/content_ECCV_2018/papers/Sanghyun_Woo_Convolutional_Block_Attention_ECCV_2018_paper.pdf>
37. arXiv:1807.06521v2 [cs.CV] 18 Jul 2018, accessed October 19, 2025, <https://arxiv.org/pdf/1807.06521>
38. Vision Transformers (ViT) Explained: Are They Better Than CNNs? | Towards Data Science, accessed October 19, 2025, <https://towardsdatascience.com/vision-transformers-vit-explained-are-they-better-than-cnns/>
39. Ocular Disease Recognition - Kaggle, accessed October 19, 2025, <https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k>
40. Deep Learning for Ocular Disease Recognition: An Inner-Class Balance - PubMed Central, accessed October 19, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9071974/>
41. Discriminative Kernel Convolution Network for Multi-Label ..., accessed October 19, 2025, <https://arxiv.org/abs/2207.07918>
42. Discriminative Kernel Convolution Network for Multi-Label Ophthalmic Disease Detection on Imbalanced Fundus Image Dataset - arXiv, accessed October 19, 2025, <https://arxiv.org/pdf/2207.07918>
43. Recognition of eye diseases based on deep neural networks for transfer learning and improved D-S evidence theory - PubMed Central, accessed October 19, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10797809/>
44. Evaluation Metrics for Imbalanced Datasets - Artificial Intelligence, accessed October 19, 2025, <https://schneppat.com/evaluation-metrics.html>
45. What evaluation metric to consider while working on imbalanced data for classfication?, accessed October 19, 2025, <https://www.kaggle.com/discussions/general/585817>
46. Interpretable AI for bio-medical applications - PMC, accessed October 19, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10074303/>
47. Interpretable Machine Learning for Image Classification with LIME - Medium, accessed October 19, 2025, <https://medium.com/data-science/interpretable-machine-learning-for-image-classification-with-lime-ea947e82ca13>
48. Hands-on XAI with SHAP and LIME - Kaggle, accessed October 19, 2025, <https://www.kaggle.com/code/yatrikshah/hands-on-xai-with-shap-and-lime>
49. Explainable Deep Learning Methods in Medical Diagnosis: A ... - arXiv, accessed October 19, 2025, <https://arxiv.org/html/2205.04766>
50. LIME vs SHAP: A Comparative Analysis of Interpretability Tools - MarkovML, accessed October 19, 2025, <https://www.markovml.com/blog/lime-vs-shap>
51. Interpreting Machine Learning Predictions with SHAP and LIME for Transparent Decision Making - IIARD, accessed October 19, 2025, <https://iiardjournals.org/get/IJCSMT/VOL.%2011%20NO.%208%202025/Interpreting%20Machine%20Learning%2022-49.pdf>
52. What is the best data augmentation for 3D brain tumor segmentation? - DiVA portal, accessed October 19, 2025, <https://www.diva-portal.org/smash/get/diva2:1588376/FULLTEXT01.pdf>
53. What is elastic transformation in data augmentation? - Milvus, accessed October 19, 2025, <https://milvus.io/ai-quick-reference/what-is-elastic-transformation-in-data-augmentation>
54. Elastic Transformations - Artificial Intelligence, accessed October 19, 2025, <https://schneppat.com/elastic-transformations.html>
55. Elastic Deformation of Optical Coherence Tomography Images of Diabetic Macular Edema for Deep-Learning Models Training: How Far to Go? - PubMed Central, accessed October 19, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10561735/>
56. data augmentation with elastic deformations - Kaggle, accessed October 19, 2025, <https://www.kaggle.com/code/ori226/data-augmentation-with-elastic-deformations>
57. A Complete Guide to Data Augmentation | DataCamp, accessed October 19, 2025, <https://www.datacamp.com/tutorial/complete-guide-data-augmentation>
58. (PDF) Survey of Image Augmentation Based on Generative ..., accessed October 19, 2025, <https://www.researchgate.net/publication/358845746_Survey_of_Image_Augmentation_Based_on_Generative_Adversarial_Network>
59. Generative Adversarial Networks (GANs) for Medical Image Synthesis and Data Augmentation | Sciety Labs (Experimental), accessed October 19, 2025, <https://labs.sciety.org/articles/by?article_doi=10.20944/preprints202506.1310.v1>
60. Generative Adversarial Networks (GANs) for Medical Image Synthesis and Data Augmentation - Preprints.org, accessed October 19, 2025, <https://www.preprints.org/manuscript/202506.1310/v1>
61. Overview of GAN Use in The Medical Field | by Aysen Çeliktaş - Medium, accessed October 19, 2025, <https://medium.com/@aysenceliktas/overview-of-gan-use-in-the-medical-field-23cb90adf51d>
62. Data augmentation using Generative Adversarial Networks (GANs) for GAN-based detection of Pneumonia and COVID-19 in chest X-ray images - PubMed Central, accessed October 19, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8607740/>
63. Multi‑label classification of biomedical data - PMC, accessed October 19, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11411592/>
64. Multi-label classification with deep learning techniques applied to the B-Scan images of GPR - Comptes Rendus de l'Académie des Sciences, accessed October 19, 2025, <https://comptes-rendus.academie-sciences.fr/physique/articles/10.5802/crphys.193/>
65. Multi-Label Contrastive Learning : A Comprehensive Study - arXiv, accessed October 19, 2025, <https://arxiv.org/html/2412.00101v1>
66. Novel loss functions for ensemble-based medical image ..., accessed October 19, 2025, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0261307>
67. Robust Asymmetric Loss for Multi-Label Long-Tailed Learning - CVF Open Access, accessed October 19, 2025, <https://openaccess.thecvf.com/content/ICCV2023W/CVAMD/papers/Park_Robust_Asymmetric_Loss_for_Multi-Label_Long-Tailed_Learning_ICCVW_2023_paper.pdf>
68. Self-Supervised Learning for Medical Imaging - Lightly, accessed October 19, 2025, <https://www.lightly.ai/blog/self-supervised-learning-for-medical-imaging>
69. Federated Learning for Medical Image Analysis: Privacy-Preserving Paradigms and Clinical Challenges - Scilight Press, accessed October 19, 2025, <https://www.sciltp.com/journals/tai/articles/2508001101>
70. When Federated Learning Meets Medical Image ... - Now Publishers, accessed October 19, 2025, <https://www.nowpublishers.com/article/OpenAccessDownload/SIP-20240048>