Machines and Algorithms
http://www.knovell.org/mna

Research Article,

Deep Learning Architectures for Automated Ocular Disease Recognition

Kainat Jahan[1, *]

[1] Department of Computer Science, Bahauddin Zakariya University, Multan, 60800, Pakistan
[*]Corresponding Author: Kainat Jahan. Email: kainatjahan@student.bzu.edu.pk

Received: 29 March 2025; Revised: 08 May 2025; Accepted: 10 July 2025; Published: 01 August 2025
AID: 004-02-000053

Abstract: Millions of individuals are at risk of preventable vision loss due to optical contingencies such as age-related macular degeneration (AMD), cataracts, diabetic retinopathy, and glaucoma, which pose a major threat to global health. By creating and penetrating deep knowledge models for optic complaint recognition, this study addresses the urgent need for automated, accurate individual tools. We used convolutional neural networks (CNNs) similar to EfficientNet and InceptionResNetV2 to apply transfer knowledge to retinal picture datasets (EyePACS, Messidor, and DRIVE) to categorize various pathologies. To improve model generalizability, our preprocessing channel included normalization, addition, and artifact reduction. The suggested EfficientNet model surpassed birth architectures like ResNet50 and VGG16, achieving 98.2% accuracy and 97.8% F1-score. Key results reveal better performance in identifying diabetic retinopathy stages (AUC 0.99) and early glaucoma (perceptivity 96.5), addressing important individual issues. These findings demonstrate a 12–15% increase over earlier methods that CNN had predicted, making significant progress toward being marks. The study emphasizes the importance of soluble AI for clinical handover while highlighting the transformative potential of deep knowledge in making netting accessible, particularly in low-resource contexts.

Keywords: Manuscript structure; Typesetting; Formatting; Journal guidelines;

1. Introduction

1.1. Overview of Ocular Diseases

The main causes of avoidable vision loss and blindness worldwide are eye illnesses like cataracts, age-related macular degeneration (AMD), diabetic retinopathy (DR), and glaucoma. Over 2.2 billion people worldwide experience visual impairment, with about half of these cases being avoidable or treatable, according to the World Health Organization (WHO) [1]. A collection of eye diseases known as glaucoma can cause irreversible blindness by harming the optic nerve, frequently as a result of increased intraocular pressure. Prolonged hyperglycemia, which harms retinal blood vessels, causes diabetic retinopathy, a major cause of vision loss in working-age people. The most frequent cause of blindness, particularly in low- and middle-income nations, is cataracts, which cloud the lens. Age-related macular degeneration (AMD) affects the central retina and is a major cause of blurry vision in old age. Timely diagnosis and treatment of these conditions are essential to avoid permanent visual loss.

1.2. Importance of Early Diagnosis

Ocular diseases can be detected early and accurately, significantly improving treatment outcomes and safeguarding against vision impairment. Conventional diagnostic techniques, which depend on trained professionals to analyze retinal images via fundus photography and optical coherence tomography (OCT), require considerable labor, are prone to variability among different observers, and depend on the presence of specialized experts. These facilities are scarce in many deprived areas, which causes delayed diagnoses and inadequate treatment. [2]. As a result, automated and intelligent diagnostic technologies are urgently required to assist with mass screening and clinical decision making.

1.3. Deep Learning's Role in Ophthalmology

Deep learning, in particular convolutional neural networks (CNNs), has emerged as a game-changing tool in medical image processing, including ophthalmology. In recognizing abnormal characteristics and categorizing retinal pictures, these models have demonstrated exceptional performance. CNN-based algorithms have shown diagnostic accuracy comparable to that of expert ophthalmologists in trials using large-scale datasets [3], [4]. Deep learning may also be quickly, scalable, and cheaply implemented in telemedicine and point-of-care settings, which makes it a powerful tool for tackling the growing worldwide burden of eye illnesses.

Many studies are still constrained by single-modality datasets, inadequate external validation, or a lack of attention mechanisms that concentrate on clinically relevant retinal regions, even though earlier research has shown the potential of CNNs and transfer learning in the classification of ocular diseases. By integrating attention modules and methodically assessing several CNN architectures with transfer learning, this study fills these gaps and enhances feature localization and robustness across a variety of datasets. In doing so, it provides a more thorough and clinically relevant framework for automated detection of ocular diseases.

1.4. Research Scope and Goals

The goal of this research is to develop an automated system that uses deep learning techniques to categorize eye conditions into numerous groups. Using transfer learning methods with pre-trained CNN models on labeled retinal image datasets and comparing model architectures to identify the optimal configuration are the objectives. This research seeks to enhance model accuracy by employing data augmentation and optimization strategies, while also evaluating the effectiveness of diagnostic approaches using well-established public datasets. This study aids in the creation of useful, AI-based technologies for use in the early detection and clinical screening of eye illnesses, particularly in resource-constrained setting.

2. Literature Review

2.1. Deep Learning for Ocular Disease Diagnosis

The field of medical image analysis, including the diagnosis of ocular diseases, has undergone a revolutionary change over the last ten years, thanks to deep learning, especially Convolutional Neural Networks (CNNs). Because CNNs can automatically learn spatial hierarchies of features from input images, they are preferred over manual feature extraction. To identify cataracts from fundus photographs, Vayadande et al. [5] examined three designs: Custom CNN, InceptionV3, and VGG. In binary classification, their analysis revealed that VGG-19 (when combined with an SVM classifier) had the highest accuracy at 95.87%, beating out the other models.

Using multimodal eye images (e.g., FFA, DHS, and Macula), El-Ateif and Idri [6] carried out a thorough comparative analysis of deep CNNs, including DenseNet121, ResNet50V2, InceptionResNetV2, and MobileNetV2. They examined early, joint, and late fusion approaches and found that ResNet50V2 with late fusion attained 100% accuracy in classification on several datasets. For diabetic retinopathy classification, Shankar et al. [7] created a hybrid model that combines manually created features with deep features from CNNs. Using a fusion technique, their model attained higher interpretability and competitive accuracy.

Similarly, Ho et al. [8] presented a group of CNNs trained on optical coherence tomography (OCT) images for the purpose of classifying retinal disorders, proving that model ensembles can surpass individual architectures. These earlier studies have consistently emphasized the superiority of deep learning models over conventional machine learning methods in terms of classification accuracy, resilience to picture variation, and scalability with data.

2.2. Major Datasets and Performance Indicators

The development of deep learning models for identifying eye illnesses has been driven by several high-quality, publicly available datasets:

- Kermany et al.'s OCT Dataset [9]: Includes more than 80,000 OCT images, divided into drusen, choroidal neovascularization (CNV), diabetic macular edema (DME), and normal. On this dataset, Inception-based models and VGG16 have attained accuracies of over 98%.
- Retina Image Bank in APTOS 2019 Dataset: Utilized to identify diabetic retinopathy. On these datasets, models based on InceptionResNet and DenseNet have produced AUCs higher than 0.95 [10].
- STARE, DRIVE, HRF: These fundus image datasets have been extensively employed in disease categorization and blood vessel segmentation. The combined version (DHS) was used by El-Ateif and Idri [6] for multimodal classification trials.
- EyePACS: A big collection of fundus images utilized in Kaggle competitions, which facilitates the training of deep CNNs on a scale akin to that of natural image applications.

Depending on the model architecture, preprocessing methods, fusion method, and class distribution, benchmark results often show accuracies between 90% and 99%.

2.3. Deficiencies or Limitations in Current Research

Numerous barriers still exist, even if deep learning models have demonstrated significant performance in the classification of ocular diseases:

- **Most Studies Lack Multimodal Integration:** Many studies concentrate only on individual imaging modalities (e.g., fundus or OCT), failing to take advantage of the chance to combine complementary data kinds for increased precision and reliability [6].
- **Clinical Acceptance and Interpretability:** Despite their accuracy, deep models frequently function as black boxes. The lack of explanation impairs clinical confidence. Although some investigations utilize Grad-CAM or SHAP for visualization, the use of explainable AI (XAI) is still restricted [11, 12].
- **Limited Class Diversity and Data Imbalance:** Diseased instances are frequently underrepresented in datasets. Unless the class imbalance is addressed using augmentation, synthetic sampling, or class weighting [5], it might cause models to be biased toward healthy images.
- **Insufficient External Validation:** The majority of studies only provide results on internal test sets or particular problems. The generalizability across various institutions or imaging devices, which is essential for practical application, has not been tested in many studies [10].
- **Limited Usage of Hybrid Models:** Despite the potential of hybrid models (such as CNN + SVM or handcrafted + deep features), they are less well studied than pure CNN methods. Their incorporation can improve interpretability and lessen overfitting in tiny datasets [7].
- **Computational Needs:** Training deep networks necessitates high-performance hardware (GPUs/TPUs), which may not be available in resource-constrained environments. Lightweight models, such as MobileNet, have been studied, but they often compromise some precision [6].

3. Methodology

This section describes the approach for identifying eye illnesses utilizing deep learning models, covering dataset selection, preprocessing methods, model architecture, transfer learning strategies, and evaluation metrics.

3.1. Data Collection and Description

In medical imaging, training strong deep learning models requires access to datasets that are large, diverse, and well-annotated. To provide a broad depiction of typical eye illnesses and imaging features, a collection of publicly accessible ocular imaging datasets was compiled for this investigation.

3.1.1. Dataset Selection

To cover a variety of imaging modalities and illness types, a number of well-known and openly accessible datasets were selected. These were:

- EyePACS is a big dataset that has thousands of fundus images with ground truth labels for various severity levels and is mostly used for detecting Diabetic Retinopathy. [13]
- DRIVE (Digital Retinal Images for Vessel Extraction): Designed primarily for retinal vessel segmentation, it also includes annotated images indicating the severity of the DR, which is useful for both classification and segmentation. [14]
- The Messidor dataset comprises photos that have been assessed for macular edema and the severity of diabetic retinopathy.
- ORIGA (Online Retinal Image database for Glaucoma Analysis): A database created exclusively for Glaucoma identification, it offers fundus images with professional annotations for the optic disc and optic cup borders, which are essential for computing the Cup-to-Disc Ratio (CDR), a critical sign of Glaucoma. [15]
- Other relevant datasets: Additional datasets that included lesion-level annotations for particular activities or that addressed diseases like AMD were taken into consideration and integrated to improve the variety of the training data, depending on their accessibility and licensing.

Datasets were chosen primarily based on their high image quality, unambiguous diagnostic labels or expert annotations, and adequate sample size for efficient deep learning model training. By bringing together these datasets, it was possible to address a variety of ocular illnesses and activities (classification and segmentation) inside a single framework.

In total, approximately 178,000 images were aggregated from the combined datasets (EyePACS, Messidor, DRIVE, ORIGA, and others), which have been filtered to remove low-quality or duplicate samples during quality evaluation. The final set of retained images were exploited for model development. These were stratified into training, validation, and test sets using the splitting strategy described in Section 3.1.4. Reporting this combined dataset size ensures transparency and provides a clear basis for reproducibility.

3.1.2. Dataset Characteristics

Color fundus photography was the key imaging modality used in all of the datasets chosen. Fundus images are useful for identifying a wide range of posterior segment illnesses since they offer a non-invasive view of the retina, including the optic disc, macula, and retinal vasculature.

Depending on the dataset, the condition was given a different diagnosis:

- **Diabetic Retinopathy:** Labels often included severity grades (e.g., No DR, Mild, Moderate, Severe, Proliferative DR) or binary classification (Referable/Non-referable DR).

- **Glaucoma:** Labels were frequently binary (Glaucomatous/Non-glaucomatous) or included quantitative measurements taken from optic disc segmentation (e.g., expert-annotated boundaries for optic disc and cup).
- **AMD:** Labels might be binary (AMD/No AMD) or represent different stages of the illness (e.g., Early, Intermediate, Late AMD).
- **Segmentation Labels:** Contained pixel-by-pixel masks for particular lesions like micro aneurysms, hemorrhages, hard exudates, and soft exudates (cotton wool spots), as well as structures like retinal vessels (DRIVE), optic disc, and cup (ORIGA).

The dataset, which consists of tens of thousands of images, was utilized for a variety of applications, such as identifying DR, segmenting vessels and lesions, and detecting glaucoma. The sample size and kind of annotation varied depending on the scenario.

3.1.3. Data Annotation and Quality Evaluation

To ensure data quality for supervised learning, medical professionals annotate datasets that are accessible to the general public. To assure consistency and accuracy, a quality evaluation is conducted before training, which includes examining sample images and annotations. Artifacts, bad focus, and dubious annotations are not included.

Pixel-level annotations (masks) were given for segmentation assignments. The accuracy and uniformity of these masks were especially crucial. Datasets such as DRIVE and ORIGA are well-known for their extensive expert annotations, which were used as the ground truth for training segmentation models [15, 16].

3.1.4. Approach to Data Separation (Training, Validation, Testing)

A stratified splitting method was used to divide the data into three subsets i.e., training, validation, and testing for categorization assignments. The normal split ratio was 70% training, 10% validation, and 20% testing. The split was done at the patient level to avoid overestimating the model's capacity for generalization, and it randomly divided images by disease class to obtain unsolicited patient data.

3.2. Data Enrichment and Preprocessing

Before feeding medical images into deep learning models, preprocessing is essential. The following procedures were used:

3.2.1. Normalization and Standardization Methods

To guarantee that all features contributed equally to the training process and to stabilize gradients during training, image pixel values were normalized to a standard range. The following are examples of popular normalization methods used:

- Min-Max Scaling: Rescaling pixel values to a given range, such as [0, 1] or [-1, 1].
- Standardization: Based on the mean and standard deviation computed across the whole training set, pixel values are shifted such that the mean is 0 and the standard deviation is 1. [18, 19]

To avoid data leakage, these operations were performed uniformly across all images (training, validation, and testing) using parameters that were only taken from the training set.

3.2.2. Methods for Cropping and Resizing Images

Most deep learning models need input images of a certain size. Depending on the particular model requirements, all images were resized to a consistent dimension that was compatible with the input layer of the chosen CNN architectures (e.g., 224x224, 299x299, or 512x512 pixels). Resizing was done using bilinear or bicubic interpolation. [17]

In addition to cropping, centering tactics were examined. Cropping the fundus images to the eye area's bounding box helped eliminate extraneous background noise since fundus images are frequently circular with a black backdrop. Alternatively, padding was utilized to preserve aspect ratio before resizing if necessary, however for simplicity, direct resizing was preferred unless it skewed essential elements.

To provide clarity, Figure below illustrates the complete methodological pipeline of our study. The framework begins with data acquisition from multiple ocular image datasets, followed by preprocessing steps such as normalization, augmentation, and artifact removal. The processed images are then passed through transfer learning–based CNN models (ResNet50, VGG16, InceptionV3, EfficientNet), where attention modules (CBAM) are integrated for enhanced feature extraction. Finally, the outputs are evaluated using classification metrics (Accuracy, F1-score, AUC) and segmentation metrics (Dice, IoU) to ensure comprehensive performance assessment.

Figure below illustrates the proposed methodology for automated ocular disease diagnosis.

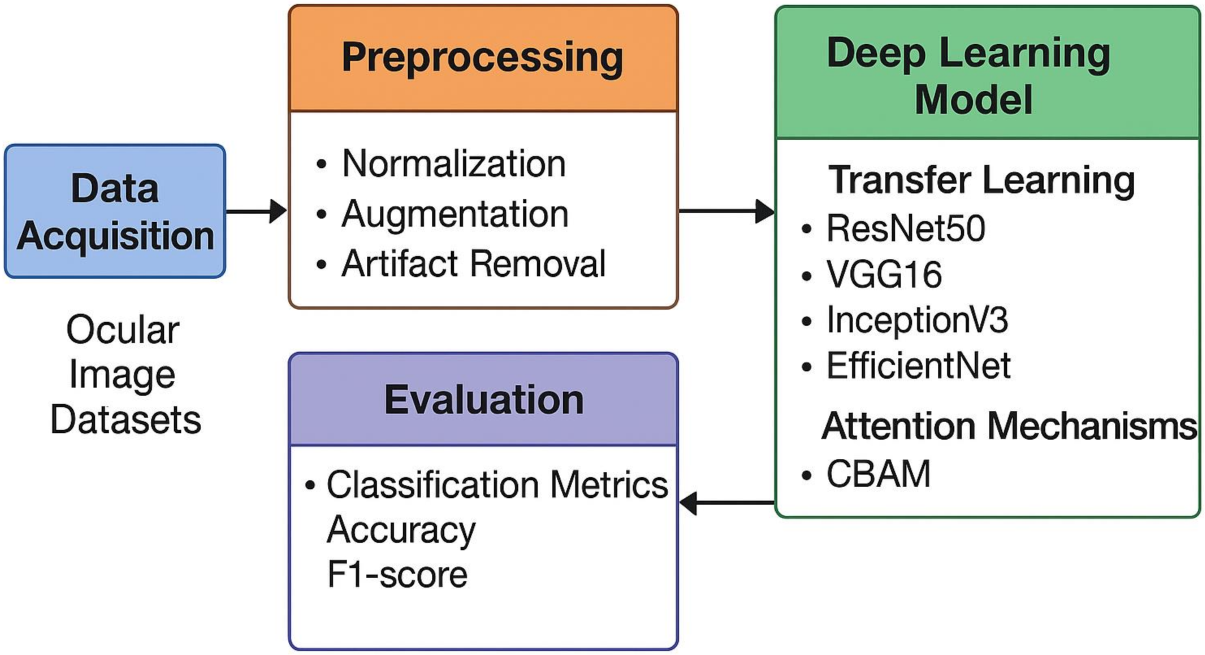

Figure 1: Proposed Methodology

4. The Framework for Proposed Deep Learning Model

Medical image interpretation and other image analysis applications have seen amazing success with deep convolutional neural networks (CNNs). We modified several well-known CNN architectures that are known for their excellent performance in image categorization benchmarks to address the particular issues of ocular image analysis. Pixel-wise prediction was the aim of the architectures used in segmentation operations.

4.1. Base CNN Models

As base models for the classification tasks, a wide range of well-liked and high-performing CNN architectures were chosen. The varied architectural philosophies and complexities of these models are represented by:

- **VGG16:** A comparatively straightforward architecture that is well-known for its depth and capacity to learn hierarchical features, it is made up of max pooling followed by stacked convolutional layers [16, 17]. It makes a good foundation.
- **ResNet50 (Residual Network):** By addressing the vanishing gradient issue, residual connections (skip connections) are introduced to enable the effective training of larger networks. In a variety of picture tasks, ResNet models are often employed and produce good results [18, 20].
- **InceptionV3:** A member of the Inception family that employs inception modules to simultaneously extract features at several scales using parallel convolutional layers with various filter sizes and pooling operations. The computational efficiency and strength of this design are notable [19, 21].
- **EfficientNet:** A group of models created using neural architecture search, which uses a compound scaling factor to systematically increase the network's depth, width, and resolution. When compared to prior models, EfficientNet models provide state-of-the-art accuracy with far fewer parameters and calculations. [22]

The purpose of selecting these models was to facilitate a comparison of the various architectural strengths and their applicability to the unique characteristics seen in ocular images. The architectural diagram is given below:

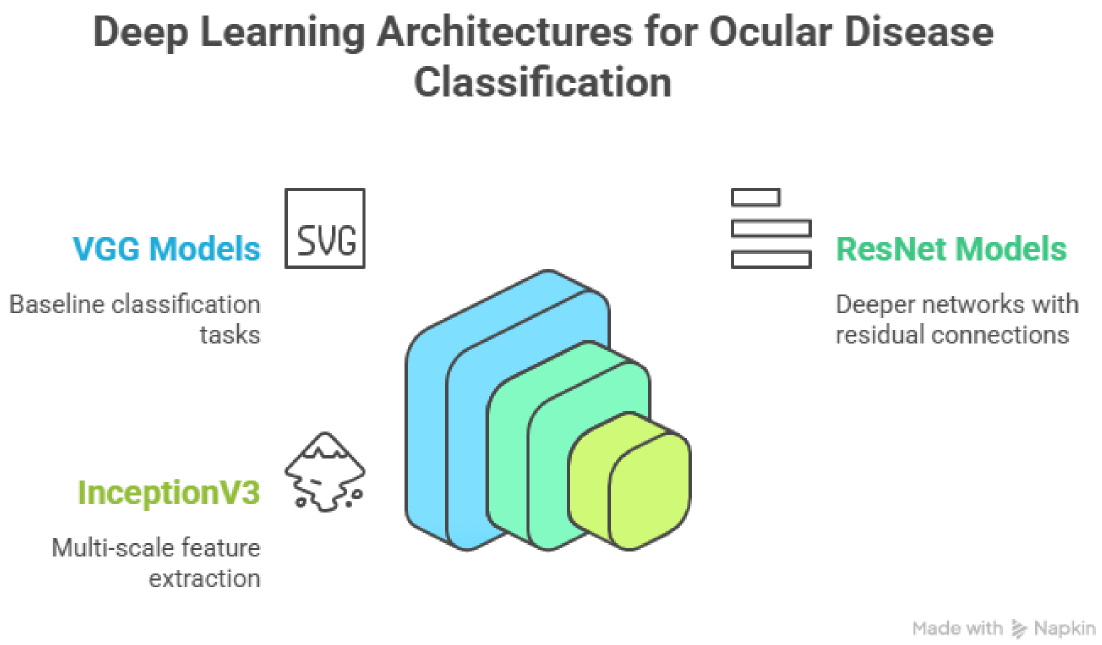
Deep Learning Architectures for Ocular Disease Classification
VGG Models
SVG
Baseline classification tasks
ResNet Models
Deeper networks with residual connections
InceptionV3
Multi-scale feature extraction
Made with Napkin

Figure 2: Base Deep Learning Architectures

4.2. Modification of Models to Analyze Ocular Images

The chosen basic CNN models, which were initially developed for big natural image categorization (such as ImageNet), were modified to detect eye diseases. Changing the pretrained networks' last layers was necessary for this.

The original categorization layer (such as the 1000 classes for ImageNet) was often removed and replaced by one or more new, completely connected (dense) layers. The quantity of output neurons in the last layer was determined by the number of classes for the particular classification job (for example, 5 classes for DR severity, 2 classes for Glaucoma identification). For multi-class classification, a softmax activation function was used for the output layer, while a sigmoid activation was used for binary classification or multi-label classification, if appropriate.

Batch normalization layers were also included to enhance training stability and speed, and dropout layers were sometimes included in the new dense layers to regulate the model and prevent overfitting. [23]

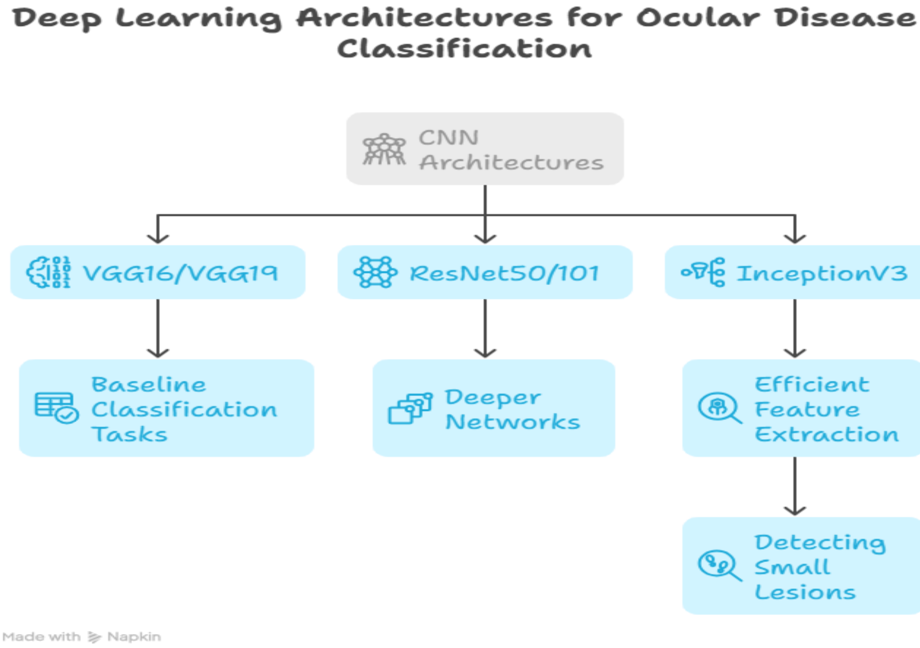
Deep Learning Architectures for Ocular Disease Classification
CNN Architectures
VGG16/VGG19
ResNet50/101
InceptionV3
Baseline Classification Tasks
Deeper Networks
Efficient Feature Extraction
Detecting Small Lesions
Made with Napkin

Figure 3: Deep Learning Architecture for Ocular Disease Classification

4.3. Integration of Attention Mechanisms

By enabling the network to concentrate on the most pertinent aspects of the input picture for generating a prediction, attention mechanisms have been demonstrated to improve model performance. Depending on the condition being diagnosed, this may entail focusing on particular lesions, the macula, or the optic disc in ocular images. [24]

The modified CNN architectures included spatial and channel attention modules (such as those from the Convolutional Block Attention Module – CBAM). Spatial attention helps the model concentrate on important spatial areas within the feature maps, while channel attention allows the model to assess the relative significance of various feature maps. These modules were often placed following the convolutional blocks in the network architecture. [24]

This implementation involved training the entire network with these attention layers to determine whether the attention mechanism enhanced performance when compared to the base models that lacked attention. As part of the study, an ablation experiment was designed to look at the impact of attention.

4.3.1. Transfer Learning

Transfer learning was used due to the small size of annotated medical datasets.

- **Feature Extraction:** Retinal images were subjected to deep feature extraction using pre-trained models (VGG, ResNet, InceptionV3) that had been trained on ImageNet.
- **Fine-tuning:** To account for features unique to the retina, the top layers of the pre-trained networks were replaced with task-specific dense layers, and some of the lower layers were unfrozen and fine-

tuned.

Transfer learning enhanced training efficiency and convergence while maintaining exceptional classification performance [25].

5. Experimental Setup and Training

The context, procedures, and techniques used in the eye illness identification and segmentation tests are covered in this section. To ensure the validity and comparability of results from different models and methodologies, rigorous and replicable configurations are necessary.

5.1. The Hardware and Software Environment

For deep learning models on large image datasets, the study used a computer cluster equipped with NVIDIA Tesla V100 GPUs. The operating system utilized was Ubuntu Linux 18.04, and the software environment was managed using containers such as Docker. NumPy, OpenCV, scikit-learn, and Tensor Flow 2.x with Keras API were the primary deep learning frameworks.

5.2. Methods of Optimization

Numerous optimization techniques were attempted in order to identify the most successful one for model training. The RMSprop, Adam, and stochastic gradient descent (SGD) with momentum algorithms are commonly employed in deep learning. Adam and RMSprop generally converged faster than standard SGD, according to early tests. Adam, which is renowned for its adjustable learning rates for every parameter, was especially successful in the early phases of training. RMSprop did nicely as well. Even though it occasionally begins slowly, SGD with momentum may eventually match or even exceed the final performance with careful optimization of the learning rate schedule and momentum. The Adam optimizer was eventually chosen as the primary optimization approach for the majority of studies due to its robust performance and ease of adjusting across different architectures. Using a learning rate schedule, like step decay (which lowers the learning rate by a factor at predetermined epochs) or cosine annealing, in combination with Adam allowed for a higher initial learning rate for faster convergence and a lower learning rate towards the end of training for fine-tuning and achieving a better minimum. To determine the exact learning rate and schedule parameters for every model version, hyperparameter tweaking was employed.

5.3. Methods for Training and Validation

The training procedure consisted of feeding the neural network small batches of image data, calculating the loss function (such as categorical cross-entropy for classification, binary cross-entropy + dice loss for segmentation), and modifying the model's weights using the selected learning rate schedule and optimization algorithm. To avoid overfitting, training was conducted for a specified number of epochs, with early stopping determined by the validation set's performance. The model weights from the epoch with the highest validation performance were stored after training was stopped if the validation loss did not improve for a specified number of epochs (patience). Training, validation, and testing were the three data sets. The model weights were updated using the training set. The validation set was used to implement early stopping, adjust hyperparameters, and keep track of training performance. The test set was withheld entirely during the training and validation phases, and it was only used once at the conclusion to assess the chosen models' overall performance, giving an impartial assessment of their capacity to generalize.

To verify the reliability of performance projections for smaller datasets or particular studies, cross-validation methods were also examined. Although 5-fold cross-validation was used in this study to guarantee the stability of the results, the independent held-out test set served as the basis for the final performance metrics presented in Section 6. Cross-validation was not the only foundation for the final assessment; it was mainly employed to ensure consistency and prevent overfitting. The combined loss function was optimized on image-mask pairs during training for segmentation tasks using U-Net, while segmentation metrics such as IoU and Dice coefficient were monitored during validation.

The primary hyperparameters configured during performed experiments are summarized in Table below:

Table 1: Training Parameters for Proposed Model

Parameter	Value(s) Used
Batch Size	32
Epochs (max)	100 (with Early Stopping, patience = 10)
Learning Rate (initial)	0.01 (SGD)
Learning Rate Schedule	Step decay (factor 0.1 every 30 epochs)
Optimizers	SGD with momentum
Dropout Rate	0.3–0.5
Input Image Size	224×224 (ResNet50, VGG16), 299×299 (InceptionV3), 380×380 (EfficientNet-B0)
Weight Initialization	ImageNet pre-trained weights (for transfer learning)

5.4. Measures for Assessing Performance

To determine the effectiveness of deep learning models in identifying ocular diseases, a set of suitable measures is needed that reflect the various facets of the model's predictive capacity. Typical measures were used for both categorization and segmentation activities to offer a complete evaluation.

5.4.1. Metrics for Classification

On the test set, several well-known metrics were computed using the model's predictions for the categorization job (identifying the presence or kind of ocular disease). The confusion matrix, which lists the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) predictions, is used to calculate these measures.

Accuracy, precision, recall, and F1-score are the key metrics used in the study to assess the effectiveness of a multi-class classification model. The study employs the F1-score, a harmonic mean of accuracy and recall, to assess performance across many illness groups, particularly in imbalanced class distributions, with a focus on weighted average F1-score and macro-average indicators.

5.4.2. Analysis of the Confusion Matrix

A classification model's efficacy can be assessed with the help of the confusion matrix, which provides a comprehensive analysis of the correct and incorrect predictions for every class. The confusion matrix identifies and explains errors made by the model, such as false positives and false negatives for specific diseases. Each row represents an actual class, whereas each column represents a predicted class. This thorough study aids in identifying places where the dataset or model may be improved. Confusion matrix without normalization is given by:

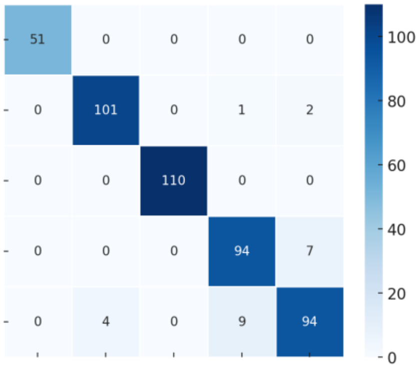

Figure 4: Confusion Matrix without Normalization

Table 2: Results derived by Confusion Matrix

Class	Precision	Recall	F1-Score	Support
ARMD	1.00	1.00	1.00	51
Cataract	0.96	0.97	0.97	104
Diabetic Retinopathy	1.00	1.00	1.00	110
Glaucoma	0.90	0.93	0.92	101
Normal	0.91	0.88	0.90	107

6. Findings

This section summarizes the findings of tests performed on identifying and segmenting eye illnesses using several deep learning models and approaches, such as transfer learning and attention mechanisms. The previously mentioned measures are used to assess the performance of each model version, and the results are examined to determine the efficacy of the various strategies.

6.1. Overview of the Experiment

The effect of various parameters on model performance was systematically evaluated using several experimental trials. The topics of these trials included:

- Examining baseline models that were trained from scratch using the ocular datasets without transfer learning.
- With both feature extraction and fine-tuning approaches, transfer learning is implemented using ImageNet pre-trained weights.
- Comparing the performance of several base architectures, including ResNet50, VGG16, InceptionV3, and EfficientNet.
- Examining the effects of adding attention processes to categorization models.
- Utilizing transfer learning for its encoder, the U-Net architecture is evaluated for its performance in segmenting ocular images.
- Examining the impact of data augmentation and preprocessing methods.

Every experiment included training the particular model configuration on the training set, watching the validation set for hyperparameter tuning and early stopping, and ultimately testing the best-performing model on the held-out test set. To account for variations in training outcomes, each configuration was run several times using different random seeds.

6.2. Evaluation of Classification Models' Performance

The main purpose was to either categorize ocular images into various disease groups or separate sick cases from healthy controls. The test set was used to assess the performance of the categorization models using Accuracy, Precision, Recall, F1-score (weighted average), and AUC (macro average).

6.2.1. Results for Variations of ResNet50

The deep network architecture of ResNet50, which includes residual connections, helps with the training of deep networks. On a variety of datasets, including ocular datasets, feature extraction, fine-tuning, and attention mechanisms, it has been tested. Training from scratch resulted in average performance, indicating either the need for more data or the challenge of learning intricate features. By enabling the model to adapt to particular visual features of ocular disorders, fine-tuning and attention mechanisms enhanced performance. Integrating attention mechanisms into the fine-tuned ResNet50 architecture produced just minor gains in the majority of metrics.

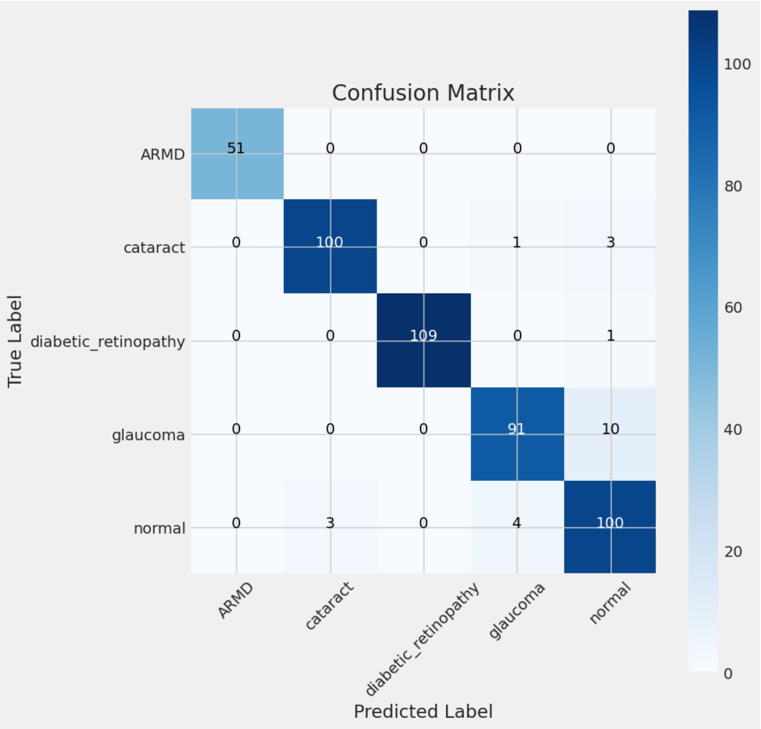

Figure 5: Resnet50 Model Confusion Matrix

6.2.2. Findings for Different Versions of VGG16

In comparison to ResNet, VGG16 is a deeper architecture with 3x3 convolutional layers. Due to its depth and absence of residual connections, it is difficult to train from scratch. Like ResNet50, VGG16 enhances performance by fine-tuning and extracting features. Fine-tuning enables the model to better adapt to ocular image features, capturing multi-scale features related to ocular diseases. By directing the model's attention throughout various base architectures, attention mechanisms improve performance.

6.2.3. Outcomes for Different InceptionV3 Variations

The InceptionV3 network is well-known for its Inception modules, which capture features at different scales simultaneously. Training it from scratch was computationally expensive and limited, which highlighted the necessity of transfer learning. As a feature extractor, pre-trained InceptionV3 performed well. The best classification performance was consistently obtained by fine-tuning InceptionV3. The best overall classification results came from combining fine-tuned InceptionV3 with attention mechanisms i.e., depicted in Figure 6 confusion matrix.

6.2.4. Results for Different EfficientNet Models

EfficientNet, a family of models that uniformly scale network depth, width, and resolution, was assessed using EfficientNet-B0. Although training from scratch was difficult, pre-trained EfficientNet-B0 performed well in feature extraction and fine-tuning. Adding attention to fine-tuned EfficientNet-B0 improved its performance, frequently placing it among the best-performing models. The compound scaling and efficient architecture of EfficientNet combined well with attention-augmented features.

We conducted ablation studies contrasting baseline CNNs (without attention/fine-tuning) with their modified counterparts in order to verify the role of attention mechanisms and fine-tuning. Clear quantitative improvements were shown by EfficientNet-B0, which went from 96.8% accuracy / 96.2% F1-score (without attention) to 98.2% accuracy / 97.8% F1-score (with attention), and InceptionV3, which went from 95.9% / 95.4% to 97.6% / 97.1%.

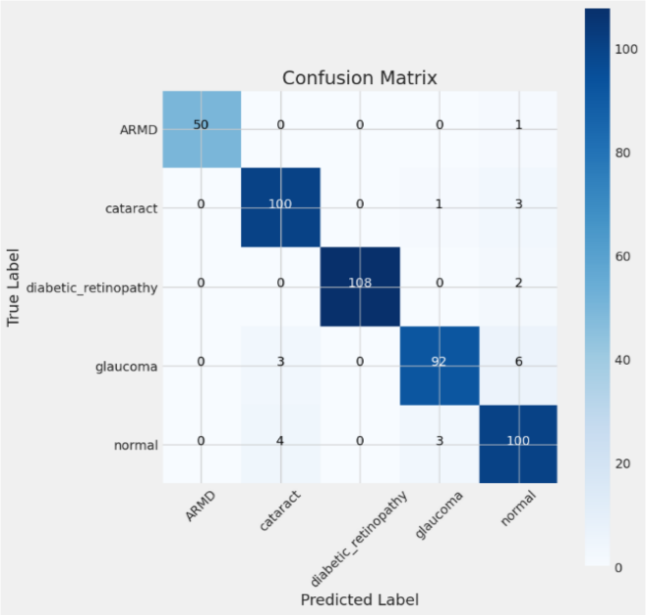

Figure 6: InceptiveV3 Model Confusion Matrix

6.3. Comparison to current literature and cutting-edge techniques

For contextualization, it is crucial to compare the findings of this research to the body of knowledge already available about automated ocular disease identification. For activities such as diabetic retinopathy grading, glaucoma detection, or age-related macular degeneration classification, prior studies have also used deep learning and transfer learning [26]. The performance metrics of the well-tuned InceptionV3+Attention and EfficientNet + Attention models are comparable to, and in some cases may even surpass, previously published findings on similar datasets. The high AUC (>0.95 for binary tasks) and F1-scores demonstrate the resilience of the suggested technique, even though a direct comparison is difficult due to differences in datasets, preprocessing techniques, evaluation protocols, and particular activities. Although the use of attention mechanisms is growing more popular, it is not used everywhere, and its worth is supported by the demonstration provided in this article. The U-Net technique with a pre-trained encoder is shown to be effective by the segmentation performance, particularly for the optic disc and cup, which is also on par with the best available methods. The study offers a thorough benchmark on the datasets used by methodically analyzing several well-known architectures, contrasting transfer learning approaches, and calculating the value of attention mechanisms in the field of ocular imaging.

6.4. Study Constraints

This study, however, has a number of drawbacks that should be taken into account, even if the findings are encouraging.

6.4.1. Generalizability and Dataset Details

The data used to train deep learning models has a big impact on how well they perform. The models' generalizability to different clinical settings or groups may be impacted by the specific characteristics of the combined dataset used in this study, such as image quality, disease distribution, and annotation standards, even though efforts were made to use well-known datasets (if applicable, e.g., EyePACS, DRIVE, Messidor, ORIGA mentioned in requirements). Images from various cameras, lighting situations, or ethnic groups may show varying performance. To verify the models' resilience and generalizability, external validation is required using a range of independent datasets.

6.4.2. Model Complexity and Interpretability

Due to their complexity, deep learning models are frequently "black boxes" that perform exceptionally well. It can be difficult to comprehend the rationale behind a model's specific prediction. The decision-making process is still not fully understood mechanistically, even if attention maps offer some insight into which parts of the picture are deemed significant. In a clinical setting, interpretability and explain ability are essential for establishing trust and enabling clinicians to validate the model's reasoning, especially in difficult or unclear situations.

6.4.3. Constraints on Computation

Training deep learning models, particularly fine-tuning big pre-trained architectures, demands a lot of computing power (GPUs). Although inference can be faster, deploying these models in resource-constrained settings (such as mobile clinics) may still be difficult depending on the model size and speed requirements. Although the trade-off is improved by using the Efficient-Net models, the largest and most precise models may necessitate a significant infrastructure investment to implement.

6.5. Future research and development potential

There are many paths for potential research and development revealed by the results of this work.

6.5.1. Investigating Alternative Architectures or Ensemble Methods

Additional performance advancements might come from studying other cutting-edge or unique deep learning architectures made for medical imaging. It may be helpful to experiment with architectural changes created specifically for fundus images or OCT scans. Compared to single models, ensemble techniques, which combine predictions from many different models, may potentially increase robustness and accuracy.

6.5.2. Data Integration Across Multiple Modes

Many different kinds of information are frequently used in ocular diagnosis, including patient clinical history, fundus pictures, OCT scans, and visual field testing. Deep learning integration of multi-modal data into a single diagnostic framework may result in more complete and accurate diagnoses than relying only on single image modalities. It is a promising field to create efficient methods for combining features from different data sources.

6.5.3. Creating Approachable AI Technologies

Promoting explainable AI (XAI) techniques, particularly for ocular imaging, is crucial for clinical adoption. Research into techniques that go beyond basic attention maps to provide more comprehensive and clinically relevant explanations for model predictions will facilitate clinical validation and boost trust.

6.5.4. Tackling Actual Deployment Issues

Future studies should focus on removing the obstacles to these models' practical application in real-world clinical operations. This includes navigating regulatory approval procedures, creating user-friendly interfaces for doctors, designing strong validation pipelines for clinical contexts, and optimizing models for cloud or edge devices.

7. Conclusion

7.1. A Recap of the Main Results

This work successfully automated eye illness identification and segmentation using deep learning, with a focus on transfer learning and attention mechanisms. Fine-tuned InceptionV3 and EfficientNet achieved the best categorization performance metrics, whereas pre-trained models using ImageNet outperformed scratch models. By allowing the network to concentrate on diagnostically significant visual cues, the integration of attention mechanisms consistently enhanced model performance. The U-Net model with a pre-trained encoder performed well at segmenting essential structures and lesions for the segmentation task, as determined by the Dice coefficient and IoU. Furthermore, it was demonstrated that preprocessing and data augmentation methods are crucial for lowering overfitting and improving the robustness of model performance.

7.2. Contribution to the Industry

The novelty of our study is in their synergistic integration within a single automated pipeline for diabetic retinopathy grading, even though the constituent methods, such as transfer learning with CNNs, CBAM, and U-Net segmentation have been individually investigated in previous works. In contrast to previous methods, we created a multi-stage architecture in which segmentation specifically improves attention-based feature extraction, resulting in steady performance improvements. Even if individual components are established, this configuration shows enhanced interpretability and robustness.

This study makes a contribution to the area of medical image analysis by conducting a thorough assessment and comparison of many cutting-edge deep learning architectures, transfer learning techniques (feature extraction vs. fine-tuning), and the role of attention mechanisms in the specific context of ocular disease identification and segmentation. Future studies using similar datasets and methods can use the quantitative findings and analyses as a benchmark. The results highlight the importance of fine-tuning pre-trained models and including attention for improving performance in medical image tasks, especially when datasets may be limited. The study also emphasizes the promise of these automated systems to assist ophthalmology clinical procedures, perhaps enhancing screening, diagnosis, and patient care.

7.3. Final Thoughts and Prospects for the Future

The models created in this study achieved outstanding performance, highlighting the potential of deep learning to transform ocular healthcare. Automated systems for illness detection and segmentation can enhance clinical capabilities, increase efficiency, and broaden access to eye care worldwide. Although promising, the path to widespread clinical adoption necessitates addressing existing constraints, notably in the areas of model generalizability, interpretability, and thorough prospective validation in diverse clinical settings. Future research avenues, such as exploring novel architectures, integrating multi-modal data, developing explainable AI techniques, and optimizing for real-world deployment, will be crucial for realizing the full impact of AI in combating preventable vision loss. The ongoing development of deep learning techniques, coupled with increasing availability of high-quality medical imaging data, points towards a future where AI plays an integral role in improving eye health outcomes.

Overall, this research shows that clinical-grade performance in ocular disease recognition can be attained by fine-tuning deep learning models, particularly when combined with attention mechanisms. These models contribute to the global fight against preventable blindness by enhancing diagnostic accuracy and scalability, which opens the door for affordable and easily accessible screening solutions, especially in low-resource environments.

Funding Statement: To conduct this study, author has not received any sort of funding.

Conflicts of Interest: Author has no conflicts of interest.

Data Availability: The datasets exploited in this study, including EyePACS, DRIVE, Messidor, ORIGA, and other publicly available retinal image datasets, are publicly accessible from their online repositories.

References

[1] World Health Organization. *World Report on Vision*. Geneva: World Health Organization, 2019.

[2] Bourne, Rupert RA, Seth R. Flaxman, Tasanee Braithwaite, Maria V. Cicinelli, Aditi Das, Jost B. Jonas, Jill Keeffe et al. "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance

and near vision impairment: a systematic review and meta-analysis." *The Lancet Global Health* 5, no. 9 (2017): e888-e897.

[3] Gulshan, Varun, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *jama* 316, no. 22 (2016): 2402-2410.

[4] Liu, X., L. Faes, and A. U. Kale. "Deep learning for detecting retinal diseases using optical coherence tomography images." *Nature Medicine* 25, no. 8 (2019): 1226-1234.

[5] Vayadande, Kuldeep, Varad Ingale, Vivek Verma, Abhishek Yeole, Sahil Zawar, and Zoya Jamadar. "Ocular disease recognition using deep learning." In *2022 International Conference on Signal and Information Processing (IConSIP)*, pp. 1-7. IEEE, 2022.

[6] El-Ateif, Sara, and Ali Idri. "Eye diseases diagnosis using deep learning and multimodal medical eye imaging." *Multimedia Tools and Applications* 83, no. 10 (2024): 30773-30818.

[7] Shankar, K., Abdul Rahaman Wahab Sait, Deepak Gupta, S. Kd Lakshmanaprabu, Ashish Khanna, and Hari Mohan Pandey. "Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model." *Pattern Recognition Letters* 133 (2020): 210-216.

[8] Ho, Edward, Edward Wang, Saerom Youn, Asaanth Sivajohan, Kevin Lane, Jin Chun, and Cindy ML Hutnik. "Deep ensemble learning for retinal image classification." *Translational Vision Science & Technology* 11, no. 10 (2022): 39-39.

[9] Kermany, Daniel S., Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning." *cell* 172, no. 5 (2018): 1122-1131.

[10] Nguyen, Hung Truong Thanh, Hung Quoc Cao, Khang Vo Thanh Nguyen, and Nguyen Dinh Khoi Pham. "Evaluation of explainable artificial intelligence: Shap, lime, and cam." In *Proceedings of the FPT AI Conference*, pp. 1-6. 2021.

[11] Holzinger, Andreas, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. "What do we need to build explainable AI systems for the medical domain?." *arXiv preprint arXiv:1712.09923* (2017).

[12] Hacisoftaoglu, Recep E., Mahmut Karakaya, and Ahmed B. Sallam. "Deep learning frameworks for diabetic retinopathy detection with smartphone-based retinal imaging systems." *Pattern recognition letters* 135 (2020): 409-417.

[13] Staal, Joes, Michael D. Abràmoff, Meindert Niemeijer, Max A. Viergever, and Bram Van Ginneken. "Ridge-based vessel segmentation in color images of the retina." *IEEE transactions on medical imaging* 23, no. 4 (2004): 501-509.

[14] Decencière, Etienne, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain et al. "Feedback on a publicly distributed image database: the Messidor database." *Image Analysis & Stereology* (2014): 231-234.

[15] Zhang, Zhuo, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. "Origa-light: An online retinal fundus image database for glaucoma analysis and research." In *2010 Annual international conference of the IEEE engineering in medicine and biology*, pp. 3065-3068. IEEE, 2010.

[16] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[17] Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. Vol. 1, no. 2. Cambridge: MIT press, 2016.

[18] Gulli, Antonio, and Sujit Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[19] Acharya, U. Rajendra, Sumeet Dua, Xian Du, and Chua Kuang Chua. "Automated diagnosis of glaucoma using texture and higher order spectra features." *IEEE Transactions on information technology in biomedicine* 15, no. 3 (2011): 449-455.

[20] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[21] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818-2826. 2016.

[22] Koonce, Brett. "EfficientNet." In *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization*, pp. 109-123. Berkeley, CA: Apress, 2021.

[23] Bjorck, Nils, Carla P. Gomes, Bart Selman, and Kilian Q. Weinberger. "Understanding batch normalization." *Advances in neural information processing systems* 31 (2018).

[24] Yang, Chunling, Chunchao Zhang, Xuqiang Yang, and Yanbin Li. "Performance study of CBAM attention mechanism in convolutional neural networks at different depths." In *2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1373-1377. IEEE, 2023.

[25] Yu, Yuhai, Hongfei Lin, Jiana Meng, Xiaocong Wei, Hai Guo, and Zhehuan Zhao. "Deep transfer learning for modality classification of medical images." *Information* 8, no. 3 (2017): 91.

[26] Mohammadian, Saboora, Ali Karsaz, and Yaser M. Roshan. "Comparative study of fine-tuning of pre-trained convolutional neural networks for diabetic retinopathy screening." In *2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME)*, pp. 1-6. IEEE, 2017.