Discover Artificial Intelligence

Research

A novel CatractNetDetect deep learning model for effective cataract classification through data fusion of fundus images

Walaa N. Ismail[1] · **Hessah A. Alsalamah**[2,3]

Received: 29 February 2024 / Accepted: 30 July 2024
Published online: 13 August 2024

Abstract
Cataracts are common eye disorders characterized by the clouding of the lens, preventing light from passing through and impairing vision. Various factors, including changes in the lens's hydration or alterations in its proteins, may contribute to their development. Regular eye examinations conducted by an ophthalmologist or optometrist are imperative for detecting cataracts and other ocular conditions early on. Manual checks by caregivers pose several problems, including subjectivity, human error, and a lack of expertise. Biomedical fusion involves combining or linking various characteristics specific to certain diseases from different medical imaging resources. The primary objectives of this approach in disease classification are to reduce the error rate and increase the number of retrieved features. The aim of this study is to evaluate the outcomes associated with fusing visual features related to left and right eye cataract characteristics. Additionally, we investigate the impact of limited variability in deep learning models, specifically in the classification of cataract fundus versus normal fundus images. To address this issue, this study introduces CataractNetDetect, an innovative multi-label deep learning classification system that fuses feature representations from pairs of fundus images (e.g., left and right eyes) for the automatic diagnosis of various ocular disorders. Our focus is on achieving improved performance by stacking discriminative deep feature representations to combine two fundus images into a unified feature representation. Several deep learning architectures are utilized as feature descriptors, including ResNet-50, DenseNet-121, and Inception-V3, enhancing the resilience and quality of representations. Fine-tuning of these DL architectures is conducted using the ImageNet dataset, followed by an integrated stacking approach combining ResNet-50, DenseNet-121, and Inception-V3 models. The model is trained on the publicly available ODIR-5k dataset, which includes 5000 left/right eye images depicting eight different ocular conditions, ranging from healthy states to uncommon ailments such as cataracts, glaucoma, age-related macular degeneration (AMD), diabetes, hypertension, and myopia abnormalities. Moreover, extensive preprocessing of the images is performed, including data augmentation, noise reduction, contrast enhancement, scaling, and circular border cropping. The CataractNetDetect system demonstrates F1-scores, AUC, and maximum validation scores of 98.0%, 97.9%, and 100%, respectively. This ensemble-based model distinguishes itself by surpassing the performance of conventional established methodologies, including ResNet-50, DenseNet-121, and Inception-V3, thereby underscoring its efficacy in diagnostic applications.

Keywords Posterior capsule rupturer · Image fusion · DNN · Bilateral fundus images · Stacking ensemble · DenseNet-121 · ResNet-50 · Inceptionv3 · CAD · Computer aided systems

✉ Walaa N. Ismail, w_abdelfattah@yu.edu.sa; Hessah A. Alsalamah, halsalamah@ksu.edu.sa | [1]Department of Management Information Systems, College of Business, Al Yamamah University, Riyadh 11512, Saudi Arabia. [2]College of Computer and Information Sciences, King Saud University, Riyadh 4545, Saudi Arabia. [3]Computer Engineering Department, College of Engineering and Architecture, Al Yamamah University, Riyadh 11512, Saudi Arabia.

1 Introduction

A cataract is an ocular condition commonly associated with aging, characterized by clouding of the natural lens of the eye [1, 2]. The term "cataract" originated in Latin and refers to a "waterfall" or "portcullis," illustrating the visual metaphor of a cloudy or obstructed lens. A majority of individuals experiencing this condition are in their later stages of life, particularly those over the age of 50 [3]. Additionally, various factors contribute to the development of this disease, including genetic predisposition, exposure to ultraviolet light, diabetes, or the adverse effects of certain medications. As a cataract progresses, it gradually obstructs the transmission of light, resulting in reduced clarity, contrast, and color perception, as well as diminished vision. Untreated cataracts can cause blindness in advanced stages, affecting tasks such as reading, driving, and recognizing faces, among other daily activities [1–3].

Recent advances in surgical techniques have greatly improved cataract surgery, making it a minimally invasive procedure with rapid visual recovery and favorable visual outcomes. The posterior capsule rupture (PCR), which results in a breach of the posterior capsule of the crystalline lens, has been reported to occur in a small percentage of cataract surgeries, between 0.2% and 1.8% [3, 4]. A serious consequence of this complication is the inability to implant an intraocular lens (IOL) successfully, the risk of endophthalmitis, and the occurrence of cystoid macular edema [3, 4]. Surgeons typically evaluate the likelihood of PCR occurring before surgery to minimize the risks associated with this complication. Their expertise guides this evaluation and often involves a scoring system [3, 4].

Computer-aided design (CAD) systems may enhance the accuracy and objectivity of these risk assessments by providing automated and objective means to reduce the incidence of PCR [5–7]. These systems approach may reduce the incidence of PCR, alleviate the limitations associated with manual assessment methods, and benefit both caregivers and healthcare PCR detection by CAD systems [5–7]. A diagnosis of glaucoma is the first step in determining the cause of the retinal disorder, which may aggravate optic nerve damage and cause progressive, irreversible vision loss. Multiple deep learning (DL) architectures have been developed for the classification problem, each differing in terms of its processing philosophy and the outcome of the classification process. To achieve competitive accuracy, specific DL approaches, such as pre-trained convolutional neural networks (CNN) models like AlexNet, VGG-16, and ResNet-50, have been utilized [8–10]. It is important to note, however, that these architectures are deep and complex, resulting in lengthy training and testing processes. According to the study presented by Khan et al. [11], the potential for developing an integrated system using CNN for learning features in glaucoma detection has been explored. CNN automatically fuses images of the left and right eyes into VGG following the removal of specific properties. Khalil et al. [12] proposed a computer-based algorithm for glaucoma recognition consisting of two modules: a hybrid textures feature-set (HTF) and a hybrid structural feature-set (HSF). The HSF module integrates hybrid structural and textual features to distinguish glaucoma from fundus images and OCT images. Gautam [13] recommended using flexible analytical wavelet transform and texture features for glaucoma diagnosis, achieving dimensionality reduction through PCA. Additionally, the SVM classifier is suggested for diagnosing reduced retinal features. As et al. [14] presented an approach to select a subset of ocular features using a KNN classification method on the Glaucoma dataset. Acharya et al. [15] introduced a new approach for glaucoma identification based on Gabor transformations, emphasizing feature elimination and classifier construction. Singh et al. [16] introduced a novel methodology utilizing Grey Wolf Optimization (GWO) and Whale Optimization Algorithm (WOA) for feature selection in glaucoma diagnosis, presenting a hybrid approach of these algorithms. They extracted 65 features from the ORIGA dataset using this method. In [17], a multiclass discriminant analysis approach was utilized to classify cataracts. The authors employed three CNN-based pre-trained models (Inception-V3, MobileNet-V2, and NasNet-Mobile) as base classifiers to generate complementary predictions for four glaucoma classes. To ensure robust cataract grading, fundus image preprocessing and data augmentation were performed, followed by training of the DL architectures. The study utilized a dataset of 590 fundus images sourced from two public databases, where predictions from base classifiers were stacked using an extreme learning machine. In [18], a bioinspired computation-based approach was applied to glaucoma classification using fundus images as input. The framework employed four machine learning models for classification, selecting half of the key features based on scores derived from both the feature significance and univariate approaches to construct a feature vector. While these efforts were successful, they encountered challenges such as inadequacies in feature selection, resulting in suboptimal accuracy and increased computational intensity [19, 20]. One primary challenge is to ensure more accurate classification of images depicting various stages of cataract development. Much of the literature assumes a single-eye occurrence of ocular issues in patients, focusing solely on classification without considering the potential presence of glaucoma [17, 21–23]. Additionally, evaluation procedures often relied on a small number of fundus images with inadequate grading, which could compromise reliability. Furthermore, some

techniques employed single-machine learning for feature selection to identify the best subset from the initial set; however, removing unnecessary features might impact the diagnostic accuracy of the condition. Consequently, this approach may lead to reduced diagnostic accuracy due to limitations in the selected feature set. Additionally, imbalanced class distributions, an abundance of parameters, and limited data all contribute to reduced accuracy when using less efficient architectural designs [1, 24]. This has led to some models performing well on certain tasks but struggling in challenging scenarios, such as identifying fusion of fundus images. It is crucial to develop an effective and comprehensive fundus screening technology that can identify various ocular illnesses simultaneously. Another characteristic of CAD systems is their reliance on training data to formulate and refine algorithms. The image dataset is typically randomly divided into a training dataset and a test dataset, with the model's performance evaluated on the latter [26]. Authors often emphasize significant model performance, highlighting high accuracy, sensitivity, and specificity. Model accuracy is frequently reported to exceed 0.99, with some studies suggesting that these models can outperform experienced medical professionals on specific test sets. However, certain studies may be susceptible to overfitting due to their design. While models may demonstrate superior performance on specific test sets used in research, the limited variability in both training and testing datasets compared to real-world data complicates the analysis. The efficacy of trained models may be compromised when confronted with novel real-world data that depicts the same diagnosis but with distinctive features, due to the lack of datasets comprising authentic fundus photos annotated for various eye conditions [25–27].

When CAD systems encounter unfamiliar or intricate cases, they may struggle to generalize to real-world contexts if the training dataset contains biases or insufficiently represents the diversity of "lens opacity" within a specific locale [24, 25]. Consequently, the accuracy of the classification can decrease when faced with unfamiliar or complex cases. Additionally, most CNN-based models typically perform the classification task by examining fundus images from a single eye, whereas ophthalmologists often use both eyes to diagnose patients. Multimodal image analysis involves examining and extracting information from various types of images to achieve a comprehensive understanding of a specific subject or issue. The analysis of bilateral eye disease progression plays a pivotal role in identifying ocular disease disorders [24, 26]. This approach allows researchers and practitioners to capitalize on the complementary characteristics of bilateral fundus images to enhance the analysis and diagnosis of ophthalmic patients [24, 25]. Furthermore, it facilitates obtaining more accurate and comprehensive results compared to independently analyzing single fundus images. Various data fusion techniques may be employed in this scenario to merge the data obtained from bilateral fundus images, thereby improving the overall performance of the diagnostic system [25–27].

The present article introduces a novel fusion-based deep convolutional neural network classification system (CatractNetDetect) designed for analyzing multi-label PCR patterns. Covers eight different ocular conditions, ranging from healthy states to uncommon diseases such as cataracts, glaucoma, diabetes, hypertension, myopia, and age-related macular degeneration (AMD), using a fusion of images from the left and right fundus. This system not only generates discriminant features from training data but also proficiently categorizes data.

Furthermore, the proposed system is built using CNN-based pre-trained models such as DenseNet-121, ResNet-50, and Inception-v3. These models are meticulously constructed, fine-tuned, and their hyperparameters optimized. To address challenges associated with cataract images, marked by a lack of distinct visual features and a limited number of training examples, we employed an ensemble of these models to enhance robustness and facilitate performance evaluation. This approach enables analysis of various feature representations of input images to assist the classification model in accurately predicting labels associated with diverse diseases, including cataracts, glaucoma, and age-related macular degeneration (AMD). A comprehensive assessment was developed through a series of experiments that yielded insight into the efficacy of these models in handling various scenarios. To mitigate concerns related to overfitting, various data augmentation techniques, including translation, flipping, rotation, scaling, shading, cropping, and multidimensional transformations, were effectively applied. Empirical evidence demonstrates that this strategic augmentation effectively prevents overfitting. In summary, the significant contributions of this paper can be outlined as follows:

1. An effective automated stacking CatractNetDetect classification model is presented for extracting discriminative deep feature representations from bilateral fundus images. This framework comprises three principal components: an efficient lightweight DenseNet-121 and ResNet-50 to extract global features from the left and right fundus images. The model's accuracy is further enhanced through layer freezing and unfreezing tuning, contributing to improved detection of multiple ocular diseases.
2. Utilizing discriminative feature maps developed in the feature extraction phase, Inception-V3 is employed to integrate, refine, and converge these features efficiently, generating the probability distribution of eight distinct eye

disorders. By incorporating the outputs of the hybrid models (a combination of different machine learning models), Inception-V3 acquires new high-level feature representations to distinguish between different lesion parts.
3. For the purpose of reducing noise and enhancing contrast in fundus images, different pre-processing and training techniques were used to enhance classification performance. Preprocessed fundus image data can be used to train deep learning models rather than raw data to enhance the models' ability to learn more important feature representations. Furthermore, this method can reduce the amount of time and effort required for generating an optimally trained model.
4. The performance of the proposed framework was compared with that of single and hybrid techniques on the same dataset and experimental configurations. A number of categorization metrics were used to evaluate its effectiveness, including precision, recall, accuracy, and f1-score.

This paper is organized as follows: The first section provides an overview of cataract disease, emphasizing its significance, prevalence, and current challenges in diagnosis and treatment. Section 2 offers a comprehensive summary of the literature on cataract detection techniques, discussing the limitations of existing methods and identifying the research gap that necessitates an integrated framework. Section 3 introduces the proposed stacking framework, beginning with data collection and preprocessing, and then detailing the DenseNet-121, ResNet-50, and InceptionV3 architectures used for cataract detection. This section also covers the proposed deep ensemble approach, which integrates these architectures into a robust prediction mechanism. Section 4 presents an analysis of three different experiments designed to evaluate CatractNetDetect's generalizability, including a description of the experimental setup, methods, results, and analysis. Finally, Sect. 5 summarizes the main findings, discusses the value of the proposed framework, and makes recommendations for further study.

2 Related work

The advent of Artificial Intelligence (AI) in ophthalmology, particularly in cataract detection and management, signifies a transformative leap in medical technology. These AI applications aim to enhance early detection, refine classification accuracy, and streamline management processes, significantly elevating patient care standards. Yet, alongside these promising advancements are inherent limitations and challenges that necessitate attention.

One pivotal study [28] reported a DenseNet201-based model that enhanced cataract classification by 10.

Similarly, a comprehensive review [29] underscored the potential of AI in managing anterior segment ocular diseases but highlighted the lack of standardized datasets and the necessity for extensive validation. This gap in standardized protocols and benchmarks can hinder the widespread adoption and trust in AI systems within ophthalmology. Another study [30] explored AI's potential in aiding the visually impaired, including those affected by cataracts. While promising, the study highlighted gaps such as ensuring data privacy, integrating AI into existing healthcare systems, and making the technology universally accessible. Additionally, research into assessing cataract surgery skill levels using machine learning [31] achieved high accuracy with an AUC ranging from 93.3.

Other studies [32, 33] developing models for cataract detection reported high accuracy rates but faced limitations related to unseen data performance and computational demands. These challenges point to a prevalent gap in the scalability and real-world application of AI models. The introduction of novel algorithms [34, 35] for cataract detection and the use of convolutional neural networks with digital camera images discussed sensitivity to image quality and the need for comprehensive validation. This indicates a limitation in models' ability to perform consistently across varied quality inputs, a crucial factor for reliability in clinical practice. These challenges, along with the need for extensive data for training as noted in several studies [36, 37], highlight common gaps in AI research related to data diversity, representativeness, and the computational demands of sophisticated models.

Furthermore, the challenges of acquiring large, annotated, diverse datasets and ensuring model transparency as highlighted in research [11, 38, 39] are essential for building trust among clinicians and patients. Additional studies [40–42] underscored the importance of real-world testing, continuous model updates, and addressing ethical considerations, pointing to gaps in regulatory frameworks, continuous learning mechanisms for AI systems, and ethical guidelines. These are vital for the responsible and beneficial implementation of AI in healthcare.

In summary, while AI and deep learning have shown considerable promise in improving the detection, classification, and management of cataracts, there are notable gaps and challenges. These include:

1. As the number of classes rises, the model performs inconsistently, especially when there are not enough training samples and inevitable picture noise.
2. Some methods are too conservative to be used in practical settings due to data diversity, model generalizability, computational costs, and incomplete datasets.
3. The majority of CNN-based research on categorizing ocular disorders frequently trains their models using unprocessed single type of fundus images, which may limit how well the adopted CNN model can generalize.

To overcome the limitations found in existing literature and accurately identify various eye disorders using colored fundus photos, we have devised a deep learning system. This system utilizes a stacking ensemble model combining ResNet 50, DenseNet-121, and Inception-V3. The objective of this approach is to improve diagnostic efficiency in distinguishing between healthy individuals and those with ophthalmic conditions.

3 Integrated stacking framework

Developing our proposed model has the fundamental goal of automatically distinguishing between individuals with eye disorders and healthy individuals while minimizing the time it takes for the detection to be completed. The present procedures will also be enhanced in terms of their effectiveness. Here, we outline our proposed technique and methodology for our proposed CatractNetDetect model (Fig. 1).

CatractNetDetect is an integrated stacking of deep convolutional neural networks (DCNNs) capable of detecting normal and different ocular disorders based on bilateral fundus images. Integrated stacking in neural networks

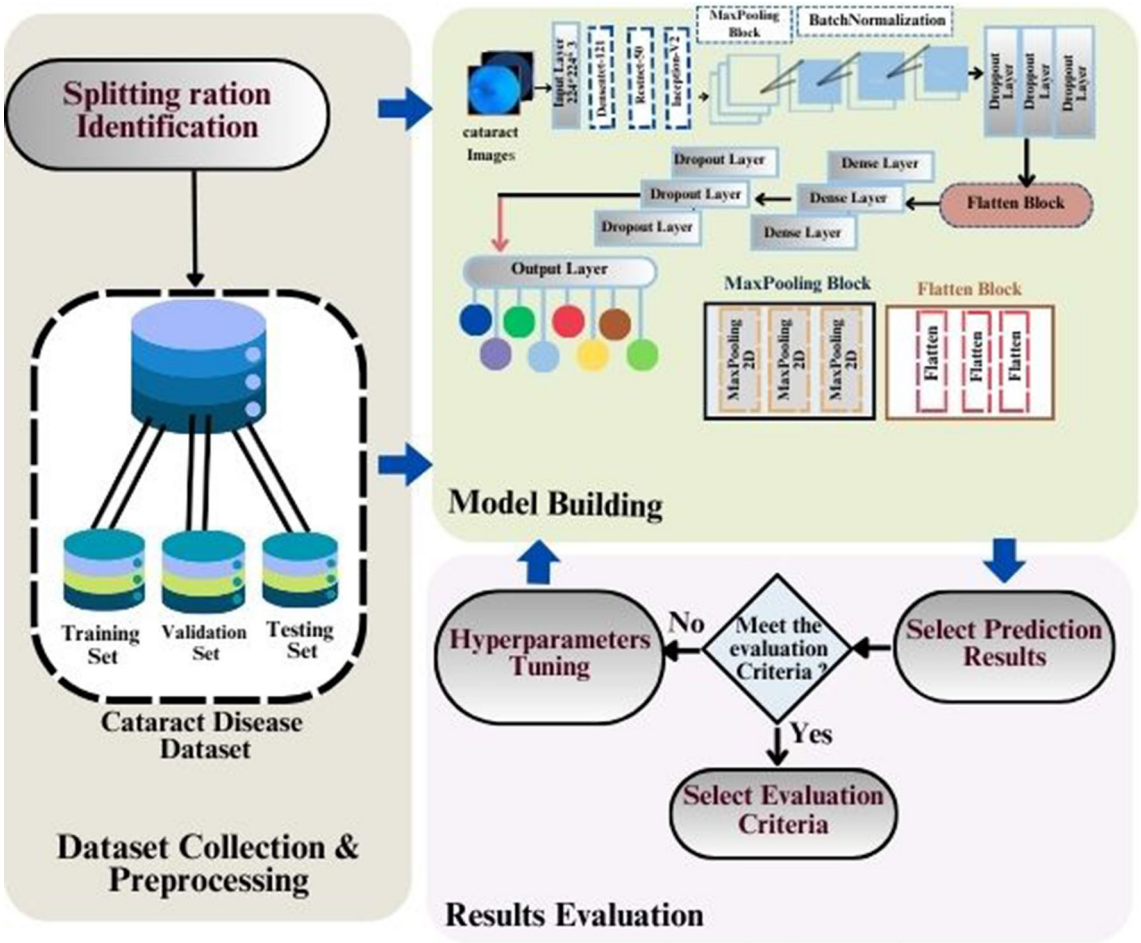


Fig. 1 Proposed CatractNetDetect framework

involves using neural networks as sub-models to categorize a dataset and then using these predictions as features for an estimator neural network. Estimator-learning determines which projections from each input submodel should be connected. In this study, three pre-trained DCNNs were used for fine-tuning: DenseNet [43], Inception-v3 [44], and ResNet-50 [45].

A rigorous experiment was conducted to select the above-mentioned pre-trained models, which concluded that each significantly improved classification performance as a result of the unique feature extraction techniques used by each of these models, which are described in detail below. Using the pre-trained weight matrices of these models, we first loaded them into ImageNet. Afterward, the models were fine-tuned to fit our dataset. Following fine-tuning, the models were stacked using the integrated stacking technique, resulting in a model that is larger and more robust. An advantage of the stacked model is the immediate provision of projections of the sub-models to the estimator. A further benefit of the estimator model is the ability to fine-tune the weights of the sub-models when averaging the results.

The study was divided into the following phases:

1. Data collection and preprocessing: During preprocessing, missing and incorrect values were removed from cases. This is followed by the splitting of the data. Data sets were randomly divided into three sections: training sets, verification sets, and test sets.
2. Feature extraction: Models are pre-trained on a standard dataset, such as ImageNet. This is followed by the elimination of the model's classification component. Afterward, the rest of the network can be viewed as an extractor of features, which is capable of running any classification algorithm.
3. Model building: First, images are processed through the input layer of the pre-trained models. A series of models is then applied to the images, which are then processed accordingly. A final convolution layer is added to each pre-trained model and fine-tuned once more. This produces predictions that are then combined via a stacking layer.
4. Fine-tuning: Using this method, we can both freeze and train the classifier layers. In addition, modify the parameters of the pre-trained model by repeating the training process on all layers.

In this section, a detailed description of the data set utilized, as well as a detailed explanation of the proposed methodology, are provided.

3.1 Dataset collection

In this context, the dataset [46] is used to prepare image files and integrate data to fulfill our research objectives. Generally, each patient in the dataset is diagnosed for both the right and left eyes, potentially indicating multiple diseases per patient. Some diseases may affect one eye differently from the other. Therefore, specific diseases are identified and represented by the number (1), while unaffected eyes are denoted by the number (0) .

In the present study, we utilized fundus images from the ODIR-5K dataset [46], which comprises approximately 5000 patients and includes attribute information such as color fundus images of both eyes, patient ages, and specialist keywords. This dataset contains multiple images of individual cataracts, captured at different times and under varying lighting conditions, contributing to its variability. The dataset was compiled by Shanggong Medical Technology Co., Ltd. in China, using cameras from Zeiss, Canon, and Kowa. Patient identification information is intentionally excluded and replaced with labels overseen by quality control management. Each patient is categorized into one of eight distinct groups: normal cases, diabetes, glaucoma, cataracts, AMD, hypertension, myopia, and other diseases/abnormalities. This categorization is based on examinations of both color fundus photographs (CFPs) and supplementary clinical features conducted by trained human readers. For a detailed description of the dataset, please refer to Table 1.

Figure 2 provides a visual example of cataract infection. After analyzing the collected dataset, we confidently observe that men tend to have a higher incidence of cataract disease compared to women. This trend begins around the age of 33 and continues until approximately 60 years of age. This observation underscores the significance of cataract disease in younger women, as it typically affects older individuals but can significantly impact the visual health of those who develop it prematurely. Early intervention and treatment are crucial in managing this condition (Fig. 3).

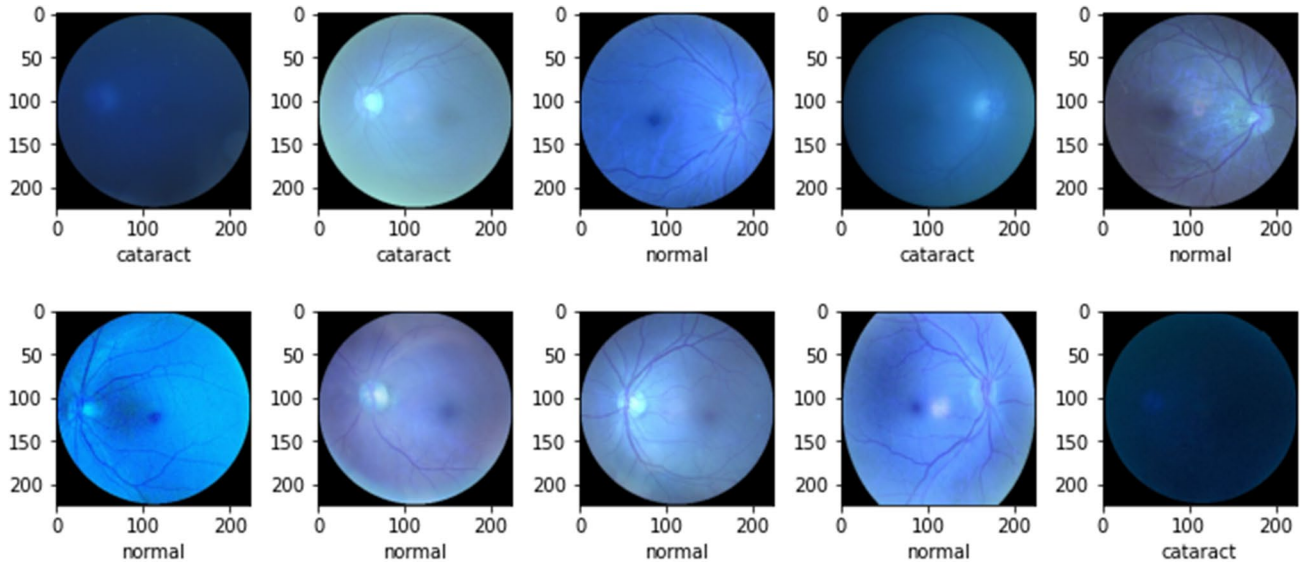


Fig. 2 An example of cataract infection classes

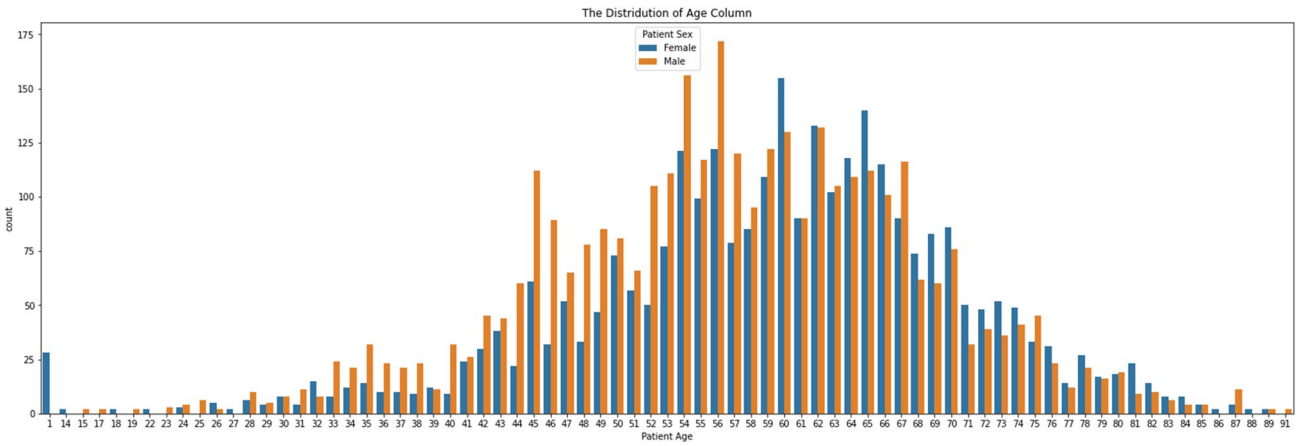


Fig. 3 Gender-specific cataract infection rates across age groups

Table 1 Class distribution of ODIR-5K dataset

Categories	Total cases
Normal	1146
Glaucoma	215
AMD	164
Cataract	594
Normal	1140
Diabetes	1128
Hypertension	103
Myopia	174

3.2 Dataset preprocessing

Deep learning models are noted for their ability to generalize and perform well based on their training-to-test ratios, also known as train-test split ratios. The correlation between overfitting and underfitting is influenced by this ratio;

overfitting occurs when a model captures too much noise and intricacies in the training data, hindering its ability to generalize to new data. The scarcity of adequate data for training medical imaging algorithms poses a significant challenge in deep learning development. To address these limitations, we apply three geometric transformations to the training samples, including rescaling, random-angle image rotations, zooming, and horizontal flipping, across three different training-to-test ratios.

This paper underscores the importance of a three-way data split (training, validation, and test) as a critical methodology in machine learning model development. These distinct augmentations enhance model performance, increase robustness, and provide a clearer understanding of the model's real-world capabilities. By segregating the training data, the model is prevented from memorizing the data and instead learns meaningful patterns. The validation set aids in selecting optimal hyperparameters to improve model accuracy and reliability. Finally, the test set serves as the benchmark to determine the model's readiness for deployment.

3.3 Cataracts disease identification

In this section, we conduct a detailed exploration of the hybrid deep learning architectures implemented in CataractNetDetect. Specifically, we delve into the traditional ResNet-50, DenseNet-121, and Inception-V3 models, alongside our proprietary ensemble architecture algorithm.

3.3.1 DenseNet-121 architecture

DenseNet-121 [43] belongs to the DenseNet family and is characterized by its 121 layers, including 117 convolutional layers, three transition layers, and one classification layer. The architecture features four dense blocks, each separated by transition layers. Within each dense block, a 1×1 convolutional layer is followed by a 3×3 convolutional layer. Outputs from each dense block are concatenated with those of the preceding block before being passed on. After reducing channel numbers with 1×1 convolutions, feature maps are downsampled using 2×2 average pooling layers. In our study, the last fully connected layer with softmax activation in DenseNet-121 was replaced with a custom classifier. Figure 4 illustrates the DenseNet-121 architecture, while Table 2 provides a detailed description of the fine-tuned DenseNet-121 utilized. For the diagnosis of cataract disease, our adaptation of DenseNet-121 operates with input dimensions of (224, 224, 3). After DenseNet-121 processes the input, the output shape transitions from (7, 7, 1024) to (3, 3, 1024) through MaxPooling2D. Dropout layers and batch normalization are employed to enhance stability and mitigate overfitting. Following flattening, dense layers with 512 neurons each are incorporated, followed by additional dropout layers. This configuration contributes to robust learning capabilities for classifying cataract conditions, utilizing a total of 11,675,521 trainable parameters.

3.3.2 RestNet-50 architecture

ResNet-50 [45] is constructed from several key building blocks. This architecture consists primarily of residual blocks, which introduce skip connections to mitigate the vanishing gradient problem and facilitate the training of very deep neural networks. In order to extract features efficiently, bottleneck layers consisting of 1×1, 3×3, and 1×1 convolutions are used within these residual blocks. An identity block is used when the dimensions of the input and output are identical, whereas a projection block is used when the dimensions of the input and output are different. A

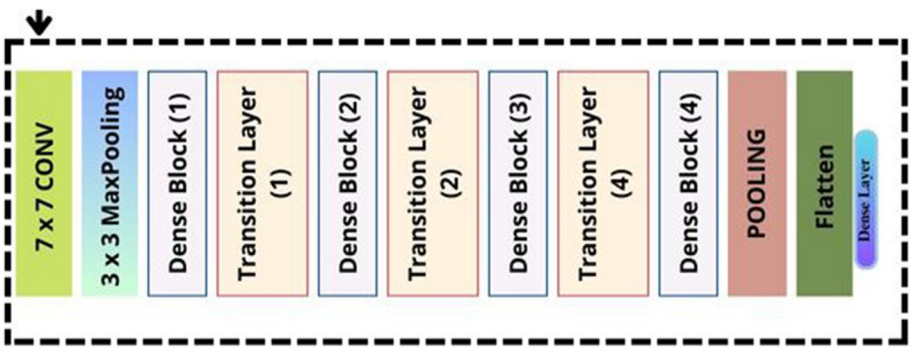


Fig. 4 DenseNet-121 architecture

Table 2 DenseNet architecture designed for cataract disease identification

Layer (type)	Output shape	Param #
input_2 (InputLayer)	(None, 224, 224, 3)	0
densenet121 (Functional)	(None, 7, 7, 1024)	7,037,504
max_pooling2d (MaxPooling2D)	(None, 3, 3, 1024)	0
batch_normalization (BatchNormalization)	(None, 3, 3, 1024)	4,096
dropout (Dropout)	(None, 3, 3, 1024)	0
flatten (Flatten)	(None, 9216)	0
dense (Dense)	(None, 512)	4,719,104
dropout_1 (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 1)	513
Total params	11,761,217	
Trainable params	11,675,521	
Non-trainable params	85,696	

max-pooling layer downsamples feature maps, and a final fully connected layer performs classifications. ResNet-50 is able to capture intricate features through these building blocks, which are fundamental to its ability to classify images and perform other computer-aided tasks. Figure 5 shows the RestNet-50 architecture, with the model's configuration used in this study illustrated in Table 3. ResNet-50 analyzes input images of size (224, 224, 3) and reduces dimensions to (7, 7, 2048). The result is further downsampled by MaxPooling2D layers. As part of the model, residual connections (Add) are used to enhance learning, as well as dense layers of 512 neurons each. With a total of 44,794,817 parameters, the model exhibits high performance in disease classification.

3.4 InceptionV3 architecture

The InceptionV3 [44] is a well-known convolutional neural network (CNN) architecture developed by Google, primarily designed for image classification. InceptionV3 incorporates a distinctive inception module that utilizes parallel convolutional filters of varying widths within the same layer (shown in Fig. 6). This method enables the model to discern complex patterns in photos by capturing characteristics of different dimensions and complexities. The architecture involves the use of convolutional layers for initial feature extraction, followed by inception modules to capture diverse feature sets, and ultimately, fully connected layers to amalgamate high-level information for the final classification. The output layer is a softmax layer, responsible for generating a probability distribution across classes and determining the model's ultimate prediction.

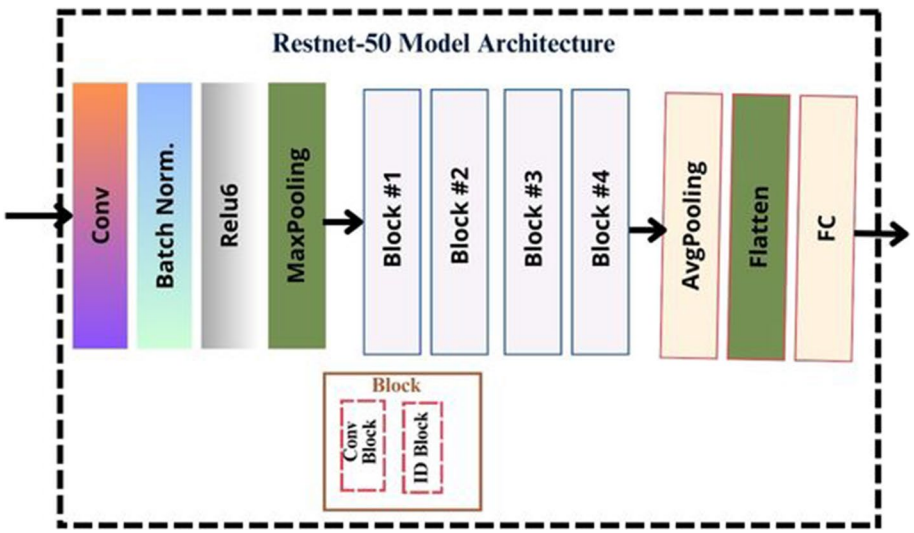


Fig. 5 RestNet-50 architecture

Table 3 ResNet-50 architecture designed for cataract disease identification

Layer (type)	Output shape	Param #
input_4 (InputLayer)	[(None, 224, 224, 3)]	0
resnet50 (Functional)	(None, 7, 7, 2048)	23,587,712
max_pooling2d_1 (MaxPooling2D)	(None, 3, 3, 1024)	0
max_pooling2d_2 (MaxPooling2D)	(None, 3, 3, 2048)	0
batch_normalization_1 (BatchNormalization)	(None, 3, 3, 1024)	4,096
batch_normalization_2 (BatchNormalization)	(None, 3, 3, 2048)	8,192
dropout_2 (Dropout)	(None, 3, 3, 1024)	0
dropout_4 (Dropout)	(None, 3, 3, 2048)	0
flatten_1 (Flatten)	(None, 9216)	0
flatten_2 (Flatten)	(None, 18432)	0
dense_2 (Dense)	(None, 512)	4,719,104
dense_3 (Dense)	(None, 512)	9,437,696
dropout_3 (Dropout)	(None, 512)	0
dropout_5 (Dropout)	(None, 512)	0
add (Add)	(None, 512)	0
dense_4 (Dense)	(None, 1)	513
Total params	44,794,817	
Trainable params	21,117,313	
Non-trainable params	23,677,504	

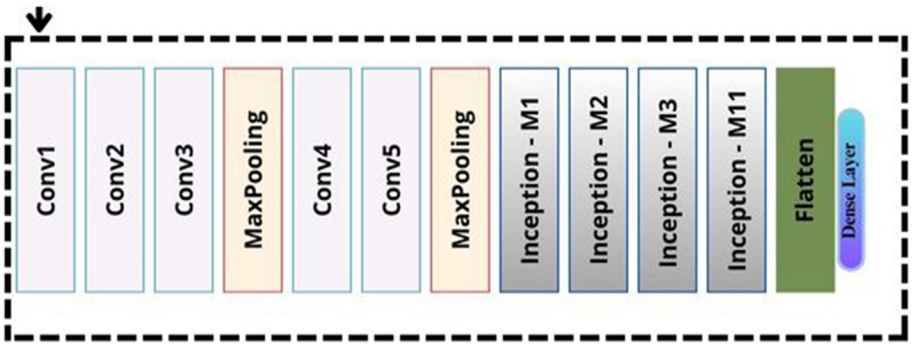


Fig. 6 Inception-V3 architecture

3.5 Proposed deep ensemble-based model for cataract disease identification

In this approach, a CNN-based ensemble model is developed to improve overall performance, robustness, and generalization of cataract detection. Using three advanced convolutional neural networks (CNNs) such as DenseNet-121, ResNet-50, and InceptionV3, the predictions of both methodologies are combined in the system, and the class with the majority of votes is deemed the final prediction.

DenseNet-121 and ResNet-50 are known for their ability to classify images. These convolutional neural network-based architectures stand out as a classification method with the capability to adeptly categorize a multitude of images spanning over a thousand distinct classes. They enjoy significant popularity for image classification tasks and lend themselves to facile implementation, especially when leveraged in conjunction with transfer learning. However, the base DenseNet-121 and ResNet-50 models exhibit limitations, manifesting in suboptimal accuracy and requiring protracted training periods. Furthermore, DenseNet-121 suffers from the challenge of vanishing gradients, mostly related to the complexity of the million-parameter architecture.

As a first step, we loaded these models into ImageNet using their pre-trained weight matrix for multi-class classification of objects. Secondly, each training sample in the ImageNet database is associated with a unique image. In line with the proposed CataractNetDetect method, both left and right eye fundus images serve as inputs for

pre-trained Convolutional Neural Networks (CNNs). To facilitate this, a stacking ensemble is employed as an image generator for concatenated inputs from DenseNet-121 and ResNet-50 models.

The two models, namely DenseNet-121 and ResNet-50, process input left and right images independently, capturing distinctive features particular to their architectures. The ensemble model leverages these features and combines them into a unified feature vector to create a comprehensive representation of the input image with the Inception-V3 model. As a result of the combination of features, often by averaging or concatenating, a broader spectrum of image characteristics is encapsulated, enhancing the ability of the model to detect cataract-related changes in the image.

To create a comprehensive representation of the input image, the input image is resized to dimensions smaller than the filter size within the convolution layer. In this layer, it is recommended to use M x M-sized filters. The primary objective of this layer is to generate feature maps. The convolutional layer is initialized as per Eq. (1):

$$W_{ij} = \sum_{m=1}^{M} \sum_{n=1}^{N} I_{i-m,j-n} \cdot K_{mn}, \tag{1}$$

W_{ij} represents the value at position (i,j) in the feature map. K_{mn} represents the value at position (m,n) in the convolution kernel. M and N denote the dimensions of the convolution kernel. The ensemble model combines these features into a unified feature vector as follows:

$$y_{(i,j,k)} = \sigma\left(\sum_{u=1}^{U} \sum_{v=1}^{V} \sum_{c=1}^{C} W_{(u,v,c,k)} X_{((i+u-1),(j+v-1,c))}\right). \tag{2}$$

According to Eq. (2), the activation values $y_{[}(i, j, k)]$ represent the activation value of each feature map in the kth feature set. A σ activation function adds weighted input values to each other, where $W_{[}(u, v, c, k)]X_{[}((i + 1), (j + v - 1, c)]$ represents the input channel value at position (i, j) and $W_{[}(u, v, c, k)]$ represents the weight value for the kth feature map at position (u, v) of the cth input channel. A filter's width, height, and channel count are specified by the parameters U, V, and C. The challenge arises in selecting patient images while ensuring a well-distributed representation across classes. Therefore, the subsequent step involves fine-tuning these convolutional neural network-based architectures for multi-label classification on the ODIR fundus database. Optimizing a pre-trained model involves the iterative process of enhancing the model's capabilities using the insights and representations acquired from an existing dataset. Unlike training a model entirely from the beginning, fine-tuning enables the transfer of knowledge from the pre-existing model to the specific task at hand, leading to enhanced performance and efficiency. The fine-tuned models were then stacked together using the Integrated Stacking technique, providing a larger and more robust model. Throughout the training process, the parameters of this model are meticulously adjusted and fine-tuned to achieve optimal performance. The pre-trained models are used as inputs to a larger stacked model once all the preceding models have been fine-tuned. The typical integrated stacking process identifies all sub-model layers as untrainable. Testing has shown, however, that training the final convolution layers of all sub-models is advantageous for classification. All submodel layers, except for the final convolution layer, have been set as non-trainable.

This fine-tuning ensures that the ensemble model becomes proficient at recognizing cataract disease by utilizing a wide range of features extracted from the input image. This highlights its effectiveness in accurately classifying intricate and complex images, particularly those related to cataract disease. Furthermore, robustness is enhanced through majority voting among the individual models, proving particularly effective when one of the individual models underperforms on a given input. The class with the highest cumulative votes from the individual models in the ensemble is selected as the final prediction, as determined by the following equation:

$$\hat{y} = \text{argmax}_k\left(\sum_{i=1}^{n} \delta(y_i, k)\right), \tag{3}$$

where n =2 (individual assigned models), $\hat{y}$ represents the final ensemble prediction for class k. $\delta(y_i, k)$ is an indicator function that returns 1 if $y_i = k$, and 0 otherwise. Individual models are trained independently using the same dataset. The output layer of each model produces its own set of class probabilities by using the softmax activation as described in the following equation:

$$P(y_i) = \frac{e^{(w_i \cdot x)}}{\sum_{j=1}^{N} e^{(w_j \cdot x)}}, \tag{4}$$

where w_i represents the weight vector for class i and the input vector x.

Adam optimizer is one of the most popular optimization algorithms used for deep neural network training. To update it, the following guidelines are used:

$$\begin{aligned} m_t &= \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \\ v_t &= \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \\ \theta_t &= \theta_{t-1} - \frac{\alpha \cdot m_t}{\sqrt{v_t} + \epsilon}, \end{aligned} \tag{5}$$

where θ represents the model parameters at time t. m_t, and v_t are the first and second moments of the gradients. α is the learning rate, β_1 and β_2 are exponential decay rates, and ϵ is a small constant for numerical stability.

The final architecture represented in Table 4 with a total number of parameters involved in this ensemble model is substantial, amounting to 70,800,609 in total. Among these parameters, there are 48,850,817 that are specifically adjusted during training to enhance the model's ability to identify cataract disease accurately (Fig. 7). Additionally, there are 21,949,792 non-trainable parameters that contribute to the model's overall complexity. This indicates the significance of the entire system, which is needed to handle the complexity of image classification tasks, particularly in terms of identifying cataract disease.

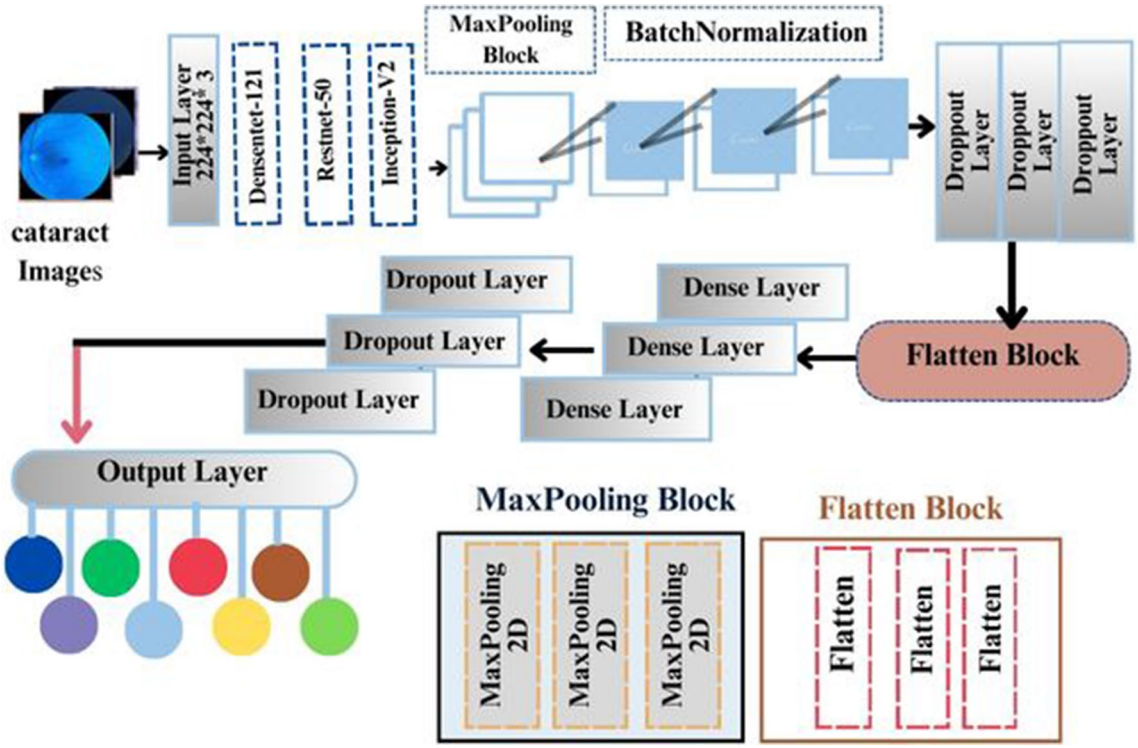


Fig. 7 Proposed CatractNetDetect model

Table 4 Ensemble model architecture summary

Layer (type)	Output shape	Param #	Connected to
input_26 (InputLayer)	[(None, 224, 224, 3)]	0	
densenet121 (Functional)	(None, 7, 7, 1024)	7037504	input_26[0][0]
resnet50 (Functional)	(None, 7, 7, 2048)	23587712	input_26[0][0]
inception_v3 (Functional)	(None, 5, 5, 2048)	21802784	input_26[0][0]
max_pooling2d_67 (MaxPooling2D)	(None, 3, 3, 1024)	0	densenet121[2][0]
max_pooling2d_68 (MaxPooling2D)	(None, 3, 3, 2048)	0	resnet50[1][0]
max_pooling2d_69 (MaxPooling2D)	(None, 2, 2, 2048)	0	inception_v3[0][0]
batch_normalization_211 (BatchNormalization)	(None, 3, 3, 1024)	4096	max_pooling2d_67[0][0]
batch_normalization_212 (BatchNormalization)	(None, 3, 3, 2048)	8192	max_pooling2d_68[0][0]
batch_normalization_213 (BatchNormalization)	(None, 2, 2, 2048)	8192	max_pooling2d_69[0][0]
dropout_64 (Dropout)	(None, 3, 3, 1024)	0	batch_normalization_211[0][0]
dropout_66 (Dropout)	(None, 3, 3, 2048)	0	batch_normalization_212[0][0]
dropout_68 (Dropout)	(None, 2, 2, 2048)	0	batch_normalization_213[0][0]
flatten_29 (Flatten)	(None, 9216)	0	dropout_64[0][0]
flatten_30 (Flatten)	(None, 18432)	0	dropout_66[0][0]
flatten_31 (Flatten)	(None, 8192)	0	dropout_68[0][0]
dense_62 (Dense)	(None, 512)	4719104	flatten_29[0][0]
dense_63 (Dense)	(None, 512)	9437696	flatten_30[0][0]
dense_64 (Dense)	(None, 512)	4194816	flatten_31[0][0]
dropout_65 (Dropout)	(None, 512)	0	dense_62[0][0]
dropout_67 (Dropout)	(None, 512)	0	dense_63[0][0]
dropout_69 (Dropout)	(None, 512)	0	dense_64[0][0]
add_9 (Add)	(None, 512)	0	dropout_65[0][0] dropout_67[0][0] dropout_69[0][0]
dense_65 (Dense)	(None, 1)	513	add_9[0][0]
Total params	70,800,609		
Trainable params	48,850,817		
Non-trainable params	21,949,792		

Algorithm 1 The suggested CataractDetectNet framework pseudocode

Input: ODIR-5K Dataset, {L1, L2, L3} set of learners, Objective Function.
Output: Ensemble CataractDetectNet, Performance Evaluation Metrics.

1: Read ophthalmic disease fundus images in batches of 16 batch of size (224, 224, 3).
2: Determine the data split percentages.
3: Create Dtest: Test Dataset, Dtrain: Train Dataset, and Dvalid: Validation Dataset.
4: Objective function = $Accuracy$.
5: L1 = Restnet-50.
6: L2 = Densenet-121.
7: **for** Tt in range(train_size, Dtrain) **do**
8: Pass image data after preprocessing to L1 model.
9: Pass image data after preprocessing to L2 model.
10: **end for**
11: Stack the predictions made from L1 and L2 with L3 to obtain the feature matrix.
12: Create solution matrix $SL_0 \leftarrow$ features (L1, L2, L3).
13: Pass SL_0 to a dense layer of 512 nodes.
14: Use feature vector W_{ij} to predict feature [i] as $\sum_{m=1}^{M}\sum_{n=1}^{N} I_{i-m,j-n} \cdot K_{mn}$, activation function: $\hat{y} = \text{argmax}_k \left(\sum_{i=1}^{n} \delta(y_i, k)\right)$, with $P(y_i) = \frac{e^{(w_i \cdot x)}}{\sum_{j=1}^{N} e^{(w_j \cdot x)}}$.
15: Validate $L3$ using (SL, DValid).
16: Calculate performance metrics based on DTest.
17: Classify Cataracts features. **return** CataractNetDetect architecture and Performance Metrics.

3.6 Experimental setting

During the experimentation phase, the Colab environment was utilized, employing the Google Compute Engine backend (GPU) in conjunction with Tensorflow using Python 3. The system configuration encompassed a total of 12.7 GB of RAM, of which 2.7 GB were available, a GPU RAM capacity of 15.0 GB, with 0.4 GB currently in use, and a disk storage of 78.2 GB, with 23.6 GB currently allocated. In this study, data were collected and monitored under controlled conditions. The data was analyzed using three different methods to conduct the experiment Afterward, the experiment data were analyzed and interpreted. First, the training, testing, and validation sets comprise (50–10–40)% of the collected data, so divide it into training, testing, and validation sets, then train a model on the training set and verify its accuracy on the validation set. The same split will be repeated two times using (30–10–60)%, and (70–10–20)%.

3.7 Hyperparameters used

In Sect. 3, the Adam optimizer was employed to train and optimize the utilized DNN-based models. A strategy was developed to adjust the learning rates of the models as the validation loss increased.

To dynamically adjust the learning rate during training, the ReduceLROnPlateau callback was utilized. This callback automatically reduces the learning rate upon detecting a plateau in validation accuracy, which helps improve model performance and mitigate overfitting. The learning rate is reduced by a factor of 0.5 (i.e., multiplied by 0.5) upon plateau detection, with a minimum learning rate set to 0.00001. A batch size of 32 was also used.

For fine-tuning the pre-trained models, the final five layers of each model were kept trainable while the additional layers were frozen. ResNet-50 trained a total of 44,794,817 parameters, with 21,117,313 parameters being trainable. Densenet-121 trained a total of 11,761,217 parameters, with 11,675,521 parameters being trainable.

3.8 Performance parameters and evaluation metrics

The proposed CataractNetDetect was evaluated based on several criteria, including accuracy, F1-score, sensitivity, and specificity. Accuracy in a seven-class classification problem measures the fraction of samples correctly classified out of the total number of samples across all classes (Eqs. (6)–(9)). True positives (TP) for each class indicate the number of samples

correctly classified as belonging to that category, while true negatives (TN) indicate the number of samples correctly classified as not belonging to a particular category.

A confusion matrix (CM) was employed to evaluate the performance of CataractNetDetect. These matrices provide insights into the performance characteristics associated with classification during the method's execution.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad . \tag{6}$$

$$Sensitivity\ (Recall) = \frac{TP}{(TP + FN)} \quad . \tag{7}$$

$$Specificity(Precision) = \frac{TP}{(TP + FP)} \quad . \tag{8}$$

$$F1\_score = \frac{TP}{(TP + 0.5(FP + FN))}. \tag{9}$$

4 Experimental results and disscusion

In this section, we aim to evaluate CatractNetDetect's generalization ability across different data divisions. The subsequent subsections will concisely detail the experimental setup and results. The primary objective of this investigation is to assess the outcomes achieved through feature fusion. We will begin by constructing a single model, proceed to develop a hybrid model, and conclude the analysis with an ensemble-based model.

4.1 Experiment A: assessment of multi-label feature fusion using the first division (50%, 10%, 10%)

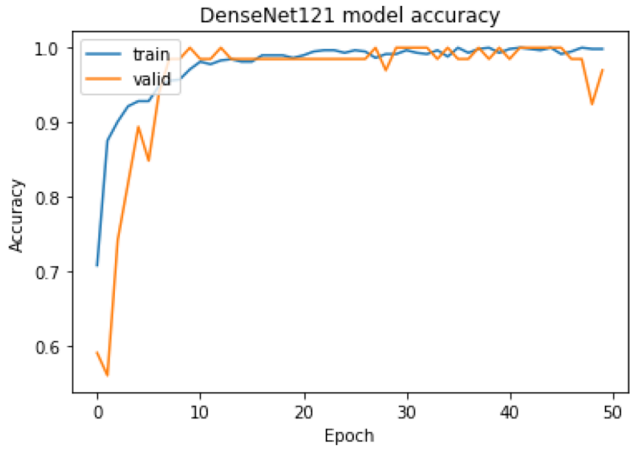
In the first investigation, we assessed the performance of the developed model using the following information: training (50%), validation (10%), and testing (40%). Table 5 illustrates the performance of the CataractNetDete architecture. During the testing and validation stages, DenseNet-121 achieved 94.0% and 97.0% accuracy, respectively, and had a high training accuracy of 99.6%. The Hybrid model achieved 100% validation accuracy and 92.2% testing accuracy but had a somewhat lower training accuracy (96.9%). The ensemble model demonstrated 100% validation accuracy, 92.0% testing accuracy, and 94.4% training accuracy. Considering all of these measures, DenseNet-121 is the model that exhibits the highest accuracy and the best balance between sensitivity and specificity. A 50-epoch training was utilized for the ensemble model, resulting in 100% validation accuracy. Based on the test data, as depicted in Fig. 8c, The model initially achieved an accuracy level of 95%, demonstrating subsequent progress after 100 epochs.

Additionally, the model demonstrated sensitivity of 88.1% and a recall of 96.2%. According to Fig. 9c, CataractNetDete achieved a loss level of 2% after 100 epochs of training.

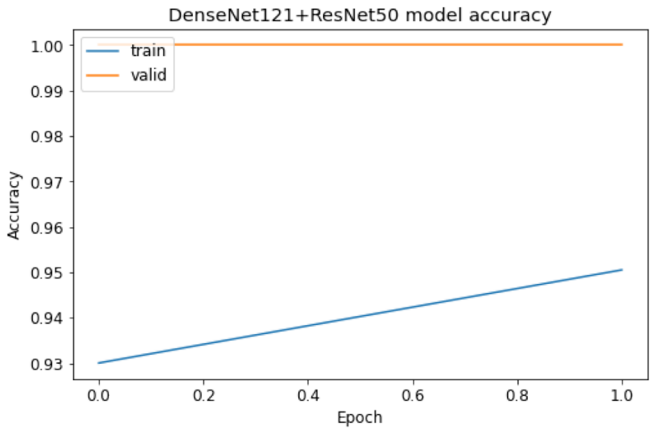
Following the same data division, the model's accuracy, loss, precision, and recall were compared with those of two other models: a single model (DenseNet-121) and a hybrid model utilizing both DenseNet-121 and ResNet-50. The comparison results reveal that the hybrid model outperformed (Figs. 8b and 9b) the single model (DenseNet-121) in terms of validation accuracy (Fig. 8a) and loss (Fig. 9a). This underscores the effectiveness of the hybrid model in predicting the outcomes of interest. The findings emphasize the significance of incorporating diverse models in the development

Table 5 Performance metrics for multi-label classification with first data division

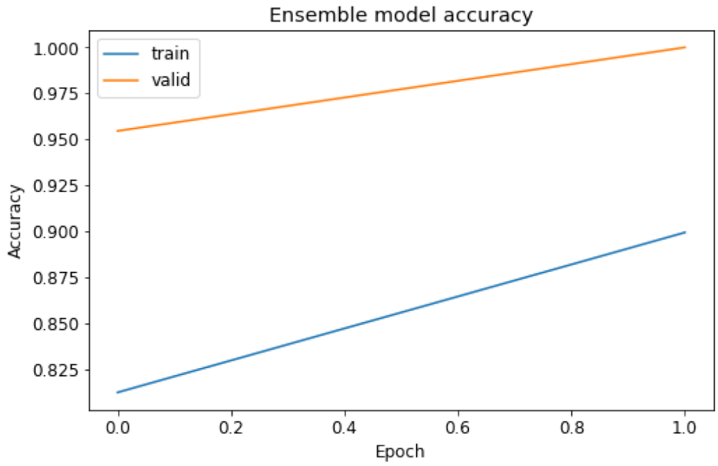
Model	Training accuracy (%)	Val accuracy (%)	Testing accuracy (%)	F1_score (%)	Sensitivity (%)	Specificity (%)
Densenet-121	99.6	97.0	94.0	93.6	96.4	91.3
Hybrid	96.9	100	92.2	92.0	89.5	94.9
Ensemble	94.4	100.0	92.0	92.0	88.1	96.2



(a) Densenet121 accuracy



(b) Densenet121 and Restnet-50 accuracy.



(c) Ensemble Densenet121, Restnet-50 and InceptionV3 accuracy.

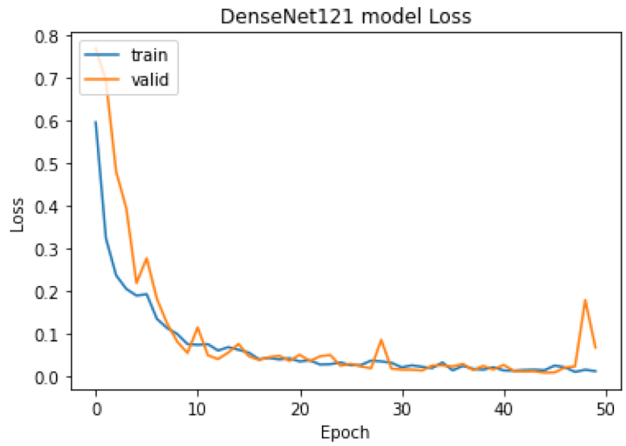
Fig. 8 Multi-label classification accuracy in data division #1: singular, hybrid, and ensemble-based models

of AI-based applications. Compared to a single model, hybrid models can provide a more extensive set of features and achieve higher accuracy. Moreover, hybrid models exhibit reduced susceptibility to overfitting and demonstrate improved adaptability to new datasets.

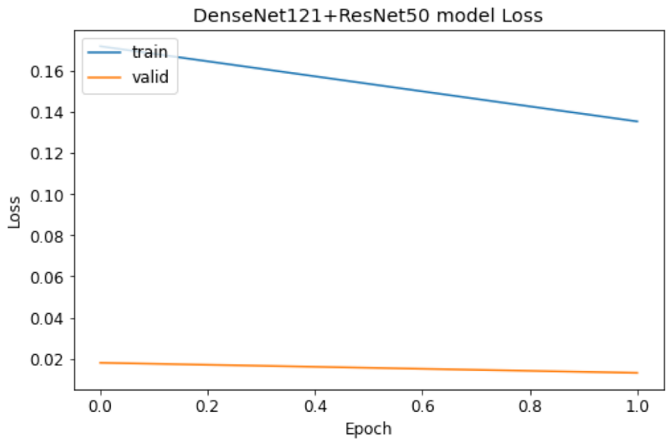
When evaluating the performance of a machine learning model, ROC curves are frequently employed to quantify accuracy and performance. In Fig. 10, the ROC curves for the single, hybrid, and ensemble models are depicted. In terms of adaptability to new data, a hybrid model often demonstrates superior performance, making it more reliable compared to other methods.

4.2 Experiment B: assessment of multi-label feature fusion using the second division (60%, 10%, 30%)

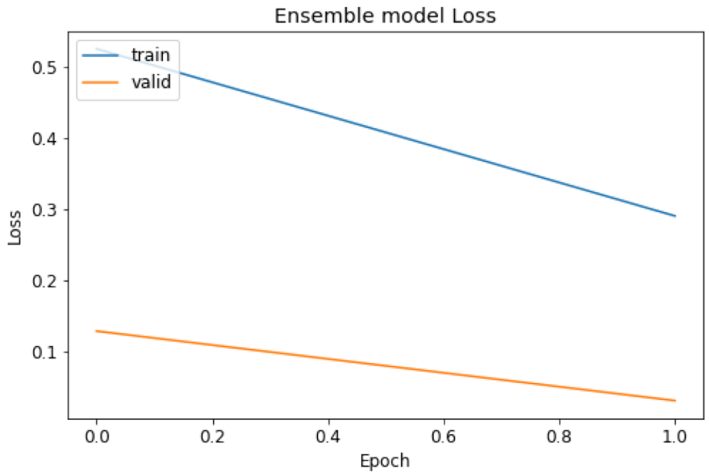
To enhance the performance of the devised model, it becomes imperative to retrain it using a larger training dataset. Additionally, adjusting hyperparameters provides a mechanism for fine-tuning the model and optimizing its overall effectiveness. Consequently, the devised model can undergo upgrades over time through the addition of new features or the modification of existing ones. As a result, the CataractNetDetect exhibits improved accuracy and enhanced performance, rendering it well-suited for real-world applications dealing with dynamically changing data.



(a) Densenet121.



(b) Densenet121 and Restnet50.



(c) Ensemble Densenet121, Restnet-50, and InceptionV3.

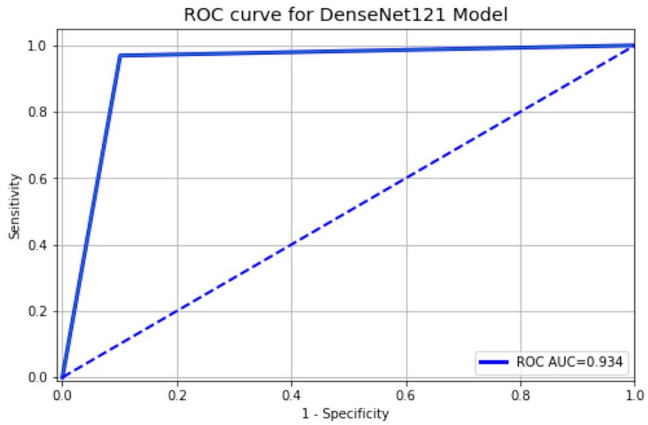
Fig. 9 Multi-label classification loss in data division #1 for singular, hybrid, and ensemble-based models

In Figs. 11 and 12, the training and test results are depicted after 50 training iterations. On average, the training and validation accuracy is the same for single, hybrid and ensemble-based catractNetDetect accuracy becomes more similar, both training and validating accuracies approached 97%. A comparison of the performance metrics of the three models, DenseNet-121, Hybrid, and Ensemble, is presented in Table 6 for the second division. In the validation tests, DenseNet-121, Hybrid, and Ensemble demonstrated 97.7% accuracy, with Ensemble showing 97.7% accuracy, while all models achieved flawless training accuracy of 100%. Testing accuracy was achieved by DenseNet-121 at 96.0%, Hybrid at 95.6%, and Ensemble at 95.3%. The F1 scores for DenseNet-121, Hybrid, and Ensemble were consistently high, at 96.0% and 95.0%.

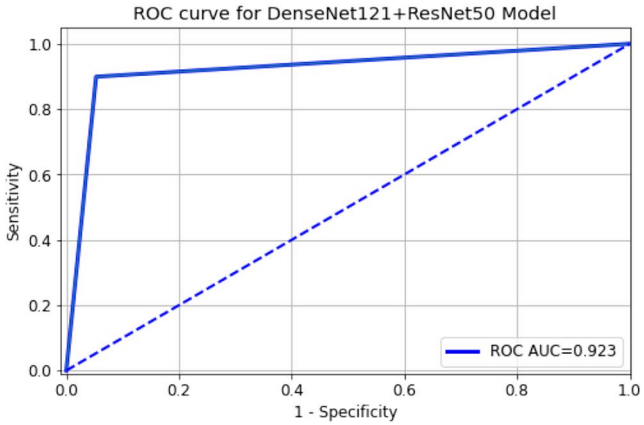
A 100% training accuracy for the three tested models (Fig. 11c) indicates that the models are capable of learning training data accurately. Furthermore, the model's ROC curves are provided in Fig. 13. Test and validation accuracies of 95.3% and 97.7%, respectively, provide evidence that the model can generalize well to unknown data with a 3.0% loss. In comparison to other architectures, the ensemble-based model is better suited for this use case due to the absence of overfitting.

4.3 Experiment C: assessment of multi-label feature fusion using the third division (70%, 10%, 20%)

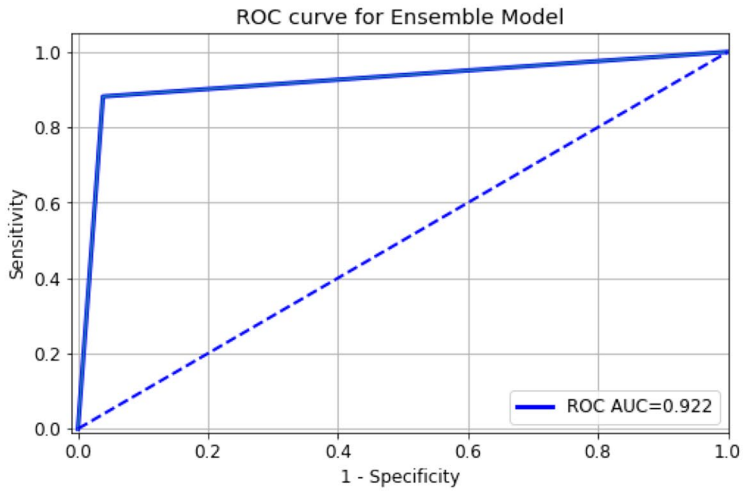
In the last evaluation, we assessed the performance of the developed model using the following information: training (70%), validation (10%), and testing (20%). This comprehensive approach ensures that the model is better equipped to handle diverse data and improve its generalization capabilities.



(a) Densenet121.



(b) Densenet121 and Restnet-50.



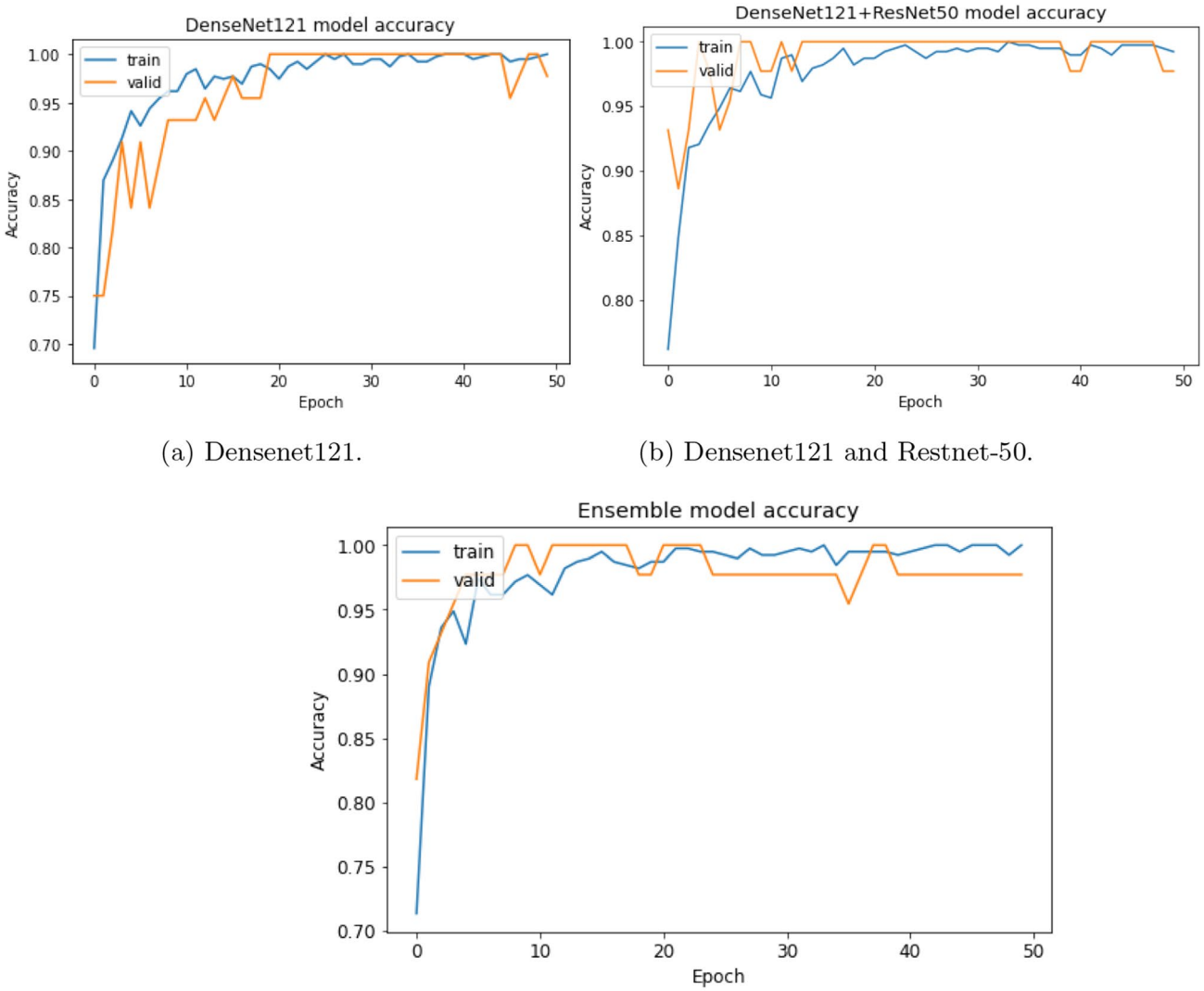
(c) Ensemble Densenet121, Restnet-50, and InceptionV3.

Fig. 10 ROC curve analysis: performance of multi-label classification across singular, hybrid, and ensemble-based approaches in data division #1

Table 6 Performance metrics for multi-label classification with second data division

Model	Training accuracy (%)	Val accuracy (%)	Testing accuracy (%)	F1_score (%)	Sensitivity (%)	Specificity (%)
Densenet-121	100	97.7	96	96.0	97.6	94.8
Hybrid	100.0	97.7	95.6	96.0	97.5	94.0
Ensemble	100.0	97.7	95.3	95.0	95.6	95.0

The Figs. 14 and 15 illustrate the performance of the proposed model during training and evaluation of singular, hybrid, and the proposed ensemble Model. An illustration of the model's learning progress during the training phase is provided in Figs. 14c and 15c. This graph illustrates the evolution of a metric over a 100 epochs period through a training curve. A summary of the accuracy and other evaluation metrics for each model is shown in Table 7. Figure 16 shows the projected and actual output images obtained from the models developed in this study. Compared to the single model, the accuracy rates of the hybrid and ensemble models exhibited a more significant increase (as shown in Figs. 14, 16), underscoring that hybrid and ensemble models are more reliable in predicting findings. Furthermore, the ensemble



(a) Densenet121.

(b) Densenet121 and Restnet-50.

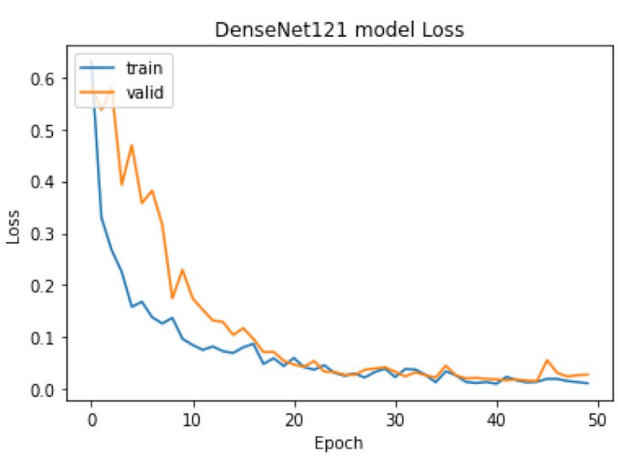
(c) Ensemble Densenet121, Restnet-50, and InceptionV3.

Fig. 11 Multi-label classification accuracy in data division #2: singular, hybrid, and ensemble-based models

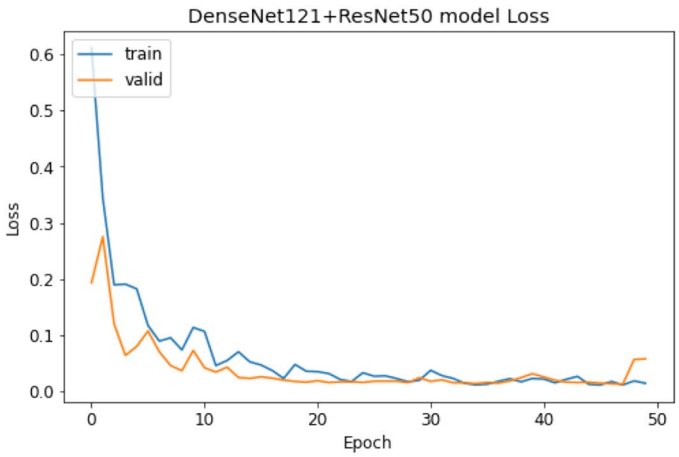
model boasts a high sensitivity rate.For the third division of DenseNet-121, Hybrid, and Ensemble, performance metrics are provided in Table 7. A 100% accuracy rate was achieved by each model during training. Both DenseNet-121 and Hybrid achieved validation accuracy of 96.6% and 97.7%, respectively. In terms of testing accuracy, DenseNet-121 and Ensemble ranked highest (98.2%), while Hybrid ranked second (97.2%). In all models, F1-scores consistently exceeded the expected range: Ensemble and DenseNet-121 obtained 98.0%, while Hybrid obtained 97.0%. Specificity for all models held steady at about 96.8%, indicating good performance in multi-label classification problems.

4.4 Comparison study

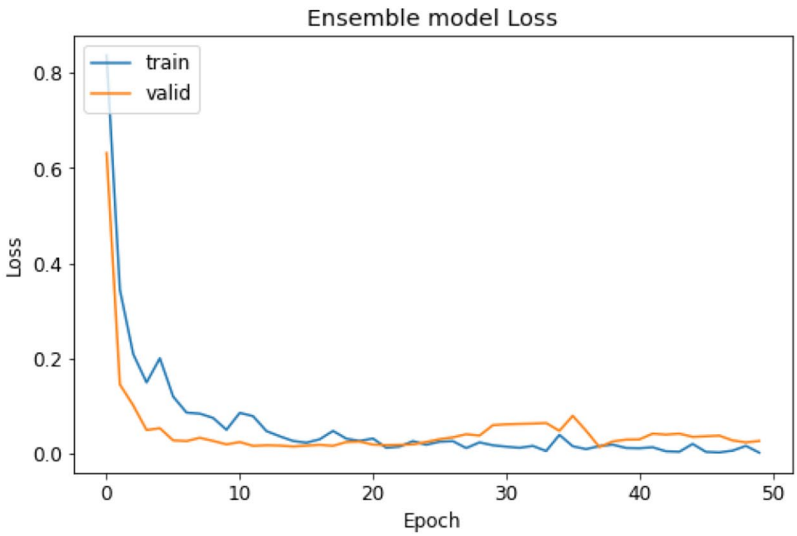
In order to assess the effectiveness of the proposed CataractNetDetect system, a comparative analysis of its performance against state-of-the-art methods has been conducted. Table 8 presents comprehensive details of the comparative results. To ensure a fair evaluation of CataractNetDetect's performance, metrics such as F1-score, AUC, and the final score were used, representing their averages. In comparison with current investigations, the proposed bilateral multiclass classification method demonstrates exceptional efficiency and efficacy. For instance, our method outperformed other approaches, such as CPSO with four machine learning classifiers [47], which achieved 99% training accuracy but did



(a) Densenet121.



(b) Densenet121 and Restnet50.



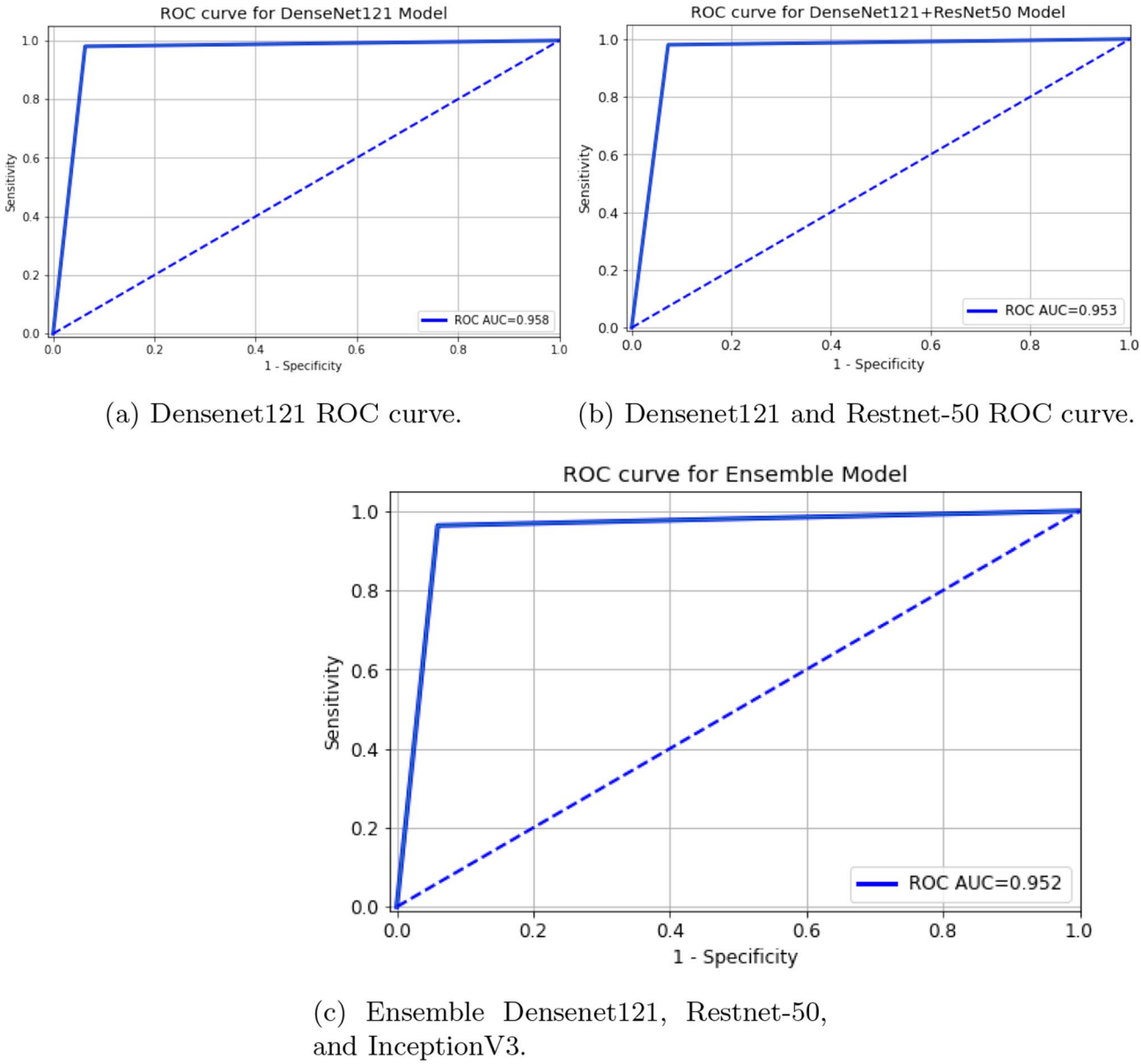
(c) Ensemble Densenet121, Restnet-50, and InceptionV3.

Fig. 12 Multi-label classification loss in data division #2: singular, hybrid, and ensemble-based model

not report validation accuracy. Experiments B and C showed perfect training accuracy of 100% and high validation accuracy of 97.7% under various conditions. Furthermore, compared with methods like BAM with SqueezeNet [50], which demonstrated high training and validation accuracy rates of 98.9% and 98.1% respectively but did not report AUC or F1 scores, our proposed method achieved F1 scores of 98% and Area Under Curve values of 97.9%, particularly in Experiment C. In contrast to Elloumi et al. [17], where a combination of three CNNs with ensemble learning achieved a 95.0% AUC and 94.07% training accuracy, our bilateral multiclass classification methodology shows significant improvements in performance and originality. However, their study did not provide F1 Scores or validation accuracy, limiting a thorough evaluation of their approach. Moreover, bilateral image fusion is a critical component of our approach, enhancing its efficacy and efficiency. This aspect was not discussed or provided by Elloumi et al.[17], highlighting a critical gap in their study. Our bilateral multiclass classification framework addresses this gap, showcasing its uniqueness and sophisticated features.

According to the results (Table 8), Fundus-DeepNet [24] slightly outperforms the proposed system in the off-site test set, however, it exhibits comparatively less effectiveness in all other metrics.

The proposed CatractNetDetect system surpasses existing methodologies in detecting multiple ocular diseases by incorporating an integrated stacking ensemble mechanism that amplifies feature information through the integration



(a) Densenet121 ROC curve.

(b) Densenet121 and Restnet-50 ROC curve.

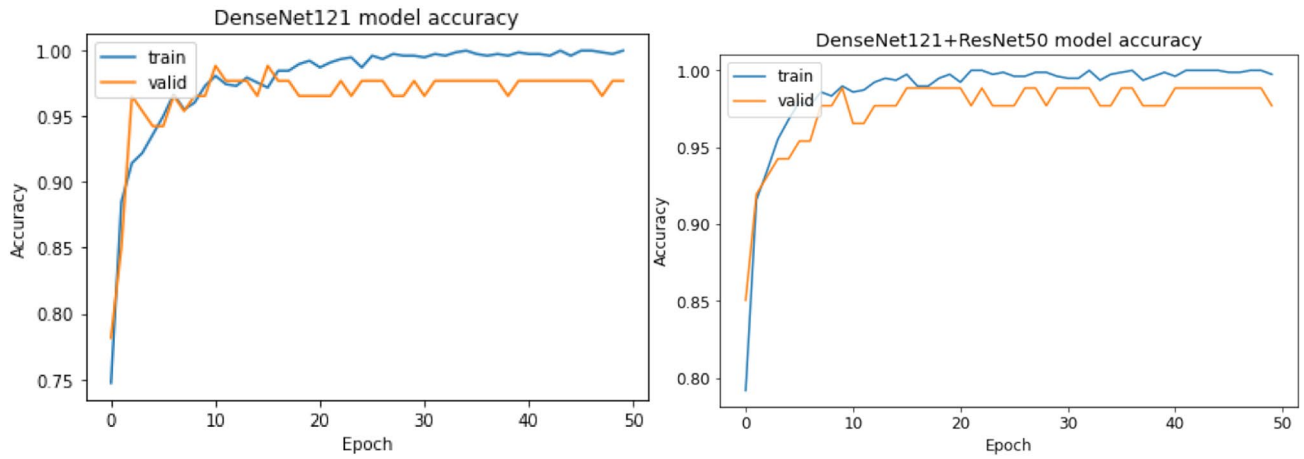
(c) Ensemble Densenet121, Restnet-50, and InceptionV3.

Fig. 13 ROC curve analysis: performance of multi-label classification across singular, hybrid, and ensemble-based approaches in data division #2

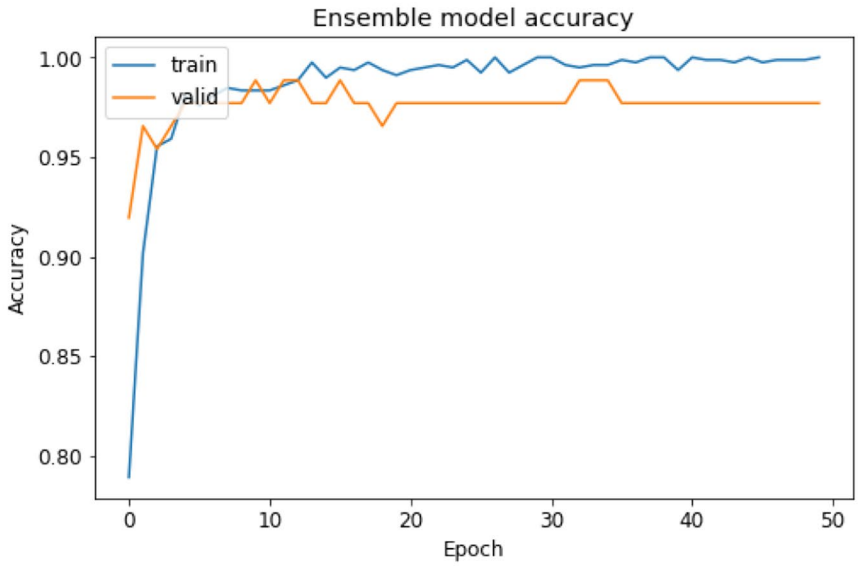
of a variety of feature representations within a two-stream interactive architecture. Consequently, the Fundus-DeepNet system achieves superior performance in detecting multiple ocular diseases compared to the prevailing methods.

4.5 Disscussion

An intraoperative complication known as posterior capsule rupture (PCR) occurs when the posterior capsule of the crystalline lens is disrupted. PCR is concerning due to its potential severity, including complications with intraocular lenses, the risk of endophthalmitis, and the development of cystoid macular edema. To assess the risk, comprehensive risk assessments are crucial, drawing insights from preoperative evaluations and patient history. These assessments guide decisions about case assignments to junior or more experienced surgeons and contribute to effective communication of potential risks to patients. Recently, a growing interest has emerged in the use of computer-aided design (CAD) systems, incorporating machine learning and probabilistic classifiers, to improve the accuracy and objectivity of these risk assessments. This interest is particularly relevant in the analysis of fundus images, which capture the interior surface of



(a) Densenet121. (b) Densenet121 and Restnet-50.

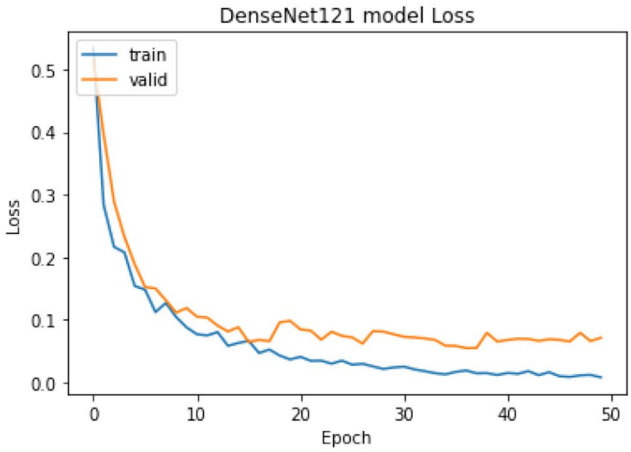


(c) Ensemble Densenet121, Restnet-50 and InceptionV3.

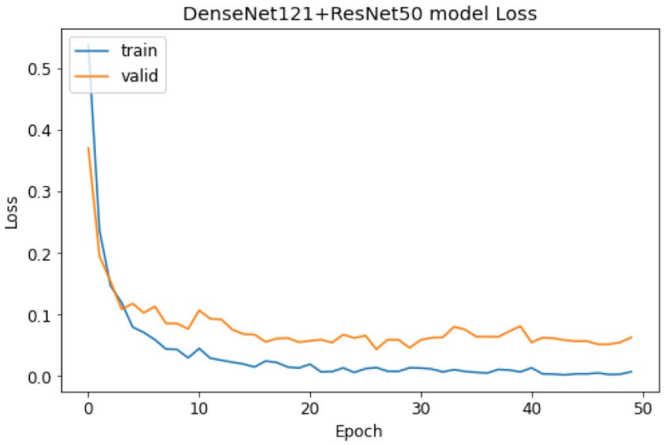
Fig. 14 Multi-label classification accuracy in data division #3: singular, hybrid, and ensemble-based model

the eye, depicting various eye conditions such as normal right/left eyes or affected left/right eyes. Despite the promising applications of deep learning models to fundus images, a key challenge faced by these models is overfitting. Overfitting occurs when a model is excessively trained, resulting in high performance on the training dataset but suboptimal results on new data. Additionally, using only one source of images to train the model can lead to a lack of variability between the test or validation data set and real-world data.

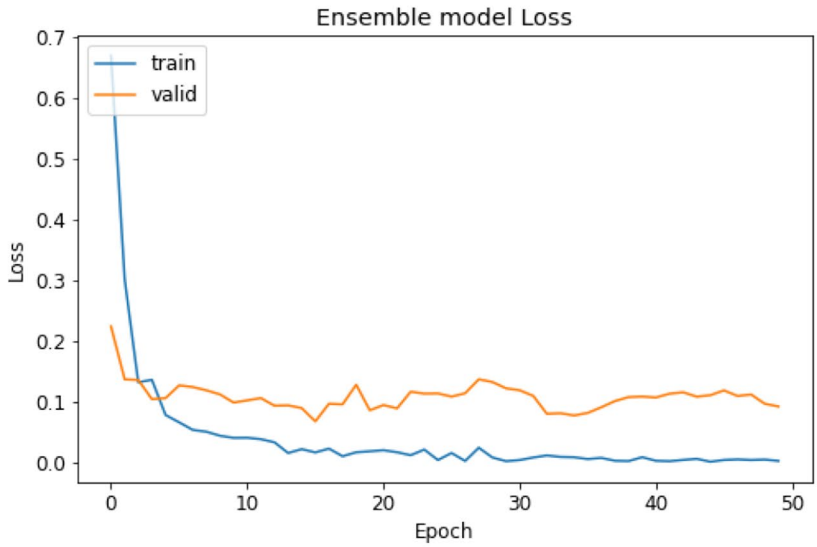
An important objective of this study is to assess the outcomes associated with the fusion of visual features related to left and right eye cataract characteristics. Additionally, we aim to investigate the impact of limited variability in deep learning models, specifically in the classification of cataract fundus versus normal fundus images. Fusion involves combining or linking various characteristics specific to cataracts in both the left and right eyes. Through bilateral feature fusion, researchers aim to effectively combine pertinent data and information from both eyes, resulting in a unified dataset. This dataset can then be utilized to comprehensively examine cataract-related factors. The incorporation of bilateral feature fusion addresses the challenge of overlooking relevant but less conspicuous features, providing a more nuanced understanding of cataract-related aspects. As part of Experiments A, B, and C, three distinct models will be sequentially developed. Initially, a singular model will be constructed, followed by the



(a) Densenet121 Model.



(b) Densenet121 and Restnet50 Model.

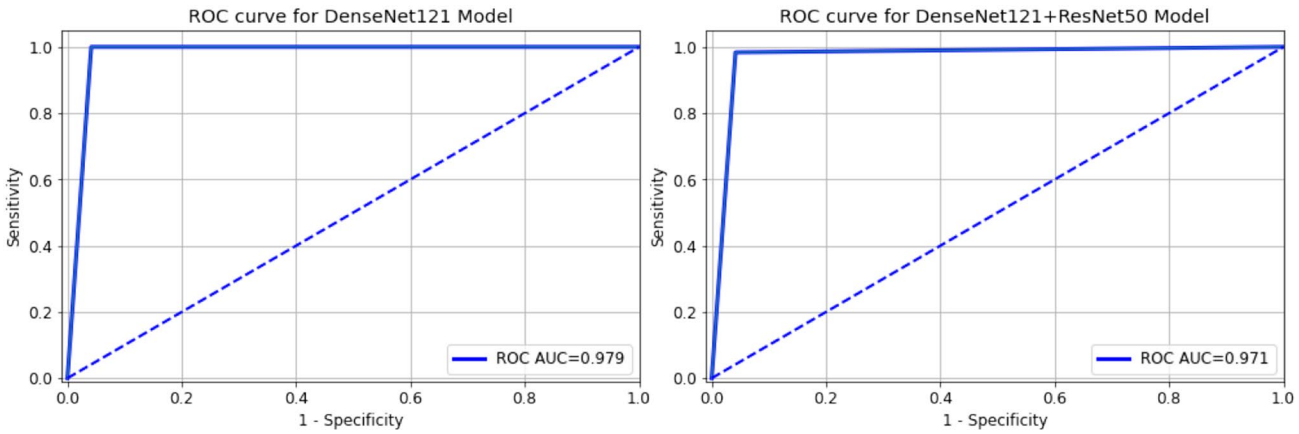


(c) Ensemble Densenet121, Restnet-50 and InceptionV3.

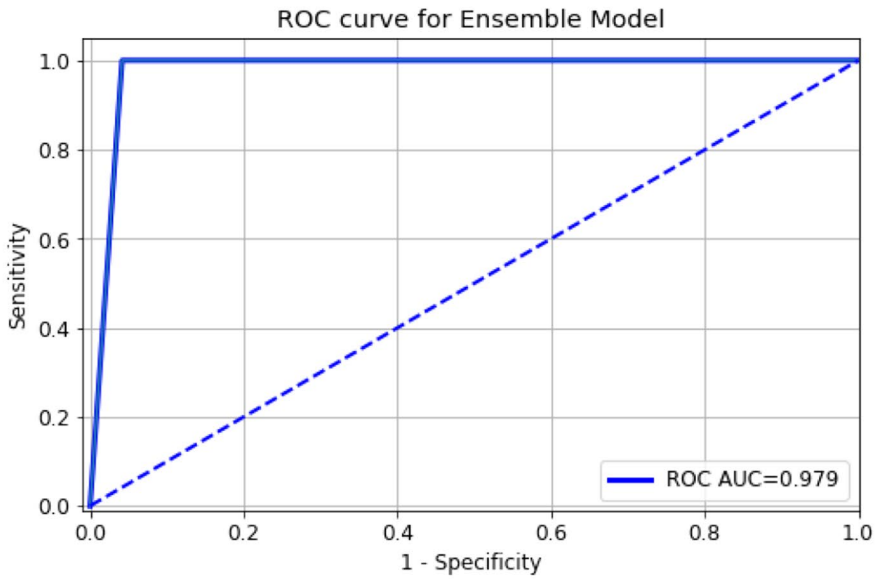
Fig. 15 Multi-label classification loss in data division #3: singular, hybrid, and ensemble-based models

creation of a hybrid model, and finally, the development of an ensemble-based model. The stacking ensemble method offers the advantage of allowing the estimator to directly receive projections from the sub-models, facilitating the investigation of various feature representations of input images and fine-tuning of variable weights.

In Experiment A, the ensemble model exhibited commendable performance on the validation data, with an AUC of 0.922 (Figs. 8, 9, 10). In Experiment B, an AUC of 0.952 was achieved (Fig. 11, 12, 13). The training and testing datasets exhibit excessive overlap due to inherent similarities among images taken from the same patient. Thus, the elevated performance observed in the test dataset is often inaccurately estimated, and the model is susceptible to a modest degree of overfitting. To address this issue, we suggest constructing a more rigorously designed test dataset to ensure that similar data from the training dataset does not appear in the test dataset. Accordingly, in Experiment C, where the model was trained on 70% of patients and tested on the remaining 20%, performance increased significantly, yielding an average AUC of 0.979 (Fig. 14, 15, 16). It appears that the model can accurately predict the disease based on this discrepancy. Additionally, Figs. 8 and 11 illustrate the training process in Experiments A and B, demonstrating a stable and consistent pattern between the training and testing sets. The discrepancy between the results of Experiments A and C suggests that the ensemble-based model has generalized well beyond the images it was trained on.



(a) Densenet121. (b) Densenet121 and Restnet-50.



(c) Ensemble Densenet121, Restnet-50 and InceptionV3.

Fig. 16 ROC curve analysis: performance of multi-label classification across singular, hybrid, and ensemble-based approaches in data division #3

Table 7 Performance metrics for multi-label classification with third data division

Model	Training accuracy (%)	Val accuracy (%)	Testing accuracy (%)	F1_score (%)	Sensitivity (%)	Specificity (%)
Densenet-121	100	96.6	98.2	98.0	100	96.8
Hybrid	100.0	97.7	97.2	97.0	97.9	96.8
Ensemble	100.0	97.7	98.2	98.0	100	96.8

In Experiments A and B, insufficient variability in training data has a substantial impact on the performance of the model. Model performance improves significantly when the training data reflects the complexity of real-world data.

Research studies have demonstrated that the incorporation of feature fusion enhances the resilience of models by augmenting data variability [52, 53]. This experiment indicates that a higher degree of variability in the training

Table 8 Summary of recent studies: a comparison of methods performance metrics

Refs.	Method	No. of samples	Training accuracy (%)	Validation accuracy (%)	AUC (%)	F1 Score (%)
[17]	Stacking 3 CNNs with Ensemble learning	590	94.07	–	95.0	95.04
[47]	CPSO and 4 ML classifiers	110	99	–	90.5	97
[48]	Stacking of SVM and BPNN	1239	84.5	–	–	–
[49]	4 pre-trained CNN with two different optimizers	5000	99.78	89.06	68.88	85.57
[24]	DRBM with a Softmax layer	10,000	92.66	–	99.76	89.13
[50]	BAM with SqueezeNet	4166	98.9	98.1	–	–
[46]	multiple deep neural networks	10,000	75.16	–	86.91	87.93
[51]	A two-stream CNN architecture	10,000	76.97	–	90.3	88.6
Proposed	Experiment A setting	5000	94.4	100	92.2	92
Proposed	Experiment B setting	5000	100	97.7	95.2	95
Proposed	Experiment C setting	5000	100	97.7	97.9	98

BAM :Bottleneck Attention Module, *DRBM:* Discriminative Restricted Boltzmann Machine, CPSO: Particle swarm optimization.

[a] Experiment A, B, C setting are given in Sect. 4

dataset proves advantageous for establishing a robust model applicable to real-world scenarios. Our study delineates the effect of insufficient variability in the training data on the model's generalization capability, providing a framework for the development of future diagnostic pathology models highly resilient to machine learning. The collection of samples from diverse subjects plays a pivotal role in enhancing data variability. Beyond subject-level variability, the inclusion of pathological knowledge grounds provides an additional avenue for increasing data variability. In the future, this can be accomplished by, for instance, incorporating different stages or types of a particular cataract-affected disease. Using the example of a cataract-diagnosing model, it is imperative to include all different subtypes of cataracts to augment data variability. However, such an approach requires unrestricted access to a substantial number of images of various types from a broad spectrum of patients. It is noteworthy that patient confidentiality constraints often restrict access to images in many published studies.

5 Conclusions

A high level of pressure inside the eyeball leads to glaucoma, a condition characterized by progressive, irreversible, and slow vision loss. The incorporation of biomedical feature fusion addresses the challenge of overlooking relevant but less conspicuous features, providing a more nuanced understanding of disease-related aspects. In this study, we aim to introduce a robust model designed to detect PCR in fundus images with enhanced accuracy and minimal false negatives. This objective is being accomplished through the implementation of a multi-headed ensemble model to classify instances of PCR cataract disease using ResNet-50, DenseNet-121, and Inception-V3. Through a series of experiments, the proposed method consistently demonstrates superior performance compared to single-based or hybrid DNN algorithms, reaching a detection accuracy of 100%. The trained and fine-tuned deep learning model exhibits enhanced generalization capability and reduced computation cost. In contrast to prior works which explicitly delineate feature sets, we have developed a fully automated multi-label neural-based machine learning classification system. In order to address issues such as overfitting and imbalanced data, we employ a variety of augmentation procedures. Additionally, the suggested CataractNetDetect system is trained using previously processed fundus images rather than raw images directly, aiming to reduce generalization error and avoid overfitting. As future work, innovative strategies that integrate data from the fundus and visual fields, reducing the noise inherent in both tests, will improve our ability to monitor the progression of glaucoma. A variety of advanced methods of image processing and classification may be used to improve the accuracy of classification. Nature-inspired computing methodologies can be used for feature reduction or selection, and machine learning optimization for classifying fundus retinal images under examination. Additionally, different stages of glaucoma infection, such as mild, moderate, and severe, can be incorporated.

Acknowledgements Acknowledgments are not compulsory. Where included they should be brief. Grant or contribution numbers may be acknowledged. Please refer to Journal-level guidance for any specific requirements.

Author contributions Walaa N. Ismail conceptualized and designed the study, conducted data analysis, interpreted the results, and drafted the manuscript. Hessah A. A. contributed to the study design, data analysis, interpretation of results, and critically revised the manuscript for important intellectual content. Both authors approved the final version of the manuscript to be published and agree to be accountable for all aspects of the work.

Data availability The dataset used in this study is publicly available and can be accessed at https://www.kaggle.com/code/gpreda/ocular-disease-recognition-eda/input?select=ODIR-5K. Further details on data availability and access can be found in the original publication [46].

Declarations

Competing interests The authors declare no competing interests.

References

1. Ang MJ, Afshari NA. Cataract and systemic disease: a review. Clin Exp Ophthalmol. 2021;49(2):118–27.
2. Lam D, Rao SK, Ratra V, Liu Y, Mitchell P, King J, Tassignon M-J, Jonas J, Pang CP, Chang DF. Cataract. Nat Rev Disease Primers. 2015;1(1):1–15.
3. Ahn SJ, Woo SJ, Hyon JY, Park KH. Cataract formation associated with ocular toxocariasis. J Cataract Refract Surg. 2013;39(6):830–5.
4. Curi ALL, de-la-Torre A, Schlaen A, Mahendradas P, Biswas J. Pediatric posterior infectious uveitis. Ocul Immunol Inflamm. 2023;31(10):1944–54.
5. Mk SV. Computer-aided diagnosis of anterior segment eye abnormalities using visible wavelength image analysis based machine learning. J Med Syst. 2018;42(7):128.
6. Sanghavi J, Kurhekar M. Ocular disease detection systems based on fundus images: a survey. Multimedia Tools Appl. 2023;83:21471–96.
7. Varma N, Yadav S, Yadav JKPS. Fundus image-based automatic cataract detection and grading system. In: AIP Conference Proceedings, 2023;2724. AIP Publishing.
8. Lu Z, Miao J, Dong J, Zhu S, Wu P, Wang X, Feng J. Automatic multilabel classification of multiple fundus diseases based on convolutional neural network with squeeze-and-excitation attention. Transl Vis Sci Technol. 2023;12(1):22–22.
9. Sengar N, Joshi RC, Dutta MK, Burget R. Eyedeep-net: a multi-class diagnosis of retinal diseases using deep neural network. Neural Comput Appl. 2023;35:10551–71.
10. Selvathi D. Classification of ocular diseases using transfer learning approaches and glaucoma severity grading. In: Computational methods and deep learning for ophthalmology. Amsterdam: Elsevier; 2023. p. 1–15.
11. Khan MS, Tafshir N, Alam KN, Dhruba AR, Khan MM, Albraikan AA, Almalki FA, et al. Deep learning for ocular disease recognition: an inner-class balance. Comput Intell Neurosci. 2022;2022:5007111.
12. Khalil T, Usman Akram M, Khalid S, Jameel A. Improved automated detection of glaucoma from fundus image using hybrid structural and textural features. IET Image Proc. 2017;11(9):693–700.
13. Gautam D. Improved machine learning-based glaucoma detection from fundus images using texture features in fawt and ls-svm classifier. Multimedia Tools Appl. 2024;1–16.
14. Singh LK, Garg H. Detection of glaucoma in retinal images based on multiobjective approach. Int J Appl Evol Comput (IJAEC). 2020;11(2):15–27.
15. Acharya UR, Ng E, Eugene LWJ, Noronha KP, Min LC, Nayak KP, Bhandary SV. Decision support system for the glaucoma using gabor transformation. Biomed Signal Process Control. 2015;15:18–26.
16. Singh LK, Khanna M, Thawkar S, Singh R. A novel hybridized feature selection strategy for the effective prediction of glaucoma in retinal fundus images. Multimedia Tools Appl. 2024;83(15):46087–159.
17. Elloumi Y. Cataract grading method based on deep convolutional neural networks and stacking ensemble learning. Int J Imaging Syst Technol. 2022;32(3):798–814.
18. Singh LK, Khanna M, Thawkar S. A novel hybrid robust architecture for automatic screening of glaucoma using fundus photos, built on feature selection and machine learning-nature driven computing. Expert Syst. 2022;39(10):13069.
19. El-Hoseny HM, Elsepae HF, Mohamed WA, Selmy AS. Optimized deep learning approach for efficient diabetic retinopathy classification combining vgg16-cnn. Comput Mater Continua. 2023;77(2).
20. Gour N, Tanveer M, Khanna P. Challenges for ocular disease identification in the era of artificial intelligence. Neural Comput Appl. 2023;35(31):22887–909.
21. Park S-J, Ko T, Park C-K, Kim Y-C, Choi I-Y. Deep learning model based on 3d optical coherence tomography images for the automated detection of pathologic myopia. Diagnostics. 2022;12(3):742.

22. Zhang X, Xiao Z, Higashita R, Hu Y, Chen W, Yuan J, Liu J. Adaptive feature squeeze network for nuclear cataract classification in as-oct image. J Biomed Inform. 2022;128: 104037.
23. Sebastian A, Elharrouss O, Al-Maadeed S, Almaadeed N. A survey on deep-learning-based diabetic retinopathy classification. Diagnostics. 2023;13(3):345.
24. Al-Fahdawi S, Al-Waisy AS, Zeebaree DQ, Qahwaji R, Natiq H, Mohammed MA, Nedoma J, Martinek R, Deveci M. Fundus-deepnet: multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images. Inf Fusion. 2024;102: 102059.
25. Gu Y, Fang L, Mou L, Ma S, Yan Q, Zhang J, Liu F, Liu J, Zhao Y. A ranking-based multi-scale feature calibration network for nuclear cataract grading in as-oct images. Biomed Signal Process Control. 2024;90: 105836.
26. Ferris FL, Davis MD, Clemons TE, Lee L-Y, Chew EY, Lindblad AS, Milton RC, Bressler SB, Klein R. A simplified severity scale for age-related macular degeneration: Areds report no. 18. Archives of ophthalmology (Chicago, Ill.: 1960) 2005;123(11):1570–1574.
27. Koh JE, Ng EY, Bhandary SV, Laude A, Acharya UR. Automated detection of retinal health using phog and surf features extracted from fundus images. Appl Intell. 2018;48:1379–93.
28. Çetİner H, Çetİner İ. Classification of cataract disease with a densenet201 based deep learning model. J Inst Sci Technol. 2022;12(3):1264–76.
29. Heidari Z, Baharinia M, Ebrahimi-Besheli K, Ahmadi H. A review of artificial intelligence applications in anterior segment ocular diseases. Medical Hyp Discov Innov Optometry. 2022;3(1):22–33.
30. Wang J, Wang S, Zhang Y. Artificial intelligence for visually impaired. Displays. 2023;77: 102391.
31. Ruzicki J, Holden M, Cheon S, Ungi T, Egan R, Law C. Use of machine learning to assess cataract surgery skill level with tool detection. Ophthalmol Sci. 2023;3(1): 100235.
32. Simanjuntak RBJ, Fuâ Y, Magdalena R, Saidah S, Wiratama AB, Daâ I. Cataract classification based on fundus images using convolutional neural network. JOIV Int J Inf Visual. 2022;6(1):33–8.
33. Tripathi P, Akhter Y, Khurshid M, Lakra A, Keshari R, Vatsa M, Singh R. Mtcd: cataract detection via near infrared eye images. Comput Vis Image Underst. 2022;214: 103303.
34. Padalia D, Mazumdar A, Singh B. A cnn-lstm combination network for cataract detection using eye fundus images. arXiv preprint arXiv: 2210.16093 2022.
35. Lai C-J, Pai P-F, Marvin M, Hung H-H, Wang S-H, Chen D-N. The use of convolutional neural networks and digital camera images in cataract detection. Electronics. 2022;11(6):887.
36. Junayed MS, Islam MB, Sadeghzadeh A, Rahman S. Cataractnet: an automated cataract detection system using deep learning for fundus images. IEEE Access. 2021;9:128799–808.
37. Raju M, Shanmugam KP, Shyu C-R. Application of machine learning predictive models for early detection of glaucoma using real world data. Appl Sci. 2023;13(4):2445.
38. Alaeddini Z. A review of the latest machine learning advances in cataract diagnosis. J Ophthal Optometr Sci. 2021;4(4):46–60.
39. Uppamma P, Bhattacharya S, et al. Deep learning and medical image processing techniques for diabetic retinopathy: a survey of applications, challenges, and future trends. J Healthc Eng. 2023;2023:2728719.
40. Triepels RJ, Segers MH, Rosen P, Nuijts RM, Biggelaar FJ, Henry YP, Stenevi U, Tassignon M-J, Young D, Behndig A, et al. Development of machine learning models to predict posterior capsule rupture based on the Eurequo registry. Acta Ophthalmol. 2023;101:644.
41. Goutam B, Hashmi MF, Geem ZW, Bokde ND. A comprehensive review of deep learning strategies in retinal disease diagnosis using fundus images. IEEE Access. 2022;10:57796–823.
42. Gao X, Lin S, Wong TY. Automatic feature learning to grade nuclear cataracts based on deep learning. IEEE Trans Biomed Eng. 2015;62(11):2693–701.
43. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017;4700–4708.
44. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2017;31.
45. Koonce B, Koonce B. Resnet 50. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization. 2021. p. 63–72.
46. Li N, Li T, Hu C, Wang K, Kang H. A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In: Benchmarking, Measuring, and Optimizing: Third BenchCouncil International Symposium, Bench 2020, Virtual Event, November 15–16, 2020, Revised Selected Papers 3, 2021;177–193. Springer
47. Singh LK, Khanna M, Garg H, Singh R. Efficient feature selection based novel clinical decision support system for glaucoma prediction from retinal fundus images. Med Eng Phys. 2024;123: 104077.
48. Yang J-J, Li J, Shen R, Zeng Y, He J, Bi J, Li Y, Zhang Q, Peng L, Wang Q. Exploiting ensemble learning for automatic cataract detection and grading. Comput Methods Progr Biomed. 2016;124:45–57.
49. Gour N, Khanna P. Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network. Biomed Signal Process Control. 2021;66: 102329.
50. Zia A, Mahum R, Ahmad N, Awais M, Alshamrani AM. Eye diseases detection using deep learning with bam attention module. Multimedia Tools Appl. 2023;1–24.
51. Ou X, Gao L, Quan X, Zhang H, Yang J, Li W. Bfenet: a two-stream interaction cnn method for multi-label ophthalmic diseases classification with bilateral fundus images. Comput Methods Progr Biomed. 2022;219: 106739.
52. Mayya KUSDKV. Acharya UR, An empirical study of preprocessing techniques with convolutional neural networks for accurate detection of chronic ocular diseases using fundus images. In: Applied Intelligence, 2023;53:1548–1566. Springer.
53. Veturi YA, Woof W, Lazebnik T, Moghul I, Woodward-Court P, Wagner SK, Guimarães TAC, Varela MD, Liefers B, Patel PJ. Syntheye: investigating the impact of synthetic data on artificial intelligence-assisted gene diagnosis of inherited retinal disease. Ophthalmol Sci. 2023;3(2): 100258.