Multimedia Tools and Applications (2024) 83:30773–30818
https://doi.org/10.1007/s11042-023-16835-3

Eye diseases diagnosis using deep learning and multimodal medical eye imaging

Sara El-Ateif[1] · **Ali Idri**[1,2]

Received: 22 October 2022 / Revised: 24 May 2023 / Accepted: 31 August 2023 /
Published online: 13 September 2023

Abstract
The present study carries out an empirical evaluation and comparison of the seven most recent deep Convolutional Neural Network (CNN) techniques (VGG19, DenseNet121, InceptionV3, InceptionResNetV2, Xception, ResNet50V2, MobileNetV2) for eye disease classification into normal and diseased using mono-modality and late fusion multimodality over three datasets. All empirical evaluations were carried out using: (1) six classification metrics, (2) the Scott-Knott Effect Size Difference (SK-ESD) statistical test, (3) Borda count method, and (4) publicly available datasets: DHS (combining three datasets: DRIVE, STARE, and HRF), FFA, and Macula. The results showed that DenseNet121 and ResNet50V2 were the top performing and less sensitive techniques on the three datasets using mono-modality with accuracy values of 99.57% and 99.51% respectively. As for late fusion techniques, they outperformed mono-modality across the three datasets, regardless of the mono-modality used. Moreover, ResNet50V2 late fusion was the best late fusion technique and scored 100% in accuracy across all three datasets. Additionally, our proposed ResNet50V2 late fusion model trained on the Macula dataset outperforms current state-of-the-art models trained on more than one eye disease (accuracy of: proposed ResNet50V2 late fusion = 100%, New CNN = 81.33%), and is similar to best ranking feature-level fusion one eye disease model with accuracy equal to 100%.

Keywords Eye diseases · Diabetic eye diseases · Binary classification · Late fusion · Multimodality · Deep convolutional neural networks

✉ Ali Idri
ali.idri@um5.ac.ma

Sara El-Ateif
sara_elateif@um5.ac.ma

[1] Software Project Management Research Team, ENSIAS, Mohammed V University, Rabat, Morocco

[2] Mohammed VI Polytechnic University, Benguerir, Morocco

1 Introduction

Eye diseases (ED), especially Diabetic Eye Diseases (DED) mainly Diabetic Retinopathy (DR), Diabetic Macular Edema (DME), Glaucoma (Gl), and Cataracts are all conditions caused by diabetes. These DED can cause vision loss and blindness in adults between the ages of 20 and 74. In Europe, the number of patients with DED is expected to rise from 6.4 million today to 8.6 million in 2050 [1]. Therefore, early detection of these disorders is essential for slowing their progression and lowering the risk of blindness.

Recently, deep learning (DL) techniques have shown great progress at diagnosing diseases using mono-modality medical imaging and even more so by using fundus images to identify diabetic retinopathy and its severity level [2–4]. However, in some cases, it may be difficult to diagnose a patient by examining only one modality, which requires physicians to acquire another medical modality such as an optical coherence tomography (OCT). Along with the fundus modality, the OCT provides 3-dimensional structural information that helps in the diagnosis of age-related macular degeneration (AMD). By fusing these two modalities and feeding them to a pretrained deep neural network such as VGG-19, and transfer learning using random forest, the performance improved in comparison with trained mono-modality models [5].

Multimodality fusion methods have been explored for several tasks in the medical field, such as prognosis prediction of breast cancer [6], Mild Cognitive Impairment (MCI) detection [7], and glaucoma classification [8]. Multimodality learning can be classified into three categories [9]: early, joint, and late fusion. Most multimodal previous studies in eye diseases [5, 10–16] concentrated on comparing proposed multimodal fusion models with mono-modality by focusing on one eye diseases (AMD, DR, and Glaucoma). For AMD, Yoo et al. [5], trained VGG19 with Random Forest, InceptionV3, Restricted Boltzmann Machine, and Deep Belief Networks with mono-modality and different combinations of OCT and fundus images. Vaghefi et al. [10] detected intermediate dry AMD by training Inception-ResNetV2 on OCT, OCT angiography (OCTA), fundus, and different combinations of these modalities. Jin et al. [11] trained a feature-level fusion model for multiview learning to assess choroidal neovascularization in AMD by combining OCT and OCTA image features. For DR, Tseng et al. [12] proposed two DL architectures: late fusion and two stage early fusion. Late fusion using a post-processing process combines lesion and severity classification models in parallel. Two stage early fusion combines lesion detection and classification models sequentially. El-Ateif and Idri [13] trained and evaluated seven deep learning CNN models (VGG19, ResNet50V2, DenseNet121, InceptionV3, InceptionResNetV2, Xception, and MobileNetV2) using a joint fusion approach with the fundus and preprocessed fundus (WGBF) as multimodality. Hervella et al. [14] pretrained a neural network using the proposed Multimodal Image Encoding with the unlabeled multimodal retinal image pairs (color retinography, and fluorescein angiography), fine-tuned it, and used it for DR grading using fundus images. For Glaucoma, An et al. [15] trained five VGG19 models on specific regions from fundus and OCT images using transfer learning. The outputs of these five models were then fed into a random forest (RF) model to distinguish between healthy and glaucomatous disc fundus images. In order to detect referable from non-referable Glaucoma, Lee et al. [16] trained a U-Net combined with DenseNet121 to segment the optic disk from the fundus images using weakly supervised learning. Then, they trained a multimodal network comprised of two DenseNet121 models, concatenated at the dropout level, and took as input the fundus and extracted optic disk. All of these studies

concluded that multimodality outperformed mono-modality for eye disease classification. However, to the best of our knowledge, none of the primary studies dealing with mono and multimodality have evaluated and compared advanced DL models trained on both mono and multimodal data for several eye diseases classification, which motivated the present study.

The purpose of this study was to assess and compare the performance of both mono-modality and late fusion multimodality strategies when using most recent and frequent seven DL techniques (VGG19 [17], ResNet50V2 [18], DenseNet121 [19], InceptionV3 [20], InceptionResNetV2 [21], Xception [22], and MobileNetV2 [23]) in ED classification based on accuracy, sensitivity, specificity, precision, F1-score, and AUC over three datasets (DHS built from three existing datasets: DRIVE [24], STARE [25], and HRF [26]; FFA [27]; and Macula [28]. Moreover, we used the Scott-Knott Effect Size Difference (SK-ESD) [29] statistical test, and the Borda Count voting method [30] for ED classification. The SK-ESD test was used to determine the best statistically different groups of DL techniques with non-negligible differences. The Borda Count voting technique is used to rank the best techniques selected by the SK-ESD test [31].

This study explored five research questions:

(RQ1): What is the overall performance of DL models using mono-modality data in ED classification? Is there any mono-modality DL architecture that outperforms its pairs?
(RQ2): How does a modality impact the diagnostic performance of a DL architecture?
(RQ3): What is the overall performance of DL models using multimodality data in ED classification?
(RQ4): How do late fusion architectures perform in comparison with the mono-modality?
(RQ5): How do the best model(s) compare with the state-of-the-art?

The key contributions of this empirical study are as follows:

(1). Training seven mono-modality and late fusion multimodality DL architectures (VGG19, DenseNet121, InceptionV3, InceptionResNetV2, Xception, ResNet50V2, MobileNetV2) for ED classification.
(2). Evaluating both mono-modality and late fusion multimodality DL techniques over three datasets: DHS [24–26], FFA [27], and Project Macula [28].
(3). Training and evaluating the impact of different eye diagnostic modalities (i.e., fundus, vessel segmentation, fundus fluorescein angiography photography (FFAP), OCT and histology) using one modality at a time, and several modalities' combinations depending on the dataset (i.e., fundus and vessel segmentation; fundus and FFAP; fundus, OCT, and histology) through late fusion on ED diagnosis.
(4). Comparing the performances of mono-modality and late fusion multimodality when using the seven DL architectures by means of SK-ESD statistical test and Borda Count voting technique.

The remainder of this paper is organized as follows. Section 2 presents the medical and technical concepts used in this study. Section 3 highlights related work. Section 4 presents the data preparation. Section 5 outlines the experimental process. Section 6 presents and discusses the results. Section 7 reports the threats to the validity of the study. Finally, Section 8 concludes the paper and introduces the future direction of this study.

2 Background

This section provides an overview of the medical and technical concepts used in this work. First, we introduce common retina eye diseases and imaging modalities. Second, we explain multimodality fusion and its types. Finally, we describe the seven DL approaches used in this study.

2.1 Retina eye diseases and imaging modalities

According to the Centers for Disease Control and Prevention (CDC), age-related eye diseases such as age-related macular degeneration, diabetic retinopathy, and glaucoma are the leading causes of blindness and low vision in the United States, most of which manifest themselves in the retina. Hereafter, we primarily describe eye diseases that can be diagnosed by retina imaging modalities [32]:

(1) *Diabetic retinopathy (DR).* The retina detects and interprets light as electrical impulses, which are then transmitted to the brain via the optic nerve. Diabetic retinopathy is an outcome of diabetes-related damage to the retina's small blood vessels, which can lead to blood and fluid leaks. It is the leading cause of blindness and vision loss in the United States. Its progression can be divided into four stages: mild, moderate, and severe non-proliferative retinopathy, and proliferative diabetic retinopathy.
(2) *Age-related macular degeneration (AMD).* The macula, the center region of the retina that allows the eye to discern small details, is affected by age-related eye conditions. Reading and fine or close vision are permanently impaired because of AMD. The two main types of AMD are dry AMD and wet AMD, with dry AMD often causing a progressive decrease of visual acuity. The most visually hazardous type of AMD is wet AMD, also known as choroidal neovascularization (CNV), which is characterized by the ingrowth of a choroidal vascular structure into the macula, as well as increased vascular permeability. It is worth noting that severe and mild AMD diminish the likelihood of finding work by 61%, and 44%, and salary by 39% and 32%, respectively [32].
(3) *Glaucoma (Gl).* Occurs when the natural fluid or eye pressure inside the eyes progressively increases. It can harm the eye optic nerve, resulting in vision loss.
(4) *Cardiovascular disease.* One of the ways cardiovascular disease appears in the retina is through changes in the A/V ratio which measures the distance between the diameters of the retinal arteries and veins.

Most of these diseases can be treated if they are diagnosed earlier. To diagnose these diseases, doctors use the following imaging modalities: fundus photography, OCT and fluorescein angiography. Fundus photography is commonly used to diagnose diabetic retinopathy, age-related macular degeneration, and glaucoma on a broad scale (i.e., population based). Inflammatory retinal disorders, diabetic retinopathy, and macular degeneration are widely diagnosed and treated using OCT and fluorescein angiography [32].The following is a quick introduction to each of these imaging techniques.

(1) *Color fundus photography.* A diagnostic procedure that uses a fundus camera to record colored images of the interior surface of the eye. It captures the retina, optic disc, macula, retinal blood vessels, and posterior pole, commonly known as the fundus.

(2) *Fluorescein angiography.* The posterior layers of the eyes were imaged using this technique. It involves injecting a small amount of liquid dye into a vein to help physicians better comprehend the disease process.
(3) *Optical coherence tomography (OCT).* Light waves are used to take cross-sectional photographs of the retina in this noninvasive imaging method. OCT allows ophthalmologists to evaluate each layer of the retina and map and measure their thickness.

In most cases doctors rely on color fundus photography to diagnose diseases, but in some cases such as early detection of AMD, it can be quite challenging to determine using only one modality, so physicians run an OCT for a better view and detailed diagnosis [33].

2.2 Multimodality fusion

Multimodality fusion, as defined in [9], is the process of combining data from many modalities to derive complementary and more complete information to enhance the performance of machine learning models. Huang et al. [9] highlighted three fusion strategies: early fusion, joint fusion, and late fusion. Early fusion, or feature level fusion, is the process of merging several input modalities into a single feature vector before putting it into a single machine learning model for training. In joint fusion, learned feature representations from intermediate layers of neural networks are combined as the input to a final model with features from other modalities. Late fusion, also known as decision-level fusion, is the process of combining the predictions of multiple models trained independently to generate a final conclusion. In this study, we compared seven multimodal fusion DL models to seven mono-modal DL models. Mono-modality refers to the process of feeding a DL model only one modality (i.e., retinal fundus images) during its training.

2.3 Deep learning techniques

Deep learning using medical imaging for automatic disease diagnoses has been explored by several researchers [34–40] and has been found to be extremely helpful in assisting doctors during the diagnosis process of some DED using mono-modality imaging. Very few studies have been conducted on the combination of several modalities to improve diagnosis [5, 10, 13, 15, 41–45]. Seven ImageNet-trained DL techniques were chosen for training on three datasets: VGG19, DenseNet121, InceptionV3, InceptionResNetV2, Xception, ResNet50V2, MobileNetV2. These seven DL architectures were selected from the most commonly used techniques in previous diabetic eye diseases diagnostic studies [10, 15, 34, 35, 37–39]. The seven DL architectures are briefly described as follows:

VGG19 VGG19 [17] is a 19 layers CNN model, built by stacking convolutional layers together with a limited depth owing to the diminishing gradient problem.

ResNet50V2 ResNet [46] or residual network, is an extremely deep CNN composed of 152 layers. ResNet solves the diminishing gradient problem and introduces the skip connections principal. This model has several variations and we chose ImageNet-trained ResNet50V2 with 50 layers [18] which is more lightweight than the original architecture to speed up the training process. Note that ResNet50V2, in comparison with v1, uses batch normalization before each weight layer. As part of the model architecture, batch normalization [47] entails normalizing for each training mini-batch. Batch normalization also

functions as a regularizer, allowing significantly higher learning rates to be used without the need for precise parameter settings.

DenseNet121 DenseNet is a similar model to ResNet, but instead of residual blocks, DenseNet is composed of densely connected dense blocks. We used the DenseNet121 [19] as it has 121 layers and so is much faster to train.

InceptionV3 InceptionV3 [20] is a convolutional neural network architecture with 42 layers, which is part of the Inception models. It introduces Label Smoothing Factorized 7×7 convolutions along with the use of batch normalization for layers in the sidehead, as well as an auxiliary classifier to propagate label information throughout the network.

InceptionResNetV2 InceptionResNetV2 [21], with 164 layers, is part of the Inception family architectures with the difference that it replaces the Inception architecture's filter concatenation stage.

Xception Xception [22], with 71 layers, is a CNN architecture that relies on depth-wise separable convolution layers instead of the Inception module.

MobileNetV2 MobileNetV2 [23], is a CNN architecture with 53 layers that aims to outperform on mobile devices. Additionally, it is built on an inverted residual structure with residual connections between bottleneck layers.

3 Related work

This section provides a high-level overview of the automatic diagnosis of diabetic eye diseases using mono-modality and multimodality learning. First, we present a summary of some primary studies on the diagnosis of mono-modality automatic diabetic eye diseases. Second, we summarize the findings of some primary studies on multimodal fusion models for automatic medical disease diagnosis.

3.1 Mono-modality automatic diabetic eye diseases (DED) diagnosis

Deep learning advancements have resulted in significant efforts to develop convolutional neural networks for automatic eye disease recognition [13, 34–42]. Sarki et al. [48] reviewed the automatic detection of DED using DL and transfer learning (TL) techniques. The review included 65 selected studies, published between 2014 and 2020, that were analyzed according to the following factors: (1) accessible datasets, (2) image preprocessing approaches, (3) deep learning models, and (4) performance evaluation measures. The main findings of this review are as follows.

- The most frequently used DR datasets are Kaggle and Messidor [49].
- The most frequently used preprocessing techniques are: green channel extraction, contrast enhancement, image resizing, eradication and masking of blood vessels and optical discs, segmentation of the meaningless black border of fundus images, image augmentation and RGB images transformation into grayscale images.

- 38 of the 65 studies employed transfer learning, 21 created their own DL architectures, and six used both DL and ML classifiers.
- Most DL techniques used were part of the VGGNet, Inception, and ResNet families.
- The most frequently used criteria for evaluating classification performance were specificity, sensitivity, accuracy, and AUC.

Table 1 summarizes the information on some selected studies of [48] along with some recent studies from 2021 to 2023 [13, 41–45], including DL techniques used, type of classification, performance metrics used, and the findings of each study. Note that all the studies of Table 1 used the fundus modality. Additionally, most of the recent studies used transformer architectures, specifically vision transformers variants.

3.2 Automatic medical diagnosis with multimodality fusion models

The principle of leveraging the power of multiple modalities to diagnose diseases has been widely used in several medical tasks [5–7, 10, 15]. This section provides an overview of multimodal learning in ED. Table 2 presents information on some selected studies of [9] pertaining to AMD, Gl and DR [5, 10–16] including the fusion method, modalities, DL techniques and performance metrics used. From Table, 2 we observe the following.

- Accuracy was the most commonly used criterion for evaluating the performance of multimodality DL approaches, while it was supplemented by other metrics such as precision, sensitivity, specificity, and statistical tests such as the Duncan, Cohen kappa, Delong tests, and Quadratic Weight Kappa.
- In all reported studies, multimodality outperformed mono-modality techniques.
- The most commonly used multimodality fusion type in the literature is late fusion [5–7]. However, in ED diagnosis early [5, 11, 12, 15, 42] and joint fusion [5, 10, 13, 16] are most commonly used.
- Most studies used OCT and fundus modalities [5, 10–12, 14, 15].
- The most frequently trained model is VGG-19 [5, 13, 15].

4 Data preparation

This section explains how we prepared the data for the five datasets: DRIVE, STARE, HRF, FFA, and Project Macula, which includes:

(1) Datasets description
(2) Data oversampling
(3) Data preprocessing

4.1 Datasets description

We used five publicly available ED datasets containing more than one imaging modality: Digital Retinal Images for Vessel Extraction (DRIVE) [24], STructured Analysis of the Retina (STARE) [25], High-Resolution Fundus (HRF) Image Database [26], Fundus

Table 1 Summary of the literature review of the use of mono-modality DL techniques in DED classification

Authors	Techniques	Classification type	Performance Metrics	Findings and results
Choi et al. [34]	VGG19, VGG19-TL-RF	Binary classification (normal and abnormal)	AUC, Sensitivity, Specificity	Perform a ten retinal diseases classification into normal or abnormal on the STARE dataset using VGG19-TL-RF. The performance results were: AUC = 90.3%, sensitivity = 80.3%, and specificity = 85.5%.
Sayres et al. [35]	Inception version 4	Multi-class classification (five different stages of DR)	Accuracy, Sensitivity, Specificity	Train and tune the Inception version 4 model on 1.6 million retinal fundus images to differentiate between the five stages of DR with accuracy equal to 96.9% on normal images and an accuracy of 57.9% on mild and severe NPDR.
Phan et al. [36]	DenseNet201, ResNet152, VGG19	Binary classification (Glaucoma, non-glaucoma eyes)	AUC, Sensitivity, Specificity	Apply three DCNNs (i.e., DenseNet201, ResNet152, VGG19) to study the glaucoma discrimination and the influence of the quality of the images on the model's performance to detect glaucoma from non-glaucoma and found that poor quality of the images reduced the AUC by 0.1 to 0.2.
Pinto et al. [37]	VGG16, VGG19, InceptionV3, ResNet50 and the Xception	Binary classification (glaucoma and healthy)	AUC, Accuracy, Sensitivity, Specificity	Evaluate VGG16, VGG19, InceptionV3, ResNet50, and Xception ImageNet-trained models for the assessment of the glaucoma disease using fundus images on five public datasets. The average AUC/specificity/sensitivity achieved were 96.05, 85.80%, and 93.46% using the Xception model.

Table 1 (continued)

Authors	Techniques	Classification type	Performance Metrics	Findings and results
Gómez-Valverde et al. [38]	VGG19, RESNET50, GOOGLENET and DENET DISC	Binary classification (Glaucoma, non-glaucoma eyes)	Balanced Accuracy, AUC, Sensitivity, Specificity	Identify glaucoma and non-glaucoma using transfer learning with VGG19 on the RIM-ONE [50], DRISHTI-GS [51] datasets, and one private screening dataset from Hospital de la Esperanza. As results, AUC equal to 94% was obtained, sensitivity and specificity ratio similar to medical experts were achieved.
de Moura Lima et al. [39]	VGG16, VGG19, ResNet50, InceptionResNetV2, InceptionV3 With 4 classifiers: Logistic Regression, Random Forest, SVM, Naive Bayes.	Binary classification (normal or glaucoma)	Accuracy, Sensitivity, Specificity, AUC	Use RIM-ONE dataset to diagnose Gl with VGG-16, VGG-19, ResNet50, InceptionV3 and InceptionResNetV2. Combine ResNet and Logistic Regression that yielded promising results on RIM-ONE-r2 [50] (AUC of 95.7%) and InceptionResNet using a similar classifier resulted in AUC of 86% on RIM-ONE-r3 [50].
Umapathy et al. [40]	Decision Tree (Gini), Decision Tree (Information Gain), SVM (RBF Kernel), SVM (Linear Kernel), Feedforward Artificial Neural Network, Logistic Regression	Binary classification (DR, Normal)	Mean Accuracy, Mean Precision, Mean Recall, F1-score	Conduct two approaches using images from MESSIDOR [49], HRF, STARE, and images from the Retina Institute of Karnataka datasets. In the first approach, using retinal image processing and textual feature extraction with a decision tree classifier, the presence of DR was predicted with an accuracy of 94.4%. In the second one, retraining of the last layer of Inception-V3 was performed and achieved an accuracy of 88%.

Table 1 (continued)

Authors	Techniques	Classification type	Performance Metrics	Findings and results
Sarki et al. [41]	Their own CNN model	Multi-class (DR, Macular Edema, Gl, Cataract and Healthy)	Accuracy, Sensitivity, Specificity	Train a newly created CNN model on the assembled dataset from Messidor [49], Messidor-2 [49, 52], DRISHTI-GS, and kaggle cataract datasets to distinguish between five classes: DR, Macular Edema, Gl, Cataract and Healthy. The proposed model, achieves 81.33% in accuracy, 100% in sensitivity and specificity.
Liu et al. [42]	AlexNet, DenseNet-152, ResNet-50, ResNet-152	Binary classification (Snellen better than 20/40, Snellen worse than 20/40)	AUC, Precision, Recall	Train and compare pretrained AlexNet, DenseNet-152, ResNet-50 and ResNet-152 for classification of Snellen better or worse than 20/40 using the infrared and OCT modalities as mono-modalities. The best model was ResNet-152 that scored 87% in AUC on both their internal Johns Hopkins University (JHU) dataset and the external Amsterdam University Medical Centers (AUMC) dataset on the OCT modality and scored in AUC 83% and 78% on the infrared modality, on the internal and external datasets respectively. The best results were obtained while training on the OCT modality.

Table 1 (continued)

Authors	Techniques	Classification type	Performance Metrics	Findings and results
El-Ateif et al. [13]	VGG19, ResNet50V2, DenseNet121, InceptionV3, InceptionResNetV2, Xception, and MobileNetV2	Binary classification (non-referable DR, referable DR)	AUC, Accuracy, Sensitivity, Specificity, F1-score, Precision	Train and evaluate seven deep learning CNN models using the fundus and preprocessed fundus as single modalities from the APTOS 2019 [53] and Messidor-2 [49, 52] datasets. The best models DenseNet121 (fundus accuracy = 90.63%, preprocessed fundus accuracy = 89.49%) for APTOS 2019 across both modalities and fundus MobileNetV2 (accuracy = 71.80%) and preprocessed fundus Xception (accuracy = 76.79%) for Messsidor-2.
Canayaz [43]	EfficientNet, DenseNet, SVM, Random Forest	Multi-class classification (five different stages of DR)	Accuracy, Quadratic Weight Kappa (QWK)	Train EfficientNet and DenseNet to extract features from the fundus images (512 features) using the APTOS 2019 dataset. Then, feed these features to the wrapper methods (Binary Bat Algorithm, Equilibrium Optimizer, Gravity Search Algorithm, and Gray Wolf Optimizer), to choose the best features from the 512 extracted features. Finally, these best features were input into the classifiers SVM and Random Forest. The best approach scored an accuracy of 96.32% and QWK of 0.98.

Table 1 (continued)

Authors	Techniques	Classification type	Performance Metrics	Findings and results
Gu et al. [44]	Vision Transformer and Residual Attention	Multi-class classification (five different stages of DR)	AUC, Accuracy, Sensitivity, Specificity	Train a Vision Transformer and Residual Attention model composed of two blocks: FEB, feature extraction block, used for feature extraction of fundus images; and GPB, grading prediction block, used to classify the five stages of DR. The fine-grained attention in FEB transformer detects retinal hemorrhage and exudate areas. The residual attention in GPB captures spatial regions occupied by different classes. The model is trained and tested on the DDR [54] and IDRiD [55] datasets. The model scored an accuracy of 82.35%.
Yao et al. [45]	FunSwin (their own transformer model)	Detect Healthy, DR, and Macular Edema (ADM)	Accuracy, Sensitivity, Specificity, F1-score	Train a proposed FunSwin transformer constituted of the Swin Transformer model to detect Healthy, DR, and AMD. The fundus images from Messidor [49] are oversampled through rotation and mirroring and augmented through CutMix and MixUp. FunSwin scored an accuracy of 84.12%.

Table 2 Summary of the literature review of the use of modality fusion in disease classification

Authors	Fusion method	Modalities	Disease	Techniques	Classification type	Performance Metrics	Findings and results
Yoo et al. [5]	Early Fusion, Joint Fusion, Late Fusion	OCT, Fundus images	AMD	VGG-19, Random Forest, Inception-v3, RBM (restricted Boltzmann machine), DBN (Deep Belief Networks)	Binary classification (AMD classification)	AUC, Accuracy, Duncan test, Delong test, Specificity, Sensitivity	Train VGG-19 with Random Forest, Inception-v3, RBM and DBN with single and multimodality using different combinations of OCT and retinal fundus images. The highest performing models were those trained utilizing multimodal data, as the combination of fundus and OCT boosted the diagnostic result with AUC of 96.9% (0.956–0.979) and 90.5% (89.2–91.8%) accuracy rate, according to Duncan's multiple range test and Delong test.
Vaghefi et al. [10]	Joint Fusion	Optical coherence tomography (OCT), OCT angiography (OCTA), color fundus photography (CFP)	Intermediate dry age-related macular degeneration (AMD)	INCEPTION-RESNET-V2	Multi-class classification (Intermediate dry AMD diagnosis (Old Healthy (OH), Young Healthy (YH), and patients with dry AMD))	Accuracy, Sensitivity, Specificity	Detect intermediate dry AMD by training an INCEPTION-RESNET-V2 on OCT, OCTA, CFP and different combinations of these modalities. The combination of these three modalities resulted in an increase in the model accuracy, which reached 96%, whereas for OCT and OCTA, the accuracy was 94% and 91%, respectively.

Table 2 (continued)

Authors	Fusion method	Modalities	Disease	Techniques	Classification type	Performance Metrics	Findings and results
Jin et al. [11]	Feature-level fusion	OCT, OCT angiography (OCTA)	Neovascular AMD	Feature-level fusion for multiview learning	Assessment of choroidal neovascularization (CNV) in AMD	Accuracy, AUC	Train a feature-level fusion (FLF) model for multiview learning using an attention mechanism. This model combines OCT and OCTA image features for feature extraction. They introduced two FLF methods: the unidirectional fusion network (UFNet) and bidirectional fusion network (BFNet). UFNet comprises feature pathways, assistant modules, and residual fusion. Both the principal pathway and the assistant pathway utilize ResNet-50 as the backbone. The assistant module served to connect the main road and the auxiliary road, ensuring consistency in the fusion process. The model was trained on internal and external datasets, and scored an accuracy of 100% over the Ningbo dataset.
El-Ateif et al. [13]	Joint Fusion	Fundus, Preprocessed fundus (Weighted Gaussian Blur Fundus: WGBF)	Diabetic Retinopathy	VGG19, ResNet50V2, DenseNet121, InceptionV3, Inception-ResNetV2, Xception, and MobileNetV2	Binary classification (non-referable DR, referable DR)	AUC, Accuracy, Sensitivity, Specificity, F1-score, Precision	Train and evaluate seven deep learning CNN models using the joint fusion approach with the fundus and preprocessed fundus (WGBF) as multimodality from the APTOS 2019 [53] and Messidor-2 [49, 52] datasets. The best model is Joint Fusion VGG19, with an accuracy of 97.49% for APTOS 2019 and an accuracy of 91.20% for Messsidor-2. Find that the Joint Fusion multimodality models outperformed the single-modality models.

Table 2 (continued)

Authors	Fusion method	Modalities	Disease	Techniques	Classification type	Performance Metrics	Findings and results
Tseng et al. [12]	Early and Late Fusion	Lesion information, Disease severity	Diabetic Retinopathy	Late fusion, two stage early fusion	Multi-class classification (five different stages of DR)	Accuracy, QWK	Propose two DL architectures: late fusion and two stage early fusion. The late fusion using a post-processing process combines lesion and severity classification models in parallel. Two stage early fusion combines lesion detection and classification models sequentially lesion detection and classification models. The models were trained on three Taiwanese hospital datasets and evaluated on Messidor-2. The hybrid architecture scored an accuracy of 84.29%.
Hervella et al. [14]	No specific fusion approach	Fundus (color retinography), fluorescein angiography	Diabetic Retinopathy	Multimodal image encoding (MIE)	Multi-class classification (five different stages of DR)	QWK, Average Classification Accuracy (ACA), Accuracy, DR AUC, referable DR AUC	Pre-train a neural network using the proposed MIE with the unlabeled multimodal retinal image pairs: color retinography, and fluorescein angiography. Then fine-tune the pre-trained neural network using labeled color retinography images and supervised approaches. After these two steps, the neural network using color retinography images predicts the DR grade. The model scored an accuracy of 65.05% on the IDRiD dataset [55].

Table 2 (continued)

Authors	Fusion method	Modalities	Disease	Techniques	Classification type	Performance Metrics	Findings and results
Liu et al. [42]	Early Fusion	Infrared, OCT	Visual impairment in retinitis pigmentosa	ResNet-152	Binary classification (Snellen better than 20/40, Snellen worse than 20/40)	AUC, Precision, Recall	Train ResNet-152 for classification of Snellen better or worse than 20/40 using the combined infrared and OCT modalities. The model scored 85% in AUC on both their internal Johns Hopkins University (JHU) dataset and the external Amsterdam University Medical Centers (AUMC) dataset. Demonstrate that the best results were obtained using the OCT modality, which were better than the combination with infrared, as the model tended to pay more attention to the OCT modality.
An et al. [15]	Early Fusion	Optical Coherence Tomography and Color Fundus Images	Glaucoma	VGG-19, RF	Binary classification of Glaucoma (Healthy, Glaucomatous)	AUC	Train five VGG-19 models on specific regions from fundus and OCT images using transfer learning (i.e., (1) fundus image of the optic disc in grayscale format, (2) disc retinal nerve fiber layer (RNFL) thickness map, (3) macular ganglion cell complex (GCC) thickness map, (4) disc RNFL deviation map, and (5) macular GCC deviation map). The outputs of these five models were then fed into a random forest (RF) model to distinguish between healthy and glaucomatous disc fundus images. Compared to mono-modality trained models, the RF integrating the five independent trained models enhanced the 10-fold cross validation AUC to 96.3%.

Table 2 (continued)

Authors	Fusion method	Modalities	Disease	Techniques	Classification type	Performance Metrics	Findings and results
Lee et al. [16]	Joint Fusion	Fundus, Extracted optic disk	Glaucoma	DesneNet121, U-net	Binary classification (referable and non-referable glaucoma)	Area under the receiver operator characteristics curve (AUROC), Specificity, Sensitivity, Cohen kappa	Train U-Net combined with DenseNet121 to segment the optic disk from the fundus images using weakly supervised learning. Then, train multimodal network comprised of two DenseNet121 models, concatenated at the dropout level, and takes as input the fundus and extracted optic disk. The model scored 76.35% for the partial AUROC, 61.25% in sensitivity at 95% specificity, 53.16% in Cohen's Kappa, and 80.57 in AUROC for the ungradable prediction.

Fig. 1 Sample images from the diseases represented in the DRH dataset (DRIVE, STARE and HRF)

Fluorescein Angiogram Photographs of Diabetic Patients (FFA) [27] and Project Macula [28]. Note that during the training of our models, we combined the three datasets: DRIVE, HRF and STARE and obtained only one dataset, which we referred to as DHS. Therefore, DHS contains 84 images described by two modalities: fundus and vessel segmentation. As for the FFA and Project Macula datasets (i.e., Macula), they were used as independent datasets. Figures 1, 2 and 3 shows image samples of the diseases available in the five datasets, with their description when provided.

DHS dataset [24–26]: constituted of three datasets which are the Digital Retinal Images for Vessel Extraction (DRIVE) [24], STructured Analysis of the Retina (STARE) [25] and High-Resolution Fundus (HRF) Image Database [26] containing two diagnostic modalities: fundus and blood vessel segmentation and contains all together 84 images per modality. Hereafter, the details of each dataset.

(1) **DRIVE dataset** [25]: covers 400 diabetic patients aged 25 to 90 years old who were screened for diabetic retinopathy in the Netherlands. Forty images were chosen at random, 33 of which show no evidence of DR and seven of which show signs of mild diabetic retinopathy. These photographs were taken with a Canon CR5 non-mydriatic 3CCD camera using a 45° field of view (FOV). The dataset comprises of two imaging modalities: color fundus photograph and retinal vessel segmentation images (Fig. 1). We selected 20 images, both from fundus and vessel segmentation, from the provided

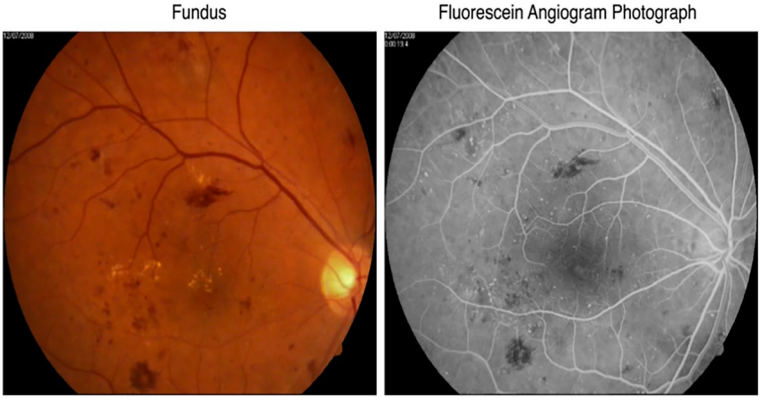

Fig. 2 FFA diabetic retinopathy example

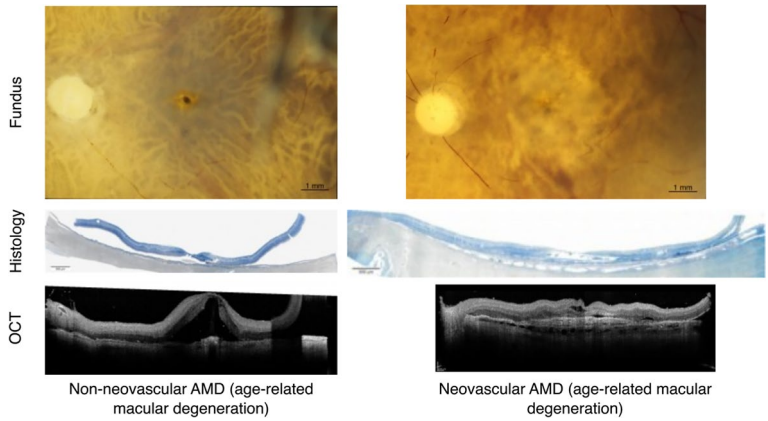

Fig. 3 Sample images from the diseases represented in the Macula dataset

training set. Three of these images were classified as diseased and the remaining 17 images as normal.

(2) **HRF dataset** [26]: is a public database for retinal fundus that was created by a collaborative group to assist comparative studies on automatic segmentation techniques. It includes 15 photos from each of the three groups: healthy, diabetic retinopathy, and glaucoma sufferers (Fig. 1). The photos were taken with a Canon CR-1 fundus camera with a 45-degree field of view and various acquisition settings. We kept all the 15 images from the healthy, diabetic retinopathy and glaucomatous diagnosed patients, which led to 45 images in total for each modality. We used the two modalities: fundus and vessel segmentation to train the models.

(3) **STARE dataset** [27]: is a publicly available dataset made available by the Shiley Eye Center at the University of California, San Diego, and by the Veterans Administration Medical Center in San Diego. It provides two professionally hand-labeled 400 raw fundus images and 40 hand-labeled photos of blood vessel segmentation (Fig. 1). For the vessel segmentation modality, we used the ones provided by the expert Adam Hoover and selected only images (i.e., fundus and vessels) that were 100% normal from the normal batch (i.e., 9 normal images) and 10 images from the ones suffering eye diseases that had corresponding vessel segmentation.

FFA dataset [27] contains fundus fluorescein angiography pictures of diabetic patients from a study at the Persian Eye Clinic at Feiz Hospital at Isfahan University of Medical Sciences. According to the International Clinical Diabetic Retinopathy Disease Severity Scale, the dataset contains retinal images of 70 patients with various stages of diabetic retinopathy: 30 normal stages and 40 abnormal stages (i.e. mild non-proliferative diabetic retinopathy (NPDR), moderate NPDR, severe NPDR, and proliferative diabetic retinopathy (PDR)) [56] (Fig. 2). We removed the 30th image from the data because it did not have the FFAP modality and used the remaining 69 images from both fundus and FFAP modalities.

Project Macula dataset [28] or Macula is a collection of annotated eyes at various phases of AMD pathology, from early to late AMD, as determined by a pathohistological study. The dataset includes three categories: normal (50 eyes), non-neovascular AMD (19 eyes), and neovascular AMD (19 eyes) (40 eyes) (Fig. 3). The fundus images were personally inspected and

preprocessed by two ophthalmologists. We picked the images which had all three modalities available (i.e., fundus, OCT and histology images) from all the three categories and ended up with 80 images classified into two classes normal and diseased from each modality.

4.2 Data oversampling

As the number of data points collected from each of the three datasets used is low, we performed oversampling by augmenting the data using several transformations [5]: horizontal flip, rotation with ranges of [1°,10°], shear with ranges of [0.1°, 2.0°], channel shift with ranges of [120, 170] and finally a brightness change with ranges of [0.1, 0.9]. All these transformations were implemented using the image data generator method from the TensorFlow v2.0 (TF) library. Note that, apart from horizontal flip and brightness change, all the range values were looped over. This process magnified the datasets sample sizes by 80, and we obtained for each modality of the datasets DHS, FFA and Macula datasets 6720, 4640 and 5520 images respectively.

4.3 Data preprocessing

The images were downsized to 224×224 pixels and preprocessed using the preprocess input function from the TF library according to each ImageNet-trained DL model during the preparation steps. After the images were converted from RGB to BGR, each color channel was zero-centered without scaling with respect to the ImageNet dataset for the VGG19 model. The pixel values were sample-wise scaled between -1 and 1 for the models that use Batch Normalization (i.e., DenseNet121, InceptionV3, InceptionResNetV2, Xception, ResNet50V2, MobileNetV2).

5 Experimental setup

The experimental setup for the current investigation is described in this section. As a result, we begin by describing the assessment metrics that were utilized to assess the DL models. Second, we show how we used the Scott Knott ESD statistical test (Borda Count method) to group the DL techniques based on their accuracy (to rank the best SK-ESD techniques) using precision, AUC, sensitivity, specificity, and F1-score. Third, we describe the experimental procedure we used to conduct all of the empirical assessments. Finally, the abbreviations we utilized to shorten the names of the DL approaches and modalities are shown.

5.1 Evaluation metrics

To evaluate the performance of the DL classifiers, we used six metrics: accuracy, AUC score, sensitivity, specificity, precision, and F1-score. Formulae defining the six metrics are given by Eqs. 1 to 5 respectively.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{F1} = 2\frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{5}$$

where: TP: diseased case identified as diseased. FP: diseased case identified as normal. TN: normal case identified as normal, and FN: normal identified as diseased.

5.2 SK-ESD and Borda count

This section introduces the SK-ESD statistical test and Borda Count approach used to compare and rank the performance of the seven DL techniques.

Scott-Knott effect size difference (SK-ESD) [29] As described by Algorithm 1, it is a statistical test that divides a set of treatment averages (e.g., means) into statistically separate groups with non-negligible differences by using hierarchical clustering. It is a Scott-Knott variation that takes into account the magnitude of the difference in treatment means within and between groups (i.e., effect size). The test's version two plus creates a ranking of treatment means while assuring that the size of the difference is: (1) insignificant for all treatments within each group; and (2) non-negligible for all treatments across groups. The input to the SK-ESD algorithm is the accuracy values of the 5-fold cross validation (CV) (5 accuracies results for the 5 different data splits) results of the seven DL models for each modality, fusion, and dataset. The output consists of the ranks of these seven DL models.

Borda count [30] As explained by Algorithm 2, it is a voting technique for selecting a winner from a group of contestants. It entails assigning points to a group of applicants depending on their ranking: 1 point for last option, 2 points for second-to-last choice, and so on until the top is achieved. After that, the candidates' point totals are added together, and the candidate with the highest total point wins. The candidates in the present study are the models belonging to the best SK-ESD cluster. The voters are the five performance indicators: precision, AUC, sensitivity, specificity, and F1-score.

Algorithm 1 Scott-Knott Effect Size Difference (SK-ESD)

Input: a list of n effect sizes (ES), where each ES is a real number
Output: a list of k clusters, where each cluster is a sub list of the input ES list, and the clusters are sorted in decreasing order of statistical significance

1. Sort the input ES list in decreasing order.
2. Set k to 1 and initialize the first cluster to be the entire input ES list.
3. Compute the p-value of the ANOVA test for the current cluster. Let the p-value be p_0.
4. For i in [2, n-1]:
   a. Move the i-th ES from the current cluster to a new cluster.
   b. Compute the p-value of the ANOVA test for the current cluster and the new cluster combined. Let the p-value be p_i.
   c. If p_i < p_0, update the current cluster to be the combined cluster and increment k by 1. Otherwise, leave the i-th ES in the current cluster and continue to the next iteration.
5. Return the k clusters sorted in decreasing order of statistical significance, where the statistical significance is measured by the p-value of the ANOVA test for each cluster.

Algorithm 2 Borda Count

Input: a list of m candidates, C = [c_1, c_2, ..., c_m], and a list of n voters, V = [v_1, v_2, ..., v_n], where each voter ranks the candidates from most preferred to least preferred.
Output: a ranked list of the candidates based on the Borda Count method.

1. Initialize an empty dictionary, S, to store the score of each candidate.
2. For each voter v_i in V:
   a. Assign m points to their most preferred candidate in C, m-1 points to their second-most preferred candidate, and so on down to 1 point for their least preferred candidate.
   b. For each candidate c_j in C, add the points assigned to c_j by v_i to the current score of c_j in S.
3. Sort the candidates in S in decreasing order of their score.
4. Return the ranked list of candidates based on the order in which they appear in the sorted list.

5.3 Training and testing processes

We choose seven ImageNet-trained DL models (DenseNet121, InceptionV3, Inception-ResNetV2, MobileNetV2, ResNet50V2, VGG19, and Xception), and we train them using the three oversampled datasets. Apart from VGG19 model, all of the six remaining models used a batch normalization (BN). For simplicity, batch normalization models (BNM) refer to: DenseNet121, InceptionV3, InceptionResNetV2, MobileNetV2, ResNet50V2, and Xception. Figure 4 briefly presents the architectures of these seven DL models and their classification layers. The models were trained for 100 epochs each using both the early stopping and reduce learning rate on plateau methods from the TF library as detailed in Table 3, including dropout with probability $p=0.5$ in the classification layer of the models to avoid overfitting. As our data was oversampled, no further data augmentation was performed during training. Furthermore, we used the binary cross entropy loss and a batch size of 32. The models were evaluated on the test set. We follow the mono-modality (with phase I and II) and late fusion processes to obtain the mono-modality and late fusion results, respectively, as explained in Fig. 5. In the following, we define the early stopping, reduce learning rate on plateau, and 5-fold cross validation methods; moreover, we present the mono-modality, and late fusion training processes.

Early stopping [57]: a method used to regulate deep neural networks, which involves stopping the training process when further updates to the model parameters no longer result in significant improvements on a validation set. During training, the current best parameters are stored and updated; once updates no longer lead to improvement within a specified number of iterations, the training process is stopped and the most recent best parameters are used. This technique effectively limits the optimization process to a smaller range of parameter space, resulting in a regularization effect.

Reduce learning rate on plateau [58]: a method commonly used as a technique for optimizing deep neural networks during training, and it involves reducing the learning rate when the validation loss fails to improve for a certain number of epochs.

5-fold cross validation (5-fold CV) [59]: or k-fold cross validation with $k=5$, an approach that involves dividing the dataset into five equal-sized subsets, and training and testing the model five times, using a different subset as the validation set each time. This technique is widely used in machine learning to evaluate a model's performance and reduce the risk of overfitting.

Mono-modality process for each modality and each dataset (Fig. 5 –Mono-modality):

- For the VGG19 model: (1) Load the model from the TF library using ImageNet weights. (2) Freeze all the layers apart from the top convolutional (conv.) layer to improve eye diseases features representation and avoid overfitting by unfreezing more layers. (3) Add the classification layer to the VGG19 architecture (Fig. 4f), and use the

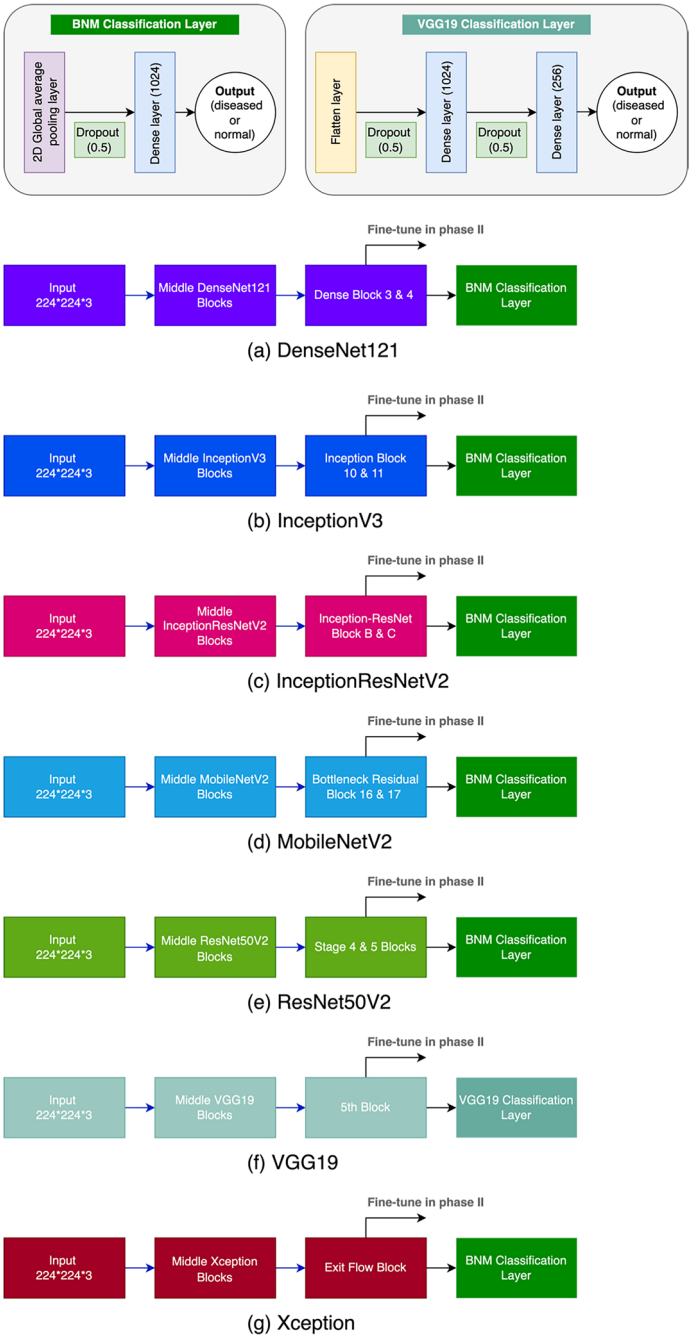

Fig. 4 Compressed architectures of the seven DL models: (**a**) DenseNet121, (**b**) InceptionV3, (**c**) Inception-ResNetV2, (**d**) MobileNetV2, (**e**) ResNet50V2, (**f**) VGG19, and (**g**) Xception

Table 3 Early stopping and reduce learning rate on plateau parameters

Technique	Parameter	Value
Early stopping	monitor	validation loss
	mode	minimum
	patience	10
Reduce learning rate on plateau	monitor	validation loss
	factor	0.6
	patience	8
	verbose	1
	mode	minimum
	minimum learning rate	1e-4

Adam optimizer with 1e-6 as learning rate. (4) Split the data into 60% train, 20% validation, and 20% test, and use the 5-fold CV approach to train the model. And (5) report the results from assessing the model's performance on the test set for every fold.

- For the BNM (DenseNet121, InceptionV3, InceptionResNetV2, MobileNetV2, ResNet50V2, and Xception):

For Phase I, pre-training: (1) Load M_i with $i = \{1, \ldots, 6\}$ from M_bn with ImageNet weights, with $M\_bn = \{M\_1, \ldots, M\_6\}$ and represent the six DL models: DenseNet121, InceptionV3, InceptionResNetV2, Xception, ResNet50V2, and MobileNetV2. (2) Add the BMN classification layer (Fig. 4a-e, g), and use the Root Mean Square Propagation optimizer (RMSProp) with the default learning rate. And (3) Split the data to 80% train, and 20% test using the stratified by class method to ensure that each subset contains a proportionate representation of the diseased and normal classes.

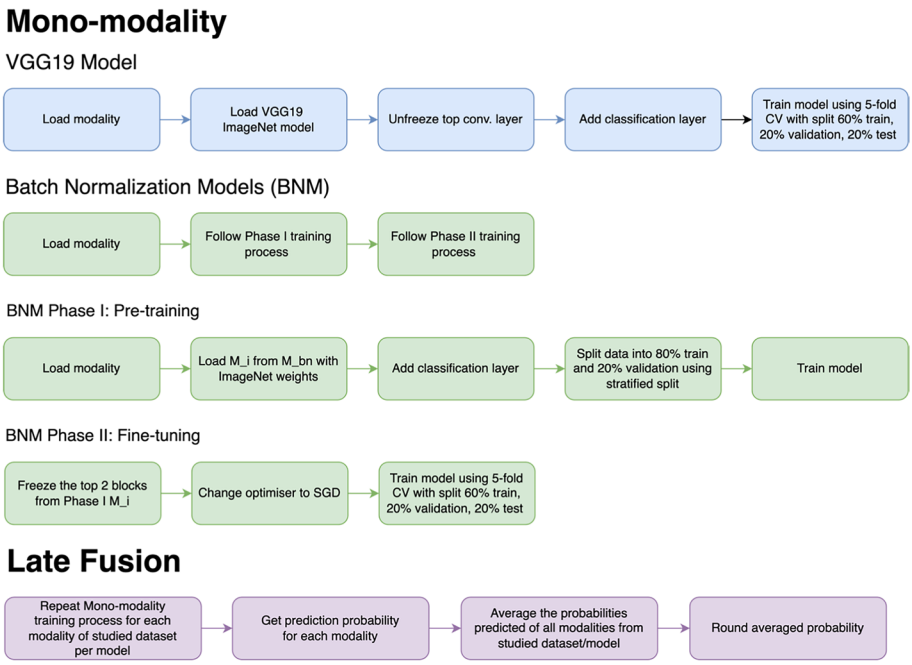

Fig. 5 Mono-modality and late fusion DL models training processes

For Phase II, fine-tuning: (1) Based on best practices for BNM fine-tunning, unfreeze the top two blocks of M_i (Fig. 4a-e, g). (2) Change the RMSProp optimizer with the Stochastic Gradient Descent (SGD) with momentum [60] with a learning rate of 1e-4 and a momentum of 0.9. (3) Split the data into 60% train, 20% validation, and 20% test, and use the 5-fold CV approach to train the model. And (4) report the results of the model's performance on the test set for every fold.

Late fusion process for each dataset (Fig. 5 -Late fusion): (1) Repeat the Mono-modality training process for all the seven DL models and on the three datasets. (2) Obtain the prediction probability for each modality-dataset respective pair. (3) Sum all of the prediction probabilities per dataset, then divide them by N, where N is the number of modalities available in the dataset. And (4) round the result to get the final prediction (i.e., normal or diseased) for each patient of the testing dataset. Note that we used the same model for each modality in the dataset. For example, the DHS dataset contains two modalities: fundus and vessels, and we proceed as follows: (i) Train the VGG19 model following the mono-modality process for each modality. (ii) Obtain the predicted probabilities for the fundus (p_f) and vessels (p_v). (iii) Sum p_f and p_v to obtain the VGG19 late fusion prediction (iv) Divide the resulting sum with 2 (2 is the number of modalities in the DHS dataset). And (v) round the average to get 1 (if the prediction probability is higher than 0.5) and 0 (if the prediction probability is strictly lower than 0.5) to determine if patient is diseased or normal.

The models were implemented using the TensorFlow v2.0 library and Keras API on Colab using its GPU for training and CPU for data preparation. As for metrics calculation, we used Scikit-learn v0.24 library to evaluate the performance of the trained models. As a side note, Colaboratory, or "Colab" for short, is a free platform that provides cloud instances to train deep learning models using GPU or TPU. Scikit-learn is a machine learning library with various classification, regression and clustering algorithms programmed with the python language.

5.4 Empirical process

We used a three-step empirical approach to carry out the empirical assessments [31] that consists of:

(1) Assessing the performances in terms of (accuracy, AUC, precision, sensitivity, specificity, F1-score) of the deep learning architectures with mono-modality and late fusion using a 5-fold CV
(2) Clustering the DL techniques using SK-ESD based on accuracy to select the best models in the first cluster
(3) Ranking the DL models of the best SK-ESD cluster using the Borda Count method based on the five performance measures (precision, sensitivity, specificity, F1-score, and AUC) and select the top deep learning model(s)

5.5 Abbreviation

To assist the reader and shorten the names of the deep learning techniques used, we follow the subsequent naming rules. We abbreviate the name of each DL technique as provided in Table 13 of the appendix A for each mono-modality trained models. As for the

comparison of the performances of a DL model on all of the modalities in each dataset, we used the abbreviation of the model coupled with a letter from the modality as shown in Table 14 of the Appendix A. Additionally, we add to the abbreviation of the models the "LF" that stands for late fusion when comparing the mono-modality models with the multimodality ones.

6 Results and discussion

In the following we present and discuss the results of the empirical evaluations of the seven DL architectures: VGG19, DenseNet121, InceptionV3, InceptionResNetV2, Xception, ResNet50V2, MobileNetV2, trained on three datasets: DHS, FFA and Macula. This section is structured in order to address RQ1, RQ2, RQ3 and RQ4 as follows:

- *(RQ1):* Using the accuracy, AUC, sensitivity, specificity, precision and F1-score we evaluate the performances of the DL models. For the mono-modality models, we first run the SK-ESD statistical test to cluster the DL models trained per modality for each dataset. Then, we select the best cluster per modality for each dataset. Furthermore, we compute the Borda Count ranks of the DL models belonging to the best SK-ESD cluster.

  To assess the performance of each DL technique across the available modalities, we calculate the sum of its obtained SK-ESD ranks (SR). To assess the sensitivity of each DL technique across the available modalities, we calculate the sum of the differences of the ranks (SDR) of each DL technique across modalities. This allows to highlight the most and less accurate/sensitive DL techniques according to existing modalities.
- *(RQ2):* Using the SK-ESD statistical test and Borda Count we evaluate and discuss the impact of each modality on the diagnosis performance of a DL architecture. At first, we use SK-ESD with accuracy values to cluster all the combinations of modalities and DL techniques for each dataset. Then, we use Borda Count to rank modality-technique combinations part of the best cluster.
- *(RQ3):* As for multimodality models, we use the same process as of the mono-modality models (RQ1) to evaluate and compare the late fusion models.
- *(RQ4):* We perform a comparison between mono-modality and multimodality models using SK-ESD test and Borda Count based on the six performance criteria to figure out what models are best to diagnose eye diseases.

6.1 Evaluating and comparing mono-modality techniques (RQ1)

Table 4 reports the mean values of the 5-fold cross validation six metrics (sensitivity, specificity, precision, F1-score, AUC and accuracy) of the seven DL techniques using each modality of the three datasets (DHS, Macula and FFA). Figures 6, 7 and 8 show the SK-ESD results based on accuracy for Macula, DHS and FFA datasets respectively. Moreover, Table 5 shows the Borda Count ranks based on specificity, sensitivity, precision, AUC and F1-score of the DL models belonging to the best cluster of SK-ESD of the three datasets (DHS, Macula, and FFA). As for Tables 6, 7 and 8, they present the

Table 4 Mean metrics results on the test set of the 5-fold CV for each model on each modality from all three datasets (Macula, DHS, FFA)

Dataset	Modality	Model	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)	AUC (%)	Accuracy (%)
Macula	Fundus	Xception	99.92	100	100	99.96	99.96	99.98
		VGG19	99.02	99.98	99.93	99.47	99.5	99.78
		InceptionResNetV2	99.86	100	100	99.93	99.93	99.97
		DenseNet121	99.93	100	100	99.96	99.96	99.98
		InceptionV3	100	100	100	100	100	100
		ResNet50V2	99.93	100	100	99.96	99.96	99.98
		MobileNetV2	100	100	100	100	100	100
	Histology	Xception	99.28	100	100	99.64	99.64	99.84
		VGG19	98.36	99.98	99.93	99.14	99.17	99.64
		InceptionResNetV2	99.71	99.98	99.93	99.82	99.85	99.92
		DenseNet121	99.64	100	100	99.82	99.82	99.92
		InceptionV3	99.86	99.98	99.93	99.89	99.92	99.95
		ResNet50V2	99.86	100	100	99.93	99.93	99.97
		MobileNetV2	99.78	100	100	99.89	99.89	99.95
	OCT	Xception	99.71	100	100	99.86	99.86	99.94
		VGG19	99.07	100	100	99.53	99.54	99.8
		InceptionResNetV2	99.79	100	100	99.9	99.9	99.95
		DenseNet121	99.93	100	100	99.97	99.97	99.98
		InceptionV3	99.67	99.98	99.92	99.8	99.83	99.91
		ResNet50V2	99.72	100	100	99.86	99.86	99.94
		MobileNetV2	99.7	100	100	99.85	99.85	99.94
DHS	Fundus	Xception	99.97	100	100	99.98	99.98	99.99
		VGG19	99.94	99.97	99.97	99.95	99.96	99.96
		InceptionResNetV2	99.97	100	100	99.98	99.98	99.99
		DenseNet121	99.94	99.97	99.97	99.95	99.96	99.96

Table 4 (continued)

Dataset	Modality	Model	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)	AUC (%)	Accuracy (%)
		InceptionV3	99.94	99.94	99.94	99.94	99.94	99.94
		ResNet50V2	100	99.97	99.97	99.98	99.99	99.99
		MobileNetV2	100	99.86	99.85	99.92	99.93	99.93
	Vessels	Xception	99.7	99.94	99.94	99.82	99.82	99.82
		VGG19	99.85	99.94	99.94	99.89	99.9	99.9
		InceptionResNetV2	100	99.65	99.64	99.82	99.82	99.82
		DenseNet121	99.76	99.62	99.6	99.68	99.69	99.69
		InceptionV3	99.97	99.91	99.91	99.94	99.94	99.94
		ResNet50V2	99.97	99.97	99.97	99.97	99.97	99.97
		MobileNetV2	99.94	99.8	99.78	99.86	99.87	99.87
FFA	Fundus	Xception	98.26	97.87	97.92	98.09	98.06	98.08
		VGG19	98.81	98.59	98.61	98.71	98.7	98.69
		InceptionResNetV2	97.69	99.19	99.17	98.42	98.44	98.43
		DenseNet121	98.81	98.76	98.75	98.78	98.78	98.77
		InceptionV3	98.62	98.24	98.25	98.44	98.43	98.43
		ResNet50V2	99.06	97.78	97.81	98.44	98.42	98.41
		MobileNetV2	97.98	98.67	98.66	98.32	98.33	98.32
	FFAP	Xception	97.68	97.92	98.05	97.86	97.8	97.84
		VGG19	98.88	98.74	98.76	98.82	98.81	98.81
		InceptionResNetV2	98.28	99.36	99.34	98.81	98.82	98.81

Table 4 (continued)

Dataset	Modality	Model	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)	AUC (%)	Accuracy (%)
		DenseNet121	98.89	98.45	98.45	98.67	98.67	98.66
		InceptionV3	98.46	98.49	98.5	98.48	98.47	98.47
		ResNet50V2	97.41	99.13	99.13	98.26	98.27	98.28
		MobileNetV2	99.06	97.07	97.24	98.14	98.07	98.06

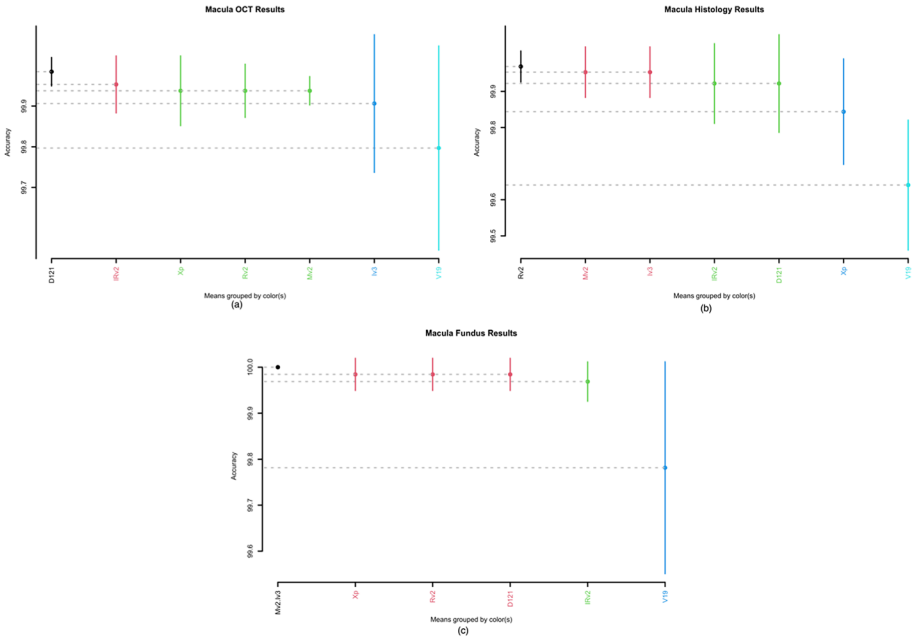

Fig. 6 Scott-Knott ESD results of mono-modality DL techniques over the Macula dataset for the (**a**) OCT, (**b**) histology and (**c**) the fundus modalities respectively

ranks of each DL technique according to each modality and the values of SR and SDR of Macula, DHS and FFA respectively. We observe that:

- For the Macula dataset, Fig. 6a-c shows that: (1) for the OCT and histology modalities, we obtained five clusters while for the fundus modality we only obtained four clusters. (2) The best cluster of the OCT modality only contains one technique: DenseNet121 with an accuracy of 99.98% as reported in Table 4. (3) The best cluster of the histology modality contains one technique: ResNet50V2 with an accuracy of 99.97%. And (4) the best cluster of the fundus modality, contains two DL

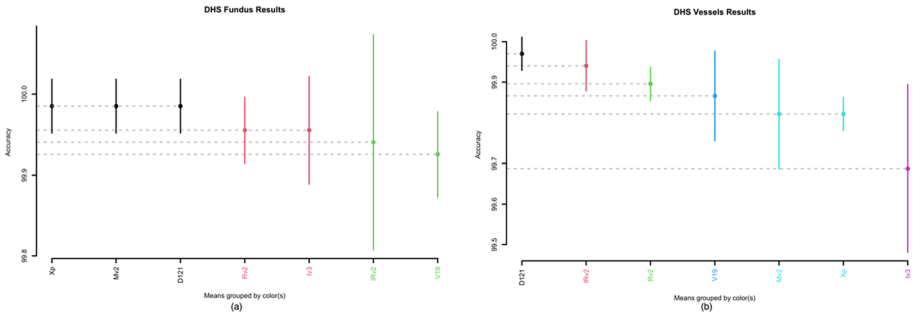

Fig. 7 Scott-Knott ESD results of mono-modality DL techniques over DHS dataset for the (**a**) fundus and (**b**) vessels modalities respectively

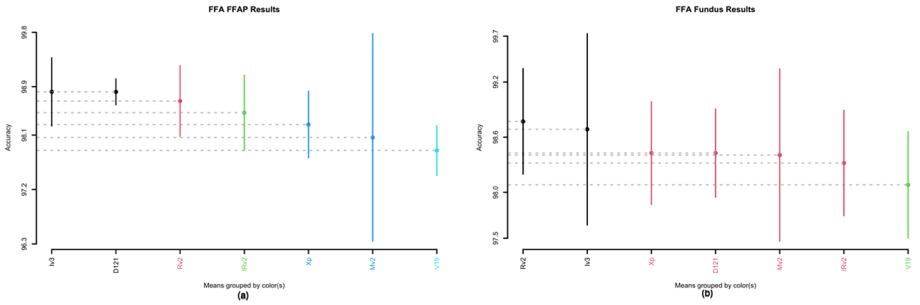

Fig. 8 Scott-Knott ESD results of mono-modality DL techniques over FFA dataset for the (**a**) FFAP and (**b**) vessels modalities respectively

techniques: MobileNetV2 and InceptionV3 with an accuracy of 100% each (with 100% present in all of the six reported metrics). However, as shown in Table 5, both MobileNetV2 and InceptionV3 were ranked first according to Borda Count for the fundus modality. Note that VGG19 was the worst DL technique since it belongs to the last cluster of all modalities. In order to compare the seven DL techniques taking into consideration the three modalities, Table 6 presents: (1) the sum of the SK-ESD ranks (SR) of each DL technique across the three modalities: for each DL technique and each modality, a technique has a score equal to its SK-ESD rank. Thereafter, the total score SR of each technique across the three modalities is the sum of its individual scores (i.e., score according to each modality). The total score SR of each DL technique determines its performance across modalities (i.e., the lower the total score the higher is the performance); and (2) the sum of the differences of the ranks (SDR) of each DL technique across modalities in order to evaluate its sensitivity to modalities (i.e., the lower the differences of ranks the lower is its sensitivity). From Table 6, we notice that MobileNetV2, DenseNet121 and ResNet50V2 have the lowest total score of SR (SR = 6); VGG19 has the highest total score of SR (SR = 14). For the total score SDR, VGG19 and InceptionResNetV2 have the low-

Table 5 Ranking of Borda Count of the models in the best Scott Knott ESD cluster for mono-modality models for each dataset

Dataset	Modality	Model	Borda Count rank
Macula	Fundus	MobileNetV2	1
		InceptionV3	1
	Histology	ResNet50V2	1
	OCT	DenseNet121	1
DHS	Fundus	Xception	2
		MobileNetV2	1
		DenseNet121	1
	Vessels	DenseNet121	1
FFA	Fundus	InceptionV3	2
		ResNet50V2	1
	FFAP	InceptionV3	2
		DenseNet121	1

Table 6 Statistics on Scott-Knott ESD ranks for the Macula dataset

Model	Scott-Knott ESD rank			Sum of SK-ESD ranks (SR)	Sum of SK-ESD ranks differences (SDR)
	Macula				
	Fundus	Histology	OCT		
Xception	2	4	3	9	4
MobileNetV2	1	2	3	6	4
DenseNet121	2	3	1	6	4
ResNet50V2	2	1	3	6	4
InceptionV3	1	2	4	7	6
InceptionResNetV2	3	3	2	8	2
VGG19	4	5	5	14	2

est one (SDR = 2). However, MobileNetV2, DenseNet121 and ResNet50V2 have an SDR score equals to 4 and were ranked first, second and third interchangeably over the three modalities. We therefore conclude that these three DL techniques are the best in terms of performance and sensitivity.

- For the DHS dataset, Fig. 7a-b shows that three clusters and six clusters were obtained for the fundus and vessels modalities respectively. The SK-ESD best cluster of the fundus modality contains three DL techniques: Xception (accuracy = 99.99%), MobileNetV2 (accuracy = 99.93%), and DenseNet121 (accuracy = 99.96%), while the best cluster of the vessel's modality contains only one DL technique: DenseNet121 (accuracy = 99,69%). Moreover, Table 5 shows that MobileNetV2 and DenseNet121 were ranked first according to Borda Count for the fundus modality. In order to compare the seven DL techniques taking into consideration the DHS modalities, Table 7 presents SK-ESD ranks of each DL techniques according to each modality, the SR and SDR values. We notice that DenseNet121 was the best DL technique in terms of both performance and sensitivity (SR = 2 and SDR = 0).
- For the FFA dataset, Fig. 8a-b shows that five clusters and three clusters were obtained for the FFAP and fundus modalities respectively. The SK-ESD best cluster of the FFAP contains two DL techniques: DenseNet121 and InceptionV3 with accuracy values equal

Table 7 Statistics on Scott-Knott ESD ranks of the DHS dataset

Model	Scott-Knott ESD rank		Sum of SK-ESD ranks (SR)	Sum of SK-ESD ranks differences (SDR)
	DHS			
	Fundus	Vessels		
Xception	1	5	6	4
MobileNetV2	1	5	6	4
DenseNet121	1	1	2	0
ResNet50V2	2	3	5	1
InceptionV3	2	6	8	4
InceptionResNetV2	3	2	5	1
VGG19	3	4	7	1

to 98.66% and 98.47% respectively; while the SK-ESD best cluster of the fundus modality contains two DL techniques InceptionV3 and ResNet50 with accuracy values equal to 98.43% and 98.41% respectively. Moreover, Table 5 shows that according to Borda Count DenseNet121 and ResNet50 were ranked first according to FFAP and fundus modalities respectively, and InceptionV3 was ranked second for both modalities. In order to compare the seven DL techniques taking into consideration the DHS modalities, Table 8 presents SK-ESD ranks of each DL techniques according to each modality, the SR and SDR values. We notice that InceptionV3 was the best DL technique in terms of both performance and sensitivity (SR = 2 and SDR = 0); DenseNet121 and ResNet50 were the second-best DL techniques in terms of performance and sensitivity (SR = 3 and SDR = 1).

To sum up: (1) the best DL technique in terms of SR and SDR values is DenseNet121 as it obtained optimum scores of SR and SDR in all the three datasets; (2) ResNet50V2 is the second as it obtained optimum scores in Macula and FFA datasets, but it has performed less in the DHS dataset; and (3) MobileNetV2 and InceptionV3 only performed better in the Macula and FFA datasets respectively.

6.2 Impacts of modalities on the performances of DL techniques (RQ2)

This section evaluates and discusses the impact of each modality on the diagnosis performance of a DL architecture. For this purpose, we used SK-ESD statistical test based on accuracy values to cluster all the combinations of modalities and DL techniques for each dataset. Thereafter, Borda Count was used to rank modality-technique combinations belonging to the best cluster.

Figures 9a-c shows the SK-ESD results based on accuracy for Macula, DHS and FFA datasets respectively. Moreover, Table 9 shows the Borda Count ranks based on specificity, precision, sensitivity, AUC, and F1-score of the modality-technique belonging to the SK-ESD best cluster of the three datasets (Macula, DHS, and FFA). Finally, to determine which imaging modality is cost-effective, we highlight: (1) the cost of each modality involved into the diagnostic of some eye diseases, and (2) the importance of the imaging modalities in diagnosing a certain disease.

From Figs. 9a-b, the fundus modality was the best to positively impact the performances of DL techniques for DED diagnosis as most of the techniques reported in the best clusters

Table 8 Statistics on Scott-Knott ESD ranks of the FFA dataset

Model	Scott-Knott ESD rank		Sum of SK-ESD ranks (SR)	Sum of SK-ESD ranks differences (SDR)
	FFA			
	Fundus	FFAP		
Xception	2	4	6	2
MobileNetV2	2	4	6	2
DenseNet121	2	1	3	1
ResNet50V2	1	2	3	1
InceptionV3	1	1	2	0
InceptionResNetV2	2	3	5	1
VGG19	3	5	8	2

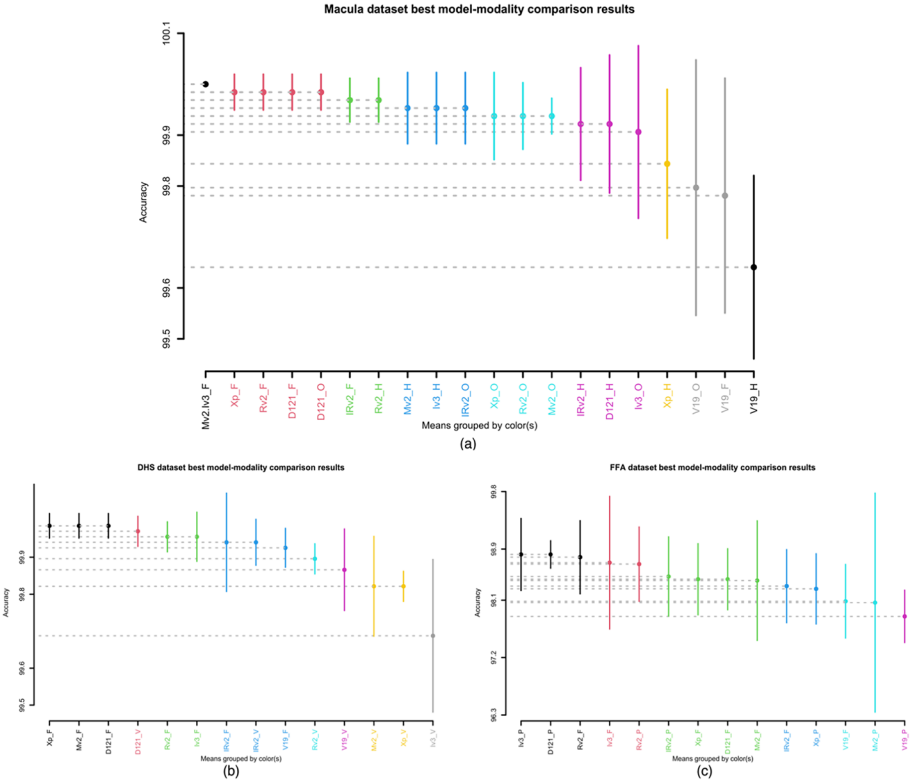

Fig. 9 Scott-Knott ESD results of mono-modality DL techniques over the three datasets: (**a**) Macula, (**b**) DHS and (**c**) FFA

are using fundus images: (1) for the Macula dataset, the models part of the best cluster were MobileNetV2 and InceptionV3 when using the fundus modality, and (2) for the DHS dataset, the three models part of the best cluster were Xception, MobileNetV2 and DenseNet121 when using the fundus modality. They were ranked 2nd, 1st and 1st based on Borda Count as shown in Table 9 respectively.

Table 9 Ranking of Borda Count of the models in the best cluster across modalities for each dataset

Dataset	Modality	Model	Borda Count rank
Macula	Fundus	MobileNetV2	1
		InceptionV3	1
DHS	Fundus	Xception	2
		MobileNetV2	1
		DenseNet121	1
FFA	Fundus	ResNet50V2	3
	FFAP	InceptionV3	2
		DenseNet121	1

As for the FFA dataset, Fig. 9c shows that the best cluster contains three DL techniques: ResNet50V2 with fundus modality and InceptionV3/DenseNet121 with FFAP modality. However, the Borda Count ranking shows that the FFAP modality led DenseNet121 and InceptionV3 to the 1st and 2nd ranks respectively, outperforming ResNet50V2 with the fundus modality.

6.3 Evaluation of multimodality DL late fusion performance (RQ3)

This section discusses the overall performance of multimodality DL late fusion techniques in ED classification. Table 10 presents the mean values of the 5-fold CV six metrics (sensitivity, specificity, precision, F1-score, AUC and accuracy) of the seven multimodality DL techniques for the three datasets (Macula, DHS and FFA). Fig. 10a-c shows the SK-ESD results based on accuracy for Macula, DHS and FFA datasets respectively. Additionally, Table 11 presents the Borda Count ranks based on the remaining five metrics (i.e., excluding accuracy from the six metrics mentioned) of the DL late fusion models part of the SK-ESD best cluster of the three datasets. We observe that:

- For the Macula dataset, Fig. 10a shows that the SK-ESD test provides three clusters and the best one contains the three DL late fusion techniques: ResNet50V2, MobileNetV2, InceptionV3 and InceptionResNetV2 with an accuracy of 100% each (with 100% present in all of the six reported metrics) as presented in Table 10. As shown in Table 11, all of these best techniques were ranked first. VGG19 was the worst performing late fusion technique as it belongs to the last cluster.
- For the DHS dataset, Fig. 10b shows that the SK-ESD test provides six clusters and the best one contains one DL late fusion technique: ResNet50V2 with an accuracy of 100% (with 100% present in all of the six reported metrics). For this dataset as well, the last ranking DL late fusion technique was VGG19.
- For the FFA dataset, Fig. 10c shows that the SK-ESD test provides five clusters and the best one contains one DL late fusion; ResNet50V2 with an accuracy of 100% (with 100% present in all of the six reported metrics). VGG19 was the worst technique as it belongs to the last cluster.

To sum up, ResNet50V2 was the best DL late fusion technique across the three datasets, while VGG19 late fusion was the worst. Additionally, InceptionResNetV2 late fusion appeared in both of the second clusters of the FFA and DHS dataset, while it was ranked first in the Macula dataset. This leads us to conclude that InceptionResNetV2 is the 2nd best late fusion technique. Finally, the DL techniques were most sensitive with the DHS and FFA modalities as the number of clusters produced was higher (six and five clusters respectively) than in the Macula dataset (three clusters).

6.4 Comparison of mono-modality DL techniques and multimodality DL late fusion techniques (RQ4)

This section compares the performances of mono-modality and multimodality late fusion DL techniques. To this aim, for each dataset, we cluster and compare the best

Table 10 Performance results of DL late fusion techniques over the three datasets (Macula, DHS, FFA)

Dataset	Modalities	Model	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)	AUC (%)	Accuracy (%)
Macula	Fundus, Histology, OCT	ResNet50V2	100	100	100	100	100	100
		InceptionV3	100	100	100	100	100	100
		Xception	100	99.98	99.93	99.96	99.99	99.98
		MobileNetV2	100	100	100	100	100	100
		InceptionResNetV2	100	100	100	100	100	100
		VGG19	99.2	96.97	90.4	94.59	98.08	97.42
		DenseNet121	100	99.98	99.93	99.96	99.99	99.98
DHS	Fundus, Vessels	DenseNet121	100	95.57	95.73	97.82	97.79	97.78
		InceptionResNetV2	100	99.3	99.32	99.66	99.65	99.66
		VGG19	100	87.18	88.6	93.96	93.59	93.56
		MobileNetV2	100	95.28	95.43	97.66	97.64	97.61
		Xception	100	99.28	99.26	99.63	99.64	99.63
		ResNet50V2	100	100	100	100	100	100
		InceptionV3	100	96.16	96.31	98.12	98.08	98.08
FFA	Fundus, FFAP	DenseNet121	98.84	93.29	93.35	96.01	96.06	96
		VGG19	98.36	93.02	93.07	95.64	95.69	95.64
		InceptionV3	98.87	93.52	93.57	96.15	96.2	96.13
		InceptionResNetV2	99.69	98.85	98.78	99.24	99.27	99.26
		MobileNetV2	100	93.11	93.26	96.52	96.56	96.47
		ResNet50V2	100	100	100	100	100	100
		Xception	100	95.44	95.44	97.66	97.72	97.66

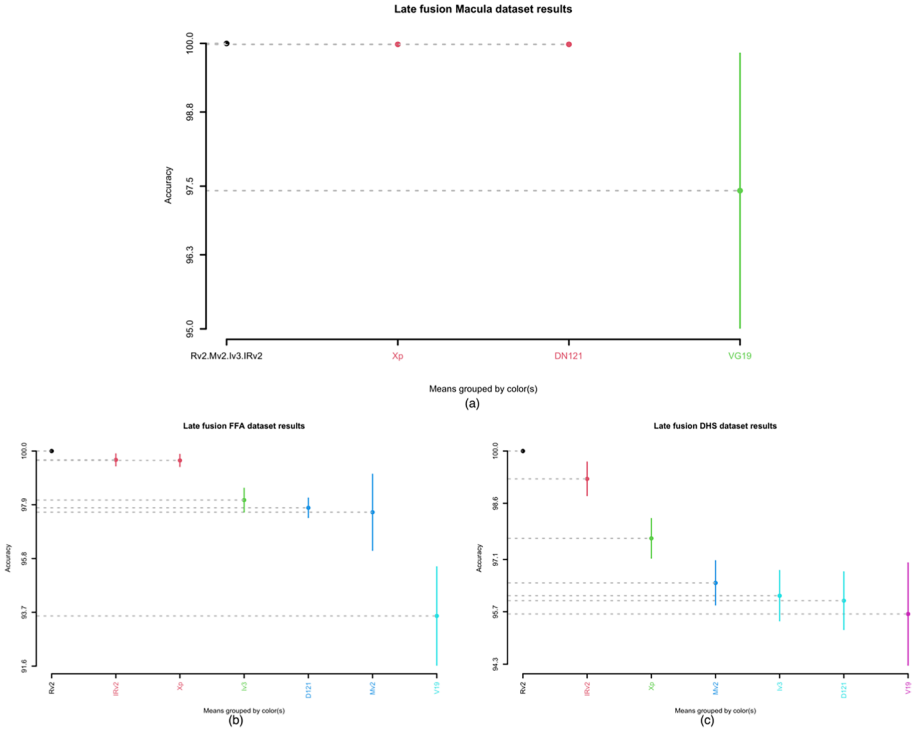

Fig. 10 Scott-Knott ESD results of Late fusion over the three datasets: (**a**) Macula, (**b**) FFA and (**c**) DHS

mono-modality DL techniques (RQ1) with the best multimodality late fusion DL techniques (RQ3) using the SK-ESD test based on accuracy. Figure 11a-f shows the SK-ESD results for the Macula, DHS and FFA datasets respectively. We observe that:

- For the Macula dataset, Fig. 11a shows the 1st cluster with the four best late fusion DL techniques (ResNet50V2, MobileNetV2, InceptionV3 and InceptionResNetV2) and the two best fundus DL techniques (MobileNetV2 and InceptionV3) with an accuracy, sensitivity, specificity, precision, F1-score, and AUC of 100%. Note that we removed

Table 11 Ranking of Borda Count of the late fusion models belonging to the best clusters of the three datasets

Dataset	Modality	Model	Borda Count rank
Macula	Fundus, Histology, OCT	ResNet50V2_LF	1
		InceptionV3_LF	1
		MobileNetV2_LF	1
		InceptionResNetV2_LF	1
DHS	Fundus, Vessels	ResNet50V2_LF	1
FFA	Fundus, FFAP	ResNet50V2_LF	1

InceptionV3 for the fundus from the graph to improve figure readability. The 2nd cluster lists the DenseNet121 OCT model, and the 3rd cluster contains the ResNet50v2 histology model. The listed results show that the late fusion ResNet50V2, MobileNetV2, InceptionV3 and InceptionResNetV2, along with fundus MobileNetV2 and InceptionV3 DL models outperform the best OCT and histology DL models.

- For the DHS dataset, Figs. 11b shows two clusters with the 1st cluster containing late fusion ResNet50v2 outperforming the fundus and vessels best DL models (fundus Xception, fundus DenseNet121, and vessels DenseNet121).
- For the FFA dataset, Fig. 11c shows two clusters: late fusion ResNet50V2 belongs to the 1st cluster while FFAP InceptionV3, DenseNet121, and fundus ResNet50V2 with InceptionV3 belong to the 2nd last cluster. Thus, late fusion ResNet50V2 outperformed mono-modality DL models over the FFA dataset.

To sum up, DL late fusion outperformed mono-modality DL techniques whatever the modality of the three datasets.

6.5 Comparison of best model(s) with state-of-the-art (RQ5)

This section compares the performances of best models trained on DHS, FFA and Macula datasets with state-of-the-art models. To this aim, we use our ResNet50V2 late fusion model since it was the best over the three datasets. It was firstly compared to the proposed CNN of Sarki et al. [41] as it was the only work treating several eye diseases with a focus on DED. Although, we mainly perform the comparison using the results from the Macula dataset as it is the only dataset with OCT and Fundus modalities (to somehow match the state-of-the-art).

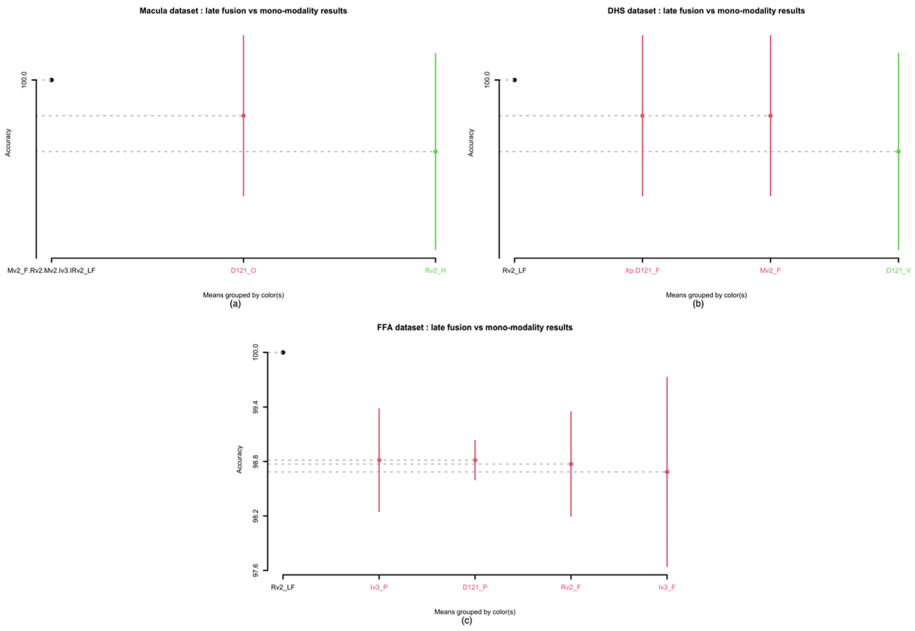

Fig. 11 Comparison of best late fusion with best mono-modality DL techniques on the (**a**) Macula dataset, (**b**) DHS dataset, and (**c**) FFA dataset

Based on the results presented in Table 12, the best results are provided by our proposed late fusion ResNet50V2 model trained on the Macula dataset using the combination of Fundus, Histology and OCT modalities with an accuracy of 100%. Secondly, we compare our ResNet50V2 late fusion model to the best multimodal model of Jin et al. [11] (feature level fusion model). Although both models have a similar accuracy (100%), the model of [11] assesses only choroidal neovascularization in AMD using OCT and OCTA modalities but our model assesses normal, non-neovascular AMD, and neovascular AMD using a combination of Fundus, Histology and OCT modalities. Thirdly, we compared our ResNet50V2 late fusion model to that one of [43]: the best mono-modality model detected DR grades using the Fundus modality with an accuracy of 96.32%. Although, our proposed model ranks first in this setting (along with [11]), it is relevant to note that the current models have been trained on disease imbalanced datasets (with DR and Gl being the most dominant diseases) with fewer instances (at most 100 true images) that have been oversampled through several data augmentation techniques; which might have considerably overfit the models. In this case, further empirical evaluations on newly acquired data from other hospitals are recommended to confirm or refute these results. But results from [11] confirm that the diagnostic of eye diseases using DL is promising and could in the future outperform ophthalmologists.

7 Threats to validity

This section discusses the three dimensions of the empirical study's validity threats.

Internal validity The batch normalization-based models were pre-trained on the data before being fine-tuned. This may cause these models to overfit on the data they were trained on. To avoid this overfitting, we used the early stopping, reduce learning rate on plateau, and the 5-fold cross validation methods. Another internal threat resides in using the RMSProp optimizer in the pretraining phase of these models and SGD with momentum during the fine-tuning phase. The reason for using RMSProp optimizer, introduced by Geoff Hinton, is that it tries to resolve the problem that gradients may vary widely in magnitudes. As for SGD with momentum, it helps accelerate gradients vectors in the right directions.

External validity The three datasets used in this study had few instances of the diseased imaging modalities to train the DL techniques; that is why oversampling using data augmentation was performed. This step could lead the models to: (1) overfit, and (2) not to generalize well if tested on datasets with specific eye diseases as the diseases included had very few unrepresentative instances. A similar study employing real-world representative datasets to confirm or refute the findings of the present study would be interesting. Another concern lays in splitting the data into 80% training and 20% test during pre-training which may increase bias. To increase variability and reduce bias we fine-tune using a split of 60% training, 20% validation, and 20% test while using the 5-fold CV approach.

Construct validity We provided the values of six metrics to evaluate the classifiers' performance (accuracy, sensitivity, specificity, precision, F1-score, AUC). The key reasons for this decision are that (1) the data is balanced, and (2) these measures were utilized in most research to evaluate the performance of DL models. In addition, to make conclusions, we employed the SK-ESD test and the Borda Count voting method with equal weights for these performance indicators. This procedure was used in order to prevent favoring one metric over another.

Table 12 Comparison of best proposed model with state-of-the-art models

Model	Task	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC (%)
Sarki et al. [41] (New CNN)	Classify DR, Macular Edema, Gl, Cataract and Healthy using Fundus	100.00	100.00	81.33	Not available
Yoo et al. [5] (Multimodal deep belief network)	Classify AMD using OCT and Fundus	88.80	91.30	89.70	96.10
Tseng et al. [12] (Proposed fusion model)	DR grading using Fundus (lesion information and disease severity classification)	86.83	94.49	92.95	95.51
Jin et al. [11] (feature level fusion)	Assessment of choroidal neovascularization in AMD using OCT and OCTA	Not available	Not available	100.00	100.00
Yao et al. [45] (FunSwin)	DR grading using Fundus	81.54	94.13	84.12	Not available
Canayaz [43] (wrapper methods with DL)	DR grading using Fundus	Not available	Not available	96.32	Not available
Proposed late fusion ResNet50V2	Classify ED using Fundus, Histology and OCT	100.00	100.00	100.00	100.00

8 Conclusion and future work

In this empirical study, seven deep learning models (VGG19, DenseNet121, InceptionV3, InceptionResNetV2, Xception, ResNet50V2, MobileNetV2) were trained using both mono-modality and multimodality images from three publicly available datasets (DHS, FFA, Macula) for ED (DED) classification. Six performance criteria were used in the empirical evaluation process, along with SK-ESD statistical test and the Borda Count voting method to asses and rank the seven trained DL models. The findings in respect to the research questions were as follows:

- **(RQ1): What is the overall performance of DL models using mono-modality data in ED classification? Is there any mono-modality DL architecture that outperforms the others?** The empirical evaluations showed that the best DL technique in terms of both criteria: performance and sensitivity (i.e., SR and SDR values) is DenseNet121 as it obtained optimum scores of SR and SDR in all the three datasets. Additionally, Resnet50V2 was found to be the second-best model as it obtained optimum scores in the Macula and FFA datasets, though it performed less in the DHS dataset.
- **(RQ2): How does a modality impact the diagnosis performance of the model?** All of the models with the fundus modality ranked as first on all of the three datasets. Furthermore, from the studies conducted on cost-effectiveness, we concluded that the fundus photography was relatively the most cost-effective especially when incorporated in portable devices and with the development of DL-based methods, we can further reduce the effect and costs of eye diseases on patients.
- **(RQ3): What is the overall performance of DL models using multimodality data in ED classification?** Late fusion best ranked models scored 100% in all of the six metrics for all of the datasets. Particularly, the ResNet50V2 model was the best DL late fusion technique across the three datasets, while VGG19 late fusion was the worst. Moreover, InceptionResNetV2 late fusion ranks 2nd as it appeared in both of the second clusters of the FFA and DHS dataset, while it was ranked first in the Macula dataset. Finally, we found that the DL techniques were most sensitive with the DHS and FFA modalities as the number of clusters produced was higher than in the Macula dataset.
- **(RQ4): How does late fusion architecture perform in comparison with mono-modality models?** Compared to mono-modality based models, late fusion using multimodality techniques achieved the best results with 100% in accuracy in comparison with the best mono-modality trained models DenseNet121 and ResNet50V2 with an average accuracy across modalities and datasets of 99.57% and 99.51% respectively.
- **(RQ5): How does best model(s) compare with state-of-the-art?** The proposed ResNet50V2 late fusion model trained on the Macula dataset (accuracy = 100%) outperforms current state-of-the-art models Multimodal deep belief network [5] (accuracy = 89.70%) and New CNN [41] (accuracy = 81.33%).

Ongoing works intend to investigate early and joint fusion methods using the same datasets and DL techniques for comparison with late fusion and mono-modality. Additionally, an efficiency analysis will be conducted regarding the number of parameters, consumed time, and memory.

Appendix A: Deep Learning abbreviation

Table 13 Abbreviations used for the DL techniques in the SK-ESD graphs

DL techniques	Abbreviation
DenseNet121	D121
InceptionV3	Iv3
InceptionResNetV2	IRv2
MobileNetV2	Mv2
ResNet50V2	Rv2
VGG19	V19
Xception	Xp

Table 14 Abbreviations used for mono-modality model comparison in the SK-ESD graphs

DL techniques	Modality	Abbreviation
DenseNet121	Fundus	D121_F
InceptionV3		Iv3_F
InceptionResNetV2		IRv2_F
MobileNetV2		Mv2_F
ResNet50V2		Rv2_F
VGG19		V19_F
Xception		Xp_F
DenseNet121	Vessels	D121_V
InceptionV3		Iv3_V
InceptionResNetV2		IRv2_V
MobileNetV2		Mv2_V
ResNet50V2		Rv2_V
VGG19		V19_V
Xception		Xp_V
DenseNet121	OCT	D121_O
InceptionV3		Iv3_O
InceptionResNetV2		IRv2_O
MobileNetV2		Mv2_O
ResNet50V2		Rv2_O
VGG19		V19_O
Xception		Xp_O
DenseNet121	Histology	D121_H
InceptionV3		Iv3_H
InceptionResNetV2		IRv2_H
MobileNetV2		Mv2_H
ResNet50V2		Rv2_H
VGG19		V19_H
Xception		Xp_H

Acknowledgments The authors would like to thank the Google Ph.D. Fellowship program for supporting Sara El-Ateif.

Funding The authors declare they have no financial interests.

Data availability Data used in this study is publicly available online:
- DRIVE: Digital Retinal Images for Vessel Extraction dataset: last accessed on 2023-09-10, https://drive.grand-challenge.org/DRIVE/
- STARE (Structured Analysis of the Retina) dataset: last accessed on 2023-09-10, https://cecas.clemson.edu/~ahoover/stare/
- High-Resolution Fundus (HRF) Image Database: last accessed on 2022-09-10, https://www5.cs.fau.de/research/data/fundus-images/
- FFA: Fundus Fluorescein Angiogram Photographs of Diabetic Patients dataset: last accessed on 2023-09-10, https://sites.google.com/site/hosseinrabbanikhorasgani/available-datasets/fundus-fluorescein-angiogram-photographs-of-diabetic-patients?authuser=0
- Project Macula dataset: access each eye condition from the eye gallery navigation bar. Last accessed on 2023-09-10, https://projectmacula.org/eye-gallery/

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Li JQ, Welchowski T, Schmid M et al (2020) Prevalence, incidence and future projection of diabetic eye disease in Europe: a systematic review and meta-analysis. Eur J Epidemiol 35:11–23. https://doi.org/10.1007/s10654-019-00560-z
2. Gargeya R, Leng T (2017) Automated identification of diabetic retinopathy using deep learning. Ophthalmol 124:962–969. https://doi.org/10.1016/j.ophtha.2017.02.008
3. Harangi B, Toth J, Baran A, Hajdu A (2019) Automatic screening of fundus images using a combination of convolutional neural network and hand-crafted features. Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS 2699–2702. https://doi.org/10.1109/EMBC.2019.8857073
4. Pratt H, Coenen F, Broadbent DM et al (2016) Convolutional neural networks for diabetic retinopathy. Procedia Comput Sci 90:200–205. https://doi.org/10.1016/j.procs.2016.07.014
5. Yoo TK, Choi JY, Seo JG et al (2019) The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. Med Biol Eng Comput 57:677–687. https://doi.org/10.1007/s11517-018-1915-z
6. Sun D, Wang M, Li A (2019) A multimodal deep neural network for human breast Cancer prognosis prediction by integrating multi-dimensional data. IEEE/ACM Trans Comput Biol Bioinforma 16:841–850. https://doi.org/10.1109/TCBB.2018.2806438
7. Qiu S, Chang GH, Panagia M et al (2018) Fusion of deep learning models of MRI scans, Mini–Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment. Alzheimer's Dement Diagnosis, Assess Dis Monit 10:737–749. https://doi.org/10.1016/j.dadm.2018.08.013
8. An G, Omodaka K, Tsuda S, et al (2018) Comparison of Machine-Learning Classification Models for Glaucoma Management https://doi.org/10.1155/2018/6874765
9. Huang SC, Pareek A, Seyyedi S, et al (2020) Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. npj Digit Med 3:. https://doi.org/10.1038/s41746-020-00341-z
10. Vaghefi E, Hill S, Kersten HM, Squirrell D (2020) Multimodal retinal image analysis via deep learning for the diagnosis of intermediate dry age-related macular degeneration: a feasibility study. J Ophthalmol 2020:1–7. https://doi.org/10.1155/2020/7493419
11. Jin K, Yan Y, Chen M et al (2022) Multimodal deep learning with feature level fusion for identification of choroidal neovascularization activity in age-related macular degeneration. Acta Ophthalmol 100:e512–e520. https://doi.org/10.1111/aos.14928

12. Tseng VS, Chen C-L, Liang C-M et al (2020) Leveraging multimodal deep learning architecture with retina lesion information to detect diabetic retinopathy. Transl Vis Sci Technol 9:41. https://doi.org/10.1167/tvst.9.2.41
13. El-Ateif S, Idri A (2022) Single-modality and joint fusion deep learning for diabetic retinopathy diagnosis. Sci African 17:e01280. https://doi.org/10.1016/j.sciaf.2022.e01280
14. Hervella ÁS, Rouco J, Novo J, Ortega M (2022) Multimodal image encoding pre-training for diabetic retinopathy grading. Comput Biol Med 143:. https://doi.org/10.1016/j.compbiomed.2022.105302
15. An G, Omodaka K, Hashimoto K, et al (2019) Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images. J Healthc Eng 2019:. https://doi.org/10.1155/2019/4061313
16. Lee YC, Cho H Bin, Choi YH (2022) Classification for referable glaucoma with fundus photographs using multimodal deep learning, pp 2–3. http://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/7b0a4c21-e8b4-4fee-9e01-f48e25b2b1b4/Classification_for_ref_sFDsTWh.pdf
17. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. 3rd Int Conf learn represent ICLR 2015 - Conf track proc, pp 1–14. https://doi.org/10.48550/arXiv.1409.1556
18. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 9908 LNCS:630–645. https://doi.org/10.1007/978-3-319-46493-0_38
19. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017 2017-Janua:2261–2269. https://doi.org/10.1109/CVPR.2017.243
20. Szegedy C, Vanhoucke V, Ioffe S, et al (2016) Rethinking the inception architecture for computer vision. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016-Decem:2818–2826. https://doi.org/10.1109/CVPR.2016.308
21. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-ResNet and the impact of residual connections on learning. 31st AAAI Conf Artif Intell AAAI 2017:4278–4284. https://doi.org/10.1609/aaai.v31i1.11231
22. Chollet F (2016) Xception: deep learning with Depthwise separable convolutions. SAE Int J Mater Manuf 7:1251–1258. https://doi.org/10.48550/arXiv.1610.02357
23. Sandler M, Howard A, Zhu M, et al (2018) MobileNetV2: inverted residuals and linear bottlenecks. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 4510–4520. https://doi.org/10.1109/CVPR.2018.00474
24. Staal J, Abramoff MD, Niemeijer M et al (2004) Ridge-based vessel segmentation in color images of the retina. IEEE Trans Med Imaging 23:501–509. https://doi.org/10.1109/TMI.2004.825627
25. Hoover A (2000) Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans Med Imaging 19:203–210. https://doi.org/10.1109/42.845178
26. Budai A, Bock R, Maier A, et al (2013) Robust vessel segmentation in fundus images. Int J Biomed Imaging 2013:. https://doi.org/10.1155/2013/154860
27. Hajeb Mohammad Alipour S, Rabbani H, Akhlaghi M (2014) A new combined method based on curvelet transform and morphological operators for automatic detection of foveal avascular zone. Signal, Image Video Process 8:205–222. https://doi.org/10.1007/s11760-013-0530-6
28. Zanzottera EC, Messinger JD, Ach T et al (2015) The project MACULA retinal pigment epithelium grading system for histology and optical coherence tomography in age-related macular degeneration. Investig Opthalmo Vis Sci 56:3253. https://doi.org/10.1167/iovs.15-16431
29. Tantithamthavorn C, McIntosh S, Hassan AE, Matsumoto K (2019) The impact of automated parameter optimization on defect prediction models. IEEE Trans Softw Eng 45:683–711. https://doi.org/10.1109/TSE.2018.2794977
30. Emerson P (2013) The original Borda count and partial voting. Soc Choice Welf 40:353–358. https://doi.org/10.1007/s00355-011-0603-9
31. Elmidaoui S, Cheikhi L, Idri A, Abran A (2020) Predicting software maintainability using ensemble techniques and stacked generalization. CEUR Workshop Proc 2725:1–16. https://doi.org/10.5277/E-INF190105
32. Abramoff MD, Garvin MK, Sonka M (2010) Retinal imaging and image analysis. IEEE Rev Biomed Eng 3:169–208. https://doi.org/10.1109/RBME.2010.2084567
33. Fenner BJ, Wong RLM, Lam W-C et al (2018) Advances in retinal imaging and applications in diabetic retinopathy screening: a review. Ophthalmol Therapy 7:333–346. https://doi.org/10.1007/s40123-018-0153-7
34. Choi JY, Yoo TK, Seo JG et al (2017) Multi-categorical deep learning neural network to classify retinal images: a pilot study employing small database. PLoS One 12:1–16. https://doi.org/10.1371/journal.pone.0187336

35. Sayres R, Taly A, Rahimy E et al (2019) Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmol 126:552–564. https://doi.org/10.1016/j.ophtha.2018.11.016
36. Phan S, Satoh S, Yoda Y et al (2019) Evaluation of deep convolutional neural networks for glaucoma detection. Jpn J Ophthalmol 63:276–283. https://doi.org/10.1007/s10384-019-00659-6
37. Diaz-Pinto A, Morales S, Naranjo V et al (2019) CNNs for automatic glaucoma assessment using fundus images: An extensive validation. Biomed Eng Online 18:1–19. https://doi.org/10.1186/s12938-019-0649-y
38. Gómez-Valverde JJ, Antón A, Fatti G et al (2019) Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. Biomed Opt Express 10:892. https://doi.org/10.1364/BOE.10.000892
39. De Moura Lima AC, Maia LB, Pinheiro Pereira RM, et al (2018) Glaucoma Diagnosis over Eye Fundus Image through Deep Features. Int Conf Syst Signals, Image Process 2018-June:1–5. https://doi.org/10.1109/IWSSIP.2018.8439477
40. Umapathy A, Sreenivasan A, Nairy DS, et al (2019) Image processing, textural feature extraction and transfer learning based detection of diabetic retinopathy. ACM Int Conf Proceeding Ser 17–21. https://doi.org/10.1145/3314367.3314376
41. Sarki R, Ahmed K, Wang H, et al (2018) Convolutional neural network for multi-class classification of diabetic eye disease. ICST Trans Scalable Inf Syst 172436. https://doi.org/10.4108/eai.16-12-2021.172436
42. Liu TYA, Ling C, Hahn L, et al (2022) Prediction of visual impairment in retinitis pigmentosa using deep learning and multimodal fundus images. Br J Ophthalmol bjophthalmol-2021-320897. https://doi.org/10.1136/bjo-2021-320897
43. Canayaz M (2022) Classification of diabetic retinopathy with feature selection over deep features using nature-inspired wrapper methods. Appl Soft Comput 128:109462. https://doi.org/10.1016/j.asoc.2022.109462
44. Gu Z, Li Y, Wang Z et al (2023) Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention. https://doi.org/10.1155/2023/1305583
45. Yao Z, Yuan Y, Shi Z et al (2022) FunSwin: a deep learning method to analysis diabetic retinopathy grade and macular edema risk based on fundus images. Front Physiol 13:1–9. https://doi.org/10.3389/fphys.2022.961386
46. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. https://doi.org/10.48550/arXiv.1512.03385
47. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. 32nd Int Conf Mach learn ICML 2015 1:448–456. https://doi.org/10.48550/arXiv.1502.03167
48. Sarki R, Ahmed K, Wang H, Zhang Y (2020) Automatic detection of diabetic eye disease through deep learning using fundus images: a survey. IEEE Access 8:151133–151149. https://doi.org/10.1109/ACCESS.2020.3015258
49. Decencière E, Zhang X, Cazuguel G et al (2014) Feedback on a publicly distributed image database: the MESSIDOR database. Image Anal Stereol 33:231. https://doi.org/10.5566/ias.1155
50. Fumero F, Alayon S, Sanchez JL et al (2011) RIM-ONE: An open retinal image database for optic nerve evaluation. In: 2011 24th international symposium on computer-based medical systems (CBMS). IEEE, pp 1–6. https://doi.org/10.1109/CBMS.2011.5999143
51. Sivaswamy J, Krishnadas SR, Datt Joshi G et al (2014) Drishti-GS: retinal image dataset for optic nerve head(ONH) segmentation. In: 2014 IEEE 11th international symposium on biomedical imaging (ISBI). IEEE, pp 53–56. https://doi.org/10.1109/ISBI.2014.6867807
52. Abràmoff MD, Folk JC, Han DP et al (2013) Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol 131:351. https://doi.org/10.1001/jamaophthalmol.2013.1743
53. APTOS 2019 Blindness Detection | Kaggle. https://www.kaggle.com/c/aptos2019-blindness-detection/overview/description. Accessed 1 Nov 2021
54. Li T, Gao Y, Wang K et al (2019) Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Inf Sci (Ny) 501:511–522. https://doi.org/10.1016/j.ins.2019.06.011
55. Porwal P, Pachade S, Kamble R et al (2018) Indian diabetic retinopathy image dataset (IDRiD). In: IEEE Dataport. https://doi.org/10.21227/H25W98
56. Wilkinson CP, Ferris FL, Klein RE et al (2003) Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmol 110:1677–1682. https://doi.org/10.1016/S0161-6420(03)00475-5
57. Gencay R, Qi M (2001) Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping, and bagging. IEEE Trans Neural Netw 12:726–734. https://doi.org/10.1109/72.935086

58. Szegedy C, Liu W, Jia Y, et al (2015) Going deeper with convolutions. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 07–12-June:1–9. https://doi.org/10.1109/CVPR.2015.7298594
59. Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugenics 7:179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x
60. Liu C, Belkin M (2018) Accelerating SGD with momentum for over-parameterized learning. https://doi.org/10.48550/arXiv.1810.13395