Deep Learning Model Evaluation and Insights in Inherited Retinal Disease Detection

Hélder Ferreira[1(✉)], Ana Marta[2,3], Inês Couto[2], José Câmara[1], João Melo Beirão[2,3], and António Cunha[1,4]

[1] University of Trás-os-Montes and Alto Douro, 5000-801 Vila Real, Portugal
helder2003ferreira123@gmail.com
[2] Department of Ophthalmology, Centro Hospitalar Universitário de Santo António, EPE (CHUdSA), Porto, Portugal
[3] Instituto de Ciências Biomédicas Abel Salazar (ICBAS), Porto, Portugal
[4] INESC TEC - INESC Technology and Science, FEUP Campus, Porto, Portugal

Abstract. Inherited retinal diseases such as Retinitis Pigmentosa and Stargardt's disease are genetic conditions that cause the photoreceptors in the retina to deteriorate over time. This can lead to vision symptoms such as tubular vision, loss of central vision, and nyctalopia (difficulty seeing in low light) or photophobia (high light). Timely healthcare intervention is critical, as most forms of these conditions are currently untreatable and usually focused on minimizing further vision loss.

Machine learning (ML) algorithms can play a crucial role in the detection of retinal diseases, especially considering the recent advancements in retinal imaging devices and the limited availability of public datasets on these diseases. These algorithms have the potential to help researchers gain new insights into disease progression from previous classified eye scans and genetic profiles of patients.

In this work, multi-class identification between the retinal diseases Retinitis Pigmentosa, Stargardt Disease, and Cone-Rod Dystrophy was performed using three pretrained models, ResNet101, ResNet50, and VGG19 as baseline models, after shown to be effective in our computer vision task. These models were trained and validated on two datasets of autofluorescent retinal images, the first containing raw data, and the second dataset was improved with cropping to obtain better results. The best results were achieved using the ResNet101 model on the improved dataset with an Accuracy (Acc) of 0.903, an Area under the ROC Curve (AUC) of 0.976, an F1-Score of 0.897, a Recall (REC) of 0.903, and a Precision (PRE) of 0.910.

To further assess the reliability of these models for future data, an Explainable AI (XAI) analysis was conducted, employing Grad-Cam. Overall, the study showed promising capabilities of Deep Learning for the diagnosis of retinal diseases using medical imaging.

A. Cunha et al. (Eds.): MobiHealth 2023, LNICST 578, pp. 295–306, 2024.
https://doi.org/10.1007/978-3-031-60665-6_22

1 Introduction

Inherited retinal diseases (IRDs) are genetic diseases that affect the normal function of light-sensitive cells (photoreceptors) and the cell layer that supports them (retinal pigment epithelium). They are the most common cause of blindness in the working-age population in some developed countries, to some degree [1].

The most common IRD, Retinitis Pigmentosa, manifests through fundoscopy, revealing characteristics like bony spicule-like pigmentations in the retina, narrowed arterioles, optic disc pallor, cataracts, and vitreous cell presence [2]. Stargardt Disease presents hiperfluorescent pisciform lesions, macular atrophy, and choriocapillaris silence due to lipofuscin accumulation. Cone-rod Dystrophy predominantly impacts the macula, causing central vision difficulties, color vision impairment, and macular atrophy.

IRDs often have unique phenotypic features that medical professionals can identify using advanced retinal imaging technology [3]. This technology allows for the rapid and non-invasive acquisition of high-resolution retinal images, requiring sometimes just a dilated pupil and causing no discomfort to the patient. Various imaging modalities, including fundus autofluorescence (FAF) and spectral-domain optical coherence tomography (SD-OCT), can be used to perform these scans [4].

For example, FAF images are critical in identifying patterns indicative of photoreceptor dysfunction and apoptosis. This is done by detecting lipofuscin and related compounds by exploiting their auto-fluorescent properties. Lipofuscin accumulates primarily in the retinal pigment epithelium (RPE) due to oxidative stress. By visualising lipofuscin distribution, FAF images provide valuable insight into the extent and location of oxidative stress-induced damage in the retina [5].

The availability of high-resolution, detailed multimodal information allows ophthalmologists to identify patterns associated with specific diseases. However, due to the rarity of these diseases, accurate clinical diagnoses require specialised expertise and experience that is limited to a small number of specialists and hospitals.

State-of-the-art machine learning models can recognise disease-specific patterns from retinal images and be used to extend the reach of accurate diagnoses to a broader population [6].

These models' importance is increased when considering the scarcity of available public data about these diseases and the existence of different modalities of retinal imaging that go from fundus photography with 30° of field of view up to Optos Ultra-Widefield with 200° [7,8].

A substantial amount of data about IRDs is generally hard to get or have access to. However, for this study, a public Portuguese hospital provided us with a private dataset involving autofluorescence retinal images, which will be used.

In this paper, standard state-of-the-art Deep Learning models were applied and evaluated to automatically classify retinal images on three inherited retinal diseases: Retinitis Pigmentosa, Stargardt Disease, and Cone-Rod Dystrophy. Additionally, Grad-CAM maps were used to identify potential biases and improve models.

2 Methodology

The methodology adopted in this study follows the standard machine learning steps. It is divided into two procedures (P1 and P2), illustrated in Fig. 1.

First, the set of retinal images, raw data (RD), underwent a preprocessing stage, where it was cleaned, restructured with proper labels and split into 2 datasets (DS1 and DS2) for model training. Then, models were chosen and trained (T1 and T2). Finally, the models are evaluated using standard classification metrics (E1 and E2) and accessed with Grad-CAM maps.

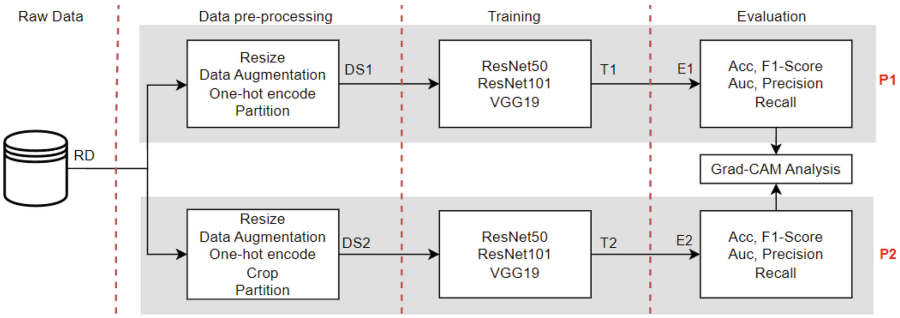

Fig. 1. Methodology pipeline

2.1 Raw Data

A collection of 491 autofluorescence retinal images was made available by a public hospital in Portugal. The images were labelled by clinical experts, with 326 identified as Retinitis Pigmentosa (RP), 81 as Cone-rod Dystrophy (CR), and the remaining 84 as indicative of Stargardt Disease (STG), as can be seen in Fig. 2.

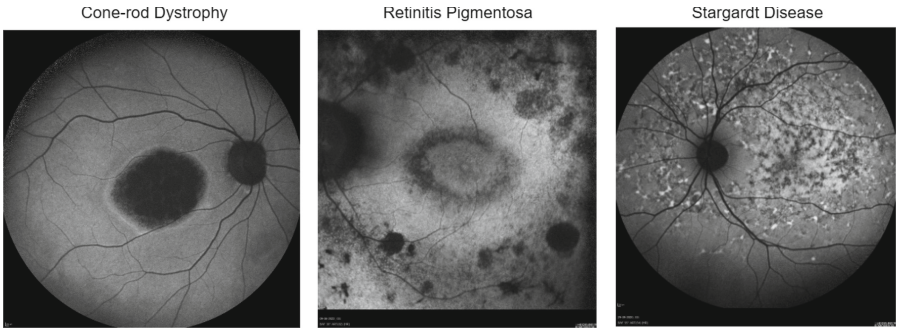

Fig. 2. Examples of images from each class

The set included images from each patient's eyes with multiple imaging modalities to capture the retinal images, adding diversity to our dataset. Some images have a field view of 30° (fundus photography), others a field of view at 55° (Wide-field retinal camera) and others have a field view of 200° (Optos Ultra-Wideview).

2.2 Data Preprocessing

The preprocessing stage is structured according to Fig. 3.

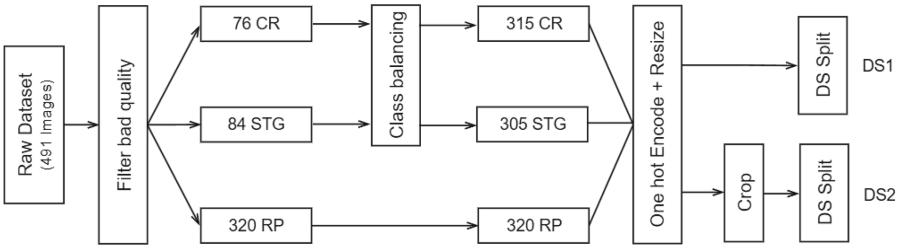

Fig. 3. Data Preprocessing Diagram

During this step, 11 images were removed for having poor quality, i.e., low resolution and too dark (unreadable), where 5 belonged to the CR category and 6 to the RP category.

To maintain data symmetry and ensure balanced class representation, we applied data augmentation techniques involving horizontal flips and image rotations within the range of -10º to 10º specifically for the Cone-rod and Stargardt classes, illustrated in Fig. 4, as those techniques allowed the increase of data while improving the efficiency of the models. This approach resulted in a balanced dataset, comprising 315 images for CR, 320 for RP, and 305 for STG. Then, all images were resized to a standard dimension of 224×224 pixels and

transformed categorical class data into numerical attributes using one-hot encoding allowing a faster model training. Finally, both datasets were partitioned into training, validation, and testing sets, with ratios of 64:16:20, respectively.

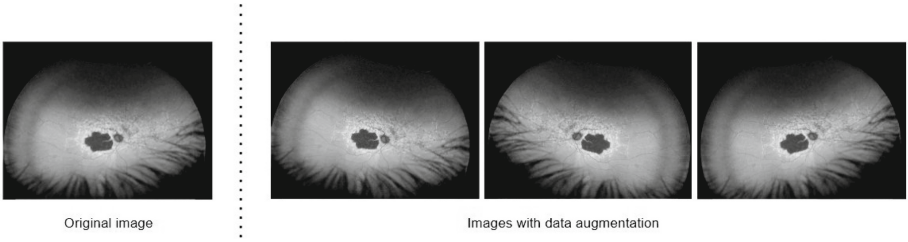

Fig. 4. Original image (on the left) and images with data augmentation (on the right)

The diagram illustrated in Fig. 3, shows all steps taken during the preprocessing of data, including removal of unreadable images, application of data augmentation on CR and STG classes, one hot encoding and resize, crop in DS2 and finally partition on both sets.

Apart from the cropping variation between DS1 and DS2, our strategies for image preprocessing remained consistent.

In the first approach, we also identified that some of the images contained bottom text, on the RP and STG classes which could lead to model memorization during training, since it's not present in the CR class. Considering this possibility, we made a strategic decision to divide our data into two distinct datasets. One dataset (DS1) contained the raw images with bottom text, while the other dataset (DS2) was edited to remove this text by cropping 15% of the image's height. This allowed us to later determine potential model bias and ensure it focuses on the right features.

2.3 Model Training

The VGG (Visual Geometry Group) and ResNet (Residual Networks) models were selected since they are state-of-the-art classifiers with exceptional performance for medical analysis tasks. Additionally, these models are suited for problems with small datasets, just like the one we're dealing with, since they have trained weight for Transfer learning, e.g., from the ImageNet [9]. ResNet101, ResNet50, and VGG19 model versions were selected for the project.

The models went through a two-step training process for P1 and P2, utilizing DS1 for both training and validation of P1, and DS2 for both training and validation of P2, respectively.

The training was conducted using 15 epochs, which proved to be sufficient to achieve convergence in terms of performance for all models.

Transfer learning was executed by loading the state-of-the-art models, freezing all layers, and appending a final dense layer for classification with 3 neurons matching our number of classes. The training optimiser employed was the Adam Optimization Algorithm, alongside the utilization of a learning rate scheduler callback to lower the learning rate over epochs to achieve a more stable convergence as the optimisation progresses.

2.4 Model Evaluation

Standard classification metrics were used to evaluate models, including Area under the ROC Curve (AUC), Accuracy (Acc), Precision (PRE), Recall (REC), and F1-Score. These metrics rely on essentially 4 values: The number of True Positive cases (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) identified by the model. AUC metric was particularly useful in assessing the performance of our model in distinguishing between positive and negative cases.

It is essential to note, however, that while the Acc metric is the most commonly employed, it may not consistently provide the most representative evaluation, particularly when data exhibits class imbalances. For that reason, other metrics such as AUC or F1-Score can be better options [10].

The Acc, PRE, REC and F1-Score metrics are calculated according to their equations:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \qquad F_1 = 2 \cdot \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Pre} = \frac{TP}{TP + FP} \qquad \text{Rec} = \frac{TP}{TP + FN}$$

All models underwent a final evaluation using a dedicated test set comprising 20% of the dataset employed during training, which involved DS1 for P1 and DS2 for P2. The dataset split was performed as part of the initial data preprocessing stage.

3 Results and Discussion

This section provides an overview of the setup used for training the models and presents the results, starting with P1 with the use of DS1, then P2 with DS2, and finally a discussion between them.

3.1 Setup

When fitting the model, we used data in batches of 16 for both training and validation, for 15 epochs, while using the learning rate callback with the formula,

learning rate * exp (-0.05). This formula exponentially decreases the learning rate during each epoch.

The machine learning models were trained on a PC with an Intel Core i5-7400 CPU, NVIDIA GeForce RTX 3060 GPU, and 12 GB of RAM. The system ran on Windows 10 and utilised the TensorFlow framework for training and evaluating the model.

3.2 P1 Results

The results for the initial procedure (P1) are presented in Fig. 5 and include metrics results and the confusion matrix, which provides details on how many test set images were correctly classified by each model (the values on the diagonal) and how many were misclassified (all other values).

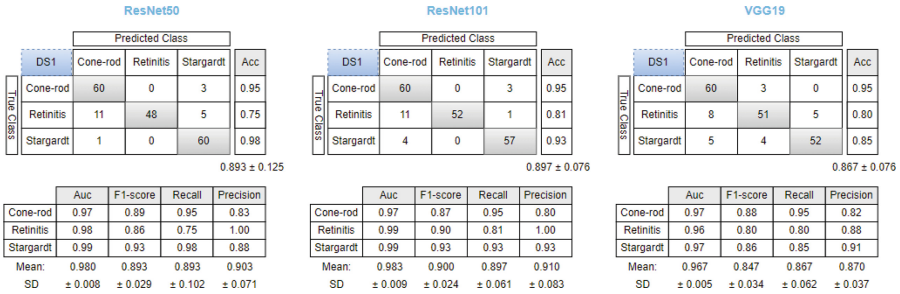

Fig. 5. ResNet101, ResNet50 and VGG19 models - Confusion matrices and metrics results when performed on DS1 with their Standard Deviation (SD)

All 3 models showed positive results in terms of mean Acc ranging from 86 to 90%. When looking at each class individually, on the Cone-rod dystrophy all models achieved an Acc of 0.95, Retinitis Pigmentosa had Accs of 0.75, 0.81 and 0.80 for ResNet50, ResNet101 and VGG19 respectively, and on Stargardt Disease these models got Accs of 0.98, 0.93 and 0.85 in the same order.

As for the other metrics, ResNet50 achieved a mean AUC of 0.980, with an F1-Score of 0.893, REC of 0.893, and PRE of 0.903. Similarly, ResNet101 obtained scores of 0.983, 0.900, 0.897, and 0.910, for mean AUC, F1-Score, REC, and PRE, respectively. On the other hand, VGG19's results were slightly lower, with values of 0.967 for mean AUC, 0.847 for F1-Score, 0.867 for REC, and 0.870 for PRE.

In addition to obtaining these results, the visualization of images for which the models made incorrect predictions is always beneficial for evaluating the justification of these errors. Some of these examples can be found in Fig. 6.

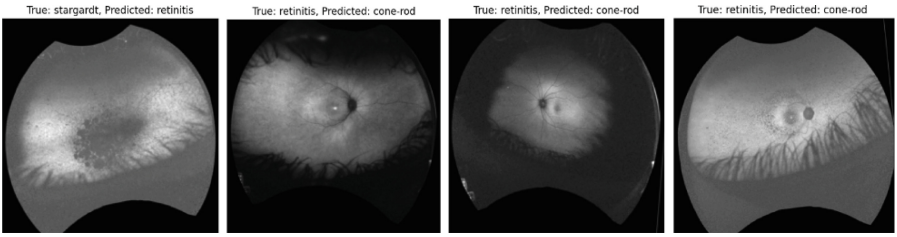

Fig. 6. Examples of images that all 3 models predicted wrong

After examining the images, it appeared they weren't that easy to predict and it's reasonable to confuse the models. It's important to remember that the dataset we used was relatively small, so we shouldn't discredit the models based solely on these challenging examples.

3.3 P2 Results

In P2, the procedure remained similar, with the only change being the utilization of DS2 to obtain results, and they can be seen in Fig. 7.

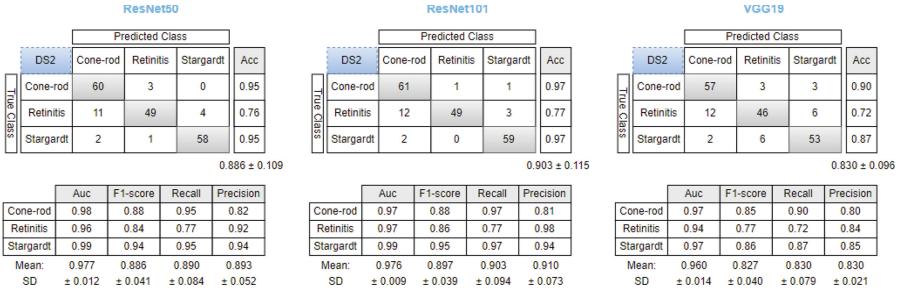

Fig. 7. ResNet101, ResNet50 and VGG19 models - Confusion matrices and metrics results when performed on DS2

P2 results were not very far apart from the P1 seen previously, this time the mean ACC ranged from 83 to 90%. On the Cone-rod dystrophy class, ResNet50, ResNet101 and VGG19 achieved an Acc of 0.95, 0.97 and 0.90 respectively. On the same order of models, Retinitis Pigmentosa had an Acc of 0.76, 0.77 and 0.72. Lastly, on Stargardt Disease these models got Accs of 0.95, 0.97 and 0.87.

In the other metrics, ResNet50 achieved a mean AUC of 0.977, with an F1-Score of 0.886, REC of 0.890, and PRE of 0.893. Similarly, ResNet101 obtained scores of 0.976, 0.897, 0.903, and 0.910, for mean AUC, F1-Score, REC, and PRE, respectively. VGG19 results were once again lower, with values of 0.960 for mean AUC, 0.827 for F1-Score, 0.830 for both REC and PRE.

The same idea of visualising images in which the models made incorrect predictions was conducted, and examples can be seen in Fig. 8.

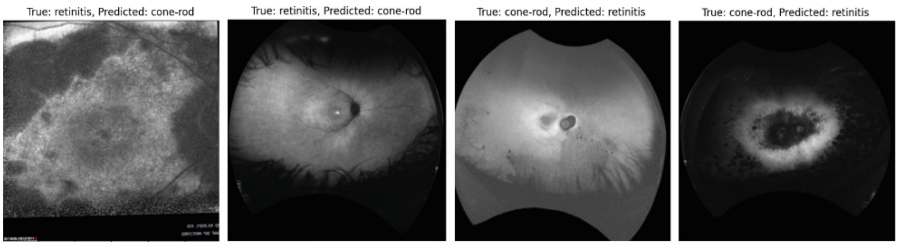

Fig. 8. Examples of images that all 3 models predicted wrong

Similarly to what happened in P1, when analysing the models' incorrect predictions, most of the images are susceptible to confusion and do not disregard the models' capabilities

3.4 Discussion and Explainability Analysis

While the results from DS1 and DS2 may not be very far off each other, their preprocessing strategies were different with the removal of extraneous text on DS2 that was present on the bottom part of images and could lead to memorization during model training, consequently affecting its generalisation. This was a real concern especially considering the text was present on the RP and STG images categories but not on the CR, so the models could use that for advantage if untreated.

To visualise if text memorization had happened in the first case (DS1), we utilised the Grad-CAM as a final analysis technique. Grad-CAM is a state-of-the-art visualization method commonly used in computer vision tasks, specifically in convolutional neural networks (CNN) for image classification [11].

By generating a heatmap, Grad-CAM allows us to identify the regions in an image that had the most impact on the network's final prediction. Unlike previous approaches, which focused solely on the last layer's feature maps, Grad-CAM analyses the gradients flowing into the final convolutional layer. This enables us to pinpoint the specific areas of the image that the model paid attention to while making its decision [11].

When visualizing these attention maps, we gain valuable insights into the decision-making process of the model. We can determine whether the model focuses on relevant image features or relies on irrelevant cues such as text, thereby validating our hypothesis about text memorization. This technique has immense potential in uncovering the inner workings of deep learning models and understanding their strengths and limitations in diverse applications, including medical image analysis.

For this evaluation, we will only be considering the DS1 and DS2 models with best results, which both correspond to ResNet101 and apply Grad-CAM on images of every category containing the bottom text, taking into account DS1 model was trained with text and DS2 model was trained without it. Results can be seen in Fig. 9.

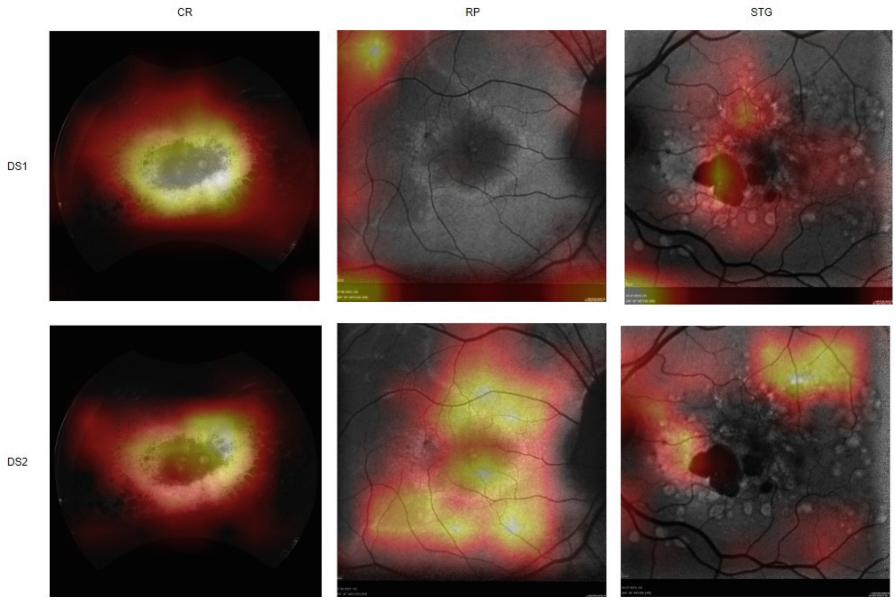

Fig. 9. Examples of Grad-CAM applications on images from each category (Cone-rod on left, Retinitis Pigmentosa on middle and Stargardt on the right) using the ResNet101 model trained with DS1 (top images) and ResNet101 trained with DS2 (lower images)

In the case of the DS1 train set, the Grad-CAM image reveals that the model seems to be memorizing irrelevant text information present in images from both RP and STG. This suggests that the model is giving excessive importance to text that may not be directly related to the task at hand and that is problematic as it may lead to overfitting and the model's inability to generalize well on unseen data.

When examining the Grad-CAM image of the DS2 train set, it is evident that the model is focusing on the important features relevant to the task. By cropping the images to only include the necessary visual elements, the model was able to disregard irrelevant information and prioritize the important aspects even when asked to make predictions on images that contain the text. This shows that the DS2 is providing more accurate and meaningful insights, allowing the model to make informed decisions.

Based on these observations, it's safe to say that the model trained with ResNet101 on DS2 is the best option for future predictions as it has good results and disregards irrelevant information allowing it to generalize well on unseen data. This highlights the importance of preprocessing and selecting relevant training data to enhance the performance and interpretability of machine learning models [12].

4 Conclusion

In our research, we have showcased the exciting potential of leveraging convolutional neural networks (CNNs) in conjunction with fundus autofluorescence (FAF) images to autonomously categorize a spectrum of inherited retinal diseases (IRDs). While it's important to acknowledge that our results may not yet attain perfection from a purely technical standpoint, they represent a significant leap forward in the realm of medical diagnostics.

As we look ahead, our research trajectory aims to explore new deep learning strategies for disease classification to improve the results. While at the same time, increasing our database with a diverse range of FAF images to enhance the robustness of our models. This expansion will undoubtedly broaden the applicability of our findings and deepen our understanding of IRDs.

Moreover, our aspiration extends beyond the confines of our work. We envision that this study will serve as a ground base for other projects with different data to enhance healthcare solutions and improve patient outcomes in the field of retinal diseases.

Acknowledgments. This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.

References

1. Rachael C., et al.: Inherited retinal diseases are the most common cause of blindness in the working-age population in Australia. Ophthalmic Genet. **42**(4), 431–439 (2021)
2. Francis, P.J.: Genetics of inherited retinal disease. J. Royal Soc. Med. **99**(4), 189–191 (2006)
3. Dockery, A., Whelan, L., Humphries, P., Farrar, G.J.: Next-generation sequencing applications for inherited retinal diseases. Int. J. Mol. Sci. **22**(11), 5684 (2021)
4. Pichi, F., Abboud, E.B., Ghazi, N.G., Khan, A.O.: Fundus autofluorescence imaging in hereditary retinal diseases. Acta Ophthalmol. **96**(5), e549–e561 (2018)
5. Heiferman, M.J., Fawzi, A.A.: Discordance between blue-light autofluorescence and near-infrared autofluorescence in age-related macular degeneration. Invest. Ophthalmol. Visual Sci. **57**(12), 25–25 (2016)
6. Pontikos, N., et al.: Eye2Gene: prediction of causal inherited retinal disease gene from multimodal imaging using AI. Invest. Ophthalmol. Visual Sci. **63**(7), 1161 (2022)

7. Liesenfeld, B., et al.: A telemedical approach to the screening of diabetic retinopathy: digital fundus photography. Diab. Care **23**(3), 345–348 (2000)
8. Nagiel, A., Lalane, R.A., Sadda, S.R., Schwartz, S.D.: Ultra-widefield fundus imaging: a review of clinical applications and future trends. Retina **36**(4), 660–678 (2016)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
10. Jordaney, R., Wang, Z., Papini, D., Nouretdinov, I., Cavallaro, L.: Misleading metrics: on evaluating machine learning for malware with confidence. Technical report (2016)
11. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
12. Famili, A., Shen, W.M., Weber, R., Simoudis, E.: Data preprocessing and intelligent data analysis. Intell. Data Anal. **1**(1), 3–23 (1997)