applied sciences

MDPI

Article

Retinitis Pigmentosa Classification with Deep Learning and Integrated Gradients Analysis

Hélder Ferreira [1,†], Ana Marta [2,3,†], Jorge Machado [1,4,†], Inês Couto [2], João Pedro Marques [5,6], João Melo Beirão [2,3] and António Cunha [1,7,*]

[1] School Sciences and Technologies, Universidade de Trás-os-Montes e Alto Douro, 5000-801 Vila Real, Portugal; helder2003ferreira123@gmail.com (H.F.); miguelj1992@outlook.com (J.M.)
[2] Department of Department of Ophthalmology, Unidade Local de Saúde de Santo António, 4099-001 Porto, Portugal; analuisamarta2@gmail.com (A.M.); inescouto12@gmail.com (I.C.); brandaobeirao@gmail.com (J.M.B.)
[3] Instituto de Ciências Biomédicas Abel Salazar (ICBAS), 4050-313 Porto, Portugal
[4] Department of Sciences and Technologies, Universidade Aberta, 1269-001 Lisboa, Portugal
[5] Department of Ophthalmology, Hospital da Universidade de Coimbra, Unidade Local de Saúde de Coimbra, 3004-561 Coimbra, Portugal; marquesjoaopedro@gmail.com
[6] Clinical and Academic Center of Coimbra (CACC), 3000-075 Coimbra, Portugal
[7] ALGORITMI Research Centre, University of Minho, 4800-058 Guimarães, Portugal
* Correspondence: acunha@utad.pt
† These authors contributed equally to this work.

Abstract: Inherited retinal diseases (IRDs) are genetic disorders affecting photoreceptors and the retinal pigment epithelium, leading to progressive vision loss. Retinitis pigmentosa (RP), the most common IRD, manifests as night blindness, peripheral vision loss, and eventually central vision decline. RP is genetically diverse and can be categorized into non-syndromic and syndromic. Advanced imaging technologies such as fundus autofluorescence (FAF) and spectral-domain optical coherence tomography (SD-OCT) facilitate diagnosing and managing these conditions. The integration of artificial intelligence in analyzing retinal images has shown promise in identifying genes associated with RP. This study used a dataset from Portuguese public hospitals, comprising 2798 FAF images labeled for syndromic and non-syndromic RP across 66 genes. Three pre-trained models, Inception-v3, ResNet-50, and VGG-19, were used to classify these images, obtaining an accuracy of over 80% in the training data and 54%, 56%, and 54% in the test data for all models. Data preprocessing included class balancing and boosting to address variability in gene representation. Model performance was evaluated using some main metrics. The findings demonstrate the effectiveness of deep learning in automatically classifying retinal images for different RP-associated genes, marking a significant advancement in the diagnostic capabilities of artificial intelligence and advanced imaging techniques in IRD.

Keywords: machine learning; deep learning; inherited retinal diseases; retinitis pigmentosa; gene classification

check for updates

Academic Editor: Gordana Žauhar

Received: 13 January 2025
Revised: 11 February 2025
Accepted: 13 February 2025
Published: 18 February 2025

Citation: Ferreira, H.; Marta, A.; Machado, J.; Couto, I.; Marques, J.P.; Beirão, J.M.; Cunha, A. Retinitis Pigmentosa Classification with Deep Learning and Integrated Gradients Analysis. *Appl. Sci.* **2025**, *15*, 2181. https://doi.org/10.3390/app15042181

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Inherited retinal diseases (IRDs) are genetic disorders that impair the function of photoreceptors and the retinal pigment epithelium, often leading to progressive vision loss and blindness [1,2]. Retinitis pigmentosa (RP) is the most prevalent, characterized by night blindness, peripheral vision loss, and eventually central vision decline [3]. RP can be categorized into non-syndromic RP or syndromic RP, and both types can be caused by various genes, demonstrating this disease's genetic complexity and variability [4].

Non-syndromic RP manifests solely in the retina, with clinical features such as bone spicule-like pigmentation, narrowed retinal arterioles, optic disc pallor, cataracts, and vitreous cell presence [5]. In contrast, syndromic RP is associated with systemic conditions affecting other body parts, such as hearing loss in Usher syndrome or kidney issues in Bardet–Biedl syndrome [2].

Treatment advancements heavily rely on identifying the specific genes responsible for these conditions. This process involves discovering new genes and screening for mutations in individuals and families affected by RP [6]. Despite its inherent complexity, significant strides have been made in identifying the genes and mutations associated with RP. Depending on the RP subtype and the diagnostic technology employed, mutations can be detected in up to 80% of cases using state-of-the-art genetic testing [7].

Retinal imaging modalities have revolutionized the diagnosis and management of IRDs, and technologies such as fundus autofluorescence (FAF) and spectral-domain optical coherence tomography (SD-OCT) enable detailed visualization of retinal structures, facilitating the identification of disease-specific patterns. FAF, for instance, highlights lipofuscin accumulation in the retinal pigment epithelium, revealing areas of photoreceptor dysfunction and oxidative stress [8,9].

Machine learning (ML) models, particularly deep learning techniques, offer promising solutions for automating the classification of retinal images and identifying genetic variations associated with RP, being able to analyze vast amounts of imaging data to detect subtle features indicative of specific genes, potentially enhancing diagnostic accuracy and accessibility [10].

One notable work in this area is the Eye2Gene study, which achieved an impressive top-5 accuracy of 85.6% in classifying retinal images based on 36 genes in IRDs from 44,817 scans. The study demonstrated how machine learning could enhance diagnostic precision, particularly when applied to large datasets. Our study builds upon this by utilizing a smaller dataset, directly comparing classification performance, and addressing the challenges posed by smaller datasets [11].

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 degrees of field of view up to Optos Ultra-Widefield with 200 degrees [12,13].

In this study, we utilized a private dataset provided by two Portuguese public hospitals, which included autofluorescence retinal images from patients with syndromic or non-syndromic RP and a total of 66 genes. State-of-the-art deep learning models were applied to classify these images according to the associated causing gene and category of RP. Additionally, we employed an Explainable AI (XAI) technique known as integrated gradients (IGs) to generate explanatory maps, helping to identify and mitigate potential biases in the models [14].

The results demonstrate the efficacy of deep learning in automatically classifying retinal images for different genes of RP. By integrating IG, we enhanced the interpretability of the models, providing insights into the decision-making process and ensuring the reliability of the results. This approach represents a significant step towards improving the diagnosis and understanding of inherited retinal diseases through advanced imaging and artificial intelligence.

The purpose of this study is to develop a neural network for genetic identification in retinitis pigmentosa. The aim is therefore to develop an artificial intelligence tool (model) for classifying FAF scans in order to identify the associated gene. This will make it possible, in the future, to predict the genes responsible for this disease quickly and in a non-evasive

way, and these models, the genetic classifiers, can be used as decision support tools for health professionals.

2. Materials and Methods

The methodology used in this study follows the standard machine learning steps as illustrated in Figure 1.

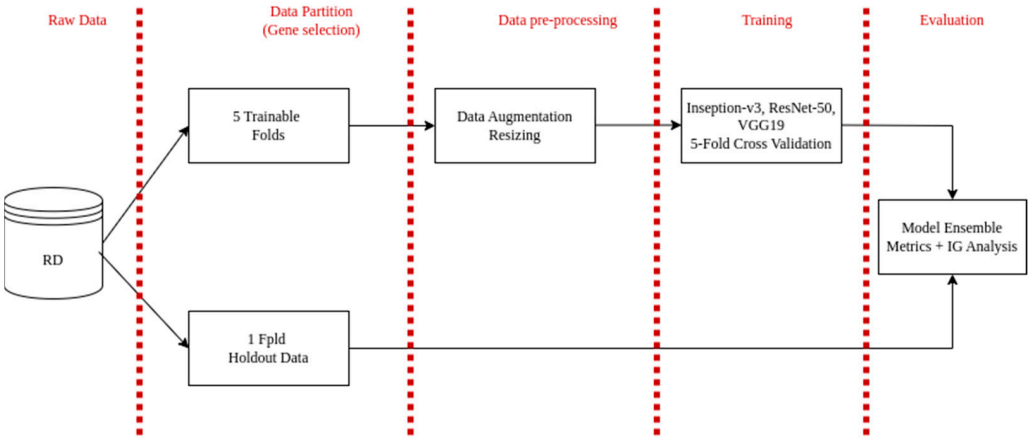


Figure 1. Methodology pipeline.

The process begins by partitioning the raw dataset (RD) into five trainable folds and one fold of holdout data, ensuring that only genes with at least five images per fold are included. This ensures sufficient representation across all folds. The five trainable folds are then subjected to a data preprocessing pipeline, which includes essential steps, such as data augmentation to enhance the diversity of the training set and resizing to fit the input requirements of the model. These preprocessed datasets are used to train the three models, the Inception-V3 model, the ResNet-50 model, and the VGG19 model, through a 5-fold cross-validation approach.

Once training is completed, the model's predictions are evaluated using model ensemble techniques, which combine the results from all five folds to generate a more robust prediction. Key metrics such as accuracy, recall, precision, F1-score, ROC-AUC, and the confusion matrix are calculated to assess the model's performance.

Additionally, integrated gradient (IG) analysis is applied to the model with the best accuracy result to help explain the model's predictions by highlighting the important features that influenced each decision. The holdout fold is kept untouched until the final evaluation phase, ensuring that the model's ability to generalize to unseen data is thoroughly tested before deployment or further use.

This dataset is the result of a collaborative effort by ophthalmology professionals from different hospitals, who managed to gather 2798 FAF images annotated for syndromic and non-syndromic RP. All questions regarding the collection and use of the data, in terms of ethics, were left to the health professionals who collected them and who maintained the confidentiality and anonymity of the patients on whom these exams were carried out. Under no circumstances were the subjects of this study known. The study complied with the tenets of the Declaration of Helsinki for biomedical research. Institutional Review Board (IRB)/Ethics Committee approval was obtained from the Departamento de Ensino, Formação e Investigação (DEFI) of Unidade Saúde Local de Santo António with reference number 2021.269(216-DEFI/224-CE).

2.1. Raw Data

Two public hospitals in Portugal provided a collection of 2798 autofluorescence retinal images. These images were labeled by clinical experts, who categorized them into two primary groups: non-syndromic retinitis pigmentosa (RP) and syndromic RP. In addition to the clinical diagnosis, each image was also annotated with the specific mutant gene, resulting in a dataset encompassing 66 genes linked to syndromic and non-syndromic RP.

The dataset featured images from each patient's eyes, captured using multiple imaging modalities to provide a more diverse and comprehensive view of the retina. Some images were taken with a 30-degree field of view using standard fundus photography, while others had a wider 55-degree field of view. Additionally, some images were captured using an Optos Ultra-Wideview camera (Optos plc, Dunfermline, UK), offering a 200-degree field of view. This variety in imaging techniques added depth to the dataset. It ensured that different areas of the retina were represented, potentially enhancing the model's ability to identify and classify features associated with different genes.

Although the dataset includes images with different fields of view (30, 55, and 200 degrees), no distinction was made based on this factor. All images were treated equally during model training, without any specific consideration for field of view.

2.2. Data Partition

In this step, the raw dataset was partitioned into five equal parts (folds), one of which was designated as the held-out data for final evaluation, while the remaining five folds were reserved for training purposes. This partitioning strategy is particularly useful for cross-validation during the model training.

Additionally, a threshold was established for gene selection, requiring each gene to have at least five images per fold for inclusion in the study. Genes with insufficient data were excluded to maintain consistency and avoid skewed results due to limited information, reducing the original 66 genes to 22 genes considered for this study. This filtering process helped ensure that only adequately represented genes contributed to the training and evaluation of the model.

The genes included and the number of samples, separated by syndromic and non-syndromic, for the training and test sets can be seen in Table 1.

Table 1. Number of samples per class.

	Gene	Train	Test
Non-Syndromic	*EYS*	419	87
	RPGR	311	58
	RHO	147	32
	RPE65	125	25
	NON_S_USH2A	111	22
	CNGB1	76	17
	NR2E3	39	7
	PROM1	37	9
	PRPF31	37	8
	CRB1	31	8
	PCARE	28	4
	CRCEP290	27	6
	PDE6B	26	7
	IMPG2	25	7
	MERTK	25	7
	HGSNAT	23	4

Table 1. *Cont.*

	Gene	Train	Test
Syndromic	*S_USH2A*	133	33
	MYO7A	81	18
	BBS1	58	18
	ARSG	49	9
	ADGRV1	48	13
	BBS10	30	8

2.3. Data Preprocessing

In the data preprocessing step, data augmentation techniques were used exclusively on the training folds to create diversity and introduce variations that the model might encounter in real-world scenarios, such as horizontal flips, brightness changes, and zoom. These transformations were applied with a 50% probability, ensuring a balance between the original and augmented images during training. Examples of these augmentations can be found in Figure 2, which showcases how they modify the images.

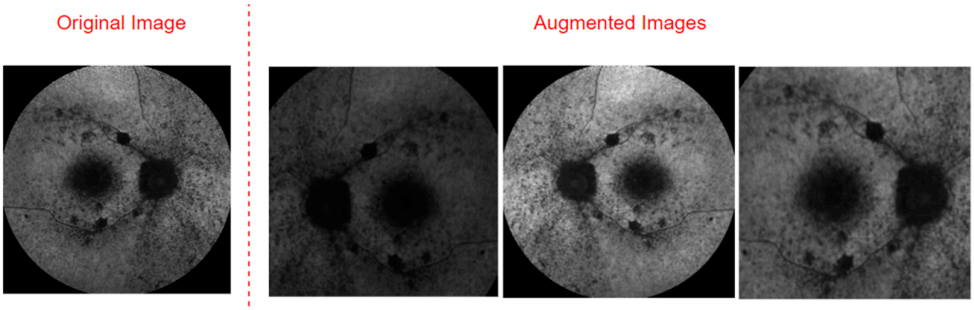


Figure 2. Original image vs. augmented images.

Importantly, no augmentation was applied to the validation fold during training to ensure that the validation fold provided an accurate representation of how the model would perform on unseen data without introducing artificial variations from augmentations.

In addition to data augmentation, all images in the training and validation folds were resized to 299 × 299 pixels. This resizing step was crucial to match the input requirements of the three architectures of the three models, which were used later in the training pipeline.

2.4. Model Training

For the training phase, we selected three architectures to select the one with the best results. The Inception-V3, VGG19, and ResNet-50 architectures were selected as base models, as they offer a relatively small model size, which makes them computationally efficient while maintaining good performance in classification tasks. These architectures have been successfully applied in other studies related to ophthalmology, where they have demonstrated their ability to handle complex image data, including retinal images, with good results [11,15].

We opted for a 5-fold cross-validation approach primarily because of the size of our dataset. In small datasets, allocating sufficient data for training and validation can be challenging without losing valuable information for model learning. Cross-validation addresses this issue by allowing all data to be used for training and validation across different iterations, ensuring data were not wasted [16].

Each of the five folds was trained for 50 epochs. The model's final layer was modified to match our specific classification task, with 22 output classes corresponding to the different

genes in our study. We used the Adam optimizer with a learning rate of 0.0001 to optimize the model's weights during training.

These three architectures (Inception-v3, ResNet-50, and VGG19) were all trained under the same conditions, which include the same training and test data. The training process was also carried out using k-folds (five folds) during the same epochs (50 epochs). The parameters used, the metrics calculated, and the graphs drawn up were of the same type to maintain the same conditions for all three architectures. These models were trained for around 20 h each, using an NVIDIA A100-SXM4-40GB graphics card (NVIDIA Corporation, Santa Clara, CA, USA) on a server with the Ubuntu 22.04.3 LTS operating system.

2.5. Model Evaluation

As for the evaluation phase of our pipeline, we applied a model ensemble approach after the 5-fold cross-validation training process. Each fold produced a trained model, and by combining their outputs, we leveraged the strengths of all five models to predict the held-out test set. Model ensembling, in this context, helps reduce prediction variance and improves generalization, as the models are trained on slightly different subsets of the data. For each test sample, the final prediction was based on an aggregation (such as averaging) of the predictions made by all five models.

The evaluation of the ensemble model was performed using several important metrics. These included top-1 accuracy and top-5 accuracy, which are critical in gene classification tasks, as some genes might be harder to distinguish, and top-5 accuracy gives a more comprehensive measure of the model's predictive ability.

We also computed recall, precision, and the F1-score. Recall (sensitivity) measures how well the model identifies all relevant genes, while precision indicates how many of the model's positive predictions were correct. The F1-score, a harmonic mean of precision and recall, provides a balanced view of the model's performance, especially in cases where there is an imbalance between the number of positive and negative cases in the dataset. This metric proved particularly insightful given the potential imbalance in our gene classification task.

3. Results

The accuracy for training and validation was monitored for all architectures over the 50 epochs. It was possible to see that for all the models, the accuracy of the training and validation data increased over the epochs. Values close to 90% (VGG19, Inception-v3) and above 90% (ResNet50) were obtained for the average training accuracy. In the validation data, lower values were obtained, with accuracy remaining stable at around 80%. This clearly distinguishes between the two accuracy metrics for both datasets.

However, the curves for each of the five folders, in all architectures, are closely aligned (remaining close to 80%), and there are not many differences between the two, showing consistency in the model's performance in different validation sets.

The model was trained for 50 epochs, reaching convergence after approximately 30 epochs.

Although there was a degree of overfitting in the training, with the training accuracy curve consistently higher than the validation accuracies, this does not detract from the validity of the model's overall performance. The model's generalization ability is still evident, as demonstrated by the strong evaluation metrics obtained for the validation data.

After all the 5-fold models had finished training, we pooled the models for the final evaluation on the retained test set. The following metrics were recorded: accuracy, recall, precision, F1-score, AUC, and ROC curve. We also used integrated gradients (IGs) with the

model with the best metric values as an explainability tool to analyze correct and incorrect predictions. The metrics obtained for each model in the test dataset can be seen in Table 2.

Table 2. Results of the metrics obtained from the test data.

Metric	Inception-v3	ResNet-50	VGG19
Accuracy	0.54	0.56	0.54
Precision	0.58	0.55	0.54
Recall	0.54	0.56	0.54
F1-Score	0.53	0.54	0.52

The metrics obtained show that no model stood out much more than the others. On the contrary, both models had similar results, judging by the metrics. However, it was the ResNet-50 model that obtained slightly better results in three of the four metrics calculated (accuracy, recall, and F1-score), losing only in precision to the Inception-v3 architecture. It was the ResNet-50 architecture that obtained the best accuracy results, with 56%. The F1-score, which obtained better results in the ResNet-50 architecture, shows a moderate balance between precision and recall but also suggests that some gene classes were more difficult for the model to recover correctly. The VGG19 architecture had the lowest metric values, although with little difference.

Despite these differences, both architectures obtained very similar behaviors and results, proving the outcome of the metrics.

Table 3 shows, for each model used, the value obtained for each metric for each gene class. The highest values obtained by each model for the respective metrics for each class are highlighted to give a better understanding of the results obtained.

Table 3. Results for each class for the different models.

		Inception-v3				ResNet-50				VGG19				
	Gene	Prec *	Rec *	F1 *	AUC *	Prec	Rec	F1	AUC	Prec	Rec	F1	AUC	Sup *
Non-Syndromic	*EYS*	0.56	0.66	0.61	0.88	**0.64**	0.68	0.66	**0.91**	0.60	**0.75**	0.66	**0.91**	87
	RPGR	**0.77**	**0.64**	0.70	**0.88**	0.63	0.57	0.60	0.84	0.62	0.62	0.62	0.87	58
	RHO	**0.48**	0.31	0.38	**0.81**	0.46	**0.34**	0.39	0.69	0.33	0.28	0.31	0.72	32
	RPE65	**0.86**	**1.00**	0.93	**1.00**	0.71	**1.00**	0.83	**1.00**	0.63	0.96	0.76	**1.00**	25
	NON_S_USH2A	0.26	**0.73**	0.38	**0.93**	0.36	**0.73**	0.48	0.92	**0.37**	0.64	0.47	**0.93**	22
	CNGB1	0.45	**0.53**	0.49	**0.88**	**0.50**	0.29	0.37	0.86	0.21	0.18	0.19	0.85	17
	NR2E3	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	7
	PROM1	**0.75**	0.33	0.46	0.85	0.50	**0.44**	0.47	0.9	0.75	0.33	0.46	**0.95**	9
	PRPF31	0.00	0.00	0.00	0.23	0.00	0.00	0.00	**0.47**	0.00	0.00	0.00	0.42	8
	CRB1	0.25	**0.12**	0.17	**0.91**	0.14	**0.12**	0.13	0.76	**0.50**	**0.12**	0.20	0.83	8
	PCARE	0.00	0.00	0.00	0.97	**0.33**	**1.00**	0.50	**0.99**	0.14	0.50	0.22	0.98	4
	CEP290	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	6
	PDE6B	0.00	0.00	0.00	0.73	0.00	0.00	0.00	0.74	0.00	0.00	0.00	**0.82**	7
	IMPG2	**1.00**	0.14	0.25	0.9	0.00	0.00	0.00	**0.96**	0.50	**0.29**	0.36	0.92	7
	MERTK	0.00	0.00	0.00	0.91	0.00	0.00	0.00	**0.95**	0.00	0.00	0.00	0.84	7
	HGSNAT	0.67	**0.50**	0.57	0.91	0.00	0.00	0.00	**0.93**	**1.00**	0.25	0.40	0.8	4
Syndromic	*S_USH2A*	0.47	0.64	0.54	0.88	0.48	**0.76**	0.59	**0.91**	**0.59**	0.61	0.60	0.89	33
	MYO7A	**1.00**	0.44	0.62	0.94	0.90	**0.5**	0.64	**0.97**	0.70	0.39	0.50	0.93	18
	BBS1	**0.54**	0.39	0.45	0.86	0.40	**0.44**	0.42	0.79	**0.54**	0.39	0.45	**0.87**	18
	ARSG	0.53	**1.00**	0.69	**1.00**	0.90	**1.00**	0.95	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	9
	ADGRV1	0.50	0.08	0.13	0.83	**0.67**	**0.31**	0.42	0.81	0.25	0.15	0.19	**0.91**	13
	BBS10	**1.00**	0.12	0.22	0.85	**1.00**	**0.38**	0.55	**0.88**	**1.00**	**0.38**	0.55	0.74	8

Note: For each gene, the highest values of each metric for each model are highlighted in bold. * Legend: Prec—precision; Rec—recall; F1—F1-score; AUC—area under the curve; Sup—support.

Some gene classes obtained maximum values for the metrics, even though these classes were trained with a small amount of training data. On the other hand, some classes were trained with a larger amount of data, but the dog models achieved an accuracy of over 77%.

This table also shows the existence of classes that obtained values of 0% for all metrics. With these metric values, these classes were trained with a small amount of data.

By analyzing the confusion matrices in Figures 3–5, it is possible to draw better conclusions from these results through deeper analysis, which provides a clear view of how each gene was classified. In the Inception-v3 model, most of the true positives are correctly aligned along the diagonal, showing that the model performed well for many classes of genes. This suggests that genes with more training samples were classified accurately.

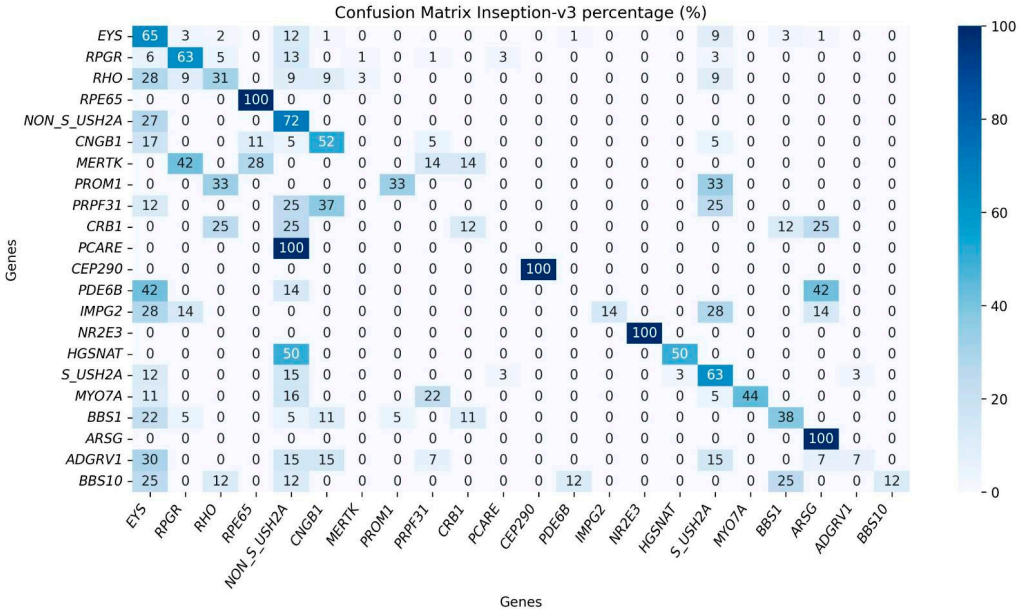


Figure 3. Confusion matrix of the Inception-v3 model on the test set in percentage (%).

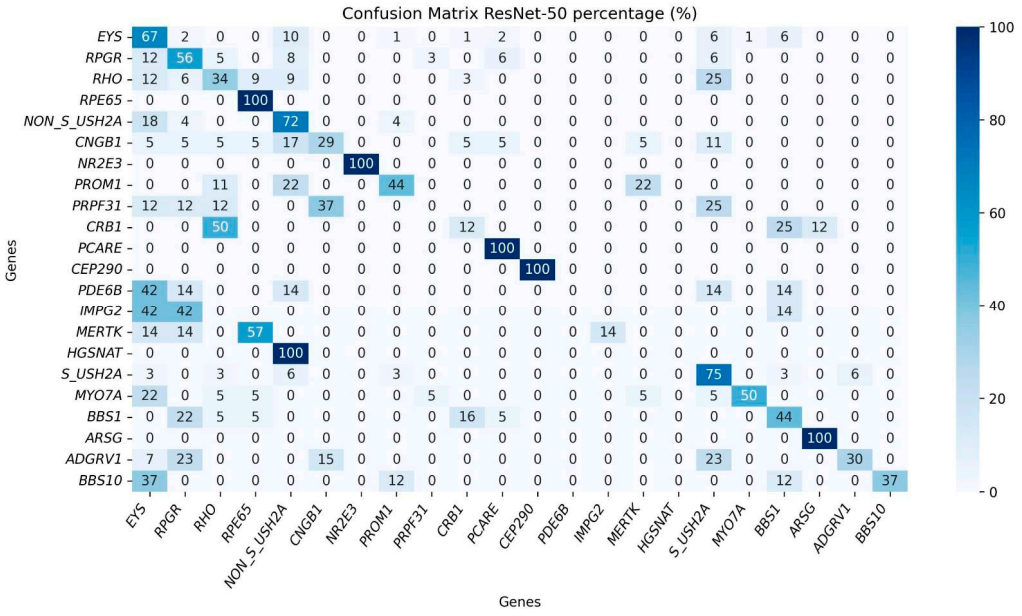


Figure 4. Confusion matrix of the ResNet-50 model on the test set in percentage (%).

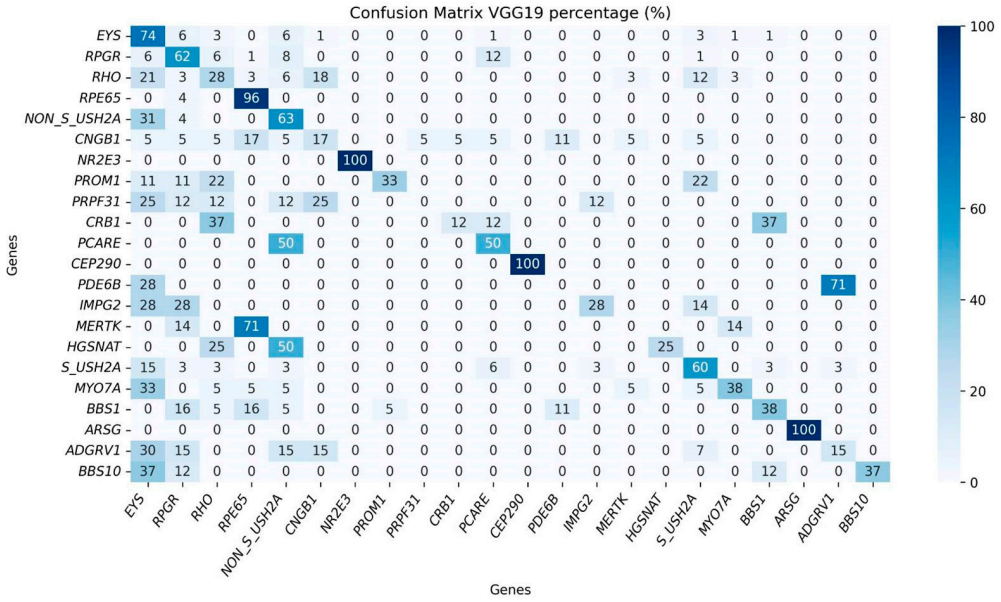


Figure 5. Confusion matrix of the VGG19 model on the test set in percentage (%).

Interpretation with Integrated Gradients (IGs)

Explainability and interpretability are crucial when developing machine learning models in healthcare, as clinicians and researchers require more than just predictive accuracy. To this end, we applied integrated gradients (IGs) to understand better how the model arrived at its predictions.

Two IG visualizations were generated: one for a correct prediction and another for a wrong prediction (Figures 6 and 7). Each of these examples compares the original image, the IG mask highlighting important features, and the overlap.



Figure 6. Integrated gradients visualization on a correct prediction.

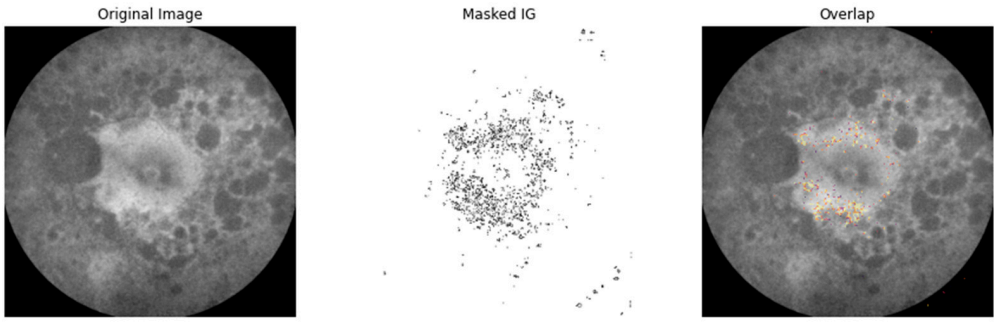


Figure 7. Integrated gradients visualization on a wrong prediction.

The IG overlay highlights the regions that the model used to decide. In the case of a correct prediction, retinal areas with pathological changes were highlighted, namely

around vascular arcades and outside the central macula. The IG visualization effectively demonstrates the model's ability to focus on the relevant retinal features associated with the target gene, giving confidence in the biological plausibility of the model's decision-making process.

However, when looking at a wrong prediction, the IG visualization still shows that the model valued areas of the retina preserved by the disease, i.e., areas without pathological changes, as seen in this example within the vascular arcades excluding the altered fovea. This suggests that the model did not use the same regions to make the predictions, although it is expected that the model will focus on potentially relevant areas of the image. This behavior may reflect the global representations learned, which did not allow the model to learn the areas it should pay attention to due to the small number of samples it had available to train. As a result, the model's difficulty in separating classes of similar genes increases. This insight further supports the need for an expanded dataset to represent underperforming gene classes better.

4. Discussion

From the results obtained in this study, it can be seen that quite satisfactory results were obtained for some classes of genes, not only with a large representation in the dataset but also some classes with a smaller representation. However, it was in the less represented classes where the worst results were obtained. Table 1 above shows the distribution of the data for each gene used in both training and testing the models. This is important to visualize because we are dealing with a very unbalanced dataset, with no fair balance between the representation of each class of genes, either in training or testing.

The model struggled with less represented genes, where classification errors are more evident. Classification performance was notably better for classes with more data, highlighting the challenge of working with unbalanced datasets and their effect on the model's overall accuracy.

In the ResNet-50 model, when we evaluated the test data, we obtained similar results to those obtained with the VGG19 model. The model learned to correctly identify some classes, obtaining 100% correct answers in the test data. However, it was not in the most representative classes of the training data (which can be verified by the number of test samples) that it obtained the best results. This could mean that the model learned more easily to identify some genes, even with a smaller training sample, than other genes with larger samples.

It can also be seen that the model's performance on some classes of genes was quite satisfactory, with an accuracy of over 60%. However, the low amount of data used in training influences the results, as in classes trained with less data, the accuracy is lower, reaching 0% accuracy in some gene classes. Thus, some genes may be more difficult to identify than others, making removing revealing features from the model more complicated.

This same behavior is also observed in the VGG19 architecture; this model could classify other genes, which was the previous model's low classification accuracy. However, the low amount of data in some training classes negatively influenced the accuracy of the hits in the test data, although in some classes with little data, the percentage of hits was 100%. This only reinforces the theory that some classes need more data for the model to extract features from the data and classify more accurately.

Another key indicator of the model's performance is the ROC curve for each of the 22 gene classes. The AUC obtained for each gene class (Table 3), most of which are close to, and even higher than, 0.9, reflects the model's ability to distinguish between most gene classes. This suggests that, for most genes, the model can reliably separate positive from negative cases. However, specific classes of genes have an AUC of less than 0.54,

revealing a significant difficulty in the model's ability to classify this class correctly. The low AUC suggests that the model could not effectively differentiate this gene from the others, probably due to a lack of distinctive features in the images or an insufficient number of representative samples.

Although these results are promising, they suggest that there is still a lot of room for improvement, not only in the methods used but also in balancing and increasing the data used to train the models. This solution is a proposal that has potential for future work, particularly in improving accuracy and, consequently, forecasting results. As such, it could become a useful tool in the future to help health professionals detect and identify the diagnosis and the gene. This will save time and resources as they will already have a prediction of the gene they will need to carry out traditional genetic tests on, avoiding the need to test them all.

The limitations found in this study influenced the results obtained from the accuracy of the model and its ability to generalize. This was reflected in the confusion matrices in which some classes of genes were very difficult to classify correctly. This limitation in this study was the small size of the dataset because retinitis pigmentosa is a rare disease, and there are not a lot of data available to train the models, and there are even less data when we try to classify into gene classes. Therefore, the small amount of data available for training and the large imbalance between gene classes was the greatest difficulty encountered in this study, which was reflected in the results obtained, which were lower than expected.

As for future work and in order to achieve better results, one approach that could be considered is the possibility of more partnerships in order to obtain more data and thus increase the dataset and try to increase the representativeness of the less represented classes, balancing it fairly. Another approach could be to use data augmentation techniques based on generative models to generate high-quality, high-fidelity synthetic data of the less represented classes. These methods could generate realistic images of less represented classes of genes and balance the dataset more fairly, increasing the representativeness of genes. As for the models, we used more recent architecture, such as transformer-based architectures, to see if they would help improve the results obtained.

5. Conclusions

This study highlights the significant progress in classifying retinal images associated with retinitis pigmentosa (RP) using deep learning techniques. Three models were used (Inception-v3, ResNet-50, and VGG19) to verify the most efficient approach for the problem in question. The results showed that the models performed similarly, with approximately 80% accuracy in training validation. However, the ResNet-50 model was the one that obtained slightly better results across all the metrics. These results underestimate the model's effectiveness in distinguishing between various genes linked to RP despite the challenges posed by the variability of the dataset and the representation of genes.

Integrating data preprocessing techniques, such as balancing and class augmentation, was crucial in improving the model's performance. In addition, the application of integrated gradients (IGs) provided valuable information on the model's decision-making process, validating its focus on relevant retinal features. This interpretability is crucial to ensure the reliability and biological relevance of the model's predictions.

The observed performance metrics indicate that while the model performs well in classifying genes with ample training data, there is room for improvement in handling less represented gene classes. The lower recall and F1-score for certain genes highlight the impact of data imbalance on the model's effectiveness. The ROC curve analysis, with a mean AUC of 0.91, further confirms the model's strong ability to differentiate between

most gene classes. However, one gene showed notably lower AUC, suggesting more representative samples were needed.

Analysis of the confusion matrices clearly showed that the model could correctly classify some of the less representative classes in the test data. Unlike other more representative classes, despite having more data to classify in the test, they were also trained with more data and obtained much lower percentages of correctness.

On the other hand, the less representative classes were also the ones with the worst results on the test data, with some of them obtaining 0% correct.

Thus, more data is needed for these results, but some of the classes will be more easily identified by the model, even with little representation in the training, than others.

Our future work will focus on expanding the dataset to include a broader range of genes and images, particularly targeting those with fewer samples. This expansion aims to improve the model's generalizability and address current limitations. Additionally, exploring more advanced deep learning architectures and further enhancing interpretability methods will be important in refining the model's accuracy and applicability.

By advancing artificial intelligence in retinal imaging, we strive to contribute to the early and accurate diagnosis of inherited retinal diseases. This progress holds promise for improving clinical practices and patient outcomes through more precise genetic analysis and personalized treatment approaches. Our continued commitment to research and collaboration with medical institutions will be essential in realizing the full potential of AI-driven tools in healthcare.

Author Contributions: Conceptualization, H.F., A.C. and A.M.; methodology, H.F. and J.M.; software, J.M.; validation, J.P.M., J.M.B. and I.C.; formal analysis, A.M.; investigation, H.F.; resources, A.M.; writing—original draft preparation, H.F.; writing—review and editing, J.M.; visualization, A.M.; supervision, A.C.; project administration, A.C.; funding acquisition, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Doctoral Scholarship of the Portuguese Ophthalmology Society, by Fundação para a Ciência e Tecnologia, IP (FCT) within the R&D Units Project Scope: UIDB/00319/2020 (ALGORITMI), and by FCT I.P. under the project Advanced Computing Project 2024.AIVLAB. 00008.VISION. It was produced with the support of Rede Nacional de Computação Avançada (RNCA).

Institutional Review Board Statement: The study complied with the tenets of the Declaration of Helsinki for biomedical research. Institutional Review Board (IRB)/Ethics Committee approval was obtained from the Departamento de Ensino, Formação e Investigação (DEFI) of Unidade Saúde Local de Santo António with reference number 2021.269(216-DEFI/224-CE), approved on 6 June 2022.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

IRDs	Inherited retinal diseases
RP	Retinitis pigmentosa
SD-OCT	Spectral-domain optical coherence tomography
FAF	Fundus autofluorescence
ML	Machine learning
XAI	Explainable AI

IG	Integrated gradients
AUC	Area under the ROC curve

References

1. Hansen, S.; McClements, M.E.; Corydon, T.J.; MacLaren, R.E. Future perspectives of prime editing for the treatment of inherited retinal diseases. *Cells* **2023**, *12*, 440. [CrossRef] [PubMed]
2. Verbakel, S.K.; van Huet, R.A.C.; Boon, C.J.F.; den Hollander, A.I.; Collin, R.W.J.; Klaver, C.C.W.; Hoyng, C.B.; Roepman, R.; Klevering, B.J. Non-syndromic retinitis pigmentosa. *Prog. Retin. Eye Res.* **2018**, *66*, 157–186. [CrossRef] [PubMed]
3. von Krusenstiern, L.; Liu, J.; Liao, E.; Gow, J.A.; Chen, G.; Ong, T.; Lotery, A.J.; Jalil, A.; Lam, B.L.; MacLaren, R.E. Changes in retinal sensitivity associated with cotoretigene toliparvovec in X-linked retinitis pigmentosa with RPGR gene variations. *JAMA Ophthalmol.* **2023**, *141*, 275–283. [CrossRef] [PubMed]
4. Cortinhal, T.; Santos, C.; Vaz-Pereira, S.; Marta, A.; Duarte, L.; Miranda, V.; Costa, J.; Sousa, A.B.; Peter, V.G.; Kaminska, K.; et al. Genetic profile of syndromic retinitis pigmentosa in Portugal. *Graefes. Arch. Clin. Exp. Ophthalmol.* **2024**, *262*, 1883–1897. [CrossRef] [PubMed]
5. Heckenlively, J.R.; Yoser, S.L.; Friedman, L.H.; Oversier, J.J. Clinical findings and common symptoms in retinitis pigmentosa. *Am. J. Ophthalmol.* **1988**, *105*, 504–511. [CrossRef] [PubMed]
6. Ferrari, S.; Di Iorio, E.; Barbaro, V.; Ponzin, D.; S Sorrentino, F.; Parmeggiani, F. Retinitis pigmentosa: Genes and disease mechanisms. *Curr. Genom.* **2011**, *12*, 238–249. [CrossRef]
7. Daiger, S.P.; Sullivan, L.S.; Bowne, S.J. Genes and mutations causing retinitis pigmentosa. *Clin. Genet.* **2013**, *84*, 132–141. [CrossRef] [PubMed]
8. Popović, P.; Jarc-Vidmar, M.; Hawlina, M. Abnormal fundus autofluorescence in relation to retinal function in patients with retinitis pigmentosa. *Graefe's Arch. Clin. Exp. Ophthalmol.* **2005**, *243*, 1018–1027. [CrossRef] [PubMed]
9. Oishi, A.; Ogino, K.; Makiyama, Y.; Nakagawa, S.; Kurimoto, M.; Yoshimura, N. Wide-field fundus autofluorescence imaging of retinitis pigmentosa. *Ophthalmology* **2013**, *120*, 1827–1834. [CrossRef] [PubMed]
10. Brancati, N.; Frucci, M.; Gragnaniello, D.; Riccio, D.; Di Iorio, V.; Di Perna, L.; Simonelli, F. Learning-based approach to segment pigment signs in fundus images for retinitis pigmentosa analysis. *Neurocomputing* **2018**, *308*, 159–171. [CrossRef]
11. Pontikos, N.; Woof, W.; Krawitz, P.; Arno, G.; Hess, K.; Varela, M.D.; Liefers, B.; Guimaraes, T.; Shah, M.; Madhusudhan, S.; et al. Eye2Gene: Prediction of causal inherited retinal disease gene from multimodal imaging using AI. *Investig. Ophthalmol. Vis. Sci.* **2022**, *63*, 1161. [CrossRef]
12. Liesenfeld, B.; Kohner, E.; Piehlmeier, W.; Kluthe, S.; Aldington, S.; Porta, M.; Bek, T.; Obermaier, M.; Mayer, H.; Mann, G.; et al. A telemedical approach to the screening of diabetic retinopathy: Digital fundus photography. *Diabetes Care* **2000**, *23*, 345–348. [CrossRef] [PubMed]
13. Nagiel, A.; Lalane, R.A.; Sadda, S.R.; Schwartz, S.D. Ultra-widefield fundus imaging: A review of clinical applications and future trends. *Retina* **2016**, *36*, 660–678. [CrossRef] [PubMed]
14. Kapishnikov, A.; Venugopalan, S.; Avci, B.; Wedin, B.; Terry, M.; Bolukbasi, T. Guided integrated gradients: An adaptive path method for removing noise. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
15. Mujahid, M.; Rustam, F.; Álvarez, R.; Luis Vidal Mazón, J.; Díez, I.D.L.T.; Ashraf, I. Pneumonia classification from X-ray images with inception-V3 and convolutional neural network. *Diagnostics* **2022**, *12*, 1280. [CrossRef] [PubMed]
16. Raschka, S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv* **2018**, arXiv:1811.1280. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.